OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Provided by the author(s) and University of Galway in accordance with publisher policies. Please cite the published version when available.

| | |
|---|---|
| Title | Pathogen genomics of methicillin resistant staphylococcus aureus and leishmania |
| Author(s) | Coughlan, Simone |
| Publication Date | 2017-02-15 |
| Item record | http://hdl.handle.net/10379/6315 |

Downloaded 2024-05-05T11:17:06Z

Some rights reserved. For more information, please see the item record link above.

# Pathogen genomics of Methicillin resistant *Staphylococcus aureus* and *Leishmania*

Simone Coughlan

A thesis submitted to the

School of Mathematics, Statistics and Applied Mathematics,
National University of Ireland, Galway

In partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Supervised by

Dr Tim Downing

&

Professor Cathal Seoighe

November 2016

# Abstract

Infectious diseases caused by the single-celled eukaryotic parasite *Leishmania* and the methicillin-resistant *Staphylococcus aureus* (MRSA) bacterium are major public health problems in many countries. In this thesis, I use genomics to explore the genomic plasticity of *Leishmania* and characterise the genomic and transcriptomic responses of MRSA treated with an antibiotic called oxacillin.

The *Leishmania* parasite is transmitted by sandflies and can be maintained in the wild by various animals, as well as in people. It causes leishmaniasis, which is often difficult to treat and can prove fatal. In order to understand the *Leishmania* spp. infecting wild animals and their relationships to human-infecting *Leishmania*, we assembled, annotated and analysed the genomes of three *Leishmania* spp. The first of these was from a rodent in Ethiopia which we identified as *Leishmania adleri* using a phylogenomic approach. This species is part of the *Sauroleishmania* subgenera, whose genomes are expected to have 36 chromosomes and can infect reptiles. We found evidence of two novel independent chromosomal fission events in *L. adleri* using both our genome and an unassembled *L. adleri* sample isolated from a lizard. This resulted in 38 chromosomes, which was a novel finding because there was no evidence of these fissions in the sole published genome from the same subgenus: *L. tarentolae* Parrot-TarII. Extensive gene amplifications and aneuploidy were discovered in all three *Sauroleishmania* samples analysed, in common with previous work on other *Leishmania* spp., highlighting the lack of differentiation between animal- and human-infecting species. This new *L. adleri* genome is a high-quality annotated draft suitable for use as a reference, is the first assembled sequence available for *L. adleri*, and is only the second species in the *Sauroleishmania* subgenus to have a published genome.

The other two *Leishmania* samples were isolated from dogs with leishmaniasis in Colombia and these were assembled and analysed using the same approach. A control genome was assembled using reads from the *L. braziliensis* genome so that we could quantify the completeness of our assemblies and identify any problems caused by our assembly approach. We classified our samples as *L. naiffi* and *L. guyanensis*, both members of the subgenus *Viannia*, whose members are only found in the Americas, predominately South America. This is the first report of *L. naiffi* in Colombia and in dogs illustrating the usefulness of genomics in disease surveillance. These genomes are also the first genomes for these two species. We compared both genomes with multiple other species from this subgenus and identified a 45 kb amplification in many *Viannia* spp. as well as a minichromosome in *L. shawi* M8408. Genes with high copy number and those unique to both species and the *Viannia* subgenus as a whole were also documented, which will aid development of diagnostics for this subgenus.

Multiple responses to drug treatment with oxacillin have been investigated in many MRSA lineages. In this thesis, colleagues and I examined the genomic and transcriptomic responses of a community acquired MRSA strain (USA300) in a continuous culture (chemostat) experiment as well as in growth on agar plates. MRSA can exhibit heterogeneous resistance (HeR) which occurs when most cells in a sample are susceptible to low levels of antibiotic and only a few cells are highly resistant. A highly homogenously resistant (HoR) can be selected from a HeR sample using high doses of oxacillin. We discovered a novel tandem amplification of *SCCmec*IV in a drug resistant sample taken from a chemostat experiment. *SCCmec*IV is a mobile genetic element that harbours the *mecA g*ene which facilitates resistance to β-lactam antibiotics, such as oxacillin.  Multiple SNPs and indels at genes previously implicated in resistance were also identified. HeR isolates treated with oxacillin had low-frequency SNPs at some genes as well as numerous differentially expressed genes, whereas HoR samples had a nonsynonymous SNP at the *gdpP* gene, but few differentially expressed genes. This demonstrated that HeR cell populations responded to oxacillin by modifying gene expression regulation, whereas HoR ones had a genetic mutation to become resistant. We also found that purine metabolism had a role in oxacillin stress response because it was highly down-regulated at all levels of oxacillin, and SNPs and indels were discovered at two genes in this pathway (*apt* and *guaA*).

Overall, we have assembled the genomes of three *Leishmania* spp., discovered novel chromosomal fission events in *L. adleri* and documented the presence of *L. naiffi* in a dog in Colombia for the first time. These genomes, coupled with that of *L. guyanensis* have extended our understanding of genome architecture and plasticity in *Leishmania* and will facilitate future research by others on these species. We have found a novel amplification of SCC*mec*IV in response to drug treatment demonstrating the need to search for copy number variation in addition to SNPs and indels, and found multiple responses to various levels of oxacillin, some of which had not been previously reported. These findings have important clinical implications for drug treatment of *S. aureus* as they demonstrate that amplification of large mobile elements can occur and that these can be maintained on the chromosome with variable copy number in response to drug pressure. Furthermore, commonly mutated genes and pathways in resistant samples show that cells converge on common solutions to survive drug treatment and these genes/pathways could serve as drug targets.

# Acknowledgements

Firstly, I would like to sincerely thank my supervisors Dr Tim Downing and Prof. Cathal Seoighe. I am so lucky to have been mentored by such knowledgeable, supportive and nice people and you have made doing this PhD such a rewarding experience. Particular thanks to Tim, who despite a move to Dublin, still made time every week to discuss PhD work and welcomed me on visits to his Dublin lab, as well as always being generous with advice and feedback. Thanks to Cathal for prompt feedback on any work I presented or gave to him and whose door was always open in Galway for any help or questions that I had, and for his regular reminders of that! I am truly grateful to both of you.

Thanks to all the lab members in both Galway and Dublin who have welcomed me and made this such an enjoyable and memorable experience. In Dublin, thank you to Arun and Ray for always welcoming me on every visit and all the games of table tennis! Thank you for all the food Arun and both of you for all the hilarious conversions. In Galway, I have so many people I need to thank. Thank you to all the bioinformatics and mathematics PhD students both past and present for your friendship and support! I particularly enjoyed our varied lunch time chats, mathematics society events, nights out and your patient and often humorous attempts to explain your mathematics research to me and others. I am also pleased that I now know how to solve a wide variety of puzzles! I have to thank Shane, Ronan and John from Maths for many evenings in your house as well as Shane in particular for joining me on many walks/swims in various parts of Galway and for introducing me to so many people (and the person who would become my fiance). In bioinformatics, thank you in particular to Peter for both bioinformatics advice and your friendship. Thank to my friends at home in Clare/Limerick for your constant support and friendship: in particular Deirdre, Yvonne, Rachel, Mary and Danielle, as well as Lynne (and Brendan) who always encouraged me to visit them in Galway and made those visits such fun. Thanks to Siobhan who always kept in touch even though you lived abroad (it was always nice to chat about bioinformatics too) and to Joey for meeting up with me in Dublin when I was around there.

Thank you to my mum and dad who have supported me in anything I have decided to do, for visiting me in Galway and always giving me a warm welcome home. Thanks to my brothers Patrick and Colm who always keep in touch. Thanks a million to my sister Claudia who gave me so much advice and encouragement and who has always been there for me. To my fiancé Allen, thank you for your unwavering love and support. You have been such a rock and always believed in me and supported me. I hope I can do the same for you when you need it. Thank you for all your visits to Galway, dragging me hiking and for always being able to make me laugh. I need to thank Connie and Sean, my future parents in-law for being such amazing people and letting me stay in their apartment in the final months of my PhD. I'm so lucky to have such kind and easy-going in-laws and I really appreciate what you have done for me and love meeting up with you. Also thank you to Allen's sister, Yvonne, and her fiance Tony for living with me now. Thanks you so much Yvonne for all the dinners you insist on making me (not that I put up much of a fight!) as well as being such a great friend.

# Declaration

I certify that this thesis is all my own work and that I have not obtained a degree in the National University of Ireland, Galway or elsewhere on the basis of any of this work. I have acknowledged and made clear any assistance or contributions and cited the published work of others where applicable.

Signed: ______________________

Date:_____________________

# Glossary

ABACAS: Algorithm Based Automatic Contiguation of Assembled Sequences

ACME: Arginine Catabolic Mobile Element

ACT: Artemis Comparison Tool

BAM: Binary Sequence Alignment/Map

BQ: Base Quality

BWT: Burrows Wheeler Transform

CA-MRSA: Community Associated Methicillin Resistant *Staphylococcus aureus*

CC: Clonal complex

CIA: Coinertia Analysis

CL: Cutaneous Leishmaniasis

CNV: Copy Number Variant

COG: Cluster of Orthologous Groups

CTn: Conjugative Transposon

DCL: Diffuse Cutaneous Leishmaniasis

DR: Direct Repeat

EFB: Error Free Bases

eQTL: Expression Quantitative Trait Loci

FCD: Fragment Coverage Distribution

FDR: False Discovery Rate

GO: Gene Ontology

HA-MRSA: Healthcare Associated Methicillin Resistant *Staphylococcus aureus*

HeR: Heterotypic Resistance

HGT: Horizontal Gene Transfer

HGT: Horizontal Gene Transfer

HoR: Homotypic Resistance

ICE: Integrative and Conjugative Element

iCORN: iterative Correction Of Reference Nucleotides

IGV: Integrative Genomics Viewer

IGV: Integrative Genomics Viewer

Indel: Insertion or deletion

IR: Inverted Repeat

IS: Insertion Sequence

KDNA: Kinetoplast DNA

KEGG: Kyoto Enclyopedia of Genes and Genomes

KO: Kyoto Enclyopedia of Genes and Genomes Ontology

LFC: Log Fold Change

$\log_2$FC: $\log_2$ fold-change

MCL: Mucocutaneous Leishmaniasis

MDK: Minimum Duration of Killing

MGE: Mobile Genetic Element

MIC: Minimum Inhibitory Concentration

MLST: Multilocus Sequence Typing

MQ: Mapping Quality

NGS: Next Generation Sequencing

OG: Orthologous Group

OLC: Overlap Layout Consensus

ORF: Open Reading Frame

PCR: Polymerase Chain Reaction

PFGE: Pulsed Field Gel Electrophoresis

PKDL: Post-Kala-azar Dermal Leishmaniasis

PTU: Polycistronic Transcription Unit

PVL: Panton-Valentine Leukocidin

QQ-plot: Quantile-Quantile plot

RATT: Rapid Annotation Transfer Tool

RDAF: Read Depth Allele Frequency

REAPR: Recognition of Errors in Assembly using Paired Reads

REAPR: Recognition of Errors in Assembly using Paired Reads

RNAi: RNA interference

RNAPII: RNA Polymerase II

rRNA: ribosomal RNA

RS: Repeat Sequence

SAM: Sequence Alignement/Map

SaPI: Staphylococcal Pathogenicity Island

SCC: Staphylococcal chromosomal cassette

SIDER: Short Interspersed Degenerate Retroposon

SL: Spliced Leader

SLACS: Spliced Leader Associated Conserved Sequence

snoRNA: small nucleolar RNA

SNP: Single Nucleotide Polymorphism

snRNA: small nuclear RNA

SQ: SNP Quality

SSPACE: SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension

SSR: Strand Switch Region

ST: Sequence Type

SV: Structural Variant

TATE: Telomere Associated Transposable Element

tRNA: transfer RNA

VL: Visceral Leishmaniasis

VSG: Variant Surface Glycoprotein

# Contents

# List of Figures

# List of Tables

# Chapter 1- Introduction

## 1.1 Pathogen genomics

A pathogen is anything that can cause a disease and the term is usually applied to infectious agents such as bacteria, single-celled eukaryotes, viruses and fungi. The search for drugs to prevent or treat infections caused by these microorganisms in humans has been a preoccupation for millennia, with hymns from an ancient Sanskrit text dated ~2,100 BC describing curing ‘invisible worms’ using plants [1,2] and it is vital today, especially in light of the growing threat of antibiotic resistance in bacteria which is reaching crisis point [3].

Pathogen genomics or pathogenomics uses data generated from high-throughput sequencing technologies such as microarrays and whole-genome sequencing to enhance our understanding of these organisms. It can be used to unravel the relationship between genotype and phenotype such as the ability of an organism to resist drug treatment, to discover virulence factors that enable it to invade and colonise a host, to study spread and transmission of an organism at local and global scales (disease surveillance), to identify novel species or strains of an organism, to track disease outbreaks [4,5] and to study the relationship between organisms.

In this chapter, I first review how next generation sequencing data can be used for genome assembly, structural variant detection and transcriptome analysis, which yield insights into pathogen genomes. I then provide an introduction to *Leishmania* and methicillin-resistant *Staphylococcus aureus* (MRSA) and their genomes.

### 1.1.1 Next generation sequencing

The first widely used method of DNA sequencing, termed Sanger sequencing or chain-termination sequencing was developed in the 1970s by Fredrick Sanger and Alan Coulsen. This was based on the incorporation of chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during replication of a template fragment of single-stranded DNA [6,7]. This technique was used in the  first DNA genome project which was completed in 1977 by Fredrick Sanger who sequenced the genome of the 5,368 base pair (bp) Phage Φ-X174 bacteriophage [8].  The first bacterial genome, of *Haemophilus influenza*, was completed in 1995 at the Institute for Genomic Research [9]. This project used a whole genome shotgun approach (DNA randomly fragmented into smaller parts) and this approach also used in the sequencing of the draft 3.3 Gb human genome [10,11] published in 2001. However, the three billion dollars and fifteen years needed to sequence the human genome was impractical for most laboratories and the demand for lower cost and less resource intensive methods lead to the development of next generation sequencing technologies (NGS) in the late 1990s.

NGS has revolutionised the study of pathogens by enabling faster and cheaper generation of sequence data and resulted in a plethora of new bioinformatics algorithms and tools for its analysis. A variety of NGS platforms are now available each with their differences in terms of read length, error rate, cost and output. NGS platforms such as those produced by Illumina are based on the *in-situ* sequencing by synthesis principle whereby primed templates are extended in sequential cycles by a DNA polymerase [12,13] or ligase enzyme [13,14]. Illumina produces a variety of sequencers that use this chemistry including the Illumina HiSeq, MiSeq, MiniSeq and NextSeq systems. The sequencers used to produce the data in this thesis are outlined in Table 1.1.

A typical sequencing by synthesis pipeline starts with random fragmentation of DNA, followed by the ligation of adaptors to the 5' and 3' ends. In the case of RNASeq, messenger RNA (mRNA) is selected from total RNA using either ribosomal RNA depletion in the case of bacterial DNA (as their mRNAs do not have polyA tails) or polyA enrichment [15]. The mRNA is reverse transcribed into double-stranded complementary DNA (cDNA) which is fragmented and adaptors ligated as for genomic DNA. These fragments are amplified by PCR and purified to form the sequencing library.

Multiple samples can be pooled and sequenced in the same lane of a flowcell in a single Illumina run (multiplexing) by adding index sequences to each fragment in the library preparation step. After sequencing, the indexed sequences can be identified and separated by sample into separate files (demultiplexed) using an algorithm, which increases the amount of samples that can be sequenced in one lane to 96.

Once prepared for Illumina sequencing, the sequencing library is loaded onto the flow cell where the fragments are immobilised via annealing of the adaptors complementary to oligonucleotides that are attached to the surface of the flowcell. Each fragment is amplified into clonal clusters using multiple rounds of bridge PCR amplification and so multiple large clusters are produced, each containing approximately a thousand copies of the original fragment. The reverse strands are cleaved from the now double-stranded DNA, the free 3' ends of fragments are blocked to prevent unwanted priming and a sequencing primer is hybridised. Clusters are sequenced in parallel over multiple cycles (cyclical sequencing). In each cycle, a single fluorescently labelled deoxyribonucleotide triphosphate (dNTP) is incorporated into the DNA template in a reaction catalysed by DNA polymerase. Each fluorescently labelled dNTP is bound with a reversible terminator sequence that ensures that only one nucleotide is incorporated per cycle and all four dNTPs are present in each cycle, creating a competition between the bases to be incorporated, minimising incorporation bias. The flurophore is cleaved after incorporation which allows the next nucleotide to be added in the following cycle. At the end of each cycle, the incorporated nucleotide is identified based on the fluorescence wavelength and intensity emitted from the cluster (base calling). The cycle is repeated 'N' times to create reads of length 'N'.

Sequencing both the 5' and 3' ends of the same fragment produces paired-end reads. One set of paired-end reads constitutes a forward and reverse read that are generally oriented towards each other. Usually the fragment length will be larger than the sum of the length of the forward and reverse reads and so there will be a gap between them. The length of the two reads and the gap is often called the insert size. Reads can also overlap each other such that they have a shared segment. DNA fragment selection by enzymatic digestion library preparation can produce a mixture of overlapping and non-overlapping paired-end reads. Thus, some fragments will be smaller than the sum of the paired-end read length e.g. if the fragment is 500 bp but the read length is 300 bp, the paired-end reads produced will overlap by 100 bp.

Nanopore sequencing, which can be undertaken using the handheld sized MinION device from Oxford Nanopore, uses a protein nanopore embedded in a synthetic polymer based membrane. A processive enzyme is bound to DNA and its step-wise movement at the pore opening controls the movement of the DNA strand through the nanopore so that only one base passes through at a time. A potential applied across the membrane results in a current flowing through the aperture of the nanopore and single molecules entering the nanopore cause disruptions to the current. These disruptions are associated with five-nucleotide DNA *k*-mers (DNA words of length *k*) [16,17]. Both strands of the DNA can be read (2D sequencing), up to 512 DNA molecules can be read at a time, and PCR amplification is not required because only a single molecule is sequenced. Much longer read lengths (range from 6,000 to 48,000 bp) are achievable although the error rate is higher, at approximately 38% (~ 1/3 bases are incorrect), compared with Illumina reads which have a 0.1% error rate (1/1000 bases are incorrect) [16,18]. The high error rate means reads are better suited to scaffolding and gap closing genomes that were originally assembled from shorter or less error prone read data, as well as solving repeat regions [19]. MinION sequencing of a DNA sample was undertaken in Chapter 4 to produce long reads in order to verify the position and orientation of a large amplification in a *S.aureus* genome (Table 1.1).

| Platform | Read types | Usage |
|---|---|---|
| Illumina HiSeq | 75 bp and 100 bp paired-end | Chapters 2 and 3: Assemble of three *Leishmania* genomes and comparison with other *Leishmania* genomes |
| Illumina MiSeq | 300 bp paired-end | Chapter 4: Differential gene expression (RNASeq), genome assembly, mapping and variant calling (DNASeq) of *S. aureus* USA300 genomes |
| MinION | 10 kb single-end | Chapter 4: Verification of tandem amplification architecture and location by assembly of reads (carried out by Mick Watson and collegues at the University of Edinburgh) |

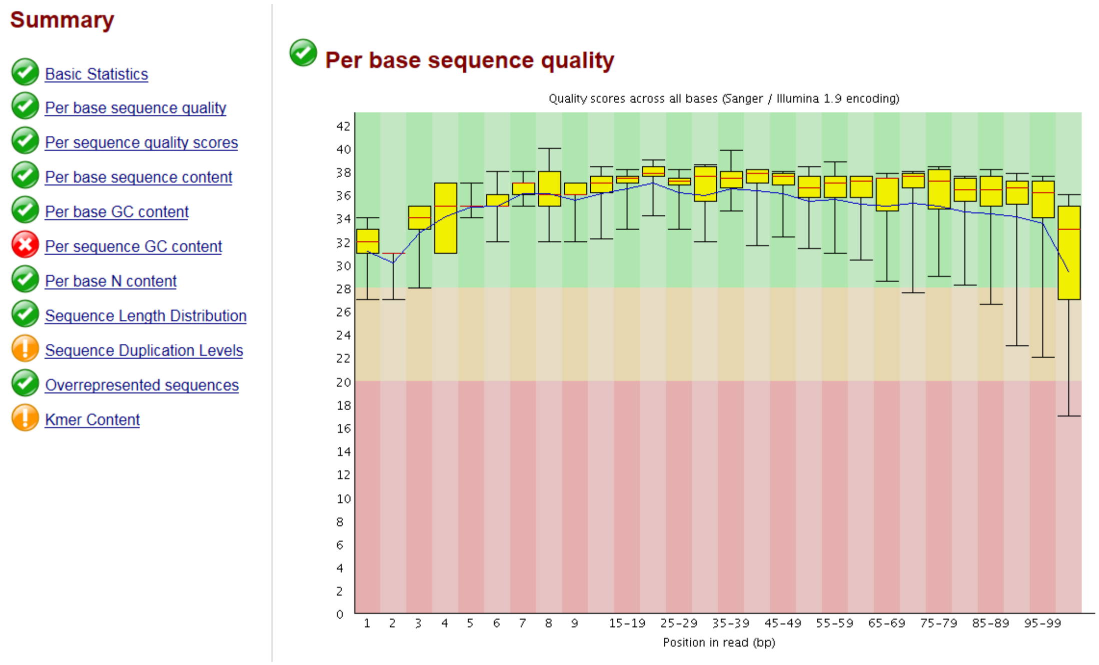
**Table 1.1:** Summary of sequencing instruments used to produce the sequence reads used in this thesis and chapters that the data was used in.

### 1.1.2 Quality Control of NGS data

The output of an NGS sequencing run is millions of nucleotide sequences (reads) as well as the associated base qualities of these reads which are represented by ASCII characters in FASTQ format files. The most common Sanger format has the ASCII codes shifted by +33. The base qualities reflect the confidence that the base has been correctly identified and these are commonly converted to Phred quality score (*Q*): this computes the probability *P*, that a base call is incorrect using the formula $Q = -10\log_{10}P$. A Phred quality score of 10 at a base means that base has a 1 in 10 chance of being incorrect whereas a Phred score of 30 indicates the base has a 1 in 1000 chance of being incorrect [20]. Illumina reads can demonstrate a drop in base quality towards the 3’ end of reads which can be visualised using a boxplot of the Phred quality values at each base in the reads using software such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This drop in quality scores occurs due to the incorporation of more than one nucleotide in a cycle (pre-phasing) or failure to incorporate a nucleotide in a cycle (off-phasing) causing loss of synchrony in the readout of molecules in a cluster [21]. The amount of sequences in a cluster affected by these issues increases with cycle number, causing noisy signals which result in the drop in base quality [22]. These low quality bases can be trimmed off prior to mapping using software such as Trimmomatic [23] or the FASTX quality trimmer (http://hannonlab.cshl.edu/fastx_toolkit/), which can improve the accuracy of downstream analysis such as single nucleotide polymorphism (SNP) calling [24]. Erroneous parts of reads can be corrected without trimming by superimposing reads’ *k*-mers on each other and correcting low frequency *k*-mers using the most common *k*-mers e.g. using tools such as Quake [25]. However, this approach typically needs uniform coverage to work and so is not suited to transcriptomic, single-cell or metagenomics datasets or those with low coverage (< 15-fold for Quake) although tools such as BayesHammer [26] can be used for datasets with non-uniform coverage [24]. Indeed, BayesHammer is coupled with the SPAdes assembler, where it error corrects reads prior to assembly, improving the resulting assemblies [27].

Some library preparation steps and platforms require specific optimisation. Enzymatic size selection during library preparation produces a range of fragment sizes such that extremely small reads could be longer than the fragment. This means the adaptor is sequenced along with the main portion of the read, leaving adaptor sequence in those reads. Automatic removal or masking of these adaptor sequences is generally performed and results in reads with either different lengths or many ‘N’ bases at the 5’ end of the read. However, it is prudent to check for adaptor sequence, which can removed using tools such as Trimmomatic [23]. Overlapping paired-end reads must also be screened and can be either merged into single-end reads based on the overlapping regions using tools such as FLASH [28], or considered later depending on the goal.

Another issue that can occur is contamination of sequencing runs with DNA from organisms outside the target organism. This can be identified using the GC content distribution in tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Figure 1.1). A GC content distribution that deviates from a Normal distribution indicates contamination or the presence of a symbiont. Contamination can be removed by either removing reads with abnormal GC content and/or aligning reads to a database using tools such as BLAST [29] to identify the species, with subsequent removal of reads that have hits outside the target species. An example of FASTQC reports produced using reads before and after contamination removal can be seen in Appendix B, Figure B1. Another approach is to perform a preliminary assembly of the reads and align the resulting contigs to database to identify non-target species. These approaches are incorporated in the software Blobtools [30] which produces taxon-annotated GC-coverage plots (TAGC) to help identify contaminating contigs based on GC content, coverage and taxonomic information. Reads mapping to these contigs can be removed and the target species assembled separately from contaminating or symbiont species.



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content

✓ Per base sequence quality

**Figure 1.1:** Screenshot showing part of a HTML report produced by FASTQC of *L. guyanensis* CL085 forward sequence reads. Green icons indicate modules with a pass status, yellow icons are warnings that indicate that a quality control issue may be present and red icons represent failures. Here the ‘Per Sequence GC content’ module had a fail status and this was corrected by removing contaminating reads in a method outlined in Chapter 3. The ‘Per base sequence quality’ plot here shows the distributions of phred quality scores at each base in the read.

### 1.1.3 Alignment and Assembly

After QC has been performed on the read data it can either be mapped to a reference genome if one is available for the species/strain, or assembled. Mapping reads facilitates the identification of structural variants as well as comparison of the results across multiple samples and is often a first step in measuring gene expression in RNASeq if a reference genome is available. However, sequences that are not present in the reference genome will not map and be lost to further analysis. If no reference genome is available, reads can be *de novo* assembled and that assembly used for further analysis and mapping. This has advantages: sequences unique to the organism will be retained unless they are repetitive and short. Such sequences cause assemblers problems in placing them on the genome which can result in either their omission or assembly as short contigs.

### 1.1.4 Genome assembly

*De novo* genome assembly is generally achieved using approaches based on overlap-layout-consensus (OLC) or De Bruijn graphs. OLC constructs a graph consisting of nodes representing the reads, and edges representing overlaps between reads. It does this by aligning reads to each other and identifying overlaps via a ‘seed-and-extend’ approach. Non-intersecting paths in the graph identify contigs and a list of such contigs will be produced, which will have gaps with both unknown sequence and length between them. OLC is used in assemblers such as Celera [31] (Table 1.2). It is typically used for assembling longer reads such as those produced Sanger sequencing as they have the long stretches of homologous sequence needed to produce unique overlaps [32,33].

For the shorter read lengths, De Bruijn graphs are more effective. Assemblers using this approach include Velvet [34], ALLPATHS [35], ABySS [36] and SPAdes [27] (Table 1.2). These use subsections of the reads of size *k* (*k*-mers) for assembly rather than whole reads. The *k*-mers values supplied to the assembler must be an odd value to avoid palindromic sequence and also shorter than the read length. The use of *k*-mers is computationally less intensive than the OLC method of aligning reads to each other. Instead, *k*-mers are summarised using a data structure called a hash table which indexes data in a non-hierarchical manner that facilitates fast computational searching. The De Bruijn graph has a node for every *k*-mer, and an edge between two nodes if two *k*-mers are adjacent in a read (Figure 1.2). Thus, edges between these nodes have only a single base difference (a *k*-1 difference). This approach results in reads being split across multiple nodes (as each *k*-mer has its own node) [32,37]. Sequence errors can create ‘bulges’ in the graph which must be corrected to produce (if present) exact overlaps between *k*-mers. Thus, error-correcting and trimming reads can help prevent this.

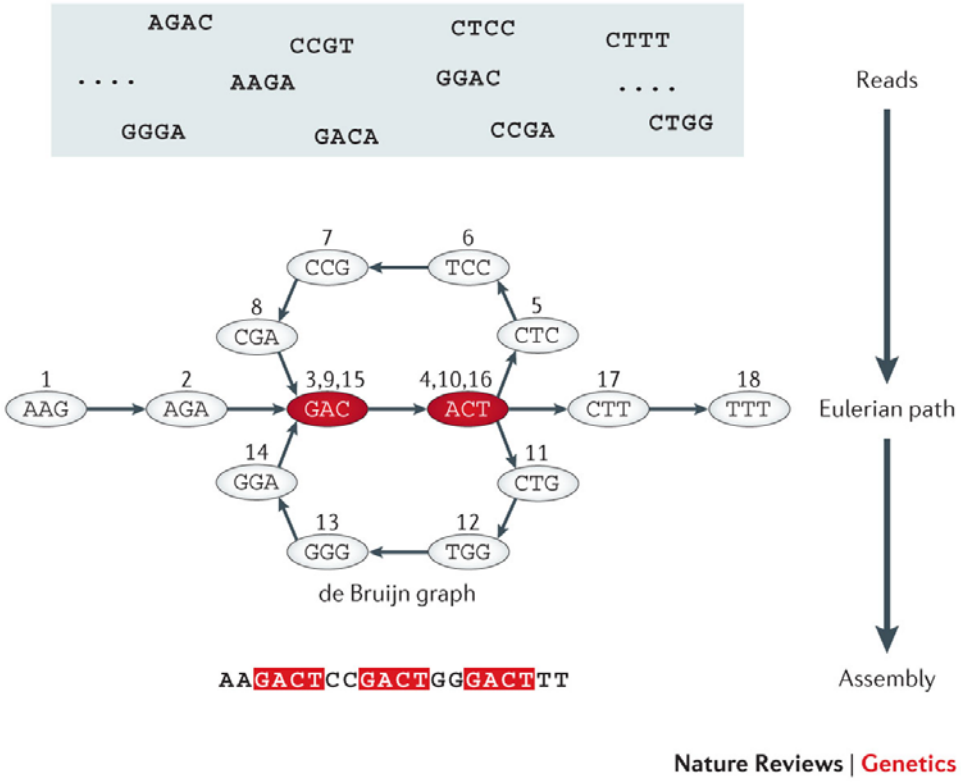


**Figure 1.2:** De Bruijn graph based assembly. "Each directed edge in a de Bruijn graph denotes a sequence read or a fragment of fixed length (4 bp in the figure); the source node of this edge is a prefix string of the read omitting the last nucleotide; the destination node of this edge is a suffix string of the same read (or sequence fragment) by omitting the first nucleotide. In the example shown in this figure, the top panel is a pool of representative short reads or fragments. In the middle panel, each node denotes a unique sequence prefix or suffix segment of length 3 bp found in the original reads of length 4 bp. The assembly of DNA sequences (segments) is thus represented as a de Bruijn graph. Assembling reads (or sequence fragments) in a de Bruijn graph reduces the problem to a fragment assembly problem that can be formulated as the goal to find a trail or Eularian path that visits each edge (read or fragment) in the (de Bruijn) graph exactly once. Nucleotides with a red background occur more than once in the sequence. Numbers on the edges represent an ordered Eulerian path through the de Bruijn graph, which can be followed to reconstruct the assembled sequence from the compact graph representation." This figure is reproduced from [38] with permission.

The use of mate-pair (larger insert size reads) or paired-end reads can help resolve repeats. If one read of the pair maps at or before the start of a repeat in the De Bruijn graph and the other path maps at or after the exit of the repeat in the graph the information can be used to resolve the correct path through that section of the graph [37]. However, if the insert size of the fragment producing the paired-end reads is shorter than the repeat sequence, it will not be possible to resolve the repeat properly. Paired-end and mate-pair reads are also used for scaffolding to join contigs into longer sequences where each read of a paired-end read overlaps a contig edge. This results in scaffolds with gaps of unknown sequence with estimated length based on the insert size of the fragment. In addition, dedicated software for scaffolding (see [39] for a comparison)  such as SSPACE (SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension) [40] can extend and scaffold contigs by mapping reads to contigs using Bowtie [41] to determine the position and orientation of each pair, before removing duplicate read-pairs. It determines putative contigs to scaffold based on the calculated distances between paired-end reads and iteratively combines these into scaffolds starting with the longest contig if a minimum number of read pairs support the connection. It can also handle multiple libraries by scaffolding them in a hierarchical manner, starting with the smallest insert size library [40].

Hybrid assembly strategies use a mixture of libraries with varying insert size and error profiles generated using different sequencing technologies. These are usually assembled separately with assemblers best-suited to the data type before being combined, although assemblers which can handle both long and short reads, e.g. hybridSPAdes [42],  Cerulean [43] and DBG2OLC [44], have also emerged (Table 1.2). Hybrid assembly approaches help to resolve repeat copy number, link contigs into scaffolds and close gaps or to improve existing assemblies [45–48].

Larger *k* values used in a De Bruijn approach produce longer contigs at high coverage sections but smaller contigs at lower coverage sections [37], so multiple *k* values are generally tested to find those producing an optimal assembly. Assemblies from multiple *k*-mers can also be combined to minimise trade-offs associated with any one *k* value, as implemented in SPAdes.

| Name | Assembler type | Data type required | Reference |
|---|---|---|---|
| Celera | Overlap consensus based assembly | Long reads only e.g Sanger sequencing reads | [31] |
| Velvet | De bruijn graph based assembly of *k*-mers | Illumina short reads (>25 bp). Sanger or 454 reads can also be added to resolve repeats if available. Paired-end and single-end reads supported | [34] |
| ABySS | De bruijn graph based assembly of *k*-mers | Illumina paired-end short reads (> 25 bp). | [36] |
| ALLPATHS | De bruijn graph based assembly of *k*-mers | Illumina short reads (25 to 50 bp). Needs two libraries, one with short insert size so that reads overlap and | [35] |

|  |  |  |  |
|---|---|---|---|
|  |  | one with large insert size (~ 4000 bp)<br>> 40X coverage also recommended and cannot be used with Sanger or 454 data. |  |
| SPAdes | De bruijn graph based assembly of *k*-mers | Paired-end, mate-pair or single-end reads. Illumina and Ion Torrent reads supported. | [27] |
| hybridSPAdes | Hybrid assembly of different data types | Short and long reads needed. PacBio, Sanger and Nanopore reads supported for hybrid assembly with shorter reads. | [42] |
| Cerulean | Hybrid assembly of different data types | Short and long reads needed. Resolves repeats by mapping long reads to assembly graph of short reads | [43] |
| DBG2OLC | Hybrid assembly of different data types using components of both de bruijn graph and overlap consensus layout approaches | Short and long reads required<br>Illumina, PacBio and Nanopore reads supported | [44] |

**Table 1.2:** Details of some assembly tools that are freely available and their requirements for use.

### 1.1.5 Assembly improvement

Several computational methods can be used to improve draft genomes output by assemblers. These include closing gaps, correcting erroneous bases and misassembled regions, determining the layout of scaffolds on chromosomes using a reference genome and performing annotation. Some are packaged as a set of tools for draft genome improvement such as PAGIT [49] (Table 1.3).

Gaps in scaffolds can be closed using tools such as IMAGE (iterative mapping and assembly for gap elimination) [50] or Gapfiller [51] (Table 1.3). IMAGE aligns paired-end reads to scaffolds and performs a local assembly (using De Bruijn graphs) of the aligned reads where one member of the read pair mapped to a contig and the other member is within a gap. This produces new contigs which are used to extend or merge the original contigs and fill part of the gap sequence. Reads are again mapped to the newly extended contigs and assembled as before to fill in more sequence, and this is repeated until the gap is closed. This approach can even be used with reads used to create the draft assembly, as *k*-mers that were too repetitive for incorporation in the genome-wide assembly can be unambiguously aligned to a contig edge reducing the search space [49,50]. Gapfiller takes a similar approach but also takes the size of the gap into account, as the gap length in each scaffold was originally estimated based on insert size of paired-end reads by the assembler. It maps reads back to the draft genome and looks for read pairs with a member mapped to a contig and a member in the gap. However, instead of local assembly of these reads, it extends the contigs at either side of the gap from each contig end using *k*-mer overlap, which helps overcome issues with tandem repeats. It also evaluates the accuracy of contig edges during gap closure as these can often be incorrectly assembled and only closes gaps if the newly assembled sequence corresponds with the estimated gap size. It then repeats the process until no more gaps can be closed [51].

iCORN (iterative correction of reference nucleotides) [52] can correct small insertion and deletion errors (up to three bp) as well as single base errors. It iteratively maps reads to the genome sequence using the SSAHA read mapper [53] and identifies the potential insertions, deletions and single base errors. The read coverage of reads that have the correct insert size and orientation (ie, perfectly mapped) is checked at each identified location using SNP-o-matic [54] and if the coverage at that location will not drop when the correction is applied, the sequence is corrected. It either continues till no new corrections can be applied [49].

Incorrectly assembled sequence (misassemblies) can be identified by mapping reads to the assembly and looking for deviations in the insert size distribution, read pair orientation and coverage. Misassemblies can also be pin-pointed by aligning the genome to a reference sequence to visualise differences in a genome browser such as the Artemis Comparison Tool (ACT) [55] or QUAST [56] (Table 1.3). QUAST aligns sequences to a reference using the NUCmer aligner from MUMmer [57] and considers misassembly breakpoints to be those where a) a position on assembled contigs in the sequence flanking the left side of the breakpoint is over 1 kb away from the sequence flanking the right side of the breakpoint on the reference, b) the flanking sequences overlap by 1 kb, or c) the flanking sequences align on opposite strands or different chromosomes on the reference [56].

| Name | Function | Reference |
|---|---|---|
| **PAGIT** | Package combining four improvement tools (ABACAS, IMAGE, iCORN and RATT) | [49] |
| **IMAGE** | Gap filling by read mapping to assembly | [50] |
| **Gapfiller** | Gap filling by read mapping to assembly | [51] |
| **iCORN** | Correct erroneous bases by mapping reads back to assembly | [52] |
| **ACT** | Visualise the comparison of multiple genomes (Comparison file needs to be produced using BLAST first) | [55] |
| **QUAST** | Quantify genome accuracy and completeness using a large amount of assembly metrics including N50, number and size of misassembled regions and gene content among others. | [56] |
| **REAPR** | Quantify genome accuracy and completeness by mapping reads back to the genome. | [58] |
| **CONTIGuator** | Contiguate contigs/scaffolds using a reference genome. Prokaryotes only. Webservice available. | [59] |
| **Mauve Contig Mover** | Contiguate contigs/scaffolds using a reference genome with the Mauve aligner.  Also allows manual movement and visualisation of sequences. Prokaryotes only. | [60] |
| **CAR** | Contiguate contigs/scaffolds using a reference genome. Prokaryotes only. | [61] |
| **Projector2** | Webservice to contiguate contigs/scaffolds using a reference genome. Prokaryotes only. | [62] |

| | | |
|---|---|---|
| **ABACAS** | Contiguate contigs/scaffolds using a reference genome. Prokaryotes and eukaryotes supported. | [63] |

**Table 1.3:** Summary of tools that can be used for assembly improvement.

GC bias describes the dependency between the proportion of G (guanine) and C (cytosine) bases in a region and the amount of fragments mapped to that region [64]. Both GC-rich and AT-rich fragments can be underrepresented in Illumina sequencing and library preparation procedures [65]. GC content varies among samples and can also be caused by PCR: fragments generated using PCR-free protocols and optimised PCR protocols tend to have reduced bias [64].

REAPR (recognition of errors in assembly using paired reads) finds misassemblies and corrects them [58]. It maps reads to the genome using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) allowing each member of a read pair to map independently so as not to force reads into mapping in proper pairs, which would reduce sensitivity. It extracts the coverage of perfect and uniquely mapped reads from the BAM file. A pre-processing step calculates the average insert size and depth of coverage from a sample of the genome and accounts for GC bias by calculating the expected fragment coverage at any given GC content value and plotting the GC content versus the fragment coverage in a scatter plot. It takes a LOWESS (locally weighted scatter plot smoothing) line through this plot and uses this to correct the fragment coverage in all downstream steps. REAPR creates a fragment coverage distribution (FCD) plot from fragments mapped to each base so that the difference in area between the theoretical FCD and observed FCD can be measured as the FCD error. Regions of the assembly with a high rate of FCD errors are marked as assembly errors. It also calculates the number of error-free bases in the assembly. For a base to be defined as error-free, it must have at least five properly paired mapped reads with no mismatches and FCD error below a threshold. It can produce a new corrected assembly by breaking the genome at gaps if an error is called at a gap or by replacing erroneous regions in contigs with 'N's [58].

Scaffolds or contigs can be contiguated (aligned, ordered and oriented) into pseudo-chromosomes if a suitable reference genome is available. Software such as CONTIGuator [59], Mauve Contig Mover [60], CAR [61], Projector2 [62] services and ABACAS [63] (Table 1.3) can be used for such ordering. However, the precision of the ordering will be determined by the completeness of the reference as well as their homology. ABACAS (algorithm-based automatic contiguation of assembled sequences) uses MUMmer, to align either amino acid or nucleotide sequences, to identify a tiling path of contigs along the reference chromosome, as well as estimating the sizes of gaps between contigs. Overlapping contigs are often caused by low quality at contig edges or low-complexity regions.

Annotation of assemblies to identify open-reading frames (ORFs) can be performed using a wide variety of software such as RATT (Rapid annotation transfer tool) [66], which transfers gene models from a reference assembly to a new one. RATT is also a component of integrated pipelines such as Companion [67] that both infers existing gene models based on a reference by alignment and discovers genes *de novo*. Tools such as Prokka [68] and RAST [69] are also informative for prokaryotic genome annotation. Transcriptome assemblies can also assist with verifying gene models and identifying issues caused by incorrect assembly. The details of these tools are summarised in Table 1.4.

| Name | Details | Organism | Type | Reference |
|---|---|---|---|---|
| **RATT** | Transfer annotation from a reference genome to a new genome | Prokaryotes and Eukaryotes | Unix | [66] |
| **Companion** | Pipeline combining *ab-initio*, gene transfer and protein sequence homology based approaches | Prokaryotes and Eukaryotes | Webservice or Unix | [67] |
| **Prokka** | Combines *ab-initio* tools and protein sequence homology | Prokaryotes | Unix | [68] |
| **RAST** | Combines *ab-initio* tools, literature review and protein sequence homology | Prokaryotes | Webservice | [69] |

**Table 1.4:** Summary of tools that can be used for genome annotation, the type of organisms that they are restricted to and whether or not the software can be run on a local Unix server or sequences must be submitted to a webservice.

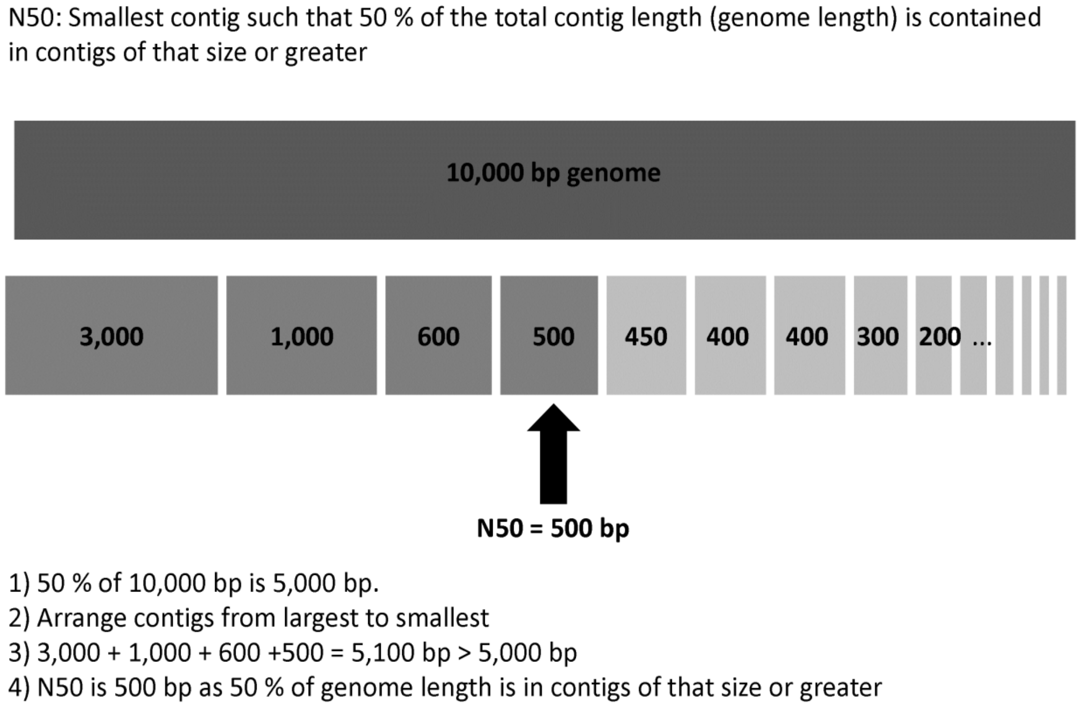
### 1.1.6 Evaluating assemblies

A large variety of factors can influence the completeness of genomes such as the read length, insert size, polymerase error rates and read coverage. Assemblers interpret low coverage as poor support for a region as they expect uniform coverage, so these low coverage regions are designated as errors and omitted from the final assembly. Similarly, high coverage can also cause fragmentation of the assembly because they it may be interpreted as evidence of repeats [65]. Genomes or parts of genomes with a high amount of repetitive sequence will result in gaps or incorrect repeat copy numbers because identical repeats may be collapsed in the assembly, or there may be incorrect arrangement of repeats. For example, *de novo* assemblies of the human genome missed 420 megabases (Mb) of repeat sequence, were 16% shorter than the reference and also missed 2,377 coding exons [70,71].

All assemblers have different parameters for estimating contig structure. For example, conservative ones that require extensive support to create contigs (or scaffolds) have lower error rates but shorter contigs, which reveal less about how a genome is organised and may make complete annotation more difficult. Less conservative assemblers produce longer sequences but may join sequences incorrectly or in the wrong orientation [72,73]. All assemblies will be incomplete and have misassembled sections or incorrect bases unless they are subjected to manual finishing, so for instance genes that appear missing in a genome can often be found by reanalysing the raw data as well as through targeted resequencing or transcriptome sequencing.

Six different levels of assembly completeness have been proposed [74]. Standard draft genomes contain sequences assembled into contigs with minimal filtering and the minimum amount of information needed for submission to a public database. High quality drafts have at least 90% of the genome assembled and contaminating sequences have been removed. Improved high quality drafts have had additional manual or automated improvement and have also undergone gap reduction to reduce the number of scaffolds or contigs as well as having with no detectable misassemblies and annotation directed improved assemblies have been annotated and the annotation checked to ensure it is as accurate as possible although repeat regions may still not be resolved. Non-contiguous finished describes genomes where most gaps, low quality regions and misassemblies have been resolved. Finished genomes are 'gold standard' genomes and are those that have no gaps with minimal errors (less than 1 error per 100 kb). These standards may overlap so that a genome fits more than one description [74].

Assemblies can be evaluated in multiple ways. The most common metric reported is the N50, which is the length of the smallest contig such that half the total contig length is contained in contigs of that size or greater (Figure 1.3 & Table 1.5). Larger sizes are generally considered better as they demonstrate that the assembly is not highly fragmented. However, incorrectly merged sequences or those containing many or long gaps will produce assemblies with high N50s even though there may be many misassembled sections or sections lacking sequence. Thus, the amount of gap sequence can help determine how much of the assembly remains to be resolved. Additionally software such as QUAST or REAPR can count and pinpoint misassembled sections. QUAST also outputs a number of metrics alongside misassembly and N50 information, which can give an idea of assembly quality [56]. Annotation can also inform on how many partial or full gene sequences are present for comparison reference datasets. Tools such as BUSCO (Benchmarking Universal Single-Copy Orthologs) [75] take this a step further by evaluating completeness using orthologous groups that are nearly universally single-copy genes in all species (Table 1.5).


N50: Smallest contig such that 50 % of the total contig length (genome length) is contained in contigs of that size or greater

1) 50 % of 10,000 bp is 5,000 bp.
2) Arrange contigs from largest to smallest
3) 3,000 + 1,000 + 600 +500 = 5,100 bp > 5,000 bp
4) N50 is 500 bp as 50 % of genome length is in contigs of that size or greater

**Figure 1.3:** An example of how to calculate the N50 of an assembly.

| Name | Details | Tools | Quantification | Reference |
|---|---|---|---|---|
| **N50** | Smallest contig length such that 50% of the total contig length is contained in contigs of that size or greater. | QUAST, REAPR | Completeness. Larger N50's indicate more complete genomes but are only meaningful when assemblies are of similar size. REAPR can also calculate a corrected N50 by breaking scaffolds at incorrectly joined regions. | [56,58] |
| **Percentage of Error Free Bases** | Percentage of genome with bases that are predicted to be correct based on reads mapped to them and locations outside of putative miss-assembled regions. | REAPR | Base level accuracy calculated by mapping reads to the genome and checking agreement between the base on the genome and the bases in the reads | [58] |
| **Gaps** | Percentage of genome that is composed of 'N' bases and the amount of large gaps that are present. | NA | Completeness. More complete genomes have a smaller percentage of gaps and longer scaffolds although this can be altered by procedures which manually reduce unknown gap lengths and so this must be considered also. | NA |
| **Number of assembled genes** | Number of genes found compared with a reference genome (QUAST) or number of genes present from a group of genes that are found as single-copy genes in most species (BUSCO). | QUAST, BUSCO | Completeness. Gene numbers close to the expected number (based on a related reference genome) indicate that most gene models are likely present. | [56,75] |

**Table 1.5:** Summary of metrics that can be used to quantify completeness or accuracy of a genome assembly and tools that can calculate these.

### 1.1.7 Mapping short reads to genomes

Aligning a short sequence read to a long reference sequence is termed mapping. Short read mapping tools generally use either a hash-based or Burrows Wheeler Transform (BWT)-based methods [76] (Table 1.6). Hash-based methods were generally used in the first short read alignment programs created to deal with NGS data and include programs such as MAQ [77], SOAP [78] and SMALT (http://www.sanger.ac.uk/science/tools/smalt-0). A hash table is a data structure that can index data in a way that allows rapid searching. Tools may either use the reference genome to scan a hash table created from the reads, or use the reads to scan a hash table created from the genome (as done in SMALT). BWT based methods typically use a data structure called the FM index (also called a compressed suffix array). This is created by first modifying the reference genome using the BWT algorithm so that sequences that appear multiple times in the genome are together in the data structure. An index is created from this and it is used to rapidly determine where reads originated in the genome. This method is considerably faster than hash-based implementations, but with a slight loss of sensitivity and an inability to align highly polymorphic reads. Examples of mappers that use this approach are Bowtie, Bowtie2 and BWA [32,79–81].

| Name | Method | Reference |
|---|---|---|
| **MAQ** | Hash-based | [77] |
| **SOAP** | Hash-based | [78] |
| **SMALT** | Hash-based | http://www.sanger.ac.uk/science/tools/smalt-0 |
| **Bowtie &Bowtie2** | BWT | [80,82] |
| **BWA** | BWT | [81] |

**Table 1.6:** Summary of short-read mapping tools that use either hash or Burrows Wheeler Transform (BWT) based approaches.

### 1.1.8 Structural variation detection using whole genome shotgun sequence data

Structural variants (SV) are changes in the arrangement of genomic regions and are generally larger than 50 bp [83]. SVs account for 1.2% of variation in human genomes [84] and include changes such as deletions, insertions, translocations, inversions, mobile element transpositions and copy number variants (CNVs). Insertions are additions of sequence to the sample genome that are not in the reference genome and deletions are regions where bases that are present in the reference genome have been removed from the sample genome. As it is often unclear as to whether an insertion or deletion is an addition to the reference or a removal from the sample, insertions and deletions are collectively known as indels. SVs may be balanced, where the total number of nucleotides does not change as in an inversion (the orientation of a locus is reversed relative to a reference), or unbalanced where the number of nucleotides increases or decreases e.g. indels. Complex structural variants can also occur which encompass a mix of individual types such as tandem duplications with nested deletions or duplication-inversion-duplication events [85,86]. Mobile genetic elements (MGEs) are sections of DNA that usually encode the enzymes enabling themselves to move either within genomes, or between bacteria via horizontal gene transfer (HGT) [87]; these are discussed in more detail later. A translocation occurs when a sequence has moved to a new location on the genome, either on the same chromosome (intrachromosomal) or a different chromosome (interchromosomal). A CNV or copy number alteration (CAN) can either be a copy number gain or a copy number loss. A copy number gain is where there are additional copies of a sequence relative to the reference sequence and a loss is a drop in copy number of a sequence on a sample genome compared with the reference. Tandem and dispersed duplications are examples of copy number gains.

There are four approaches used for detecting SVs using sequence data: read depth, read-pair mapping, split-read mapping and *de novo* assembly methods. Read depth (depth of coverage) methods are generally used for CNV detection and estimation of copy number, and are the only method that can directly detect CNVs [88]. This method assumes that sequencing produces uniform coverage so that the number of reads mapped to a region is proportional to the number of times that region is on the sample genome. Reads from a sample can be mapped to a reference genome if one is available or to their own assembly, although mapping to a reference means that sections in the sample that are not in the reference will not be considered. Increases in read coverage relative to the read coverage across a chromosome indicate amplified regions, whereas deleted regions have lower read coverage than surrounding regions. Copy number is generally estimated by comparing the coverage in non-overlapping windows, usually > 1 kb, across a chromosome and comparing this with the median coverage of the chromosome. Prior to copy number estimation, coverage can also be normalised to account for GC bias, for example by counting the number of reads mapped to windows across the genome and adjusting the count based on the GC content of the window. Mapping bias is caused by reads mapping to multiple positions due to short read length and repetitive sequences, which can inflate the read coverage at certain regions. This can be alleviated to some extent by only considering reads that unambiguously map or by randomly assigning an ambiguously mapped read to one position. However, the latter approach can increase the number of false positive CNVs. Paired case and control samples can be used so that copy number is calculated as a relative quantity between these instead of an absolute quantity e.g. the number of copies in the tumour sample compared with the parent sample [33,83,89].

Read-pair mapping methods are based on examining the insert size distribution and orientation of paired-end reads from a sample mapped to a reference genome. Paired-end reads produced by the same library preparation protocol have a specific insert size distribution, so read-pairs whose distance is outside this distribution may be mapped to a locus where a genomic rearrangement has taken place. Read-pairs mapping far apart on the reference indicate a deletion in the sample, whereas reads mapping closer than expected indicate an insertion. If the orientation of the reads is changed so that for instance both forward and reverse reads now both face the same direction but the read-pair is mapped with expected insert size, an inversion can be suspected. Translocations can be detected where members of a read pair map on different chromosomes. These events can be confirmed by viewing mapped read orientation and insert sizes using tools such as the Integrative Genomics Viewer (IGV) [90], local assembly of reads spanning the breakpoints, or long read sequencing such as Sanger, PacBio or Oxford Nanopore sequencing to generate reads spanning the rearrangement and its breakpoints (start and end positions of the variant on the chromosome). However, read mapping approaches cannot detect insertions larger than the average insert size or variants in low complexity regions and so they are generally used to find small events (< 100 bp) [83,91].

Split-read methods also use paired-end reads to identify SVs and have single base-pair resolution. As before, paired-end reads are mapped to a reference genome and read pairs where one member maps to the reference and the other is either unmapped or mapped by allowing a gap (split read) are detected. In tools such as Pindel [92], the mapped member is commonly used as an anchor point from which to attempt to map the other read. The unmapped read is split into two fragments, as it may be spanning the breakpoint of a deletion. Based on the mapped member and insert size, a direction and mapping location for the fragmented parts of the unmapped read can be discovered, revealing the breakpoint and deleted sequence. This method can find deletions < 10 kb and insertions < 20 kb using 36 bp reads. A more recent version of Pindel implements both a split-read and paired-end approach to find indels as does another tool called DELLY [88,93] (Table 1.7).

Assembly-based approaches either first assemble the reads into contigs and align the contigs to the reference before detecting variants, or use the reference genome as a guide to assemble the reads (reference-guided). However, as mentioned already, *de novo* assembly is computationally intensive and assemblers struggle with duplicated regions potentially collapsing multiple copies together, though it does have the ability to examine more complex variants and pinpoint breakpoints.

All these approaches tend to yield SV sets that have incomplete intersections even when the callers use the same approach. Another problem is that many individual callers, particularly those that use read-pair and split-read methods, have high false discovery rates (FDR). This means multiple methods need to be implemented to attempt to resolve variants and their breakpoints [88,94]. Ensemble-based methods use multiple callers and combine the output (such as SVMerge [95]) or only report variants supported by at least two callers (such as in IntanSV, https://www.bioconductor.org/packages/release/bioc/html/intansv.html and HugeSeq [96]) (Table 1.7). The ability of each caller to resolve the breakpoints of the variants is important as these allow the correct merging of results from individual callers and samples. Some studies use callers which can handle multiple samples simultaneously to prevent having to perform merging of results from multiple samples using other tools [88].

| Name | Method | Reference |
|---|---|---|
| **Pindel** | Split read and read pair | [92] |
| **DELLY** | Split read and read pair | [93] |
| **SVMerge** | Ensemble | [95] |
| **InstanSV** | Ensemble | https://www.bioconductor.org/packages/release/bioc/html/intansv.html |
| **HugeSeq** | Ensemble | [96] |

**Table 1.7:** Summary of Structural Variant calling tools discussed in this section and the approach used by these tools. Ensemble based methods integrate the output of multiple individual callers.

### 1.1.9 Single nucleotide polymorphism (SNP) detection

Traditional approaches using alignment of assembled sequences for SNP detection (such as NUCmer in the MUMmer package) [97] have long since been surpassed by read-mapping that uses consensus calling schemes that are based on the framework outlined for the 1000 Genomes project [98]. This proposed that no single caller was necessarily better, and importantly that errors are independent across callers such that valid SNPs will show a consistent signal which aids in reducing the false discovery rate (FDR) [99]. This approach is effective for samples without a database of known variants, and thus is particularly advantageous for *Leishmania* or *de novo* mutation discovery [100].

Reads are mapped to a reference, which can be a different sample or the same original sample. For the latter, few homozygous SNPs are expected from such self-mapped reads and examining the read depth coverage of different allele frequencies can help determine chromosome copy number. SNPs and small indels may be detected using tools such as GATK [101], SAMtools [102,103] and Freebayes [104] (Table 1.8). SAMtools uses a Bayesian model that includes information on sequencing error, read mapping qualities and read depth to infer the genotype with the highest probability at a site as well as estimate the allele frequency. It also incorporates a base alignment quality (BAQ) calculation which represents the probability that a base is misaligned, which can reduce false positives caused by misalignment around indels [102]. However, SAMtools assumes the prior probability of observing a heterozygote at 0.001 making it less likely to call heterozygous genotypes than callers with no such prior [105,106]. It also assumes that samples are diploid whereas some others allow variable ploidy. Variants can be filtered based on multiple criteria such as the removal of low complexity regions to reduce false positive SNPs [106].

| Name | Method | Details | Reference |
|---|---|---|---|
| **MUMmer** | Pairwise genome alignment | SNP calling based on differences between two aligned genomes | [97] |
| **GATK** | Read mapping | Maps reads to genome and calls SNPs and small indels based on Bayesian models | [101] |
| **SAMtools** | Read mapping | Maps reads to genome and calls SNPs and small indels based on Bayesian models | [102,103] |
| **Freebayes** | Read mapping | Maps reads to genome and calls SNPs and small indels based on Bayesian models | [104] |

**Table 1.8:** Summary of tools discussed in this section that can be used to call SNPs and small indels.

### 1.1.10 Measuring changes in gene expression

The first step to determine the expression level of genes is to map the reads to a genome. In the case of most prokaryotes, this is technically straightforward as there are no introns and so mappers such as Bowtie or BWA can be used. However, reads span splice junctions in most eukaroytes that have pre-mRNA splicing. In these cases, a gapped or splicing-aware aligner such as Tophat [107] must be used: these map unspliced reads to locate exons, and split unmapped reads to align them separately to identify the exon junctions [108]. If no reference genome is available, reads can be *de novo* assembled into transcripts using assemblers such as Oases [109] or Trinity [110] (Table 1.9) and either the reads mapped back to this transcriptome to quantify the abundance of transcripts, or the assembled transcripts aligned to a closely related reference genome. For reads mapped to the genome, tools such as Cufflinks [108] can be used to calculate transcript abundance and detect novel transcripts. If gene level quantification of expression is required, software such as HTSeq [111] can count the number of number of reads mapping to each gene. For paired-end reads, if both members of the pair map to the same gene the count for that gene is only increased by one, as both reads originated from the same cDNA fragment.

Changes in expression can be evaluated using the scaled difference and ratio of expression between treatments/conditions [112] typically by applying discrete probability distributions such as negative binomial or poisson distributions. A negative binomial distribution is a generalisation of the poission distribution that allows for overdispersion [113]. Such packages include such as DESeq2 [114], edgeR [115] or PoissonSeq [116] (Table 1.9). In DESeq2, the read counts are normalised to account for differences in sequencing depth between samples. Changes in expression can be assessed using a Wald test where the null hypothesis is that there is no difference between treatment and control. A threshold log2 fold change difference can also be used. The Wald test divides the shrunken estimate of the log fold change (LFC) between groups by its estimated standard error to produce a Z statistic which is compared to a standard Normal distribution. The LFC and standard error were obtained by fitting a generalised linear model to each gene. Genes with low counts can have large LFC estimates due to low numbers, so LFC estimates are shrunk towards zero based on the gene count, with lower counts experiencing more shrinkage in DESeq2 [114]. The p-values produced for each gene by the Wald test must be corrected for multiple testing, using methods such as Benjamini-Hochberg correction [117]. Within-group expression can be estimated if biological replicates are provided [113].

Batch effects are caused by subgroups of genes whose expression has consistently different behaviour across conditions independent of the biological or scientific variables being examined [118]. Variables such as processing date and lab can be included in the models used to detect differential expression in order to remove this signal. However, if the preparation is confounded with the variable of interest e.g. if we are measuring the difference between a treatment and a control and all the treated samples are prepared in lab A while all the control samples are prepared in lab B, it will be impossible to tell if genes are differentially expressed due to being prepared in different labs or if the effect is due to the treatment [118]. Unknown or unmeasured batch effects can also be uncovered using techniques such as surrogate variable analysis (SVA) as implemented in packages such as svaseq [119] or PEER [120] (Table 1.9), and the latent variables produced by this incorporated into linear models in the same way as known batch effects. Latent variables can be associated with: GC-content, uniformity of gene body coverage, insert size and base error rates. Detection and removal of latent variables which are often associated with these effects dramatically reduces the FDR [121].

| Name | Function | Reference |
|---|---|---|
| **Bowtie** | Map reads to genome | [82] |
| **BWA** | Map reads to genome | [81] |
| **Tophat** | Map reads to genome (splice aware mapper) | [107] |
| **Oases** | De-novo transcript assembly | [109] |
| **Trinity** | De-novo transcript assembly | [110] |
| **Cufflinks** | Quantify transcript abundance and find novel transcripts | [108] |
| **HTSeq** | Count reads mapped to genome | [111] |
| **DESeq2** | Detect differential gene expression (R package) | [114] |
| **edgeR** | Detect differential gene expression (R package) | [115] |
| **PoissonSeq** | Detect differential gene expression (R package) | [116] |
| **svaseq** | Detect batch effects (R package) | [119] |
| **PEER** | Detect batch effects (R package) | [120] |

**Table 1.9:** Summary of tools discussed in this section which can be combined to measure changes in gene expression.

## 1.2 *Leishmania*

*Leishmania* are protozoan parasites belonging to the *Trypanosomatidae* family (class Kinetoplastida) that cause the neglected tropical disease Leishmaniasis (or leishmaniosis). All members of the *Trypanosomatidae* family are exclusively parasitic and are found in insects, plants, invertebrates and vertebrates. Other members of the *Trypanosomatidae* that cause disease include *Trypanosoma cruzi* which causes Chagas disease and *T. brucei* species which cause African trypanosomiasis (African sleeping sickness). Kinetoplastida members are all flagellated at some point in their lifecycle and have a unique mitochondrial organelle called the Kinetoplast. This organelle is located at the base of the cell flagellum (basal body) and its DNA (kDNA) is organised as a network of interlocking rings producing a chainmail like appearance. The kDNA structure in *Leishmania* and *Trypanosoma* is composed of concatenated maxicircles and minicircles and each kinteoplast contains five to ten thousand small minicircles (0.5 to 10 kb) and 25 to 50 larger maxicircles (20 to 40 kb) [122].

*Leishmania* are transmitted by the bite of infected female phlebotomine sandflies and have a digenetic (two hosts) lifecycle, existing as flagellated promastigotes in the sandfly gut and as intracellular amastigotes in macrophages in vertebrate hosts such as humans, canids and rodents. They are found in the tropics and subtropics in areas including North, Central and South America (the New World) and also the south east of Europe, the Mediterranean basin, the Middle East, Indian subcontinent, Africa and Central and Southeast Asia (collectively known as the Old World), as well as in red kangaroos in Australia [123,124]. Leishmaniasis is endemic in 98 countries on five continents and currently infects 12 million people with another 350 million people at risk [125,126]. It can have either an anthroponotic transmission cycle where humans are the main or only hosts or a zoonotic transmission cycle in which animals are the main hosts and can transmit it among themselves and to humans.

### 1.2.1 *Leishmania* classification

There are 53 known species of *Leishmania*, 31 of which can infect mammals and 20 of which that can infect humans [123]. The classification of these species into subgenera has been fraught with debate and is still being refined and updated. These species were originally classified based on their development in sandflies and their choice of hosts. Species that developed in the mid- and foregut of sandflies were classified as the subgenera *L. (Leishmania)*, those that underwent additional development in the hindgut were classified in the subgenus *Viannia* and those that infected lizards (and mainly developed in the sandfly hindgut) were proposed to belong to a separate genus named *Sauroleishmania*. The most recent classification system based on molecular data has divided the *Leishmania* species into two major phylogenetic lineages termed sections – these are *Euleishmania* and *Paraleishmania* (Figure 1.4). *Euleishmania* has three subgenera and one complex: *Leishmania*, *Viannia*, *Sauroleishmania*, and the *L. enriettii* complex [127]. The *Paraleishmania* section includes five *Leishmania* species and species that were formerly classified in the closely related *Endotrypanum* genus of which there are only two: *E. schaudinni* and *E. monterogeii*. *Paraleishmania* has not been fully resolved and so is a polyphyletic clade (contains individuals that are not closely related as they descended from more than one ancestor) [123,127]. The *Leishmania* subgenus contains thirteen species whose members have been found in both the New and Old World. These species are distributed in four species complexes which are the *L.major* complex, the *L. donovani* species complex, the *L. tropica* complex and the *L. mexicana* complex [128] (Figure 1.4). The subgenus *Viannia* consists of nine species which are only found in the New World [123]. It's members usually cause cutaneous leishmaniasis (CL) and *L. panamensis*, *L. guyanensis*, *L. braziliensis* and *L. peruviana* can cause mucocutaneous leishmaniasis (MCL) (discussed in section 1.2.2) [129–131]. Another member, *L. utingensis,* has only been found in sandflies [132]. Within *Viannia*, *L. panamensis* and *L. shawi* are subspecies of *L. guyanensis* and all three belong to the *L. guyanensis* species complex. *L. peruviana* may also be a subspecies of *L. braziliensis* and both are in the *L. braziliensis* species complex [123] (Figure 1.4).

The *Sauroleishmania* subgenus contains 19 named and two unnamed species (Figure 1.4) whose members infect reptiles in the Old World – members of this subgenus are not found in the New World [123] (Figure 1.4). However, one member *L. adleri* can infect both reptiles and mammals, including humans where it can cause CL. It also develops in the mid-gut of the sandfly in common with *L. (Leishmania)* species [133]. *L. tarentolae* is commonly used in laboratories due to its non-pathogenic nature in humans and is the type strain for this subgenus. Once thought to exclusively infect reptiles, more recent evidence has emerged of *Sauroleishmania* infections in humans and in China [134] and also in a 300 year old Brazilian mummy, which had been infected with *L. tarentolae* possibly causing visceral leishmaniasis (VL) based on kDNA isolated from its bone marrow [135]. Little molecular data exists for many members of these species resulting in only a few such as *L. adleri*, *L. gymnodactyli*, *L. hoogstraali* and *L. tarentolae*, commonly used for phylogenetic analysis in studies.

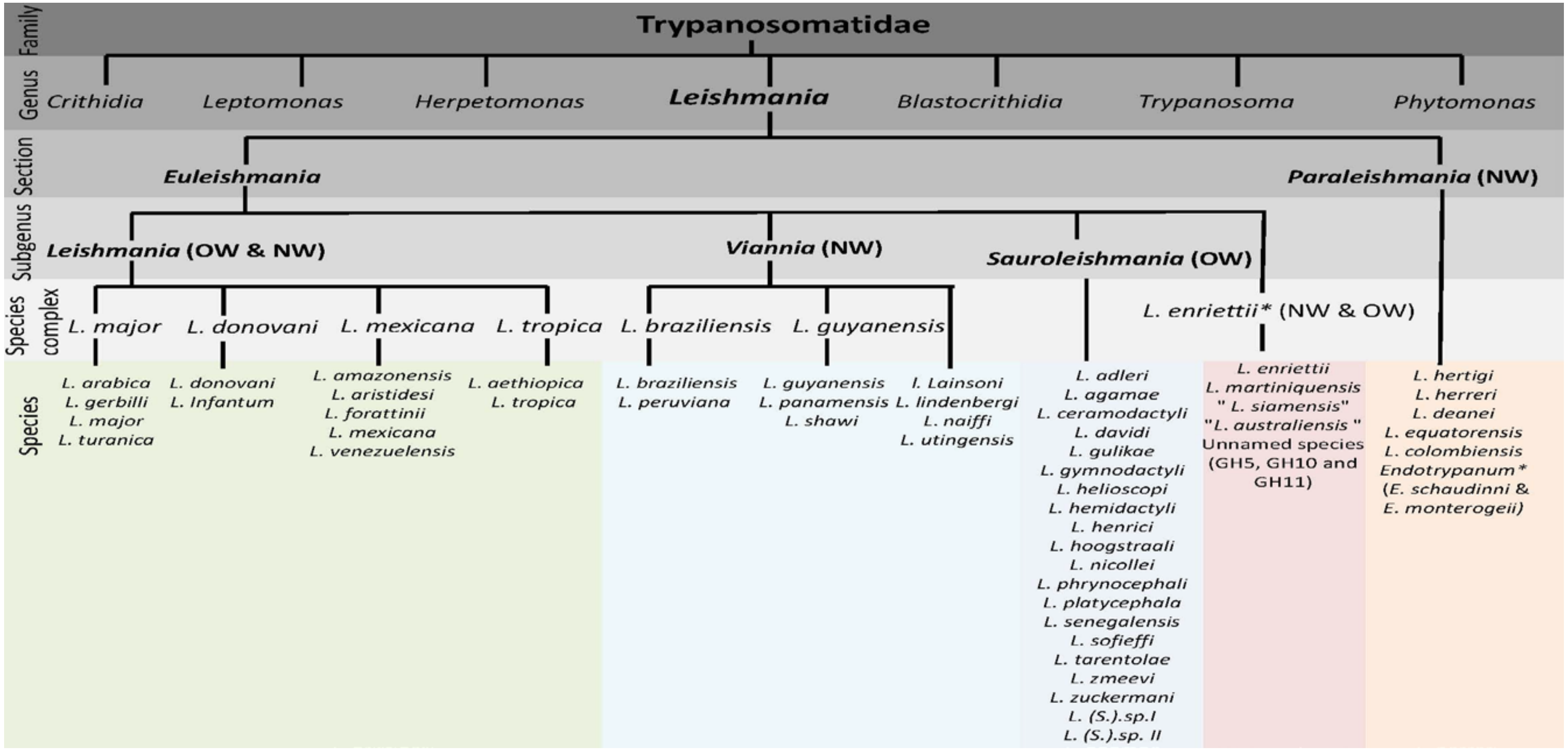


**Figure 1.4**: Classification of the *Leishmania* genus. This image is based on an image in [136] but uses information from [123,137]. *L. enretti** indicates that the *L. enriettii complex* may constitute a new subgenus [123] and *Endotrypanum** indicates that *Endotrypanum* was formerly classified as a genus but is proposed to be part of the Section *Paraleishmania* [127]. The three unmanned species in the *L. enriettii* are from [137]. "*L. siamensis*" and "*L. australiensis*" have not been formally described.

### 1.2.2 Clinical presentation and incidence

Leishmaniasis has complex, highly variable and understudied clinical presentations. Traditionally these have been classified as cutaneous leishmaniasis (CL), mucocutaneous leishmaniasis (MCL) and visceral leishmaniasis (VL, also known as kala-azar). However, advances in the molecular biology of *Leishmania* have demonstrated that these categories are no longer useful for tackling this disease. The type of leishmaniasis infection that occurs is determined by a variety of factors such as parasite species, host immune response (influenced itself by a variety of factors such as co-infection, nutritional status and genetics) and other environmental factors. For example, short-term exposure to sandfly bites prior to *Leishmania* challenge confers a degree of protection from infection in mice, though no protection is afforded by long term exposure or a long time period since the last bite [138].

CL is the most common form causing 0.7 to 1.2 million leishmaniasis cases a year. It begins at skin exposed to the sandfly bite (localised CL) resulting in pink papules (pimples or swellings) that enlarge developing into nodule (solid lesion greater than 1cm in diameter) or plaque (elevated, superficial lesion greater than 1 cm in diameter) like lesions. Multiple lesions may also be present and lesions leave lifelong scars causing significant social stigma. Diffuse cutaneous leishmaniasis (DCL) is a rare form of CL that begins as a lesion that does not ulcerate and is associated with infection by *L. aethiopica* or *L. amazonensis*. Instead, parasites disseminate to macrophages in other parts of the skin and can involve skin on the whole body. DCL tends to appear in immunocompromised hosts, such as those with human immunodeficiency virus (HIV) [125].

Another rarer form of CL is known as leishmaniasis recidivans (also lupoid leishmnaiasis, tuberculoid leishmaniasis or chronic relapsing cutaneous leishmaniasis) [125,139]. This is caused by *L. tropia* in the Old World and involves the recurrence of lesions at their original site years after they have apparently healed. It is caused by persistent infection with *Leishmania* parasites, can be triggered by trauma in the area of the original lesion [140], and commonly infects children. It often involves the face, particularly the cheek, manifesting as an enlarging papule that heals with scarring in the centre and its relentless expansion can cause significant destruction of facial tissue [125,139].

VL, also known as Dumdum fever or black fever, is a systemic disease caused by infection of the liver, spleen and bone marrow resulting in darkening of the skin, fever, significant weight loss, anaemia and swelling of the liver and spleen among other symptoms. It generally appears between two to eight months after initial infection, is fatal without treatment and is the second largest parasitic cause of fatality in the world after malaria. It is also an important opportunistic infection [125] in those with HIV and can accelerate progression of HIV to AIDS.  Up to 70% of VL cases in adults in southern Europe are in people with HIV (HIV/VL co-infection). There are approximately 0.2 to 0.4 million VL cases annually, with 50,000 deaths [125]. Secondary bacterial infection leading to sepsis, commonly caused by *Staphylococcus aureus,* is also a risk factor for death in VL patients [141,142]. 90% of VL cases occur in just six countries (India, Bangladesh, Sudan, South Sudan, Brazil and Ethiopia) compared with a more dispersed distribution for CL, where 70 to 75% of the estimated incidence occurs in the ten countries of Afghanistan, Algeria, Colombia, Brazil, Iran, Syria, Ethiopia, North Sudan, Costa Rica and Peru [126]. VL is mainly caused by *L. donovani* and *L. infantum* (which is in the *L. donovani* complex) in the India subcontinent and East Africa. A type of CL, termed post-kala-azar dermal leishmaniasis (PKDL), can also develop months to years after recovery from VL, most commonly in patients in East Africa and the Indian Subcontinent. It causes multiple chronic lesions on the face which can persist for decades and these can either resolve spontaneously or require prolonged treatment [125].

MCL, or espundia, is characterised by the destruction of the mucosa, most commonly the nose and mouth but also the upper respiratory tract, from parasites that have spread from the site of the original bite. It can cause substantial disfigurement especially if treatment is not initiated promptly, and it rarely heals spontaneously. Secondary bacterial infections are again common and pneumonia caused by bacterial infection is the most common cause of death in patients with MCL. It is typically caused by New World parasites belonging to the *Viannia* subgenus such as *L. braziliensis*, *L. panamensis* and *L. guyanensis* and about 90% of cases occur in Brazil, Bolivia and Peru [125].

### 1.2.3 Diagnosis and Treatment of Leishmaniasis

Diagnosis is made by detection of *Leishmania* parasites using light microscopy or DNA in tissue specimens, using techniques such as PCR and isoenzyme analysis. No vaccine is available for leishmaniasis in humans. A variety of treatments are available, although most of these have limitations such as toxicity, need for parenteral administration, long durations of treatment and the need for hospitalisation. These include pentavalent antimonials ($Sb^v$) compounds such as sodium stibogluconate which is highly toxic to veins (phlebotoxic), as well as amphotercin B, paromomycin, a newer oral treatment called miltefosine, and the azoles: ketoconazole**,** itraconazole, and fluconazole. Resistant to pentavalent antimonials is widespread. This is because the drug is readily available and many patients first consult unqualified medical practitioners [143] leading to substandard doses. There have also been problems with inconsistent manufacturing quality of this drug in some places [144] and the anthroponic transmission cycle of leishmaniasis in India and East Africa can lead to uptake and subsequent infection of others with resistant parasites [145]. In an attempt to prevent the appearance of lesions caused by *Leishmania* on areas such as the face, which can cause severe social stigma, people in some countries such as Iran and Israel, undertook a practice known as 'leishmanization'. This involved the inoculation of live *Leishmania* parasites from the exudates of a skin lesion into young children, particularly girls, in a hidden part of their body, leading to the development of a self-healing lesion at the inoculation site. It produced a vaccination-like effect which protected them from getting lesions from other *Leishmania* infections on more exposed areas such as the face. Although practiced in many places since the 1940s, this practice has largely been abandoned due to issues such as non-healing lesions and standardisation, although it is still practised in Uzbekistan, where a mixture of live and attenuated *L. major* is licensed for use as a vaccine in high-risk populations. However, fresh isolates must be cultured from humans each year to maintain high virulence, as successive passage of parasites in culture leads to loss of virulence [146,147]. This demonstrated that immunity could be achieved by prior infection and paved the way for current research into *Leishmania* vaccine candidates.

### 1.2.4 *Leishmania* lifecyle

When a female phlebotomine sandfly bites a host to take a blood meal, it inoculates metacyclic flagellated promastigotes of *Leishmania* into the skin of the host via its proboscis (mouthpart) (Figure 1.5). Metacyclic promastigotes are the extracellular, non–dividing, infective stage of the parasite. These are phagocytosed by host macrophages where they transform and replicate into small (3–5 µm), intracellular, aflagellated amastigotes inside the acidic phagolysosomal compartments of macrophages. The amastigotes scavenge their essential nutrients such as purines, vitamins, lipids and at least ten essential amino acids from the phagolysome via membrane transporters [148,149]. Amastigotes also periodically escape the macrophages and infect other macrophage cells. When a sandfly takes a blood meal from the infected host, the *Leishmania* amastigotes are also taken up from macrophages in the blood and released in the gut where they move to the midgut or hindgut, depending on the parasite species (*Viannia* move to the hindgut at first). The decrease in temperature and increase in pH moving from macrophages to the sandfly gut drives their transformation into procyclic promastigotes - weakly mobile forms with a short flagellum at their anterior end [150]. These move towards the anterior midgut of the sandfly, until they reach the stomodeal valve, which is at the junction between the fore- and midgut [151]. In order to avoid being excreted during digestion of the blood meal, they attach to the epithelium of the gut and this binding is has been postulated to be the main determinant of vector parasite specificity [152]. Binding and subsequent detachment of metacyclic promastigotes from the epithelium of some vectors such as *L. major* in *Phlebotomus papatasi* is mediated by stage-specific modifications in the structure of lipophosphoglycan (LPG), a glycolipid covering the entire promastogote cell surface, that binds to a galectin on the sandfly epithelium, although in other *Leishmania* spp. it may occur via an LPG independent mechanism [153–156]. They divide by binary fission becoming metacyclic promastigotes. Once attached to the gut, the promastigotes encase themselves in a gel-like plug called promastigote secretory gel (PSG), whose main component is filamentous proteophosphoglycan (fPPG) [151,154]. This plug blocks and holds open the stomodeal valve [157]. The metacyclic promastigotes move to the pharynx (still inside the PSG), and the sandfly regurgitates the promastigotes inside the PSG when it bites, along with their saliva which contains agents such as vasodilators, anticoagulants, anti-platelet agents, as well as molecules with anti-inflammatory and immunomodulatory effects that help to promote *Leishmania* infection [154,158]. The blockage of the sandfly gut with PSG also hinders the ability of the sandfly to take a blood meal, leading it to probe the skin more frequently, feed more often, spend more time feeding and deliver more bites than uninfected sandflies, which increases the transmission of the parasite [159,160].

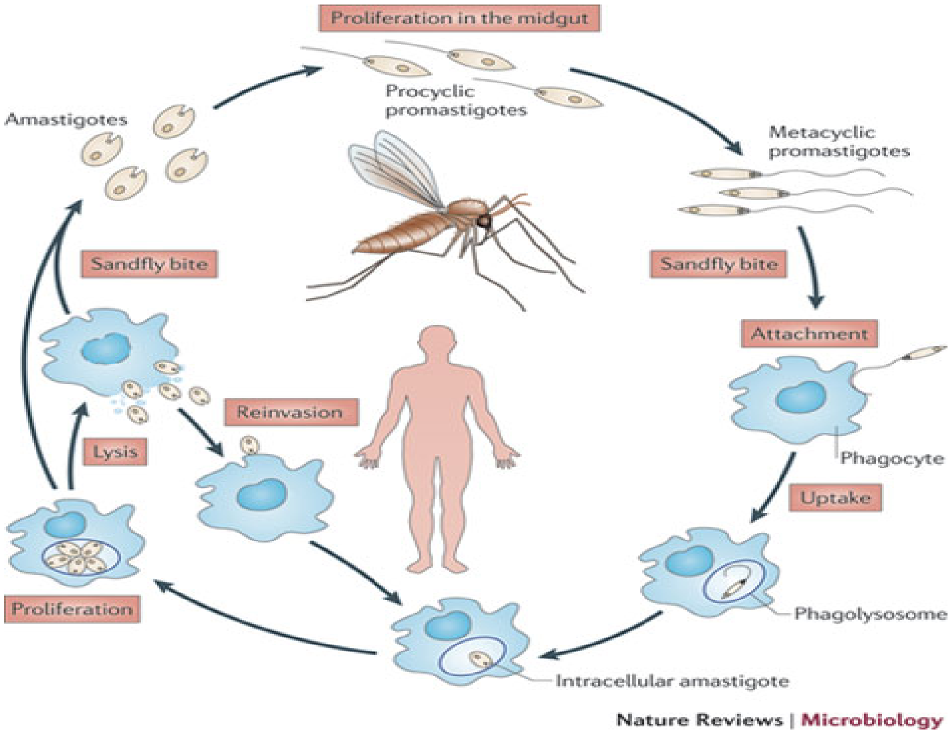


**Figure 1.5:** *Leishmania* lifecycle. This figure is reproduced from [161] with permission.

### 1.2.5 *Leishmania* vectors

There are over 800 species of phlebotomine sandfly – approximately 464 are found in the New World and 375 in the Old World. Of these, 98 are proven or suspected to be vectors of human leishmaniasis, 42 which transmit it in the New World and 56 in the Old World [123,162]. The Phlebotominae are a subfamily of the family Psychodidae, order Diptera, class Insecta. This subfamily has five genera: *Phlebotomus, Sergentomyia* and *Chinius* in the Old World and *Lutzoymia, Warileya* and *Brumptomyia* in the New World. *Lutzoymia* species transmit *Leishmania* in the New World, *Phlebotomus* transmits it in the Old World and *Sergentomyia* [163] is a vector of *Sauroleishmania* spp. in the Old World. There is also evidence that day-feeding midges (order Diptera, family Ceratopogonidae) are potential vectors of *Leishmania* in red kangaros in Australia [164], and fleas and ticks have been implicated in spreading other *Leishmania* spp. such as *L. infantum* [165,166]. Adult phlebotomine sandflies are small (< 3.5 mm in length), range in colour from almost white to almost black, and unlike mosquitos, silently attack their hosts. Female sandflies require a blood meal in order to complete the development of their eggs, and thus only females bite and transmit *Leishmania* spp. (also male sandfly mouth parts cannot pierce the skin). They are mainly active from dusk to dawn in humid conditions when there is no wind and their seasonal activity is affected by rainfall and temperature [160].

For a sandfly to be demonstrated as a vector of *Leishmania* it must satisfy five criteria: the sandfly must feed on humans (anthropophilic), it must bite the reservoir host in the case of *Leishmania* that is transmitted through zoonotic transmission cycles, it must be naturally infected in nature with the same *Leishmania* species that is found in humans using DNA or isoenzyme analysis, it must support the complete insect stage development (promastigote stage) of the *Leishmania* parasite and it must be able to transmit the parasite by biting a susceptible host when it takes a bloodmeal. Vector species can be considered as ‘proven’, ‘strongly suspected’ or ‘suspected’ based on meeting either some or all of these criteria [162,167]. Sandflies are considered to be either specific (restrictive) vectors if they only transmit one species of *Leishmania* or permissive vectors when they can transmit multiple species. Most are restrictive vectors of *Leishmania*. Other routes of *Leishmania* transmission are rare and are caused by needle sharing in drug users, blood transfusion, as well as venereal and congenital transmission [160].

### 1.2.6 Zoonotic *Leishmania* and reservoirs

Most *Leishmania* infections are zoonoses where animals are the reservoir hosts and humans are incidentally infected. Notable exceptions are Old World VL and PKDL, caused by *L. donovani* and the CL caused by *L. tropica*, which are predominantly anthroponotic, and occur in densely populated urban settings [125]. A reservoir is an animal that can maintain the parasite in nature and plays a distinct role in its transmission, although their competence at transmitting the parasite may vary [168]. In order to be defined as a reservoirs they need to be present in large enough numbers and sufficiently long lived to both maintain and transmit the parasite, the parasite should be available in a large enough quantity in the skin or blood of the host to be taken up by the sandfly, intense sandfly contact with the host must be observed, a large number of hosts can become infected in the same season. The parasites recovered from the host must also be the same as those found in humans [125].

Many reservoir hosts of *Leishmania* are found in forests or jungles (sylvatic) and these include foxes, wolves, jackals, rodents, sloths and armadillos among others [168], although some such as dogs, horses and donkeys are common in domestic and peridomestic settings [125]. Human migration into tropical areas and conflict which leads to the movement of civilians and soldiers into forested areas, as well as activities such as hunting, mining and highway construction can bring people into contact with sylvatic reservoirs. The flight range of sandflies is short (approximately 300 m) and so contact with the reservoir inevitably leads to contact with sandflies that bite it. In Mexico, CL is known as ‘ulcera de los chicleros' and signifies the association of the disease with people who travelled into forests to gather latex (‘chicle’) from trees [162].

In South America and the Mediterranean basin, dogs are considered to be the main domestic reservoir of *L. infantum,* which causes human VL, and also they serve to attract sandflies into houses. They have been implicated as reservoirs of *L. braziliensis,* and *L. guyanensis/L. panamensis* in South America (causing CL and MCL respectively). Domestic cats and opposums have also been implicated in transmission of *L. infantum* in urban areas [168–170]. However, the diversity of hosts with differing transmission abilities, has limited our understanding of reservoirs, particularly for CL [168].

Leishmaniasis in dogs is termed canine leishmaniasis, or CanL. Dogs can be infected with at least 12 species of *Leishmania* that also infect humans (*L. amazonensis*, *L. arabica*, *L. braziliensis*, *L. colombiensis*, *L. guyanensis*, *L. infantum*, *L. panamensis*, *L. major*, *L. mexicana*, *L. shawi*, *L. pifanoi* and *L. tropica*) [171–173]. CanL caused by *L. infantum* causes a VL-like disease resulting in weight-loss, muscular atrophy and skin lesions as well as dissemination of the parasite to multiple organs that can result in death [172]. Cutaneous CanL in dogs is mainly caused by *L. braziliensis* in South America. In dogs infected with *L. infantum*, parasites can be isolated from skin that does not contain lesions [174] and a majority that are infected with *L. infantum* and *L. braziliensis* can be asymptomatic [175,176], leading to dogs being targets of programs aiming to prevent transmission of *Leishmania*. These include using insecticide-treated collars to prevent vector biting and culling of dogs that test positive for *Leishmania* infection, although the latter approach has had little impact on incidence of human VL caused by *L. infantum*. Although no vaccines are available for human leishmaniasis, two vaccines are available for use in dogs (Leishmune and Leish-Tec) and these are expected to become more popular [172].

### 1.2.7 *Leishmania* genomes

Comparative analysis of *Leishmania* genomes has revealed that there is a strong degree of conserved gene content, synteny and chromosomal architecture with only a small number of differentially distributed (~200 to 400) genes and also few genes (~20) specific to one species. This is unprecedented considering the 20 to 100 million years of divergence within the *Leishmania* genus [177–179]. The approximately 8,000 genes in *Leishmania* spp. genomes lack introns so protein-coding genes are arranged in long clusters, where tens to hundreds of genes are arranged sequentially on the same strand, with the same direction of transcription; these clusters are termed polycistronic transcription units (PTUs) [180].

The number of chromosomes differs between *Leishmania* subgenera and species: *L. (Leishmania)* and *L. (Sauroleishmania)* have 36 chromosomes, with the exception of the *L. (Leishmania) mexicana* complex. Genomes in this complex have 34 chromosomes caused by fusion of chromosomes homologous to *L. major* chromosomes 8 and 29 to produce chromosome 8 of *L. mexicana* and fusion of chromosomes homolgous to *L. major* 20 and 36 forming *L. mexicana* chromosome 29 [177,181]. We have also demonstrated that *L. (Sauroleishmania) adleri* has 38 chromosomes due to fission of chromosomes 30 and 36 (Chapter 2). The *Viannia* subgenus has 35 chromosomes: its chromosome 20 is homologous to chromosomes 20 and 36 of *L. major* [181]. In common with other obligate parasites, *Leishmania* have more compact genomes (33 Mb vs 40 Mb), less than half the amount of genes in their closest free living relative *Bodo saltans* (< 8,400 genes vs 18,963) and have a more loose packing of genes caused by the to expansion of non-coding DNA sequences in *Leishmania* (~77% longer intergenic regions in *L. major*) [182–184]. Refinement of annotation and use of transcriptome sequencing may lead to a slight increase the number of

*Leishmania* genes, as in the case of *L. mexicana*, where transcriptome sequencing resulted in the annotation of 9,169 genes due to the discovery of 926 novel genes [185].

Genomes for 19 *Leishmania spp.* are available but are at different stages of completeness (scaffolds or chromosomes). For some species such as *L. donovani*, genomes from multiple isolates are available [186,187]. In the *Viannia* subgenus, the genomes of *L. braziliensis* [178], *L. panamensis* [188] and *L. peruviana* [189] are available as well as our contributions of the *L. naiffi* and *L. guyanensis* genomes (Chapter 3). In the *Sauroleishmania* subgenus, only *L. tarentolae* [190] had been assembled. We have expanded this with the addition of the *L. adleri* genome (Chapter 2). In the *L. (Leishmania)* subgenus, *L. major* [191], *L infantum* [178], *L. mexicana* [177], *L. donovani* [186] and *L. amazonensis* [136] have been published and others including *L. aethiopica, L. tropica (L. tropica* complex), *L. arabica, L. turanica, L. gerbilli* (*L. major* complex), *L. entriettii* (*L. enriettii* complex) as well as the *Paraleishmania* section species *Endotrypanum monterogeii* are being assembled by the Kinetoplastid Genomes Consortium (Bioproject accession PRJNA176381).

### 1.2.8 Transcriptional regulation in *Leishmania*

Trypansomatid species have a unique mechanism of gene expression among eukaryotes. PTUs are organised in a divergent manner, where the PTUs are transcribed towards the telomeres, or in a convergent manner. These PTUs are separated by A/T rich strand-switch regions (SSR) although there is no significant sequence similarity between SSRs on different chromosomes e.g. chromosome one of *L. major* has only two PTUs on opposing strands, one containing 29 genes and the other 50 genes, with a divergent SSR between them [192,193]. Within PTUs, genes are separated by intergenic regions rich in pyrimidine tracts that are important for mRNA processing [194]. tRNA and rRNA genes are clustered at the edge of convergent PTUs: tRNA genes are transcribed by RNA polymerase III and rRNA genes transcribed by RNA polymerase I [195,196]. In contrast to bacteria that possess polycistronic transcription in the form of operons to co-regulate functionally related genes, *Leishmania* PTUs generally contain genes with unrelated functions [191]. These PTUs are co-transcribed by RNA polymerase II (RNAPII) and the primary transcripts are trans-spliced into monocistrons by the addition of a 39 nucleotide spliced leader (SL) mini-exon cap at the 5’ end and in a coupled reaction, polyadenylation by the addition of a poly-A tail at the 3’ end of the mRNAs [197]. The SL sequence is added 30 to 100 bases upstream of the initiation codon resulting in short 5’ UTRs [198]. Transcription initiates in divergent strand-switch regions (and so is bi-directional) enriched in acetylated histone H3 [199] and terminates at strand switch regions in convergent PTUs [200]. A hypermodified modified

DNA base unique to Kinteoplastids, called base J, which is formed by hydroxylation and glucosylation of thymine bases, is found at convergent strand switch regions (RNAPII termination sites) and telomeric repeats and is necessary for proper termination of transcription at SSR termination sites, where it prevents read-through of transcriptional stops [201]. The SL-RNA gene is the only gene identified to be transcribed using an RNAPII promoter in *Leishmania* and is highly conserved in *Leishmania* spp. The polycistronic transcription of many genes and the absence of RNAPII promoters to regulate gene expression means that PTUs are constitutively expressed and that gene regulation occurs post-transcriptionally, via regulation of mRNA levels by mRNA stability and maturation controlled by 3' UTRs, rather than increases in RNAPII activity [202,203]. About 10-12% of genes expressed in both *L. infantum* and *L. major* change their expression between different life stages [204,205].

### 1.2.9 *Leishmania* have highly plastic genomes

*Leishmania* can change gene dosage through gene loss, duplication and amplification. Amplifications can occur either intrachromosomally as gene clusters of tandemly duplicated genes (tandem gene arrays) or extrachromosomally in the form of linear minichromosomes or circular episomes [177,206]. For instance, there are varying numbers of tandem arrays in each species with 200 in *L. major* Friedlin, 214 in *L. braziliensis* M2904, 132 in *L. mexicana* U1103 [177] and ~400 tandem gene arrays in *L. panamensis* PSC-1 [188]. Changes in the copy number of chromosomes (aneuploidy), as well as gene deletion and SNPs at drug targets and transporters have also been observed [207–209]. Tandem arrays, episomes and aneuploidy are intrinsic features of *Leishmania* that occur repeatedly in all tested conditions.

*Leishmania* are primarily diploid and aneuploid [177,186] and this is evidenced in cells from single cloned isolates (such as *L. major* using fluorescence *in situ* hybridization (FISH)) [210]. Analysis of read depth coverage demonstrates that chromosomes can display intermediate read depths indicative of mosaic aneuploidy [177]. Chromosome 31 (or 30 in *L. mexicana*) is most commonly tetrasomic. The reason for this is unclear but RNAseq analysis using *L. mexicana* revealed that this chromosome is enriched for genes that are upregulated in amastigotes [185] and it is also enriched for genes involved in iron metabolism [189].

Given that the amplification or deletion of chromosomes alters gene dosage (number of gene copies present), mRNA levels would be expected to tally with genetic mutations. However, comparisons of expression with chromosome copy number have shown that ploidy and mRNA levels are independent [211]. Aneuploidy in *Leishmania* can occur by asymmetric chromosome allotments during mitosis, where the total number of homologs in both dividing nuclei is an odd number e.g. 1+2 or 2+3. This was proposed to occur as a result of a defect in chromosomal replication followed by asymmetric segregation [210].

Extrachromosomal elements typically do not have a specific origin of replication (ORI) and tend not to be maintained in the absence of drug pressure [212,213], although there are exceptions especially in cases where they arise spontaneously [214–216]. *Leishmania* resistant to the antifolate methotrexate (MTX) amplify their dihydrofolate reductase-thymidylate synthase (DHFR-TS) gene on circular episomes which are formed by homologous recombination between direct repeats (DRs) that flank the amplified regions. In contrast, linear minichromosomes are formed by the annealing of inverted repeats (IRs) followed by duplications extending to the telomeres producing a minichromosome which consists of a large inverted duplication [216,217]. The repeat sequences (RS) used for DNA amplification are typically noncoding and highly conserved between species.  Over 5% of the *L. major* genome comprises these DR and IR sequences facilitating the potential formation of over 3,000 extrachromosomal DNA sequences and ~68% of RS in *L. major* are composed of Short interspersed degenerate retroposons (SIDERs). These are truncated versions of retroposons that can regulate gene expression at the post-transcriptional (SIDER2 family) level usually via mRNA degradation or at translation (SIDER1) and their presence may also facilitate copy number changes of genes via homologous recombination. Most (60 to 80%) of the predicted amplicons appear to already be present in parasite populations in laboratory cultures and are selected in response to drug pressure, increasing in abundance with increases in drug exposure and decreasing when the drug is removed. Thus, cells from a single isolate have a common core genome but have different gene copy numbers [213,217,218].

*Leishmania* was long considered to reproduce clonally but this was refuted by population studies [219,220] and the discovery of hybrid species such as *L. lainsoni*/*L. naiffi, L. donovani/L. aethiopica, L. major/L. arabica* hybrids [221–223]. However, the exact mechanism of genetic exchange has not been elucidated but a sexual cycle with meiosis and generation of haploid gametes has been proposed, as has a parasexual model (no meiosis) where haploid gametes are not produced and instead cells fuse with homologous recombination occurring at some point in this process. The parasexual model has some support, in that *Leishmania* promastigotes have previously been observed undergoing cell fusion [224]. Additionally, the mosaic aneuploidy observed in *Leishmania* makes it difficult to reconcile how gametes could be formed, given that all chromosomes would need to be reduced to a haploid state [225].

## 1.3 Methicillin-resistant *Staphylococcus aureus* (MRSA)

### 1.3.1 *Staphylococcus aureus* as a pathogenic bacterium

*Staphylococcus aureus* is a gram positive coccoid (round shape) bacterium that belong to the genus *Staphyloccoccus* (family *Staphylococcaceae* and phylum *Firmicutes*), which contains 80 species or subspecies [226]. It was discovered in 1880 by Sir Alexandar Ogston in Scotland in pus from an infected surgical wound. *S. aureus* are facultative anaerobes and are also known as 'golden staph' as they produce large (0.5-1.5 μm diameter), round, golden-yellow colonies when grown on blood agar plates. They also break down red blood cells underneath and surrounding the colonies (β-hemolysis) and grow in clusters or pairs, are non-motile and do not form spores [227].

Staphylococci are typically divided into coagulase-positive or -negative groups depending on their ability to clot blood, which is catalysed by the enzyme coagulase. Most staphylococci are coagulase-negative with the exception of seven, including *S. aureus* (the others are *S. intermedius, S. schleiferi* subsp. *coagulans, S. hyicus, S. lutrae, S. delphini,* and *S. pseudintermedius*), although some *S. aureus* defective in coagulase have been found in nature [228,229]. Coagulase-positive species are generally more pathogenic than coagulase-negative staphylococci [229,230]. *S. aureus* is also catalase-positive and this can be used to distinguish it from catalase-negative species. *S. aureus* and coagulase-negative *S. epidermidis* are common human commensals that asymptomatically colonise the skin and mucosae although *S. aureus* is mainly found in the anterior nares of the nose. Longitudinal studies, have established that approximately 20% of people have persistent nasal carriage of *S. aureus,* 30% have intermittent carriage and the other 50% do not carry it. Children have higher rates of carriage with at least 70% of newborn babies testing positive for nasal carriage of *S. aureus* at least once, although the rate drops to 21% within six months [231–233].

### 1.3.2 The origins of drug resistance in *S. aureus*

Before the advent of antibiotics, attempts at treating bacterial infections were limited to topical application of phenol and close to 80% of people with *S. aureus* blood infections (bacteremia) died [234]. The introduction of penicillin in the 1940s and its success at curing bacterial infections lead to US Surgeon General William H. Stewart declaring in the 1960s that "*it is time to close the book on infectious diseases and declare the war against pestilence won*" [235]. However, this was short-lived and evidence that strains could resist penicillin also appeared. This included reports of strains that could inactivate penicillin using the penicillinase (a type of β-lactamase) enzyme, just two years after the introduction of penicillin [234,236]. New β-lactam antibiotics were introduced but strains soon appeared that were resistant to these. β-lactamase enzymes hydrolyse the four atom β-lactam ring of β-lactam antibiotics and are encoded by the *blaZ* gene. This gene is generally found at a transposon on a plasmid and is under control of two genes: the antirepressor *blaR1* and the repressor *blaI*. The now widespread presence of penicillinase in *S. aureus* ( > 90%) has rendered penicillin almost useless for most infections [237,238].

Methicillin-resistant staphylococci were first reported in the United Kingdom in 1961 [239], soon after the introduction of methicillin, the first penicillin antibiotic that could resist β-lactamase.  Methicillin-resistant *S. aureus* (MRSA) is now defined as any strain of *S. aureus* resistant to β-lactam antibiotics e.g. penicillin, methicillin, oxacillin. Initial MRSA infections in the UK were largely limited to hospital outbreaks and gradually rose in frequency from the 1990s [240]. At present, MRSA causes more than 150,000 infections annually in the European Union [241].

### 1.3.3 Hospital-acquired and community-acquired MRSA

In the mid 1990s to 2000, MRSA emerged in otherwise healthy individuals, particularly young people, who had not been in recent contact with healthcare settings. These newly-identified strains were termed community acquired/associated MRSA (CA-MRSA) [242]. CA-MRSA is now globally disseminated [243] and can cause severe infections such as infective endocarditis and necrotizing pneumonia [244,245]. It also tends to have increased virulence [246] and a higher growth rate than hospital acquired/associated MRSA (HA-MRSA) [247]. In addition, it typically has the SCC*mec*IV [248,249] and more rarely SCC*mec*V, in comparison with HA-MRSA with SCC*mec* types I, II and III [250] – these are discussed in more detail in section 1.3.5.

CA-MRSA was initially defined by the Centers for Disease Control and Prevention (CDC) as MRSA infection in people who did not have the following risk factors in the year prior to infection: hospitalization, surgery, residence in a long-term care facility, percutaneous medical devices or in-dwelling catheters and dialysis, as well as those without previous MRSA infection, colonization or hospitalization >48 hours before MRSA culture [251,252]. HA-MRSA typically has resistance to multiple different classes of antibiotics whereas CA-MRSA, aside from their methicillin resistance [253], were generally more susceptible, although CA-MRSA strains have become increasingly resistant [245,254].  The distinction between HA-MSRA and CA-MSRA has also become blurred because strains associated with nosocomial infections also circulate in the community [255]. There is evidence that

CA-MRSA is invading hospital settings where they have become a common cause of bacteremias [256–259] and they may replace HA-MRSA in health-care settings [260]. CA-MRSA may have arisen by the HGT of SCC*mec* elements from MRSA or a coagulase-negative *Staphylococcus* into methicillin-susceptible *S. aureus* (MSSA) [261,262] with one study estimating that transfer of SCC*mec*IV into MSSA has occurred more than twenty times [263].

### 1.3.4 *S. aureus* typing and classification

Identification of *S. aureus* and examinations of diversity have used techniques such as pulsed field gel electrophoresis (PFGE) [264], multilocus enzyme electrophoresis (MLEE) [265,266], multilocus sequence typing (MLST) and genome sequencing. Other techniques include SCC*mec* typing and staphylococcal protein A gene (*spa*) typing [267,268].

PFGE compares the banding patterns (‘fingerprints’) produced by restriction enzyme digestion of DNA, producing fragments separated by size in a gel with an electric current. The United States PFGE database classifies clones using a USA100, USA200 etc system, the U.K. uses a EMRSA system, and Canada a CMRSA system [269]. However, it is difficult to compare results between laboratories and so has been replaced by DNA sequencing in many places.

MLST uses sequencing of 400 to 600 bp of fragments of seven housekeeping genes (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tip* and *ypiL* in *S. aureus*) amplified using PCR. Variation at each gene is combined to produce an allelic profile or sequence type (ST). STs that have the same alleles at least five loci are assigned to the same clonal complexes (CCs) and these CCs can be determined and visualised using software such as eBURST (http://eburst.mlst.net/). This approach facilitates a way to share results between different laboratories via the internet (http://saureus.mlst.net/) in a standard and reproducible manner [226,269,270]. 53 STs were distinguished using 61 CA-MRSA and 94 HA-MRSA isolates taken from patients in Oxford, U.K. from 1997 to 1998. Most were MSSA (81%) and only one of the MRSA isolates was CA-MRSA. One major clone identified was ST36, which had 22 MRSA isolates identical to the EMRSA-16 clone (as typed by PFGE), one of the most common MRSA clones in U.K. hospitals [270]. A later study increased the number of STs to 75 and CCs to 10 [271,272] so at present 21 STs associated with CA-MRSA have been identified [269].

The U.S. PFGE types USA100 are equivalent to CC5, USA200 to CC30, USA300 to CC8, USA400 to CC1 and USA600 to CC45 in the MLST scheme [273]. Additionally, genome sequencing of many CCs has determined that they are monophyletic and that CCs have less diversity within them (< 3,000 SNPs) than between them (> 15,000 SNPs) [274–276].

### 1.3.5 SCC*mec* and MRSA resistance

Resistance to methicillin is due to the acquisition of a MGE called the staphylococcal cassette chromosome *mec*. SCC*mec* elements consist of the *mecA* gene, a *ccr* gene complex, direct and inverted repeats (DR and IR) and joining regions (J1, J2, J3) and can range from 21 to 67 kb in size. The joining regions are nonessential components that consist of truncated transposon copies and pseudogenes although they can encode genes involved in resistance to other types of antibiotics and heavy metals. Eleven SCC*mec* types (types I to XI) have been described and are based on the type of *mec* and *ccr* complexes that they harbour [277–279] and subtypes are determined by the joining regions. The *mecA* gene encodes the low-affinity penicillin binding protein 2A, PBP2A (synonym PBP2'), a cell wall transpeptidase, which can continue to synthesise in the cell wall in conjunction with native PBP in the presence of β-lactam antibiotics [280].

SCC*mec* integrates into a specific chromosomal site called *attB* at the 3' end of the *orfX* gene [277,281] through recombination between the *att*SCC site on the circular SCC*mec* and the *attB* site. This results in the integrated SCC*mec* being flanked by the *attL* site within *orfX* and the *attR* site on the chromosome. The crystal structure of the *orfX* gene has shown it is similar to ribosomal methyltransferase. Even though there is a slight change in DNA sequence of *orfX* as a result of integration of SCC*mec* into *orfX*, its expression is unaltered, as the terminal amino acids and stop codon remain intact [282]. The integration and excision of the SCC*mec* element at the *orfX* integration site is mediated by genes (*ccrA*, *ccrB*, *ccrC*) on the cassette chromosome recombinase complex (*ccr*) which recognise DR or IR flanking *orfX*. Two *ccr* complexes have been described, one of which contains the *ccrA* and *ccrB* genes and the other containing only *ccrC*. Overexpression of *ccrAB* has been shown to trigger excision of SCC*mec* as a circular molecule [238,283].

SCC*mec* elements may also carry a gene encoding a sensory protein *(mecR1*) and a transcriptional repressor (*mecI*) gene, which regulate *mecA* expression although they are transcribed divergently from it. *The mecI-mecR1* regulatory system is homologous to the *blaR1-blaI* regulatory system for *blaZ* on the β-lactamase plasmid and the same DNA sequence is bound by their repressor genes (*mecI* or *blaI*) to activate *mecA* or *blaZ.* Additionally, one of either *mecI* or *blaI* must be functional in MRSA, possibly to prevent deleterious overexpression of the toxic protein [237,238]. The *mecA* gene, regulatory sequences and insertion sequences define the *mec* complexes, which are divided into four classes (A- D) based on the particular insertion sequences at the *mecR1* gene [281]. The *mec* gene complex likely originated in the *S. fleurettii*, a commensal of animals by combining with an SCC*mec* element lacking the *mecA* gene, potentially mediated by the *IS*431 insertion sequence element, which is present in the *mec* complex and is part of many composite transposons [284].

SCC*mec* elements can move between MRSA and MSSA strains in the laboratory via transduction, although the β-lactamase plasmid was required in the recipient strains. In one study, strains without the β-lactamase plasmid that received SCC*mec* expressed a lower level of *mecA* and had *mecA* mutations that hampered resistance, whereas those whose genetic background previously had SCC*mec* were able to maintain and express *mecA* to maintain high levels of resistance [285]. The requirement for a β-lactamase plasmid may be due to the fact that *blaR* on that plasmid can stabilise *mecA* expression [238]. SCC*mec* acquisition is confined to a small number of *S. aureus* lineages likely due to variations in sequences flanking the *attB* integration site in some strains [286] as well as restriction-modification systems [287] which prevent HGT between different lineages of *S. aureus.*

### 1.3.6 Resistance phenotypes

Although *mecA* is required for methicillin resistance, most cells expressing *mecA* have a heterogenous (or heterotypic) phenotype (HeR). HeR occurs if most cells in a culture grown from a single-cell inoculum, have low to moderate levels of resistance with minimum inhibitory concentration (MIC) values close to that of susceptible *S. aureus* and varying colony size. However, a subpopulation (0.01 to 0.1% of cells) have extremely high resistance (HoR) with high MIC values (several hundred μg/ml) and little colony size variation [288]. MIC is measured by exposing the bacteria to increasing amounts of antibiotic in standardized media and measuring the minimum concentration that growth is not detected at after a certain amount of time [289]. The HeR phenotype was originally discovered during an analysis of isolates from the first identified MRSA infection in 1961 [239]. Selection of HoR in the laboratory is usually achieved by growing HeR cells in high levels of oxacillin (≥ 100 μg/ml) [290]. The frequency at which HoR cells arise from a HeR population upon exposure to β-lactam antibiotics is reproducible, strain-specific and is measured using population analysis profiles (PAPs) where the number of bacteria that form colonies on agar plates are plotted against the concentration of antibiotic in the plates. The shape of the PAP is unique to particular clones [288,291–293]. Furthermore, HoR cells do not revert to HeR after passage in antibiotic free media [292], with some exceptions caused by chromosomal mutation [294]. In Chapter 4, we examine gene mutation, expression and CNVs associated with HeR and HoR cells and relate these findings to previous research in this area.

Another much rarer type of resistance is Eagle-type resistance in pre-MRSA: their *mecA* expression is strongly repressed by the *mecI* gene [295,296]. It has an unusual phenotype as it is resistant to high levels (64 to 512 μg/ml in strain N315) of methicillin but susceptible to low levels (2 to 16 μg/ml in strain N315). This is caused by the repression of *mecA* expression at low concentrations and its release at high concentrations. Experiments have shown that inactivation of the *mecI* gene causes Eagle-type resistance to be converted to HoR, as *mecA* is released from repression [297]. Missense *rpoB* mutations affect the encoded RNA polymerase β subunit product to cause HeR to HoR and HeR to Eagle-type conversion as well as cell wall thickening, decreased autolysis, prolonged doubling time and increased linezolid susceptibility (antibiotic for treatment of infections by gram positive bacteria that are resistant to other antibiotics).  In the case of Eagle-type resistance a *rpoB* mutation was hypothesised to enable the cell to tolerate the levels of methicillin required to enable *mecA* expression and in HoR it may combine with *mecA* to produce the HoR phenotype [298].

### 1.3.7 Persistence and tolerance

Aside from resistance, two other phenomena enable bacteria to survive antibiotic treatment. One is persistence (dormancy) in which a subpopulation (< 1%) of antibiotic-susceptible cells can survive but not reproduce (or only reproduce slowly) in the presence of the antibiotic. However, once the antibiotic is removed, they resume growth and when re-treated with the same antibiotic, the same heterogeneity in response appears instead of resistance, demonstrating that it is not heritable. This subpopulation of persisters can cause relapsing infections that require unusually long treatment regimens with multiple antibiotics to combat them, and they have a MIC similar to that of the susceptible strains [299,300].

Another phenotype that can appear is tolerance. This occurs when the bacterial population survives transient exposure to levels of antibiotic significantly above the MIC. Tolerant cells again have a low MIC, similar to the MIC of susceptible strains, but either stay in their dormant lag phase for a longer duration than unexposed cells (tolerance by lag) or have a slower growth rate (tolerance by slow growth). In tolerance induced by lag, once the antibiotic is removed the cells resume growth as normal. The killing rate of antibiotics such as β-lactams is correlated with the growth rate, so slowing growth results in fewer cell wall defects that cause cell death. In contrast to persisters, tolerant cells acquire mutations which either slow their growth or cause them to spend an extended amount of time in lag phase, although their MIC remains unchanged. Tolerance can also occur in strains that are naturally slow growing, and in the absence of inherited mutations if the cell is in an environment with poor conditions for growth [301].

Persistent, tolerant and resistant strains can be distinguished using their MIC and minimum duration of killing (MDK) values. The MDK is the duration of antibiotic treatment needed to kill a certain amount of the population at much higher concentrations than the MIC. Tolerant and persister cells need a longer duration of antibiotic exposure to be killed, whereas resistant ones have a much higher MIC. The MDK of 99% of cells in the population ($MDK_{99}$) is higher in tolerant cells compared with a resistant ones. The MIC and $MDK_{99}$ of persisters is similar to susceptible cells but the MDK for 99.99% of cells ($MDK_{99.99}$) is much higher [301].

### 1.3.8 Gene amplification in MRSA

Common mechanisms of resistance to antibiotics (or increasing concentrations) include HGT and inactivating mutations at genes targeted by the antibiotics. Bacteria can increase gene copy number in response to antibiotic stress from tetracyclines, macrolides, antifolates and β-lactams resulting in increased production of antibiotic-modifying enzymes, efflux pumps and molecules targeted by the antibiotic [302]. These amplifications tend to be unstable, and gene duplications may be present only in a subpopulation of cells.

Most studies on CNVs in bacteria have been carried out in *Samonella typhimurium.* These have found that approximately 10% of cells in non-selective culture have a gene duplication on their chromosome [303], most commonly between ribosomal RNA (rrn) operons that are directly repeated which facilitates homologous recombination (HR). Measurements of loss rates of amplifications have indicated that they are rapidly lost unless continuous selection is maintained e.g. growth in antibiotic. Loss of duplications and amplifications occurs due to HR between the amplified sequences although these can be stabilised by *recA* (a DNA repair and maintenance gene) mutations or selection. Loss can also occur if a subsequent mutation occurs that enables high levels of resistance independently of the amplification. In *S. typhimurium*, fitness costs (reduced growth rate) are independent of the duplication size, suggesting that these are determined by the genes at the duplicated region rather than the extra DNA copies. Given that variable gene copy number is relatively common in bacterial populations, selection for gene amplification instead of rarer SNPs, is likely the initial response to conditions that can be overcome by increasing the level of a gene product. Duplications can form in a RecA-dependent manner by unequal HR between 20 to 40 bp DR, or at the duplication point if there are short (< 20 bp) or no repeat sequences, followed by HR with the original repeat. RecA-independent processes include strand slippage during DNA replication or repair and pairing of single-stranded sections of sister chromatids at the replication fork [302].

In *S. aureus*, tandem duplications were discovered in clinical isolates taken from one patient during an infection [304] and we have also discovered (Chapter 4) an unstable amplification of the entire SCC*mec* element that arose after continuous culture of MRSA USA300 in oxacillin, as well as partial amplification of SCC*mec* and the ACME (arginine catabolic mobile element) element in other samples taken from this culture.

### 1.3.9 The *S. aureus* genome

*S. aureus* genomes are approximately 2.6 to 2.9 Mb, contain ~2,600 genes, and consist of one circular chromosome as well as various numbers of plasmids [305,306]. The first *S. aureus* genomes were published in 2001 from two clinical isolates, one resistant to methicillin (N315) and the other (Mu50) with decreased susceptibility to the glycopeptide antibiotic vancomycin. Both belonged to ST5 and had only a 0.08% nucleotide difference. Antibiotic resistance and virulence genes as well as a large amount of horizontally transferred elements were discovered such as transposons, bacteriophages and pathogenicity islands, highlighting the importance of HGT in shaping bacterial genomes [307]. Subsequently, the genomes of many MRSA and some MSSA strains have been published. These include HA-MRSA MRSA252, belonging to the EMRSA-16 clone that causes half of MRSA infections in the U.K. and is also a major U.S. clone (USA200). This was compared with CA-MSSA MSSA476, which was the cause of a severe infection of an immunocompromised child in the U.K. MRSA252 contained SCC*mec*II and the MSSA476 strain contained a novel 22.8 kb SCC element ($SCC_{476}$) integrated at the SCC*mec* site. $SCC_{476}$ contained a gene involved in resistance to fusidic acid, an antibiotic that inhibits protein synthesis in bacteria (through inhibition of elongation factor G) without killing them (bacteriostatic antibiotic), allowing the host immune system to eradicate them [308]. Other genomes include the genome of MRSA COL, an early methicillin resistant strain from the 1960s [305], the CA-MRSA MW2 strain [246] and the CA-MRSA USA300 strain FPR3757 [309]. In addition, the genome of the livestock-associated *S. aureus* ET3-1 revealed that some virulence factors such as antibody-binding protein A (encoded by the *spa* gene) and clumping factor A (*clfA*) are pseudogenes as a result of premature stop codons [310].

The increasing availability of *S. aureus* genomes has facilitated a wider comparison of genomes to reveal the ‘core’ and ‘accessory’ genomes.  The core genome is the set of genes that are present in all *S. aureus* (or nearly all) genome studied and excludes genes found on MGEs, whereas the accessory genome is the rest of the genes which have a patchy distribution across genomes. Most genes in the core genome are associated with central metabolism or involved in housekeeping functions [311]. There is also a strong shared synteny among genomes and high levels of nucleotide identity (> 97%) between genes on different genomes [306].

Genome sequencing has also enabled multiple studies examining the global population structure of single STs and CCs [312,313] and demonstrated its use in tracking hospital outbreaks, as well as in examining evolutionary history [314,315]. It has provided estimations of genome-wide mutation rate of *S. aureus* at $1.2\times10^{-6}$ to $2.0\times10^{-6}$ [226,316]. However, there is a strong bias for medically relevant strains: 94 % of genomes on Genbank are from only four CCs (CC5, CC8, CC30 and CC398) so the evolutionary history and structure of many CCs and STs remains unexplored [274].

### 1.3.10 *S. aureus* mobile genetic elements

*S. aureus* has a wide range of MGEs: these are DNA segments that mediate the movement of DNA within or between bacteria. These include SCC elements discussed earlier, plasmids, bacteriophages, transposons (Tn), insertion sequences (IS), pathogenicity islands and integrative conjugative elements (ICEs). This movement of DNA between bacterial cells (HGT) enables resistance, toxin and virulence genes to be passed to susceptible bacteri,a facilitating survival and colonisation in new conditions or environments.

HGT in bacteria occurs through three main mechanisms: transformation, conjugation and transduction [317]. Transformation is the transfer of DNA between bacteria through the uptake of free DNA from decomposing cells or DNA excreted from living cells (donors). The DNA is subsequently integrated into the recipient genome by HR, except in the case of plasmids that can replicate independently. For transformation to occur, the recipient cells must be competent and this is mediated by chromosomally-encoded proteins. *S. aureus* has low natural competence although subpopulations of cells can become competent at low frequency, and so most HGT in *S. aureus* is mediated by conjugation or transduction [317–319].

Conjugation is the transfer of genes on plasmids or ICEs that have genes encoding products that facilitate their transfer or the transfer of DNA on a plasmid from donors to recipients. It requires cell contact through a pilus to transfer DNA between cells [320].

Transduction is the transfer of DNA by bacteriophages (or phages), which are bacterial viruses that replicate independently. Phage genomes consist of DNA or RNA genomes. They contain replicase genes, genes encoding phage components that can take over the host cell machinery and genes encoding proteins that package DNA into a protein shell (capsid). Virulent bacteriophages replicate at a high frequency and lyse the host cell. Lysogenic bacteriophages integrate onto the host chromosome where they replicate with the rest of the chromosome as prophages, although in some cases phages can replicate independently as a plasmid. When these are activated by detecting the SOS response, they are excised from the chromosome (prophage induction) and resume the lytic cycle resulting in cell lysis, and sometimes, accidental packaging of host bacterial DNA in their capsid. This bacterial DNA is injected it into the next receipient where it can recombine and integrate into the chromosome. Transduced bacterial DNA must recombine with that of the recipient, so transduction can only occur within a species and can be restricted to certain lineages [320,321]. Phage transfer of DNA is common in *S. aureus*, including during infection and colonisation of people [322]. New phenotypes such as toxin production [323] can be conferred to the host bacterium by prophages (positive lysogenic conversion) and prophages can also recombine with other prophages or MGEs producing a mosaic phage structure. However, negative lysogenic conversion can occur when a phage integrates at a gene causing inactivation of that gene e.g. inactivation of β-hemolysin in *S. aureus* [324]. The inactivation of one gene is mitigated by the fact that these prophages can carry virulence factors such as Panton-Valentine leukocidin (PVL), enterotoxin A and exfoliative toxin A, or immune modulating proteins such as chemotaxis inhibitory protein of *S. aureus* (CHIPS), staphylococcal inhibitor of complement (SCIN) and staphylokinase (Sak) [321]. Prophages can mediate the excision of staphylococcal pathogenicity islands or move plasmids in the chromosome to other bacteria by transduction [321]. *S. aureus* typically have one to four prophages on their genomes. At least 80 genome sequences of phage and prophages that can infect *S. aureus* are available, most of which belong to the *Siphoviridae* family of temperate phages [306,320,325].

Staphylococcal pathogenicity islands (SaPIs) are MGEs, ranging in size from 12 to 27 kb, and are present in most *S. aureus*. They contain highly conserved core genes encoding transcriptional regulatory proteins, an integrase that recognises the integration site on the chromosome, a terminase and a replication initiation protein (Rep) and many also encode superantigens including enterotoxins B and C and toxic shock syndrome toxin (TSST), implicated in toxic shock and food poisoning [319,326,327]. Indeed, approximately half of the *S. aureus* toxins and virulence factors are on these islands and induction of the SaPI results in increased copy number of the genes [305]. At least 23 SaPI have been sequenced [328] and they are typically integrated site specifically into one of six attachment sites which contain short DRs on the chromosome (*att*$_{c}$). Each attachment site can be used by at least two SaPIs so if one site is deleted they tend to be able to attach at a different one. These sites are generally not used by other mobile elements, are at the 3' end of genes and are close to helper phages with which they share several functions: excision, integration and replication [327]. They do not encode genes for transferring themselves, so they rely on the phages to mediate transfer and are transferred at high frequently between cells so that they have a wide distribution in *S. aureus*. The transfer process is induced by stress which causes the helper phage to excise and replicate or can be induced by infection with a phage that is able to act as a helper. Once induced, the SaPI is excised and packed into small-phage like particles. These particles transfer to the recipient cells as the helper phage is transferring itself and once in the recipient cell they integrate into the chromosome [306,329–331].

Transposable elements are DNA sequences that can move ('jump') within a chromosome or between chromosomes. *S. aureus* transposons tend to be small and can be integrated in multiple copies into the chromosome, as well as into MGEs such as plasmids and SCC elements. They encode a transposase gene whose product catalyses the excision, replication and integration of the element and can carry resistance genes e.g. Tn*552* carrying *bla* for penicillinase.  They can also be horizontally transferred between bacteria on plasmids. Conjugative transposons (CTns), also known as ICEs, can mediate their own transfer from donor to recipient cells by conjugation. Small (< 2.5 kb usually) transposable element that only carry genes required for transposition are called insertion sequences. These do not harbour resistance or virulence genes, but they can inactivate genes by inserting into their coding sequence. A pair of IS sequences flanking accessory genes form a composite transposon [306,327].

A plasmid is a double-stranded DNA molecule that is smaller than the chromosome, can self-replicate and generally does not contain essential genes. They are usually circular, though linear plasmids have also been identified. Plasmids with the same replication machinery cannot be maintained in the same cell long-term – this is termed incompatibility (Inc). *S. aureus* clinical isolates typically contain at least one plasmid and there are three families of *S. aureus* plasmid (named I, II and III) which are based on the plasmid size and conjugation ability. Most plasmids encode resistance genes, some encode toxin genes and the function of others is still unknown (cryptic plasmid). Class I plasmids small (1.3 to 4.6 kb) and have high copy number (10–55 copies per cell), class II are larger (15 to 46 kb) and have lower copy number (four to six copies per cell) and class III range from 30 to 60 kb. Class I plasmids contain few (1 to 2 genes), usually carry resistance determinants or are cryptic, and some can be integrated into the chromosome including into mobile genetic elements on the chromosome e.g. an integrated plasmid, pUB110 carrying kanamycin and bleomycin resistance genes, is part of the SCC*mec*II element. Class II plasmids include most of those involved in producing penicillinase and these plasmids can also encode genes for resistance to antiseptics and heavy metals e.g. mercury or arsenate.  In many cases the genes are on transposons integrated into the plasmid. Most class I and II plasmids are likely transferred between bacteria via transduction. Class III plasmids are similar to class II plasmids except they carry a determinant of transfer (*tra*) which facilitates conjugative transfer of the plasmid between isolates at low frequency on solid surfaces, as they can be too big to be transferred by transduction. As with class II plasmids, they also carry resistance genes as well as transposons and insertion sequences [306,320,327].

# Chapter 2 - The genome of *Leishmania adleri* from a mammalian host highlights chromosome fission in *Sauroleishmania*

*The contents of this chapter are in review at Scientific reports as:*

*Coughlan, S, Mulhair, P., Sanders, M., Schonian, G., Cotton, J.A., Downing T. The genome of Leishmania adleri from a mammalian host highlights chromosome fission in Sauroleishmania*

*I performed all assembly and analysis in this chapter.*

*Peter Mulhair, Dublin City University, performed most of the manual correction of gene models for L. adleri HO174.*

*A submission is in progress at the European Bioinformatics Institute (EBI) for the genome sequence and annotation of L. adleri HO174.*

*Supplementary material for this chapter can be found in Appendix A.*

# 2.1 Chapter Overview

## 2.1.1 Aims and objectives

In this chapter, we sought to assemble and annotate the genome of a *Leishmania* sample from an Ethiopian rodent using only short read data. We first needed to identify the species of *Leishmania* using genomics based approaches. We then compared it's genome with those of other *Leishmania* species using alignment and read mapping to determine if they exhibited the same genomic architecture and extensive genome plasticity that characterise other *Leishmania* spp. analysed to date.

## 2.1.2 Methodology

*Leishmania* DNA from the Ethiopian rodent was sequenced on the Illumina HiSeq and produced 18,183,113 75 bp reads. An assembly and analysis pipeline (Figure 2.1) was developed that first performed quality control on sequencing libraries using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were then assembled of into contigs with Velvet [34] using a range of *k*-mer values with subsequent selection of the contig assembly with the highest N50, for contigs greater than 100 bp. Scaffolding was performed on contigs > 100 bp using SSPACE [40]. These scaffolds were improved using software that performed gapfilling (Gapfiller [51]) and base correction (iCORN [52]) via iterative read mapping. Misassembled regions were detected using REAPR [58] and scaffolds broken at these loci. These scaffolds were then aligned, ordered and oriented into pseudo-chromosomes using a reference genome (*L. tarentolae* Parrot-TarII [190]) with ABACAS [63] and all gaps greater than 100 bp were shortened to 100 bp. Unplaced scaffolds less than 1000 bp were removed as full gene models were unlikely to be recovered from these. Genes on the chromosomes and the retained bin sequences were annotated using the Companion annotation pipeline [67].

The species of *Leishmania* was determined by extracting seven genes from the *Leishmania* genome produced by this pipeline using BLASTn [29], concatenating them, and aligning the concatenated genes with those from multiple strains of *Leishmania* species using Clustal Omega [332]. This alignment was used to produce a neighbour-net network using SplitsTree [333] which indicated that the genome was *L. adleri*, a member of the *Sauroleishmania* subgenus. This identification was further confirmed by constructing a separate network using two concatenated genes from multiple *Sauroleishmania* spp.

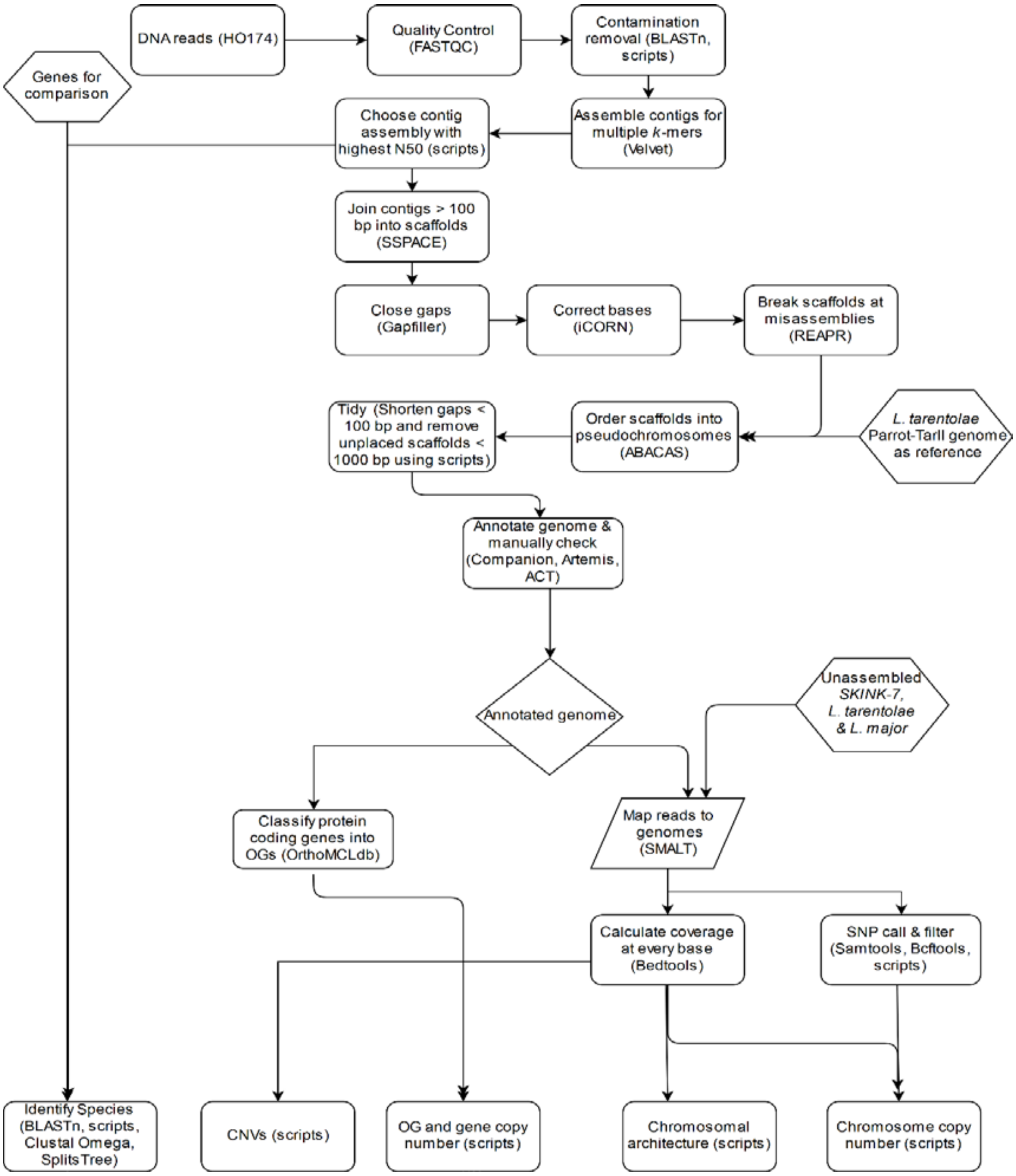

DNA reads (HO174)
Quality Control (FASTQC)
Contamination removal (BLASTn, scripts)
Genes for comparison
Choose contig assembly with highest N50 (scripts)
Assemble contigs for multiple *k*-mers (Velvet)
Join contigs > 100 bp into scaffolds (SSPACE)
Close gaps (Gapfiller)
Correct bases (iCORN)
Break scaffolds at misassemblies (REAPR)
Tidy (Shorten gaps < 100 bp and remove unplaced scaffolds < 1000 bp using scripts)
Order scaffolds into pseudochromosomes (ABACAS)
*L. tarentolae* Parrot-Tarll genome as reference
Annotate genome & manually check (Companion, Artemis, ACT)
Annotated genome
Unassembled *SKINK-7*, *L. tarentolae* & *L. major*
Classify protein coding genes into OGs (OrthoMCLdb)
Map reads to genomes (SMALT)
Calculate coverage at every base (Bedtools)
SNP call & filter (Samtools, Bcftools, scripts)
Identify Species (BLASTn, scripts, Clustal Omega, SplitsTree)
CNVs (scripts)
OG and gene copy number (scripts)
Chromosomal architecture (scripts)
Chromosome copy number (scripts)

**Figure 2.1:** Pipeline used in Chapter 2 for genome assembly, improvement, annotation and analysis. Sections of the pipeline that different to the pipeline used in Chapter 3 are in hexagonal shapes.

The chromosome copy number was determined using read depth coverage of chromosomes by mapping reads to the genome using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) and comparing the median coverage of each chromosome with the median coverage of all chromosomes, assuming that disomy was the most common chromosome copy number. SNPs were called using Samtools [103] and Bcftools, and filtered based on multiple criteria in order to reduce the number of false positives. The distribution of read depth allele frequencies for heterozygous SNPs was used to confirm the chromosome copy numbers and had the advantage of not assuming any particular chromosome copy number was common.

Visualisation of the coverage distribution across chromosomes from the *L. adleri* genome assembled here, and another sample using read mapping (*L. adleri* SKINK-7 [179]), was used to examine the chromosomal architecture. The copy numbers of genes were inferred using coverage of mapped reads, and large CNVs were identified by examining read coverage in 10 kb non-overlapping windows.

### 2.1.3 Conclusions

The research presented in this chapter has contributed to the field by providing evidence of two novel chromosome fission events that occurred independently in two strains of *L.adleri*, producing 38 chromosomes. The fission breakpoints preserve origins of replication (ORIs) on three of the four fission chromosomes and as discussed in section 5.1 of Chapter 5, these sites may also be centromeric sites, although further work is needed to determine this as well as the mechanisms involved in protecting the chromosome ends that lack telomeres. We have provided the first *L.adleri* genome which is annotated and composed of chromosome level scaffolds, demonstrated that the species of unknown *Leishmania* can be resolved using our phylogenomics based approach, and characterised chromosome number which showed that *L. adleri* and *L. tarentolae* (both *Sauroleishmania*) were aneuploid in common with other *Leishmania*. Analysis of gene copy number revealed that a putative virulence factor, elongation factor 1 alpha (*EF-1α*), had the highest copy number in *L. adleri*. A gene array with very high copy number in *Sauroleishmania* compared with species in the *Leishmania* subgenus was also discovered and domains present on these genes indicated that genes in this array may be involved in infection or signalling in *Sauroleishmania*, although further work is needed to determine the significance of the high copy number of this array in *Sauroleishmania.*

## 2.2 Abstract

Most *Leishmania* parasites infect insect vectors and mammal hosts, but *Sauroleishmania* are an exception because they primarily infect lizards. Here, we examined *L. adleri,* which infects both mammals and reptiles, and is associated with cutaneous disease in humans, but can be asymptomatic in wild animals. We assembled the *L. adleri* genome isolated from an asymptomatic Ethiopian rodent (MARV/ET/75/HO174) and verified it as *L. adleri* by comparing it with other *Sauroleishmania* spp. (*L. hoogstraali*, *L. gymnodactyli* and *L. tarentolae).* Chromosome-level scaffolding was achieved by combining *de novo* assembly, optimisation steps and reference-guided scaffolding to produce a final draft genome with contiguity comparable to other published *Leishmania* reference genomes. Annotation of the genome resulted in 7,959 gene models of which 95% were on chromosome level scaffolds. Comparison of *L. adleri* HO174 with another recently sequenced lizard-infecting *L. adleri* sample (SKINK-7) and with *L. tarentolae* Parrot-TarII, demonstrated extensive gene amplifications and two independent chromosome fission events in *L. adleri,* as well as pervasive aneuploidy in *Sauroleishmania*. There was little genetic differentiation between *L. adleri* extracted from mammals and reptiles, highlighting challenges for leishmaniasis surveillance. This is the first genome to be assembled from a mammal infecting species of *Sauroleishmania* and only the second assembled genome in the *Sauroleishmania* subgenus.

## 2.3 Introduction

### 2.3.1 *Sauroleishmania adleri*

*Sauroleishmania* are closely related to the *Leishmania* subgenus [334], with whom they shared a common ancestor ~42 million years ago (95% confidence interval of 24-65 million years ago) [179]. *Sauroleishmania* consists of 19 named and two unnamed species [123] (Figure 1.4), although few of these have been described [334], and are globally distributed with a range of sandfly vectors and animal hosts [335]. The phylogenetic position of *Sauroleishmania* between the mammal-infecting subgenera *Leishmania* and *Viannia* [137,336,337] suggests that it represents a lineage of *Leishmania* that switched from mammals to reptiles as their main hosts and *L. tarentolae* has been widely used as a non-pathogenic lab model because it rapidly replicates in lizards although it can invade human macrophages and may exist as amastigotes in mammals, with a slower replication rate [338,339]. In general, *Sauroleishmania* undergo development in the hind-gut of sandflies whereas *L. adleri* develops in the mid-gut of sandflies, in common with *Leishmania* subgenus spp. that can infect humans [340]. *L. adleri* can cause transient CL in humans [133] and cryptic infections in hamsters and mice that last up to five weeks [341].

### 2.3.2 Isolation of *L. adleri* HO174 in Ethiopia

MARV/ET/1975/HO174 was originally isolated in the Setit Humera district of north-western Ethiopia from an asymptomatic African grass rat (*Arvicanthis niloticus*) [342]. In this rural area, *Acacia* and *Balanites* forests, associated with *Phlebotomus orientalis* sandflies that feed on a range of hosts [343], are used as shelter for overnight sleeping [344]. VL is also endemic in Setit Humera [345]. HO174 was classified using multi-locus sequence typing (MLST) and multi-locus microsatellite typing (MLMT) as an unusual *L. donovani* lineage [346]. In Kenya, *L. adleri* is associated with the lizard- and mammal-feeding vector *Phlebotomus clydei* [340], whereas it is associated with *Sergentomyia* (a genus of Phlebotominae sanflies) *dentata* in Iran [347]. The African grass rat is a reservoir of several *Leishmania* species and is involved in transmission of *L. donovani* in Sudan [348] and Kenya [342], and *L. major* across Africa [348].

### 2.3.3 Leishmaniasis in Africa

Leishmaniasis is the most common neglected tropical disease in East Africa, as tropical and sub-tropical climates sustain the sandfly populations [349] that transmit it. One in eight people have undergone VL treatment in neighbouring Sudan [350], and there are an estimated 4,000 cases of VL annually in Ethiopia [351]. *Leishmania* spp. causing CL in this part of Africa are transmitted by *P. papatasi* or *dubosqui* sandflies, which are most frequently associated with *L. major* [352]. In contrast, *L. donovani* complex species, which cause VL, are transmitted by *P. orientalis* [344].

### 2.3.4 *Sauroleishmania* genomics

The only sequenced *Sauroleishmania* genome is for *L. tarentolae* RTAR/DZ/1939/Parrot-TarII that was isolated from the lizard *Tarentola mauritanica* [190]. It has 36 chromosomes, is both disomic and aneuploid, contains 8,530 genes, and gene-level orthology as well as PTU arrangement are conserved with *L. major* [190]. Here, we generated whole-genome sequence data for MARV/ET/1975/HO174 which we assigned as *L. adleri* on the bases of phylogenomic analysis. We used a combination of *de novo* assembly and reference-guided scaffolding to create an annotated-directed improved draft genome [74]. We used this *L. adleri* HO174 genome, along with short-read data from another *L. adleri* sample (RLAT/KE/1957/SKINK-7) isolated from a long tailed lizard (*Latastia longicaudata*) [179] and the genome sequence of *L. tarentolae*, to characterise *Sauroleishmania* spp., including genome rearrangements, chromosome structure and gene copy number.

## 2.4 Methods

### 2.4.1 Genome sequencing

DNA was sampled from an African grass rat (*Arvicanthis niloticus*), with no lesions on the 24th January 1975 in Humera, Ethiopia, a rural area with endemic Kala Azar (MARV/ET/1975/HO174). The sample was received by London School of Hygiene and Tropical Medicine from Liverpool University on 09/09/1980 (where it was also known as LV388) and subsequently the DNA used for sequencing was isolated at Charite University (Berlin), which was amplified with a Kapa HiFi DNA polymerase. The amplified paired-end short-read Illumina HiSeq 2000 library contained 18,183,113 75 bp paired-end reads with a median insert size of 400 and raw reads are deposited at the SRA (accession ERX180410).

### 2.4.2 Comparative Data

The *L. tarentolae* RTAR/DZ/1939/Parrot-TarII genome [190] and *L. major* Friedlin genome [191] as well as protein sequences and gff annotation files were downloaded from TriTrypDB version 6. 36 bp single end Illumina shotgun reads, originally used to correct the *L. tarentolae* genome assembly, were kindly supplied by the authors of [190]. 18,322,426 *L. adleri* 100 bp paired-end Illumina HiSeq 2000 reads (RLAT/KE/1957/SKINK-7), originally isolated from a long tailed lizard (*Latastia longicaudata),* in Kenya in 1957, were downloaded from SRA accession SRX764330 [179] and 12,680,080 76 bp paired-end Illumina Genome Analyzer II *L. major* Friedlin reads were downloaded from SRA accession ERX005636 [177].

### 2.4.3 Quality control

Quality control of the MARV/ET/1976/HO174 read library was carried out using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Additional sequence contamination was identified by viewing the distribution of mean GC content for the reads compared to the normal distribution in FastQC. To identify and remove potential contaminants, reads were searched against a the nucleotide database using blastn [29] (parameters: megablast, dust filtering on, percentage identity > 80 and evalue < 0.05). The blast hits were sorted by bitscore and E-value to keep only the top hit for each read. Read names and their corresponding genbank GI numbers were extracted from the filtered blast results and the taxonomic name for each sequence was retrieved using the blastdbcmd script and a copy of the ncbi taxonomy database (ftp://ftp.ncbi.nih.gov/blast/db/taxdb.tar.gz) in the BLAST+ [29] suite of tools with (parameters: GI numbers as input, lookup primary GI's only and taxonomid ID as output). The FASTQ files were then filtered to remove sequences not in the *Kinetoplastida* taxonomic class by the BLAST search. The resulting files were checked using FASTQC and any remaining reads not identified by the BLAST that deviated from a normal GC distribution were removed. The order of read-pairing was fixed and singletons removed using scripts from http://www.mcdonaldlab.biology.gatech.edu/seqtools_frame.htm#Instructions1.

### 2.4.4 Genome assembly

The QC processed reads were de-novo assembled into contigs using Velvet version 1.2.09 [34] with *k*-mer value of 53, a *k*-mer coverage of 16 and coverage cutoff of 8. A *k*-mer of 53 was used as it maximised the N50 value for contigs > 100 bp long, among *k*-mers ranging from 21 to 79. Contigs > 100 bp were scaffolded with the stand-alone scaffolder SSPACE v2.0 [40], which has yielded assemblies with fewer scaffolds and an improved N50 length value compared to the Abyss and SOAP assemblers [40]. This used the distance information from the paired-end reads to link contigs together and singleton reads for contig extension where possible first. SSPACE parameters were set to use a minimum of 50 overlapping bases with the seed/contig during overhang consensus build and to extend contigs using paired-end reads. All other parameters were at default. Gap closure was attempted with Gapfiller [51] which trimmed ten nucleotides off the edges of gaps, mapped reads to the scaffolds and found reads where one member of the pair matched a sequence region and the other one was partially within a gap. The reads that fell within the gap were broken into *k*-mers and overlaps between these were used to close the gap. This process was iteratively repeated for all gaps and ceased when no more gaps can be closed or it had reached ten iterations. Gapfiller parameters were set to have the minimum of 50 overlapping bases with the sequences around the gap and a maximum of 10 iterations. Erroneous bases were identified and corrected by mapping reads to the reference sequences over ten iterations using iCORN [52]. A plot of the number of corrections at each iteration was generated by summing the values of ‘INS’, ‘DEL’ and ‘SNP’ columns ( one column per iteration) in the ‘Stats. Correction.csv ’file produced by iCORN. REAPR [58] was then run on the iCORN corrected assembly to detect and break scaffolds at putative miss-assemblies. This mapped reads back to the assembly using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) to identify sites where paired-end read information disagreed with the scaffold information. Scaffolds were then broken at these sites. The broken scaffolds output by REAPR were then contiguated into 36 pseudo-chromosomes using ABACAS [63] with the *L. tarentolae* genome as a reference. The ABACAS union file of the reference genome needed to run abacas was created using ‘joinMultifasta.pl’ and the resulting pseudo-chromosomes were split up using ‘split2Multifasta.pl’ scripts from the post genome improvement toolkit (PAGIT) [49]. As true gap sizes on the pseudo-chromosomes were unknown, all gaps > 100 bp were shortened to 100 bp and gaps < 100 bp were left unaltered. A BLASTn search of the

200 bp at each unplaced scaffold edge against the 200 bp flanking all pseudo-chromosome gaps did not detect any additional scaffolds that could be placed onto pseudo-chromosomes. Finally, unplaced scaffolds (bin sequences) < 1000 bp in length were discarded. Pseudo-chromosomes for *L. adleri* were visualized along with the with the *L. tarentolae* reference [190] using the Artemis Comparison Tool (ACT) [55].

### 2.4.5 Kinetoplastid DNA assembly and annotation

Unplaced bin sequences were searched against BLAST databases of minicircles (753 sequences; kinetoplast AND minicircle AND leishmania ) and maxicircles (152 sequences; query : kinetoplast AND maxicircle AND leishmania) downloaded from Genbank (https://www.ncbi.nlm.nih.gov/genbank/) using MegaBLAST [29]. Hits were filtered to keep those with E value < 0.01 and bitscore > 100 to remove short hits and percentage identity > 40. Scaffolds that had homology to both mini- and maxicircle sequences were annotated in sequence headers by adding the '.kDNA.unassigned' tag to the ends of their headers while the others were annotated as '.kDNA.maxicircles' and '.kDNA.minicircles' accordingly

### 2.4.6 Phylogenomic characterisation

A multilocus sequence analysis (MLSA) style approach was taken to identify the genus and species of the MARV/ET/1976/HO174 genome. Nucleotide sequences of seven genes from 222 strains of ten *Leishmania* spp., isolated from infected patients, mammals and insect vectors [353], were downloaded from the NCBI (Accessions: KC158588: KC160141). Their homologs were extracted from the MARV/ET/1976/HO174 genome and *L. tarentolae* genome [190] using blastn [29] results filtered by E-value and bit score to obtain the top hit for each gene and genes were concatenated together to produce one sequence for each strain. The seven genes used were elongation initiation factor 2 alpha subunit (LbrM.03.0980; LaHO174_030250), spermidine synthase 1 (LbrM.04.0580; LaHO174_040600), zinc binding dehydrogenase-like protein (LbrM.10.0560; LaHO174_100590), translation initation factor alpha subunit (LbrM.12.0010; LaHO174_120010), nucleoside hydrolase-like protein (LbrM.14.0130; LaHO174_140120), RNA polymerase II (LbrM.31.2610; LaHO174_312240) and a hypothetical gene (LbrM.31.0280; LaHO174_310180)[353]. In order to include the *L. adleri* SKINK-7 sample in this analysis, the *L. adleri* SKINK-7 reads were assembled into contigs using Velvet [34] version 1.2.09 with *k*-mer value of 55 and expected coverage set to auto, yielding an assembly with 15,507 contigs and N50 of 4.88 kb. Orthologs of the seven genes were retrieved from the HO174 genome, *L. adleri* SKINK-7 contigs and the *L. tarentolae* version 6 TriTryDB genome using BLASTn and concatenated as before. Clustal Omega version 1.1 [332] was used to align the 225 concatenated sequences. A neighbour-net network of uncorrected p-distances was constructed with the alignment and visualised using SplitsTree v4.13.1 [333]. To further resolve the phylogenetic position of the isolate, two genes, DNA polymerase *α* catalytic polypeptide (POLA) (LaHO174_161460) and RNA polymerase II largest subunit (LaHO174_310170) were obtained for 5 *Sauroleishmania* species [334]. Their orthologs in the HO174 genome, *L. adleri* SKINK-7 contigs and the *L. tarentolae* TriTryDB v6.0 genome were retrieved using alignments with BLASTn as before and concatenated. The eight concatenated sequences were aligned and visualised as a neighbour-net network of uncorrected p-distances in SplitsTree v4.13.1 1 [333]. Sampling dates for the five sequences from [334] were retrieved from [354]. The number of substitutions between sequences was calculated by multiplying the uncorrected pairwise distance between sequences output by Omega version 1.1 [332] by the number of sites in the alignment.

### 2.4.7 Gene annotation

The *L. adleri* pseudo-chromosomes and bin sequences were submitted to the Companion webserver [67] developed by the Sanger Parasite genomics group to annotate genomes. Gene models were transferred from the *L. major* genome using RATT with species level transfer and ab-initio gene finding was also carried out by Augustus trained on *L. major* with a score of 0.8 to determine if a gene was predicted de-novo. Infernal aragon was used to predict tRNA, rRNA and ncRNA a part of the pipeline. Genes longer than 50,000 bases were discarded. Gene models and putative open reading frames > 450 bp identified by Artemis [355] were manually checked using Artemis [355] and ACT [55] by Peter Mulhair, DCU.

Candidate genes missed by Artemis were recovered by manually searching the genome for open reading frames (ORFs) > 450 bp in length. Genes which had a start and stop codon or which could be extended to the nearest start or stop codon were searched against the NCBI protein database using blastp. Genes with E-value < 0. 1 and percentage identity > 30% were considered homologous to the genes here and considered to true genes. Genes with multiple exons were caused by gaps, rearrangements, pseudogenes and stop codons in the gene body and these were corrected where possible by Peter Mulhair, DCU. Where genes extended over one or more gaps of unknown length (gaps $\geq$ 100 bp) in length, the gene was trimmed to the edge of the first gap. Genes with multiple stop codons were adjusted to the first stop codon.

### 2.4.8 Chromosome copy number

To calculate the chromosome copy number using a read depth of coverage approach, reads were mapped to the genome using SMALT version 5.7 (http://www.sanger.ac.uk/science/tools/smalt-0) with exhaustive mapping and a maximum insert size of 1000. Duplicate reads were removed using samtools rmdup [103] and the resulting BAM files were coordinate sorted. The read depth at every base was retrieved using bedtools 'genomecov' version 2.17.0 [356] on the sorted BAM files and the median coverage of every sequence in the assembly calculated (inclusive of sequence start and end coordinates). The median of chromosomal medians (bin sequences were excluded in both the *L. adleri* and *L. tarentolae* genomes) was then divided by the expected ploidy of two to produce a normalised median which represented the haploid chromosome coverage. Each chromosomal median coverage was divided by the haploid chromosome value to produce the copy number estimate for that chromosome.

The read depth allele frequency (RDAF) distribution of heterozygous SNPs was used to confirm the estimated somy of each chromosome. Allele frequencies were calculated by dividing the number of reads mapping to each of the four possible bases at a heterozygous SNP site by the total read depth at the site using samtools pileup v0.1.11 output. The allele frequencies were then binned into categories from 0.1 to 1.0 in steps of 0.05. For self-mapped reads, the number of allele frequencies in each category on each chromosome was counted and normalised by dividing the number of allele frequencies on that chromosome, producing a percentage number of SNPs in each category on each chromosome. For non-self mapped reads, homozygous SNPs (RDAF > 0.85) were excluded from the normalisation calculation to maximize the heterozygous SNP signal which would otherwise be difficult to visualise due to the large amount of homozygous SNPs. Read depth allele frequency distributions for each chromosome were plotted with R using the ggplot2 and gridExtra packages, omitting RDAFs of 0.1 and 0.15 from plots of non-self mapped reads to maximise the signal observed and omitting RDAFs > 0.85 in the case of self-mapped reads. This approach can be limited by the numbers of heterozygous SNPs observed, or by numbers of reads mapped to any location.

### 2.4.9 Copy number variation

To minimise potential false positives due to PCR duplicates and repetitive regions, the BAM files, produced by SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) were filtered as before to remove PCR duplicates and then filtered further to keep uniquely mapped reads by setting the mapping quality to >30 using samtools view [103]. Copy number over 10 kb intervals was measured for each chromosome by first creating a file of 10 kb non-overlapping windows across the genome using Bedtools ‘makewindows’ version 2.17.0 [356] with start and end coordinates of each chromosome or bin sequence supplied. In cases where the sequence was shorter than 10 kb, the window was the size of the sequence. The median coverage across each of these windows was calculated from files containing the read depth for every base and compared with the median coverage of the sequence it was on to produce a copy number estimate. Regions with $\geq$ 2 fold change from the median sequence coverage were reported. Median coverage was plotted in R using the ggplot2 package (median coverage was not filtered to remove uniquely mapping reads for plotting) and %GC content was also measured across the same windows and plotted to rule out changes in coverage due to GC bias. Each copy number variant was also manually checked and breakpoints refined by visualising the BAM files using IGV [90]

The copy number of each gene in *L. major*, *L. tarentolae* and *L. adleri* HO174 was estimated in the same manner using the gene coordinates but without removing multi-mapped reads from the BAM file. Assembled multi-copy genes have reads that map to more than one gene due to their similarity and so depleting multi-mapped reads would result in these genes having much lower coverage, which would result in inaccurate copy number estimates. However, for assemblies made using short read data this is less of a concern as highly similar genes are unlikely to be assembled into multiple copies due to the inability of the shorter read lengths to resolve the separate copies. Genes with copy number greater than or equal to two indicate that two copies of the gene are likely present (haploid copy number of two) but that only one was assembled (assembled number of one).

### 2.4.10 Variant calling and genetic divergence

To minimise false variant positives, repetitive regions and low quality regions of the genome (chromosomes and bin) were masked using three criteria. First repetitive sequences, homopolymers and tandem repeats were discovered using Tantan [357], then bases within 300 bases of the edges of assembly sequences and bases within 100 bases of gaps were marked and finally all three sets of sites were filtered out of the samtools mpileup candidate SNP files. Genomes were indexed with *k*-mer of 13 and step size of 2 using SMALT v5.7 (http://www.sanger.ac.uk/science/tools/smalt-0). Paired-end reads were mapped allowing for a maximum insert size of 1000 with exhaustive mapping enabled. SAM files were converted to BAM files; these were then coordinate sorted and PCR duplicates removed using Samtools v0.1.18 [103].

Candidate SNPs were detected where the base quality (BQ) was >25, the mapping quality (MQ) was >30, the read depth was <100 using Samtools mpileup v0.1.18, Bcftools v0.1.17-dev, and the Samtools 0.1.18 samtools.pl varFilter function [103]. Each candidate SNP was assessed using the following additional criteria on pileup files beyond that used in the SNP calling:

1) SNP Quality (SQ) >30
2) read coverage >5
3) forward-reverse read coverage ratio between 0.1 and 0.9
4) non-reference read allele frequency >0.1
5) 2+ forward reads
6) 2+ reverse reads
7) Outside a masked region

To remove SNP calls in low complexity and low quality regions, the masking file with sites to exclude was supplied to Vcftools version 0.1.12b [358], which was used to exclude positions where the reference allele in the SNP VCF file overlapped sites in the masking file (‘exclude-positions-overlap’ parameter) and produce a new filtered VCF file. Only SNPs which passed all the above criteria were retained. SNPs were considered heterozygous if RDAF >0.1 and RDAF < 0.85 and homozygous if RDAF $\geq$ 0.85. Variants were not called for *L. tarentolae* as base qualities of reads were not available.

In order to measure the genetic divergence of *L. adleri* SKINK-7 and HO174 with *L. tarentolae, the* number of homozygous SNPs per 10 kb non-overlapping window on each chromosome was calculated from the SNP results using the windows file produced by Bedtools previously. Graphs were plotted in R.

### 2.4.11 Gene ontology over-representation analysis

All available gene ontology (GO) terms for protein coding genes were extracted from the companion annotated gff file for *L. adleri* HO174 using a python script. These were supplied to the GOseq v1.18 R package [359] which uses the wallenius method to test for over-represented terms. GO terms with adjusted P < 0.1 using Benjamini-Hochberg correction were considered over-represented.

### 2.4.12 Identification of orthologous groups and arrays

Protein sequences for *L. tarentolae* (8,452 sequences) were downloaded from TriTryDB [360] version 6 and submitted to OrthoMCL [361] version 5 via their web-service. *L. adleri* protein sequences were also uploaded. This step excluded 44 *L. adleri* genes originally classified as pseudogenes by Companion: subsequent manual correction indicated these were valid protein coding genes. 11,825 orthologous groups (OGs) with associated sequence information were also downloaded from OrthoMCL by searching for groups based on a phyletic pattern using the expression ‘EUGL>=1T’ (at least one subtaxon from the Phylum *Euglenozoa* must be in the OG). 7,654 of these OGs contained gene(s) present in at least one of *L. major* strain Friedlin, *L. infantum*, *L. braziliensis* and *L. mexicana* species and 10,073 OGs contained gene(s) present in at least one of five Trypanasoma species genes: *T.vivax, T. brucei, T. brucei gambiense*, *T. cruzi* strain CL Brener and *T. congolense*. Results were parsed into tables and the copy number of each OG was estimated by summing the haploid copy number of each gene in the OG. Gene arrays in each genome were identified by finding all OGs with haploid copy number ≥ 2. Large arrays ( ≥ 10 gene copies) in *L. major, L. adleri* HO174 and *L. tarentolae* were examined and arrays with unassembled gene copies were identified by finding those with haploid gene copy number more than twice the assembled gene number.

### 2.4.13 Splitting fission chromosomes

Due to evidence of fission of chromosomes 30 and 36 in *L. adleri*, these chromosomes were divided into separate chromosomes at the fission breakpoints. The LaHO174.36 FASTA and EMBL file was split into two files, one for LaHO174.36.1 and the other for LaHO174.36.2. Gene and gap coordinates in the second EMBL file were adjusted by subtracting 989,797 (the length of LaHO174.36.1 with 99 bp added to account for a gap that was removed from end of sequence) from the original coordinates so that the coordinates in LaHO174.3.6.2 started at 1. Chromosome 30 of *L. adleri* HO174 was broken into two chromosomes (Chr 30.1 and Chr 30.2) at position 230,911 on the nucleotide sequence. Gene and gap coordinates in the second EMBL file (Chr 30.2) were adjusted by subtracting 231,011 (the length of LaHO174.30.1 with 100 bp added to account for the 100 bp gap that was removed from end of sequence) from the original coordinates so that the coordinates in LaHO174.30.2 started at 1. Sequence statistics were recalculated for the split sequences and added to the EMBL files along with the reformatted sequences. The resulting files were checked against the original chromosome 30 and 36 files using ACT.

### 2.4.14 Gene nomenclature

A series of python scripts was used to prepare the manually curated EMBL files for submission to the European Bioinformatics Institute (EBI) database. Locus tags and gene names produced by Companion were replaced with a system of identifiers where LaHO174 was the ID followed by an underscore, then the chromosome number or contig number and finally the gene number e.g. LaHO174_010110 for the first gene on chromosome 1 and

LaHO174_01bin0110 for the first gene on an unassigned contig. Gene IDs were incremented in steps of 10. *L. major Friedlin* orthologs for genes were retrieved from the gff file produced by companion where possible and added as a '/note' line for each gene in the EMBL files.

### 2.4.15 Detection of putative origin sites in *L. adleri*

MFASeq paired-end reads from early S-phase and the G2 phase of the *Leishmania major* Friedlin cell cycle (SRA accessions ERR688810 and ERR688812 respectively) were downloaded [362]. Reads were trimmed and adaptor sequence removed using Trimmomatic [23] with leading, trailing and 4 bp sliding window phred quality values of 30. These reads were mapped to both *L. major* Friedlin TriTryDB v6 genome and the *L. adleri* HO174 genome using Bowtie2 [80] version 2-2.1.0 with parameters as described in [362] (reporting up to one alignment per read and using the very-sensitive-local option). 2.5 kb non-overlapping windows were created using bedtools makewindows and the number of reads in each window counted using bedtools coverage with the counts parameter and the BAM and windows files as input. The number of reads in each 2.5 kb window was scaled by the number of reads mapped to chromosomes per million reads so that counts could be compared across cell-cycle stages. The ratio of reads per 2.5 kb window per million mapped reads in early S phase/G2 phase was then calculated for each window and plotted in R using ggplot2.

## 2.5 Results

### 2.5.1 Assembly of the draft genome of HO174

Preliminary analysis based on mapping sequence reads to existing reference genomes suggested that MARV/ET/1975/HO174 was a member of the *Sauroleishmania* subgenus. We *de-novo* assembled a genome for this mammal-infecting *Sauroleishmania* and formed chromosome level scaffolds by using *L. tarentolae* as a reference.

75 bp paired-end reads (17,644,995 reads total) remaining after the QC process (see Methods section 2.4.3) were *de-novo* assembled into 18,480 contigs with an N50 of 4.7 kb using Velvet [34], scaffolded using SSPACE [40] into 5,259 scaffolds with an N50 of 54.18 kb and gap-filled with Gapfiller [51] closing 55 % of the gaps (4,834/8,786 gaps) at that stage resulting in a drop in gaps and gap sizes (164,558 'N' bases were reduced to 48,655 'N' bases). REAPR [58] was used to document the improvement in N50's, corrected N50's and the percentage of error free bases at each stage (Table A1). Thousands of SNPs and small indel (< 3 bp) errors were corrected by running iCORN [52] for 10 iterations (Figure A1). Scaffolds were either broken at each of 627 errors predicted by REAPR if a gap was present or hard masked (bases replaced with N's) if the region did not have a gap. This resulted in 5,785 scaffolds with an N50 of 38.82 kb and 89.1% of the bases predicted to be error free. Error free bases are defined by REAPR as those which have at least five read pairs with the correct orientation and insert size mapped with no mismatches to the assembly and a sufficiently small fragment coverage distribution (FCD) error (discussed in Chapter 1). The percentage error free bases (% EFB) calculated by REAPR compares favorably with those for published *Caenorhabditis elegans* WS228 (90.3 %) *Plasmodium falciparium* v3 (94.9 %) and *Mus musculus* GRCm38 reference genomes (80.1%) [58].

The scaffolds were aligned and oriented into 36 pseudo-chromosomes using the *L.(Sauroleishmania) tarentolae* TriTryDB version 6 genome [190] as a reference using ABACAS [63]. As the true gap sizes of many gaps were unknown, all gaps were shortened to 100 bp reducing the total gap length from 2.93 Mb to 198 kb. An inversion of length 5,639 bp on chromosome 32 (breakpoints 59,906 bp and 65,545 bp) which harbored a putative protein kinase gene (LaHO174_320220) and was flanked by two gaps, was manually corrected so that the synteny of the gene was in agreement with the surrounding genes and the genes on the homologous locus on *L. tarentolae*.

4,104 short sequences (bin sequences) with a total length of 2.39 Mb that could not be incorporated into pseudo-chromosomes due to repeats or conflicting homology were filtered to keep only those with length greater than or equal to one kilobase. This reduced the final number of bin sequences (4,104 to 250 sequences) and also the final total bin length to 1.66 Mb. 17.56% of bin sequences had homology to kinteoplastid DNA (kDNA) – of these fourteen were annotated as minicircle kDNA (total length of 290,165 bp), one as maxicircle kDNA (1,075 bp in length) and one as unassigned kDNA (1,078 bp in length) as it had homology to both minicircle and maxicircle sequences.

The resulting draft genome for *L. adleri* HO174 had a length of 30.35 Mb with 94.5% of sequence in chromosomes and 69X median coverage (Table 2.1). HO174 had shorter average chromosome lengths although chromosomes 4, 8, 9, 15, 21, 24 and 28 were longer (Figure A2) and it had fewer genes on chromosomes (7,570) but more on unassigned contigs (389) than *L. tarentolae* (Table 2.1).

| **Genome statistics** | ***L. adleri* HO174** | ***L. tarentolae* ParrotII** |
|---|---|---|
| Number of chromosomes | 38 | 36 |
| All genes | 7,959 | 8,530 |
| Protein coding genes | 7,849 | 8,454 |
| Genes on chromosomes | 7,570 | 8,282 |
| Genes on bin contigs | 389 | 248 |
| Number of gaps | 4,350 | 4,568 |
| N content (%) | 0.64 | 3.77 |
| Chromosomes total length (bp) | 28,686,960 | 31,056,039 |
| Bin sequence total length (bp) | 1,664,372 | 578,687 |
| Genome length (bp) | 30,351,332 | 31,634,726 |
| GC content (%) | 56.76 | 56.66 |
| Median Coverage | 69 | 30 |

**Table 2.1:** Summary statistics for the *L. adleri* and *L. tarentolae* genomes, including unassigned (bin) contigs. *L. adleri* had two additional chromosomes due to the fission of chromosome 30 into 30.1 and 30.2 and chromosome 36 into chromosome 36.1 and chromosome 36.2.

### 2.5.2 HO174 is *L. adleri* in the *Sauroleishmania* subgenus

A multiLocus sequence analysis (MLSA) style approach was used to resolve the species of the MARV/ET/1976/HO174 isolate. Neighbournet analysis using Splitstree4 of uncorrected p-distances for this isolate, an *L. adleri* SKINK-7 isolate and *L. tarentolae* compared with 222 other sequences from [353] showed that MARV/ET/1976/HO174 was most closely related to *L. adleri* SKINK-7, which was isolated 18 years earlier from a lizard in Kenya, and then *L. tarentolae,* both members of the *Sauroleishmania* subgenus, and these three species were quite divergent from the seven Old World groups (*L. donovani*, *L. gerbilli*, *L. major*, *L. turanica*, *L. tropica*, *L. aethiopica*, *L. arabica*) (Figure 2.2(a)). There was a striking level of sequence similarity between the seven concatenated genes (4,677 sites) in HO174 and *L. adleri* SKINK-7 with only two substitutions between the isolates compared with 177 substitutions between both HO174 and SKINK-7 with *L. tarentolae* RTAR/DZ/1939/Parrot-TarII.

To refine the phylogenetic placement of HO174 within *Sauroleishmania*, a Neighbournet network was constructed using the alignments of two genes DNA polymerase *α* catalytic polypeptide (POLA) and RNA polymerase II largest subunit.  Eight species in the *Sauroleishmania* subgenus were used including two *L. tarentolae* isolates, RTAR/DZ/1939/TarVI (LV414) and the *L. tarentolae* Parrot-TarII as well as *L. adleri* SKINK-7 and an *L. major* outgroup, was used to further resolve the identity of the isolate (Figure 2.2(b)).  This determined that the HO174 isolate clustered with the two other *L. adleri* isolates and was most closely related to the SKINK-7 isolate confirming the species of HO174 as *L. adleri*. There were only two substitutions between HO174 and *L. adleri* SKINK-7 in the two concatenated genes for these isolates and both were equidistant from the other isolates in the network (Table A2).

Similar topologies were observed when individual networks were constructed for the POLA (Figure A3(a)) and RNA polymerase II largest subunit genes (Figure A3(b)) indicating that individual genes had the same signal as the concatenated genes. *L. adleri* SKINK-7 was again the most closely related to the HO174 isolate in the POLA and RNA polymerase II networks, (Tables A3 & A4). Each of the genes used in the analysis were also confirmed to be single copy genes in *L. adleri* HO174 using read depth coverage.

We also examined the genome-wide SNP rates using homozygous SNPs called from *L. adleri* HO174 and *L. adleri* SKINK-7 reads mapped to *L. tarentolae* and also SKINK-7 reads mapped to *L. adleri* HO174 to examine the genetic divergence of these samples with *L. tarentolae*. SNPs were called using samtools and extensive quality filtering and masking as described in Methods.  For *L. adleri* HO174, a total of 2,908,704 sites were masked out of results (1,865,904 repetitive bases called by tantan, 172,800 potential low quality bases around gaps and 870,000 potential low quality bases at sequence edges). For the *L. tarentolae* genome a total of 3,694,196 sites were masked out (1,975,331 from tantan, 810,600 sites near sequence edges and 908,265 sites at gap edges). Only five of 10,663 SNPs detected using self mapped *L. adleri* HO174 reads were homozygous (RDAF >=0.85) and 10,658 SNPs were heterozygous indicating that most bases in the assembly were correct. SNPs were not examined for *L. tarentolae* because base quality scores for reads were not available. These confirmed the older common ancestry of *L. tarentolae* with HO174 (999,834 SNPs or 35.8 SNPs/kb) and SKINK-7 (855,686 or 30.6 SNPs/kb) (Figure 2.3(a)) compared to that for HO174 and SKINK-7 (36,254 or 1.3 SNPs/kb - SKINK7 had 15,816 heterozygous SNPs), and SKINK-7 showed less divergence to *L. tarentolae* than HO174 (Figure 2.3(b)). We thus proposed that HO174 was a mammalian isolate of *L. adleri*.

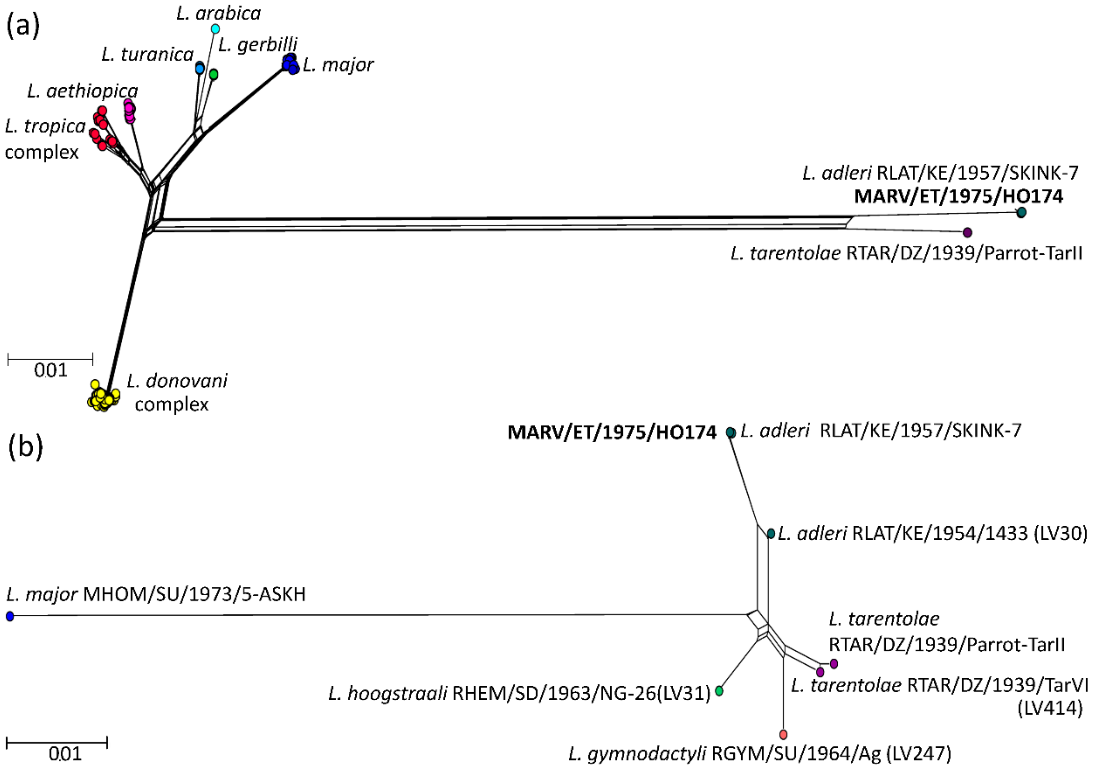

(a)
L. arabica
L. turanica
L. gerbilli
L. major
L. aethiopica
L. tropica complex
L. adleri RLAT/KE/1957/SKINK-7
MARV/ET/1975/HO174
L. tarentolae RTAR/DZ/1939/Parrot-TarII
001
L. donovani complex
(b)
MARV/ET/1975/HO174
L. adleri RLAT/KE/1957/SKINK-7
L. adleri RLAT/KE/1954/1433 (LV30)
L. major MHOM/SU/1973/5-ASKH
L. tarentolae RTAR/DZ/1939/Parrot-TarII
L. hoogstraali RHEM/SD/1963/NG-26(LV31)
L. tarentolae RTAR/DZ/1939/TarVI (LV414)
L. gymnodactyli RGYM/SU/1964/Ag (LV247)
0.01

**Figure 2.2:** *L. adleri* HO174 was a *Sauroleishmania* isolate based on: (a) alignment of seven concatenated genes with 4,677 sites for 225 strains are shown for a neighbornet network of the uncorrected p-distances. The scale bar indicates the number of substitutions per site. The *L. adleri* RLAT/KE/1957/SKINK-7 and MARV/ET/1975/HO174 nodes partially obscure each other. Compared to HO174, there are only two substitutions with SKINK-7, 177 substitutions with *L. tarentolae* RTAR/DZ/1939/Parrot-TarII, 635 with *L. major* MRHO/SU/1959/P-STRAIN, 627 with *L. infantum* MHOM/IT/1985/ISS175, 630 with *L. donovani* MHOM/YE/1993/LEM2677, and 599 substitutions with *L. tropica* MHOM/JO/1996/JH-88. (b) Alignment of two concatenated genes with 2,192 sites (genes encoding DNA polymerase $\alpha$ catalytic polypeptide and RNA polymerase II largest subunit) for six samples shown for a neighbornet network of the uncorrected p-distances. The scale bar indicates the number of substitutions per site. The SKINK-7 and HO174 nodes partially obscure each other. Compared to HO174, there are two substitutions with SKINK-7, 21 with *L. adleri* RLIZ/KE/1954/1433, 49 with *L. tarentolae* RTAR/DZ/1939/TarVI (from a *Tarentola* wall gecko) and *L.tarentolae* Parrot-TarII, 51 with *L. hoogstraali* RHEM/SD/1963/NG-26 (LV31), 55 with *L. gymnodactyli* RGYM/SU/1964/Ag (LV247) and 203 with *L. major* MHOM/SU/1973/5-ASKH. *L. adleri* SKINK-7 had the same number of substitutions as HO174 with each of these isolates.

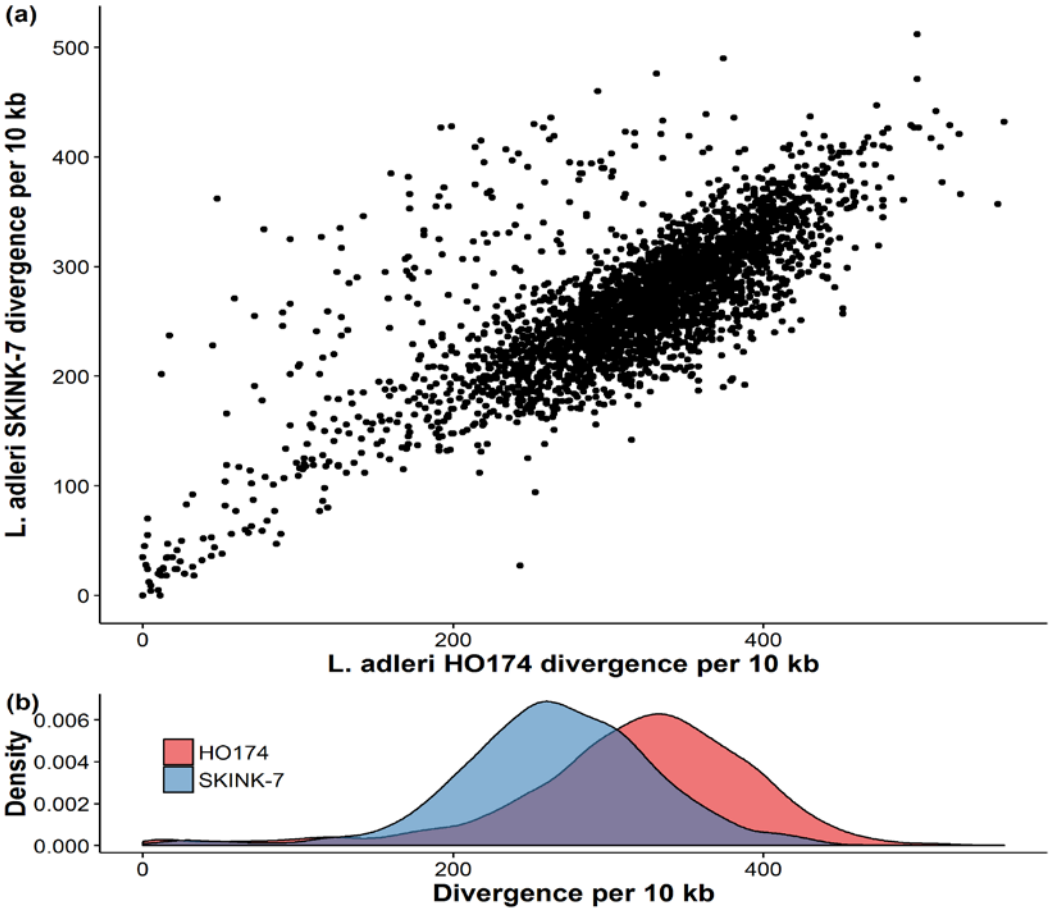


**Figure 2.3:** (a) Divergence of *L. adleri* HO174 and *L. adleri* SKINK-7 from *L. tarentolae* as the number of homozygous SNPs per 10 kb window on the genome. Loci with high divergence in both genomes are at the top right. (b) Density plot of divergence per 10 kb of HO174 and SKINK-7 indicated that HO174 was more divergent from *L. tarentolae* than SKINK-7.

### 2.5.3 Two ancestral *L. adleri* chromosome fission events produce 38 chromosomes

Evidence of two putative chromosome fission events was detected using read coverage of *L. adleri* HO174 mapped to itself, *L. adleri* SKINK-7 mapped to HO174 and both HO174 and SKINK-7 mapped to *L. tarentolae*.

The first proposed ancestral chromosome fission was identified for HO174 based on a sharp change in coverage after chromosome 36 base 989,698 (chromosome 36.1) with 62-fold median coverage that was 5' of a gap of unknown length (arbitrarily 100 bp). The median coverage increased from a uniform 62-fold to 94-fold median coverage at the end of a 100 bp gap (gap was from 989,698 bp to 989,797 bp on the original LaHO174 chromosome 36) between divergent gene clusters and was uniformly higher (at 94-fold median coverage) to the end of the chromosome (Figure 2.4(a)). This fission break separated two PTUs at a region homologous to LmjF.36.2560-LmjF.36.2570 (both hypothetical genes) and the coverage increase was unrelated to GC content ruling out GC content bias. There was also no pileup of mapped reads at the boundary of the gap which might indicate tandem duplication of this locus on the chromosome (Figure A4). This change in coverage was also evident when the HO174 reads were mapped to either the *L. adleri* (Figure A5) or *L. tarentolae* reference genomes (Figure A6) and absent for SKINK-7 reads mapped the *L. adleri* (Figure A7) and *L. tarentolae* reference genomes (Figure A8) as well as for *L. tarentolae* reads mapped to the *L. tarentolae* genome (Figure A9). This fission was supported by HO174 reads mapped to the *L. adleri* (Figure A4(a)) and *L. tarentolae* reference genomes (Figure A4(c)) which showed that no HO174 or SKINK-7 read pairs spanned this location when mapped to the HO174 (Figure A10(a)) or *L. tarentolae* (Figure A10(c)) chromosome. Consequently, the most parsimonious explanation was the existence of separate chromosomes 36.1 and 36.2 in *L. adleri* HO174 and SKINK-7 without a somy change for these chromosomes in SKINK-7. Thus, chromosome 36 was broken at the start of the 100 bp gap, with the section before the gap renamed as chromosome 36.1 with a length of 989,698 bp, and the section after the gap renamed chromosome 36.2, with a length of 1,599,953 bp. Read depth allele frequency (RDAF) distributions of heterozygous SNPs called from reads mapped to the chromosomes as well as depth of coverage analysis indicated that chromosome 36.1 was disomic and chromosome 36.2 was trisomic (Figures 2.5. & 2.6).

The second putative chromosome fission was identified for SKINK-7 chromosome 30 based on a marked shift in coverage when the SKINK-7 reads were mapped to *L. adleri* HO174 (Figure 2.4(b) & A4(b)) and *L. tarentolae* (Figures A4(d) &A8). The coverage dropped at the start of a gap (gap range was 230,911 to 231,011 bp) and remained uniformly lower across the rest of the chromosome. No HO174 or SKINK-7 read pairs spanned the breakpoint when mapped to the HO174 reference (Figure A10(b)) and the chromosome 30 break had no read pile-up. It occurred at a contig gap separating PTUs at a region homologous to LmjF.30.0710 (a cell division cycle 16 gene associated with mitosis) and hypothetical gene LmjF.30.0720. A single SKINK-7 read-pair crossed the breakpoint when mapped to the *L. tarentolae* one (Figure A10(d)) but it had a 57 kb insert size indicating that one read was incorrectly mapped (Figure A10(b)). The *L. tarentolae* chromosome also contained a homologous gap (Figure A11). The coverage change in SKINK-7 and lack of read pairs spanning the breakpoint in both HO174 and SKINK-7 reads mapped to HO174 provided evidence of a second fission creating chromosomes 30.1 and 30.2. Chromosome 30.1 spanned *L. adleri* HO174 bases 1-230,911 with 88-fold median coverage, and chromosome 30.2 spanned bases 231,011-1,197,246 (the end) with 43-fold median coverage (Figure A7). There was an increase in coverage at the last third of chromosome 30.1 which corresponded with a drop in GC content but this did not affect the calculated median value of 88 for this chromosome (monosomic chromosomes have a median read depth coverage of 21.5). Chromosome 30.1 was predicted to be tetrasomic and chromosome 30.2 is predicted to be disomic based on read coverage (Figure 2.5) and the RDAFs of heterozygous SNPs (Figure 2.6). No over-represented GO terms were identified using protein-coding genes from chromosomes 30.1, 30.2, 36.1 or 36.2 when the entire proteome or chromosomes 30 and 36 only were used as backgrounds.

In order to verify that putative origins of replication were at sites in *L. adleri* that were homologous to the predicted ORIs in *L. major,* 4,391,083 251 bp *L. major* MFAseq paired-end reads from early S-phase and 3,929,298 251 bp paired-end reads from the G2 phase of the *Leishmania major* Friedlin cell cycle [362] were downloaded from the SRA. 70.8% (3,108,560 read pairs) of early S-phase pairs and 73.8% (2,899,838 read pairs) of G2 phase pairs were retained after adaptor clipping and quality trimming. 92% of these *L. major* reads mapped to their own genome compared with only 36% of *L. major* reads to *L. adleri* limiting the coverage that could be obtained. Replicating S-phase/ non-replicating G2 phase ratios of *L. major* MFASeq promastigote derived reads mapped to *L. adleri* HO174 chromosomes exhibited the same trends at homologous positions (Figure A12) as for *L. major* mapped to itself in [362] indicating that the putative ORI sites in *L. major* are also present in *L. adleri*.

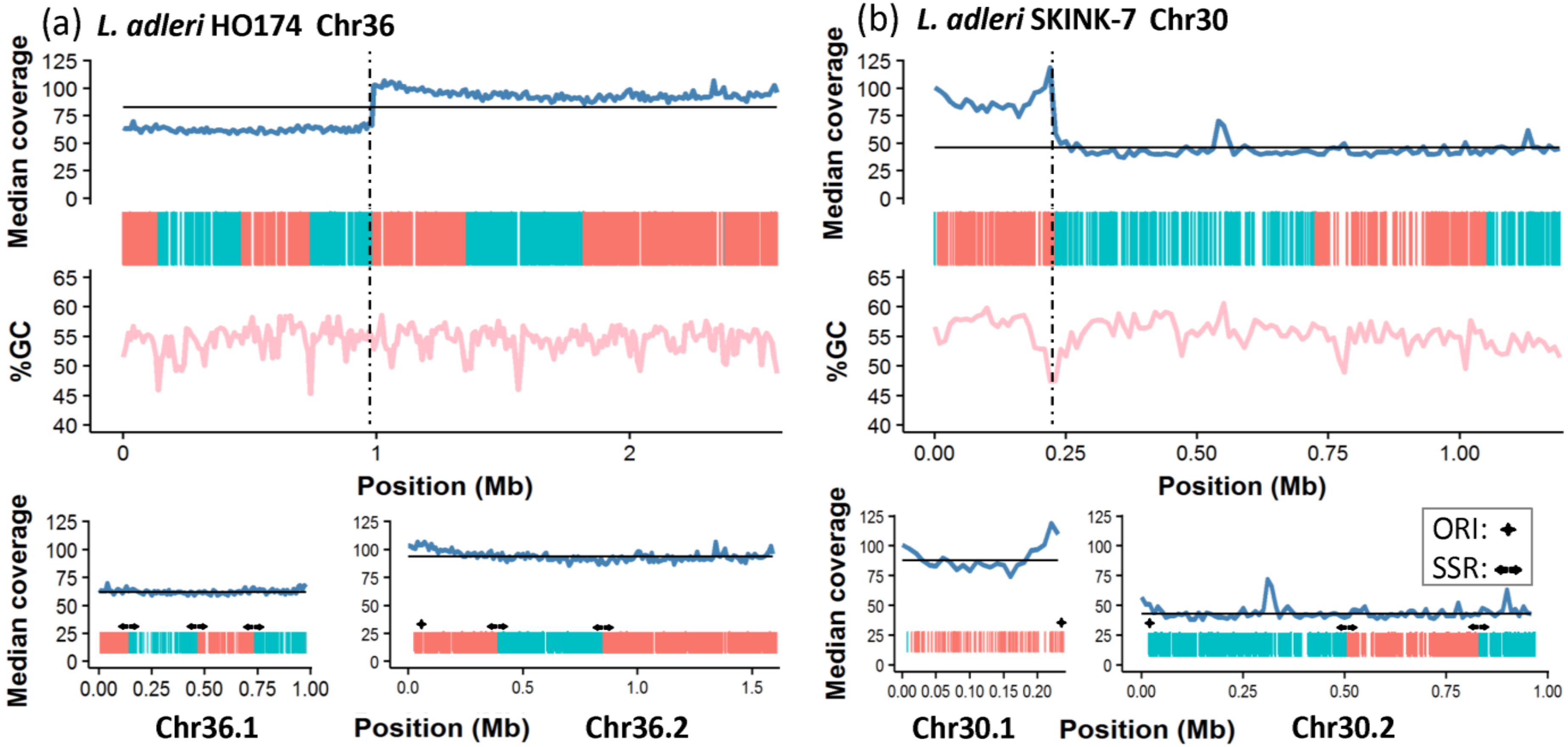


**Figure 2.4:** Evidence of chromosome fission of (a) *L. adleri* HO174 chromosome 36 into 36.1 and 36.2; and (b) *L. adleri* SKINK-7 chromosome 30 into 30.1 and 30.2. Median read coverage (blue) and GC content (pink) was measured in 10 kb blocks. Black horizontal lines indicate the median coverage of the chromosome. The dashed line indicates the fission breakpoints on the original chromosomes: at 989,697 for chromosome 36 and 230,911 for chromosome 30. Genes transcribed from left to right (green) and from right to left (red) are homologous to *L. major* polycistronic transcriptional units (PTUs) with their strand switch regions (SSRs) shown as arrows and the origins of replication (ORIs) shown as black crosses.

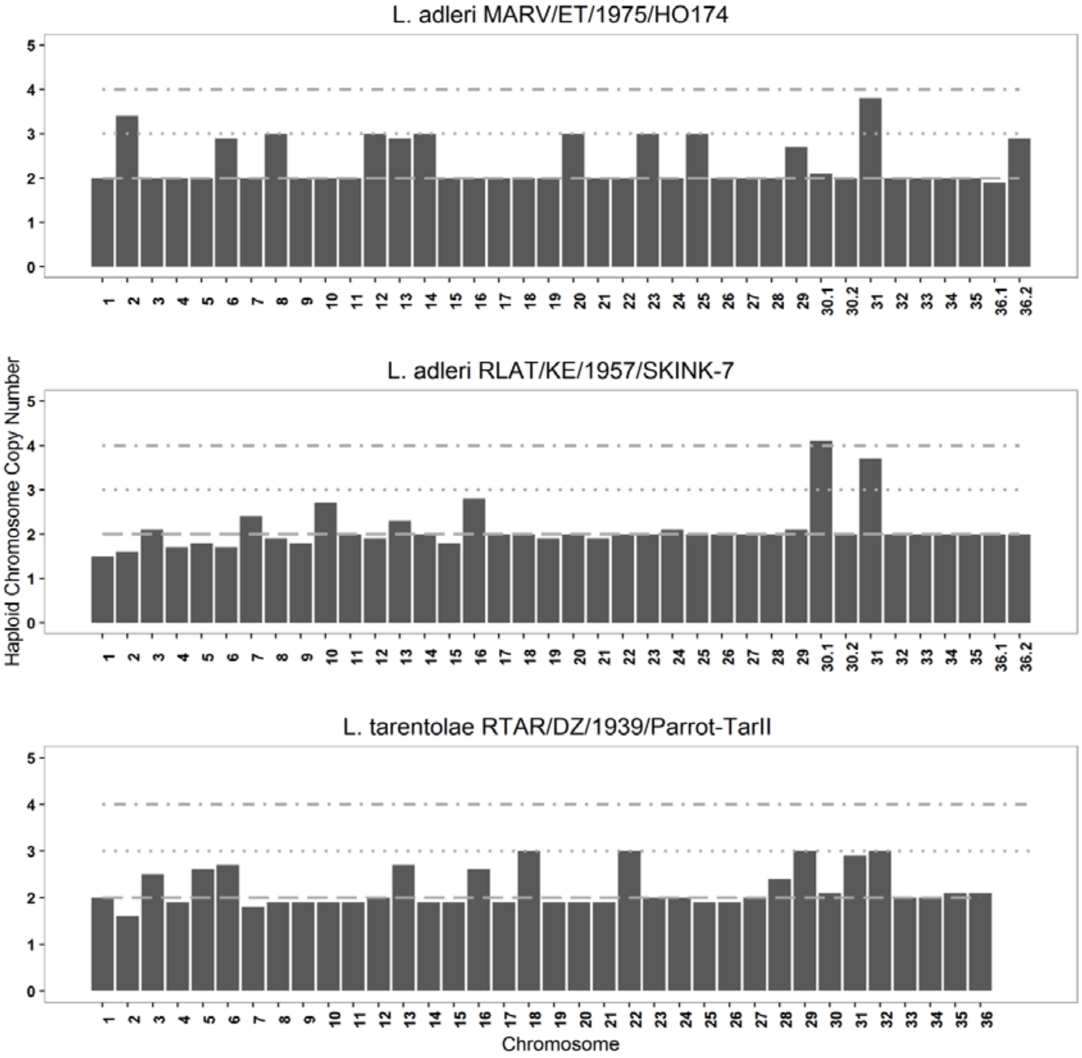


**Figure 2.5:** Chromosome copy numbers based on haploid median read coverage for *L. adleri* HO174 reads mapped to the *L. adleri* HO174 reference (top); *L. adleri* SKINK-7 reads mapped to the *L. adleri* HO174 reference (middle); and *L. tarentolae* mapped to itself (bottom).

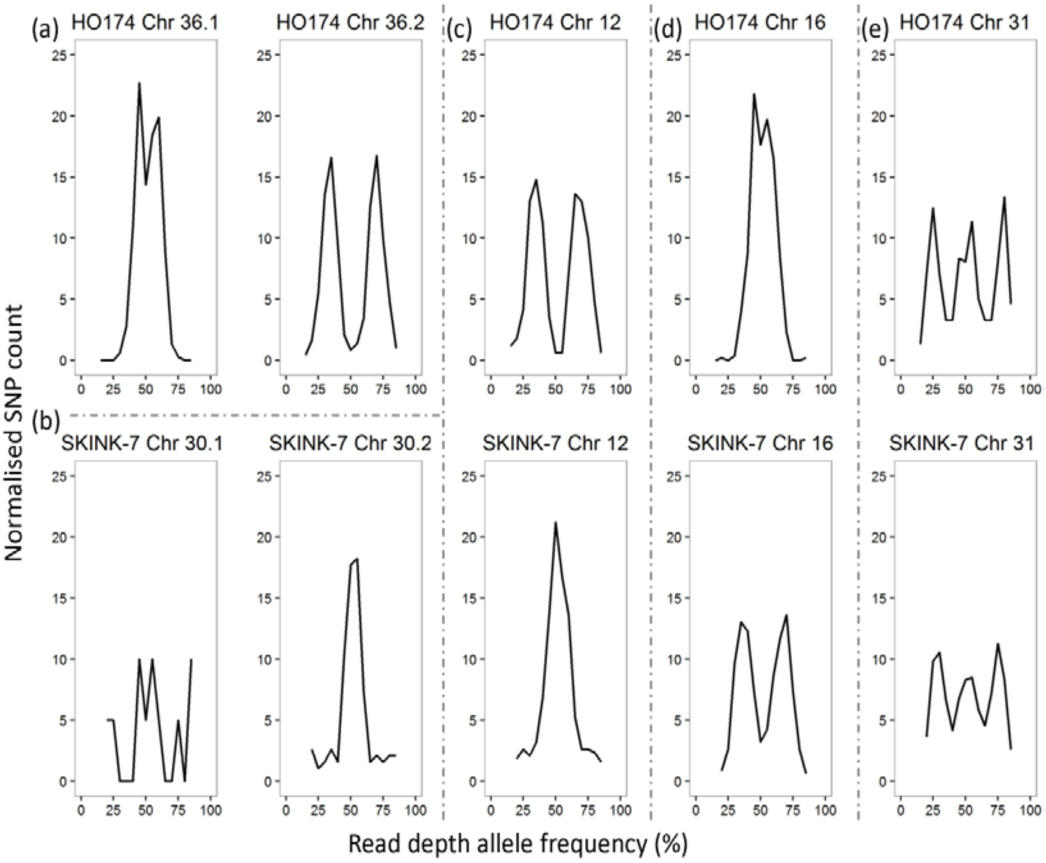


**Figure 2.6:** Read depth allele frequency (RDAF) distributions of normalised SNP counts for: (a) *L. adleri* HO174 disomic chromosome 36.1 and trisomic chromosome 36.2; (b) *L. adleri* SKINK-7 tetrasomic chromosome 30.1 and disomic chromosome 30.2; (c) chromosome 12 trisomic in HO174 and disomic in SKINK-7; (d) chromosome 16 disomic in HO174 and trisomic in SKINK-7; (e) tetrasomic chromosome 31 in HO174 and SKINK-7.

### 2.5.4 *L. adleri* and *L. tarentolae* were mainly disomic but aneuploid

Aneuploidy has also been demonstrated in a variety of *Leishmania* species and strains from different isolates and region [186,210,363]. Whole chromosome copy number variation was examined using the chromosomal median read depth, initially assuming diploidy [177,186,363]. However, this approach was limited by the need to set an expectation of disomy for normalisation. Another approach to determine chromosome copy number values uses the distribution of read-depth allele frequencies (RDAF distributions) for heterozygous SNPs. Alleles on disomic chromosomes should have approximately equal proportions of each allele (one peak at 50%), whereas those on trisomic ones have a majority of one allele, proportionate to the number of chromosome copies (peaks at 33% and 66%). Similarly, tetrasomic chromosomes will have peaks at 25%, 50% and 75% and pentasomic ones peaks at 20%, 40%, 60% and 80%. This approach can be limited by the numbers of heterozygous SNPs observed, or by numbers of reads mapped to any location.

Using the read coverage based approach, *L. adleri* HO174 was predominantly disomic but 10 chromosomes were trisomic (6, 8, 12, 13, 14, 20, 23, 25 and 29 and 36.2) (Figure 2.5) and chromosome 2 exhibited intermediate read depth values with a somy estimate of 3.4, suggestive of a mosaic cell population [210]. Chromosome 31 which was tetrasomic as expected from previous work that has illustrated its universal double-dose relative to the chromosomal average [177,186,363]. The median read depth of each chromosome was unaffected by GC content bias or local spikes in read coverage due to repetitive regions or local amplifications (Figures A5-A9). RDAF distributions plotted for each chromosome (Figure 2.5 and A13) supported the read depth coverage predictions and predicted that chromosome 2 was trisomic. A density plot of all heterozygous self-mapped SNPs also exhibited one peak close to an RDAF of 50% (Figure A14) indicating that *L. adleri* HO174 is predominantly disomic.

To examine if the same pattern of aneuploidy was observed in other sequenced *Sauroleishmania* spp., 100 bp paired-end Illumina reads from *L. adleri* SKINK-7 [179] were mapped to the *L. adleri* HO174 genome and 36 bp single end Illumina reads [190] were self-mapped to the *L. tarentolae*.

*L. adleri* SKINK-7 did not exhibit the same degree of aneuploidy as observed for *L. adleri* HO174. Only chromosome 16 was trisomic, chromosomes 30.1 and 31 were tetrasomic, chromosomes 7 and 10 were between di- and tri-somy and all others had values closest to disomy (Figures 2.5, 2.6 & A15). Chromosome 3 was predicted to be trisomic by the read depth allele frequency plots but closer to disomic by the median read coverage on chromosome 3 (copy number of 2.2).

*L. tarentolae* was also predominantly disomic based on the median read depth analysis (Figure 2.5); read allele frequency distributions could not be used to confirm this as SNPs could not be called based on our SNP calling criteria. Five chromosomes (18, 22, 29, 31, and 32) were predicted to be trisomic and five chromosomes (3, 5, 6, 13, and 16) had intermediate read depths (somy estimates of 2.5, 2.6, 2.7, 2.7 and 2.6 respectively). Interestingly chromosome 31 was predicted to be trisomic and not tetrasomic in *L. tarentolae*. Trisomy of chromosome 31 has only previously been reported in the predominantly diploid *L. donovani* LV9 (MHOM/ET/1967/HU3) strain [177] and *L. peruviana* LEM-1537 [189].

Extra chromosomes would have allowed more heterozygous SNPs to accrue over time, however there was no difference in the heterozygous SNP rate per 10 kb segment for SKINK-7 chromosome 30.1 versus 30.2. This was also true for HO174 chromosome 36.1 versus 36.2, suggesting the differences in somy were recent rather than long-term.

### 2.5.5 *L. adleri* HO174 genome annotation

A total of 7,959 genes were annotated on the *L. adleri* HO174 reference, of which 7,849 were protein-coding (Table 2.1). 7,570 genes were assigned to chromosomes (95.1%) and 389 to unassigned bin sequences. 7,845 (98.6%) of the 7,959 genes in total were annotated by Companion [67]. A further screen for candidate genes found 117 more genes, of which 110 had orthologs in *L. major*, one in *L. mexicana*, two in *L. infantum*: four genes without *Leishmania* orthologs encoded hypothetical gene products with homology to other trypanosomatid species (Table A5). Most of the manually discovered genes encoded hypothetical gene products, but two were ATP-binding cassette (ABC) gene family members: a duplicate *ABCA9* homolog (LaHO174_270850) and *ABCC2* (LaHO174_230250).

82 of the 103 RNA genes were tRNAs, which was only one fewer tRNA than *L. major* [191], fourteen were rRNAs, five were small nucleolar RNA genes (snoRNA) and two were small nuclear RNA genes (snRNA). Seven pseudogenes were also annotated. The number of snoRNA's predicted was much lower than that predicted for other *Leishmania* species such as *L. major* Friedlin, *L. infantum* JPCM5, *L. panamensis* PSC-1 and *L. braziliensis* M2904 which have 741, 50, 54 and 30 respectively [177,178,188,191]. SnoRNA genes were not annotated on *L. tarentolae* but Companion only predicted two and a test run using Companion with the published *L. braziliensis* M2904 genome with itself as a reference only found ten snoRNAs (rather than 53). Repeating this for *L. panamensis* PSC-1 using *L. braziliensis* M2904 as a reference found just three snoRNA genes (rather than 30) indicating that many snoRNAs are not discovered by that annotation pipeline. 92.5% (7,893 genes) of the gene number annotated on TriTryDB version 6 of the *L. tarentolae* genome was annotated in a test run of Companion using *L. tarentolae* chromosome and bin sequences (1,351 sequences) with *L. major* Friedlin as a reference indicating that it can discover most genes that have been found previously. Previously, *L. major* Friedlin was used as the reference to annotate *L. donovani* in a test of Companion, which found that 86% of *L. major* genes were perfectly annotated on *L. donovani* [67].

### 2.5.6 Comparative analysis of putative protein-coding orthologous genes

Protein-coding genes in *L. adleri* and *L. tarentolae* were categorised into orthologous groups (OGs) using OrthoMCL [361] for each species: 7,728 genes (98%) into 7,168 OGs for *L. adleri*; 8,113 (96%) into 7,368 OGs for *L. tarentolae* and 8,367 genes into 7,519 OGs for *L. major* (Table A6). OGs contain both orthologs and paralogs and range from single genes or gene subfamilies to whole gene families depending on the diversity of the gene family. In addition, each gene can only be assigned to one OG. 98% of *L. adleri* genes had orthologs in *L. major* (subgenus *Leishmania*) and *L. tarentolae*, indicating high gene content conservation (Figure 2.7). Previously, 250 genes were described as absent in *L. tarentolae* Parrot-TarII but present in *L. major* [190]. Analysis using OrthoMCLdb v5 and the *L. tarentolae* TriTrypDb v6 proteome found 280 protein-coding genes in 203 OGs absent in *L. tarentolae* Parrot-TarII but present in *L. major* (Table A6): 32 had orthologs in *L. adleri* HO174 (Table A7).

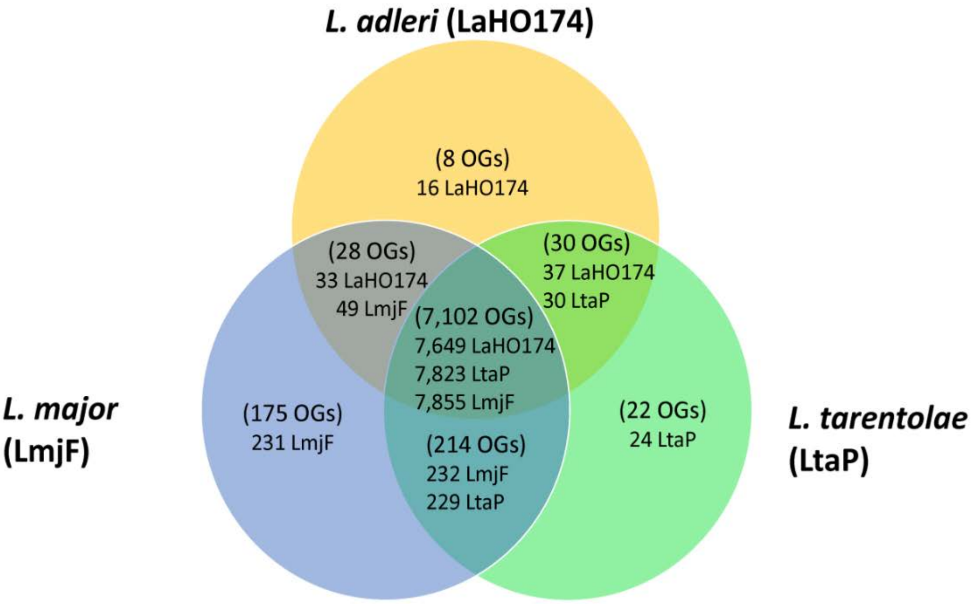


**Figure 2.7:** Numbers of genes either unique to or with orthologs in each *L. adleri* HO174, *L. major* Friedlin and *L. tarentaole* Parrot-TarII determined using OrthoMCL v5 orthologous groups (OGs). The OGs are in parentheses. The number of genes in *L. major* OGs are denoted by LmjF, *L. tarentolae* OGs by LtaP, and *L. adleri* OGs by LaHO174.

#### 2.5.6.1 Genes exclusive to *L. adleri*

Sixteen *L. adleri* genes had no orthologs in *L. tarentolae* and *L. major* (Table A8, Figure 2.7). Of these, four had orthologs in at least one of *L. infantum*, *L. mexicana* or *L. braziliensis*, and three had orthologs in one of the five *Trypanosoma* species (*T.vivax*, *T. brucei*, *T. brucei gambiense*, *T. cruzi* strain CL Brener and *T. congolense*) but not in *L.*

*major, L. infantum, L. mexicana, L. brazilinensis* or *L. tarentolae*. Nine had no orthologs in the five *Leishmania* and five *Trypanosoma* species mentioned above but eight of these had domains orthologous to variant-specific surface protein genes in parasites such as *Giardia, Entamoeba* and *Trichomonas vaginalis*, in which their protein products undergo antigenic variation to evade host immune responses and facilitate host adaptation [364]. All eight had top hits of 35-38% sequence identity with an unnamed product from *Phytomonas sp.* isolate HART1 - trypanosomatids from this genus can infect plants via an insect vector [365] (Table A9).

#### 2.5.6.2 Genes with orthologs in *L. adleri* and *L. major* but not *L. tarentolae*

There were 32 genes with orthologs in both *L. adleri* and *L. major* that were absent in *L. tarentolae* (Table A7). Four of these encoded a serine/threonine-protein phosphatase PP1, a folate/biopterin transporter, a protein kinase and DNA polymerase kappa and the rest encoded hypothetical proteins. The nucleoside diphosphate kinase B (LaHO174_323240, Table A7) gene absent in *L. tarentolae* had five copies in HO174 compared to one in *L. major*, three each in *L. infantum*, *L. braziliensis* and *L. mexicana* [177], and two in *L. panamensis* PSC-1 [188]). Overall, most (22/32) of these were also present in *L. infantum*, *L. braziliensis* and *L. mexicana* demonstrating that most of them are common to other *Leishmania* genomes.

A chromosome 19 gene array encoding autophagy-related protein 8 (ATG8/AUT7/APG8/PAZ2, OG5_137181) involved in endocytic trafficking and recycling [366] may be absent or partially assembled in *L. tarentolae* because it had two genes, a gap and collapsed repeat (Tables A6 & A7). This orthologous group contained all eight *L. major* genes assigned to the ATG family named ATG8B [366] and also gene LmjF.19.0910 with a total of nineteen predicted haploid copies in *L. major* and nine haploid copies (two assembled genes: LaHO174_190780 and LaHO174_190790) in *L. adleri*. The ATG8B family is one of four gene families, associated with autophagy, that are present in *L. major* (the others are ATG8C and ATG8) and the ATG8A and ATG8B families are unique to *Leishmania* [366].

#### 2.5.6.3 Genes with orthologs in both *L. tarentolae* and *L. adleri* but not in *L. major*

Of the 30 *L. adleri* genes with orthologs in *L. tarentolae* but not in *L .major*, eighteen had orthologs in another *Leishmania* (Table A10). Twelve *L. adleri* genes had orthlogs in *L. tarentolae* that were absent from the *L. major*, *braziliensis, mexicana* and *infantum* but were present in at least one of six *Trypanosoma* genomes. All except two were hypothetical proteins. The two genes with assigned functions encoded a putative Sedlin, N-terminal conserved region containing protein and the other encoded a putative anaphase promoting complex subunit 11 (*Apc11*) which contained a RING finger domain.  The anaphase-promoting complex/cyclosome (APC/C) is a multi-subunit E3 ubiquitin ligase consisting of 11 to 13 subunits that marks cell cycle regulator proteins and mitotic cyclins for degradation by the proteasome, initiating transition from metaphase to anaphase and promoting exit from mitosis [367].  The activity of  APC/C is regulated by phosphorylation and association with two regulatory proteins – *Cdc20* and *Cdh1*. *Cdc20* homologs have previously been reported in *L. donovani* , *L. infantum* and *L. major* [368] and several APC subunit homologs have previously been identified in *T. brucei* including *Apc11* [369]. A putative cell division cycle protein 20 (*Cdc20*) gene was also at chromosome 24 in both *L. adleri and L. tarentolae* (LaHO174_241770 and LtaP24.1870) indicating that ubiquitin-dependent proteasomal degradation may be involved in their cell cycles.

#### 2.5.6.4 Genes exclusive to *L. tarentolae*

A total of 24 assembled genes were absent in *L. adleri* and *L. major* but present in *L. tarentolae* (Table A11). Ten of these were present in at least one of *L. infantum***,** *L. braziliensis* and *L. mexicana*. The other fourteen genes were exclusive to *L. tarentolae* among *Leishmania* spp. but ten of these had orthologs in *Trypanosoma* spp. and these included two GP63 pseudogenes (LtaP10.0550 and LtaP10.0570), an expression site-associated gene (LtaP24.1490), a putative zinc-finger protein (LtaP26.0070) and a gene encoding ‘High cysteine membrane protein Group 2’ with orthologs in *Giardia* parasites (LtaP11.1290). Four other genes encoding surface antigen like protein (LtaP11.1290), malate dehydrogenase (LtaP34.0490) and Ser/Thr protein phosphatase family proteins (LtaP34.0940 and LtaP34.0890) had no *Leishmania* or *Trypanosoma* orthologs. One gene, LtaP27.2450 which had 8 orthologs in *T. congolense* had high copy number in *L. tarentolae* (30 haploid copies) (Table A11) but no domains were discovered that could be used to infer its function.

### 2.5.7 Gene arrays and copy number variation

The haploid copy number of each gene in *L. adleri* HO174, *L. tarentolae* and *L. major* was determined using read depth analysis. Gene arrays were identified using the haploid copy number of the genes and arrays were defined here as OGs that contained 2 or more haploid gene copies. Thus, gene arrays could contain either tandemly duplicated genes or genes distributed across multiple chromosomes/bin contigs. *L. adleri* had 295 gene arrays (Table A12), *L. tarentolae* had 281 (Table A13) and *L. major* had 289 arrays (Table A14). 62 arrays in *L. adleri*, 119 arrays in *L. tarentolae* and 12 in *L. major* had an array copy number (Tables A15- A17) predicted by read depth analysis to be more than two fold higher than the assembled number, indicating that *L. major* has most of its repeat gene copies resolved in its assembly. If all copies of a gene were assembled the haploid copy number would match the assembled copy number.

### 2.5.7.1 *L. adleri* copy number variation (CNV)

Six CNVs were discovered in *L. adleri* HO174 ranging in size from 5.7 kb to 19.8 kb of which four were also identified in *L. adleri* SKINK-7 (Table 2.2). The two CNVs unique to *L. adleri* HO174, included one with no genes. The other 15.9 kb CNV was present in two to three copies at chromosome 27 and had 3 genes. Two of the genes encoded ATP-binding cassette subfamily A members and the other encoded a putative cysteine peptidase, Clan CA, family C2 protein which has a calpain-like domain. Calpains are involved in cytoskeletal remodelling and signal transduction and are thus important for cellular remodelling during *Leishmania* differentiaton [370]. There are also 3 copies of this gene in *L. panamensis* PSC-1 [188] and orthologs in *L. major* and *L. infantum* are differentially up-regulated in log-phase promastigotes in vitro [205] as well as in antimony resistant *L. infantum* promastigotes compared to wild-type promastigotes [371]. Calpain inhibitors have also been tested in vitro for *L. amazonensis* as potential *Leishmania* treatments [372].

A 10.9 kb-long chr10 CNV with two to three copies in HO174 and SKINK-7 spanned three genes encoding a phosphate-repressible phosphate permease-like protein, a pteridine transporter and a delta-12 fatty acid desaturase (Table 2.2) (OG5_129265). *Leishmania* parasites are pteridine auxotrophs so multiple transporters help transfer it from the insect or host [373,374]. The antifolate drug methotrexate, used in *Leishmania* treatment, targets the folate biosynthesis pathway to inhibit cellular growth [373]. These include transmembrane pteridine transporters, and associated proteins like pteridine reductase are potential drug targets [375]. The pteridine transporter protein amplified here has a BT1 domain and both the BT1 locus and pteridine reductase 1 (PTR1) can undergo amplification either spontaneously or after selection for anti-folate resistance, as either extra-chromosomal linear or circular elements. However, the BT-1 gene itself (LaHO174_354810) is not amplified here. Amplifications and mutations of other folate transporter genes may be driven by drug pressure or pteridine limitation [374].

Phosphoglycan beta 1,3 galactosyltransferase 5 or SCG5, is on a 5.7 kb locus with three to four copies in both *L. adleri* HO174 and SKINK-7 (Table 2.2). Five haploid copies of the gene were predicted in *L. adleri* HO174 compared to just one copy in *L. tarentolae* and *L. major.* It is one of an orthologous group of genes necessary for the modification of the phosphoglycan repeats on the surface glycocongugate lipophosphoglycan (LPG) adhesion. Modification of these repeats is important for stage specific adhesion of *Leishmania* promastigotes to lectins in the midgut epithelium of the sandfly and later detachment of the infectious metacyclic promastigotes from the midgut so that they can to be transmitted during sandfly biting [376,377]. The majority of phosphoglycan 1,3 galactosyltransferase gene family members, are up-regulated in the *L. major* and *L. infantum* amastigote stage [205].

*L. adleri* possessed 18 α- and 17 β-tubulin gene copies, whereas *L. tarentolae* had zero and two, respectively - though the *L. tarentolae* assembly may be incomplete at this region. The structural subunit of microtubules is the α/β heterodimer formed from these tubulins. Microtubules dictate cell shape, flagellar motility, intracellular transport, are drug targets [378,379] and interact with ATG8 to help deliver autophagosomes to the vacuole/endosomal-lysosomal compartment [380]. This assists with nutrient recycling through autophagy [366,381], which is crucial for life cycle stage differentiation and surviving nutrient-limiting conditions [366]. Here, an *ATG8* gene array was absent in *L. tarentolae* but present in *L. major* and *L. adleri*, so these tubulin and *ATG8* gene family changes may interact to promote stress tolerance.

The gene with the highest copy number in *L. adleri* (Table 2.3 & A11) was elongation factor 1-alpha (*EF-1α*, LaHO174_170090), which had 53 copies compared to 38 in *L. tarentolae* and 22 in *L. major* (21 were reported in [177]). Only a single *EF-1α* reference copy was in the *L. adleri* and *L. tarentolae* genomes, relative to seven in *L. major*. *EF-1α* genes undergo extensive changes in copy number in *Leishmania*, including *L. infantum* JPCM5 (seven to twelve copies), *L. braziliensis* M2904 (one to ten) [177], and *L. panamensis* PSC-1 (one to fifteen) [188]. Here, it was amplified in both HO174 (five to six copies) and SKINK-7 (four copies).

*EF-1α* was within an amplified 19.8 kb region with three other genes (Table 2), all receptor-type adenylate cyclases (A, B and a putative one). Receptor type adenylate cyclases were down-regulated during the metacyclic promastigote-to-amastigote transition in *L. major* [382]. *L. adleri* HO174 had 13 copies of the OG containing the receptor type adenylate cyclase genes (Table S10) which is more than two fold higher than in other *Leishmania* (eight in *L. tarentolae,* six in *L. major,* five in *L. infantum,* two in *L. mexicana*, and seven each in *L. panamensis* [188] and *L. braziliensis* M2904 [177]. A putative receptor-type adenylate cyclase, (LmjF17.0200 and LinJ17_V3.01200) was differentially up-regulated in the promastigote stage of both *L. major* and *L. infantum* [205] and receptor type adenylate cyclases are down-regulated during the metacyclic promastigote-to-amastigote transition in *L. major* [382].

#### 2.5.7.2 Expanded gene arrays in both *L. adleri* and *L. tarentolae*

*L. tarentolae* has a highly expanded number of leishmanolysin (aka GP63) genes with 84 haploid copies (Table 2.3 & A12) whereas *L. adleri* has 37 haploid copies which was more similar to the 31 copies in *L. braziliensis,* fifteen in *L. infantum,* thirteen in *L. mexicana,* six in *L. major* [177] and 28 in *L. panamensis* [188]. 49 leishmanolysin genes were originally reported on *L. tarentolae* [190] which is possibly due to differences in group assignment using the older OrthoMCL version or the addition of 253 protein-coding genes in the TriTryDB v6 genome. Leishmanolysin is a surface zinc-dependent metalloprotease and virulence factor involved in involved in cleaving the VAMP8 membrane fusion regulator to evade phagocytosis by host macrophages [383]. It is abundantly expressed on the promastigote cell surface [384] and is up-regulated during the metacyclic promastigote-to-amastigote transition in *L. major* [382].

There were also 30 predicted copies of genes (13 assembled copies) coding for hypothetical proteins in one array in *L. adleri* (OG5_139233) and a staggering 156 copies of *L. tarentolae* genes predicted in the same array (51 assembled copies) with only two haploid copies predicted for *L. major* (Table 2.3). The proteins in this array mainly contained one to two 'Leucine-rich repeats domain, L-domain' domains (Interpro ID: IPR032675) and some proteins also contained one to two 'Growth factor receptor cysteine-rich' domains (Interpro ID: IPR009030) e.g. genes LaHO174_bin830010 and LaHO174_bin2490010. Leucine-rich repeats (LRR) are generally two to 29 amino acids long, contain two to 45 motifs and are thought to provide a structural framework for protein-protein interactions [385]. They are also the largest repeat class found in *Leishmania* [178] and these domains are found in proteins involved in interactions between macrophage complement receptors and the parasite surface [386]. Cysteine rich growth factor receptor domains are found in a range of eukaryotic proteins involved in signal transduction by receptor tyrosine kinase. Together, the presence of these domains suggested that these genes may be involved in *Sauroleishmania* signalling or infection.

| ***L. adleri* HO174** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Copy number variant information** | | | | | | **Gene information** | |
| **Chr** | **Copy number** | **Start (bp)** | **End (bp)** | **Length (bp)** | **Gene number** | **Gene ID** | **Gene product** |
| 10 | 2 | 490,001 | 500,889 | 10,888 | 3 | LaHO174_101360 | Phosphate-repressible phosphate permease-like protein |
| | | | | | | LaHO174_101370 | Pteridine transporter (folate/biopterin transporter) |
| | | | | | | LaHO174_101380 | Delta-12 fatty acid desaturase |
| 17 | 5.5 | 22,673 | 42,500 | 19,827 | 4 | LaHO174_170090 | Elongation factor 1 -alpha |
| | | | | | | LaHO174_170100 | Receptor-type adenylate cyclase a |
| | | | | | | LaHO174_170110 | Receptor-type adenylate cyclase b |
| | | | | | | LaHO174_170120 | Receptor-type adenylate cyclase |
| 26 | 3.5 | 931,670 | 941,785 | 10,115 | 3 | LaHO174_262430 | Protein kinase |
| | | | | | | LaHO174_262440 | Conserved hypothetical protein |
| | | | | | | LaHO174_262450 | Paraquat-inducible protein-A (PqiA) |
| 27* | 2.4 | 437,425 | 453,344 | 15,919 | 3 | LaHO174_271110 | ATP-binding cassette subfamily A, member 8 (ABCA8) |
| | | | | | | LaHO174_271120 | ATP-binding cassette subfamily A, member 9 (ABCA9) |
| | | | | | | LaHO174_271130 | Cysteine peptidase, Clan CA, family C2 |
| 31 | 3.9 | 1,181,426 | 1,187,096 | 5,670 | 2 | LaHO174_312740 | Conserved hypothetical protein |
| | | | | | | LaHO174_312750 | Phosphoglycan beta 1,3 galactosyltransferase 5 |
| 33* | 2.2 | 1,034,500 | 1,040,973 | 6,473 | None | | |
| | | | | | | | |
| ***L. adleri* SKINK-7** | | | | | | | |
| 10 | 2.5 | 490,001 | 500,889 | 10,888 | 2 | LaHO174_101360 | Phosphate-repressible phosphate permease-like protein |
| | | | | | | LaHO174_101370 | Pteridine transporter (folate/biopterin transporter) |
| | | | | | | LaHO174_101380 | Delta-12 fatty acid desaturase |
| 17 | 3.6 | 22,673 | 42,500 | 19,827 | 2 | LaHO174_170090 | Elongation factor 1 -alpha |
| | | | | | | LaHO174_170100 | Receptor-type adenylate cyclase a |
| | | | | | | LaHO174_170110 | Receptor-type adenylate cyclase b |
| | | | | | | LaHO174_170120 | Receptor-type adenylate cyclase |
| 17* | 2 | 518,715 | 521,099 | 2,384 | None | | |
| 26 | 2.9 | 931,670 | 941,785 | 10,115 | 3 | LaHO174_262430 | Protein kinase |
| | | | | | | LaHO174_262440 | Conserved hypothetical protein |
| | | | | | | LaHO174_262450 | Paraquat-inducible protein-A (*PqiA*) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 31 | 3.34 | 1,181,426 | 1,187,096 | 5,670 | 4 | LaHO174_312740 | Conserved hypothetical protein |
| | | | | | | LaHO174_312750 | phosphoglycan beta 1,3 galactosyltransferase 5 |

**Table 2.2:** Putative amplifications and corresponding gene information for those detected in *L. adleri* HO174 and *L. adleri* SKINK-7.

| | | | **Assembled gene number** | | | **Read depth gene number** | | |
|---|---|---|---|---|---|---|---|---|
| **ID** | **Chromosomes** | **Gene product(s)** | ***L. adleri*** | ***L. tarentolae*** | ***L. major*** | ***L. adleri HO174*** | ***L. tarentolae*** | ***L. major*** |
| | | | **All** | | | | | |
| OG5_126558 | 13,14,22,23,25,26,27,28,34,36 | dynein heavy chain, cytosolic, putative | 13 | 13 | 13 | 13 | 14 | 13 |
| OG5_126631 | 17 | elongation factor 1-alpha | 1 | 1 | 7 | 53 | 38 | 22 |
| | | | ***L. adleri* and *L. tarentolae*** | | | | | |
| OG5_126568 | 2,15,27,29,11 | ATP-binding cassette protein subfamily A, member 7, putative | 10 | 10 | 10 | 14 | 13 | 9 |
| OG5_126585 | 14,16,22,25,34,35 | kinesin K39, putative | 10 | 11 | 7 | 19 | 14 | 8 |
| OG5_126749 | 10, bin | GP63, leishmanolysin | 25 | 54 | 5 | 37 | 84 | 6 |
| OG5_139233 | 31, bin | hypothetical protein | 13 | 51 | 1 | 30 | 156 | 2 |
| OG5_140928 | 31, bin | sodium stibogluconate resistance protein, putative | 2 | 4 | 4 | 11 | 19 | 6 |
| OG5_144952 | 15 | tb-292 membrane associated protein-like protein | 3 | 1 | 1 | 43 | 23 | 1 |
| | | | ***L. adleri* and *L. major*** | | | | | |
| OG5_126605 | 13 | alpha tubulin | 1 | 0 | 12 | 18 | 0 | 10 |
| OG5_126611 | 8,21,33 | beta tubulin | 1 | 1 | 18 | 17 | 2 | 18 |
| | | | ***L. tarentolae* and *L. major*** | | | | | |
| OG5_133076 | 12, bin | surface antigen protein, putative | 3 | 8 | 18 | 4 | 18 | 24 |
| | | | ***L. adleri*** | | | | | |
| OG5_126617 | 17, bin | receptor-type adenylate cyclase | 4 | 6 | 6 | 13 | 8 | 6 |
| OG5_126703 | 9,31,36 | polyubiquitin, putative | 2 | 1 | 1 | 11 | 1 | 9 |
| OG5_133827 | 7, 34, bin | hypothetical protein, unknown function | 8 | 8 | 4 | 10 | 7 | 4 |
| OG5_164370 | 34 | hypothetical protein, conserved | 2 | 2 | 1 | 11 | 2 | 1 |
| OG5_183275 | 10, bin | hypothetical protein, conserved | 2 | 3 | 1 | 24 | 4 | 1 |
| | | | ***L. tarentolae*** | | | | | |
| OG5_126561 | 23, 26, 31, 36 | p-glycoprotein-like protein ,pentamidine resistance protein 1 | 7 | 11 | 7 | 7 | 11 | 6 |
| OG5_129265 | 6, 10 ,19 | folate/biopterin transporter, putative | 7 | 13 | 10 | 9 | 17 | 10 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| OG5_158273 | 27 | No description | 0 | 1 | 0 | 0 | 30 | 0 | |
| ***L. major*** | | | | | | | | | |
| OG5_126623 | 29,33, bin | lipophosphoglycan biosynthetic protein, putative | 2 | 1 | 18 | 2 | 1 | 13 | |
| OG5_130729 | 34, bin | amastin-like surface protein, putative | 2 | 1 | 24 | 2 | 1 | 21 | |
| OG5_130987 | 8 | amastin-like protein | 0 | 0 | 16 | 0 | 0 | 18 | |
| OG5_134066 | 34, 36 | tuzin-like protein | 1 | 1 | 22 | 1 | 1 | 16 | |
| OG5_134607 | 32 | hypothetical protein, conserved | 0 | 0 | 17 | 0 | 0 | 15 | |
| OG5_135520 | 12, bin | hypothetical protein, conserved | 1 | 3 | 12 | 6 | 2 | 22 | |
| OG5_137181 | 19 | ATG8/AUT7/APG8/PAZ2, putative | 2 | 0 | 9 | 9 | 0 | 19 | |
| OG5_173557 | 12 | promastigote surface antigen protein 2 PSA2 | 0 | 0 | 5 | 0 | 0 | 18 | |
| OG5_184212 | 30 | class i nuclease-like protein | 0 | 0 | 4 | 0 | 0 | 24 | |

**Table 2.3:** Gene arrays with >=10 haploid gene copies as predicted by read depth analysis (grey shading) for each species. Arrays with >=10 haploid gene copies in more than one species are shown in separate parts of the table and shaded in grey. Genes were assigned to arrays based on their OrthoMCL group ID.

## 2.6 Discussion

### 2.6.1 High quality annotated draft of *L. adleri* HO174

We produced a high quality annotated draft genome of *L. adleri* MARV/ET/1975/HO174, originally isolated from an asymptomatic African grass rat. Reads that passed quality control were *de novo* assembled into 5,785 scaffolds, which were contiguated initially into 36 chromosomes using the lizard-infecting *L. tarentolae* Parrot-TarII genome. The final 30.4 Mb *L. adleri* genome has 38 chromosomes with 94.5% of assembled sequence on chromosomes and 69-fold median coverage. Despite the inevitable gaps, misassemblies and low-quality regions, comparison with other genomes demonstrates that it is largely complete.

### 2.6.2 Comparative analysis reveals species–specific and shared gene content

*L. adleri* HO174 has 7,959 genes, with 7,849 protein-coding ones on 38 chromosomes and 389 on unassigned contigs, the vast majority of which were computationally assigned (98.6%). 32 gene models were absent in the nearest relative (*L. tarentolae*) but present in *L. major* and 22 of these were present in *Leishmania* belonging to other subgenera (*L. mexicana*, *L. infantum* and *L. braziliensis*) demonstrating that either they are not truly absent from *L. tarentolae* (as they could be just unassembled or not annotated) or that they are more important in amastigote (which infect human macrophages) infectious stages than in lizard infections.

Among the 32 genes identified, *ndK* and calretulin have been associated with infection. *NdK* had five copies in *L. adleri,* at least two each in *L. mexicana, L. infantum* and *L. braziliensis* and none in *L. tarentolae.* It is a housekeeping gene that catalyses the transfer of phosphate from NTP (nucleoside triphosphate) to NDP (nucleoside diphosphate) for homeostasis of NTP levels in cells [387] and also participates in purine salvage pathways in trypanosomatids [388]. It is important for the successful infection of macrophages in *L. amazonensis* LV78 by preventing ATP-mediated lysis of host macrophages, preserving the cells for use by *Leishmania* [387]. Promastigotes progressively release NDK leading to its accumulation in stationary phase, when they become metacyclic promastigotes (the most infective stage). Concordant with this, *ndk* expression has been shown to be up-regulated in the metacyclic promastigote and amastigote stages in *L. major* [389] and in the promastigote stage of *L. mexicana* [390] and down-regulated in antimony resistant *L. infantum* (strain Sb2000.1) promastigotes [212]. *Leishmania ndk* can also potentially use ATP to produce different NTPs such as GTP and UTP which regulate gene expression in signal pathways and may further alter macrophages to make them more hospitable to intracellular parasites [387].

Calrectulin (LaHO174_312230), which is absent in *L. tarentolae* but present in one copy in *L. adleri* and *L. major, L. mexicana*, *L. braziliensis* and *L. infantum*, is an endoplasmic reticulum chaperone protein involved in quality control of synthesized glycoproteins by assisting the retention of misfolded proteins which are then targeted for proteasome degradation. Overexpression of the truncated form of calreticulin in *L. donovani* results in decreased secretion of acid phosphatase glycoproteins and a reduction in parasite survival inside macrophages, suggesting that its alteration affects trafficking of virulence associated proteins through the secretory pathway [391]. It is also highly overexpressed in *L. infantum* amastigotes compared with promastigotes [392] .

Further work is required to characterise the eight genes that appeared to be unique to *L. adleri* when compared with five *Leishmania* and five *Trypanosoma* spp; these had VSG like domains and were orthologous to genes present in the trypanosomatid *Phytomonas sp* HART isolate, although the function and importance of these genes remains unknown.

### 2.6.3 Large gene arrays are present in *L. adleri*

Large gene arrays formed by gene duplication in *Leishmania* and *Trypanosoma* provide a means of increasing gene expression in the absence of gene regulation at the level of transcription initiation [178,393]. We found 295 gene arrays in *L. adleri* HO174, 281 in *L. tarentolae* and 289 in *L. major* using our approach. The Elongation factor 1 alpha (*EF-1α*) gene array had the highest gene copy number in *L. adleri* HO174 and leishmanolysin array had the highest copy number in *L. tarentolae,* compared with *L. major*, *L. infantum*, *L. mexicana* and *L. braziliensis*.

*EF-1α* encodes a highly conserved GTP-binding protein involved in protein translation, and is a candidate virulence factor in *L. donovani* because it deactivates infected macrophage cells by binding and stimulating the host Src homology 2 domain containing tyrosine phosphatase-1 (SHP-1) [394]. This activation of host SHP-1 is associated with leishmaniasis disease pathogenesis, and SHP-1 is inhibited by sodium stibogluconate [395]. *EF-1α* is up-regulated in antimony-resistant *L. infantum* promastigotes compared with wild type [371] and a *L. donovani* EF-1α homolog (LdBPK_354220) had a reduced copy number in a sample with resistance to antimonials during miltefosine-resistance induction [396]. We also found that the *L. tarentolae* genome had more ‘collapsed’ gene copies (119 arrays incompletely assembled) than *L. adleri* HO174 (62 arrays) and *L. major* (12 arrays) although this approach was limited by the numbers of genes from each array annotated on the genomes.

The large number of hypothetical proteins in an array in *L. adleri* (OG5_139233) (30 haploid copies) and *L. tarentolae* (156 haploid copies) compared with only two haploid copies in *L. major* and three in *L. mexicana* are also of interest and it is tempting to speculate that these may have some importance in *Sauroleishmania* survival or infections given their high copy number.

### 2.6.4 Identification of HO174 as *L. adleri*

Alignment of seven *L. adleri* HO174 genes with those of 224 other *Leishmania* isolates from infected patients, mammals and insects [353] showed that HO174 was a *Sauroleishmania* isolate most closely related to *L. adleri* RLAT/KE/1957/SKINK-7 and *L. adleri* RLIZ/KE/1954/1433 compared to *L. tarentolae*, *L. hoogstraali* and *L. gymnodactyli*. Mapping and SNP calling SKINK-7 reads using the HO174 reference also confirmed HO174 as *L. adleri*, rather than *L. tarentolae*. Consequently, HO174 is the first genome sequenced from the *Sauroleishmania* subgenus isolated from a mammal. Previous multilocus microsatellite typing (MLMT) could not classify HO174 clearly [346] and so our work demonstrates the usefulness of whole genome sequencing for resolving uncertain taxonomic classification.

### 2.6.5 Two chromosome fission events discovered in *L. adleri*

Most *Leishmania* species have 36 chromosomes, including all members of the *Leishmania* subgenus – except for 34 in the *L. mexicana* complex, in which chromosomes 8 and 29 are a single fused chromosome 8, and chromosomes 20 and 36 are a single fused chromosome 29 [177] and the 35 chromosomes in the *Viannia* subgenus where chromosome 20 is homologous to fused chromosomes 20 and 36 [181]. In contrast, chromosome 30 is largely conserved as a single unit in trypanosomatids [184]. Gene copy number fluidity is common in *Leishmania* because transcription is controlled at PTUs rather than individual genes and so chromosomal fission may be more common at SSRs to preserve RNA polymerase II promoters, which are enriched in acetylated histone H3 (H3Ac) [199], because transcription is initiated at SSRs [362]. In *Leishmania*, these fissions may be functionally neutral and have arisen from erroneous chromosome replication [210]. The fission breakpoints on *L. adleri* chromosomes 30 and 36 were at SSRs in *L. adleri*, yielding three SSRs at chromosome 36.1 and two for 36.2 (Figure 2.4(a)). The chromosome 30 break was at a SSR, suggesting 3’ to 5’ transcription of chromosome 30.1 with no SSRs, and two SSRs for 30.2 (Figure 2.4(b)).

Viable new chromosomes must retain a single origin of replication (ORI) for viability, and these are at SSRs for 30 out of 36 *L. major* chromosomes. Only one single bidirectional ORI is typically retained after chromosomal fusion in *Leishmania,* such as the single ORIs that are now present on the fused *L. mexicana* chromosomes 8 and 29, even though the ancestral chromosomes originally had one on each chromosome [362]. *L. major* and *L. mexicana* had their chromosome 30 ORI at a site equivalent to the *L. adleri* chromosome 30 fission breakpoint, indicating that replication may proceed from this origin in a 3'-5' direction for chromosome 30.1 and a 5'-3' direction for 30.2. The chromosome 36 origin (position 1,110,127 to 1,116,528 bp in *L. major*) [362] was at the 5' end of the *L. adleri* chromosome 36.2 suggesting that that a different origin may be used for chromosome 36.1.

*L. adleri* HO174 and SKINK-7 as well as *L. tarentolae* were shown here to be primarily disomic and also aneuploid, in common with *Leishmania* spp. in other subgenera [363]. Chromosome 36.1 was disomic whereas 36.2 was trisomic in *L. adleri* HO174 and chromosome 30.1 was disomic and 30.2 was trisomic in *L. adleri* SKINK-7.

Only a single early firing origin has been reported for each *Leishmania* chromosome using MFASeq of *L. major* and *L. mexicana* [362]. However, MFASeq does not have adequate resolution to detect a large number of low or variable efficiency firing origins especially if they are contained within more prominent replication domains with similar replication timing in different cells or exhibit heterogeneous origin firing in each cell [397]. Indeed, a study using DNA combing, which can measure inter-origin distances and replication rate, of promastigote forms of *L. donovani, L. major* and *L. mexicana* and procyclic forms of *T. brucei,* estimated that there are 168 to 180 active origins per haploid genome for *L. major* and *L. donovani* and ~150 for *L. mexicana* [398] but replication origins cannot be localised to specific chromosomes using DNA combing alone so the precise location of the additional origins remains to be discovered. Additionally, a mean fork velocity of 2.6 kb/min for *L. mexicana* indicates that large 1-3 Mb chromosomes would take 6-20 hours to replicate suggesting that there must be multiple origins on each chromosome, at least for larger chromosomes [398]. Thus, chromosomes 30 and 36 of *L. adleri* may have multiple origins and so the split chromosomes may retain at least an origin each.

Spontaneous chromosomal fission has been observed in another *Sauroleishmania* isolate, the *L. tarentolae* strain LEM115, isolated from a gecko in southern France [399]. During routine subculture and after single-cell cloning a chromosome of 365 kb in length (named chromosome 4 but similar length to *L. tarentolae* chromosome 3) underwent fission to produce a truncated chromosome of length 340 kb (chromosome 4A) [400]. One cloned line had cells with both chromosomes 4 and 4A where the chromosome 4A line outgrew the wild-type after re-cloning, suggesting no fitness loss [400]. Chromosome 4A may have been due to a contraction of the mini-exon gene array and expansion/contraction of this array has been shown to occur spontaneously on chromosome 2 of an *L. major* LT252 delta line [401].

Chromosome 11 from a *Trypanosoma cruzi* Y strain also exhibited a similar read depth coverage increase to that of *L. adleri* HO174 chromosome 36 and its coverage change occurred at a SSR (at 248 kb) resulting in the segment 5’ of the SSR having a uniformally lower coverage than the 3’ segment. This indicated either loss of the 5’ 248 kb region in one copy of the chromosome with perhaps partial fixation in the cell population or fission of the chromosome into a monosomic chromosome 11.1 and a disomic chromosome 11.2, which may explain variable bands seen on PGFE for *T. cruzi* strains [402].

### 2.6.6 Animals are important *Leishmania* hosts

Increasing evidence suggests that *Sauroleishmania* spp. such as *L. tarentolae*, which was considered non-pathogenic to humans, infected people [135,403] and dogs, causing VL from 1984 to 1990 in China, and that traditionally human infecting species such as *L. donovani* can infect lizards [404]. Thus, increased sampling to find these reservoirs as well as characterisation of their *Leishmania* spp. will be important to understand how leishmaniasis is spread which will assist in its control e.g. bats have been shown to be naturally infected with *L. tropica* and *L. major* in Ethiopia [405] and dogs are reservoirs of *L. infantum* and possibly *L. donovani*, there [406]. The *L. adleri* genome produced here will assist in the study of *Sauroleishmania* infections.

# Chapter 3 - Comparative analysis of the genomes of *L. naiffi* and *L. guyanensis* provides insights into the *Viannia* subgenus and the first evidence of *L. naiffi* causing canine leishmaniasis in Colombia

*I performed all analyses in this chapter.*

*Ali Shirley Taylor and Eoghan Feane, both of Dublin City University, manually checked and corrected most of the gene annotation for L. naiffi CL223 and L. guyanensis CL085 genomes.*

*A submission is in progress at the EBI for the genome sequence and annotation of L. naiffi CL223 and L. guyanensis CL085.*

*Supplementary material for this chapter can be found in Appendix B.*

# 3.1 Chapter Overview

## 3.1.1 Aims and objectives

In this chapter, we aimed to assemble the genomes of two *Leishmania* samples taken from dogs in Colombia and identify the species using a similar approach to that used in Chapter 2. We decided to assemble a control genome using short-read data from a published genome to quantify the ability of our assembly pipeline to complete the genome and also to assist with quality control of the short-read assemblies. We aimed to characterise gene and chromosome copy number variation in both genomes that we assembled, as well as in other publically available *Leishmania* samples.

## 3.1.2 Methodology

*Leishmania* DNA sampled from two dogs in Colombia was previously sequenced on the Illumina HiSeq producing 15,272,969 100 bp paired-end reads for one sample (CL085) and 8,131,246 100 bp paired end reads for the other (CL223). The same pipeline developed for assembly, species identification and characterisation in Chapter 2 was used in this chapter with some changes (Figure 3.1). These differences included the identification of contaminant reads in one of our read libraries (CL085) and in a 100 bp paired-end *L. braziliensis* M2904 library [177] using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and their subsequent removal based on GC content and BLASTn [29] results. These steps were also undertaken as part of the *L. adleri* HO174 assembly pipeline in Chapter 2 in case of low level contamination, even though no contamination was identified using FASTQC 'Per Sequence GC content plots' there. The CL223 library was quality trimmed using Trimmomatic [23] to remove low quality bases at the 3' end of reads and the CL085 scaffolds were also searched against the nucleotide database using BLASTn [29] to find contamination that was not effectively removed at the read QC stage. A control genome was assembled using the *L. braziliensis* M2904 reads which enabled us to measure the level of genome completeness achieved with our pipeline by comparing the gene content and genome length of the control assembly with same published genome assembled from long reads. It also aided in the discovery of misassemblies in the CL085 and CL223 genomes by finding false joins shared by the control and CL223/CL085 genomes that were absent in the *L. braziliensis* reference genome created using long reads [178]. A test run of the Companion annotation pipeline [67] was also performed, using the published *L. braziliensis* reference genome as both the input genome and reference genome, to determine its ability to rediscover known genes. Protein coding genes were assigned to OGs and the gene content and copy number was examined in the same manner as in Chapter 2, with the addition of *L. panamensis* PSC-1 [188] genes to the OGs comparison

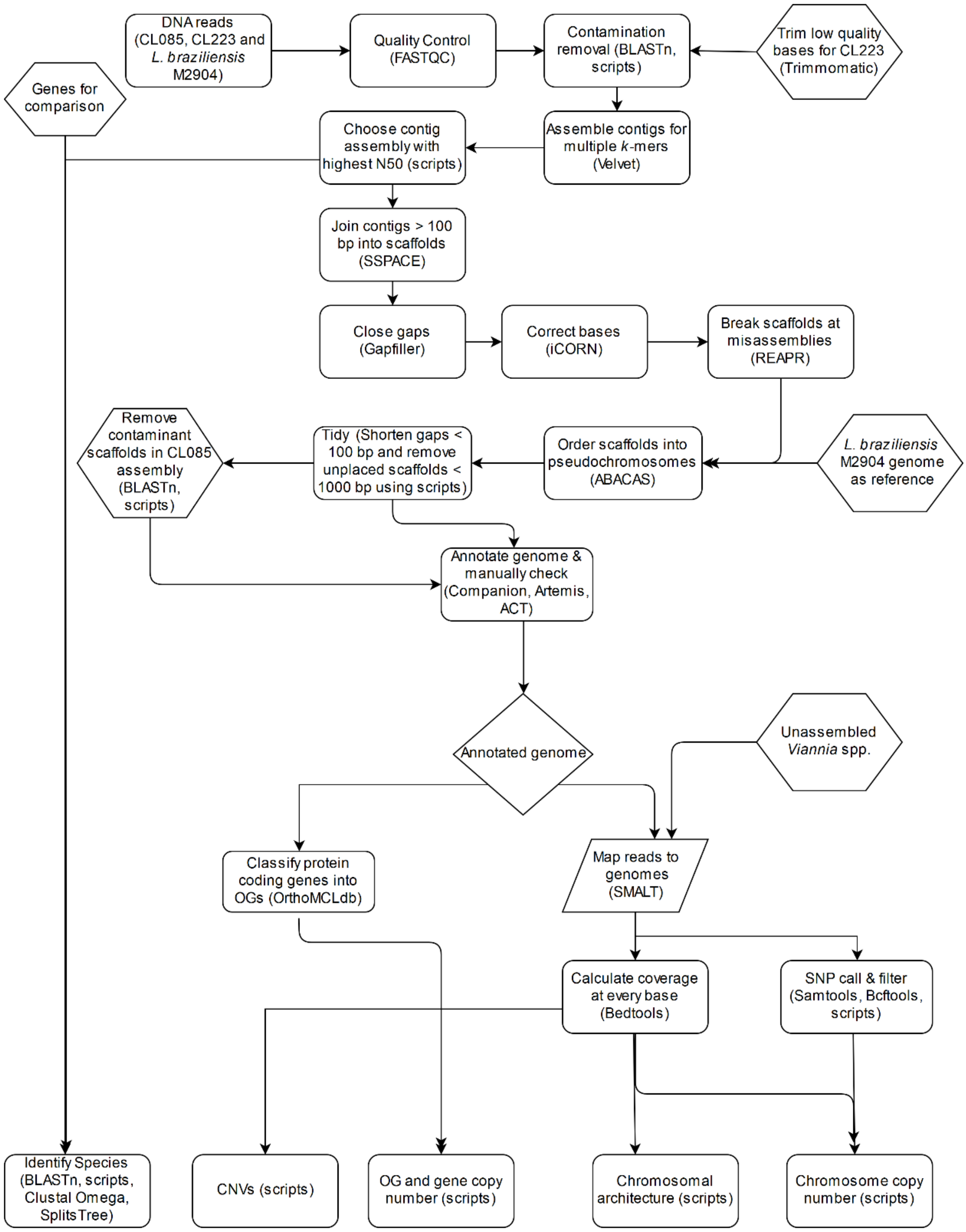

DNA reads (CL085, CL223 and *L. braziliensis* M2904)
Quality Control (FASTQC)
Contamination removal (BLASTn, scripts)
Trim low quality bases for CL223 (Trimmomatic)
Genes for comparison
Choose contig assembly with highest N50 (scripts)
Assemble contigs for multiple *k*-mers (Velvet)
Join contigs > 100 bp into scaffolds (SSPACE)
Close gaps (Gapfiller)
Correct bases (iCORN)
Break scaffolds at misassemblies (REAPR)
Remove contaminant scaffolds in CL085 assembly (BLASTn, scripts)
Tidy (Shorten gaps < 100 bp and remove unplaced scaffolds < 1000 bp using scripts)
Order scaffolds into pseudochromosomes (ABACAS)
*L. braziliensis* M2904 genome as reference
Annotate genome & manually check (Companion, Artemis, ACT)
Annotated genome
Unassembled *Viannia* spp.
Classify protein coding genes into OGs (OrthoMCLdb)
Map reads to genomes (SMALT)
Calculate coverage at every base (Bedtools)
SNP call & filter (Samtools, Bcftools, scripts)
Identify Species (BLASTn, scripts, Clustal Omega, SplitsTree)
CNVs (scripts)
OG and gene copy number (scripts)
Chromosomal architecture (scripts)
Chromosome copy number (scripts)

**Figure 3.1:** Pipeline used in Chapter 3 for genome assembly, improvement, annotation and analysis. Sections of the pipeline that are either different or newly added compared to the pipeline in Chapter 2 are in hexagonal shapes.

Other changes to the methodology used in Chapter 2 involved the use of *Viannia* spp. for species identification instead of *Leishmania* and *Sauroleishmania* spp. and thus the use of different genes to those used in the Chapter 2 phylogenomics step (as the strains and genes used for comparison with our samples were taken from published data [407]). The phylogenomics step used here used four genes from 102 sequences and revealed that the CL223 sample was *L. (Viannia) naiffi* and the CL085 sample was *L. (Viannia) guyanensis*. Based on this, the *L. (Viannia) braziliensis* M2904 published reference genome was used as a reference for chromosome level scaffolding as it was closely related to these samples.

Chromosomal architecture and copy number as well as SNPs and CNVs were examined for *L. guyanensis* CL085 and *L. naiffi* CL223 genomes as well as two *L. (Viannia) peruviana* genomes [189], two *L. (Viannia) panamensis* genomes and five unassembled *Viannia* spp. (*L. shawi* M8408, *L. guyanensis* M4147, *L. naiffi* M5533, *L. lainsoni* M6426 and *L. panamensis* WR120) [179].

### 3.1.3 Conclusions

In this chapter, we have provided the first draft genomes of *L. naiffi* and *L. guyanensis* which we have characterised extensively, as well as the first evidence of *L. naiffi* in Colombia and in dogs. We have also discovered a minichromosome in a strain of *L. shawi* commonly used in laboratories, as well as an unstable ~45 kb amplification in many *Viannia* genomes. Genes that have high copy number such as the NADH-dependent fumarate reductase gene, tuzins, TATE transposons in *L. guyanensis* CL085 and leishmanolysin in *L. naiffi* CL223 were also identified, as were genes in twenty-three OGs that are possibly unique to *Vianna* spp., which could aid investigations into the ability of *Viannia* spp. to cause the highly disfiguring MCL. The control *L. braziliensis* M2904 genome was found to be mostly complete as 93.5 % of reference sequence length was assembled and most of the protein coding genes that were annotated on the reference genome were also annotated on the control genome only (2.8 % of genes present in reference genome OGs were absent in the control genome OGs). Additonally, 70 protein coding genes that were not annotated on the reference genome were identified on the control. Taken together, these results indicate that our short-read assemblies are generally complete, although as discussed in Chapter 5, section 5.4, long reads and transcriptomics data would still be useful to further improve the assemblies and annotation.

## 3.2 Abstract

*Leishmania* is a protozoan parasite that causes the neglected tropical disease leishmaniasis which infects 12 million people in 98 countries. In many cases it is a zoonosis with many of the *Leishmania* species capable of infecting humans also able to infect animals such as dogs where they cause canine leishmaniasis. *Leishmania* DNA was sampled from two dogs with the cutaneous form of canine leishmaniasis in Colombia in 1985/86 and whole genome sequenced. We assembled the genomes to chromosome level with short-read data, using both de-novo assembly and reference based alignment, and annotated them, finding over 8,000 genes on each genome. A control assembly created from short-reads, using the same pipeline, was also used to validate results. The genomes were identified as *L. guyanensis* CL085 and *L. naiffi,* using a phylogenomics approach. These are both members of the *Viannia* subgenus of *Leishmania* which is only found in the Americas, mainly South America, and *L. naiffi* had not previously been reported to cause disease in dogs or humans in Colombia, making this the first report of this species in Colombia. Using depth of coverage from mapped reads as well as SNP calling, both species were found to exhibit mosaic aneuploidy. Comparisons of these genomes with other assembled and unassembled *Viannia* genomes, revealed the presence of a minichromosome in an unassembled *L. shawi* sample and a 45 kb amplification common to many *Viannia* genomes including *L. naiffi* and *L. guyanensis*. High copy number of TATE transposons, a *Vianni*a specific feature, was found in *L. guyanensis* CL085 and high copy number of leishmanolysin, a putative virulence factor, was found in *L. naiffi* CL223. NADH-dependent fumarate reductase, which is involved in oxidative stress responses, was found to have increased copy number in all *Viannia* genomes examined compared with orthologs in other subgenera. Few genes specific to *L. naiffi* and *L. guyanensis* or restricted to the *Viannia* subgenus were uncovered. Our results demonstrate that genome sequencing and assembly of short-read data has an important role to play in disease surveillance and the characterisation of *Leishmania* spp. and subgenera.

## 3.3 Introduction

### 3.3.1 Human and canine leishmaniasis

In many cases, leishmaniasis is a zoonosis - 12 of the 21 human infecting *Leishmania* spp. are also capable of infecting domestic dogs (*Canis familiaris*) and causing canine leishmaniasis (CanL). CanL is prevalent in ~50 countries [173] and is caused by *L. infantum* (syn. *L. chagasi*) in the Old World and mainly *L. donovani (infantum)* and *L. (Viannia) braziliensis* in the New World. *Lutzomyia longipalpis* is the main vector of *L. donovani* *(infantum)* in the New World [408], though it can also transmit other *Leishmania* spp [156]. In general, the majority of infected dogs - over 80% of seropositive dogs in some foci in Brazil for instance - are asymptomatic [175,176] though they can act as carriers and are capable of transmitting *Leishmania* to sandflies [409,410].

Human CL is considered to be a zoonosis in the Americas and Mediterranean basin with transmission from sylvatic and peridomestic mammalian reservoirs, including domestic and stray dogs [129], via sandflies of the genus *Lutzomyia* in the New World and *Phlebotomus* in the Old World [167]. In Colombia, dogs are reservoirs of *L. braziliensis* and *L. panamensis* [411–413] as well as *L. infantum* which causes VL, making them targets of control measures for human disease [172,335]. Seven *Leishmania* spp. are present in Colombia: *L. braziliensis*, *L. guyanensis*, *L. panamensis*, *L. colombiensis*, *L. amazonensis*, *L. infantum* and *L. mexicana* [412] and *L. lainsoni* and *L. equatoriensis* infect humans in Colombia [414]. CL is endemic in Colombia [415], comprising 96.35% of 58,897 cases recorded in Colombia from 1990 to 1999, compared with 2.65% of MCL [414] which can be caused by *L. panamensis*, *L. guyanensis* or *L. braziliensis* [416]. *L. infantum* causes 1% of VL cases [414]. *L. panamensis* and *L. braziliensis* are frequently associated with human CL in Colombia [411,417–419] and less often with *L. guyanensis* and *L. mexicana* [417].

The two *Leishmania* samples assembled in this study, typed as *L. guyanensis* and *L. naiffi,* both members of the *Viannia* subgenus, were isolated from dogs with cutaneous leishmaniasis in Colombia in 1985 and 1985, respectively.

### 3.3.2 *L. naiffi*

*L. naiffi* was first described in 1989 in the Pará state, northern Brazil [420] where it was isolated from the liver and spleen (indicative of VL) of the nine banded armadillio (*Dasypus novemcinctus*), which is the only known reservoir of *L. naiffi* [168,420,421]. Its initially uncertain taxonomical position was clarified by [422] who demonstrated that it belonged to the *Viannia* subgenus, and placed it between *L. braziliensis* and *L. guyanensis.* It has been found as a cause of human localised CL in Brazil, French Guiana, Ecuador, Peru and Surinam [423,424], and is the second most common cause of human CL after *L. guyanensis* in Brazil [425]. *L. naiffi* vectors include *Lu. (Psathyromyia) ayrozai* and *Lu.(Psychodopygus) paraensis* in Brazil [426], *Lu.(Psathyromyia) squamiventris, Lu. tortura* in Ecuador [427] and *Lu. trapidoi* and *Lu. gomezi* in Panama [428].

*L. naiffi* has only ever been isolated from humans, armadillos [420,421] and the rodent *Thrichomys pachyurus* which lives in the same habitat as the nine-banded armadillo in Brazil [429], making this the first report of both *L. naiffi* causing cutaneous CanL in a dog and also the first evidence of *L. naiffi* in Colombia. Although *L. naiffi* causes VL in armadillos, it only causes localised CL with small discrete lesions on the hands, arms or legs in humans [421,425]. It usually responds to treatment [421,423] and can be self-limiting [430], though poor response to antimonial or pentamidine therapy was reported in two patients in Manaus, Brazil [425].

### 3.3.3 *L. guyanensis*

*L. guyanensis* was first described in 1954 [431]. Its reservoirs in Colombia have not been elucidated, although the forest dwelling two-toed sloth (*Choloepus didactylus*) is the principal reservoir of *L. guyanensis* in neighbouring Brazil [412,432]. Potential secondary reservoirs of *L. guyanensis* include the lesser anteater *Tamandua tetradactyla*, rodents *(Proechimys* sp) [432] and *Marmosops incanus* in Brazil and *Didelphis marsupialis* in Venezuela [433,434]. It has also been found in the reservoir host of *L. naiffi*, *Dasypus novemcinctusn*, *in* Brazil [435]. *Lu. umbratilis* is the principal vector of *L. guyanensis* [436] and is the only confirmed vector of *L. guyanensis* in Colombia [437]. *Lu. anduzei and Lu. whitmani* are secondary vectors [438,439] and all three vectors are prevalent in forests in South America [437]. *L. guyanensis* has been found in French Guiana, Bolivia, Brazil, Colombia and Suriname [437,440–444]. Infection of humans, dogs and *Lutzomyia ovallesi* with *L. guyanensis/L. braziliensis* hybrids has been reported in Venezuela [445,446] and an *L. shawi*/*L. guyanensis* hybrid was isolated from a human CL lesion in Amazonian Brazil [447].

### 3.3.4 Cutaneous leishmaniasis in Colombia

CL is known as ‘guerrilla’s sore’ in Colombia and Venezuela and prior to the 1960s was mainly confined to sylvatic habitats [160]. Its transmission in Colombia and other South American countries has been shifting to domestic and peridomestic habitats as a result of migration, new settlements and deforestation of primary forests [166,448,449] resulting in cases increasing from approximately 6,500 cases per year in the 1990s to 16,098 in 2006 with the highest number in the Andean region of Colombia [450].

### 3.3.5 *Leishmania* genomes

*Leishmania* genomes have genes organised in polycistronic transcription units (PTUs) and these exhibit a high degree of shared synteny across *Leishmania* [178]. These PTUs are co-transcribed by RNA polymerase II as polycistronic pre-mRNAs which are then transpliced and polyadenylated [192,197]. The genomes of *L. infantum*, *L. donovani*, and *L. major* have of 36 chromosomes each [451] and *Viannia* spp. genomes have 35 chromosomes due to a fusion of chromosome 20 and 34 [178,181]. Two *Viannia* genomes are assembled to the level of chromosomes and annotated - these are *L. panamensis* MHOM/PA/94/PSC-1 [188] and *L. braziliensis* MHOM/BR/75/M2904 [177,178]. Additional *L. panamensis* (MHOM/COL/81/L13) and *L. braziliensis* (MHOM/BR/75/M2903) assemblies are available as scaffolds in TriTrypDB and Genbank and two *L. peruviana* genomes are also available which are assembled into chromosomes although these are not annotated [189].

### 3.3.6 Comparative genomics of *Viannia* genomes

Here, we assembled and annotated the genomes of *L. guyanensis* (MCAN/CO/1985/CL085) and *L. naiffi* (MCAN/CO/1986/CL223) from short-read data, identifying over 8,000 genes and resolving more than 28 Mb of sequence into 35 chromosomes for each genome. We examined these samples in comparison to other *Viannia* spp. to examine structural variation and sequence divergence as well as gene and chromosome level copy number changes. We also documented aneuploidy in five unassembled *Viannia* short read datasets isolated from humans, armadillos and primates [179] (*L. shawi* MCEB/BR/1984/M8408, *L. guyanensis* MHOM/BR/1975/M4147, *L. naiffi* MDAS/BR/1979/M5533, *L. lainsoni* MHOM/BR/1981/M6426 and *L. panamensis* MHOM/PA/1974/WR120) and uncovered evidence of a mini-chromosome in *L. shawi* M8408. These five strains are commonly used in studies of *Viannia*, including [452,453,407]. This study highlights the utility of whole genome sequencing for the identification, characterisation and comparison of *Leishmania* spp. and reports for the first time, to our knowledge, both the presence of *L. naiffi* in Colombia and *L. naiffi* as an etiologic agent of cutaneous leishmaniasis in dogs.

## 3.4 Methods

### 3.4.1 Genome Sequencing

DNA was taken from cloned samples of two dogs with cutaneous leishmaniasis in Colombia in 1985/6. Subsequent library preparation, DNA sequencing and read quality verification was conducted using methodology outlined previously [186]. Paired-end short-read Illumina HiSeq 2000 libraries were prepared for both specimens. The amplified library for *L. guyanensis* contained 15,272,969 100 bp reads with a median insert size of 400 and the *L. naiffi* library contained 8,131,246 100 bp reads with a median insert size of 500. Raw reads are deposited at the NCBI Short Read Archive (SRA) at ERX180458 for *L. guyanensis* and ERX180449 for *L. naiffi*.

### 3.4.2 Comparative data

The *L. braziliensis* reference genome (MHOM/BR/75/M2904) was a positive control examined using the same methods: this was originally amplified using an Illumina Genome Analyzer II [177]. It had 26,007,384 76 bp paired-end reads obtained from the SRA with a median insert size of 250 bp (ERX005631). The *L. braziliensis* reference genome and annotation (version 3) EMBL files were downloaded from ftp://ftp.sanger.ac.uk/pub/project/pathogens/L_braziliensis/Archives/LbrM_v3_20110311/artemis/EMBL/Lbraziliensis/1/. Protein sequences were retrieved from the EMBL files using Artemis [454].

Two *L. panamensis* genomes and two *L. peruviana* genomes were used for comparison. The *L. panamensis* MHOM/PA/94/PSC-1 genome and annotation files [188] were downloaded from the NCBI (accessions: CP009370:CP009404) and 5,875,837 100 bp paired-end whole genome shotgun Illumina HiSeq 2000 reads used for part of its assembly were downloaded in SRA accession SRX681913. Its 7,748 protein coding sequences and genome annotation were downloaded from RefSeq via the NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000755165.1_ASM75516v1/).

The *L. panamensis* MHOM/CO/1981/L13 assembly contains scaffolds rather than chromosomes and was downloaded from TriTrypDB release 4.2 (credit: Stephen M. Beverley and the Genome Institute, Washington University School of Medicine; Genbank accession: AOND00000000.1). The *L. peruviana* PAB-4377 genome [189] was downloaded from NCBI Bioproject PRJEB7263, and its 16,117,316 100 bp Illumina HiSeq 2000 paired-end reads from SRA accession ERX556165. The *L. peruviana* LEM1537 (MHOM/PE/1984/LC39) genome [189] was downloaded from the same NCBI Bioproject, and its 9,378,317 100 bp (Illumina HiSeq 2000) paired-end reads from SRA accession

ERX556164. Five 100 bp paired-end Illumina HiSeq 2000 read libraries of isolates from the *Viannia* subgenus used in [179] were also retrieved (Table 3.1). These were:

1. *L. shawi* MCEB/BR/1984/M8408 from a primate (5,110,479 reads; SRX764331)
2. *L. guyanensis* MHOM/BR/1975/M4147 from a human (6,225,035 reads; SRX767379)
3. *L. naiffi* MDAS/BR/1979/M5533 from an armadillo (9,646,461 reads; SRX764332)
4. *L. lainsoni* MHOM/BR/1981/M6426 from a human (4,630,952 reads; SRX764333)
5. *L. panamensis* MHOM/PA/1974/WR120 from a human (4,536,341 reads; SRX767384)

| **Species** | **Data Source** | **Data Type** | **WHO Number** | **SRA Accessions (Genome Accession or TriTrypDB information)** | **Number of reads** | **Reference** |
|---|---|---|---|---|---|---|
| ***Subgenus Viannia*** | | | | | | |
| *L. braziliensis* | Sanger ftp | Genome & Reads | MHOM/BR/1975/M2904 | ERX005631 (Sanger ftp site- LbrM2904 version 3) | 26,007,384 (76 bp paired-end) | [177] |
| ***L. guyanensis*** | **SRA** | **Reads** | **MCAN/CO/1985/CL085** | **ERX180458** | **15,272,969 (100 bp paired-end)** | **This study** |
| *L. guyanensis* | SRA | Reads | MHOM/BR/1975/M4147 | SRX767379 | 6,225,035 (100 bp paired-end) | [179] |
| *L. lainsoni* | SRA | Reads | MHOM/BR/1981/M6426 | SRX764333 | 4,630,952 (100 bp paired-end) | [179] |
| *L. naiffi* | SRA | Reads | MDAS/BR/1979/M5533 | SRX764332 | 9,646,461 (100 bp paired-end) | [179] |
| ***L. naiffi*** | **SRA** | **Reads** | **MCAN/CO/1986/CL223** | **ERX180449** | **8,131,246 (100 bp paired-end)** | **This study** |
| *L. panamensis* | TriTrypDB | Genome | MHOM/CO/1981/L13 | NA (AOND00000000.1; TriTrypDB version 4.2) | NA | Stephen M. Beverley and The Genome Institute, Washington University School of Medicine |
| *L. panamensis* | Genbank & SRA | Genome & Reads | MHOM/PA/1994/PSC-1 | SRX681913; (CP009370: CP009404) | 5,875,837 (100 bp paired-end) | [188] |
| *L. panamensis* | SRA | Reads | MHOM/PA/1974/WR120 | SRX767384 | 4,536,341 (100 bp paired-end) | [179] |
| *L. shawi* | SRA | Reads | MCEB/BR/1984/M8408 | SRX764331 | 5,110,479 (100 bp paired-end) | [179] |
| *L .peruviana* | Genbank & | Genome & | PAB-4377 | ERX556165 | 16,117,316 | [189] |

| | | | | | | |
|---|---|---|---|---|---|---|
| | SRA | Reads | | (Bioproject ID: PRJEB7263) | (100 bp paired end) | |
| *L. peruviana* | Genbank & SRA | Genome & Reads | LEM1537 (MHOM/PE/1984/LC39) | ERX556164 (Bioproject ID: PRJEB7263) | 9,378,317 (100 bp paired end) | [189] |

**Table 3.1:** Data used in this study. In the World Health Organisation (WHO) IDs, M = mammal and R = reptile. HOM = human, CAN = canine, DAS = armadillo, CEB = primate, ARV = rodent , TAR and LAT = lizards. Data in bold was used to assemble the *L. guyanensis* and *L. naiffi* genomes in this work.

### 3.4.3 Quality Control

Quality control of the *L. naiffi* CL223, *L. guyanensis* CL085, *L. braziliensis* M2904, the five *Viannia* libraries from [179], two *L. peruviana* libraries and *L. panamensis* PSC-1 read library was carried out using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). No corrections were required for the [179] or *L. panamensis* PSC-1 libraries. Two Illumina paired-end PCR primer sequences identified by FastQC in the forward (_1.fastq) read pairs of *L. braziliensis* were removed using fastx_clipper (http://hannonlab.cshl.edu/fastx_toolkit/index.html) to remove sequences less than 50 bp after clipping. Sequence contamination was identified and removed in *L. guyanensis* CL085, *L. braziliensis* M2904 and *L. naiffi* CL223 as for *L. adleri* reads by removing reads contributing to either an abnormal GC content distribution or those identified as sequence contaminants using BLASTn [29]. Additionally, quality trimming was carried out on the *L. naiffi* library to remove low quality bases at the 3’ end of reads using Trimmomatic [23] with parameters set to phred 33 quality encoding, remove trailing bases with quality score of 30 and retain reads with a minimum length of 95 bp after trimming.

### 3.4.4 Genome evaluation, assembly and optimisation

Processed reads were assembled into contigs using Velvet v1.2.09 [34]. Assemblies for all odd numbered *k*-mer lengths in the range 21 to 75 were evaluated using default parameters. The expected *k*-mer coverage was determined for each assembly using the mode of a *k*-mer coverage histogram determined by the velvet-estimate-exp_cov.pl script in Velvet in order to optimise resolution of repetitive and unique regions of the genome [34]. This process produced *k*-mers of 61 for *L. guyanensis* and 43 for both *L. naiffi* and *L. braziliensis,* which produced assemblies with the highest N50s. Each assembly was assembled with this expected coverage, and contigs were removed if their average *k*-mer coverage was less than half the expected value because these had low sequencing coverage and may have numerous errors. An expected coverage of 16 and a coverage cutoff of 8 was applied to *L. naiffi,* an expected coverage of 19 and coverage cutoff of 8.5 to *L. guyanensis*, and an expected coverage of 28 and coverage cutoff of 14 to *L. braziliensis*.

The assembly with the highest N50 for contigs > 100 bp (longer than one read) was scaffolded using SSPACE [40]. Gaps in scaffolds were closed where possible using Gapfiller [51], and erroneous bases corrected with iCORN [52]. Mis-assemblies detected and broken using REAPR [58] were aligned to the *L. braziliensis* M2904 reference [177] excluding the bin chromosome 00 with parameters as described in Chapter 2. Contiguation of scaffolds into pseudo-chromosomes using the *L. braziliensis* genome was tested using these broken scaffolds relative to their original unbroken state to determine if removing mis-assemblies prior to contiguation of scaffolds resulted in more accurate chromosomes.

Gaps > 100 bp were reduced to 100 bp. 200 bp at the edge of each unplaced scaffold was searched against the 200 bp flanking all pseudo-chromosome gaps using BLASTn [29] to evaluate if unplaced scaffolds might bridge gaps. Unplaced bin scaffolds < 1,000 bp were discarded and the resulting assemblies were visualised and compared to *L. braziliensis* using the Artemis Comparison Tool (ACT) [55]. *L. guyanensis* CL085 bin sequences were searched against the non-redundant nucleotide database using BLASTn. Sequences with E-value < 1e-05 and percentage identity > 40% to non-*Leishmania* species were considered to be contaminants and removed.

### 3.4.5 kDNA assembly

Sequences ≥ 1,000 bases in length that could not be assigned to chromosomes (bin sequences) were searched against BLAST databases of minicircle (753 sequences; "*kinetoplast AND minicircle AND leishmania*") and maxicircle (152 sequences; query: "*kinetoplast AND maxicircle AND leishmania*") sequences downloaded from Genbank using MegaBLAST [29]. Hits were filtered to keep those with E-value < 0.01, bitscore > 100 and percentage identity > 40 to remove short hits. Sequences that had homology to both minicircle and maxicircle sequences were annotated in the bin sequences by adding '.kDNA.unassigned' or '.kDNA.maxicircles' or '.kDNA.minicircles' to the ends of their headers.

### 3.4.6 Phylogenomic characterisation

To identify the *Leishmania* species sampled from the two Colombian dogs using published datasets, a MLSA (multi-locus sequence analysis) approach was adopted using the assembled genomes. Four housekeeping genes: glucose-6-phosphate dehydrogenase (G6PD), 6-phosphogluconate dehydrogenase (6PGD), mannose phosphate isomerase (MPI) and isocitrate dehydrogenase (ICD) from 95 *L. Vianna* strains [407] including *L. guyanensis* M4147, *L. shawi* M8404, *L. naiffi* M5533 and *L. lainsoni* M6426 were downloaded from GenBank (accessions JN996517:JN996708 and JQ181608:JQ181801) and used as references to retrieve orthologs. The *L. guyanensis, naiffi, braziliensis* control and *braziliensis* reference [177] assemblies as well as *L. panamensis* MHOM/CO/1981/L13 TriTryDBv4.2, *L. panamensis* PSC-1 [188], *L. peruviana* LEM-1537 and *L. peruviana* PAB4377 [189] assemblies were searched against the reference database using BLASTn with thresholds of E-value < 0.05 and percentage identity > 70%. It should be noted that the ICD gene could not be found in TriTrypDB versions 6 or 7 of the *L. panamensis* L13 genome and so the version 4.2 genome was used. The best BLAST hit for each sequence based on E-value and bitscore was used to extract the ortholog from each assembly. The *L. peruviana* LEM-1537 genome has gaps at the MPI and 6PGD genes and so was not included in the analysis. The four housekeeping genes were concatenated in the order G6PD, 6PGD, MPI and ICD, and aligned using Clustal Omega v1.1 [332] to produce a distance matrix exported to create a Neighbour-Net network of uncorrected p-distances using SplitsTree v4.13.1 [333]. The number of differences between each sample pair was estimated by multiplying the number of sites in the alignment (2,902) with the number of substitutions per site between each pair of samples in the distance matrix.

### 3.4.7 Variant calling

The reads of each genome were mapped to its own assembly and to other genomes with a maximum insert size parameter of 1000 and exhaustive mapping enabled using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0). Duplicate reads were removed from BAM files using samtools "rmdup" and SNPs called using samtools pileup and mpileup [103] and quality-filtered as for *L. adleri*. Only SNPs found in both pileup and mpileup results were retained. Low quality and repetitive regions of the assemblies were identified and variants in these regions were masked as described in Chapter 2, section 2.4.10. SNPs were classed as homozygous if their RDAF $\geq$ 0.85 and heterozygous if it was > 0.1 and < 0.85. The genetic divergence of *L. naiffi* and *L. guyanensis* compared to L. braziliensis was quantified using the density of heterozygous and homozygous SNPs per 10 kb non-overlapping window on each chromosome using Bedtools [356] and visualised with R.

### 3.4.8 Chromosome copy number

Read depth at every base was calculated using Bedtools 'genomecov' v2.17.0 [356] using the mapped reads. The chromosome copy number and RDAF distribution of heterozygous SNPs was calculated and plotted for each chromosome as for *L. adleri*.

### 3.4.9 Gene and 10 kb loci copy number variation

To minimise potential false positives due to PCR duplicates and repetitive regions, BAM files were filtered as before to remove PCR duplicates, and filtered further for mapping quality (MQ) > 30 using Samtools view [103]. 10 kb windows across each chromosome or bin contig were assessed for copy number variation (CNV). The copy number of 10 kb windows was determined by dividing the median coverage of each window by the chromosome's (or contig's) median as in Chapter 2, section 2.4.9. Gene copy numbers were calculated in the same manner except using BAM files not filtered for MQ > 30. Loci with a copy number $\geq$ 2 were analysed for *L. naiffi* CL223, *L. guyanensis* CL085 and the *L. braziliensis* control using self-mapped reads and reads mapped to the *L. braziliensis* M2904 reference [177] for *L. guyanensis* M4147, *L. naiffi* M5533, *L. shawi* M8408, *L. lainsoni* M6426 [179], *L. panamensis* WR120, *L. panamensis* PSC-1 [188], *L. peruviana* LEM1537 and *L. peruviana* PAB-4377 [189]. *L. panamensis* PSC-1 was also self-mapped to verify that we could find previously identified amplified loci in this genome [188], including mapping *L. panamensis* WR120 to it so that any loci shared by both *L. panamensis* genomes could be obtained. BAM files of *L. naiffi* CL223*, L. guyanensis* CL085 and *L. braziliensis* M2904 self-mapped reads were visualised in Artemis to confirm and refine the boundaries of amplified loci.

### 3.4.10 Gene annotation

Annotation of the *L. guyanensis*, *L. naiffi* and *L. braziliensis* control genomes was completed using the Companion webserver [67] with parameters as for *L. adleri* except using *L. braziliensis* M2904 as the reference. A control run with the *L. braziliensis* M2904 reference genome using itself as a reference was also performed. Gene models were manually checked and corrected in the same manner as for *L. adleri*.

### 3.4.11 Identification of orthologous groups and gene arrays

Protein-coding genes from *L. guyanensis* CL085, *L. naiffi* CL223 and the *L. brazilinesis* M2904 control genome were produced from the EMBL files for each genome and these were submitted to the OrthoMCLdb v5 webserver [361] to identify their orthologous groups (OGs). 11,825 OGs with associated gene IDs in at least one of four *Leishmania* species (*L. major* strain Friedlin, *L. infantum*, *L. braziliensis* and *L. mexicana)* or five *Trypanosoma* species (*T.vivax*, *T. brucei*, *T. brucei gambiense*, *T. cruzi* strain CL Brener and *T. congolense*) were retrieved from the OrthoMCL database and compared with OGs for each genome. The copy number of each OG was estimated by summing the haploid copy number of each gene in the OG. Gene arrays in each genome were identified by finding all OGs with haploid copy number $\geq$ 2. Large arrays ($\geq$ 10 gene copies) were examined and arrays with unassembled gene copies were identified by finding those with haploid gene copy number more than twice the assembled gene number.

### 3.4.12 RNAi pathway analysis

Six genes involved in the RNAi pathway in *L. braziliensis*, viz., *tudorSN* [455], *dcl2* [455], *dcl1* [455], *rif4* [456], *rif5* [456] and *ago1* [455] were downloaded from GeneDB [457] using IDs: LbrM.32.1040, LbrM.25.1020, LbrM.23.0390, LbrM.35.6220, LbrM.33.0190 and LbrM.11.0360, respectively. Their orthologs in annotated genomes were retrieved from the OrthoMCLdb v5 webserver and OG assignments. Orthologs in genomes without clear annotation were extracted as the top BLASTn hit with E-value < 1e-5 and percentage identity > 70%.

## 3.5 Results

### 3.5.1 Quality Control and Genome Assembly

The genomes of *L. naiffi* CL223 and *L. guyanensis* CL085 were assembled from short-read data along with the *L. braziliensis* M2904 short reads as a positive control [177] (Table 3.1). This facilitated comparison with the published *L. braziliensis* M2904 genome, assembled from long and short reads [178], and enabled us to  quantify the ability of short reads to correctly and comprehensively resolve genome architecture. The three genomes were assembled and analysed in the same manner.

#### 3.5.1.1 Contamination removed from *L. braziliensis* M2904 and *L. guyanensis* CL085 reads

All sequencing reads (Table 3.1) were quality checked using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). An abnormal distribution of GC content per read, caused by an extra GC content peak outside the normal peak, for *L. braziliensis* M2904 and *L. guyanensis* CL085 reads indicated sequence contamination that was removed (Figure S1). Two Illumina paired-end PCR primers were also identified in 0.7% of 200,000 sampled forward reads in *L. braziliensis* M2904: these were removed from the initial set of 26,007,384 76 bp reads, leaving 22,006,185 for investigation (Table B1). Further evaluation using GC content filtering and the non-redundant nucleotide database with BLASTn [29] to remove contaminant sequences (Figure B1) with subsequent correction of read pairing arrangements reduced this to 17,296,309 *L. braziliensis* read pairs (67% of the initial reads).

The *L. braziliensis* M2904 reads used to assemble a control genome were used for read mapping, error correction and SNP calling [177] with the assembly created by [178] and so the contamination would not have affected the published reference assembly. However, it did reduce the number of reads mapped as can be seen in [177] where only 84% of the *L. braziliensis* M2904 short reads mapped to the *L. braziliensis* assembly, compared with 92% of reads for self-mapped *L. infantum* reads, 93% of self-mapped *L. major* reads and 97% of self-mapped *L. mexicana* reads [177].

The 8,131,246 100 bp paired-end *L. naiffi* reads and 15,272,969 100 bp paired-end *L. guyanensis* reads were filtered (Table B1) in the same manner using BLASTn and GC content results to remove putative contaminants. Low quality bases were trimmed at the 3' end of *L. naiffi* reads to remove trailing bases with a phred base quality < 30 using Trimmomatic [23] (Table B1, Figure B2). This resulted in 13,033,846 paired-end *L. guyanensis* sequences and 6,989,814 paired-end *L. naiffi* sequences – 85% and 86% of the initial reads, respectively (Table B1). 44 contigs spanning 4,566,791 bp were also excised from the *L. guyanensis* contig set using BLASTn: half had high similarity to bacterium *Niastella koreensis* (Table B2).

#### 3.5.1.2 Short-read assembly and optimisation yielded chromosomes of comparable length and completeness to other *Leishmania* genomes

Filtered reads for *L. guyanensis*, *L. naiffi* and *L. braziliensis* M2904 were *de-novo* assembled into contigs using Velvet [34]. This resulted in 10,308 contigs with an N50 of 9.6 kb for *L. guyanensis*, 14,682 contigs with an N50 of 5.7 kb for *L. naiffi,* and 13,601 contigs with an N50 of 5.1 kb for *L. braziliensis*. Scaffolds formed from the contigs and read pair information using SSPACE [40] yielded 2,800 scaffolds with an N50 of 95.4 kb for *L. guyanensis,* 6,530 with an N50 of 24.3 kb for *L. naiffi* and 3,782 with an N50 of 20.6 kb for *L. braziliensis*.

76% of gaps (3,592/4,754) in *L. guyanensis* were closed, 63% (4,096/6,530) in *L. naiffi* and 67% (4,834 /8,786) in *L. braziliensis*. The resulting number of ‘N’ bases in scaffold gaps was 17,273 for *L. guyanensis*, 56,241 for *L. naiffi* and 11,273 for *L. braziliensis*. Substitutions and short insertion/deletion (indel) errors were corrected by mapping reads to the references (Figure B3). Each step of assembly improved the corrected N50 and percentage of error free bases (EFB%) assessed using REAPR [58] (Table B3), with the sole exception of *L. braziliensis* control at the error-correction stage (Table B3), likely due to its higher heterozygosity.

Subsequently, scaffolds were evaluated and broken at putative miss-assemblies detected from the fragment coverage distribution (FCD) error and regions with low fragment coverage. Erroneous regions without a gap were replaced with N bases. This corrected 444 errors in *L. naiffi* of which 59 were caused by low fragment coverage, 206 in *L. guyanensis* of which eight were also caused by low fragment coverage, and 232 in the *L. braziliensis* control of which 57 were caused by low fragment coverage.

*L. guyanensis* and *L. naiffi* are closely related to *L. braziliensis,* so the corrected scaffolds for *L. guyanensis, L. naiffi* and *L. braziliensis* control were contiguated (aligned, ordered and oriented) using the *L. braziliensis* reference using ABACAS [63]. The output was split into 35 pseudo-chromosomes and the unincorporated segments were labeled as unassigned “bin” contigs. Alignment of the 200 bp of the end of each such bin contig against all chromosomal gap edges did not find areas of unique homology so these sequences could not fill any gaps. The pseudo-chromosome lengths approximated the length of the *L. braziliensis* reference chromosomes with the exceptions of shorter *L. guyanensis* pseudo-chromosomes 2, 4, 12 and 21, and a longer *L. naiffi* chromosome 1 (Figure B4).

Retaining bin contigs ≥ 1,000 bases reduced their number from 812 to 186 for the *L. braziliensis* control; from 3,654 to 247 for *L. naiffi*; and 1,401 to 234 for *L. guyanensis* after contaminant screening.

The final genomes had median coverage values of 56-fold for *L. guyanensis* CL085, 36-fold for *L. naiffi* CL223 and 75-fold for the *L. braziliensis* M2904 control which compares favorably with the 74-fold median coverage of short reads mapped to the *L. brazileinsis* M2904 genome [177,178] (Table B4). In addition, genome lengths were similar to those of other genomes with more than 28 Mb of sequence on pseudo-chromosomes (Table B4).

The coverage per 10 kb segment showed more amplified loci in the *L. braziliensis* M2904 control (11 totalling ~117 kb) relative to the reference (two only) assemblies [178]. This demonstrates that the control genome had few large unresolved loci and that long-read data is desirable for the resolution of amplified loci in high-quality assemblies.

#### 3.5.1.3 High nucleotide quality of *L. guyanensis* CL085, *L. naiffi* CL223 and the *L. braziliensis* M2904 control genome

The SNP sets inferred excluded variants at low quality regions (within 300 bp of scaffold edges or 100 bp around gaps) or repetitive regions detected by tantan [357], and also those failing any of the filtering criteria detailed previously. The number of bases masked ranged from 2.2 Mb for *L. panamensis* PSC-1 to 7.6 Mb for *L. peruviana* LEM1537. The LEM1537 assembly had a large number of gaps (29,202 gaps), so ~5.8 Mb of sequence at gap edges was masked (Table B5).

Self-mapped reads for *L. guyanensis* CL085, *L. naiffi* CL223 and the *L. braziliensis* M2904 control genome showed a low number of homozygous SNPs indicating high nucleotide accuracy (M2904; Table 3.2). *L. peruviana* PAB-4377 had more homozygous SNPs than *L. peruviana* LEM1537 indicating that the nucleotide level accuracy of *L. peruviana* LEM1537 was higher. A slightly higher number of homozygous SNPs were observed for the *L. braziliensis* reference and no homozygous SNPs were observed in the *L. panamensis* PSC-1 genome indicating very high nucleotide accuracy (Table 3.2). 346 homozygous variants out of 44,935 SNPs were called for the same *L. braziliensis* M2904 genome by [177]. Differing methods including masking, read filtering, SNP filtering, aligner versions and SNP-calling software versions may account for slight divergence in the output values.

| Species | Total | Homozygous | Heterozygous | SNPs per kb* |
|---|---|---|---|---|
| **Self-mapped** | | | | |
| *L. panamensis* PSC-1 | 85 | 0 | 85 | 0.003 |
| *L. peruviana* PAB-4377 | 521 | 40 | 481 | 0.019 |
| *L. guyanensis* CL085 | 717 | 12 | 705 | 0.025 |
| *L. peruviana* LEM1537 | 1,017 | 7 | 1,010 | 0.039 |
| *L. naiffi* CL223 | 14,789 | 50 | 14,739 | 0.550 |
| *L. braziliensis* M2904 | 26,043 | 68 | 25,975 | 0.884 |
| *L. braziliensis* M2904 control | 25,478 | 4 | 25,474 | 0.945 |
| **Mapped to *L. braziliensis* M2904 genome** | | | | |
| *L. panamensis* PSC-1 | 61,785 | 59,789 | 1,996 | 2.097 |
| *L. peruviana* PAB-4377 | 63,825 | 62,502 | 1,323 | 2.167 |
| *L. peruviana* LEM1537 | 65,594 | 64,050 | 1,544 | 2.227 |
| *L. panamensis* WR120 | 297,616 | 294,459 | 3,157 | 10.103 |
| *L. shawi* M8408 | 304,115 | 296,095 | 8,020 | 10.324 |
| *L. guyanensis* M4147 | 335,641 | 326,491 | 9,150 | 11.394 |
| *L. guyanensis* CL085 | 358,020 | 355,267 | 2,753 | 12.154 |
| *L. naiffi* CL223 | 567,230 | 548,256 | 18,974 | 19.256 |
| *L. lainsoni* M6426 | 654,589 | 632,285 | 22,304 | 22.222 |
| *L .naiffi* M5533 | 657,338 | 633,560 | 23,778 | 22.315 |

**Table 3.2:** Number of SNPs called for each sample when reads were self-mapped or mapped to *L. braziliensis* M2904. Self-mapped reads indicate reads mapped to their own genome. * SNPs per kb is number of SNPs/(genome length - number of masked sites).

### 3.5.1.4 Correction of scaffolds improves completeness and accuracy of contiguation

Scaffolds were broken at putative missassemblies and were contiguated using the *L. braziliensis* M2904 chromosomes with ABACAS [63]. The output was compared with chromosomes created from scaffolds that had not been broken. *L. guyanensis* CL085 chromosome 30 had an additional ~150 kb of sequence after scaffolds were broken (Figure B5). Gene models from *L. braziliensis* M2904 [177,178] were transferred to it using the Rapid Annotation Transfer Tool (RATT) [66] to verify that the incorporated sequence contained chromosome 30 genes. 36 *L. braziliensis* chromosome 30 genes were transferred to the reverse strand and 11 to the forward strand (LbrM.30.0390 to LbrM.30.0950, Figure B5), illustrating that a large amount of information was recovered by correcting scaffolds before contiguation. In addition, a single inversion on the reverse strand of the *L. guyanensis* chromosome 30 containing four *L. braziliensis* gene orthologs (LbrM.30.1320, LbrM.30.1330, LbrM.30.1340 and L.brM.30.1350) formed from unbroken scaffolds was corrected in the revision, placing the genes on the forward strand like the *L. braziliensis* reference (Figure B6).

Breaking scaffolds before contiguation also resulted in removal of an erroneous region at the 3’ end of *L. guyanensis* chromosome 1 (Figure B7) that contained two genes transferred from *L. braziliensis* chromosome 34 but had a drop in coverage at 195,199-195,210 bp where it was joined to the rest of the chromosome. BLAST alignments of *L. braziliensis* chromosomes 33 and 34 with the 3’ end of the corrected chromosome 1 showed that it had similarity to parts of chromosome 33 only (Figure B8).

#### 3.5.1.5 One large contiguation error identified and corrected

The *L. guyanensis, L. naiffi* and *L. braziliensis* control pseudo-chromosomes were visualized with the *L. braziliensis* M2904 [177] chromosomes using the Artemis Comparison Tool (ACT) [55]. This manual verification process identified only one mistake in contiguity: a 101 kb section at the 3’ end of chromosome 18 in *L. guyanensis* CL085, *L. naiffi* CL223 and the *L. braziliensis* M2904 control genome that was comprised of a part homologous to the start of chromosome 11 and a section similar to the start of chromosome 19 in the *L. braziliensis* reference. This was confirmed by alignment of the annotated genes using BLASTn [29] and corresponded to a gap in coverage at the 3’ end of the reference chromosome 18 (Figure B9). On the basis that the *L. braziliensis* reference is the most accurate genome as it was created using long Sanger sequenced reads [178], each assembly was broken at this gap and joined to the end of chromosome 18 by a 100 bp gap. Likewise, the corresponding segment of chromosome 19 was transferred to the start of that chromosome with a 100 bp gap. The chunk of chromosome 11 was transferred to the genome bin because its 5’ end did not match *L. braziliensis* M2904 chromosome 11. Gene model annotation and refinement at chromosome ends and variable regions remains an ongoing community challenge.

#### 3.5.1.6 Three putative inversions identified in *L. naiffi* and *L. guyanensis* genomes

An inversion spanning 6.7 kb was discovered on chromosome 23 of *L. guyanensis* CL085 when compared with *L. panamensis* PSC-1, *L. braziliensis* M2904 and *L. naiffi* CL223. This inversion contained two genes encoding a DHHC zinc finger domain-like protein (LgCL085_23150) and a putative palmitoyl acyltransferase 12 protein (LgCL085_231550), and had a 99 bp gap at its 3’ end. Assuming that the scaffolder was correct, the absence of a gap at the 5’ end signifies that there was support for that join by reads, meaning the inversion may be a true inversion (Figure B10). Short Interspersed Degenerated Retroposons (SIDERs) repeats have high synteny between genomes [458] and flank the homologous locus in *L. panamensis* PSC-1 (a 298 bp SIDER2 5’ of the locus and a 1,105 bp SIDER1 3’ of the locus), and they may have confounded de-novo assembly if they are also present at the corresponding locations in *L. guyanensis*.

A 2.7 kb inversion on chromosome 33 of *L. naiffi* CL223 had a 99 bp gap at the 3' end and none at the 5' end and contained a hypothetical gene (LnCL223_331450, Figure B11). The homologous loci on *L. panamensis* PSC-1, *L. braziliensis* M2904 and *L. guyanensis* CL085 all have the opposite orientation to *L. naiffi* CL223 even though *L. guyanensis* CL085 also had a 100 bp gap 3' of this locus (Figure B11). SIDER1 and SIDER2 repeats were at either side of the homologous locus in *L. panamensis* PSC-1, which suggested that repeats could have confounded its assembly in *L. naiffi* CL223.

A 4.8 kb locus on chromosome 5 of *L. naiffi* CL223 and the *L. braziliensis* M2904 reference was inverted compared with *L. panamensis* PSC-1 and *L. guyanensis* CL085 (Figure B12). It contained one gene encoding a CYC2-like Cyclin (LbrM.05.0890 on *L. braziliensis* and LnCL223_050780 on *L. naiffi)*. There were 100 bp gaps on either side of the locus in *L. braziliensis*, 99 bp (5') and 100 bp (3') gaps in *L. naiffi* and a 99 bp gap 3' in *L. guyanensis* CL085, but none 5' in CL085. *L. panamensis* PSC-1 has no gaps on either side of this locus (Figure B12) suggesting it was properly oriented in that genome and in *L. guyanensis* CL085, and so the *L. braziliensis* and *L. naiffi* orientations may be incorrect.

#### 3.5.1.7 kDNA annotation

A large number of bin contigs were annotated as kDNA minicircles: 24 for *L. guyanensis* CL085 (total length 961,641 bp) and 4 as unassigned kDNA (total length 5,090 bp); 23 for *L. naiffi* CL223 (98,031 bp); and 10 for the *L. braziliensis* M2904 control assembly (158,544 bp). No bin contigs had homology to maxicircles.

### 3.5.2 Species Identification and Divergence

#### 3.5.2.1 Identification of CL085 and CL223 samples as *L. guyanensis* and *L. naiffi*

To identify the species of MCAN/CO/1985/CL085 and MCAN/CO/1986/CL223, four housekeeping genes, glucose-6-phosphate dehydrogenase (G6PD), 6-phosphogluconate dehydrogenase (6PGD), mannose phosphate isomerase (MPI) and isocitrate dehydrogenase (ICD) from 95 *Viannia* complex samples including *L. braziliensis, L. lainsoni, L. lindenbergi, L. utingensis, L. guyanensis, L. shawi* and *L. naiffi* [407] were compared with orthologs of each gene extracted from assemblies of CL085, CL223, *L. braziliensis* M2904 reference and control, *L. panamensis* L13 TriTryDBv4.2, *L. panamensis* PSC-1 and *L. peruviana* PAB-4377. Four samples for which sequence reads were available [179] are also examined as part of the 95 samples in [407]: *L. shawi* MCEB/BR/1984/M8408, *L. guyanensis* MHOM/BR/1975/M4147, *L. naiffi* MDAS/BR/1979/M5533 and *L. lainsoni* MHOM/BR/1981/M6426 .

The *L. braziliensis* reference sequence had no sequence differences with the control genome sequence at any of the four loci. The *L. braziliensis* control G6PD gene was on a bin contig (LbrM2904.bin.473) rather than chromosome 20.1 (as for *L. braziliensis reference*, *L. naiffi* CL223 and *L. guyanensis* CL085). The concatenated sequences were aligned using Clustal Omega [332] to create a network (Figure 3.2) with SplitsTree v4.13.1 [333]. This recreated highly reticulated genetic groups as published in Figure 3 of [407] where the *L. braziliensis* M2904 was in the *L. braziliensis* cluster as was *L. peruviana* PAB-4377 (Figure 3.2). CL085 clustered within the *L. guyanensis* species complex closest to *L. panamensis* PSC-1 (Figure 3.2) with no substitutions between them and only one substitution between each of CL085 and *L. panamensis* PSC-1 compared with *L. panamensis* L13. A comparison of CL085 with the *L. shawi* samples yielded seven substitutions between MCEB/BR/1984/M8408, five with MHOM/BR/1999/L17998 and six with MHOM/BR/1999/L17997. Comparing with selected *L. guyanensis* samples yielded 10 substitutions with isolate MHOM/BR/2002/NMT-RBO 013, 21 with isolate MHOM/BR/2007/AC and 12 with isolate MHOM/BR/1975/M4147. Based on this information, CL085 was categorized as *L. guyanensis* because *L. panamensis* and *L. shawi* are not considered to be distinct species from *L. guyanensis* [128,407].

CL223 clustered in the *L. naiffi* species complex, closest to *L. naiffi* ISQU/BR/1994/IM3936 and there were only two substitutions between these two. Other related samples were MHOM/BR/1994/IM4000 with three substitutions between it and CL223, and MDAS/BR/1987/IM3280 with 13 substitutions compared to CL223. Thus, CL223 was typed as *L. naiffi* (Figure 3.2). The CL085 and CL223 samples from Colombian dogs were interspersed with samples from humans, monkeys, insects, and armadillos illustrating no clear association of host and parasite species type.

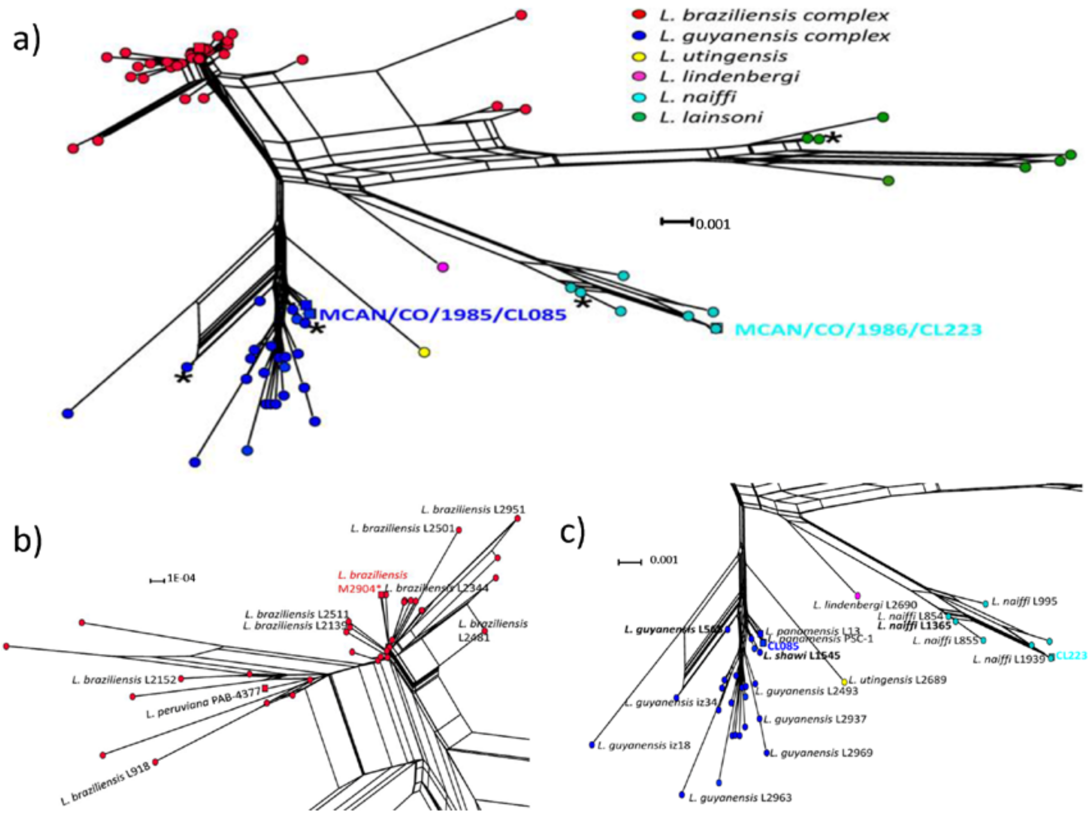


**Figure 3.2:** a) Neighbor-Net network of the concatenated nucleotide sequences of 4 housekeeping genes from 102 sequences based on uncorrected p-distances. MCAN/CO/1985/CL085 is in the *L. guyanensis* cluster in this network while MCAN/CO/1986/CL223 is placed within the *L. naiffi* cluster. The genes examined were glucose-6-phosphate dehydrogenase (G6PD), 6-phosphogluconate dehydrogenase (6PGD), mannose phosphate isomerase (MPI) and isocitrate dehydrogenase (ICD). The scale bar indicates the number of substitutions per site with 2,902 sites in the Clustal Omega v1.1 alignment. Black stars indicate samples which were also examined for aneuploidy and copy number variants in sections 3.5.3 and 3.5.4 using publically available sequencing data [179]. b) Close-up of *L. braziliensis* control and reference genome (same node whose text is marked in red) position with the *L. braziliensis* complex. *L. peruviana* PAB-4377 can also be seen. c) Close-up of CL085 within the *L. guyanensis* complex and CL223 within the *L. naiffi* complex. Nodes with bold text represent samples examined in sections 3.5.3 and 3.5.4.

#### 3.5.2.2 *L. naiffi* CL223 and *L. lainsoni* M6426 are the most divergent species from *L. braziliensis* based on SNP analysis

The divergence of *L. guyanensis* CL085 and *L. naiffi* CL223 samples as well as *L. peruviana* LEM1537, *L. peruviana* PAB-4377 [189], *L. panamensis* PSC-1 [188], *L. shawi* M8408, *L. naiffi* M5533, *L. guyanensis* M4147, *L. panamensis* WR120 and *L. lainsoni* M6426 [179] from *L. braziliensis* M2904 [177,178] was examined by mapping their reads to *L. braziliensis* M2904 (Table 3.2, Table B6). *L. braziliensis* were both closely related to *L. peruviana* PAB-4377 with 2.17 SNPs per kb of SNP callable genome (63,825 SNPs), and to *L. peruviana* LEM1537 (2.23 SNPs per kb, 65,594 SNPs). This agreed with the phylogenomic analysis and other work [459] that placed *L. peruviana* in the *L. braziliensis* species complex. The number of SNPs called for *L. peruviana* LEM1537 and PAB-4377 reads mapped to *L. braziliensis* M2904 was lower here than in [189] (144,079 for PAB-4377 and 136,946 for LEM1537 in [189]). This was likely to be caused by differences in SNP calling, filtering and masking (Table B5).

*L. panamensis* PSC-1 had only 61,785 SNPs (2.1 SNPs per kb) with *L. braziliensis* which was the lowest number of SNPs called of all the samples mapped to *L. braziliensis* (Table 3.2). *L. panamensis* WR-120 was more distant from *L. braziliensis* having 10.1 SNPs per kb and so was closer to *L. shawi* M8408 which had 10.3 SNPs per kb as well as *L. guyanensis* M4147 and *L. guyanensis* CL085 with 11.4 and 12.2 SNPs per kb, respectively (Table 3.2). *L. naiffi* CL223 had 19.3 SNPs per kb, *L. naiffi* M5533 had 22.3 SNPs per kb and *L. lainsoni* M6426 had 22.2 SNPs per kb corroborating previous findings that these two species were the most divergent in *Viannia* (Figure 3.2) [407]. The number of homozygous SNPs per 10 kb in *L. naiffi* CL223 and *L. guyanensis* CL085 when compared with *L. braziliensis* M2904 also illustrates this with *L. naiffi* CL223 appearing more divergent from *L. braziliensis* than *L. guyanensis* CL085 (Figure B13).

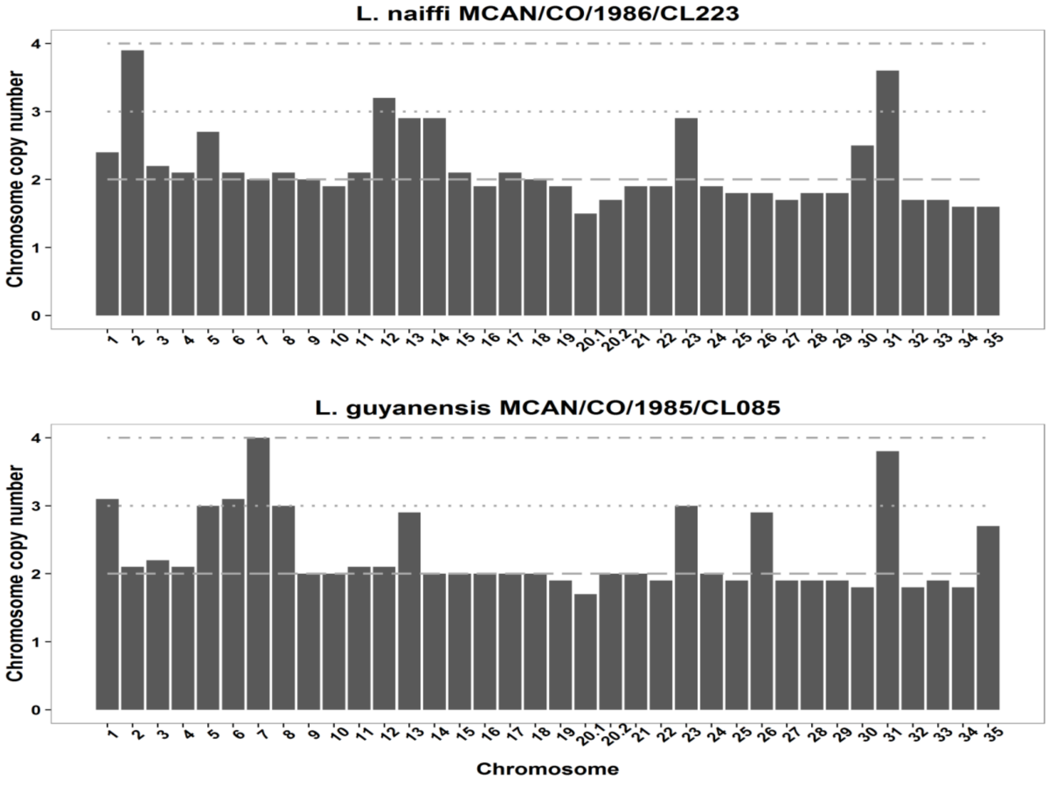
### 3.5.3 Chromosome copy number variation

#### 3.5.3.1 *Leishmania Viannia* isolates were aneuploid

Reads for the unassembled *L. shawi* M8408, *L. naiffi* M5533, *L. guyanensis* M4147, *L. panamensis* WR120 and *L. lainsoni* M6426 [179] samples were mapped to *L. braziliensis* M2904 [177,178] to estimate chromosome copy number. *L. guyanensis* CL085 and *L. naiffi* CL223 were self-mapped and also mapped to *L. braziliensis* M2904 [177,178]. Reads from the published genomes of three *Viannia* species, *L. peruviana* LEM1537, *L. peruviana* PAB-4377 [189] and *L. panamensis* PSC-1 [188] were mapped to *L. braziliensis* M2904 [177,178] to verify that mapping to a different closely related genome does not affect ploidy estimation. The chromosome copy number results of  [189] and [188] were replicated using the non-self mapped reads (Figure B14) indicating that this approach was valid. The results of [177] in which the *L. braziliensis* reference was predominantly triploid, were also replicated using the control assembly (Figure B15) demonstrating that assemblies from short read data were sufficient for the estimation of chromosome copy number. This was supported by the RDAF distributions, which had modal peaks at approximately 33% and 67% rather than a single peak at 50% at the majority of chromosomes (Figure B16).

*L. guyanensis* CL085 and *L. naiffi* CL223 were predominantly diploid (Figure 3.3). This was also supported by density plots of the read-depth allele frequency of all SNPs called from self-mapped reads for these genomes which showed a peak at ~50% for *L. naiffi* and *L. guyanensis* and a peak at ~33% for *L. braziliensis* reference and control genomes (Figure B17).  Typically, the lowest copy-number chromosomes in each were unimodal with a peak at 0.5, though the low number of heterozygous SNPs for *L. guyanensis* CL085 (623 heterozygous chromosomal SNPs) reduced power. Chromosomes 2 and 31 in *L. naiffi* CL223 were tetrasomic (Figure 3.3). The RDAF distribution for chromosome 2 of *L. naiffi* showed a large peak at 50% and two smaller peaks at 25% and 75% in comparison with that of chromosome 31, which had three approximately equal sized peaks (Figure B18). Six chromosomes (5, 12, 13, 14, 23 and 30) were trisomic and the remaining 27 were disomic (Figure 3.3 & B18). In *L. naiffi* M5533, only chromosome 2 was trisomic, as seen in both chromosome copy number and RDAF distribution plots (Figure 3.4 & B19), and all others except for chromosome 31, which was tetrasomic,  were disomic. Chromosome 31 was also tetrasomic in *L. shawi* M8408, *L. guyanensis* M4147, *L. guyanensis* CL085, *L. panamensis* WR120 and *L. lainsoni* M6426 (Figure 3.4).  In *L. shawi* M8408, only chromosome 18 appeared trisomic (copy number of 2.8) based on coverage (Figure 3.4), but its RDAF profile indicated disomy (Figure B20). *L. panamensis* WR120 had six trisomic chromosomes (1, 5, 8, 23, 25 and 26) with all others disomic, bar tetrasomic chromosome 31 (Figure 3.4). *L. lainsoni* M6426 chromosome 33 appeared to be trisomic and chromosomes 20.1, 20.2, 23 and 27 exhibited values between disomic and trisomic states (2.5-3.0), but only chromosomes 20.1 and 20.2 showed RDAF peaks indicating trisomy, whereas the remaining ones were disomic (Figure 3.4 & B21). RDAF distribution plots were largely uninformative for *L. panamensis* WR120 and *L. guyanensis* CL085 due to a lack of heterozygous SNPs, as plots here were only informative for mappings producing more than ~8000 heterozygous SNPs.

Chromosomes 1, 5, 6, 8, 13, 23, 26 and 35 of *L. guyanensis* CL085 were trisomic (total of 8 trisomic chromosomes) while chromosomes 7 and 31 were tetrasomic (copy numbers of 4 and 3.8 respectively) and the other 25 chromosomes were disomic (Figure 3.3). Although RDAF distributions of most chromosomes in this sample were uninformative, peaks indicating trisomy could be observed for chromosomes 13, 26 and 35 (Figure B22). In *L. guyanensis* M4147, 8 chromosomes were also trisomic based on read depth analysis (Figure 3.4) and the RDAF distributions (Figure B23); these were chromosomes 8,10,11,16,19,22,23 and 26, so only chromosomes 8 and 26 were also trisomic in *L. guyanensis* CL085 (Figure 3.4). All other chromosomes in *L. guyanensis* M4147, with the exception of chromosome 31, were disomic (Figure 3.4 and B23).



L. naiffi MCAN/CO/1986/CL223

L. guyanensis MCAN/CO/1985/CL085

**Figure 3.3:** Chromosome copy number of *L. guyanensis* CL085 and *L. naiffi* CL223 based on depth of coverage analysis of self-mapped reads. Dashed lines indicate 2, 3 and 4 chromosome copies.

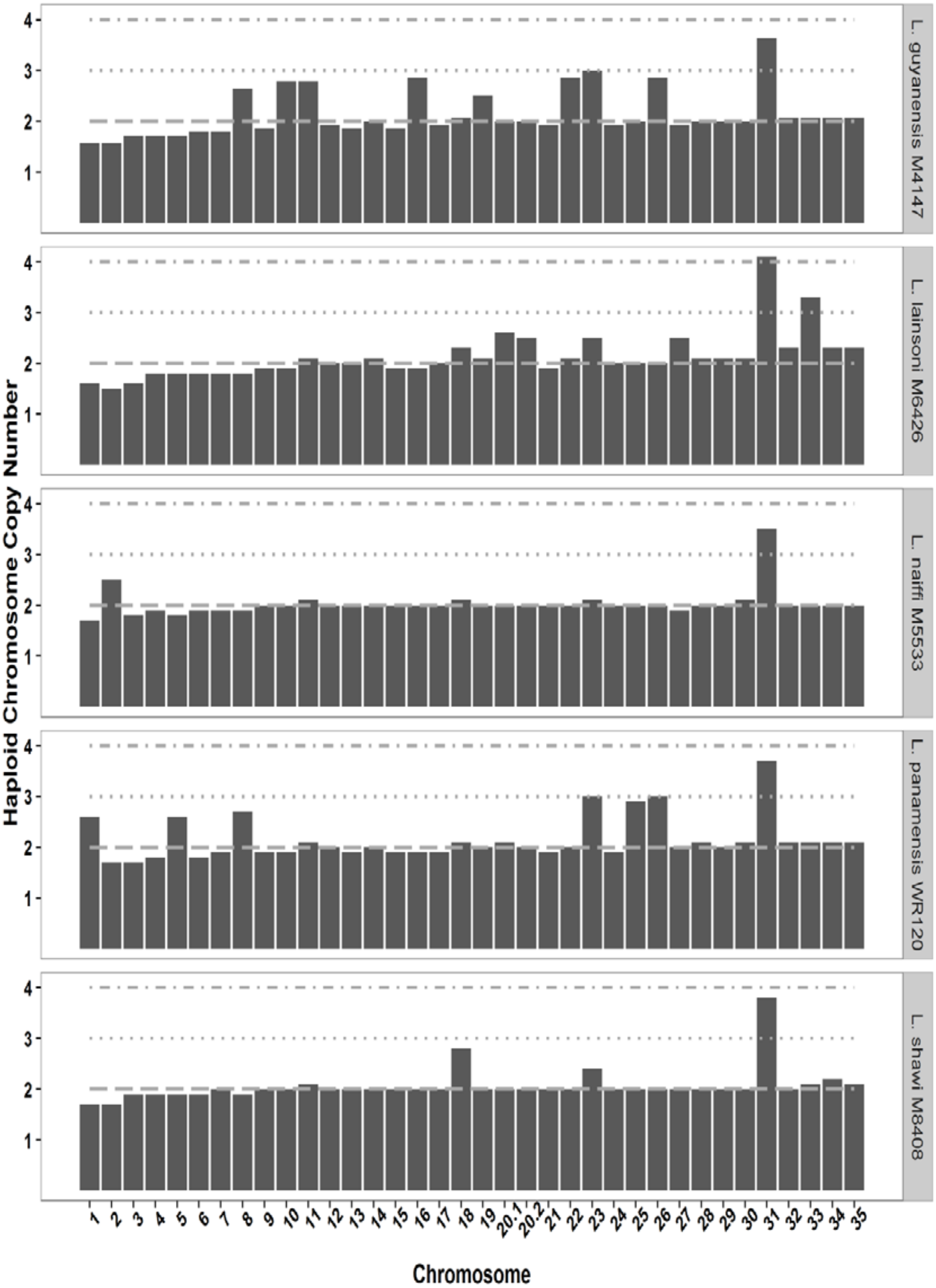


**Figure 3.4:** Chromosome copy number *L. guyanensis* M4147, *L. lainsoni* M6426, *L. naiffi* M5533, *L. panamensis* WR120 and *L. shawi* M8408 based on depth of coverage analysis of their reads mapped to *L. braziliensis* M2904. Dashed lines indicate 2, 3 and 4 chromosome copies.

### 3.5.4 Copy number variation and evidence of a minichromosome in *L. shawi*

#### 3.5.4.1 *L. shawi* M8408 has a 245 kb minichromosome previously found in *L. panamensis* PSC-1 and *L. braziliensis* M2903

We discovered a putative 245 kb minichromosome based on a uniform coverage increase of a ~130 kb locus (Figure B24 and Table B7) at the 3' end of *L. shawi* M8408 chromosome 34, which is orthologous to a 100 kb amplified locus on chromosome 34 of *L. panamensis* PSC-1 [188]. This locus is predicted to produce a minichromosome when amplified and contains a commonly amplified region in *Leishmania* called the LD1 (*Leishmania* DNA 1) region [460]. This minichromosome has previously been identified in *L. panamensis* PSC-1 [188] although the amplified locus was approximately 30 kb larger in *L. shawi* M8408 compared with *L. panamensis* PSC-1 [188]. However, it is closer to the length of the amplified locus (~120 kb) which was the source of a 245 kb minichromosome on *L. braziliensis* M2903 [461].

#### 3.5.4.2 A 45 kb locus was amplified in most *Viannia* genomes

A single 45 kb locus spanning a gene encoding a protein of the structural maintenance of chromosome (SMC) family and ten hypothetical genes on chromosome 34 of *L. panamensis* PSC-1 [188] was amplified by between two and four copies in all samples except *L. lainsoni* M6426, *L. guyanensis* M4147 and *L. peruviana* LEM1537 (Table B7). This suggested that this hypervariable amplification was common in *Viannia* but may not be ancestral to the subgenus. Using the *L. guyanensis* CL085 annotation, putative functions were assigned to five of the ten hypothetical genes. They encoded a FYVE zinc finger containing protein, a NLI interacting factor like phosphatase, a pre-rRNA processing protein PN01, a CRAL/TRIO domain containing protein and a fusaric acid resistance protein (Table B8). *L. naiffi* CL223 had two additional hypothetical genes at this locus. CRAL/TRIO domains are common in lipid binding proteins and are involved in actin remodelling, which is important for cell growth, and phospholipid transfer [462]. Proteins containing FYZE domains can bind to phosphatidylinositol 3-phosphate which is a lipid involved in the regulation of many cell processes such as cytoskeletal remodelling, cell survival, cell signalling and membrane trafficking [463].

#### 3.5.4.3 Two loci amplified in *L. adleri*, *L. guyanensis* CL085 and *L. braziliensis*

A 11 kb region on chromosome 10 and a 20 kb one on chromosome 17 with high copy numbers in two *L. adleri* samples (see Chapter 2) had amplified homologous loci on *L. guyanensis* CL085 chromosome 9 (13 kb) and 17 (19 kb) (Table B8). The 20 kb locus on *L. adleri* spanned two independent CNVs, one containing an elongation factor 1 alpha (EF-1 alpha) gene, and the other three receptor type adenylate cyclase genes. This chromosome 17 region was also amplified in the *L. braziliensis* control (Table B7) and had a partial CNV on chromosome 17 of the *L. braziliensis* reference genome (Figure B25). Consequently, this locus may not be completely assembled in the *L. guyanensis* CL085 or *L. adleri* HO174 genomes created here, and perhaps the *L. braziliensis* reference.

#### 3.5.4.4 Amplification of *L. guyanensis* CL085 transposable elements involved in RNAi

A 6 kb locus with an estimated seven copies on chromosome 16 and a 7 kb locus with approximately two copies on chromosome 20.2 of *L. guyanensis* CL085 contained the site-specific SLACS (Spliced Leader Associated Conserved Sequence) retrotransposable element and three TATE (Telomere-Associated Transposable Element) DNA transposons, respectively (Table B8). Ten haploid gene copies of SLACS were predicted in an array (OG5_127518) which contained the SLACS gene and two genes on bin sequences compared with only one SLACS copy in *L. naiffi* (Table 3.3). The three TATE sequences are in an array (OG5_132061), which also contains TATE sequences from other chromosomes and has a total of 50 haploid copies in *L. guyanensis* (Table 3.3). SLACS are found between tandem arrays of spliced leader RNA genes while TATE sequences are at telomere sequences inserted in the hexameric sequence GGGTTA [178] although TATE-like sequences have also been reported at internal positions of chromosomes in *L. braziliensis* and *L. panamensis* [188]. Both are the source of most small interfering RNAs (siRNA) in *L. braziliensis* involved in the RNA interference pathway [464] and both are absent in *L. major, L. infantum and L. mexicana genomes* [178] as well as *L. (Sauroleishmania) adleri* and L. *(Sauroleishmania) tarentolae* genomes [190]. TATE sequences were also expanded in *L. braziliensis* [465], *L. panamensis* PSC-1 [188], *L. peruviana* PAB4377 and *L. peruviana* LEM-1537 [189] and SLACs are expanded in *L. braziliensis* [465] but were not found in *L. panamensis* or *L. peruviana* genomes [188,189].

| **OG** | | **Number of assembled genes in OG** | | | **OG haploid copy number** | | |
|---|---|---|---|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. braziliensis control*** | ***L. guyanensis*** | ***L. naiffi*** | ***Lbraziliens*is control** | ***L. guyanensis*** | ***L. naiffi*** |
| **All three genomes** | | | | | | | |
| OG5_132061 | TATE DNA Transposon | 2 | 14 | 3 | 21 | 50 | 11 |
| OG5_126605 | alpha tubulin | 1 | 2 | 1 | 17 | 36 | 13 |
| OG5_130729 | amastin-like surface protein, putative | 8 | 24 | 20 | 24 | 26 | 33 |
| OG5_126631 | elongation factor 1-alpha | 1 | 1 | 1 | 12 | 18 | 20 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| OG5_126703 | polyubiquitin, putative | 2 | 2 | 1 | 21 | 16 | 43 |
| OG5_126558 | dynein heavy chain, cytosolic, putative | 14 | 13 | 14 | 13 | 13 | 14 |
| OG5_129265 | pteridin transporter, putative, folate/biopterin transporter, putative | 8 | 10 | 9 | 14 | 13 | 13 |
| OG5_143904 | amastin-like surface protein, putative | 5 | 4 | 6 | 48 | 12 | 14 |
| OG5_126623 | lipophosphoglycan biosynthetic protein, putative,heat shock protein 90, putative,glucose regulated protein 94, putative ,heat shock protein 83-1 ,lipophosphoglycan biosynthetic protein, putative | 2 | 2 | 2 | 11 | 10 | 12 |
| ***L. guyanensis* and *L. naiffi*** | | | | | | | |
| OG5_126749 | GP63, leishmanolysin, | 4 | 8 | 9 | 5 | 33 | 56 |
| OG5_126617 | receptor-type adenylate cyclase, putative | 3 | 5 | 5 | 5 | 14 | 13 |
| OG5_126611 | beta tubulin | 1 | 1 | 2 | 0 | 14 | 27 |
| OG5_128620 | NADH-dependent fumarate reductase, putative | 4 | 3 | 4 | 2 | 14 | 16 |
| OG5_126585 | kinesin K39, putative ,hypothetical protein | 7 | 10 | 10 | 7 | 12 | 11 |
| OG5_126573 | histone H4 | 1 | 7 | 4 | 3 | 11 | 10 |
| ***L. naiffi*** | | | | | | | |
| OG5_173495 | hypothetical protein | 0 | 1 | 1 | 0 | 2 | 15 |
| OG5_126568 | ABC1 transporter, putative | 9 | 10 | 12 | 10 | 9 | 14 |
| OG5_127342 | peptidase m20/m25/m40 family-like protein | 2 | 2 | 2 | 6 | 3 | 10 |
| ***L. guyanensis*** | | | | | | | |
| OG5_173452 | tuzin, putative | 2 | 3 | 1 | 1 | 19 | 1 |
| OG5_145872 | ATG8/AUT7/APG8/PAZ2, putative | 1 | 2 | 1 | 8 | 19 | 1 |
| OG5_143922 | ATP dependent DEAD-box helicase, putative | 1 | 2 | 0 | 6 | 17 | 0 |
| OG5_148241 | hypothetical protein, conserved in *Leishmania* | 1 | 1 | 1 | 0 | 14 | 1 |
| OG5_137181 | ATG8/AUT7/APG8/PAZ2, putative | 1 | 1 | 0 | 0 | 13 | 0 |
| ***L. braziliensis* control** | | | | | | | |
| OG5_126588 | heat-shock protein hsp70, putative ,glucose-regulated protein 78, putative | 4 | 3 | 3 | 14 | 7 | 8 |
| OG5_138994 | tuzin, putative | 5 | 3 | 2 | 13 | 4 | 2 |
| OG5_129839 | phosphoglycan beta 1,3 galactosyltransferase | 2 | 4 | 1 | 12 | 5 | 2 |
| OG5_12706 | thimet oligopeptidase, | 3 | 3 | 3 | 12 | 9 | 7 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | putative,metallo-peptidase, Clan MA(E), Family M3 | | | | | | |
| ***L. guyanensis* and *L. braziliensis* control** | | | | | | | |
| OG5_169610 | surface antigen-like protein | 1 | 1 | 0 | 16 | 11 | 0 |
| OG5_127518 | SLACS like gene retrotransposon element | 2 | 3 | 1 | 18 | 10 | 1 |

**Table 3.3:** Arrays with ≥10 gene copies as predicted by read depth analysis for each species. OG = orthologous group. Protein-coding genes were assigned to orthologous groups using the OrthoMCLdb version 5 webserver.

### 3.5.5 Genome annotation

#### 3.5.5.1 Automated and manual annotation identifies over 8,000 genes for *L. guyanensis* and *L. naiffi*

A total of 8,262 genes were annotated on *L. naiffi* CL223: 8,104 were protein coding genes (46 of these were manually added), 78 were tRNAs, two were snRNAs, four were snoRNAs, fifteen were rRNA genes and 59 were pseudogenes. 13 genes and one pseudogene were removed because they overlapped existing superior (often longer) gene models that had improved sequence identity with *L. braziliensis* M2904 orthologs. Most (96.2% ) of genes were located on chromosomes (Table B4).

8,376 genes were annotated on *L. guyanensis* CL085: 14 rRNAs, 2 snRNAs, 4 snoRNAs, 75 tRNAs, 51 pseudogenes and 8,230 protein coding sequences. 34 of the protein coding genes were manually added and one protein-coding gene was removed. As with *L. naiffi*, most genes were on chromosomal sequence with only 7.4 % of the total genes on unassigned bin sequence (Table B4).

269 gene models on *L. naiffi* and 198 on *L. guyanensis* with multiple joins mainly caused by the presence of short gaps were corrected by extending the gene model across the gap where the gap length was known (< 100 bp). If the gap length was unknown (> 100 bp), the gene was extended to the nearest start or stop codon.

Protein-coding genes were clustered into OGs (Table B9) using the OrthoMCLDB v5 webserver. 7,719 of the 8,001 (96.5%) genes on the *L. braziliensis* control genome clustered into 7,244 OGs. 8,137 of the 8,375 (97.2%) on the *L. braziliensis* reference grouped into 7,383 OGs. 7,961 of the 8,230 (96.7%) on *L. guyanensis* were divided into 7,381 OGs. 7,893 of the 8,104 (97.4%) on *L. naiffi* were partitioned into 7,324 OGs and 7,692 of the *L. panamensis* PSC-1 7,748 (99.3%) were split into 7,245 OGs. 6,835 OGs were shared by all

9 species considered here which were *L. major*, *L. mexicana*, *L. infantum*, *L. guyanensis*, *L. naiffi*, *L. braziliensis*, *L. panamensis*, *L. adleri* and *L. tarentolae*.

#### 3.5.5.2 Annotation of the *L. braziliensis* control identified 70 new protein-coding genes

For the control *L. braziliensis* genome, 8,161 genes (8,001 protein-coding) were computationally transferred. 76 of the annotated genes on the control genome were tRNAs, two were snRNAs, four were snoRNA's, thirteen were rRNA and 65 were pseudogenes. Manual verification of the control genome annotation was not required due to previous work on this [177,178]. 2.8% of protein-coding genes (235) in 201 OGs in the published *L. braziliensis* annotation were absent in the control, demonstrating that it recovered > 97% of protein-coding genes in OGs on the published assembly. Most un-annotated genes in OGs on the control were hypothetical or encoded ribosomal proteins (Table B10).

70 protein-coding genes in 62 OGs were present on the *L. braziliensis* control but not the published annotation – these all have a copy number of at least one on the control (Table B11).

#### 3.5.5.3 Automated computational transfer identified 99% of *L. braziliensis* protein-coding genes

The ability of the Companion pipeline to recover genes on the *L. braziliensis* genome was tested by using itself as both the input and reference genome for the pipeline. 8,661 genes were computationally transferred to the *L. braziliensis* M2904 reference genome using itself as a reference. Of these, 59 were pseudogenes, 85 were tRNAs, two were snRNAs, ten were snoRNAs, 12 were rRNAs and 8,493 were protein-coding. This process discovered 99% of the published *L. braziliensis* protein-coding genes that are in OGs, and only 81 protein-coding genes were absent (Table B12). Most of these encoded hypothetical proteins. In the *L. braziliensis* reference annotation (version 3), there were 8,620 genes in total and 66 of these were tRNAs, 7 were rRNAs, 156 were pseudogenes and 34 were non-coding RNAs.

136 extra protein-coding genes were identified compared to the current *L. braziliensis* genome. 38 of these had orthologs in other *Leishmania spp.* and 36 genes had 33 orthologs on the *L. braziliensis* control assembly (Table B13) indicating that these may be valid gene models. In at least some cases the locus may not have been fully assembled on the *L. braziliensis* reference e.g. Figure B26. Two genes on the Companion annotated *L. braziliensis* reference (one in OG5_130626 and one in OG5_164954), encoding a putative epsilon-adaptin and a hypothetical protein, were not present on any other genomes but the

epsilon-adaptin gene has orthologs in *Trypanosoma brucei* and *T. cruzi* strain CL Brener and the other gene has an ortholog in *T. cruzi* strain CL Brener.

#### 3.5.5.4 *L. naiffi* CL223 and *L. guyanensis* CL085 had few species specific genes

Only four genes in OGs unique to *L. naiffi* were identified (Table B14). Hypothetical genes LnCL223_312570 and LnCL223_292920 had orthologs in *T. brucei* and *T. vivax* respectively. LnCL223_341350 encoded a putative transferase and LnCL223_352070 a methylenetetrahydrofolate reductase (OG5_128744), but these had no orthologs in the other eight *Leishmania* spp. or five *Trypanosoma* species investigated here. The LnCL223_341350 transferase inferred amino acid sequence had 45% and 44% identity with a transferase family protein in *Leptomonas pyrrhocoris* and *Leptomonas seymouri,* repectively. A gene encoding a methyltransferase-domain containing protein (LnCL223_2021430 in OG5_129552) was in the updated but unpublished *L. braziliensis* reference genome annotation and had orthologs in other eukaryotes indicating it may be valid (Table B9, Table B13, Table B14).

*L. guyanensis* had 31 unique genes in 30 OGs (Table B15). Six were chromosomal and four of these genes were in *Trypanosoma*, encoding two hypothetical proteins, a tuzin and a poly (ADP-ribose) glycohydrolase. 28 of the 31 had orthologs in eukaryotes and 3 of these had orthologs in *Tetrahymena thermophile*, a free living freshwater ciliate protozoan (Table B15) [466].

Two genes in *L. adleri* HO174 (LaHO174_170020 and LaHO174_3623670; OGs OG5_185212 and OG5_206778), encoding a hypothetical protein and a selenoprotein that had orthologs in *T. brucei*, *T. brucei* gambiense and *T. vivax* but not *L. braziliensis*, *L. major*, *L. mexicana*, *L. infantum* or *L. tarentolae* had orthologs in *L. guyanensis*, *L. naiffi*, the control *L. braziliensis* genome and the updated reference *L. braziliensis* genome (Table B16).

#### 3.5.5.5 Twenty-three OGs were exclusive to *Viannia* genomes

The availability of OGs for nine genomes (Table B9), four in the *Viannia* subgenus, two in the *Leishmania* subgenus and two in *Sauroleishmania* allowed us to detect genes that are exclusively present in *Viannia* genomes. We found a total of 23 OGs that fit this criterion (Table B17). Aside from 4 OGs which contained DCL1, DC2 and RIF4 genes, which are involved in the RNAi pathway and TATE DNA transposons, all of which are already known to be *Viannia* specific features [177,178,453], and thirteen OGs with hypothetical genes, we identified six OGS with annotated genes. These included genes encoding eukaryotic translation initiation factor-like protein which has one copy in each genome and is also otherwise only present in *Trypanosoma brucei* and *Caenorhabditis elegans*, ABC protein PRP1 (pentamidine resistance protein 1), which has one copy in each *Viannia* genome and an OG containing both a beta tubulin and amastin-like surface protein genes with six genes in *L. naiffi*, four in *L. braziliensis* and one in the other *Viannia* genomes that has no orthologous genes in any other species outside *Leishmania*. The other three genes encoded a diacylglycerol kinase-like protein, a nucelobase transporter and an iron/zinc transporter protein-like protein and these genes all had a haploid copy number of one in the genome.

#### 3.5.5.6 Few genes were unassembled in *L. guyanensis* and *L. naiffi*

The haploid copy number for each gene is one if there is only one copy of a gene predicted based on coverage and haploid copy numbers of two indicate that there are two copies of a gene predicted based on coverage etc. Genes that have haploid copy number that is at least twice that of their assembled copy number indicate that only half the gene copies predicted by read coverage are actually assembled on the genome sequence. Thus, we looked for all OGs that had haploid copy number that was at least twice as high as the assembled copy number of genes in the OGs to quantity completeness of the assembly. Only 145 genes in 92 OGs on *L. guyanensis* (Table B18), 142 genes in 90 OGs on *L. naiffi* (Table B19) and 102 genes in 71 OGs (Table B20) on the *L. braziliensis* control met this criterion, indicating few unassembled genes in each assembly.

#### 3.5.5.7 RNAi pathway genes were present in *L. guyanensis* CL085 and *L. naiffi* CL223

RNA interference (RNAi) is a post-transcriptional gene silencing mechanism initiated by short double stranded RNA (dsRNA). RNAi is present in *Trypanosoma brucei* [467,468], *L. braziliensis* [178,453,455], *L. guyanensis*, *L. panamensis* and the non-parasitic *Crithidia fasciculata* [453], but not in the *Leishmania or Sauroleishmania* subgenera. In the RNAi pathway dsRNA is converted to small interfering RNA (siRNA) by Dicer. Five genes are in this pathway in *L. braziliensis*: cytoplasmic Dicer like 1 (DCL1, LbrM.23.0390), nuclear Dicer like 2 (DCL2, LbrM.25.1020) [178,467], Argonaute 1 (AGO1, LbrM.11.0360) [455] and RNA Interference Factors 4 and 5 (RIF4 and RIF5, LbrM.35.6220 and LbrM.33.0190) [456]. Each of these genes had orthologs in *L. guyanensis* CL085, *L. naiffi* CL223, *the L. brazilensis* control assembly, *L. panamensis* PSC-1, *L. panamensis* L13, *L. peruviana* PAB-4377 and *L. peruviana* LEM1537 (Table B21) and none in *L. adleri* or *L. tarentolae* as expected. Hypothetical gene, LbrM.29.0560, was also in an orthologous group with the AGO1 gene and has been inferred to be involved in siRNA production (GO: 0030422). The

RIF5 gene had orthologs in *L. major*, *L. infantum*, *L. mexicana* and *L. donovani* and AGO1 had mutated orthologs in *L. major* (LmjF.11.0570) and *L. infantum* (LinJ.11.0500) [178].

*L. guyanensis* and *L. naiffi* orthologs with complete ORFs were present at homologous chromosomal regions with *L. braziliensis* and had one haploid copy of each gene. The *L. panamensis* L13 AGO1 and RIF5 genes were disrupted by gaps (199 bp and 140 bp gaps respectively), and so these ORFs could not be fully examined. The *L. braziliensis* M2904 control genome had three copies of DCL1 in contrast with one copy in the published assembly. This was caused by a single 'N' base and an 11 bp gap before the 5' end of the gene that resulted in three separate annotated genes instead of just one (Figure B27).

### 3.5.6 Gene arrays in *L. guyanensis* and *L. naiffi*

#### 3.5.6.1 *L. naiffi* and *L. guyanensis* had over 300 gene arrays

There were 327 gene arrays on *L. naiffi* CL223 (Table B22), 334 arrays on *L. guyanensis* CL085 (Table B23) and 255 arrays on the control *L. braziliensis* M2904 genome (Table B24) although approximately half the arrays on each genome contained only two copies of each gene. 22 of the *L. guyanensis*, eighteen of the *L. naiffi* and fifteen of the control *L. braziliensis* genome arrays contained at least ten haploid gene copies (Table 3.3). Gene arrays were defined here as genes in the same OG with more than two haploid gene copies and so can be located in *cis* or in *trans*. The *L. panamensis* PSC-1 genome had approximately 400 tandem arrays where 71% had only two gene copies [188]. The *L. braziliensis* M2904 genome had 615 arrays using OrthoMCL v4 [177] which corresponded to 763 OGs in OrthoMCL v5. Thus, the control genome underestimated the number of gene arrays due to either the absence of an annotated gene at a locus which may contain the gene or incomplete assembly of the locus.

#### 3.5.6.2 TATE Transposons had the highest copy number *in L. guyanensis*

Nine arrays had at least ten haploid gene copies in all of *L. guyanensis* CL085, *L. naiffi* CL223 and the *L. braziliensis* M2904 control genomes (Table 3.3). TATE DNA Transposons (OG5_132061) were the most expanded array on *L. guyanensis* CL085 with 50 haploid gene copies compared with eleven copies on *L. naiffi* and 21 on the control *L. braziliensis* and 16 on *L. panamensis* PSC-1 [188]. Forty TATE DNA transposons in this array are annotated on the *L. braziliensis* M2904 assembly [177] but only two were annotated on the control genome, illustrating that the control underestimated the number of copies due to incomplete assembly of these sequences. Thus, this initial estimate of the

TATE transposon copy number for *L. guyanensis* could be improved by longer reads to assemble the transposons more completely.

#### 3.5.6.3 Leishmanolysin genes had the highest copy number in *L. naiffi*

A leishmanolysin (GP63) array (OG5_126749) had the highest haploid gene copy number in *L. naiffi* CL223 with 56 haploid gene copies compared with 33 in *L. guyanensis* CL085, 28 in *L. panamensis* PSC-1 [188] and 31 in *L. braziliensis* M2904 although this family is not expanded in *L. peruviana* LEM1537 or PAB4377. This finding was consistent with previous work on *L. guyanensis* leishmanolysin genes [469]. Leishmanolysin is a virulence factor that is highly expressed at the promastigote stage [470],  and is also important for survival of *Leishmania* during the initial stages of infection [470–473]. *Sauroleishmania g*enomes also had high haploid gene copy numbers of this array with 37 haploid gene copies on *L. adleri* and 84 on *L. tarentolae* (Table B9). *Leishmania* subgenera genomes had lower copy numbers, with thirteen haploid gene copies in this array in *L. mexicana*, fifteen in *L. infantum* and five in *L. major* (OG4_10176 for *L. braziliensis, L. mexicana*, *L. infantum* and *L. major* [177] ) .

#### 3.5.6.4 The NADH-dependent fumarate reductase gene is amplified in *Viannia*

*L. guyanensis* CL085 and *L. naiffi* CL223 had fourteen and sixteen haploid copies of a putative NADH-dependent fumarate reductase gene (OG5_128620), respectively. This gene also had sixteen copies on *L. panamensis* PSC-1 [188], 23 on *L. peruviana* PAB4377, fourteen on *L. peruviana* LEM1537 [189], and twelve on *braziliensis* M2904. However, it only has three to four copies in *L. infantum*, *L. mexicana* and *L. major* [177] and *Sauroleishmania* subgenus genomes (*L. adleri* and *L. tarentolae*) (see Chapter 2) illustrating that its copy number was high in all *Viannia* genomes sequenced to date.  It is implicated in enabling parasites to resist oxidative stress potentially aiding metastasis in mucocutaneous leishmaniasis [474,475]

#### 3.5.6.5 Tuzins had highest copy number in *L. guyanensis* and *L. panamensis*

An array of tuzin genes (OG5_173452) had a higher haploid copy number on *L. guyanensis* (nineteen copies) compared with *L. naiffi*, *L. mexicana*, *L. infantum*, *L. major*, *L. braziliensis*, *L. adleri* and *L. tarentolae* (all had one to two copies). *L. panamensis* PSC-1 also had an elevated number of tuzin copies (22) [188]. Tuzins are conserved transmembrane proteins in *Trypanosoma* and *Leishmania* with a function likely related to surface glycoprotein expression [476]. They are often contiguous with *δ*-amastin genes, whose products are abundant cell surface transmembrane glycoproteins potentially involved in infection or survival within macrophages, as they are absent in *Crithidia* and *Leptomonas* spp., who lack a stage in a vertebrate host [476]. Tuzin genes became associated with amastin genes in a common trypanosmatid ancestor and this synteny has persisted during diversification of $\delta$-amastin genes [476].

Other high copy number arrays in all three *Viannia* genomes included alpha tubulin genes, a polyubiquitin gene, an elongation factor 1-alpha gene and a putative pteridine transporter array (Table 3.3). These genes also had multiple copies in *L. infantum* JPCM5, *L. major* Friedlin, *L. mexicana* [177], *L.adleri*, *L. tarentolae* (see Chapter 2) and *L. panamensis* PSC-1 [188]. Hypothetical gene LnCL223_272760 in *L. naiffi* CL223 had a haploid copy number of 15 (OG5_173495), whereas all other genomes examined here had zero to two copies. No domains were discovered on its protein product so its function remains unknown.

# 3.6 Discussion

## 3.6.1 High-quality draft *L. guyanensis* and *L. naiffi* reference genomes

We assembled genomes for isolates *L. naiffi* MCAN/CO/1986/CL223 and *L. guyanensis* MCAN/CO/1985/CL085 from short read sequence libraries to produce high-quality draft references for studying genomic diversity in the *Viannia* subgenus. This process combined *de novo* assembly with a reference-guided approach using the published *L. braziliensis* M2904 genome. The *L. naiffi* CL223 and *L. guyanensis* CL085 were assembled into 35 chromosomes each from 6,530 and 2,800 scaffolds, respectively. The final 30.34 Mb *L. naiffi* CL223 has 96.2% of assembled sequence on chromosomes and 36-fold median coverage and the final 31.01 Mb *L. guyanensis* CL085 has 91.2% of assembled sequence on chromosomes and 56-fold median coverage.

A fundamental feature of this process was to identify and remove contamination in read libraries (*L. guyanensis* and *L. braziliensis* here) and to trim low-quality bases (*L. naiffi* here) to ensure that the reads used were informative and free of exogenous impurities. The secondary screen for contamination in unassigned contigs also removed several *L. guyanensis* ones, which also improved resulting annotation and gene copy number estimates.

The effectiveness of our strategy was tested by applying the same protocol to the *L. braziliensis* short read sequence library, which acted as a positive control and quantified the precision of the final output. This facilitated the detection of structural variation or annotation problems resulting in underestimated copy numbers at certain genes, and the observed incorrect assembly of some loci that were manually corrected. The resulting genomes were largely complete: for comparison, the control *L. braziliensis* genome covered 93.5% of the total length of the reference one, had few homozygous SNPs, and had 97% of the same protein coding genes.

There *L. guyanensis* genome has 8,230 protein coding genes and the *L. naiffi* one has 8,104: these were similar to the *Viannia* genomes of *L. panamensis* PSC-1 (7,748) [188] and *L. braziliensis* M2904 (8,357) [177]. The vast majority of protein coding gene models were computationally transferred [67] from the *L. braziliensis* M2904 reference with perfect matching, which was subsequently verified and improved manually. As for other *Leishmania* genomes [177], both *L. guyanensis* and *L. naiffi* contain unassigned bin contigs and chromosomal regions homologous to multiple chromosomal loci or containing partially collapsed gene arrays. However, the copy number of these arrays could be estimated from the read coverage where some genes of the array were present. Putative collapsed arrays were identified where they had haploid gene copy numbers more than twice the gene copy number. 90 (*L. naiffi*) and 92 (*L. guyanensis*) arrays were partially assembled based on the haploid copy number. These collapsed arrays as well as the remaining gaps and structurally repetitive regions could be resolved with the use of longer reads and additional libraries with more extensive insert size variation.

*Leishmania* genomes also have a strong conservation of gene content with few species-specific genes [177,178]. In line with this, *L. naiffi* (4) and *L. guyanensis* (31) had few species-specific genes, some of which are in *Trypanosoma* spp. Two genes previously thought to be exclusive to *L. adleri* (see Chapter 2) were discovered in *L. naiffi*, *L. guyanensis* and *L. braziliensis*. The presence of five genes involved in the RNAi pathway as well as the presence of SLACs and TATE sequences on *L. guyanensis* and *L. naiffi* also adds to the evidence for the presence of this pathway in all *Viannia*. In the case of one isolate (*L. guyanensis* M4147) also examined here, RNAi activity has been demonstrated [453].

### 3.6.2 Verification of CL085 and CL223 as *L. guyanensis* and *L. naiffi*

MCAN/CO/1985/CL085 was identified as a member of the *L. guyanensis* species complex and MCAN/CO/1986/CL223 as *L. naiffi* firstly on the basis of phylogenetic analysis of four housekeeping genes from 99 *Viannia* isolates, and secondly using genome-wide diversity from mapped sequence reads. *L. guyanensis* CL085 was found to cluster closest to *L. panamensis* PSC-1 within the *L. guyanensis* species complex. *L. panamensis* was previously proposed as a *L. guyanensis* subspecies based on internal transcribed spacer (ITS) variation [477,478], and MLEE and RAPD data indicated that *L. panamensis* and *L. guyanensis* did not constitute distinct monophyletic lines [459]. As a result CL085 was designated as *L. guyanensis* because *L. panamensis* is not a phylogenetically distinct group.

### 3.6.3 First identification of *L. naiffi* in Colombia and in dogs

*L. naiffi* has not been previously identified in Colombia or in dogs. Given that this sample was isolated in 1986, *L. naiffi* has been circulating undetected for at least 30 years in Colombia. In fact, *L. naiffi* was not formally described until 1989 [420] although *Leishmania* DNA was first isolated from 14 armadillos in Brazil in 1979 and found to be infective to hamsters [435]. It causes lesions in humans that tend to remain small [421,423,425,479] and can produce small or non-apparent infected lesions in hamsters [420]. Due to its often self-limiting and benign nature, *L. naiffi* may be underdiagnosed especially in isolated areas [424,425] and in dogs. Human CL caused by *L. naiffi* is widespread in South America, which likely results from the continent-wide range of its vectors [423] that include *Lu. paraensis, Lu. trapidoi, Lu. gomezi* and *Lu. ayrozai,* which are found in Colombia [160,480].

The nine-banded armadillo is the primary host of *L. naiffi*: these are hunted, handled and consumed in Colombia and other parts of the Americas [168,481,482]. Armadillos forage for insects at golf-clubs, gardens, sports fields and other urban and suburban areas and so are chronic pests [482]. Humans and dogs in the same vector range as nine-banded armadillos could be exposed to sandflies with *L. naiffi*: for example, three CL cases caused by *L. naiffi* followed contact with armadillos in Surinam [424].

### 3.6.4 *L. guyanensis* and *L. panamensis* epidemics in Colombia

*L. guyanensis* was the main cause of an epidemic in Chaparral County, Colombia that peaked at 2,800 human CL cases between 2003 and 2004, in contrast with the endemic infection rate of less than 10 cases per year in this area prior to 2003. It was also found in 94.6% (53 of 56) of human samples isolated in 2004 in the Chaparral, Ortega, and Rovira municipalities in 2006 [437] as well as being isolated from a dog in the same area in Chaparral in 2007 [483] and in six dogs in a rural area of Villavicencio, Colombia in 2008 [484]. Domestic transmission was suspected due to the high proportion of women and children infected in the epidemic: *Lu. longiflocosa* was implicated as the vector due to its high abundance indoors [437,485,486], although potential mammalian reservoirs remain unstudied [437]. *L. panamensis* and *L. braziliensis* caused CL in 72 dogs assisting soldiers in jungles during the largest CL epidemic in Colombian history that caused ~40,000 CL cases in 2005 to 2009 [419]. Thus *L. guyanensis* infection is present in dogs in Colombia and may be transmitted by *Lu. longiflocosa* in at least some areas.

### 3.6.5 Gene arrays are abundant in *L. naiffi* and *L. guyanensis*

We documented here over 300 arrays in *L. guyanensis* and *L. naiffi*. The largest array in *L. guyanensis* was the TATE DNA transposon array with 50 haploid gene copies, and the largest in *L. naiffi* was leishmanolysin gene array with 56 copies. Tuzin genes were present at high copy numbers in *L. guyanensis* CL085 (19) and *L. panamensis* PSC-1 (22) compared with only one to two copies of the gene(s) in this array in *L. naiffi*, *L. mexicana, L. infantum, L. major*, *L. braziliensis*, *L. adleri* and *L. tarentolae*. The high copy number of this array could be specific to the *L. guyanensis* species complex: more extensive sampling is essential to verify this. Although the function of tuzins has yet to be elucidated, it has been suggested that they may have a role in pathogenesis [487], which may be related to the ability of the *L. guyanensis* complex spp. to cause the highly destructive MCL infections, although this has not been tested.

### 3.6.6 No *Leishmania* chromosomes have stable copy number in all species

We have shown that *L. naiffi* CL223, *L. naiffi* M5533, *L. guyanensis* CL085, *L. guyanensis M4147*, *L. lainsoni M6426, L. panamensis* WR120 and *L. shawi* M8408 were aneuploid and most chromosomes in each were disomic. This was verified using RDAF distributions of both self-mapped and non-self-mapped reads. Including the other *Leishmania* sequenced to date, we found that no chromosome had a stable copy number (disomic in predominantly disomic strains or trisomic in *L. braziliensis* M2904) across all *Leishmania,* indicating that copy number variation can be tolerated for all chromosomes [177,186,188,189,211,488]. Aneuploidy has been documented in all *Leishmania* species and strains analysed to date, although the majority of chromosomes in each were disomic with the exception of *L. braziliensis* M2904 where most chromosomes are trisomic. Our observation of chromosomes with intermediate read depth values has previously been reported, and indicated that the common nature of mosaic aneuploidy in *Leishmania* in which individual parasite cells from the same isolate have differing chromosome copy numbers [363].

### 3.6.7 A 245 kb minichromosome is present in *L. shawi* M8408

The discovery of a mini-chromosome in *L. shawi* M8408 demonstrated the usefulness of our broader genomic examination of additional *Viannia* isolates. Linear minichromosomes are formed by the annealing of inverted (head-to-head) repeats to form palindromic sequences, in contrast with circular amplicons that are created by homologous recombination between direct (head-to-tail) repeat sequences [216]. Amplifications can arise as a result of drug pressure, or spontaneously as is generally the case for LD1 amplicons [460]: the type found here. The LD1 region contained a *BT1* gene encoding a biopterin transporter, so this amplification may improve pterin uptake because *Leishmania* are pteridine auxotrophs, although the function of pteridines themselves remains unclear [489]. Minichromosomes ranging in size from 180 to 250 kb have been found in *L. major, L. mexicana, L. donovani* and *L. braziliensis* [216]. This minichromsosome was homologous to 245 kb ones predicted in *L. panamensis* PSC-1 [188] and found in *L. braziliensis* M2903 [489]. This phenomenon is not confined to a particular host or species, as suggested by the sources of these isolates: *L. panamensis* PSC-1 was from a human host in Panama in 1994, and *L. shawi* M8408 was isolated in 1984 in the Pará state of Brazil from a capuchin monkey (*Cebus*).

# Chapter 4 –Identification of genomic and transcriptomic changes associated with drug resistance in Methicillin Resistant *Staphylococcus aureus* including a large tandem amplification of the SCC*mec*IV element

*I performed all analysis in this chapter.*

*Laboratory work, including both DNA and RNA sample preparation and oxacillin selection, as well as qPCR assays, were completed by Justine Rudkin, Nikki Black, Laura Gallagher and James O'Gara in Microbiology at NUIG.*

*Nanopore sequencing and assembly of one sample was carried out by Mick Watson at the University of Edinburgh.*

*RNASeq analysis of heterogenously resistant cells exposed to 0.5 and 2 µg/ml of oxacillin, which forms part of this chapter, is published in 'Waters, E.M., Rudkin, J.K., Coughlan, S., Clair, G.C., Adkins, J.N., Gore, S., Xia, G., Black, N.S., Downing, T., O'Neill, E., Kadioglu, A., O'Gara, J.P. (2016). Redeploying β-lactam antibiotics as a novel anti-virulence strategy for the treatment of MRSA infections. Journal of Infectious Diseases*, p.jiw461.

*The above publication examined the use of oxacillin to attenuate virulence using in vitro assays and in vivo models of CA-MRSA infection. It was found that oxacillin significantly attenuates virulence via down-regulation of the agr quorum sensing system and altered cell wall architecture.*

*The discovery of a tandem amplification of the SCCmecIV element in a chemostat sample forms the basis of a manuscript that has been submitted for publication as*

*Gallagher L.A., Coughlan S., Black, N.S., Lalor, P., Wee, B., Watson,M., Downing, T., Fitzgerald J.R., Fleming G.T. A., O'Gara, J.P Tandem amplification of SCCmec can drive high level methicillin resistance in MRSA.*

*Supplementary material can be found in Appendix  C*

## 4.1 Chapter Overview

### 4.1.1 Aims and Objectives

In this chapter, we sought to investigate the genomic and transcriptional responses of heterogeneously and homogenously resistant CA-MRSA isolates to various levels of oxacillin in different conditions (on agar plates and in a chemostat). To do this we examined gene expression using RNASeq data and detected mutations such as SNPs, indels and copy number variation.

### 4.1.2 Methodology

Data from two experiments was analysed in this chapter. The first experiment (referred to as the low dose experiment in this chapter) was conducted to measure the DNA and gene expression changes in *S. aureus* USA300 cells grown on agar plates, as a result of treatment using both low and high doses of the antibiotic oxacillin, compared to cells that were not treated with oxacillin. Heterotypically resistant (HeR) cells were grown in 0, 0.5 and 2 μg/ml of oxacillin and three biological replicates of each were DNA and RNA sequenced on the Illumina MiSeq producing 300 bp paired-end reads with an average of 2.4 million reads per DNA sample and 4.9 million reads per RNA sample. Homotypically resistant (HoR) cells were selected from HeR cells by growing them in 100 μg/ml of oxacillin. These HoR cells were then grown in 0, 0.5 and 2 μg/ml of oxacillin as before and again three biological replicates of each were sent for sequencing.

The second experiment involved the growth of USA300 cells in a chemostat up to 130 μg/ml of oxacillin over 13 days producing HoR cells. DNA was extracted from the parent isolate and also from sixteen samples taken at day 13 that grew on high levels of oxacillin (MIC > 800 μg/ml). These samples were sequenced on an Illumina MiSeq, producing 300 bp paired-end reads with an average of 2.7 million reads per sample.

Two analysis pipelines were developed (Figure 4.1), one to analyse RNASeq data to find differentially expressed genes at the various oxacillin doses in the first experiment and the other to analyse DNA sequencing reads from both the first (low dose) and second (high dose chemostat) experiment. The first step in the DNA and RNA analysis pipelines was quality control using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), as well as removal of adaptor sequence and low quality bases from all reads using Trimmomatic [23].

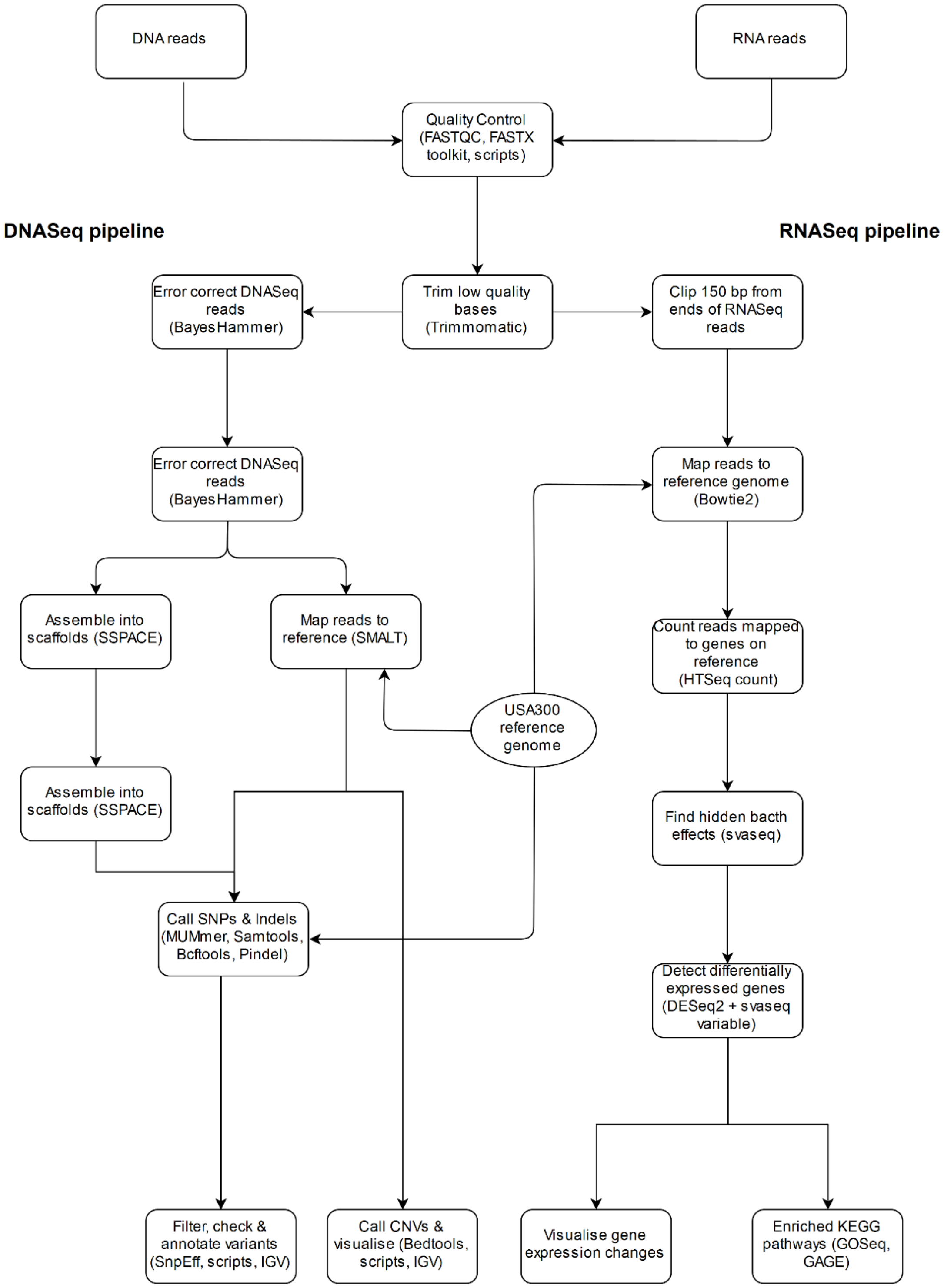

DNA reads
RNA reads
Quality Control (FASTQC, FASTX toolkit, scripts)
DNASeq pipeline
RNASeq pipeline
Error correct DNASeq reads (BayesHammer)
Trim low quality bases (Trimmomatic)
Clip 150 bp from ends of RNASeq reads
Error correct DNASeq reads (BayesHammer)
Map reads to reference genome (Bowtie2)
Assemble into scaffolds (SSPACE)
Map reads to reference (SMALT)
Count reads mapped to genes on reference (HTSeq count)
USA300 reference genome
Assemble into scaffolds (SSPACE)
Find hidden bacth effects (svaseq)
Call SNPs & Indels (MUMmer, Samtools, Bcftools, Pindel)
Detect differentially expressed genes (DESeq2 + svaseq variable)
Filter, check & annotate variants (SnpEff, scripts, IGV)
Call CNVs & visualise (Bedtools, scripts, IGV)
Visualise gene expression changes
Enriched KEGG pathways (GOSeq, GAGE)

**Figure 4.1:** RNASeq and DNASeq analysis pipelines used in this chapter. Tools used to accomplish each task are in brackets within each shape.

In the RNASeq data, ‘A’ nucelotides were overrepresented starting at approximately 150 bp on the reads and so all RNASeq reads were clipped to remove the last 150 bp. Reads were then mapped to a publically available USA300 reference genome and its three plasmids [490]. The number of reads mapping to genes on the reference genome was quantified and these counts used as input to the DESeq2 [114] R package in order to find differentially expressed genes. One surrogate variable representing latent batch effects was identified using svaseq [119] and this was included in the DESeq2 model as a covariate. Genes with adjusted p-value < 0.1 and log fold change > 0.5 were considered differentially expressed and gene expression results were visualised by grouping genes into broad functional classes. Over-represented KEGG pathways were found using GOSeq [359] where the adjusted p-value was < 0.1.

The quality processed DNA libraries were error corrected by BayesHammer [26] in order to reduce false positive variant calls and improve assemblies. These reads were then assembled using the SPAdes [27] assembler. The assembled scaffolds were aligned to the USA300 reference genome and SNPs and indels called using MUMmer [57]. Next, reads were mapped to to the USA300 reference genome using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) and SNPs called using Samtools and Bcftools [103] while indels were called with Pindel [92]. SNPs and indels called by either the assembly based strategy or the mapping based stragey were pooled into candidate SNPs. Candidate SNPs were searched for in all samples and only SNPs which passed filtering control criteria detailed in section 4.4.3.2 were considered to be true SNPs. Indels were manually checked with IGV [90] and all variants were then annotated using SnpEff [491]. CNVs were detected by examining read coverage of mapped reads in 10 kb non-overlapping windows and comparing the median coverage in these windows with the median chromosomal coverage. Read coverage across chromosomes was visualised using R.

### 4.1.3 Conclusions

This chapter contributed to the field by discovering that a novel tandem amplification of the mobile SCC*mec*IV element on the chromosome is associated with HoR in a USA300 isolate and that partial amplifications of this element and the adjoining ACME (arginine catabolic metabolic element) element may also be involved in HoR. Few gene expression differences were found when HoR groups were compared with the untreated HeR (HeR0) whereas many differentially expressed genes were found when comparing the HeR group grown in 2 μg/ml of oxacillin (HeR2) to HeR0, indicating that HeR is associated with gene expression changes whereas HoR is associated with mutation.  Mutations at three different sites on the *gdpP* gene were found to be common in HoR cells both in the low dose plate experiement cells and the chemostat, showing that it is commonly mutated in HoR cells. The purine metabolism pathway was also commonly targeted by both gene expression changes in HeR and HoR cells as its KEGG pathway was over-represented for differentially expressed genes in all comparisons with HeR0. Mutations in some HoR cells were also associated with this pathway: partial deletion of the *apt* gene was found in ten chemostat samples and the *guaA* gene had two SNPs in one chemostat sample. Further work is needed to examine the effect of the partial SCC*mec*IV and ACME amplifications as discussed in Chapter 5, section 5.7. In addition, the tool used to detect differential gene expression here (DESeq2) has been associated with a number of false positives compared with similar tools by some groups (see Chapter 5, section 5.6) although the methodology and metrics used in comparing tools can also affect how any one tool ranks amongst others and lead to different conclusions.

## 4.2 Abstract

In this study we examined the genomic and transcriptomic responses of methicillin-resistant *Staphylococcus aureus* (MRSA) to different levels of oxacillin in a continuous culture (chemostat) experiment, and homogenous and heterogeneous resistance using agar plates. This revealed a diversity of responses and adaptations, most notably a tandem amplification of the mobile genetic element SCC*mec*IV in one drug-resistant sample: this mutation included the *mecA* gene product that promotes resistance by transpeptidasing and transglycosylating peptidoglycan. We also found multiple SNPs and indels, some of which have previously been implicated in methicillin/oxacillin resistance. Heterogeneously resistant isolates exposed to low doses of oxacillin showed a mosiac of genetic changes, such as a low frequency mutation at the *rpoB* in some samples and high frequency mutations at the *pbpA* gene in other samples, in addition to differential expression of numerous genes. In contrast, few genes were differentially expressed in homogenously resistant samples and a *gdpP* mutation was present in all. Purine metabolism genes were consistently down-regulated at all doses of oxacillin in both heterogeneously reistant and homogenously resistant isolates. Moreover, purine metabolism genes including *apt* and *guaA* were mutated in a subset of chemostat samples, which highlights that this pathway may play a more significant role in the *S. aureus* response to oxacillin stress of than previously appreciated.

## 4.3 Introduction

### 4.3.1 USA300 CA-MRSA

*Staphylococcus aureus* is a well-studied, globally disseminated pathogen and human commensal that we used to explore the evolutionary basis of antimicrobial resistance (AMR) and virulence using genomics and bacteriology. Pulsed Filed Gel Electrophoresis (PFGE) type USA300 is the main cause of CA-MRSA infections in the US, with most USA300 isolates belonging to a single subtype called USA300-0114 [492,493]. USA300 is also found in Canada [494] and Europe [309] although as single cases or in small clusters in Europe where clonal lineage ST80 predominates [268,495,496]. In the US, the CDC first recognised USA300 in 2000 in an outbreak in football players in Pennsylvania [497]. USA300 is associated with skin and soft tissue infections, appears to prefentially colonise extranasal sites [498,499] and also cause more invasive disease such necrotising pneumonia [244], necrotising fasciitis [500] and severe septicaemia [501,502]. It has also become a nosocomial infection and has displaced ST5-II as the primary cause of bloodstream infections in the US [503].

### 4.3.2 The USA300-0114 genome

The USA300 FPR3757 genome (USA300-0114; ST8) is composed of a 2,872,769 bp chromosome and three plasmids (pUSA01, pUSA02 and pUSA03) and has a 23.7 kb SCC*mecIV* element [309] linked to faster growth rates [247] and no significant fitness cost [504]. It also has a mobile genetic element called ACME that encodes an arginine deiminase pathway which converts L-arginine to carbon dioxide, ATP, and ammonia, as well as an oligopeptide permease system. ACME integrates at 3’ of *orfX* on the chromosome where SCC*mec* also integrates and is flanked by repeat sequences. Its integration is most likely mediated by the cassette chromosome recombinases *ccrA/ccrB* on the SCC element [286]. Other mobile genetic elements on this genome include a staphylococcal pathogenicity island (SaPI5) encoding two enterotoxins SEQ2 and SEK2, a prophage ϕSA2usa with two genes: *lukF-PV* and *lukS-PV*, which together encode a cytotoxin called Panton-Valentine leucocidin (PVL) that targets white blood cells and a prophage ϕSa3usa encoding staphylokinase, a plasminogen activator and also a chemotaxis inhibiting protein, which is an anti-inflammatory agent. It also has non-mobile νSAα and νSAβ islands and contains gene clusters that may contribute to pathogensis [246,309].

### 4.3.3 Heterogenous resistance

Homogenously resistant (HoR) cells are selected from a heterogenous (HeR) cell population by exposure to high doses of antibiotic. Some HoR cells have been found to have mutations in the *relA* gene. RelA induces the stringent stress response in bacteria through continous over-production of (p)ppGpp, which slows down growth and dramatically represses protein production with the exception of PBP2a produced by *mecA* which has increased production [505]. However, mutations in genes involved in nutrient uptake or usage or RNA polymerase (a (p)ppGpp target) are also expected to trigger the stringent stress response [506]. Another study [507] found mutations in 27 genes where mutation of a single gene was enough to produce a HoR isolate in most cases. Six of the 27 genes carried at least two mutations, including RNA polymerase subunits *rpoB* and *rpoC* [507]. Mutations at *fem* (factor essential for methicillin resistance) [508] and *aux* (auxiliary) genes [509] have also been implicated in the conversion to HoR. Also associated with this change is the oxacillin-induced SOS stress response mediated by *lexA/recA* genes, where genes involved in DNA repair and cell survival are upregulated along with an increased mutation rate [510], dependant on the accessory gene regulator (*agr*) gene [511].

### 4.3.4 Resistance and virulence

Activation of *mecA* expression upon methicillin exposure leads to repression of the *agr* gene regulator operon [512,513], a quorum-sensing system that coordinates expression of products involved in infection, virulence and accessory functions [514]. This causes the repression of extracellular cytolytic toxin expression and an increase in expression of surface proteins [515] as well as attenuating virulence [513]. Lower virulence and toxin production associated with increased resistance is seen in HA-MRSA when compared with the more virulent and less resistant CA-MRSA [516,517].

### 4.3.5 Uncovering genetic and transcriptional controls of resistance

To discover the genetic and transcriptional controls of oxacillin resistance a HeR USA300 CA-MRSA *S. aureus* isolate was grown *in vitro* at 100 µg/ml oxacillin so that it became HoR. The original HeR isolate and HoR sample were grown with 0, 0.5, and 2 µg/ml oxacillin and their genomes and transcriptomes sequenced using the Illumina Miseq with 300 bp paired-end reads. Read-depth allele frequencies at SNPs, structural variants and differentially expressed genes were examined in three replicates at each stage to separate stochastic fluctuations from partial drug-driven selective sweeps. A USA300 isolate was grown in a chemostat for thirteen days with the concentration of oxacillin increasing from zero µg/ml to 130 µg/ml to examine genetic changes in response to high oxacillin levels. Fifteen samples surviving this and subsequent plating on 800 µg/ml of oxacillin were genome-sequenced along with the parent isolate to investigate their genomic diversity

## 4.4 Methods

### 4.4.1 Sample preparation and oxacillin selection

All work in section 4.4.1 was completed by Justine Rudkin, Nikki Black, Laura Gallagher and James O’Gara in Microbiology at NUIG.

#### 4.4.1.1 Low oxacillin dose experiment

A wild-type HeR strain of USA300 LAC CA-MRSA was streaked onto an antibiotic-free agar plate and was grown overnight to produce a stock displaying heterotypic resistance (Figure 4.2). A colony of cells was then selected from this and grown overnight (~20 hours) on an agar plate with 100 μg/ml oxacillin to produce a HoR phenotype. Nine colonies were extracted from the resulting plate and streaked onto three plates with no antibiotic, three plates with 0.5 μg/ml oxacillin and three plates with 2 μg/ml oxacillin (one colony of the HoR cells for each plate). The same procedure was carried out for the HeR cells originally grown up without any antibiotic: three plates with no antibiotic, three plates with 0.5 μg/ml oxacillin and three plates with 2 μg/ml oxacillin.

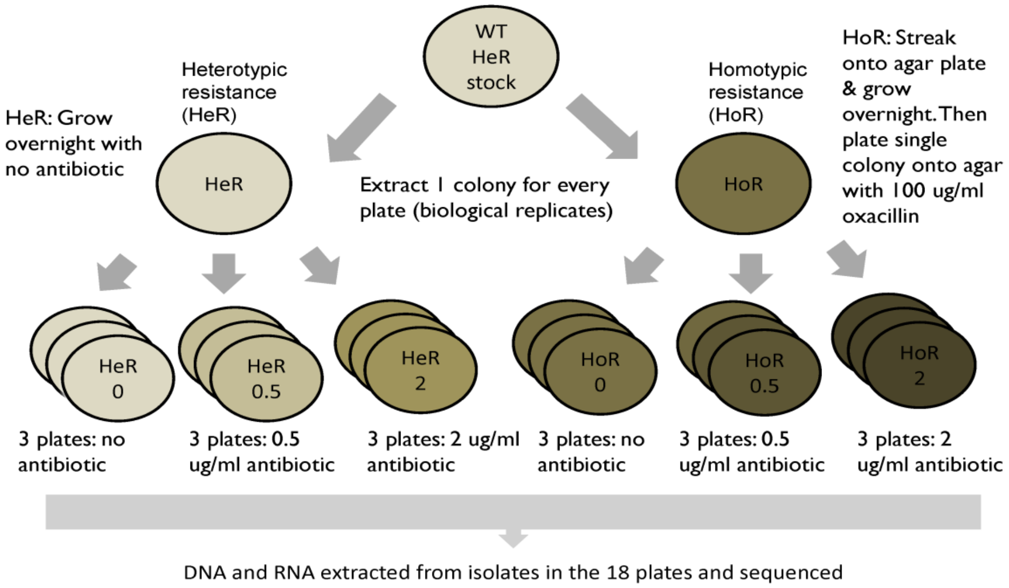

WT
HeR
stock
Heterotypic resistance (HeR)
HeR: Grow overnight with no antibiotic
HeR
Extract 1 colony for every plate (biological replicates)
Homotypic resistance (HoR)
HoR
HoR: Streak onto agar plate & grow overnight. Then plate single colony onto agar with 100 ug/ml oxacillin
HeR 0
HeR 0.5
HeR 2
HoR 0
HoR 0.5
HoR 2
3 plates: no antibiotic
3 plates: 0.5 ug/ml antibiotic
3 plates: 2 ug/ml antibiotic
3 plates: no antibiotic
3 plates: 0.5 ug/ml antibiotic
3 plates: 2 ug/ml antibiotic
DNA and RNA extracted from isolates in the 18 plates and sequenced

**Figure 4.2:** Experimental design of the low-dose experiment. Colonies from a HeR sample were grown independently in no oxacillin, 0.5 μg/ml and 2 μg/ml of oxacillin. A HeR sample was also grown overnight in 100 μg/ml oxacillin producing a HoR phenotype. Colonies selected from this were grown separately in no oxacillin, 0.5 μg/ml and 2 μg/ml of oxacillin.

#### 4.4.1.2 Bioreactor oxacillin experiment

A chemostat ran for 13 days starting at 0 ug/ml oxacillin in brain-heart infusion medium (BHI) with increasing concentrations of 8, 16, 32, 64, 100 and a final concentration of 130 µg/ml oxacillin in BHI (Figure 4.3). A USA300 HeR isolate was grown up in BHI overnight before inoculation in the chemostat. On day thirteen, 52 samples where taken from the chemostat and minimum inhibitory concentrations (MICs) checked. Fifteen of the samples grew on plates containing 800 µg/ml of oxacillin and so were selected for sequencing. These were further grown in 130 µg/ml of oxacillin before sequencing. Consequently, all samples had the HoR phenotype, the exception of the parent sample.

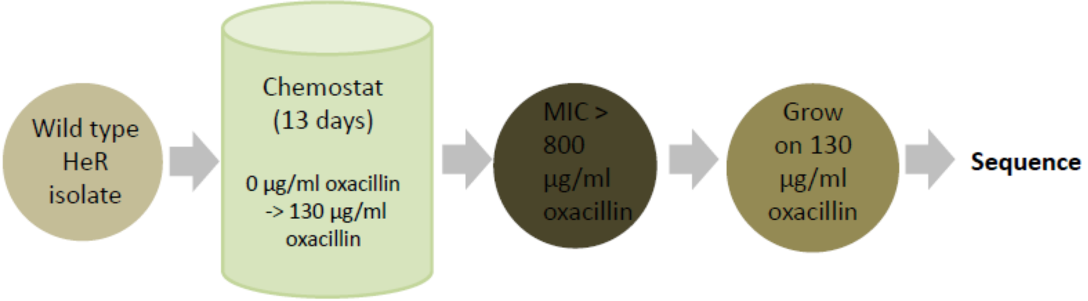

Wild type HeR isolate
Chemostat (13 days)
0 µg/ml oxacillin -> 130 µg/ml oxacillin
MIC > 800 µg/ml oxacillin
Grow on 130 µg/ml oxacillin
Sequence

**Figure 4.3:** Experimental design of the bioreactor experiment. A HeR isolate was grown for 13 days in a chemostat starting with no oxacillinand increasing in concentrations of 8, 16, 32, 64, 100 up to 130 µg/ml oxacillin. Samples were removed from the chemostat on the final day and those that grew in plates of with 800 µg/ml of oxacillin were selected and further grown on 130 µg/ml of oxacillin before being sent for DNA sequencing.

### 4.4.2 Quality control of data

DNA and RNA for 18 samples from the low oxacillin dose experiments was sequenced separately on two lanes of an Illumina MiSeq platform by Fugu (Helsinki) producing 301 bp paired-end reads with one to three million reads per sample. RNA sequencing for four was ineffective, leaving 14 in total (Table 4.1). For the bioreactor experiement, DNA from 16 samples (Table 4.2) was sequenced in the same manner producing one to four million reads per sample.

| Type | DNA | RNA |
|---|---|---|
| **HeR0** | 1a_S1 | 1_S1 |
| **HeR0** | 1b_S2 | None |
| **HeR0** | 1c_S3 | 2_S2 |
| **HeR0.5** | 2a_S4 | 4_S3 |
| **HeR0.5** | 2b_S5 | None |
| **HeR0.5** | 2c_S6 | 6_S4 |
| **HeR2** | 3a_S7 | 7_S5 |
| **HeR2** | 3b_S8 | 8_S6 |
| **HeR2** | 3c_S9 | 9_S7 |
| **HoR0** | 4a_S10 | 10_S8 |
| **HoR0** | 4b_S11 | 11_S9 |
| **HoR0** | 4c_S12 | 12_S10 |
| **HoR0.5** | 5a_S13 | 13_S11 |
| **HoR0.5** | 5b_S14 | 14_S12 |
| **HoR0.5** | 5c_S15 | 15_S13 |
| **HoR2** | 6a_S16 | None |
| **HoR2** | 6b_S17 | None |
| **HoR2** | 6c_S18 | 16_S14 |

**Table 4.1:** DNA and RNA sample names in the low-dose experiment. The 'Type' column indicates the type of resistance and 0, 0.5 and 2 are the µg/ml of oxacillin that samples were grown in: HeR0 indicates the wild type isolate as no oxaillin was used, HeR0.5 indicates that the wild-type was grown in 0.5 µg/ml and HeR2 that the wild-type was grown in 2 µg/ml of oxaillin. HoR samples were cultured from HeR0 by growing cells in 100 µg/ml of oxacillin: HoR0 indicates that the cells were not not grown in any additional oxacillin, HoR0.5 HoR in 0.5 µg/ml of oxaillin, and HoR2 HoR in 2 µg/ml of oxacillin.

| Type | DNA sample name |
|---|---|
| **HeR parent** | 1A_S1 |
| **HoR** | 2A_S2 |
| **HoR** | 3A_S3 |
| **HoR** | 4A_S4 |
| **HoR** | 5A_S5 |
| **HoR** | 6A_S6 |
| **HoR** | 7A_S7 |
| **HoR** | 8A_S8 |
| **HoR** | 9A_S9 |
| **HoR** | 10A_S10 |
| **HoR** | 11A_S11 |
| **HoR** | 12A_S12 |
| **HoR** | 13A_13 |
| **HoR** | 14A_14 |

| | |
|---|---|
| **HoR** | 15A_S15 |
| **HoR** | B1_S16 |

**Table 4.2:** DNA sample names for samples in the Bioreactor experiment. Sample 1A_S1 is the HeR parent isolate and the 15 HoR samples were removed from the chemostat at the end of a 13 day exposure to oxacillin.

Read quality was assessed by screening the read length, nucleotide and quality score distributions using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The DNA and RNA reads were trimmed based on quality scores. Potential adaptor sequence was removed using Trimmomatic v0.32 [23], which scanned reads using a four-base sliding window and trimmed reads where the average Phred base quality of the window was below 30. All ambiguous 'N' bases and reads shorter than 35 bp were removed. The first 20 bases of the DNA reads in both experiments were removed because they had a nucleotide content that deviated from the expected 25% rate for each base. The RNA sequences showed an increase in 'A' nucleotides after approximately 150 bases (Figure C1), so only the first 150 bases of each read were retained, which corrected this effect (Figure C2). The DNA reads were corrected using BayesHammer [26] to reduce sequencing errors that can reduce the alignment quality, increase false positive SNP rates and reduce the number of valid SNPs [25].

These steps retained 87% and 86% of the initial reads on average for the DNA and RNA in the low-dose experiment respectively (Tables C1 & C2), and 84% of the DNA reads in the bioreactor experiment (Table C3). This yielded median quality values > 30 across the reads. Insert sizes were an average of 227 bp for the DNA and 111 bases for the RNA in the low-dose experiment (Tables C4 and C5), and an average of 185 in the bioreactor experiment (Table C6). Read lengths after trimming and filtering averaged 170 bp for the DNA and 102 bp for the RNA in the low dose experiment, and 185 bp for the DNA reads in the bioreactor experiment (Tables C1, C2 & C3). The average coverage per sample on the chromosome, calculated using the Bedtools genomecov function [356] on mapped reads, ranged from 35 to 194 in the DNA samples (Table C7) in the low-dose experiment and 47 to 197 in the DNA samples in the bioreactor experiment (Table C8).

### 4.4.3 Structural variant discovery

#### 4.4.3.1 Genome assembly

The error-corrected paired and unpaired reads for each DNA sample were assembled using SPAdes v3.1.1 [27] with *k*-mers 21, 33, 55, 77, 99 and 127 and the 'careful' parameter, which minimized the number of mismatches in the contigs [27]. The resulting assemblies were compared to the reference USA300_FPR3757 [309] chromosome using QUAST v2.3 [56]. The GC content of each assembly was 32.6%, and there were between 31 and 51 scaffolds per assembly, with N50 values > 200 kb. One or two short gaps (<500 bp) were found in each assembly that could not be fully closed using Gapfiller [51].

#### 4.4.3.2 SNP calling using assembly and read-mapping

The chromosome and three plasmids (GenBank accessions NC_007790-NC_007793) were indexed with *k*-mer of thirteen and step size of two using SMALT v5.7 (http://www.sanger.ac.uk/science/tools/smalt-0). The error-corrected DNA reads were mapped to the genome with SMALT, which applied a Smith-Waterman sequence alignment algorithm. The SAM (sequence alignment/map) files were converted to BAM (binary alignment/map) files using Samtools v0.1.18 [103]. The BAM files were then coordinate-sorted, the paired and unpaired files were merged, and PCR duplicate reads were removed. Candidate SNPs were detected where the base quality (BQ) was >25, the mapping quality (MQ) was >30, and the read depth was <100 using Samtools Mpileup v0.1.18, Bcftools v0.1.17-dev, and the Samtools v0.1.11 vcfutils.pl function. The read depth allele frequency of the non-reference allele (RDAF) and local coverage were estimated using Samtools Pileup v0.1.11.

To call SNPs using an assembly-based approach, the scaffolds produced by SPAdes were aligned to the USA300 reference genome using nucmer in MUMmer v3.23 [57]. This was followed by eliminating conflicting repeat copies using the 'delta-filter' command and the 'show-snps' comand to call SNPs and indels. The union of SNPs called by nucmer and SNPs called by Bcftools was used as a candidate SNP set. These sites were queried across all samples using the Samtools Pileup files to find false negative SNPs uncalled by nucmer or Bcftools. The RDAF of the non-reference alleles was reported for each SNP using Samtools Pileup output. Each candidate SNP was assessed using the following additional criteria:

1) SNP Quality (SQ) >30

2) read coverage >5

3) forward-reverse read coverage ratio between 0.1 and 0.9

4) non-reference read allele frequency >0.1

5) 2+ forward reads

6) 2+ reverse reads

Results were converted to variant call format (VCF) and annotated. SNPs were homozygous if the RDAF was ≥ 0.85 and heterozygous if 0.1 < RDAF < 0.85. Insufficient read depth coverage was present to predict SNPs with RDAF < 0.1. Candidate RNA SNPs were called in the same manner using Samtools and Bcftools, omitting the Nucmer step.

#### 4.4.3.3 Indel calling using split-read mapping

Deletions and short insertions (indels) were called using the samtopindel script to convert the BAM files, and then with Pindel [92] to only keep indels with at least ten supporting reads. The RDAF of the indels smaller than the read length were calculated using the BAM files in IGV (number of reads with indel at locus / all reads at the locus). For indels greater than one bp in length, the sum of the number of reads with the indel was divided by the sum of the number of reads at each site in the indel. This approach may be limited by uneven coverage at a locus. If the indel was longer than the read length, then a lack of read coverage at the sites predicted to have the mutation was considered evidence of the deletion and the RDAF was set to one.

#### 4.4.3.4 Variant annotation

The functional effect of SNPs and indels was estimated by annotation with SnpEff v4.0e [491] using the 'Staphylococcus_aureus_USA300_FPR3757_uid58555' database file from the SnpEff database. Results were manually checked using the reference genome annotation.

#### 4.4.3.5 Copy number variation detection using read coverage

Copy number variants (CNVs) were screened using the BAM files containing reads with MQ > 30 to reduce false positive rates [89,518,519]. Coverage was calculated for every base using genomecov in Bedtools with the '-d' flag [356] so that the median chromosomal coverage could be calculated for each sample. Genome-wide coverage levels were analysed in 10 kb and 25 kb windows and plotted as 5 kb sliding windows with a 2.5 kb step using the

Bedtools makewindows function [356]. Coverage for each window was normalised by dividing it by the median coverage of the chromosome to produce a copy number estimate. Windows with copy number $\geq$ 2 were reported. The copy number of plasmids was determined by dividing the median read coverage of the plasmid by the median read coverage of the chromosome.

To test if additional SCC*mec*IV cassette copies were tandemly repeated in the chromosome, the first 300 bp and last 300 bp of the SCC*mec*IV sequence were joined in four different orientations (end to end, start to start, start to end, and end to start) to test for reads spanning putative junctions of one SCC*mec*IV element with the other. Reads from every sample were mapped to the joined sequence using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) and Samtools (as above for CNVs) to determine the copy number, which was examined in IGV [90]. Coverage across the 600 bp repeat was measured using the median of 50 bp windows with a 25 bp step size normalised by the median coverage across the 600 bp, and visualised using ggplot2 in R.

To test for the presence of amplifications in an *S. aureus* genome with high oxacillin MIC, two sets of *S. aureus* COL genome paired-end Miseq reads and two sets of single-end Ion torrent reads [520] were downloaded from the ENA using accessions in Table C9. The genome sequence and associated gff file was downloaded from ftp://ftp.cebitec.uni-bielefeld.de/pub/GABenchToB/references/ [520]. The reads were mapped to the genome separately for each set, and copy number in 10 kb non-overlapping windows was estimated as before.

### 4.4.4 Gene expression

#### 4.4.4.1 Differential gene expression

Quality-screened paired-end RNA reads were mapped to the reference chromosome using Bowtie2 v2-2.1.0 [80]. SAM files were converted to BAM files and sorted by name (samtools sort –n) using Samtools v0.1.18 [103]. Reads mapping to genes were counted using the htseq-count script from the HTSeq Python package [111] with counting mode ‘intersection non-empty’ and ‘stranded = no’ parameters and the GFF3 file for the reference chromosome (accession: NC_007793.1). This yielded a count for each of the 2,632 genes in each of the 14 RNA samples – the 16 ribosomal RNA genes were not evaluated.

One surrogate variable was discovered using the surrogate variable analysis (SVA) package v3.12.0 [119,521], which was included as an adjustment factor in the differential expression model to remove technical variation. Evidence for differential expression was tested with DESeq2 v1.6.1 [114]. Naively, p-values and adjusted p-values would reflect the evidence that $\log_2$ fold-change ($\log_2$FC) >0 [114]. Applying first a false discovery rate (FDR) threshold on p values for genes with $\log_2$FC >0 and then second applying a more strigent FC cut off (such as $\log_2$FC >0.5) would cause the FDR to decrease by an unknown amount [522]. Instead, genes were differentially expressed between conditions if there was: (i) a $\log_2$FC >0.5 using a Wald test in DESeq2 [114]; and (ii) a FDR <0.1 (Benjamini-Hochberg adjusted p value). These criteria ensured that the effect size was large enough for the results to be both statistically unexpected and biologically of interest. The p-value distribution of the test for each comparison was approximately uniform, as expected where the null hypothesis of $\log_2$FC <0.5 was generally accepted (Figure C3).

#### 4.4.4.2 GO and pathway over-representation analysis

Pathway and gene ontology (GO) over-representation analysis was executed for the differentially expressed genes using GOseq v1.18 in R [359]. Eight differentially expressed genes (*lacC, lacB, lacE, lacD, lacA, lacG, lacF* and *nuc*) were removed due to conflicting direction of expression in HeR2 vs HeR0.5 and HeR0.5 vs HeR0 comparisons, suggesting that their changes were likely to be variation unrelated to oxacillin exposure. The Python bioservices package [523] was used to download GO terms for protein-coding genes from the QuickGO database (http://www.ebi.ac.uk/QuickGO) using their RefSeq protein IDs. Genes were adjusted for their lengths using GOSeq [359]. Genes present in every KEGG pathway were retrieved using the ‘kegg.gsets’ function in the GAGE R package v2.16.0. Over-represented GO terms and KEGG pathways had Benjamini-Hochberg adjusted p-value < 0.1 using the wallenius method in GOseq.

#### 4.4.4.3 Visualisation of differentially expressed genes

Genes were classified into broad functional groups using Clusters of Orthologous Groups (COGs) annotation [524] and the KEGG Orthology (KO). KO annotation was retrieved from the htext file in the KEGG database (http://www.kegg.jp/kegg-bin/get_htext?saa00001.keg) and gene to pathway to pathway-group mappings were tabulated from this file with a Python script. COG categories were retrieved for each gene by searching its protein sequence using the NCBI conserved domain database live search tool (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) against the COG database with default parameters (E-value < 0.01 and maximum number of domain hits set to 500). Hits with the lowest E-value and highest bitscore were retained if there were multiple matches. COGs were assigned to their broader COG categories. Genes were categorised based on both COG and KO annotation. The term 'Unknown function' was applied to genes with no KO or COG assigned and to those assigned to COG categories 'R' and 'S', which contained COGs with only general (category R) or unknown (category S) functions.

KEGG pathway images were rendered using the R pathview package v1.6.0 [525] using the differentially expressed genes and their $\log_2$FC values. KEGG pathways for *S. aureus* USA300_FPR757 (denoted 'saa' in KEGG) were retrieved using the '*kegg.gsets'* function in the R GAGE [526] v2.16.0 package: pathways with at least one differentially expressed gene present in them were retained and visualised.

#### 4.4.4.4 Measuring effects of mutations on gene expression

Multiple methods were used to examine the effects of chromosomal DNA SNPs on gene expression rates. 14 samples were examined, excluding four with no RNA partner (HeR0_1b_S2, HeR0.5_2b_S5 HoR2_6a_S16 and HoR2_6b_S17). Unexpressed and ribosomal RNA genes (n=11) were excluded. The remaining 2,637 genes were re-normalised using an rlog transformation as $\log_2$ counts that minimised differences between samples for low-expression genes by scaling the expression relative to the average across the 14 samples.

The correlation of the genomic and transcriptomic changes across the 14 samples was evaluated using co-inertia analysis (CIA), which identified covariance between individual SNPs, their frequencies and gene expression differences [527]. The scaled DNA-RNA correlation was calculated as an RV coefficient (covariance(x,y)$^2$ / variance(x).variance(y)) using Made4 v1.40 [528]. CIA transforms all negative values in the expression data to positive ones by adding an integer and next performs non-symmetric correspondence analysis (NSCA) to compute an RV value: this does not require complete overlaps between gene sets in the different samples [529]. Permutation tests were implemented using the 'RV.rtest' function with 100 Monte Carlo test replicates.

A second dataset classified the genotypes as either homozygous (RDAF $\geq$ 0.85) or heterozygous (0.1< RDAF <0.85): this was used for expression quantitative trait (eQTL) analysis. Five latent factors representing putative mutational effects were estimated using PEER v1.0 [120] for the normalised expression data with R v2.15.0. PEER tolerates negative values in the rlog-adjusted normalised expression data, whereas SVA requires positive (read count) values. The residuals of latent factors were used as expression values in a linear model (expression = $\alpha + \sum_k \beta_k \cdot \text{covariate}_k + \gamma \cdot \text{genotype\_additive}$) during eQTL analysis with Matrix eQTL v2.1.0 [530] in R v3.1.1 where the maximum cis-eQTL distance was 1 kb. The deviation of the observed $\gamma$ from the expected value was quantified using a t-statistic with thresholds of $p<0.02$ for cis-eQTLs and $p<0.01$ for trans-eQTLs with Benjamini-Hochberg-adjusted $p<0.05$. P value distributions and quantile-quantile (QQ) plots of the test statistics were examined to ensure the assumptions of the tests were met.

## 4.5 Results

### 4.5.1 Genetic background of the USA300 strain in the low-dose experiment

Eighteen SNPs, seven deletions and one insertion were fixed (homozygous) in all eighteen DNA samples compared to the reference sequence (USA300_FPR3757). Of the SNPs, eight were intergenic, seven were missense and three were synonymous (Table C10). Seven SNPs were polymorphic in all eighteen samples (0.1 < RDAF < 0.85, Table C11). No polymorphic indels were discovered in any sample.

### 4.5.2 RNA yielded SNPs absent in some DNA samples in the low dose experiment

96 SNPs were found in at least one of the DNA samples, in addition to the 18 fixed SNPs and 7 polymorphic in all (Table C12). Of the 96, 46 were called by both assembly alignment and by read-mapping, 30 exclusively by alignment, and 20 exclusively by read-mapping (Table C12). All 18 fixed SNPs were called by both alignment and read-mapping (Table C10), and five of the seven SNPs polymorphic were called only by read-mapping; the other two called by both alignment and read-mapping (Table C11). The large number of SNPs exclusively called by each caller demonstrates the usefulness of combining multiple callers to pin-point potential false positive SNPs.

A search of these SNPs across all RNA samples yielded six SNPs present in the RNA absent from a DNA sample: the RDAF of these RNA SNPs was used. Two of these were intergenic and fixed in all samples, and so part of the genetic background (Table C10). The other four RNA SNPs were polymorphic in different samples.

### 4.5.3 A *gdpP* mutation enabled resistance to high levels of oxacillin

Two SNPs were fixed in all HoR samples and absent in all HeR samples (Table 4.3). One was a nonsense mutation at the *gdpP* (SAUSA300_0014) gene, and the other a missense mutation at a hypothetical gene (SAUSA300_0749). A missense mutation at a putative transposase (SAUSA300_0028) had insufficient sequence quality (BQ=20, MQ=30) to be considered valid: this SNP was present in all three HoR0 samples, one of three HoR2, and one of three HeR0 (1b_S2 with an RDAF of 0.93).

The HoR SNP (*gdpP*-R540*) was confirmed in the corresponding HoR RNA and was absent in the HeR RNA (Table 4.3). No evidence of differential expression of the whole gene nor for the gene partitioned into two regions (18,345-19,964 and 19,965-20,312) corresponding to the 5’ and 3’ ends of this SNP, indicating R540* did not change *gdpP* expression.

| **Gene symbol** | ***gdpP*** | **-** |
|---|---|---|
| **Locus tag** | SAUSA300_0014 | SAUSA300_0749 |
| **Product** | GGDEF domain protein containing phosphodiesterase activity | hypothetical protein |
| **Mutation** | Nonsense | Missense |
| **Codon Change** | Cga/Tga | gGa/gAa |
| **Amino Acid effect** | R540* | G21E |
| **DNA effect** | 1618C>T | 62G>A |
| **Position on chromosome** | 19,962 | 836,014 |
| **REF** | C | G |
| **ALT** | T | A |

**Table 4.3**: Mutations associated with high dose oxacillin exposure (≥ 100 µg/ml) and the HoR phenotype. *GdpP* and a hypothetical protein each had a fixed SNP in all HoR that were absent in all HeR.

### 4.5.4 *RpoB* and *pbpA* mutations were associated with low dose oxacillin stress

21 SNPs were present in at least one HeR sample and absent in all HoR (Table C12). Of these, five were at least two samples in either HeR0.5 or HeR2: one was at *rpoB* (RNA polymerase beta subunit), two were at *pbpA* (pencillin-binding protein A, SAUSA300_1075), resulting in a 2 bp substitution, one was intergenic, and one was at a hypothetical gene (SAUSA300_1240) (Table 4.4).

| Annotation | **Gene symbol** | ***rpoB*** | - | ***pbpA*** | - |
|---|---|---|---|---|---|
| | **Locus tag** | SAUSA300_0527 | Intergenic | SAUSA300_1075 | SAUSA300_1240 |
| | **Product** | DNA-directed RNA polymerase subunit beta | - | penicillin-binding protein 1 | hypothetical protein |
| | **Mutation** | Missense | Intergenic | Missense - 2bp substitution | Missense |
| | **Codon Change** | aCa/aTa | - | TGg/AAg | atG/atC |
| | **Amino Acid effect** | T518I | - | W351K | M58I |
| | **DNA effect** | 1553C>T | - | 1051TG>AA | 174G>C |
| | **Position on chromosome** | 587,106 | 671,174 | 1,174,871-1,174,872 | 1,358,918 |
| | **REF** | C | A | T | G |
| | **ALT** | T | G | A | C |
| **HeR0.5** | **2a_S4** | 0 | 0.17 | 0.95 | 1 |
| **HeR0.5** | **2b_S5** | 0 | 0.2 | 0.97 | 1 |
| **HeR0.5** | **2c_S6** | 1 | 0 | 0 | 0 |
| **HeR2** | **3a_S7** | 0.21 | 0 | 0 | 0 |
| **HeR2** | **3b_S8** | 0.2 | 0 | 0 | 0 |
| **HeR2** | **3c_S9** | 0.18 | 0 | 0 | 0 |

**Table 4.4:** SNPs and insertions associated with low dose oxacillin exposure (≤ 2 µg/ml on HeR samples). The values in each cell of the table are the RDAFs of the alternative allele in each sample and zero values indicate that the SNP was not present in a particular sample.

*RpoB*-T518I was homozyous in the HeR0.5 sample and was heterzygous (RDAF~0.2) in the three HeR2 samples, and was absent in any HeR0 or HoR0/0.5/2 samples (Table C12). *RpoB*-T518I may represent a mechanism to improve growth at low (0.5-2 ug/ml) oxacillin concentrations, however, its absence in two HeR0.5 samples suggested that it was not necessary to survive at low oxacillin levels.

A pair of adjacent SNPs, representing a 2 bp substitution that caused a W351K missense mutation, occurred at the *pbpA* gene in two of the three HeR0.5 samples. *PbpA*-W351K was coupled with a nonsynonymous SNP at a gene (M58I at SAUSA300_1240) encoding a hypothetical protein belonging to the UPF0154 superfamily of short bacterial proteins whose function is unknown.

### 4.5.5 Purine metabolism was down-regulated by oxacillin exposure

Purine metabolism pathway genes were down-regulated by oxacillin exposure as indicated by pairwise comparisons of HeR0 with all other groups (HeR0.5, HeR2, HoR0, HoR0.5 and HoR2) (Table 4.5). Genes down-regulated in all comparisons were: *purF*, *purM, purN, purH* and *purD*. Six other genes were down-regulated in some comparisons (*purE, purK, purC, purS, purQ* and *purL*) (Table C13). Genes involved in purine metabolism also had the largest drops in expression of all genes examined (6.15 to 15.59-fold). The decrease in expression of *purM* and *purH* was ~20-fold at 0.5 mg/ml oxacillin using LightCycler RT-qPCR (qPCR was performed by Justin Rudkin, NUIG) (Figure 4.4b), validating these results. In addition, the purine operon repressor *(purR)* had higher expression in all comparisons with HeR0, including HeR2 (2.7-fold, Table C13).

### 4.5.6 Three global regulators were affected by low dose oxacillin

Three major global regulators were discovered to be involved in the response to low doses of oxacillin: the accessory gene regulator (*agr*) (HeR2 vs HeR0, Table C13), the repressor of toxins (*rot*) and vancomycin resistance regulator (*vraR*) (Figure 4.4). *VraR* has been shown to activate a cell wall stress stimulon (CWSS) in reponse to β-lactam and glycopeptide antibiotics [531–533]. *MecA* had a 6.4-fold increase for HeR2 compared with HeR0 (Table C13) and a 3.11-fold increase in HeR0.5 versus HeR0, although the latter increase was not statistically significant (p-value of 0.24). This trio were identified from a larger set of 38 down-regulated and 31 up-regulated genes in HeR0.5 compared with HeR0, and 25 were down-regulated and 174 were up-regulated in HeR2 compared with HeR0 (Figure C13). A subset of genes with significantly increased (16) or reduced (9) expression at both oxacillin

(HeR0.5 and HeR2) concentrations was identified (Figure 4.4a) and confirmed by qPCR (Justin Rudkin, NUIG) (Figure 4.4b). Eight pathways over-represented among the up-regulated genes (Table 4.5) indicating that they are involved in response to oxacillin.

This agreed with previous work showing that oxacillin exposure causes concomitant repression of the *agr* system and induction of *mecA* [512,513]. Repression of *agr* was consistent with the five-fold increase in *rot* expression [534,535]. Rot controls toxin expression and negatively regulates expression of virulence genes, including *hla* and *hlb* producing alpha and beta hemolysin, respectively [534,535]. Here *hla* had 3.8-fold lower expression for HeR2 compared with HeR0 (Table C13). Rot tends to negatively regulate genes that are positively regulated by *agr* [534,535] and its translation is inhibited by RNAIII, a small RNA produced by the *agr* locus (encoded by SAUSA300_1988) [536,537]. The *dlt* operon (genes *dltA/B/C/D*, [538]) is positively regulated by Rot [535] and had a four-fold increase in expression. *TarS and tagH* had three- and four-fold increased expression (respectively, Table C13): these are involved in the modification and export of wall teichoic acids (WTA) along with the products of the *dlt* operon, [539–541].

### 4.5.7 Little differential expression after conversion of HeR to HoR

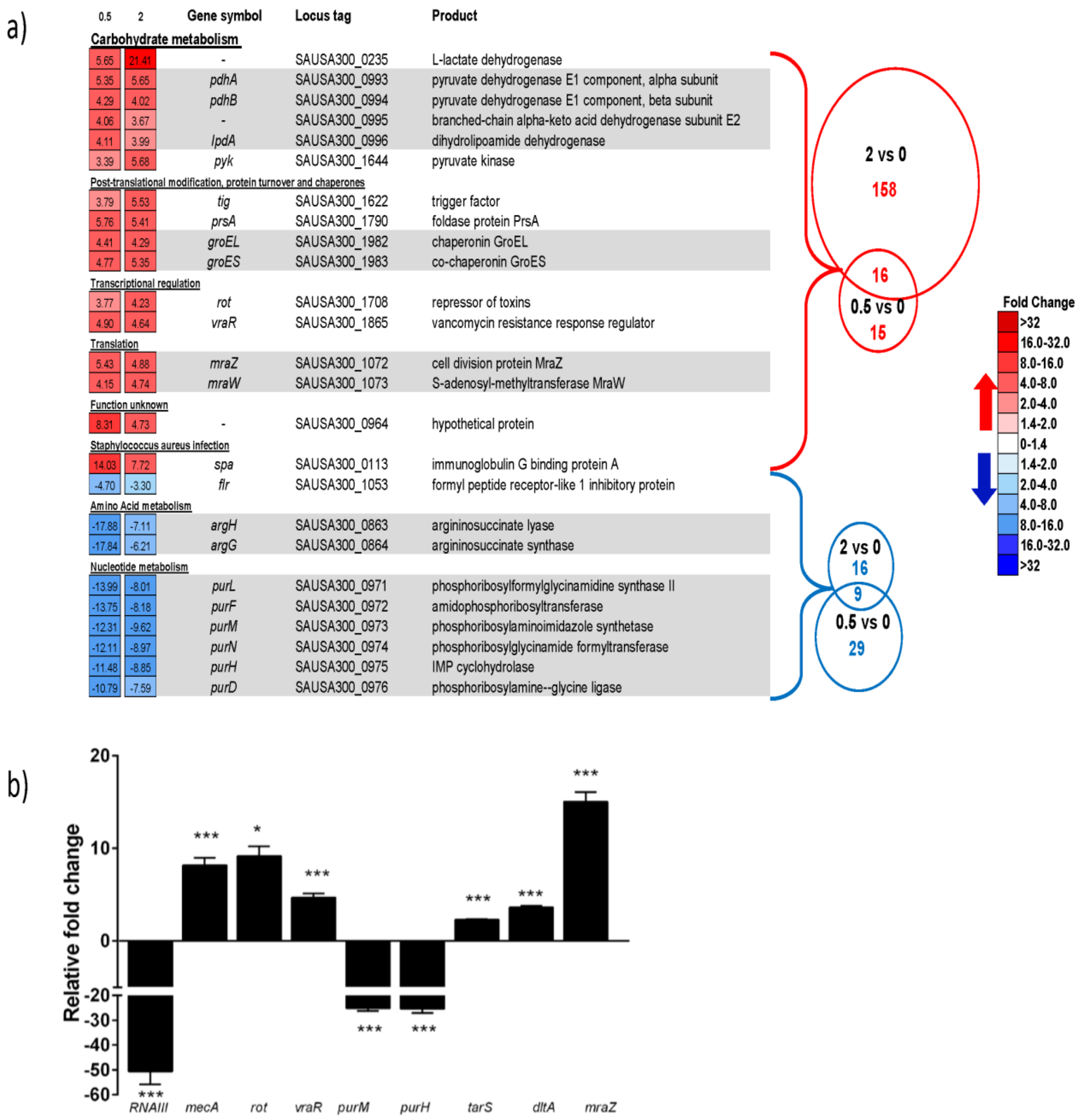
Few genes (8) were differentially expressed in HoR0 compared with HeR0 (Table C13): six were down regulated and the other two encoded up-regulated tRNAs. Four of the down-regulated genes were involved in purine metabolism and had a large (average 12-fold) decreases in expression compared with HeR0. Two of the four (*purN* and *purH*) genes in the purine metabolism pathway were also part of the one carbon pool by folate pathway which is involved in nucleotide and amino acid metabolism, both of which were over-represented in HoR0 (Table 4.5). The other down-regulated gene was *scn* (~4-fold), which encodes a staphylococcal complement inhibitor (SCIN). This gene was also down-regulated in HeR0.5, HoR0.5 and HoR2 compared with HeR0 and had decreased expression (~2.5-fold) in HeR2. SCIN is a virulence factor involved in immune evasion that prevents C3b (a product of human immune system protein complement component C3 cleavage) deposition on the surface of the bacterium and phagocytosis by human neutrophils by inactivating C3 convertases [542].

### 4.5.8 Few genes were differentially expressed in HoR exposed to additional low oxacillin doses

Four genes had changed expression rates in HoR0.5 and eight for HoR2. This indicated that additional oxacillin induced few additional expression changes beyond DNA mutations induced by the 100 μg/ml of oxacillin used to HoR cells (Figure 4.5). No change in purine metabolism gene expression was evident when comparing HoR0.5 and HoR2 cells with HoR0 cells (Tables C14 & C15).

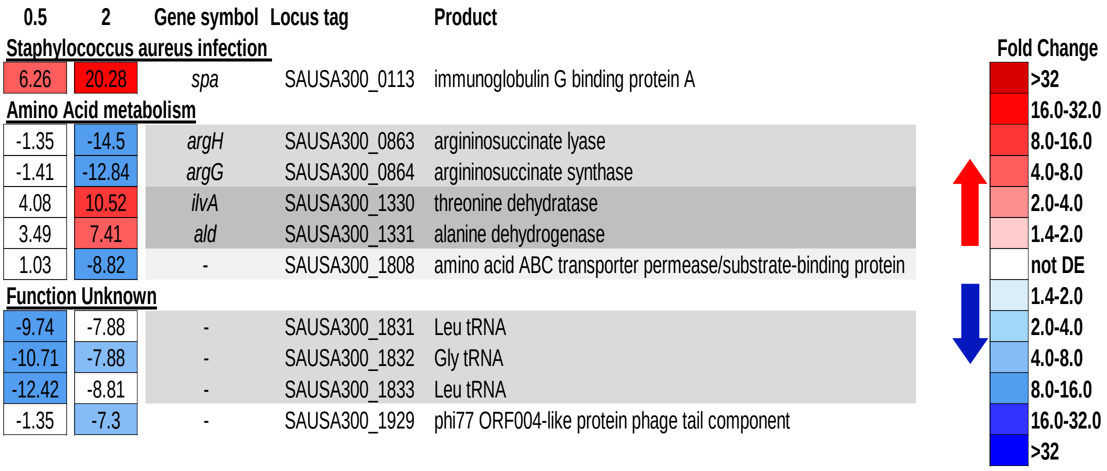
| Pathway ID | Pathway Name | numDEInCat | numInCat | p-value | Adjusted p-value |
|---|---|---|---|---|---|
| **HeR0.5 vs HeR0** | | | | | |
| **Over-represented KEGG pathways for down-regulated genes** | | | | | |
| saa00230 | Purine metabolism | 10 | 59 | 6.36E-08 | 5.98E-06 |
| **Over-represented KEGG pathways for up-regulated genes** | | | | | |
| saa00620 | Pyruvate metabolism | 6 | 37 | 1.62E-06 | 9.94E-05 |
| saa00010 | Glycolysis / Gluconeogenesis | 6 | 39 | 2.12E-06 | 9.94E-05 |
| saa00020 | Citrate cycle (TCA cycle) | 4 | 22 | 9.92E-05 | 0.0031091 |
| **HeR2 vs HeR0** | | | | | |
| **Over-represented KEGG pathways for down-regulated genes** | | | | | |
| saa00230 | Purine metabolism | 6 | 59 | 6.35E-04 | 5.97E-02 |
| **Over-represented KEGG pathways for up-regulated genes** | | | | | |
| saa00010 | Glycolysis / Gluconeogenesis | 18 | 39 | 5.08E-07 | 5.41E-09 |
| saa00620 | Pyruvate metabolism | 13 | 37 | 1.91E-03 | 4.06E-05 |
| saa03010 | Ribosome | 13 | 57 | 1.14E-02 | 3.62E-04 |
| saa00680 | Methane metabolism | 7 | 19 | 4.47E-02 | 1.90E-03 |
| saa00190 | Oxidative phosphorylation | 7 | 23 | 8.05E-02 | 5.12E-03 |
| saa00061 | Fatty acid biosynthesis | 5 | 13 | 8.05E-02 | 5.89E-03 |
| saa02060 | Phosphotransferase system (PTS) | 7 | 23 | 8.05E-02 | 6.19E-03 |
| saa00051 | Fructose and mannose metabolism | 6 | 18 | 8.05E-02 | 6.85E-03 |
| **HoR0vsHeR0** | | | | | |
| **Over-represented KEGG pathways for down-regulated genes** | | | | | |
| saa00230 | Purine metabolism | 5 | 59 | 9.94E-07 | 9.35E-05 |
| saa00670 | One carbon pool by folate | 2 | 9 | 8.91E-04 | 4.19E-02 |
| **HoR0.5vsHeR0** | | | | | |
| **Over-represented KEGG pathways for down-regulated genes** | | | | | |
| saa00230 | Purine metabolism | 6 | 59 | 8.07E-08 | 7.59E-06 |
| saa00670 | One carbon pool by folate | 2 | 9 | 1.39E-03 | 6.55E-02 |
| **HoR2vsHeR0** | | | | | |
| **Over-represented KEGG pathways for down-regulated genes** | | | | | |
| saa00230 | Purine metabolism | 9 | 59 | 3.06E-08 | 2.88E-06 |

**Table 4.5:** Over-represented KEGG pathways for differentially expressed genes in pair-wise comparisons with HeR0. Description of column names: ‘numDEInCat’: Number of differentially expressed genes in this category, ‘numInCat’: Total number of genes in this category, ‘p-value’: p-value for over-represented terms, ‘Adjusted p-value’: Benjamini Hochberg adjusted p-value for over-represented terms.



| 0.5 | 2 | Gene symbol | Locus tag | Product |
|---|---|---|---|---|
| **Carbohydrate metabolism** | | | | |
| 5.65 | 21.41 | - | SAUSA300_0235 | L-lactate dehydrogenase |
| 5.35 | 5.65 | *pdhA* | SAUSA300_0993 | pyruvate dehydrogenase E1 component, alpha subunit |
| 4.29 | 4.02 | *pdhB* | SAUSA300_0994 | pyruvate dehydrogenase E1 component, beta subunit |
| 4.06 | 3.67 | - | SAUSA300_0995 | branched-chain alpha-keto acid dehydrogenase subunit E2 |
| 4.11 | 3.99 | *lpdA* | SAUSA300_0996 | dihydrolipoamide dehydrogenase |
| 3.39 | 5.68 | *pyk* | SAUSA300_1644 | pyruvate kinase |
| Post-translational modification, protein turnover and chaperones | | | | |
| 3.79 | 5.53 | *tig* | SAUSA300_1622 | trigger factor |
| 5.76 | 5.41 | *prsA* | SAUSA300_1790 | foldase protein PrsA |
| 4.41 | 4.29 | *groEL* | SAUSA300_1982 | chaperonin GroEL |
| 4.77 | 5.35 | *groES* | SAUSA300_1983 | co-chaperonin GroES |
| Transcriptional regulation | | | | |
| 3.77 | 4.23 | *rot* | SAUSA300_1708 | repressor of toxins |
| 4.90 | 4.64 | *vraR* | SAUSA300_1865 | vancomycin resistance response regulator |
| Translation | | | | |
| 5.43 | 4.88 | *mraZ* | SAUSA300_1072 | cell division protein MraZ |
| 4.15 | 4.74 | *mraW* | SAUSA300_1073 | S-adenosyl-methyltransferase MraW |
| Function unknown | | | | |
| 8.31 | 4.73 | - | SAUSA300_0964 | hypothetical protein |
| Staphylococcus aureus infection | | | | |
| 14.03 | 7.72 | *spa* | SAUSA300_0113 | immunoglobulin G binding protein A |
| -4.70 | -3.30 | *flr* | SAUSA300_1053 | formyl peptide receptor-like 1 inhibitory protein |
| Amino Acid metabolism | | | | |
| -17.88 | -7.11 | *argH* | SAUSA300_0863 | argininosuccinate lyase |
| -17.84 | -6.21 | *argG* | SAUSA300_0864 | argininosuccinate synthase |
| Nucleotide metabolism | | | | |
| -13.99 | -8.01 | *purL* | SAUSA300_0971 | phosphoribosylformylglycinamidine synthase II |
| -13.75 | -8.18 | *purF* | SAUSA300_0972 | amidophosphoribosyltransferase |
| -12.31 | -9.62 | *purM* | SAUSA300_0973 | phosphoribosylaminoimidazole synthetase |
| -12.11 | -8.97 | *purN* | SAUSA300_0974 | phosphoribosylglycinamide formyltransferase |
| -11.48 | -8.85 | *purH* | SAUSA300_0975 | IMP cyclohydrolase |
| -10.79 | -7.59 | *purD* | SAUSA300_0976 | phosphoribosylamine--glycine ligase |

**Figure 4.4:** USA300 LAC genes with evidence of differential expression (DE) when grown in the presence of both 0.5 µg/ml and 2.0 µg/ml oxacillin. a) The genes are grouped by expression change (red for up, blue for down) and functional category. DE genes had a SVA-adjusted log2 fold-change > 0.5 and corrected p<0.1. The total number of DE genes between comparisons is shown by the circle areas. The numbers of downregulated genes between 0 µg/ml to 0.5 µg/ml (38) and 0 µg/ml to 2 µg/ml (25) were about the same as those upregulated from 0 µg/ml to 0.5 µg/ml (31), whereas those with elevated expression between 0 µg/ml and 2 µg/ml was much higher (174). b) A comparison of relative gene expression by LightCycler RT-qPCR in USA300 LAC grown for 20 hours in BHI or BHI supplemented with 0.5 µg/ml oxacillin. Experiments were repeated at least three times and the standard error of the mean is shown. Student's two-tailed t-test p-value thresholds are indicated: *** for p < 0.0001 and * for p < 0.05. All work in part b) of this figure was carried out by Justin Rudkin, NUIG.



| 0.5 | 2 | Gene symbol | Locus tag | Product |
|---|---|---|---|---|
| **Staphylococcus aureus infection** | | | | |
| 6.26 | 20.28 | *spa* | SAUSA300_0113 | immunoglobulin G binding protein A |
| **Amino Acid metabolism** | | | | |
| -1.35 | -14.5 | *argH* | SAUSA300_0863 | argininosuccinate lyase |
| -1.41 | -12.84 | *argG* | SAUSA300_0864 | argininosuccinate synthase |
| 4.08 | 10.52 | *ilvA* | SAUSA300_1330 | threonine dehydratase |
| 3.49 | 7.41 | *ald* | SAUSA300_1331 | alanine dehydrogenase |
| 1.03 | -8.82 | - | SAUSA300_1808 | amino acid ABC transporter permease/substrate-binding protein |
| **Function Unknown** | | | | |
| -9.74 | -7.88 | - | SAUSA300_1831 | Leu tRNA |
| -10.71 | -7.88 | - | SAUSA300_1832 | Gly tRNA |
| -12.42 | -8.81 | - | SAUSA300_1833 | Leu tRNA |
| -1.35 | -7.3 | - | SAUSA300_1929 | phi77 ORF004-like protein phage tail component |

**Figure 4.5:** Few differentially expressed genes were in HoR0.5 (0.5) and HoR2 (2) comparisons with HoR0. Genes are grouped by locus tag within their functional categories. Red indicates an increase in expression of the gene and blue indicates a decrease in expression, compared with its expression in HoR0 samples. The fold change is shown. White boxes indicate that the gene was not differentially expressed in that comparison.

### 4.5.9 No evidence of an effect of mutation on expression rates

The correlation quantified as a RV coefficent [528] between the genetic variation and the gene expression changes was not greater than background effects associated with experimental noise based on the CIA results (Figure C4): the absence of power to detect eQTLs (Figure C5) was likely due to the small sample size. Permutation tests to test the significance of the RV values showed that no group of RNA and DNA samples had correlations with greater than the 5% significance indicating that the observed correlations (Table C16) were not different from those expected by chance.

### 4.5.10 Genetic background of USA300 in the bioreactor experiment

We screened for SNPs in the 15 samples that survived culturing with a high oxacillin dose (130 µg/ml) in a chemostat and subsequent plating on 800 µg/ml of oxacillin with reference to the genetic background of the unexposed parent isolate. Seventeen SNPs, six deletions and one insertion were in all samples (Table C17). Nine SNPs were intergenic, seven were missense and one was synonymous. Sixteen of these SNPs were also fixed changes in the low dose experiment (Table C10). The other SNP (intergenic T>C change at base 556,291) was present in most samples of the low dose experiment (Table C12), and the four samples in which it was absent had this SNP (1a_S1, 1b_S2, 1c_S3 and 2a_S4) but the base quality values were inadequate for analysis (20, 25, 23 and 3, respectively). The eight deletions and one insertion were also found as part of the genetic background of the samples grown on agar plates (low dose experiment). Eight SNPs polymorphic in all samples were excluded from further analysis along with the SNPs and indels present in the parent isolate (and thus fixed in all samples) (Table C18).

### 4.5.11 Mutations associated with resistance to high oxacillin doses in the bioreactor experiment

64 SNPs were found in at least one of the 15 chemostat samples in comparison to the parent isolate (Table C19). Independently from this, 20 SNPs were found in all 16 chemostat samples, indicating the magnitude of changes associated with genetic drift. Thirteen of the 64 were fixed in at least one sample, excluding four intergenic SNPs (Table 4.6 & 4.7, Table C19) and four deletions were fixed in different samples (Table C21). Six samples had one fixed deletion: four (7A_S7, 10A_S10, 12A_S12, 14A_S14) had the same *dacA* gene missense mutation and the other two (3A_S3, 6A_S6) had the same missense mutation at a gene encoding a tandem lipoprotein (SAUSA300_0410) (Table 4.6). Sample 8A_S8 had an amplification of SCC*mec*IV, and also the highest number of SNPs (nine): six were missense, two were nonsense, and one was synonymous (Table 4.6 & 4.7, Table C19).

### 4.5.12 Multiple inactivating mutations target *gdpP*

*GdpP* had a missense mutation T260P in sample 8A_S8 (*gdpP*-T260P) and a two-base deletion of AT at position 1372-1373 bp causing a frameshift at M458 in sample B1_S16 (*gdpP*-Δ458) (Table 4.6, Table C21 and Figure C6). *GdpP*-T260P was at the GGDEF domain, whereas *gdpP*-Δ458 was at the DHH domain. SNPs at *guaA*, *clpX*, *dacA* and *camS*, *gdpP*-T260P and *gdpP*-Δ458 were confirmed by Sanger capilliary sequencing by collegeues (Laura Gallagher and Nikki Black, NUIG).

### 4.5.13 A large in-frame deletion removed a stop codon at the *apt* gene

There was a 403 bp in-frame deletion in ten samples spanning the 3’ end of *apt* (SAUSA300_1591) that encodes adenine phosphoribosyltransferase (Table C21). This protein catalyses the formation of AMP from adenine and is part of the purine metabolism pathway. This 403 bp deletion (Δ*apt*) would cause the loss of the stop codon, which would likely prevent a functional protein being formed. Six samples (2A_S2, 4A_S4, 5A_S5, 11A_S11, 13A_S13, 15A_S15) had no fixed SNPs at genes but all these samples did have the *apt* gene deletion (Table C21).

| **Gene symbol** | ***gdpP*** | - | ***guaA*** | ***guaA*** | - | - | - | ***dacA*** |
|---|---|---|---|---|---|---|---|---|
| **Locus tag** | SAUSA300_0014 | SAUSA300_0293 | SAUSA300_0389 | SAUSA300_0389 | SAUSA300_0410 | SAUSA300_0417 | SAUSA300_1240 | SAUSA300_2113 |
| **Product** | GGDEF domain protein containing phosphodiesterase activity | Hypothetical protein | GMP synthase | GMP synthase | Tandem lipoprotein | Tandem lipoprotein | UPF0154 protein | Diadenylate cyclase |
| **Mutation** | Missense | Missense | Missense | Missense | Missense | Missense | Missense | Missense |
| **Codon Change** | Aca/Cca | cAt/cTt | gCa/gTa | gaG/gaT | aTa/aAa | gAg/gCg | atG/atC | tCt/tTt |
| **Amino Acid Effect** | T260P | H94L | A314V | E511D | I54K | E52A | M58I | S67F |
| **DNA effect** | 778A>C | 281A>T | 941C>T | 1533G>T | 161T>A | 155A>C | 174G>C | 200C>T |
| **Position on chromosome** | 19,122 | 346,921 | 440,787 | 441,379 | 463,755 | 469,692 | 1,358,918 | 2,288,896 |
| **REF** | A | A | C | G | T | A | G | G |
| **ALT** | C | T | T | T | A | C | C | A |
| **Samples** | 8A_S8 | B1_S16 | 8A_S8 | 8A_S8 | 3A_S3, 6A_S6, 8A_S8, B1_S16 | 8A_S8 | 8A_S8 | 7A_S7, 10A_S10, 12A_S12, 14A_S14 |

**Table 4.6:** Fixed SNPs causing missense mutations in at least one sample in the bioreactor experiment.  SNPs are ordered from left to right based on their position (lowest to highest) on the USA300 chromosome. Samples that these SNPs occurred in are on the bottom row.

| **Gene symbol** | ***clpX*** | ***camS*** |
|---|---|---|
| **Locus tag** | SAUSA300_1621 | SAUSA300_1884 |
| **Product** | ATP-dependent protease ATP-binding subunit ClpX | CamS sex pheromone cAM373 |
| **Effect** | Stop gain | Stop gain |
| **Codon Change** | Gag/Tag | Caa/Taa |
| **Amino Acid Effect** | E37* | Q305* |
| **DNA Effect** | 109G>T | 913C>T |
| **Position on chromosome** | 1,775,825 | 2,046,530 |
| **REF** | C | G |
| **ALT** | A | A |

**Table 4.7:** Fixed SNPs at genes caused stop gain mutations in sample 8A_S8 in the bioreactor experiment.

### 4.5.14 Tandem amplification of SCC*mec*IV

Structural variation was inferred from variation in read coverage relative to the chromosomal median and was present in nearly all samples at the SCC*mec*IV locus (Figure C8, Table C22). The exceptions were parent 1A_S1 and three other bioreactor samples (4A_S4, 11A_S11, 15A_S15) that had no fixed SNPs but did have Δ*apt* (Table C21). The SCC*mec*IV CNV had variable length: 8A_S8 and 12A_S12 had ~25 kb loci amplified (Figure 4.6, Table C23), whereas the samples in the low dose experiment had no 10 kb loci amplified (Figure C7), and in 8A_S8 SCC*mec*IV had approximately eight copies of the entire locus (Figure 4.6, Table C25). Mapping reads to configurations of putative boundaries (SCC*mec*IV start/end and flanking regions) did not have sufficient resolution to determine the potential orientation of SCC*mec*IV and whether or not it was chromosomal due to the small read length.

Assembly of the 8A_S8 chromosome from Nanopore sequence data from a MinION platform (by Mick Watson, University of Edinburgh) revealed ten additional SCC*mec*IV copies as consecutive tandem chromosomal repeats. This was a slightly different copy number from estimates with short reads, indicating an unstable copy number.

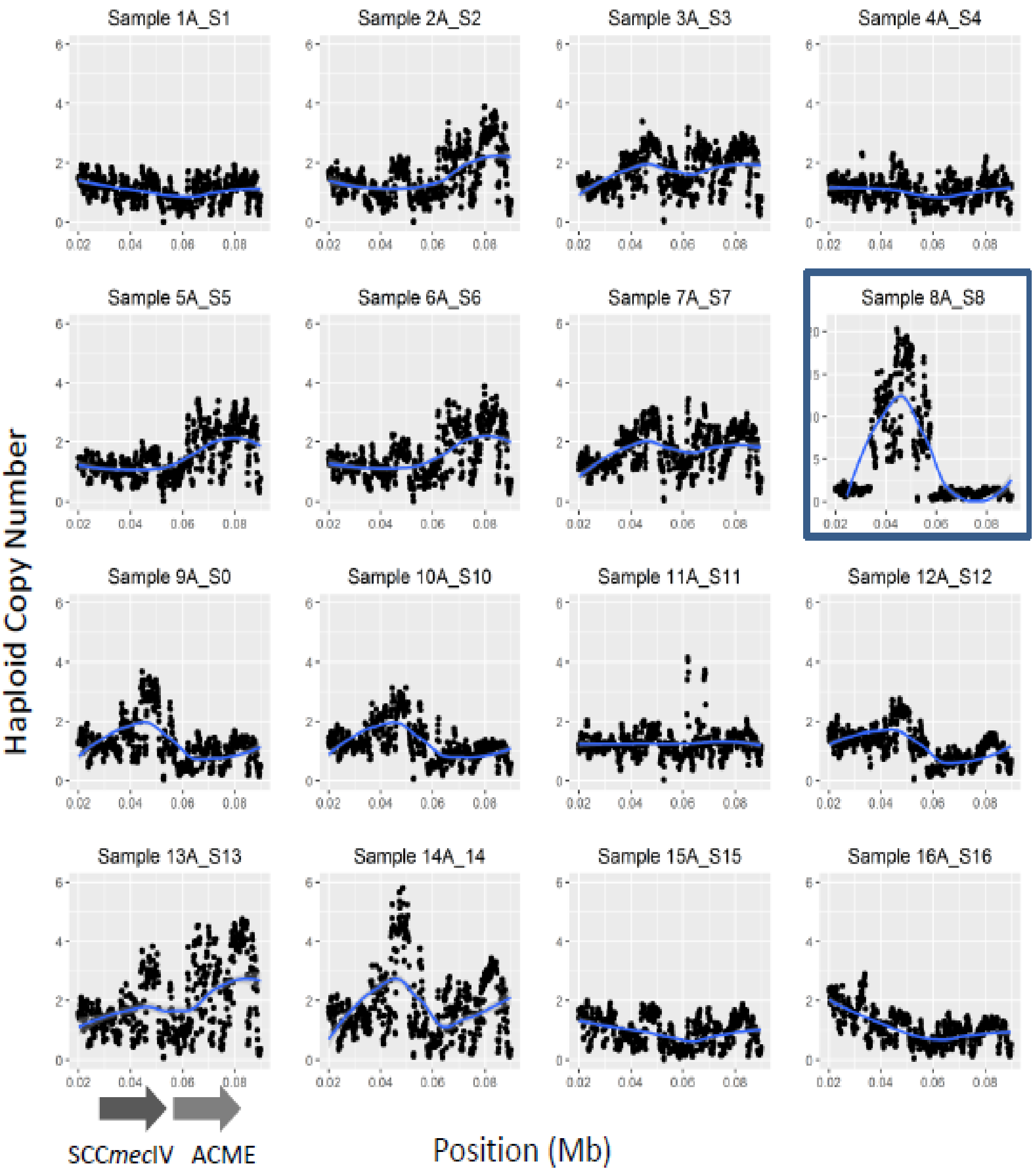


**Figure 4.6:** Copy number measured in 300 bp sliding windows with 100 bp step size across SCC*mec*IV and ACME for all samples in the bioreactor experiment. The position indicates the position on the chromosome e.g. 0.2 Mb is at base 20,000 on the chromosome (SCC*mec*IV coordinates are 0.034 – 0.057 and ACME coordinates are 0.0579 to 0.0889 Mb). The dark grey arrows indicate the positioning of *SCCmec*IV and ACME on the plots here. The blue lines are lowess smoothers applied to the data points (black) with the grey area surrounding each blue line showing the standard error. Sample 8A which has an amplification of the full SCC*mec*IV element only is inside a blue box. The y-axis for this sample also differs from the others

### 4.5.15 Parts of the ACME and SCC*mec*IV elements were amplified in many samples

Four samples (3A_S3, 7A_S7, 13A_S13 and 14A_S14) had both a 10 kb part of SCC*mec*IV and either a 10 or 20 kb section of the ~31 kb ACME element amplified (Table C22). 5a_S5 and 6A_S6 had the same 20 kb ACME locus amplification, and 9A_S9 and 10A_S10 had the same 10 kb SCC*mec*IV section amplified (Table C22). A 10 kb locus near the origin of replication amplified in B1_S16 included a gene encoding hyperosmolarity resistance protein Ebh (SAUSA300_1327) and another encoding virulence factor C (SAUSA300_1322).

#### 4.5.15.1 *S. aureus* COL had a previously undetected 40 kb rRNA operon amplification

The tetracycine-resistant *S. aureus* COL strain (Network on Antimicrobial Resistance in *Staphylococcus aureus* accession no. NRS100) has a MIC > 16 µg/ml for oxacillin and a type I SCC*mec* element, and yet COL from [520] had no SCC*mec* amplification (Table C9). COL had 40 kb duplication of 40 genes bounded by rRNA operons containing 5, 16 and 23S rRNA genes, which are capable of duplicating by homologous recombination [543] and can provide a growth advantage (Figure C8).

## 4.6 Discussion

Two distinct possible drivers of resistance to 0.5 and 2 µg/ml oxacillin were detected by measuring the gene sequence (DNA) and activity (RNA) changes in methicillin-resistant *Staphylococcus aureus* (MRSA) USA300. These correspond to resistance and tolerance. Low doses of oxacillin (0-2 ug/ml) produced substantial expression changes at multiple genes, as well as a single SNP at *rpoB* and a 2 bp substitution at *pbpA* in some samples. These genetic and transcriptional responses to oxacillin both occurred at known factors mediating treatment sensitivity, highlighting that a restricted number of genes can change during oxacillin exposure. Doses of oxacillin sufficient for conversion from heterotypic (HeR) to homotypic (HoR) resistance (100 µg/ml) induced a *gdpP* gene mutation in all HoR isolates but few differences in gene activity, which was restricted to genes involved in purine metabolism and *scn*. Continuous culturing in the chemostat to 130 µg/ml of oxacillin and subculturing in plates at 800 µg/ml revealed a wider range of mutations including SNPs at *gdpP*, *dacA* and *guaA* as well a 403 bp deletion (Δ*apt*) in many samples. The SCC*mec*IV tandem amplification and amplifications of the ACME element highlights the role of homologous recombination in refining gene copy number.

### 4.6.1 Amplification of SCC*mec*IV drives oxacillin resistance

Tandem amplification of the entire SCC*mec*IV locus has not previously been seen in resistant *S. aureus* samples, though an estimated 48 kb chromosomal amplification associated with methicillin-resistance that likely contained *mecA* and had unstable copy number in the absence of methicillin has been discovered [544]. Additionally, a search of 404 *S. aureus* genomes using read coverage of the *mecA* gene normalised with read coverage of three single copy genes and using sample 8A_S8 as a positive control did not find the SCC*mec*IV amplification in any genome suggesting that it has not been overlooked in other samples (work by Bryan Wee University of Edinburgh).

In general, gene duplication and amplification (GDA) is a common mechanism of adaption to antibiotics and drugs that has been seen in unicellular eukaryotes such as *Plasmodium* [545] and *Leishmania* [213], in prokaryotes including β-lactam resistance in *Escherichia coli* and macrolide resistance in *Streptococcus pneumoniae* [546], and multicellular eukaryotes such as human tumors in reponse to drug treatment [547–550]. Another example stems from intermediate vancomycin-resistant *S. aureus* (hVISA) isolates taken from one patient over the course of an infection that had a ~98 kb tandem duplication creating a hybrid *trpB/recA* gene as well as a ~35 kb duplication spanning bacteriophage phiSA3 present integrated into the chromosome and in episomal form. As found here, the copy number, lengths and exact boundaries of the amplified regions varied [304].

The SCC*mec*IV amplification had a variable copy number, though the number of copies of the amplification increased with oxacillin dose and fell to two to three copies when grown in the absence of oxacillin (measured using qPCR of *mecA* by Laura Gallagher, NUIG). The thicker cell wall as a result of increased *mecA* expression of PBP2A appears to slow growth on antibiotic-free media compared to the parent isolate, but once subcultured in oxacillin it grew faster (findings by Laura Gallagher, NUIG). The SCC*mec*IV amplification, *gdpP*-T260P and *clpX*-E37* were first detected in cells at 8 μg/ml oxacillin (the first dose of oxacillin added to the bioreactor), determined from *mecA* expression of other isolates not sequenced here (work by Laura Gallagher, NUIG). However, *gdpP*-T260P alone did not have any effect on resistance in contrast with other strains [507], suggesting that the SCC*mec*IV amplification alone was sufficient for the resistance phenotype (work by Laura Gallagher, NUIG). *GdpP* and *clpX* were not mutated in the parent, suggesting that the mutations in these genes most likely happened at the same 8 μg/ml oxacillin dose as the SCC*mec*IV amplification. These two mutations were not in other samples with partial SCC*mec*IV amplifications, indicating that they are functionally independent.

The mechanism by which parts of both ACME and SCC*mec*IV were amplified is currently unknown. SCC*mec*IV harbors *ccr* recombinase genes that can excise SCC*mec* and (probably) ACME [281,504]. Promoters of *ccr* genes are overexpressed in some β-lactam- and vancomycin-treated strains [551], potentially leading to increased excision of SCC*mec*IV. However, cells revert to oxacillin-susceptibilty if SCC*mec*IV is fully excised, so generally only a small fraction of cells exise the SCC*mec*IV element for transfer to MSSA cells [552]. An elevated rate of SCC*mec*IV amplification driven by increased *ccr* expression could yield copies that replicate extrachrmosomally and subsequently reintegrated into the chromosome, increased the SCC*mec*IV copy number further. The presence of partial amplifications of SCC*mec*IV and ACME elements from the bioreactor (chemostat subpopulations at 130 μg/ml oxacillin) suggests that sometimes only parts of these loci were excised or amplified. Previous work showing partial excision of a SCC*mec* region spanning *mecA* [553] parallels our observation of partial amplifications in the seven samples (Table C22 and Figure 4.6), though we found a normal copy number of *mecA*. This suggests that these varied sections at 40-50 kb to 50-57 kb could have been amplified originally on a full SCC*mec* with subsequent excision of parts of the amplified loci, and that the copy number subsequently may have changed at numerous unobserved points across this experiment as a result of this copy number instability. Consequently, the benefit of amplifying different SCC*mec*IV region remains confounded by the extensive variation and their occasional co-incidence with other chromosomal mutations elsewhere. Chromosomal amplifications have a fitness cost if the drug is removed and so tend to be unstable [554], whereas SNPs or horizontally transferred genes may be more stable due to no loss of fitness [547].

Amplifications can arise through RecA-dependent non-equal homologous recombination between 20-40 base length directly-orientated repeat sequences, or through RecA-independent mechanisms such as recombination between single-stranded repetitive sequence on sister chromatids at the replication fork [555]. A RecA-independent mechanism such as rolling circle replication is implicated here due to the absence of repeat sequences flanking the SCC*mec*IV amplification here. This mechanism starts with an initial double-strand break (DSB) followed by RecA-dependent DSB repair followed by rolling circle replication to generate the amplification, and can produce long tandem arrays in a single generation, which could facilitate fast adaption to drug treatment [547,556].

### 4.6.2 *GdpP* is commonly mutated in HoR cells

Independent mutations *gdpP*-R540*, *gdpP*-T260P and *gdpP*-Δ458 in HoR cells demonstrated that *gdpP* is a key target for oxacillin resistance. The *gdpP* gene encodes a c-di-AMP phosphodiesterase enzyme with three domains: GGDEF, DHH and DHHA1. GGDEF domains are homologs of adenylate cyclase domains (also known as diguanylate cyclase domains) that synthesise c-di-GMP from GTP [557], though the GGDEF domain in *GdpP* may not have this activity [558]. The DHH/DHAA1 domain hydrolyses c-di-AMP and c-di-GMP producing linear dinucleotides 5'-pApA and 5'-pGpG [557].

*GdpP*-R540* was repeated nine times and R540* was also in the HoR RNA reads. A USA300 HoR derivative assessed by [513] also reported the same *gdpP*-R540* C>T SNP. The *gdpP*-Δ458 frameshift was in the DHH domain, whereas *gdpP*-R540* was not in any domain (Figure C6). The HoR phenotype is associated with: *gdpP*-R602H in the DHHA1 domain of HoR 8325-4 laboratory strains; *gdpP*-G308D in seven independently isolated 8325-4 HoR strains; and deletions Δ382–504 and Δ80–174 in two other HoR 8325-4 strains [513] (Figure C6). Other substitutions in the gene were identified in HoR derivatives of clinical isolates 15981 (P392S, D105N), MSSA476 (V52I, D105N, P392S) and DAR26 (D105N, P392S) [513]. Two other HoR derivatives of clinical isolates (DAR176 and DAR9) lacked *gdpP* mutations, suggesting that the HoR phenotypes arose through a separate mechanism or through unidentified changes at c-di-AMP target genes [513].

*GdpP* homolog *yybT* is inhibited by the stringent response alarmone ppGpp in *Bacillius subtilis* [557] and increased levels of this alarmone have been seen in the HoR phenotype [505], so mutations inhibiting GdpP could assist the stringent response. High levels of c-di-AMP (cyclic diadenosine monophosphate) as a result of *gdpP* mutations are implicated in tolerance/resistance to β-lactam antibiotics, acid and heat stress [559] as well as influencing cell wall structure and biofilm formation [513]. *GdpP*-R540* in the GGDEF domain may affect the phosphodiesterase activity of this domain that degrades c-di-AMP. However, the stringent stress response results in large changes in gene expression as cells divert resources away from growth towards survival [560], which were not seen in HoR cells here. This suggests that there may be genetic differences in the cells used for RNA sequencing or that mutated *gdpP* and down-regulated purine metabolism genes were acting through an unknown resistance mechanism.

### 4.6.3 *RpoB and pbpA* were mutated in a subpopulation of cells exposed to low dose oxacillin

There are four native pencilin-binding proteins in *S. aureus* (PBP1-4) as well as PBP2A encoded by *mecA*, which catalyses peptidoglycan synthesis in the cell wall [561]. PBP1 is produced by *pbpA* and forms a monofunctional transpeptidase essential for growth and survival of MSSA and MRSA, and is hypothesised to be involved in cell divison [280,561,562]. *PbpA*-W351K here is in the transpeptidase domain of PBP1, which is near the active site (site at 337 bases in the domain) acylated by β-lactam antibiotics.

*RpoB*-A477V and *rpoB*-Y1056N have been implicated in the conversion from HeR to HoR (Figure C6) [507]. In general, most mutations at *rpoB,* particularly H481Y and D471Y, confer resistance to rifampicin by reducing the binding affinity for RNA polymerase [563]. In addition, *rpoB*-H481Y reduces susceptibility to vancomycin, as have other *rpoB* mutations in the rifampicin resistance determining region (RRDR) [304,564]. The RRDR encompasses amino acids ~463 to 550 and mutations within it reduce the hydrohphobic interaction between *rpoB* and rifampicin, which lowers the binding affinity of rifampicin for RNA polymerase [565]. Moreover, *rpoB*-T518I is at the domain fork, proximal to the catalytic site of RNA polymerase. Thus, *rpoB*-T518I in ~20% of cells in all three HeR2 samples and fixed in one HeR0.5 sample is capable of causing rifampicin resistance. *RpoB*-

A621E has also resulted in dual heteroresistance to both daptomycin and vancomycin in *S. aureus in vitro*, and was accompanied by a reduction in negative charge of the cell surface, a thickened cell wall, repression of metabolic pathways (purine and pyrimidine metabolism, the urea cycle and the lac operon), and higher biosynthesis of vitamin B2, K1, and K2 and cell wall metabolism [566].

PpGpp is the product of *relA* and targets RNA polymerase, causing it to modify the expression of various stress-related genes [567]. Structural evidence suggests that ppGpp binds at a single place close to the active centre of RNA polymerase via interaction with the β and β1 subunits [567]. RNA polymerase is the product of *rpoB* and *rpoC*, so mutation of *rpoB* could mimic changes in gene expression typically induced by the ppGpp binding, as demonstrated in *Streptomyces coelicolor* [568] and *E. coli* [569].

*RpoB* mutations in the absence of drug selection also are induced by environmental stress. *RpoB* mutations that conferred resistance to rifampicin were found in response to thermal stress in the absence of antibiotic [570]. Mutation of *rpoB* also enabled cells to cope with variety of stressors including osmotic, acid stress and n-butanol [571]. Another study that examined the mutations conferring a selective growth advantage in *E. coli* to glycerol-based minimal media found that *rpoB* and *rpoC* mutations conferred the greatest increase in the growth rate of all 13 *de novo* mutations, and facilitated extended growth in this media [572]. However, *rpoB* is much more commonly mutated in laboratory conditions than it is in nature [573], indicating that the mutation may be associated adaption to growth in laboratory conditions rather than resistance.

### 4.6.4 Purine metabolism is controlled transciptionally at low oxacillin doses and genetically at high oxacillin doses

Further support for the hypothesis that there are two distinct responses to oxacillin leveraging transcriptional control at low oxacillin levels (tolerance) and DNA mutation at high oxacillin doses (resistance) were found in the control of purine metabolism. Daptomycin- and vancomycin-resistant MRSA with the *rpoB*-A621E previously implicated purine metabolism [566], as does a *purL* mutation that attenuated virulence in *S. aureus* strain RN6390 in mice [574]. Five purine metabolism genes (*purB*, *purF*, *purH*, *purM*, SAUSA300_0147) are involved in *S. aureus* tolerance of rifampicin, but this also leads to attenuated virulence and a reduction in biofilm formation [575,576]. Missense SNPs *guaA*-A314V and *guaA*-E511D were in one bioreactor sample: *guaA* is involved in the (p)ppGpp-mediated stringent stress response. *GuaA*-P142L and a single-base *guaA* deletion (ΔS27-S28) have been observed in HoR isolates (oxacillin MIC ≥ 400 µg/ml) [507]. *GuaA* encodes a GMP (guanosine mono phosphate) synthase that synthesises GMP from IMP (inosine monophosphate), and is essential for growth and survival of *S. aureus* [577]. This gene is part of the GMP biosynthesis pathway, a component of the purine metabolism pathway.

Purine metabolism regulates the stress response in many gram-positive and gram-negative bacteria: most *E. coli* single-gene deletion mutants had a growth defect in human serum and had a deletion at a gene involved in either purine (genes: *purF*, *purD*, *purL*, *purM*, purK, *purE*, *purC*, *purH*, *guaA*, *guaB*) or pyrimidine biosynthesis. The *B. anthracis purE* and *purK* deletion mutants also exhibited severe growth defects in human and murine serum. *B. anthracis purE* knockout mutants were also avirulent and increased survival of mice compared to those challenged with a wild-type strain - but this is not effective in every mutant, such as *purK* [578]. *Salmonella enterica* serovar *Typhimurium* showed attenuated virulence when nucleotide biosynthesis genes were inactivated [579,580]. Fourteen *Salmonella Typhimurium* mutants with inactivated *pur* or *pyr* genes showed growth defects in human serum, demonstrating that *de novo* nucleotide biosynthesis is essential for growth in human serum and attenuated virulence in mutants may be partly due to an inabilty to grow in host blood [578]. Thus, the down-regulation of purine metabolism at low oxacillin doses suggests that these cells are likely to be less virulent.

The *apt* gene involved in purine metabolism and is upstream of *relA*: *apt* had a 403-base deletion here. In *Streptomyces coelicolor*, *relA* is transcribed from transcriptional read-through of *apt*, thus linking the stringent response and purine metabolism [581]. *Apt* gene expression is increased by the stringent response [582], and mutations at this gene have also found in vancomycin-intermediate (VISA) *S. aureus* [583].

The absence of (type I) SCC*mec* changes and the presence of a large rRNA operon duplication in tetracycine-resistant *S. aureus* COL illustrates firstly that growth rates are regulated by other mechanisms separate from purine metabolism [584], and secondly that homologous recombination can promote oxacillin resistance at other loci beyond SCC*mec* [543]. An increase in rRNA genes copies was likely to elevate the reproductive rate of COL, though it is unclear if this can be applied to environments with reduced resources (such as *in vivo* or *in vitro* lag growth stage), and thus may facilitate tolerance rather than being a conserved genetic mutation.

# Chapter 5 – Conclusions

Leishmaniasis and MRSA are significant public health problems whose control can be assisted using pathogen genomics approaches. In this chapter, I review our key findings, novel contributions to *Leishmania* and MRSA research and discuss the limitations of our work, as well as future work stemming from these findings.

## 5.1 Chromosome fission in *L. adleri*

Our first research topic addressed the identification and assembly of a *Leishmania* sample from a rodent (HO174), taken in Ethiopia in 1975. We identified the sample as *L.* (*Sauroleishmania*) *adleri* (HO174) using a phylogenetic scheme, making this assembled genome only the second high-quality reference available for one of the nineteen species in this subgenus. Subsequent mapping of the reads from this sample, as well as another *L. adleri* sample (SKINK-7), to this reference genome demonstrated that a novel fission of chromosomes 30 and 36 had occurred in *L. adleri* and that this coincided with differences in somy between the separated chromosome pairs in one sample (HO174). These fission events occurred at strand switch regions (SSRs), preserving polycistronic transcriptional units (PTUs) and putative origins of replication on the fission chromosomes (chromosomes 30.1, 30.2, 36.1, 36.2). Aneuploidy and extensive gene amplifications were also documented in the *L. adleri* samples and the only other assembled species in this subgenus - *L. tarentolae* ParrotTarII [190]. This chapter demonstrated that there are no major differences in genome structure, gene complement and sequence divergence between *Sauroleishmania* isolated from humans and animals.

The compelling evidence of chromosome fission based on read coverage in *L. adleri* could be complemented by additional laboratory work such as FISH to validate both the fission and somy changes documented in *L. adleri*. Previous work by [177] has shown that aneuploidy predicted by read coverage and heterozygous SNP allele frequencies correlates well with somy determined from flow cytometry of DNA content and the number of gene-specific deletions needed to generate a null mutant [177]. However, given that only two *L. adleri* samples were available, more *Sauroleishmania* isolates are required to determine if these fission events are common to all *L. adleri* samples and if they are present in any other *Sauroleishmania* spp. As spontaneous chromosomal fission has been documented in *L. tarentolae* previously [400], this may be a more widespread phenomenon in

*Sauroleishmania* than previously realised. We did not find any enriched gene functions on chromosomes 30 or 36 that could explain their fission and somy changes and so chromosome fission and fusion could be another feature of genomic instability in *Leishmania*. Experiments on *L. adleri* HO174 strain to induce chromosome fission at homologous sites at *L. tarentolae* chromosomes 30 and 36 using a technique such as that outlined by [585] could determine if there are any changes in growth, somy, virulence or function as a result of the fission. A putative chromosomal fission has also been observed in *T. brucei* CL Brener strain [402] and given that fusion events have occurred between *Leishmania* chromosomes causing a reduction in total chromosome number in the *L. mexicana* (Chr 34) complex and *L.* (*Viannia*) subgenus, fission may be another method of enhancing transcriptional regulation in some *Leishmania* spp. The origins of replication have been mapped in *L. major* using MFAseq [362] although DNA combing analysis suggests that this method is biased towards high-efficiency origins [398]. Nevertheless, three of the four *L. adleri* chromosomes resulting from the fission of chromosomes 30 and 36 retain sites homologous to *L. major* early-firing origins detected using MFASeq, indicating that they could replicate independently. The other chromosome (Chr 36.1) may have one or more lower or more variable efficiency firing origins that have yet to be discovered.

It has also been suggested that the single early-firing origins identified by MFASeq in *L. major* could represent centromeric regions, as these are also replicated in early S-phase in other eukaryotes [362,397]. The fission of chromosome 30 in *L. adleri* results in the splitting of the origin, so that there is one origin at the end of chromosome 30.1 and one origin at the start of chromosome 30.2 and so if a centromere is present at this site it would be split into two parts, with one functional part on each fission product (centric fission) [586]. In order for the four chromosomes resulting from the fission of chromosomes 30 and 36 to be maintained, the broken chromosomes ends need to be protected from degradation. This can occur through the addition of telomeres, structural rearrangements to protect the chromosomes ends, development of ring chromosomes through fusion of the broken end onto the telomere of the intact end, or translocation of the chromosomes produced by fissions onto the end of other intact chromosomes [586]. We did not find any evidence of structural rearrangements near the ends of the chromosomes produced by fission, although long reads may be needed to determine this, especially if a complex rearrangement occurred. Additionally, telomeric sequences are highly repetitive and so were not assembled fully here, limiting our investigation. Long reads and higher coverage would also be required to effectively uncover a fusion of the fission chromosomes onto other chromosomes.

## 5.2 *L. naiffi* can infect dogs and is present in Colombia

My second research topic focused on the identification and assembly of two *L. (Viannia)* spp. that were originally isolated from dogs with cutaneous canine leishmaniasis in Colombia in 1985/86. I identified the samples as *L. naiffi* and *L. guyanensis* using a similar scheme to that used for *L. adleri* and also assembled these samples in the same manner. The genome of a previously published *L. braziliensis* sample (M2904) [177] was also assembled using the same approach and served as a positive control to optimise our protocols. These genomes were compared with others in the *Viannia* subgenus (*L. panamensis* and *L. peruviana*) and read-mapping was performed for sets of unassembled reads (*L. shawi*, *L. naiffi*, *L. lainsoni*, *L. guyanensis* and *L. panamensis*). The identification of one sample as *L. naiffi* (CL223) is novel as it represents, to the best of our knowledge, the first finding of *L. naiffi* in Colombia and in dogs. This has implications for disease surveillance programs in Colombia as it demonstrates that *L. naiffi* has been circulating in dogs for at least three decades and is likely to still be circulating. It is possible that it is not reported to health professionals or veterinarians because *L. naiffi* infection is relatively benign. In addition, although *L. naiffi* DNA may have been detected previously, it may have been typed as a different species. Indeed, this was the case for *L. adleri* that we assembled here which was originally typed as an unusual *Leishmania* lineage using multi-locus microsatellite typing (MLMT). Although techniques such as PCR and multilocus enzyme electrophoresis (MLEE) [587] are used in *Leishmania* disease surveillance, identification of infected animals is often performed with anti-*Leishmania* antibodies against particular species [588,589], therefore missing infections caused by other species. In addition, even if immunological tests can only detect *Leishmania* antibodies and not an exact species, the infection is often assumed to have been caused by the endemic *Leishmania* species. In places such as Brazil where dogs are a reservoir for *L. infantum,* dogs are screened as part of control programs and those found to be seropositive are culled. However, this does not always adequately control human infection and this has been hypothesised to be caused in part by the sacrifice of dogs that were incorrectly assumed to be infected with *L. infantum* on the basis of such tests [590].

## 5.3 A minichromosome and 45 kb amplification are present in some *Viannia* spp.

By assessing read mapping, alignment and coverage distributions, we discovered a mini-chromosome in *L. shawi* M8408 for the first time. This minichromosome was previously reported in *L. panamensis* PSC-1 [188] and *L. braziliensis* [489]. We also found a variable copy number amplification of 45 kb present in many *Viannia* genomes. It will be important to determine if the minichromosome in *L. shawi* M8408 is a continuous presence in culture or if it only arose spontaneously in the culture that was used for sequencing, because this strain of *L. shawi* is widely used in laboratories studying *Leishmania*.

## 5.4 Limitations of short-read assembly and future work

The creation of three draft genomes from short-read data demonstrates the usefulness of this type of data for species identification and evolutionary analysis. The removal of contaminating sequence was also important to prevent incorrect inference of particular functions or lateral gene transfer [591]. However, gaps remain in all three genomes which could be resolved with either targeted resequencing strategies, or longer reads to bridge gaps and repetitive regions. The *L. braziliensis* control genome recovered most of the sequence content (~93%) and protein-coding gene content (97%) of the published reference genome indicating that short-read data can provide comprehensive draft assemblies; the use of a control genome was also a useful aid for detecting poorly assembled sections in the absence of longer reads. Although we manually improved the annotation and visually checked the genomes through comparison with other genomes and read-mapping, errors undoubtedly remain. Transcriptome data would be useful to refine gene models: for example, to define untranslated regions (UTRs), find incorrect gene models and also smaller gene models overlooked in our annotation. Furthermore, we showed that the annotation of non-coding RNAs was incomplete and is partly due to the annotation pipeline used, which was not designed to annotate non-coding RNAs as accurately as protein-coding regions.

## 5.5 Gene expression changes are associated with HeR whereas mutation is required for HoR

My last research question addressed the mechanisms involved in resistance to different doses of oxacillin in MRSA across different conditions. To examine this we looked at changes gene expression as well as DNA sequence at 0, 0.5 and 2 µg/ml of oxacillin in HeR and HoR cell populations grown on agar plates. We discovered that HeR is promoted by transcriptional regulation at low doses (0.5 and 2 µg/ml), whereas genetic mutation enables resistance to high doses (exceeding 100 µg/ml). The differential expression of genes involved in the purine metabolism pathway as well as global regulators of gene expression in *S. aureus* are important for drug responses: many genes were already implicated in previous work, though often with different mutations. This highlighted how tolerance and resistance appear to take effect through multiple pathways and mechanisms and yet converge on the same genes such as *gdpP*.

## 5.6 Limitations of our approach to examine differential gene expression

All HoR samples had a SNP at the *gdpP* gene but few gene expression changes compared to HeR, which was surprising given that *gdpP* mutations previously associated with the stringent response drastically altered gene expression levels [592]. This may have been caused by differences in cells sent for RNA sequencing because we did not find any relationship between gene expression and SNPs using either coinertia analysis or eQTL detection. We lacked RNAseq biological replicates for one group (HoR2), which reduced our power to find differentially expressed genes for this group. Although we had two or three biological replicates for each group, more replicates in each group would have improved our ability to detect differentially expressed genes as well as facilitating identification and removal of outlier samples. In fact, at least twelve replicates per condition have been recommended to detect most differentially expressed genes and help to control the false positive rate [593]. In addition, issues with either the sequencing preparation or run resulted in the incorporation of ‘A’ bases without a base quality drop for close to half of each read in some cases, which had to be removed prior to mapping reads. Screening for hidden batch effects identified one latent variable for which we corrected: additional experimental information could have been incorporated into the differential expression testing to adjust for these effects in DESeq2 more precisely. Finally, DESeq2 has a higher false positive rate to elevate sensitivity [114] compared with tools such as edgeR [593,594], suggesting that consensus detection of expression changes using multiple approaches may be an option for consideration in future work.

## 5.7 Amplification of SCC*mec*IV by oxacillin

In order to understand resistance to high doses of oxacillin after a step-wise increase in dose, we examined DNA samples that had survived exposure to high doses (130 μg/ml) of oxacillin in a chemostat. Our work is novel because we discovered that *S. aureus* can amplify the SCC*mec*IV element intrachromosomally in an unstable manner in response to varying doses of oxacillin, which has important implications for patient treatment. We also discovered that partial amplifications of both ACME and SCC*mec*IV appeared in a subpopulation of cells but the effect of these CNVs on resistance remains to be examined. Most *S. aureus* studies examine SNP and indel variation but few look at copy number variation. Our approach can easily be co-opted for other applications to resolve the exact orientation and copy number of amplified sections, assisted ideally by long read data and RT-qPCR verification.

## 5.8 Limitations and future work on MRSA

This research identified three major tasks for future work: firstly, to measure tolerance and resistance at oxacillin doses between 2-100 µg/ml, and extend this to higher read coverage measurements that would allow persistence to be accurately examined across multiple replicates of individual colonies [595]. Persistence by MRSA was not explored here but is pertinent to growth in environments depleted of resources [596] and persistent subpopulations could be identified using the minimum duration of killing assay (MDK) [301]. Secondly, re-assembly of *SCCmec*IV chromosomal and episomal copies using an RNAseq-style approach to resolve the *SCCmec*IV isoforms and their relative abundances could be facilitated by the development of reference graphs for *S. aureus* [597]. Thirdly, the reduced virulence driven by the stress response associated with purine metabolism found here points to therapeutic avenues for combination treatments to ameliorate symptom severity with just low antibiotic doses. For instance, further work could examine the rate of phenol soluble modulin (small cytolytic toxin) expression between the *agr*-active and -deficient samples. The release of membrane phospholipids from *agr*-defecient samples inactivates daptomycin but this effect can be by prevented by using oxacillin in combination with daptomycin [598].

## 5.9 Final thoughts

In summary, we have contributed to the field by providing evidence of chromosome fission in *L. adleri* as well as aneuploidy in the *Sauroleishmania* subgenus. We have found that *L. naiffi* is present in Colombia and that it can infect dogs as well as humans, that a minichromosome is present in *L. shawi* M8408 and that a copy number alteration of a 45 kb region is common in many *Viannia* genomes. This demonstrates the use of short-read assembly and mapping for the characterisation and identification of *Leishmania* spp. and adds to our understanding of genome plasticity in the *Viannia* and *Sauroleishmania* subgenera. We have also underscored the importance of copy number variation in the development of resistance to oxacillin in MRSA by finding an amplification of the mobile *SCCmec*IV element which is associated with resistance. Coupling this result with the finding of CNVs associated with resistance in clinical *S. aureus* isolates by others [304] will encourage the use of our short-read CNV detection approach. Our findings have also shown that mutations in *gdpP* are common, that the purine metabolism pathway is commonly targeted by both gene expression and indel changes in response to oxacillin and that HeR is associated with gene expression changes whereas HoR is mutational in nature.

# Appendix A -The genome of *Leishmania adleri* from a mammalian host highlights chromosome fission in *Sauroleishmania*

*Some tables in this section were too large to include and so are deposited on Figshare at https://figshare.com/s/0f4aca2891b891c0b3fc. The legends of each of these tables are included in this appendix as well as an indication that the particular table can be found online.*

| Assembly Stage | #Scaffolds | N50 | Corrected N50 | % Error Free Bases |
|---|---|---|---|---|
| Velvet & contigs > 100 bp | 18,480 | 4,701 | 4,701 | 66.06 |
| SSPACE | 5,259 | 54,178 | 32,259 | 87.26 |
| Gapfiller | 5,259 | 54,077 | 38,613 | 89.09 |
| iCORN | 5,259 | 54,085 | 38,817 | 89.13 |
| REAPR | 5,785 | 38,817 | 38,817 | 89.13 |

**Table A1:** Statistics calculated by REAPR for each stage of the assembly of HO174

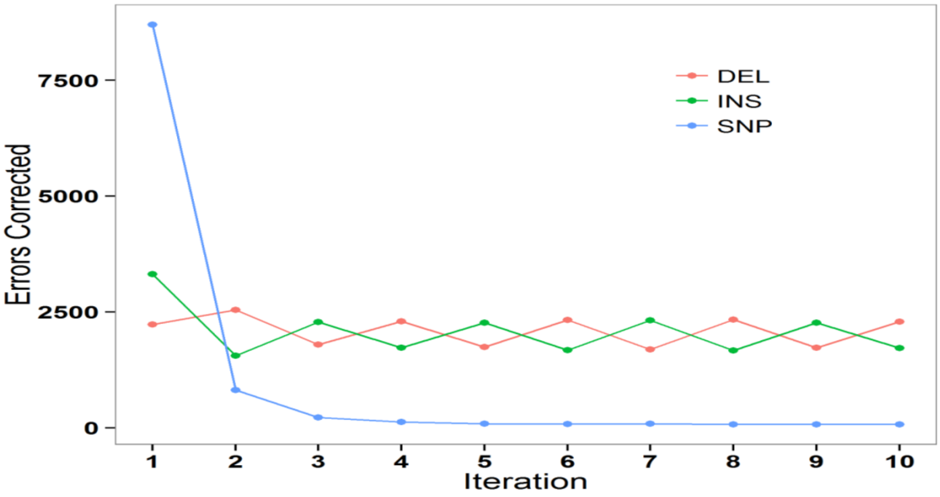


**Figure A1:** Number of iCORN corrections over 10 iterations.

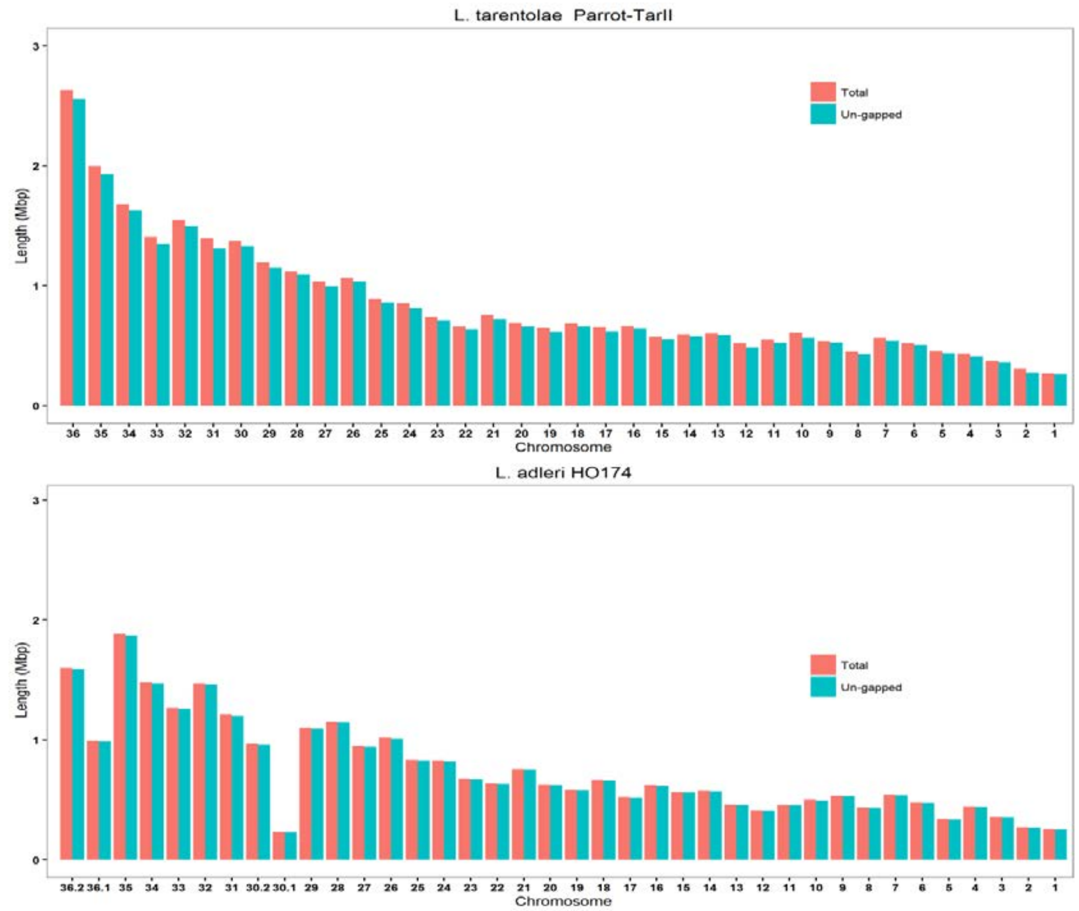


**Figure A2:** Chromosome lengths of *L. tarentolae* and *L. adleri* HO174 showing lengths with gaps included (red) and lengths with gaps excluded (green).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***L. adleri* SKINK-7** | 0 | | | | | | | |
| ***L. tarentolae* Parrot-TarII** | 49 | 0 | | | | | | |
| **HO174** | 2 | 49 | 0 | | | | | |
| ***L. major* 5-ASKH** | 203 | 193 | 203 | 0 | | | | |
| ***L. tarentolae* TarVI** | 49 | 4 | 49 | 190 | 0 | | | |
| ***L. hoogstraali* NG-26** | 51 | 33 | 51 | 187 | 28 | 0 | | |
| ***L. gymnodactyli* Ag** | 55 | 26 | 55 | 196 | 22 | 32 | 0 | |
| ***L. adleri* 1433** | 21 | 35 | 21 | 191 | 30 | 30 | 34 | 0 |
| | ***L. adleri* SKINK-7** | ***L. tarentolae* Parrot-TarII** | **HO174** | ***L. major* 5-ASKH** | ***L. tarentolae* TarVI** | ***L. hoogstraali* NG-26** | ***L. gymnodactyli* Ag** | ***L. adleri* 1433** |

**Table A2:** Number of substitutions between concatenated genes for each isolate used in Figure 2.2(b)

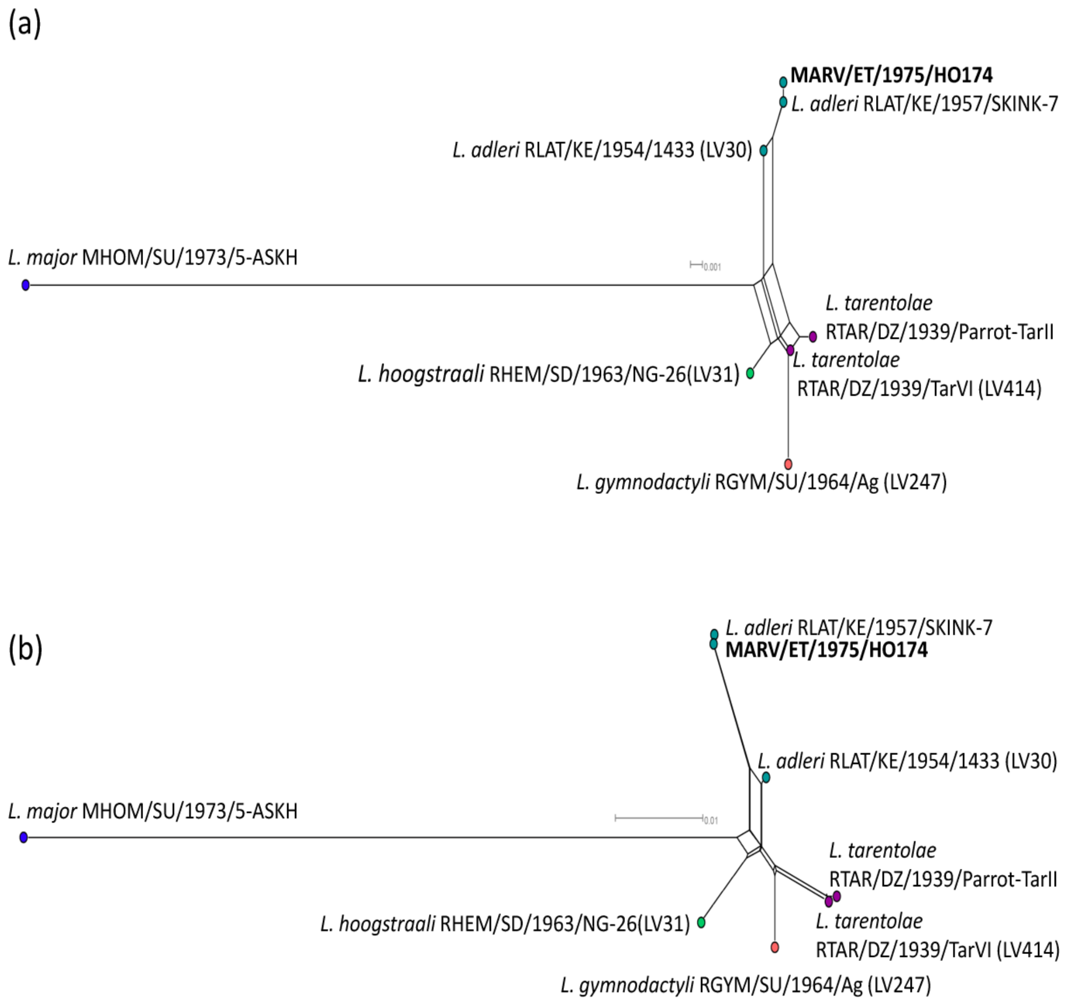

(a)
MARV/ET/1975/HO174
L. adleri RLAT/KE/1957/SKINK-7
L. adleri RLAT/KE/1954/1433 (LV30)
L. major MHOM/SU/1973/5-ASKH
0.001
L. tarentolae
RTAR/DZ/1939/Parrot-TarII
L. hoogstraali RHEM/SD/1963/NG-26(LV31)
L. tarentolae
RTAR/DZ/1939/TarVI (LV414)
L. gymnodactyli RGYM/SU/1964/Ag (LV247)
(b)
L. adleri RLAT/KE/1957/SKINK-7
MARV/ET/1975/HO174
L. adleri RLAT/KE/1954/1433 (LV30)
L. major MHOM/SU/1973/5-ASKH
0.01
L. tarentolae
RTAR/DZ/1939/Parrot-TarII
L. hoogstraali RHEM/SD/1963/NG-26(LV31)
L. tarentolae
RTAR/DZ/1939/TarVI (LV414)
L. gymnodactyli RGYM/SU/1964/Ag (LV247)

**Figure A3:** (a)NeighborNet network based on uncorrected p-distances of the DNA polymerase $\alpha$ catalytic polypeptide (924 sites in alignment) gene sequences from [334] with *L. major* as an outgroup. The scale bar indicates the number of substitutions per site. There is 1 substitution between HO174 and *L. adleri* SKINK-7, 4 substitutions between *L. adleri* 1433 and HO174, 15 between HO174 and each of the two *L. tarentolae* species, 17 between *L. hoogstraali* NG-26 and HO174, 21 between *L. gymnodactyli* Ag, 11 between *L. adleri* 1433 (LV30) and *L. tarentolae* TarVI and 73 between *L. major* 5-ASKH and HO174. (b) NeighborNet network based on uncorrected p-distances of the RNA polymerase II largest subunit (1,268 sites in alignment) gene sequences from [334] with *L. major* as an outgroup. There is 1 substitution between HO174 and *L. adleri* SKINK-7, 17 substitutions between *L. adleri* 1433 and HO174, 34 each between *L. tarentolae* TarVI, *L. tarentolae* ParrotTar-II, *L. hoogstraali* NG-26 and *L. gymnodactyli* Ag compared with HO174, 19 between *L. adleri* 1433 and *L. tarentolae* TarVI and 130 between *L. major* 5-ASKH and HO174.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***L. tarentolae* Parrot-TarII** | 0 | | | | | | | |
| ***L. adleri* SKINK-7** | 14 | 0 | | | | | | |
| **HO174** | 15 | 1 | 0 | | | | | |
| ***L. major* 5-ASKH** | 69 | 72 | 73 | 0 | | | | |
| ***L. tarentolae* TarVI** | 2 | 14 | 15 | 67 | 0 | | | |
| ***L. hoogstraali* NG-26** | 6 | 16 | 17 | 67 | 4 | 0 | | |
| ***L. gymnodactyli* Ag** | 8 | 20 | 21 | 72 | 6 | 10 | 0 | |
| ***L.adleri* 1433** | 13 | 3 | 4 | 69 | 11 | 13 | 17 | 0 |
| | ***L. tarentolae* Parrot-TarII** | ***L. adleri* SKINK-7** | **HO174** | ***L. major* 5-ASKH** | ***L. tarentolae* TarVI** | ***L. hoogstraali* NG-26** | ***L. gymnodactyli* Ag** | ***L.adleri* 1433** |

**Table A3:** Number of substitutions between each isolate's DNA polymerase $\alpha$ catalytic polypeptide gene used in Figure A3(a).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***L. tarentolae* Parrot-TarII** | 0 | | | | | | | |
| ***L. adleri* SKINK-7** | 35 | 0 | | | | | | |
| ***L. tarentolae* TarVI** | 2 | 35 | 0 | | | | | |
| ***L. hoogstraali* NG-26** | 26 | 35 | 24 | 0 | | | | |
| ***L. gymnodactyli* Ag** | 18 | 35 | 16 | 22 | 0 | | | |
| ***L.adleri* 1433** | 21 | 18 | 19 | 17 | 17 | 0 | | |
| **HO174** | 34 | 1 | 34 | 34 | 34 | 17 | 0 | |
| ***L. major* 5-ASKH** | 123 | 131 | 123 | 120 | 124 | 122 | 130 | 0 |
| | ***L. tarentolae* Parrot-TarII** | ***L. adleri* SKINK-7** | ***L. tarentolae* TarVI** | ***L. hoogstraali* NG-26** | ***L. gymnodactyli* Ag** | ***L.adleri* 1433** | **HO174** | ***L. major* 5-ASKH** |

**Table A4:** Number of substitutions between each isolate's RNA polymerase II largest subunit gene used in Figure A3(b).

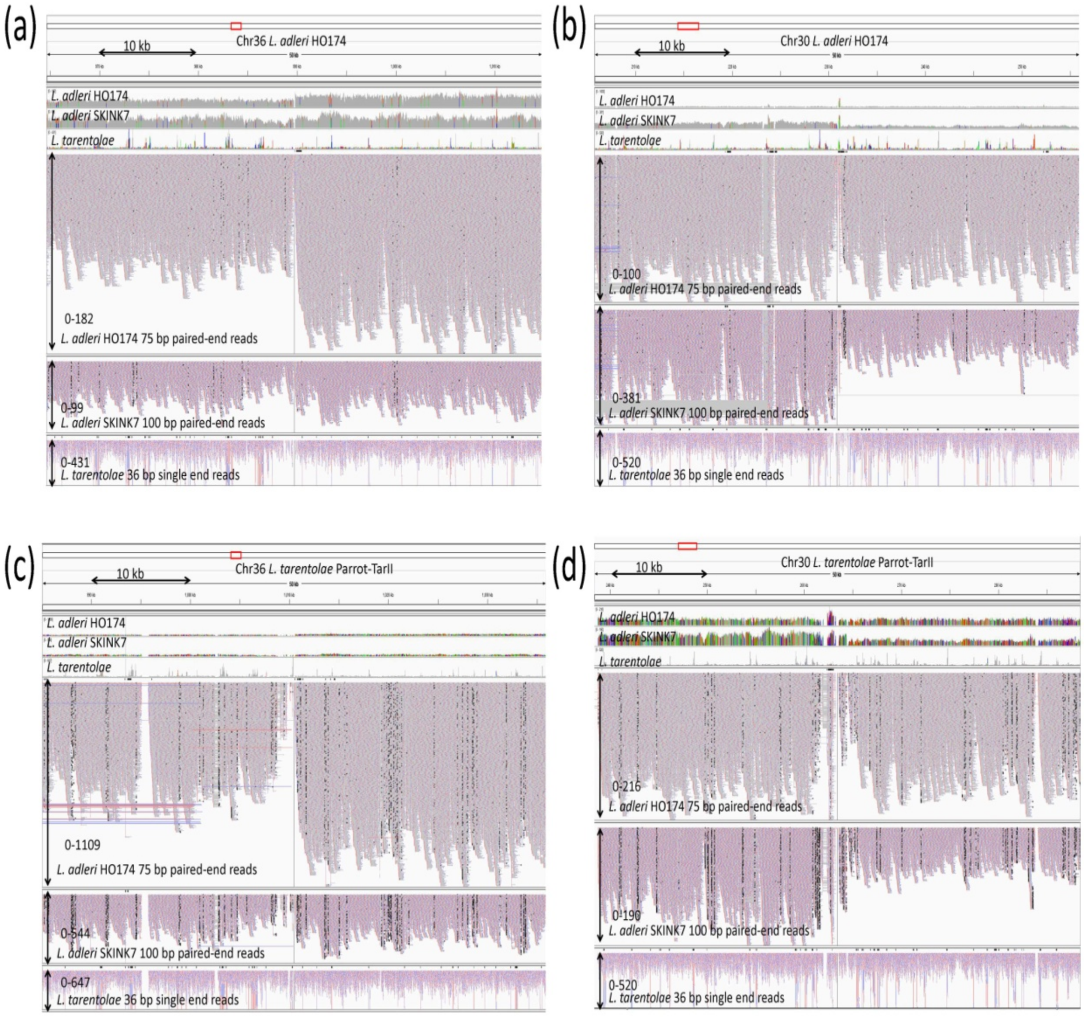

(a)
10 kb
Chr36 L. adleri HO174
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-182
L. adleri HO174 75 bp paired-end reads
0-99
L. adleri SKINK7 100 bp paired-end reads
0-431
L. tarentolae 36 bp single end reads
(b)
10 kb
Chr30 L. adleri HO174
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-100
L. adleri HO174 75 bp paired-end reads
0-381
L. adleri SKINK7 100 bp paired-end reads
0-520
L. tarentolae 36 bp single end reads
(c)
10 kb
Chr36 L. tarentolae Parrot-TarII
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-1109
L. adleri HO174 75 bp paired-end reads
0-544
L. adleri SKINK7 100 bp paired-end reads
0-647
L. tarentolae 36 bp single end reads
(d)
10 kb
Chr30 L. tarentolae Parrot-TarII
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-216
L. adleri HO174 75 bp paired-end reads
0-190
L. adleri SKINK7 100 bp paired-end reads
0-520
L. tarentolae 36 bp single end reads

**Figure A4:** Screenshot of the IGV browser showing chromosomes 30 and 36 over 50 kb windows with *L. adleri* HO174, *L. adleri* SKINK-7 and *L. tarentolae* Parrot-TarII reads mapped to reference sequences for the: (a) HO174 chromosome 36 with a gap at bases 989,697-989,797; (b) HO174 chromosome 30 with a gap at bases 230,911-231,011; (c) Parrot-TarII chromosome 36 with a 150 bp gap at bases 1,010,425-1,010,575; and (d) Parrot-TarII chromosome 30 with a 150 bp gap at bases 264,379-264,529. The regions shown are homologous, including the gaps on chromosomes 30 and 36 in HO174 and Parrot-TarII. The coverage of each set of reads is represented in the upper three panels of each figure and on the left-hand scale with the numbers representing the minimum and maximum coverage for the view, such as 0-182 for *L. adleri* HO174 in (a). Read colours indicate the read strand with red for positive strand reads (5' to 3') and blue for reverse strand (reverse complement) reads and grey lines join the mates of mapped read pairs. Vertical lines are midway through each gap and the 5' ends of these gaps are the fission breakpoints. (a) and (c) show a coverage increase 3' end of the chromosome 36 gap in HO174, but not for SKINK-7 or Parrot-TarII. (b) and (d) show a coverage increase 3' end of the chromosome 30 gap in SKINK-7, but not in HO174 or Parrot-TarII.

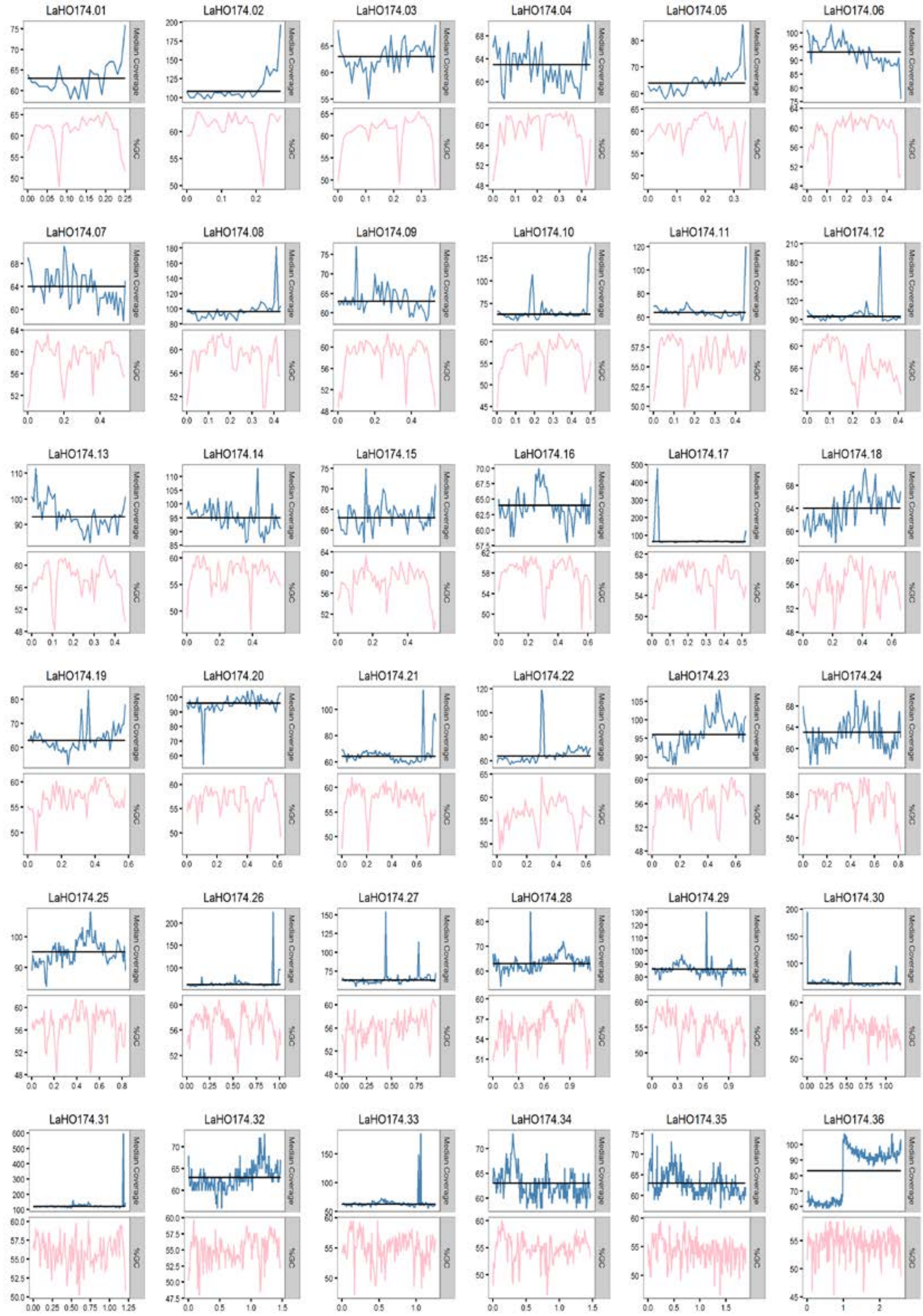


**Figure A5:** Median coverage of *L. adleri* HO174 mapped to itself with unbroken chromosomes 30 and 36 measured in 10 kb intervals (blue lines) for each chromosome. Black lines indicate median chromosomal coverage for that chromosome and pink lines show GC content in 10 kb intervals.

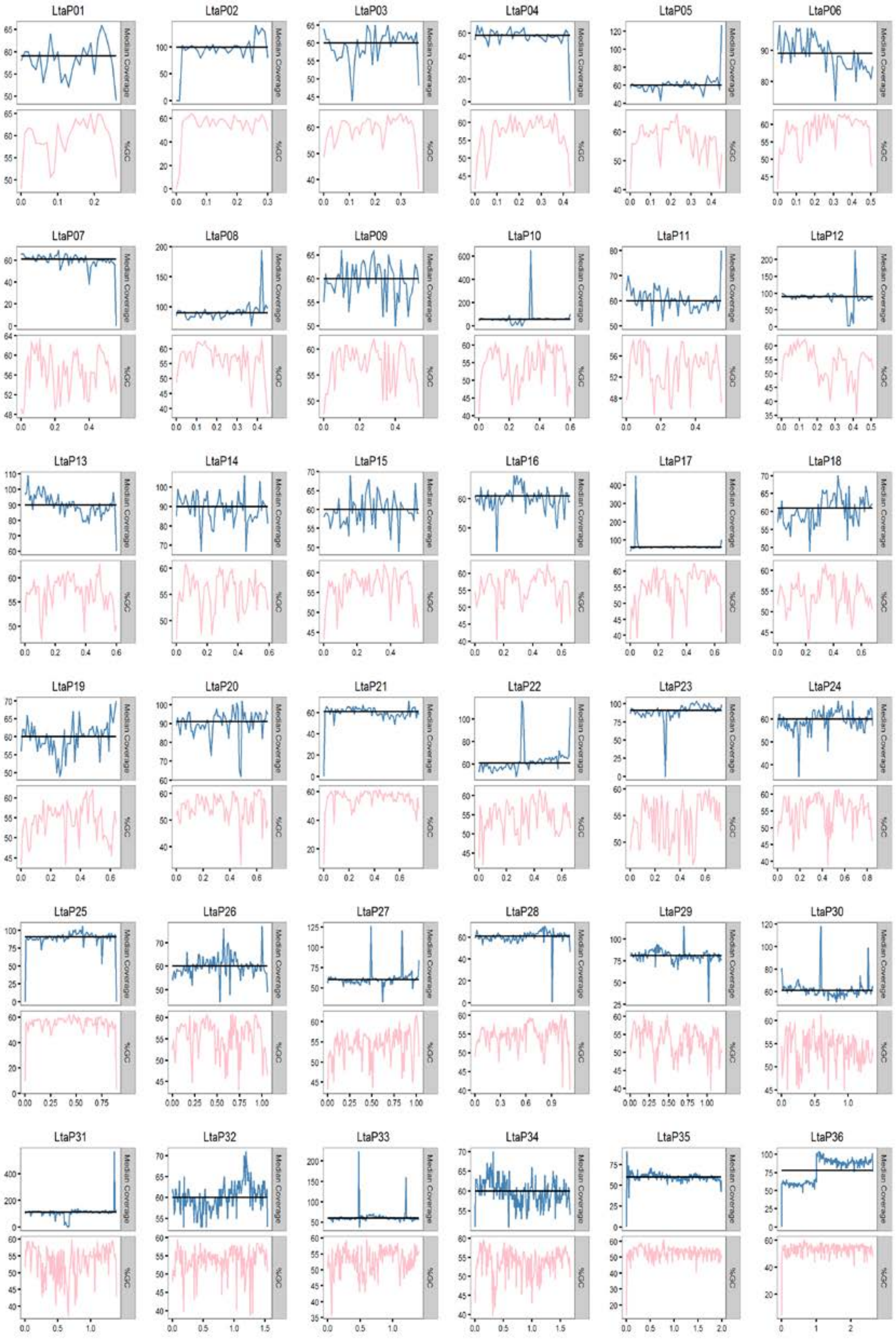


**Figure A6:** Median coverage of *L. adleri* HO174 mapped to *L. tarentolae* measured in 10 kb intervals (blue lines) for each chromosome. Black lines indicate median chromosomal coverage for that chromosome and pink lines show GC content in 10 kb intervals.

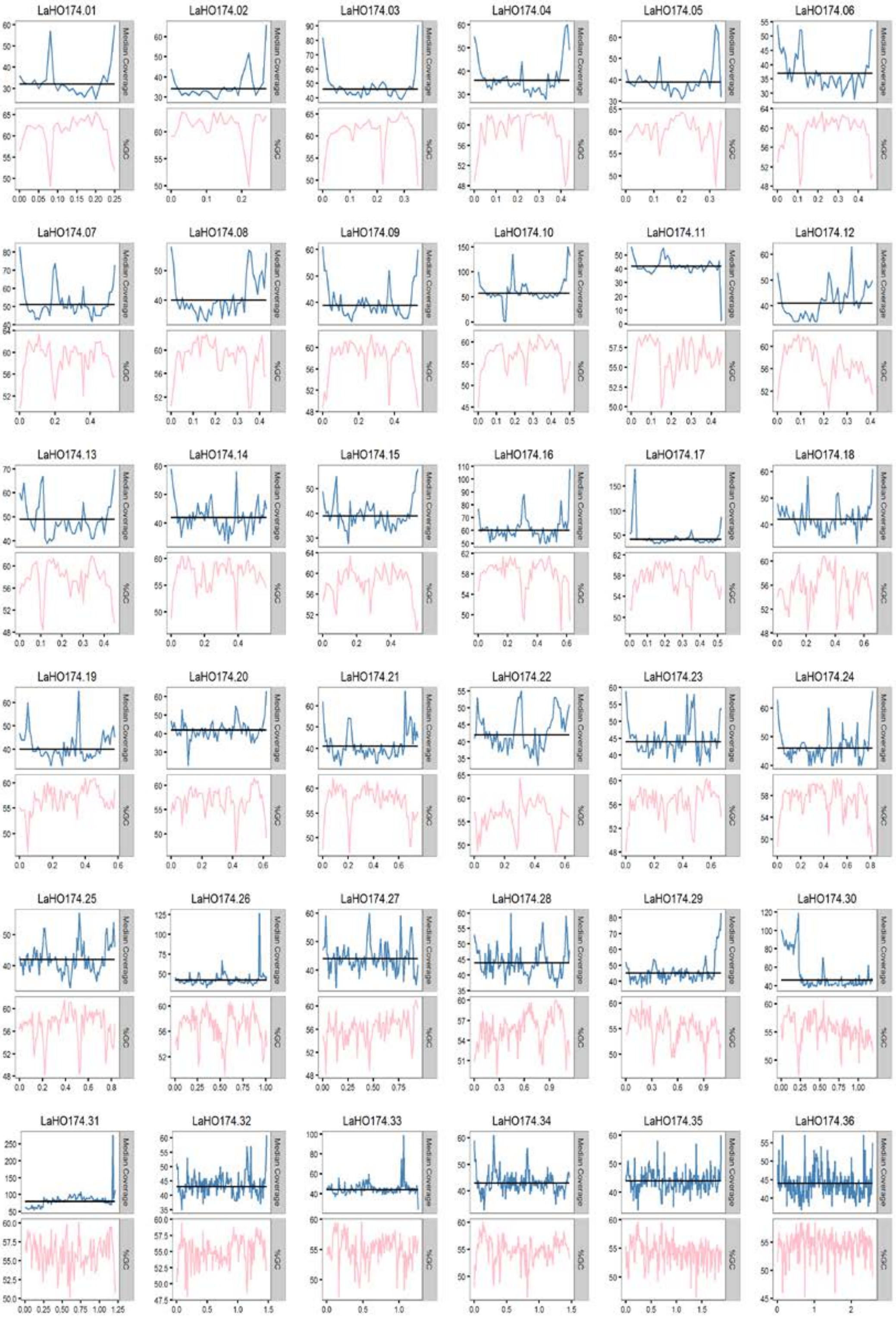


**Figure A7:** Median coverage of *L. adleri* SKINK-7 mapped to *L. adleri* HO174 with unbroken chromosomes 30 and 36 measured in 10 kb intervals (blue lines) for each chromosome. Black lines indicate median chromosomal coverage for that chromosome and pink lines show GC content in 10 kb intervals.

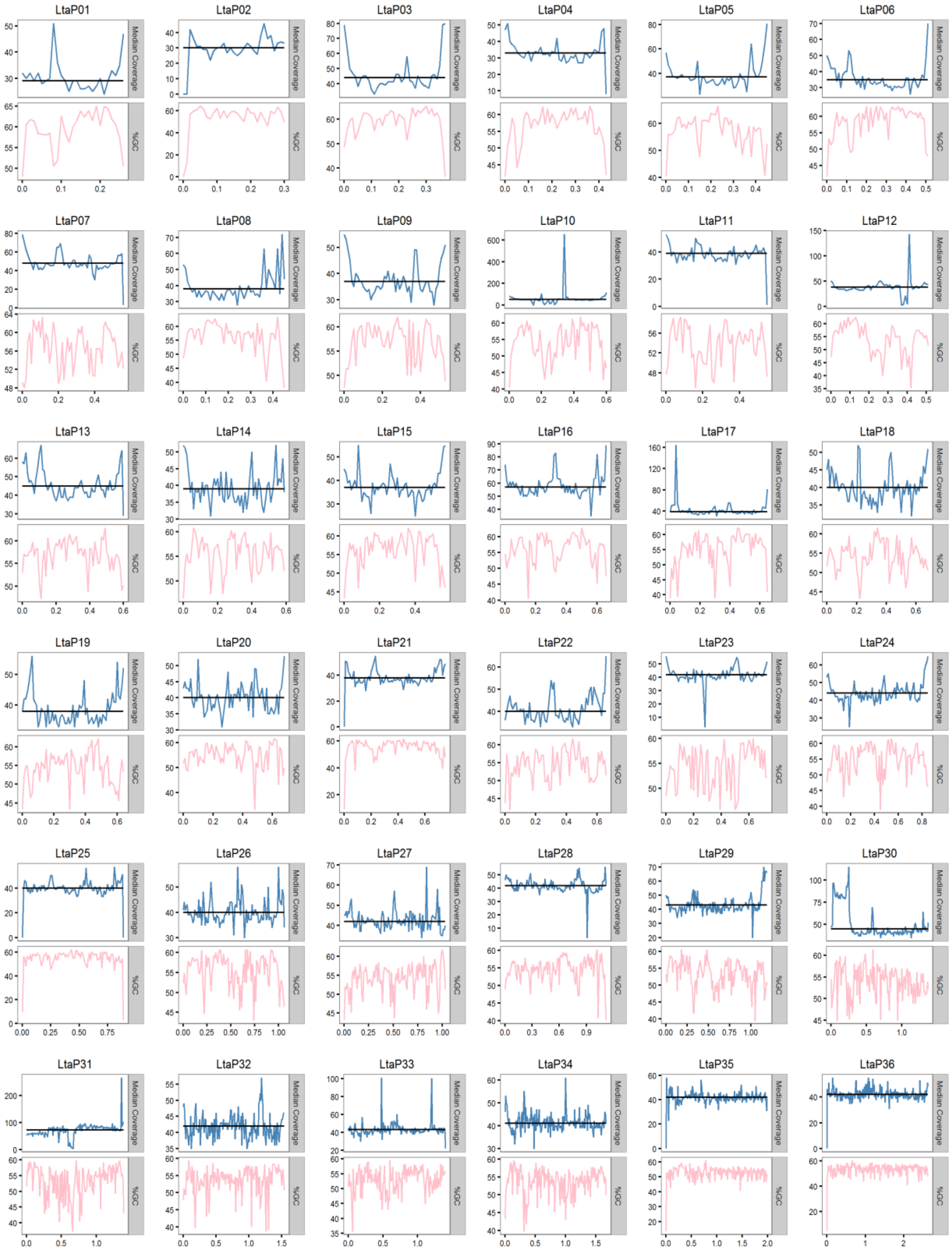


**Figure A8:** Median coverage of *L. adleri* SKINK-7 mapped to *L. tarentolae* measured in 10 kb intervals (blue lines) for each chromosome. Black lines indicate median chromosomal coverage for that chromosome and pink lines show GC content in 10 kb intervals.

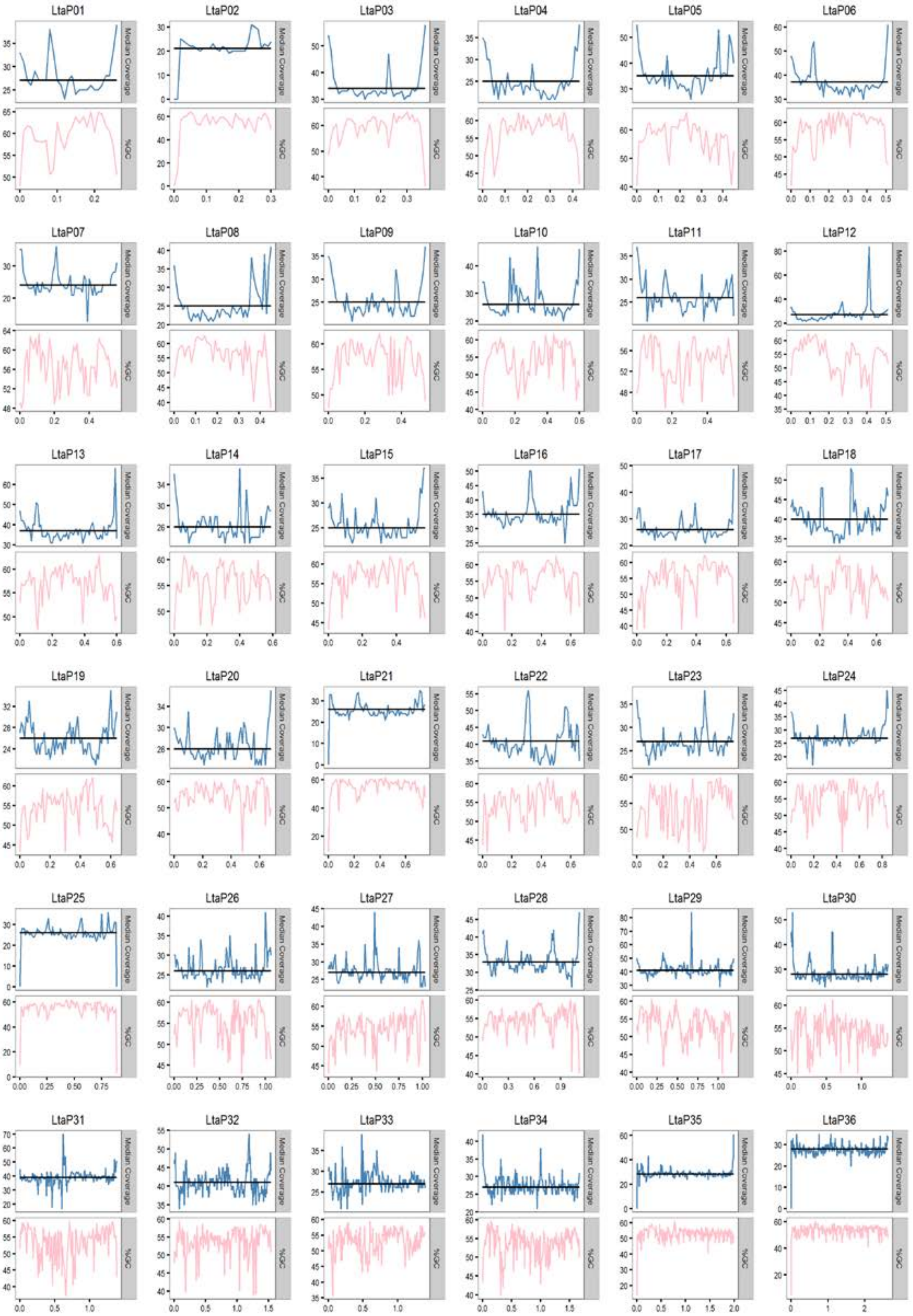


**Figure A9:** Median coverage of *L. tarentolae* mapped to itself measured in 10 kb intervals (blue lines) for each chromosome. Black lines indicate median chromosomal coverage for that chromosome and pink lines show GC content in 10 kb intervals.

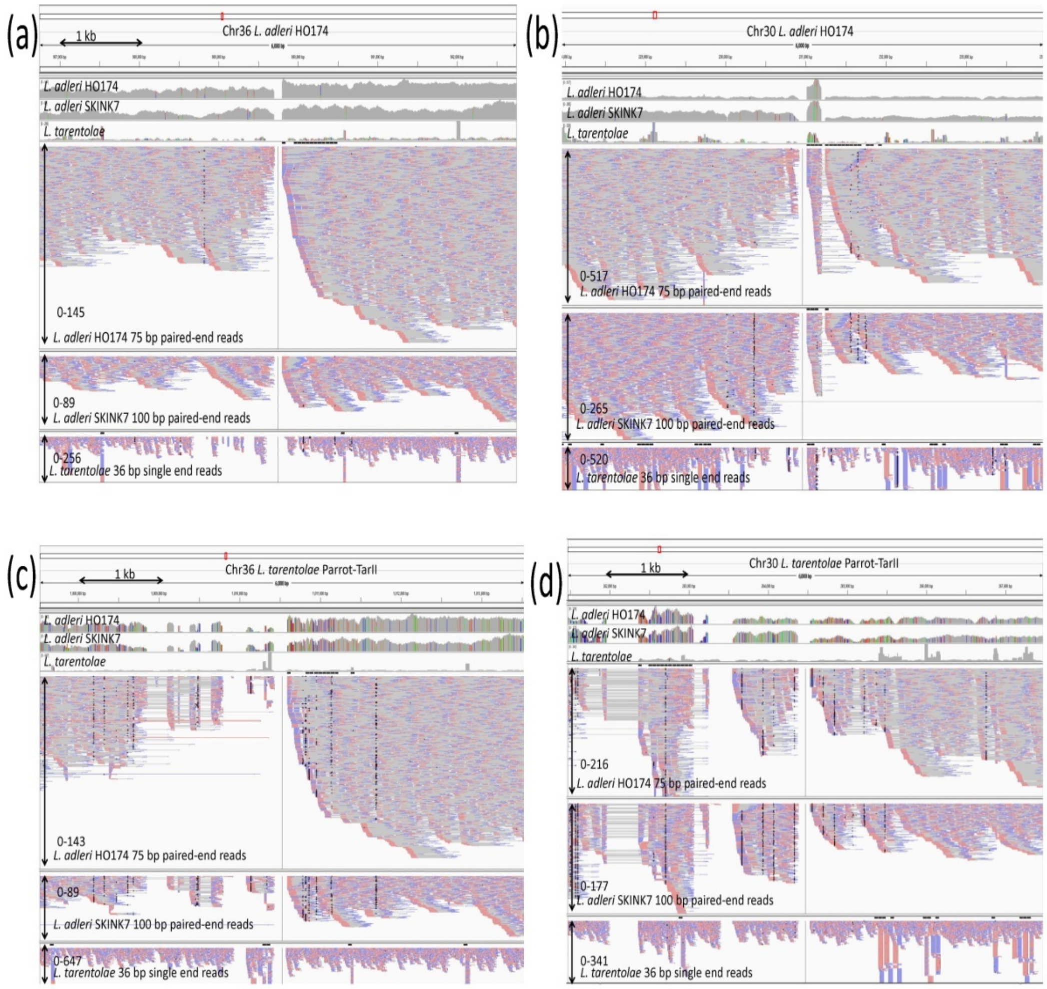

(a)
1 kb
Chr36 L. adleri HO174
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-145
L. adleri HO174 75 bp paired-end reads
0-89
L. adleri SKINK7 100 bp paired-end reads
0-256
L. tarentolae 36 bp single end reads
(b)
Chr30 L. adleri HO174
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-517
L. adleri HO174 75 bp paired-end reads
0-265
L. adleri SKINK7 100 bp paired-end reads
0-520
L. tarentolae 36 bp single end reads
(c)
1 kb
Chr36 L. tarentolae Parrot-TarII
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-143
L. adleri HO174 75 bp paired-end reads
0-89
L. adleri SKINK7 100 bp paired-end reads
0-647
L. tarentolae 36 bp single end reads
(d)
1 kb
Chr30 L. tarentolae Parrot-TarII
L. adleri HO174
L. adleri SKINK7
L. tarentolae
0-216
L. adleri HO174 75 bp paired-end reads
0-177
L. adleri SKINK7 100 bp paired-end reads
0-341
L. tarentolae 36 bp single end reads

**Figure A10:** Screenshot of the IGV browser showing chromosomes 30 and 36 over 6 kb windows with *L. adleri* HO174, *L. adleri* SKINK-7 and *L. tarentolae* Parrot-TarII reads mapped to reference sequences for the: (a) HO174 chromosome 36 with a gap at bases 989,697-989,797; (b) HO174 chromosome 30 with a gap at bases 230,911-231,011; (c) Parrot-TarII chromosome 36 with a 150 bp gap at bases 1,010,425-1,010,575; and (d) Parrot-TarII chromosome 30 with a 150 bp gap at bases 264,379-264,529. The regions shown are homologous, including the gaps on chromosomes 30 and 36 in HO174 and Parrot-TarII. The coverage of each set of reads is represented in the upper three panels of each figure and on the left-hand scale with the numbers representing the minimum and maximum coverage for the view, such as 0-145 for *L. adleri* HO174 in (a). Read colours indicate the read strand with red for positive strand reads (5' to 3') and blue for reverse strand (reverse complement) reads and grey lines join the mates of mapped read pairs. Vertical lines are midway through each gap and the 5' ends of these gaps are the fission breakpoints. (a) and (c) show a coverage increase 3' end of the chromosome 36 gap in HO174, but not for SKINK-7 or Parrot-TarII. (b) and (d) show a coverage increase 3' end of the chromosome 30 gap in SKINK-7, but not in HO174 or Parrot-TarII. *L. adleri* HO174 and SKINK-7 paired end reads have their mates joined by grey lines to show the location of both mapped mates of a pair: this shows no read pairs cross the fission breakpoints in HO174 (a) or Parrot-TarII chromosome 36 (c). For Parrot-TarII chromosome 30, no read pairs cross the fission breakpoint but for HO174

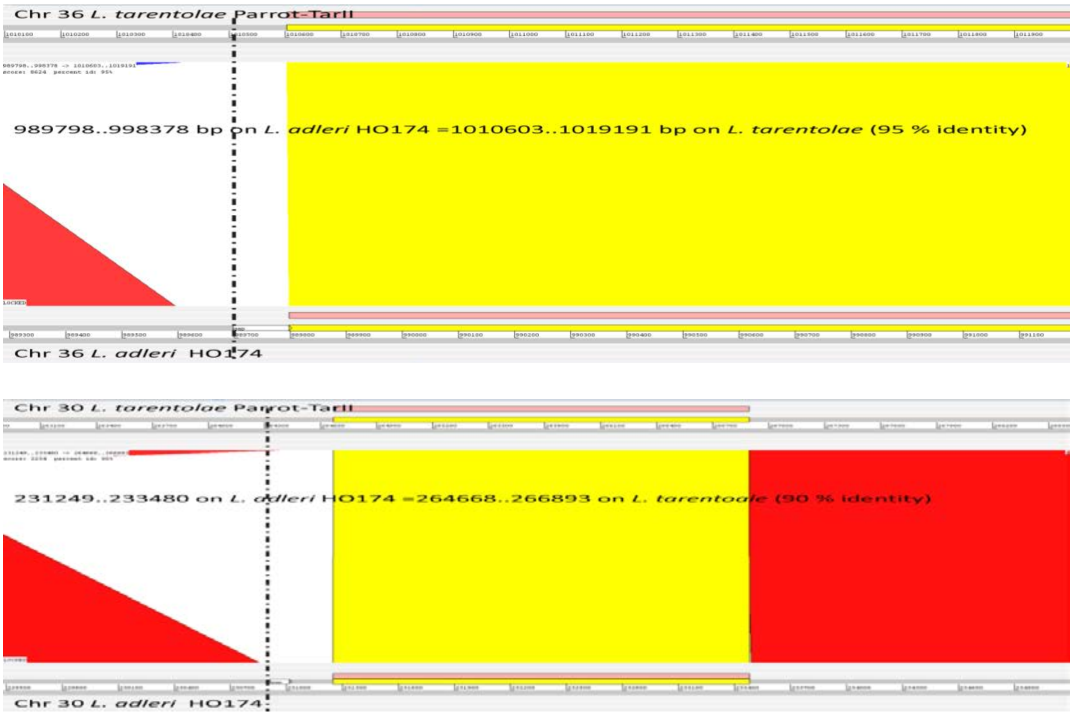

Chr 36 L. tarentolae Parrot-TarII
989798..998378 bp on L. adleri HO174 =1010603..1019191 bp on L. tarentolae (95 % identity)
Chr 36 L. adleri HO174
Chr 30 L. tarentolae Parrot-TarII
231249..233480 on L. adleri HO174 =264668..266893 on L. tarentoale (90 % identity)
Chr 30 L. adleri HO174

**Figure A11:** Artemis Comparison Tool screenshots of homologous segments of *L. tarentolae* Parrot-TarII chromosomes 30 (top) and 36 (bottom) aligned with *L. adleri* HO174. The dashed black lines indicate the fission breakpoints. The yellow segment in each plot indicates the nearest homologous *L. tarentolae* segments.

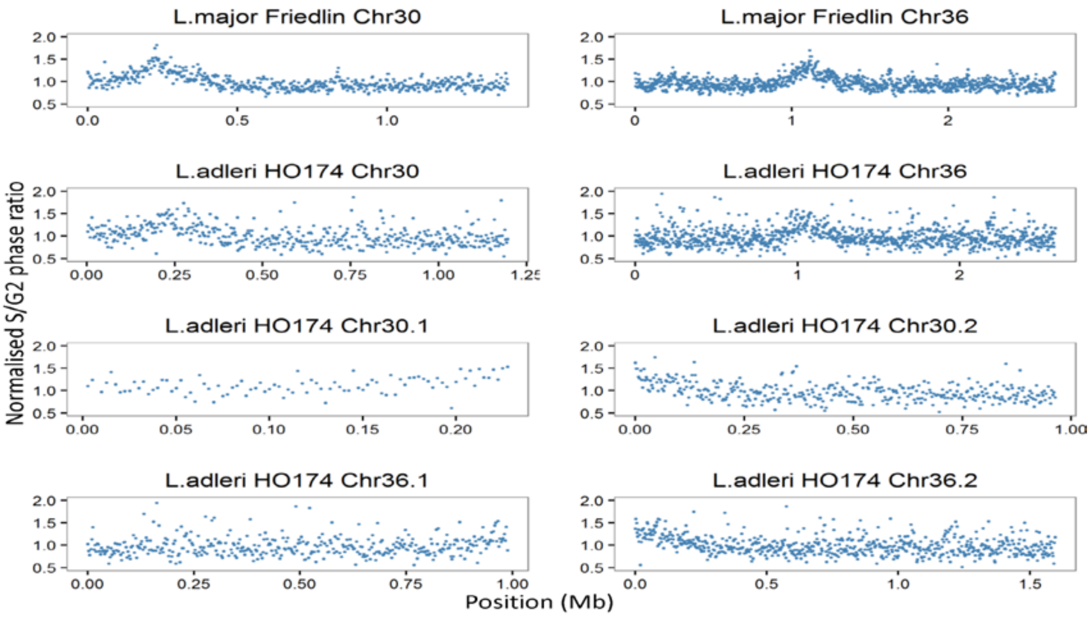


**Figure A12:** The early S/G2 ratio of read coverage across chromosomes of *L. major* MFAseq data from [362] mapped to the *L. major* Friedlin (top row) and *L. adleri* HO174 (other three rows). Peaks in coverage indicate putative origins of replications and the y-axis shows the early S/G2 ratio for the *L. major* reads mapped to *L. major* genome or *L. adleri* HO174. The y-axis is scaled to 0.5-2.0 for each chromosome for ease of comparison; however, *L. adleri* chromosomes had some values that were close to 4 on the y-axis in its plots.

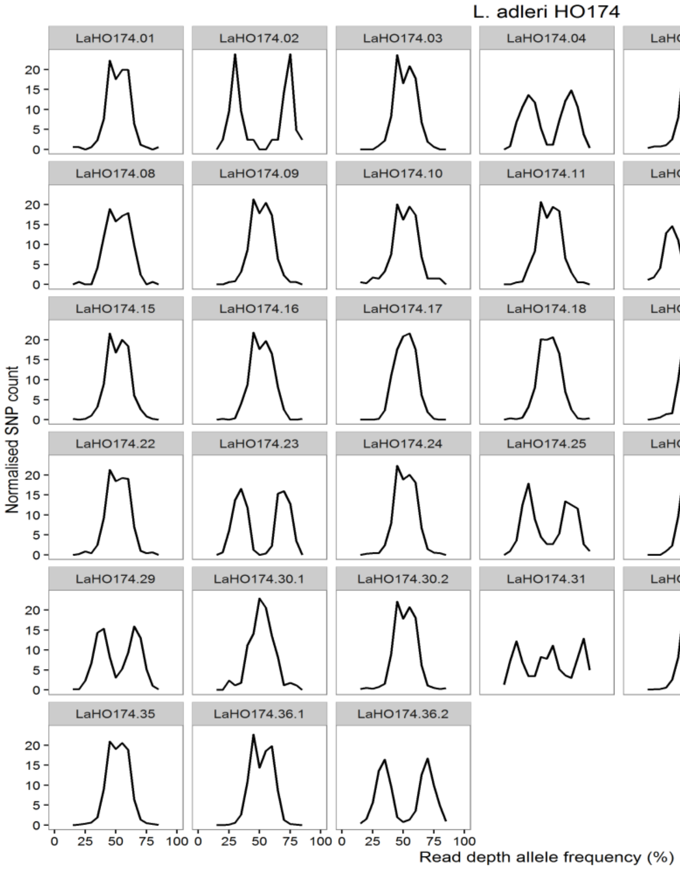


**Figure A13:** Read depth allele frequency distributions for each chromosome of *L. adleri* HO174 based on heterozygous SNPs called from self mapped reads

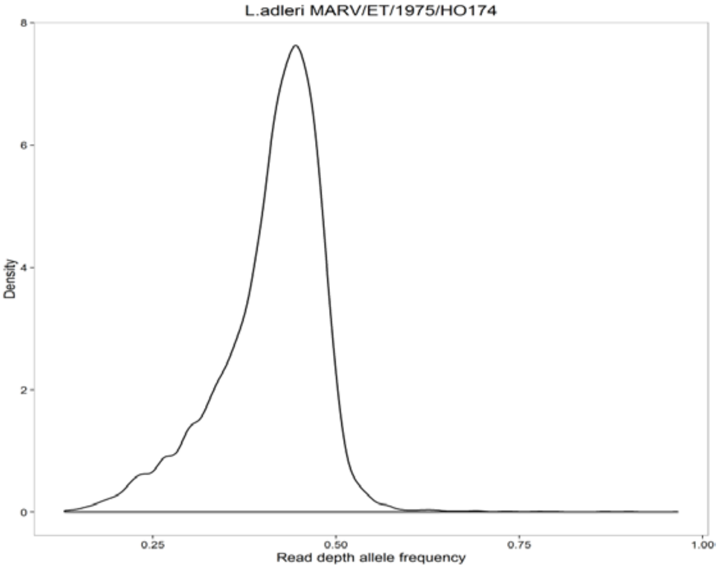


**Figure A14:** Density plot of read depth allele frequency distribution of heterozygous SNPs called from self-mapped reads for *L. adleri* HO174

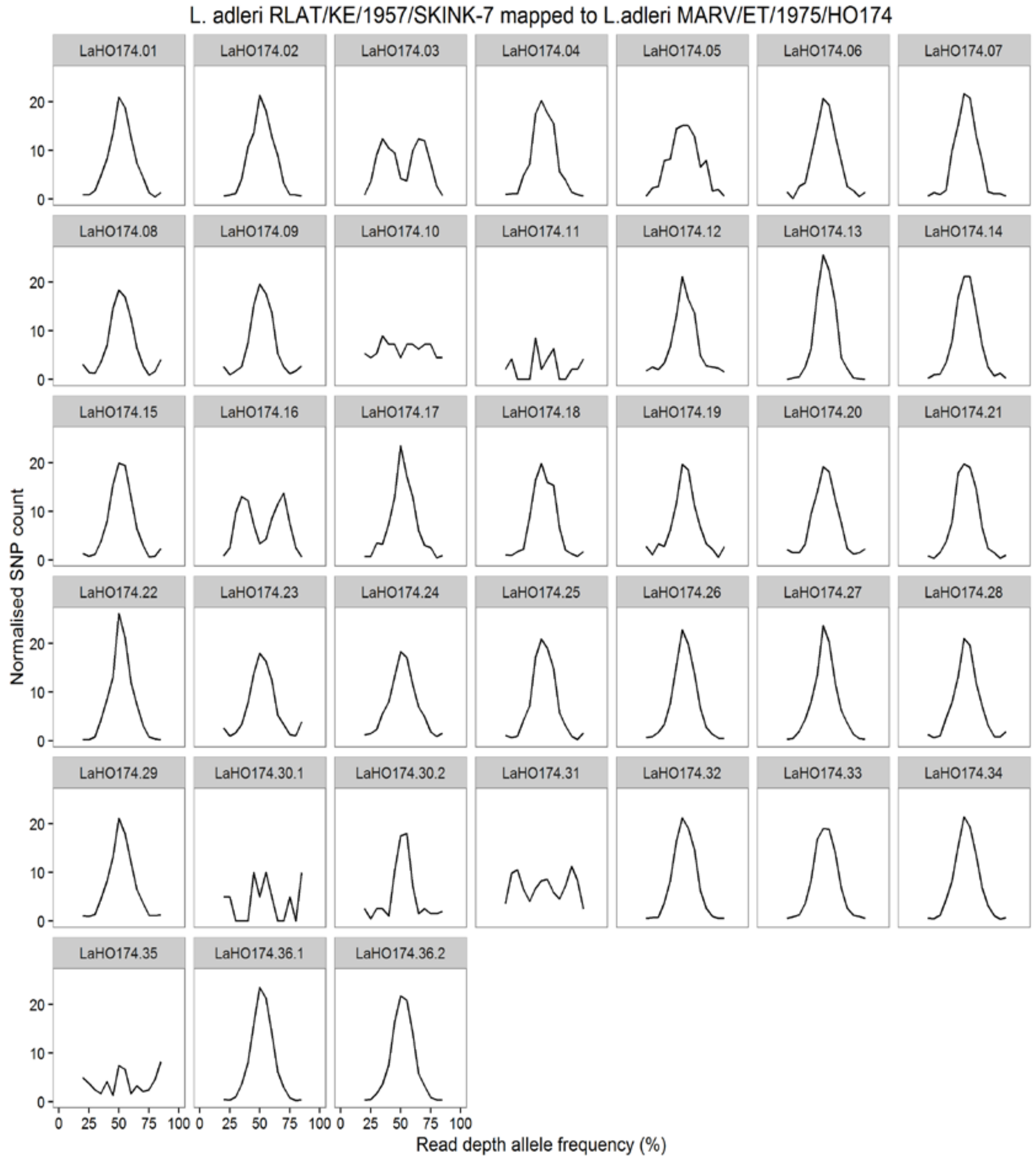


**Figure A15:** Read depth allele frequency distributions for each chromosome of *L. adleri* SKINK-7 based on heterozygous SNPs called from its reads mapped to *L. adleri* HO174

**Table A5:** Annotation manually added to the *L. adleri* HO174 genome. BLASTP hits are to *L. major* strain Friedlin and not top hits. If there was no hit to *L. major* then the best BLASTP hit was used. Partial genes have 3' ends shortened to the edge of gaps of unknown length (100 bp gaps). This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

**Table A6:** Orthologous groups present in >=1 of 6 *Leishmania* species. This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

| OG | | Number of assembled genes in OG | | | OG haploid copy number | |
|---|---|---|---|---|---|---|
| ID | Gene product(s) | *L. adleri* | *L. tarentolae* | *L. major* | *L. adleri HO174* | *L. major* |
| OG5_126605 | alpha tubulin | 1 | 0 | 12 | 18 | 10 |
| OG5_126708 | nucleoside diphosphate kinase b | 1 | 0 | 1 | 5 | 1 |
| OG5_126854 | glycoprotein 96-92, putative ,kinetoplast-associated protein-like protein | 1 | 0 | 1 | 1 | 1 |
| OG5_126940 | 60S ribosomal protein L19, putative | 2 | 0 | 2 | 2 | 2 |
| OG5_127051 | 60S acidic ribosomal subunit protein, putative | 2 | 0 | 2 | 2 | 2 |
| OG5_128780 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_129181 | calreticulin, putative | 1 | 0 | 1 | 1 | 1 |
| OG5_131531 | hypothetical protein, unknown function | 1 | 0 | 1 | 1 | 1 |
| OG5_132180 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_137181 | ATG8/AUT7/APG8/PAZ2, putative | 2 | 0 | 9 | 9 | 19 |
| OG5_144949 | folate/biopterin transporter, putative ,pteridine transporter (truncated), putative | 1 | 0 | 1 | 1 | 1 |
| OG5_145575 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_154529 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |
| OG5_166586 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |
| OG5_166721 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_173498 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_177991 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |
| OG5_178818 | protein kinase-like protein | 1 | 0 | 1 | 1 | 1 |
| OG5_183311 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |
| OG5_183583 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_184030 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_184115 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |
| OG5_204147 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |
| OG5_204150 | serine/threonine-protein phosphatase PP1, putative | 1 | 0 | 1 | 1 | 1 |
| OG5_204190 | hypothetical protein, conserved | 1 | 0 | 1 | 1 | 1 |
| OG5_204216 | hypothetical protein | 2 | 0 | 1 | 4 | 1 |
| OG5_204225 | DNA polymerase kappa, putative | 1 | 0 | 1 | 1 | 1 |
| OG5_204247 | hypothetical protein | 1 | 0 | 1 | 1 | 1 |

**Table A7:** Genes with orthologs in *L. adleri* HO174 and *L. major* that are absent in *L. tarentolae.* OG stands for orthologous group and OG haploid copy number is the number of genes in the OG predicted using read depth analysis.

**Table A8:** Genes exclusive to *L. adleri* HO174 compared with *L. tarentolae* and *L. major.* This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

| Gene ID | Chromosome/ Bin Sequence ID | E-value | Species | Accesion | Product | Number of domains | Domain Name | Domain ID |
|---|---|---|---|---|---|---|---|---|
| LaHO174_3010020 | 30 | 6.00 E-79 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 3 | VSP (X2 domains)(Giardia variant-specific surface protein);Chaperone protein DnaJ; Provisional | pfam03302,PRK14286 |
| LaHO174_bin710010 | Bin_71 | 7.00 E-77 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 2 | VSP (X2 domains)(Giardia variant-specific surface protein) | pfam03302 |
| LaHO174_bin2140010 | Bin_214 | 1.00 E-58 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 2 | VSP(Giardia variant-specific surface protein); Chaperone protein DnaJ; Provisional | pfam03302, PRK14276 |
| LaHO174_bin2140020 | Bin_214 | 2.00 E-61 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 1 | VSP(Giardia variant-specific surface protein) | pfam03302 |
| LaHO174_bin440010 | Bin_44 | 4.00 E-70 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 2 | VSP(Giardia variant-specific surface protein); Chaperone protein DnaJ; Provisional | pfam03302, PRK14276 |
| LaHO174_bin2090010 | Bin_209 | 6.00 E-78 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 2 | VSP (X2 domains)(Giardia variant-specific surface protein) | pfam03302 |
| LaHO174_bin1690010 | Bin_169 | 1.00 E-57 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 1 | VSP(Giardia variant-specific surface protein) | pfam03302 |
| LaHO174_bin1420010 | Bin_142 | 8.00 E-62 | Phytomonas sp. isolate Hart1 | CCW72337.1 | Unnamed protein product | 1 | VSP(Giardia variant-specific surface protein) | pfam03302 |

**Table A9:** BLASTP results for hypothetical proteins in the OG5_136043 orthologous group.

| OG | | Number of assembled genes in OG | | | Assembled genes in OG | | OG haploid copy number | |
|---|---|---|---|---|---|---|---|---|
| ID | Gene product(s) | *L. adleri* | *L. tarentolae* | *L. major* | *L.adleri* | *L. tarentolae* | *L. adleri* | *L. tarentolae* |
| OG5_127311 | GTPase activator protein, putative | 1 | 1 | 0 | LaHO174_291540 | LtaP29.1700 | 2.02 | 0.9 |
| OG5_127897 | cyclopropane-fatty-acyl-phospholipid synthase | 1 | 1 | 0 | LaHO174_080550 | LtaP08.0530 | 1.04 | 1.12 |
| OG5_128381 | Sedlin, N-terminal conserved region containing protein,putative,hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_161590 | LtaP16.1750 | 1.03 | 1.26 |
| OG5_128947 | lectin, putative | 1 | 1 | 0 | LaHO174_bin770170 | LtaP13.0560 | 1.06 | 0.95 |
| OG5_129607 | hypothetical protein | 1 | 1 | 0 | LaHO174_180630 | LtaP18.0640 | 1.05 | 0.88 |
| OG5_134909 | surface antigen-like protein | 1 | 1 | 0 | LaHO174_211420 | LtaP21.1350 | 1.11 | 1 |
| OG5_151269 | RNA binding protein | 1 | 1 | 0 | LaHO174_190280 | LtaP19.0270 | 1.08 | 0.88 |
| OG5_154744 | hypothetical protein, | 1 | 6 | 0 | LaHO174_bin1350010 | LtaP29.1170,LtaP31.1520,LtaPcontig04212-2,LtaP31.1630,LtaP31.1450, LtaP31.1490 | 1.07 | 8.2 |
| OG5_155317 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_231680 | LtaP23.1710 | 1.06 | 1.11 |
| OG5_155868 | RING-H2 zinc finger/Anaphase-promoting complex subunit 11 RING-H2 finger/Ring finger domain containing protein, putative,hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_160590 | LtaP16.0620 | 1.11 | 0.77 |
| OG5_160079 | hypothetical protein ,hypothetical protein | 1 | 1 | 0 | LaHO174_030010 | LtaP03.0010 | 1.06 | 1.79 |
| OG5_161893 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_302222000 | LtaP30.2100 | 0.89 | 0.96 |
| OG5_162207 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_331480 | LtaP33.1840 | 1.13 | 1.33 |
| OG5_162271 | hypothetical protein, | 1 | 1 | 0 | LaHO174_331830 | LtaP33.2230 | 1.06 | 0.93 |
| OG5_162295 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_bin010240 | LtaP34.1370 | 0.97 | 0.59 |
| OG5_162319 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_262670 | LtaP25.0070 | 1.56 | 1.31 |
| OG5_162394 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_352930 | LtaP35.3180 | 0.97 | 0.82 |
| OG5_167675 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_bin050150 | LtaP33.1400 | 1.09 | 0.85 |
| OG5_16 | hypothetical | 1 | 1 | 0 | LaHO174_362 | LtaP36.6610 | 1.07 | 1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 7687 | protein | | | | 3990 | | | |
| OG5_173551 | hypothetical protein,conserved | 1 | 1 | 0 | LaHO174_290320 | LtaP29.0320 | 0.85 | 0.9 |
| OG5_174212 | hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_140910 | LtaP14.0880 | 1.08 | 1 |
| OG5_186763 | hypothetical protein | 1 | 1 | 0 | LaHO174_3022520 | LtaP30.2600 | 1.02 | 0.89 |
| OG5_204133 | hypothetical protein, | 1 | 2 | 0 | LaHO174_040310 | LtaP04.0330,LtaP04.0340 | 1.03 | 1.8 |
| OG5_204146 | hypothetical protein | 1 | 2 | 0 | LaHO174_150460 | LtaP15.0440,LtaP17.0930 | 1.02 | 1.73 |
| OG5_204148 | hypothetical protein | 1 | 1 | 0 | LaHO174_151290 | LtaP15.1260 | 0.98 | 1 |
| OG5_204162 | hypothetical protein | 1 | 1 | 0 | LaHO174_321590 | LtaP32.1610 | 1.03 | 0.85 |
| OG5_236226 | hypothetical protein | 1 | 1 | 0 | LaHO174_241430 | LtaP24.1560 | 1.02 | 1 |
| OG5_236274 | multi drug resistance protein-like ,hypothetical protein, conserved | 1 | 1 | 0 | LaHO174_241460 | LtaP24.1590 | 1.08 | 1.04 |
| OG5_236303 | hypothetical protein | 1 | 1 | 0 | LaHO174_3621490 | LtaP36.4110 | 1.03 | 1.04 |
| OG5_247477 | hypothetical protein, | 1 | 1 | 0 | LaHO174_250500 | LtaP25.0630 | 0.99 | 0.85 |

**Table A10:** Genes with orthologs in *L. adleri* and *L. tarentolae* but not *L. major*

| **OG** | | **Number of assembled genes in OG** | | | **Assembled genes in OG** | **OG haploid copy number** |
|---|---|---|---|---|---|---|
| **ID** | **Gene product(s)** | ***L. adleri*** | ***L. tarentolae*** | ***L. major*** | ***L. tarentolae*** | ***L. tarentolae*** |
| OG5_129777 | surface protease GP63, putative | 0 | 1 | 0 | LtaP10.0550 | 0.46 |
| OG5_131800 | surface protease GP63 (pseudogene), putative, ,surface protease GP63, putative | 0 | 1 | 0 | LtaP10.0570 | 0.54 |
| OG5_132303 | hypothetical protein, conserved | 0 | 1 | 0 | LtaP21.1880 | 1.58 |
| OG5_132413 | Surface antigen like protein, High cysteine membrane protein Group 2,Matrilin-2 precursor.-related , hypothetical protein,conserved | 0 | 1 | 0 | LtaP11.1290 | 3.65 |
| OG5_133767 | zinc finger protein, putative | 0 | 1 | 0 | LtaP26.0070 | 1.12 |
| OG5_137449 | hypothetical protein, | 0 | 1 | 0 | LtaP09.1020 | 1.24 |
| OG5_141868 | expression site-associated gene,putative,expression site-associated gene 8 (ESAG8), pseudogene,chrIX | 0 | 1 | 0 | LtaP24.1490 | 0 |
| OG5_143904 | amastin-like surface protein, putative | 0 | 2 | 0 | LtaP34.0550,LtaP34.0560 | 4.04 |
| OG5_148243 | hypothetical protein, conserved , | 0 | 1 | 0 | LtaP04.0450 | 0.96 |
| OG5_152437 | hypothetical protein, conserved | 0 | 1 | 0 | LtaP33.3470 | 0.96 |
| OG5_154295 | malate dehydrogenase,mitochondrial malate dehydrogenase, conserved hypothetical protein | 0 | 1 | 0 | LtaP34.0490 | 1 |
| OG5_157987 | helicase-like protein,DNA repair and recombination protein, mitochondrial precursor, putative | 0 | 1 | 0 | LtaP32.1700 | 0.95 |

| | | | | | | |
|---|---|---|---|---|---|---|
| OG5_158273 | No description | 0 | 1 | 0 | LtaP27.2450 | 30.41 |
| OG5_165000 | protein kinase-like protein | 0 | 1 | 0 | LtaP35.1910 | 1.07 |
| OG5_169610 | surface antigen-like | 0 | 1 | 0 | LtaP09.1040 | 0.88 |
| OG5_174131 | hypothetical protein, | 0 | 1 | 0 | LtaP36.4850 | 0.71 |
| OG5_179378 | protein phosphatase-1, putative,Ser/Thr protein phosphatase family protein | 0 | 2 | 0 | LtaP34.0940,LtaP34.0890 | 0.63 |
| OG5_184203 | histone H2A, putative | 0 | 1 | 0 | LtaP29.1870 | 1.27 |
| OG5_204164 | hypothetical protein | 0 | 1 | 0 | LtaP33.0140 | 0.96 |
| OG5_236209 | GP63, leishmanolysin | 0 | 1 | 0 | LtaP10.0470 | 1.67 |
| OG5_236254 | hypothetical protein, conserved | 0 | 1 | 0 | LtaP05.1330 | 1 |
| OG5_247458 | hypothetical protein | 0 | 1 | 0 | LtaP36.6950 | 1.04 |

**Table A11:** *L. tarentolae* genes that do not have orthologs in *L. adleri* and *L. major*

**Table A12:** *L. adleri* gene arrays. This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

**Table A13:** *L. tarentolae* gene arrays. This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

**Table A14:** *L. major* gene arrays. This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

| **OG** | | **Number of assembled genes in OG** | **OG haploid copy number** |
|---|---|---|---|
| **ID** | **Description** | ***L. adleri*** | ***L. adleri*** |
| OG5_126631 | elongation factor 1-alpha | 1 | 52.59 |
| OG5_144952 | hypothetical protein | 3 | 43.22 |
| OG5_139233 | hypothetical protein | 13 | 30.36 |
| OG5_183275 | hypothetical protein, conserved | 2 | 23.5 |
| OG5_126605 | alpha tubulin | 1 | 18.49 |
| OG5_126611 | beta tubulin | 1 | 16.52 |
| OG5_126617 | receptor-type adenylate cyclase, putative, receptor-type adenylate cyclase b | 4 | 13.07 |
| OG5_164370 | hypothetical protein, conserved | 2 | 11.27 |
| OG5_126703 | polyubiquitin, putative | 2 | 11.05 |
| OG5_140928 | sodium stibogluconate resistance protein, putative | 2 | 10.81 |
| OG5_137181 | ATG8/AUT7/APG8/PAZ2, putative ,ATG8/AUT7/APG8/PAZ2, putative | 2 | 8.71 |

|  |  |  |  |
|---|---|---|---|
| OG5_135520 | hypothetical protein, conserved | 1 | 6.44 |
| OG5_130385 | paraflagellar rod protein 2C ,paraflagellar rod protein 1D | 2 | 6.39 |
| OG5_126923 | elongation factor 2 | 2 | 6.39 |
| OG5_145872 | ATG8/AUT7/APG8/PAZ2, putative | 1 | 6.37 |
| OG5_146058 | hypothetical protein, conserved | 1 | 5.46 |
| OG5_126910 | phosphoglycan beta 1,3 galactosyltransferase 5 , | 1 | 5.35 |
| OG5_126823 | ribonucleoside-diphosphate reductase small chain, putative | 1 | 5.14 |
| OG5_126708 | nucleoside diphosphate kinase b | 1 | 5.05 |
| OG5_148865 | hypothetical protein, conserved | 1 | 5 |
| OG5_129633 | glucose transporter, lmgt2 ,glucose transporter/membrane transporter D2, putative | 2 | 4.66 |
| OG5_126607 | cysteine peptidase A (CPA) ,cathepsin L-like protease | 2 | 4.33 |
| OG5_128109 | glycerol uptake protein, putative | 2 | 4.27 |
| OG5_126570 | histone H2A ,histone H2A, putative | 2 | 4.23 |
| OG5_126636 | ATP-dependent Clp protease subunit, heat shock protein 100 (HSP100), putative,serine peptidase, putative ,ATP-dependent Clp protease subunit, heat shock protein 100 (HSP100), putative,serine peptidase, putative | 2 | 4.16 |
| OG5_148058 | hypothetical protein, conserved | 1 | 4.11 |
| OG5_132183 | carboxypeptidase, putative,metallo-peptidase, Clan MA(E), family 32 | 2 | 4.05 |
| OG5_166764 | expression-site associated gene (ESAG3), putative | 1 | 4 |
| OG5_143928 | hypothetical protein, conserved | 1 | 3.98 |
| OG5_142234 | protein kinase, putative | 1 | 3.87 |
| OG5_146064 | hypothetical protein, conserved ,hypothetical protein, conserved ,hypothetical protein, conserved ,hypothetical protein, conserved ,hypothetical protein, conserved | 1 | 3.73 |
| OG5_128204 | fatty acid desaturase | 1 | 3.44 |
| OG5_128316 | myo-inositol-1-phosphate synthase,inositol-3-phosphate synthase, putative | 1 | 3.26 |
| OG5_127090 | 60S ribosomal protein L5, putative | 1 | 3.03 |
| OG5_126641 | 60S ribosomal protein L2, putative | 1 | 2.98 |
| OG5_157974 | surface protein amastin, putative | 1 | 2.95 |
| OG5_126769 | 10 kDa heat shock protein, putative | 1 | 2.95 |
| OG5_127099 | ATPase beta subunit, putative | 1 | 2.83 |
| OG5_127269 | 40S ribosomal protein S30, putative | 1 | 2.7 |
| OG5_135896 | hypothetical protein, conserved | 1 | 2.63 |
| OG5_126731 | branched-chain amino acid aminotransferase, putative | 1 | 2.61 |
| OG5_183887 | hypothetical protein, conserved | 1 | 2.59 |

|  |  |  |  |
|---|---|---|---|
| OG5_127995 | 60S ribosomal protein L28, putative | 1 | 2.56 |
| OG5_126800 | calmodulin, putative | 1 | 2.53 |
| OG5_127180 | nascent polypeptide associated complex subunit- like protein, copy 2 | 1 | 2.35 |
| OG5_183915 | hypothetical protein | 1 | 2.2 |
| OG5_127617 | IgE-dependent histamine-releasing factor, putative | 1 | 2.19 |
| OG5_127918 | hypothetical protein, conserved | 1 | 2.17 |
| OG5_127093 | D-lactate dehydrogenase-like protein | 1 | 2.17 |
| OG5_145959 | hypothetical protein, conserved | 1 | 2.16 |
| OG5_142209 | pumilio protein, putative,RNA-binding regulatory protein, putative | 1 | 2.13 |
| OG5_126734 | S-adenosylmethionine synthetase | 1 | 2.11 |
| OG5_173548 | hydrophilic acylated surface protein a | 1 | 2.1 |
| OG5_130243 | 3,2-trans-enoyl-CoA isomerase, mitochondrial precursor, putative | 1 | 2.09 |
| OG5_126957 | 60S ribosomal protein L13a, putative | 1 | 2.08 |
| OG5_126951 | 60S ribosomal protein L10, putative | 1 | 2.06 |
| OG5_147206 | hypothetical protein, conserved in leishmania | 1 | 2.03 |
| OG5_126776 | PGKC, phosphoglycerate kinase C | 1 | 2.02 |
| OG5_142235 | hypothetical protein, conserved | 1 | 2.02 |
| OG5_127311 | GTPase activator protein, putative ,GTPase activator protein, putative | 1 | 2.02 |
| OG5_142220 | ribonucleoprotein p18, mitochondrial precursor, putative ,ribonucleoprotein p18, | 1 | 2.02 |
| OG5_151771 | hypothetical protein, conserved | 1 | 2.02 |

**Table A15:** OGs with a more than two fold difference between the haploid copy number of *L. adleri* genes in the OG and the assembled number of genes in the OG.

**TableA16:** OGs with a more than two fold difference between the haploid copy number of *L. tarentolae* genes in the OG and the assembled number of genes in the OG. This table is online at https://figshare.com/s/0f4aca2891b891c0b3fc

| .OG | | Number of assembled genes in OG | OG haploid copy number |
|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. major*** | ***L. major*** |
| OG5_184212 | class i nuclease-like protein | 4 | 24.34 |
| OG5_126631 | elongation factor 1-alpha | 7 | 21.6 |
| OG5_137181 | ATG8/AUT7/APG8/PAZ2, putative | 9 | 18.65 |
| OG5_173557 | promastigote surface antigen protein 2 PSA2 | 5 | 18.15 |
| OG5_126588 | heat-shock protein hsp70, putative | 4 | 9.4 |
| OG5_126703 | polyubiquitin, putative | 1 | 9.18 |
| OG5_142729 | ama1 protein, putative | 2 | 4.9 |
| OG5_126923 | elongation factor 2 | 2 | 4.41 |
| OG5_126731 | branched-chain amino acid aminotransferase, putative | 1 | 2.84 |
| OG5_161018 | hypothetical protein, conserved | 1 | 2.16 |
| OG5_236281 | amastin, putative | 1 | 2.11 |
| OG5_138263 | calpain-like cysteine peptidase, putative | 1 | 2.06 |
| OG5_184207 | amastin-like surface protein, putative | 1 | 2.02 |

**Table A17:** OGs with a more than two fold difference between the haploid copy number of *L. major* genes in the OG and the assembled number of genes in the OG.

# Appendix B - Comparative analysis of the genomes of *L. naiffi* and *L. guyanensis* provides insights into the *Viannia* subgenus and the first evidence of *L. naiffi* causing canine leishmaniasis in Colombia

*Some tables in this section were too large to include and so are deposited on Figshare at https://figshare.com/s/0f4aca2891b891c0b3fc. The legends of each of these tables are included in this appendix as well as an indication that the particular table can be found online.*

| **Stage** | ***L. guyanensis* CL085** | ***L. naiffi* CL223** | ***L. braziliensis* M2904** |
|---|---|---|---|
| **Total number of paired-end reads before filter*** | 30,545,938 | 16,262,492 | 52,014,768 |
| **After PCR primers removed: Forward Reads - File 1** | NA | NA | 22,006,185 |
| **After PCR primers removed: Reverse Reads - File 2** | NA | NA | 26,007, 384 (unchanged) |
| **After BLAST and GC filter: Forward Reads - File 1** | 13,626,049 | 8,131,048 | 18,358,010 |
| **After BLAST and GC filter: Reverse Reads - File 2** | 13,725,181 | 8,131,042 | 21,941,567 |
| **Total Reads*** | 27,351,230 | 16,262,090 | 40,299,577 |
| **After Correcting Pairing*** | **26,067,692** | 16,261,930 | **34,592,618** |
| **Singleton reads left over** | 1,506,049 | 160 | 5,706,959 |
| **Paired after trimming** | NA | **13,979,628** | NA |
| **Unpaired after trimming** | NA | 1,068,086 | NA |
| **% Paired-end Reads left after filtering** | 85 | 86 | 67 |

**Table B1:** Read statistics at each stage of quality filtering. * Indicates that number of reads counts both forward and reverse mates of a pair separately. The numbers of reads shown in bold text are the reads that were assembled.

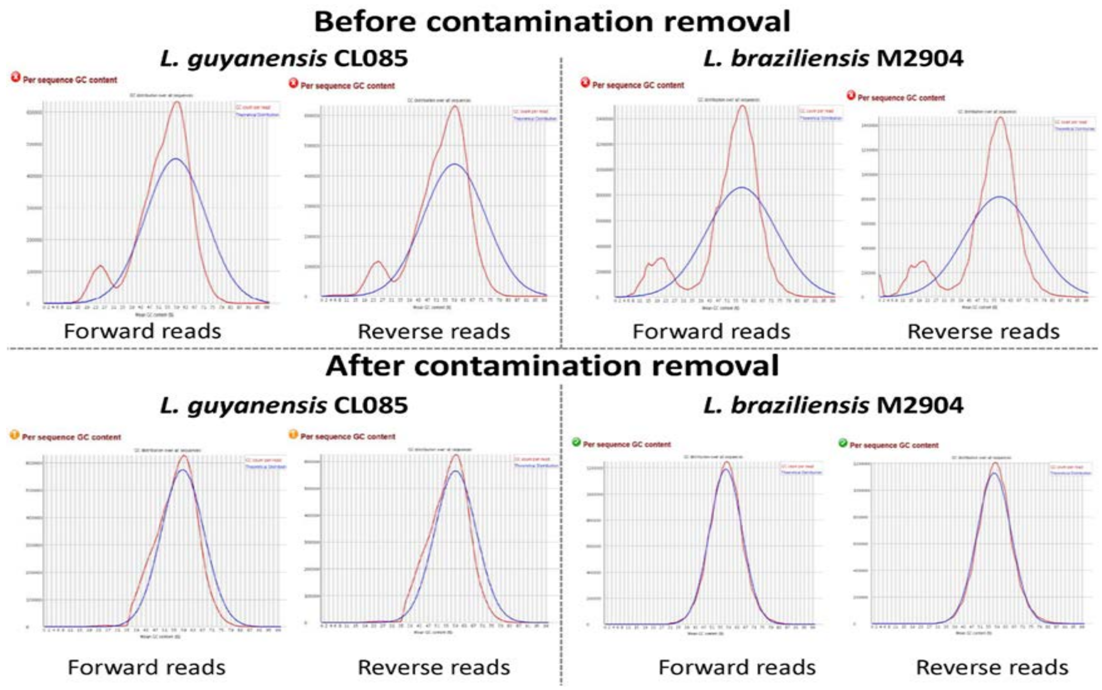


**Figure B1:** GC content plots produced by FASTQC of *L. guyanensis* CL085 and *L. braziliensis* M2904 Illumina short reads before and after contamination removal.

| **Contaminant Name** | **Number of Bin Sequences in assembly** |
|---|---|
| *Niastella koreensis* GR20-10 | 24 |
| *Chitinophaga pinensis* DSM2588 | 4 |
| *Runellas lithyformis* DSM19594 | 2 |
| *Pedobacter heparinus* DSM2366 | 2 |
| *Niabella soli* DSM19437 | 2 |
| *Serratia marcescens* SM39 | 1 |
| *Sphingobacteriaceae bacterium* 27AAV | 1 |
| *Flavobacterium johnsoniae* UW101 | 1 |
| *Flavobacteriaceae bacterium* 3519-10 | 1 |
| *Dyadobacter fermentans* DSM18053 | 1 |
| Mediterranean fruitfly | 1 |
| *Lactobacillus brevis* KB290 | 1 |
| *Paenibacillus polymyxa* M1 | 1 |
| *Dickeya dadantii* Ech586 | 1 |
| Uncultured bacterium | 1 |
| **Total number of non-*Leishmania* sequences:** | **44** |

**Table B2:** Species and number of contaminant bin sequences removed from the *L. guyanensis* CL085 assembly. Bin sequences are scaffolds or contigs that could not be incorporated into chromosomes.

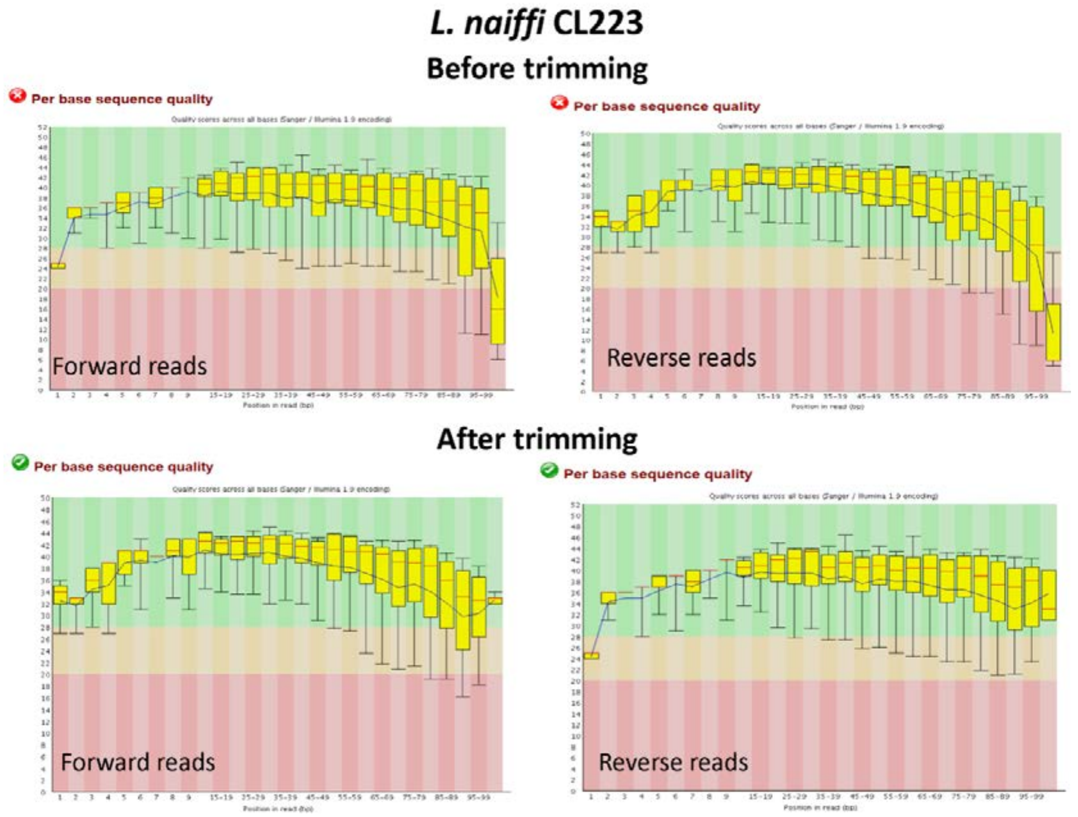


**Figure B2:** Per base sequence quality reported by FASTQC for *L. naiffi* CL223 reads before and after read trimming

| Genome | Stage | Assembly N50 (bp) | Corrected N50 (bp) | Error Free Bases (%) |
|---|---|---|---|---|
| *L. guyanensis* CL085 | Velvet | 10,308 | 10,308 | 84.7 |
| | SSPACE | 64,249 | 46,219 | 93.01 |
| | Gapfiller | 64,018 | 59,438 | 94.17 |
| | iCORN | 64,019 | 59,440 | 94.59 |
| | | | | |
| *L. naiffi* CL223 | Velvet | 5,762 | 5,762 | 69.37 |
| | SSPACE | 26,104 | 23,816 | 85.81 |
| | Gapfiller | 26,032 | 24,271 | 86.09 |
| | iCORN | 26,029 | 24,369 | 86.14 |
| | | | | |
| *L. braziliensis* M2904 control | Velvet | 5,128 | 5,128 | 76.73 |
| | SSPACE | 21,022 | 20,260 | 86.86 |
| | Gapfiller | 21,034 | 20,679 | 87.03 |
| | iCORN | 21,033 | 20,574 | 85.9 |

**Table B3:** N50, Corrected N50 and the percentage of error free bases as determined by REAPR for assembly and assembly improvement stages of the pipeline

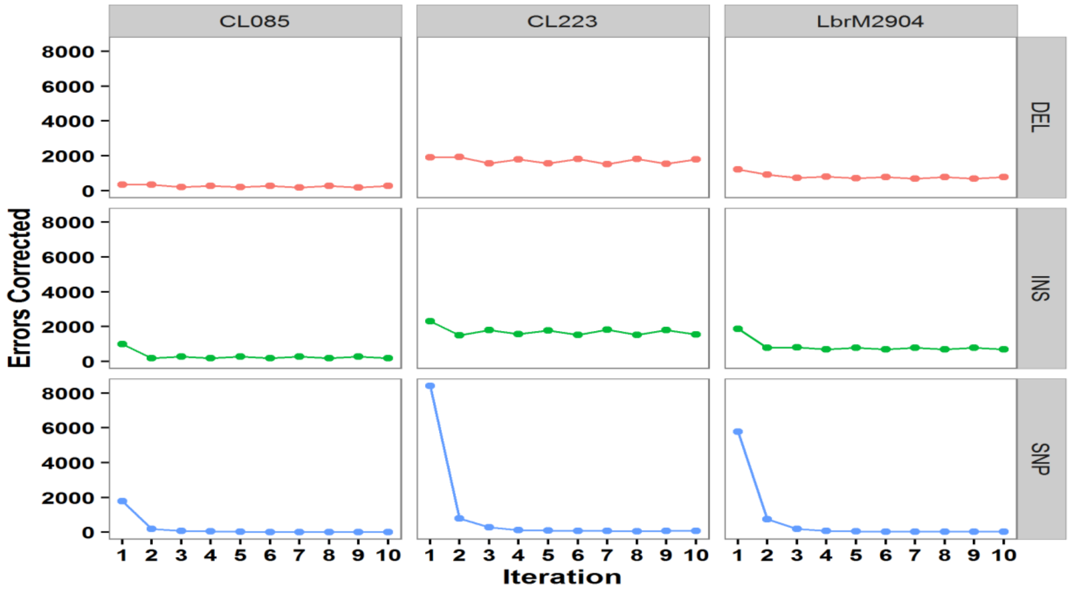


**Figure B3:** Number of deletions (DEL), short insertions (INS) and single nucleotide polymorphisms (SNPs) errors corrected at each iteration of iCORN for *L. guyanensis* CL085 (CL085), *L. naiffi* CL223 (CL223) and the control *L. braziliensis* genome (LbrM2904).

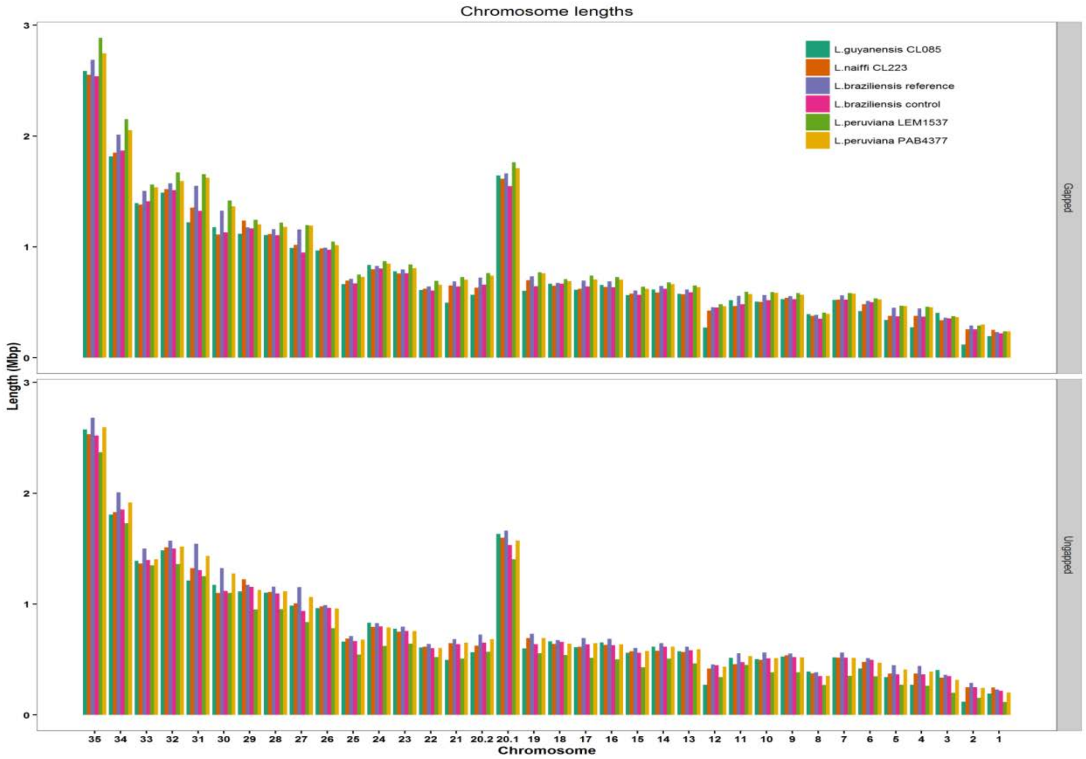


**Figure B4:** Chromosome lengths of *L. guyanesnis* CL085, *L. naiffi* CL223 and the *L. braziliensis* M2904 control compared with the *L. braziliensis* M2904 reference genome, *L. peruviana* PAB-4377, *L. peruviana* LEM1537 and *L. panamensis* PSC-1. Lengths are examined both including gaps in the chromosome length (top) and excluding gaps from chromosome lengths (bottom).

| | *L. braziliensis* M2904 | *L. braziliensis* control | *L. guyanensis* CL085 | *L. naiffi* CL223 | *L. panamensis* PSC-1 | *L. peruviana* PAB-4377 | *L. peruviana* LEM1537 |
|---|---|---|---|---|---|---|---|
| **Number of chromosomes** | 35 | 35 | 35 | 35 | 35 | 35 | 35 |
| **All genes** | 8,620 | 8,161 | 8,376 | 8,262 | 8,048 | NA | NA |
| **Protein coding genes** | 8,357 | 8,001 | 8,230 | 8,104 | 7,748 | NA | NA |
| **Genes on chromosomes** | 8,432 | 7,873 | 7,757 | 7,952 | 8,048 | NA | NA |
| **Genes on bin contigs** | 188 | 288 | 619 | 310 | NA | NA | NA |
| **Number of gaps** | 919 | 3,352 | 1,557 | 3,853 | 553 | 16,598 | 27,836 |
| **N content (%)** | 0.29 | 0.99 | 0.45 | 1.07 | 2.28 | 8.49 | 26.37 |
| **Chromosomes total length (bp)** | 31,238,104 | 28,985,156 | 28,274,008 | 29,179,723 | 30,688,794 | 32,026,188 | 33,012,161 |
| **Bin sequence total length (bp)** | 850,747 | 1,024,497 | 2,740,314 | 1,161,372 | NA | 881,593 | 878,039 |
| **Genome length (bp)** | 32,088,851 | 30,009,653 | 31,014,322 | 30,341,095 | 30,688,794 | 32,907,781 | 33,890,200 |
| **GC content (%)** | 57.77 | 57.6 | 57.34 | 57.53 | 57.56 | 57.15 | 55.87 |
| **Median Coverage** | 75 | 74 | 56 | 36 | 143 | 75 | 42 |

**Table B4:** Summary of *L. guyanensis*, *L. naiffi* and *L. braziliensis* control genomes assembled here compared with other genomes in the *Viannia* subgenus.

| Genome | Gap edges | Sequence ends | Repetitive sequence | Total masked |
|---|---|---|---|---|
| *L. guyanensis* CL085 | 354,952 | 162,000 | 2,066,621 | 2,583,573 |
| *L. naiffi* CL223 | 995,942 | 169,800 | 2,091,587 | 3,257,329 |
| *L. braziliensis* M2904 | 196,214 | 22,200 | 2,413,094 | 2,631,508 |
| *L. braziliensis* M2904 control | 887,600 | 133,200 | 2,023,885 | 3,044,685 |
| *L. panamensis* PSC-1 | 110,600 | 21,000 | 2,028,434 | 2,160,034 |
| *L. peruviana* LEM-1537 | 5,840,400 | 22,200 | 1,769,476 | 7,632,076 |
| *L. peruviana* PAB-4377 | 3,727,200 | 22,200 | 2,062,909 | 5,812,309 |

**Table B5:** Masking information for SNP filtering

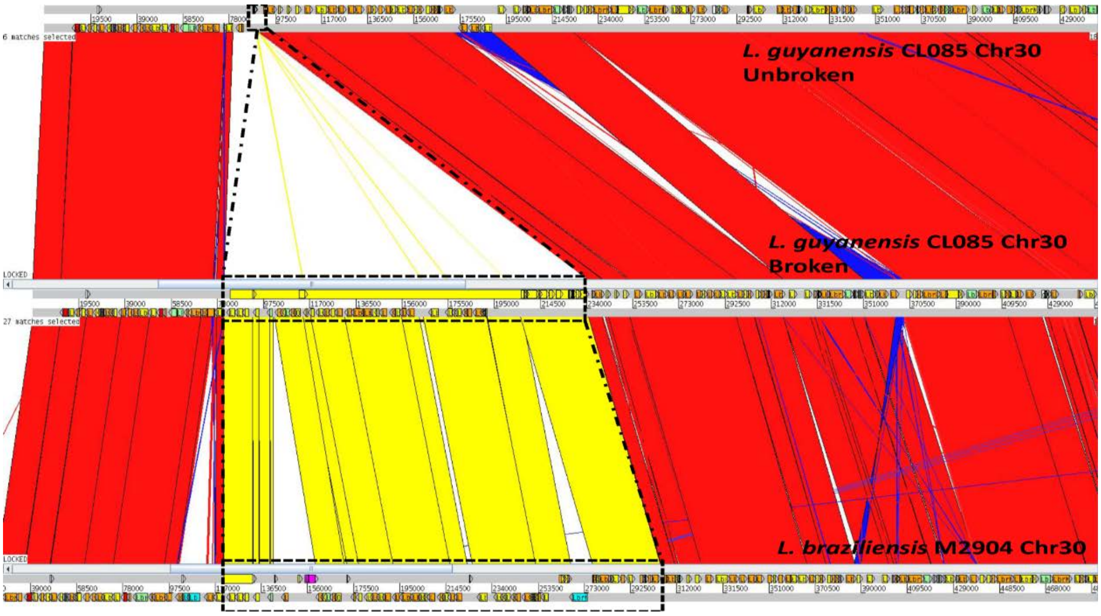

6 matches selected
L. guyanensis CL085 Chr30
Unbroken
L. guyanensis CL085 Chr30
Broken
LOCKED
27 matches selected
L. braziliensis M2904 Chr30
LOCKED

**Figure B5:** Use of REAPR before contiguation of *L. guyanensis* CL085 scaffolds with ABACAS incorporated additional sequence into chromosome 30 (middle). This additional sequence shares synteny with a locus on chromosome 30 of *L. braziliensis* M2904.

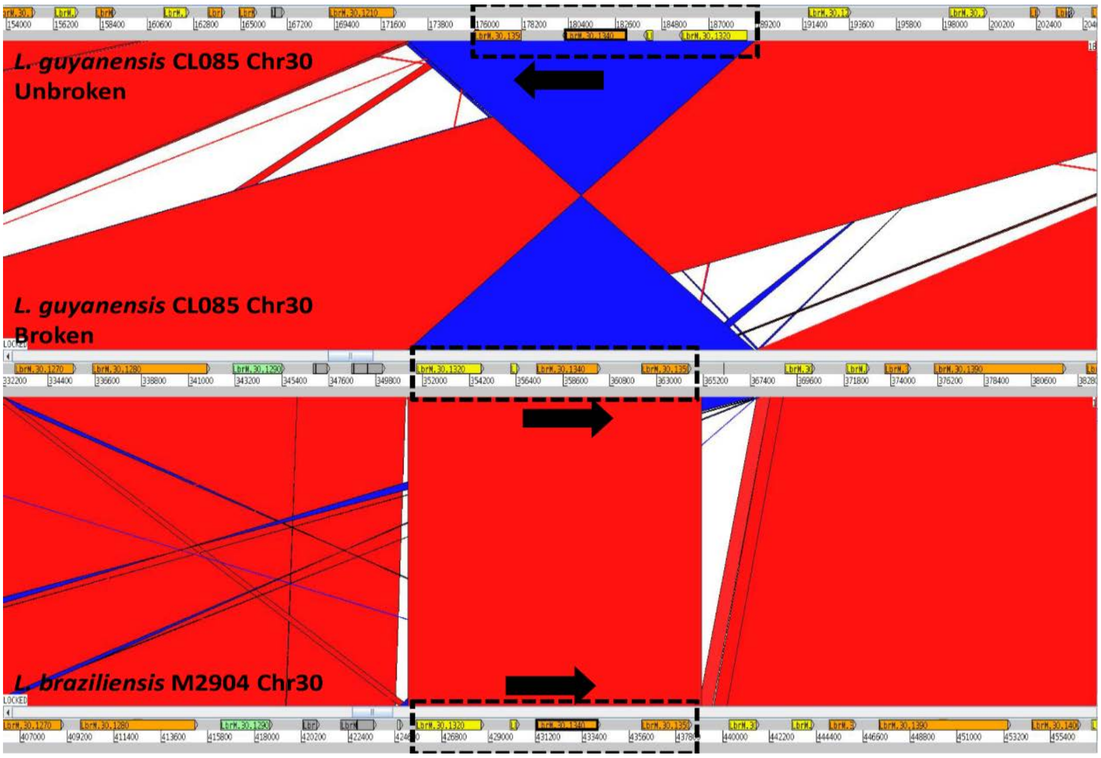

L. guyanensis CL085 Chr30
Unbroken
L. guyanensis CL085 Chr30
Broken
LOCKED
L. braziliensis M2904 Chr30
LOCKED

**Figure B6:** Correction of inversion on chromosome 30 of *L. guyanensis* CL085, as seen using a comparison with chromosome 30 of *L. braziliensis* M2904 (bottom) by using REAPR corrected scaffolds for contiguation (middle) instead of un-corrected scaffolds (top). Black arrows indicate the direction of genes on the locus.

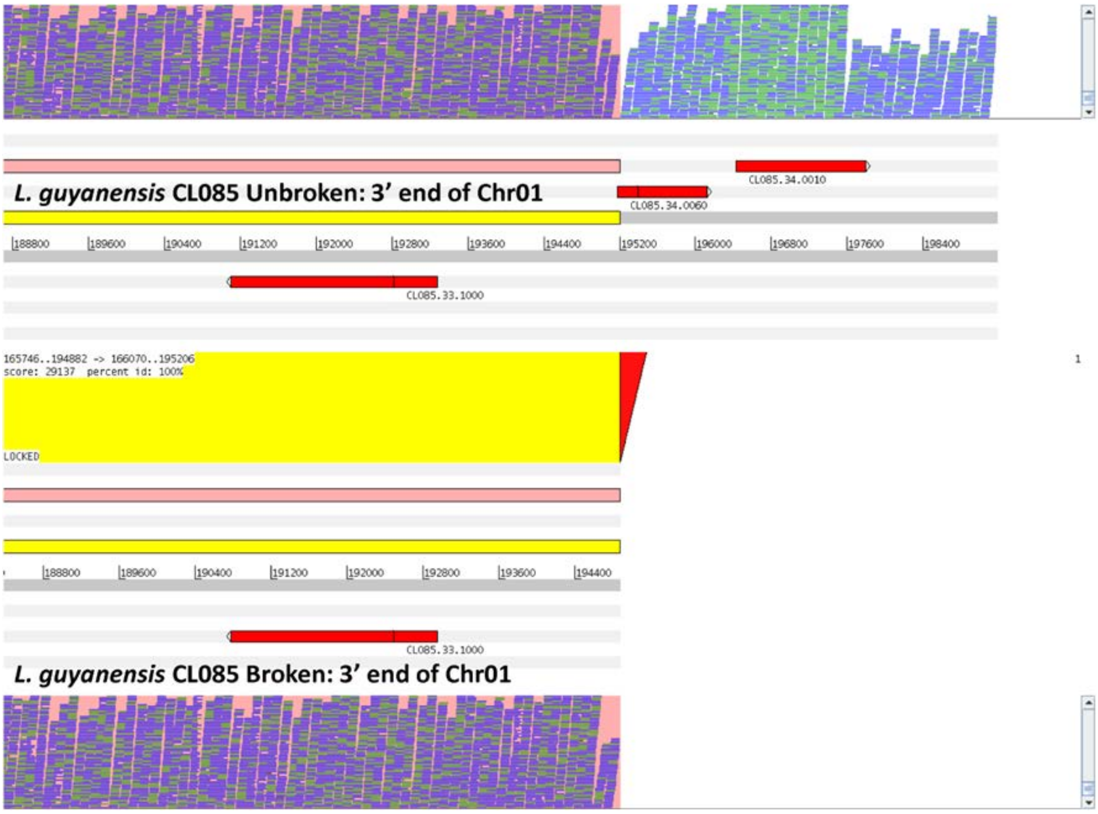


**Figure B7:** Use of REAPR before contiguation of *L. guyanensis* CL085 scaffolds with ABACAS removed a section (see top sequence) that did not have homology with chromosome 1 of *L. braziliensis* M2904.

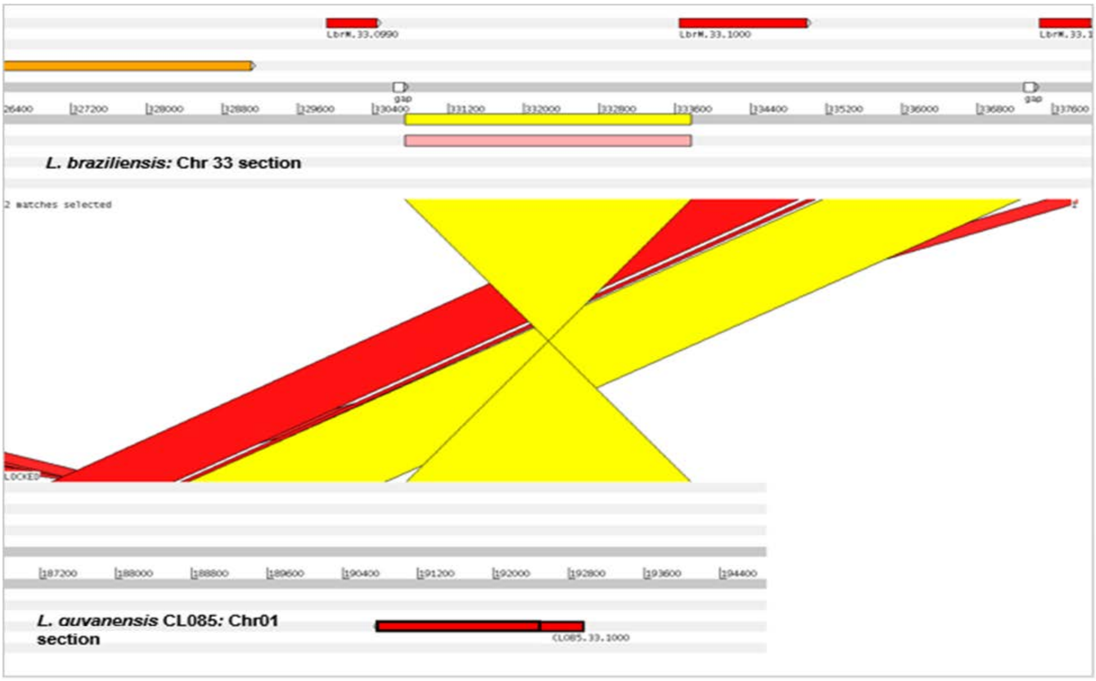


**Figure B8:** Section at the end of chromosome 1 of *L. guyanensis* CL085 with homology to chromosome 33 of *L. braziliensis* M2904

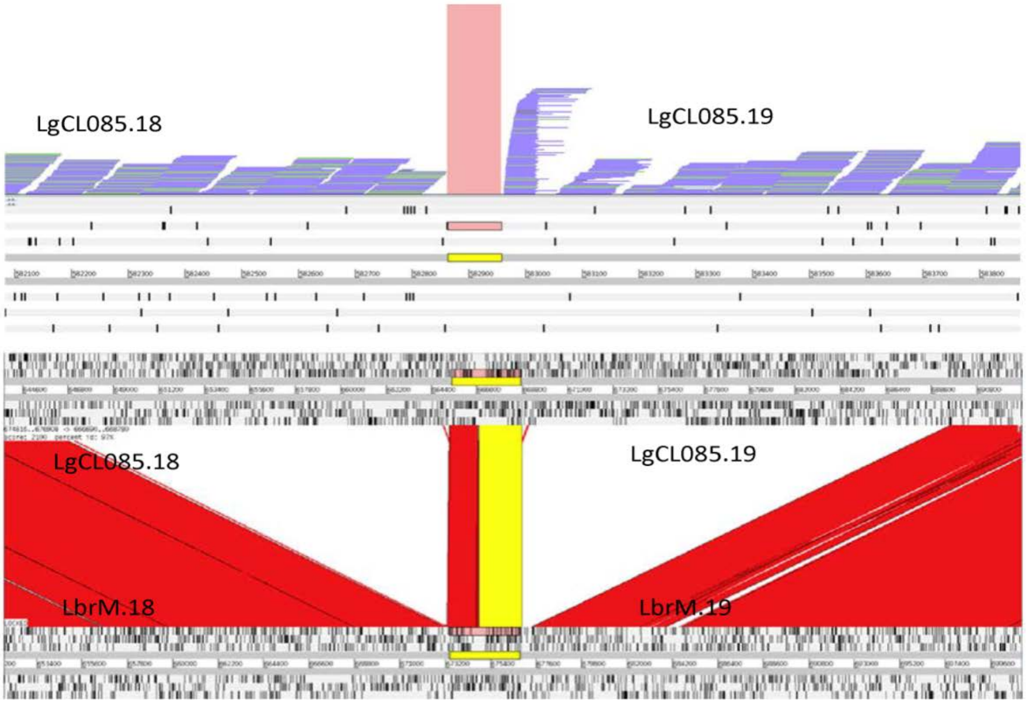


**Figure B9:** Alignment of *L. guyanensis* CL085 chromosome 18 and 19 (joined) with *L. braziliensis* M2904 chromosome 18 and 19. A gap in coverage is seen at the end of chromosome 18 of *L. guyanensis* CL085 where it was attached to a chromosome 19 and so the chromosome 18 was broken at this point.

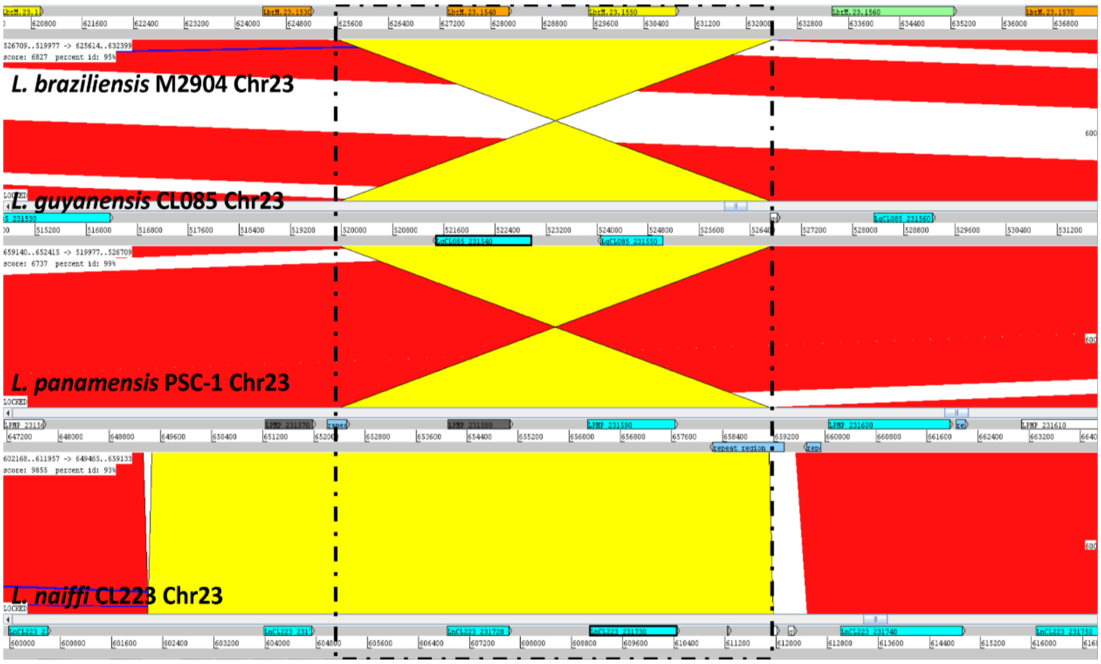

L. braziliensis M2904 Chr23
L. guyanensis CL085 Chr23
L. panamensis PSC-1 Chr23
L. naiffi CL223 Chr23

**Figure B10:** An inversion on chromosome 23 of *L. guyanensis* CL085 which is not present in either *L. braziliensis* M2904, *L. panamensis* PSC-1 or *L. naiffi* CL223.

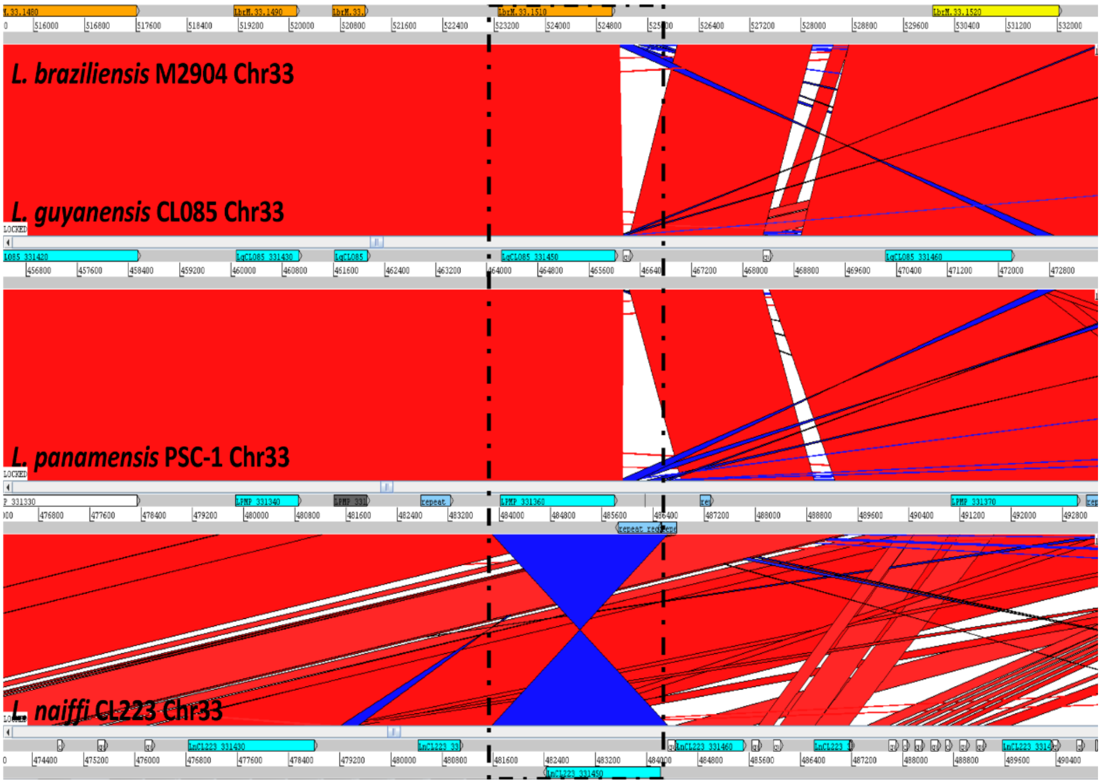

L. braziliensis M2904 Chr33
L. guyanensis CL085 Chr33
L. panamensis PSC-1 Chr33
L. naiffi CL223 Chr33

**Figure B11:** Putative inversion on chromosome 33 of *L. naiffi* CL223 which is not present in *L. braziliensis* M2904, *L. panamensis* PSC-1 or *L. guyanensis* CL085.

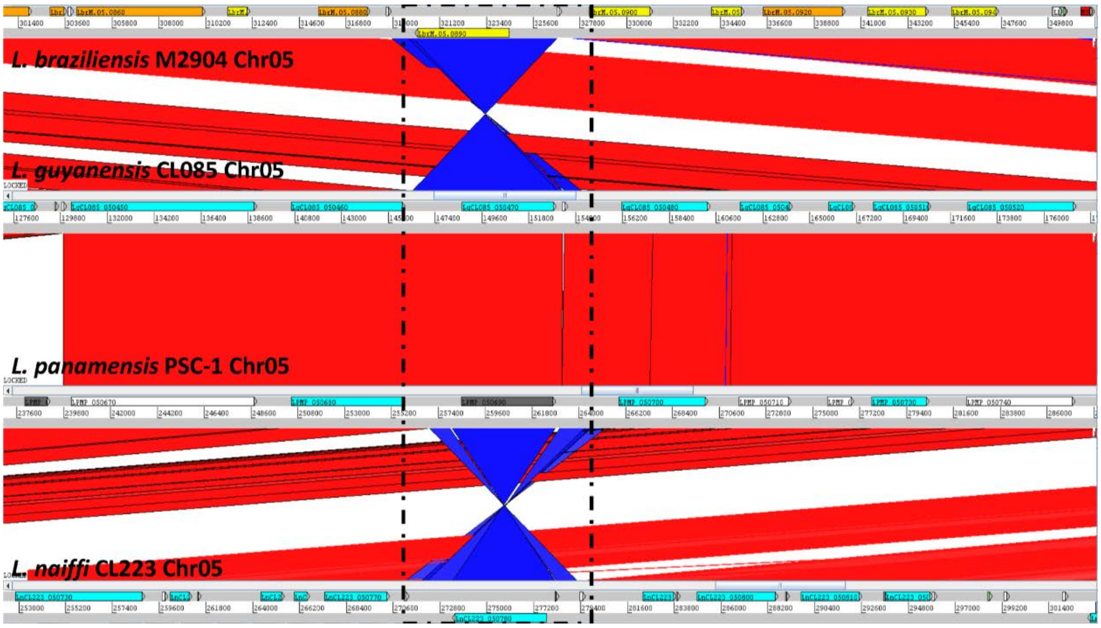

L. braziliensis M2904 Chr05
L. guyanensis CL085 Chr05
L. panamensis PSC-1 Chr05
L. naiffi CL223 Chr05

**Figure B12:** Inversion on chromosome 5 of *L. naiffi* CL223 (bottom sequence) and *L. braziliensis* M2904 (top sequence) that is not present on chromosome 5 of *L. guyanensis* CL085 (2nd sequence) or *L. panamensis* PSC-1 (3rd sequence).

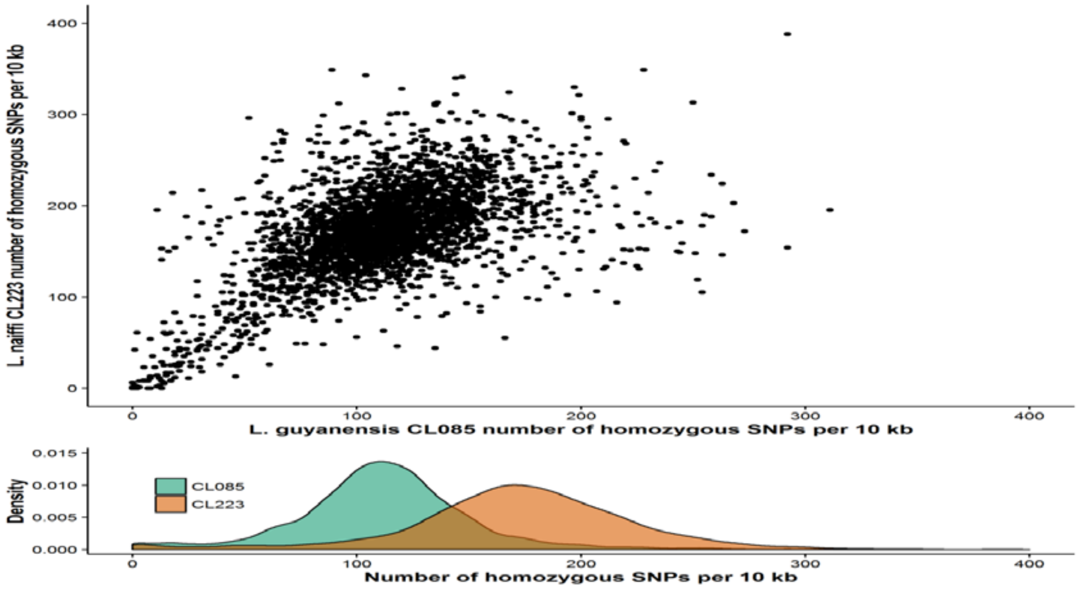


**Figure B13:** Divergence of *L. naiffi* CL223 (green in bottom part) and *L. guyanensis* CL085 (orange in bottom part) from *L. braziliensis* M2904 measured by counting the number of homozygous SNPs in 10 kb non-overlapping blocks for *L. naiffi* CL223 reads mapped to *L. braziliensis* M2904 and *L. guyanensis* CL085 reads mapped to *L. braziliensis* M2904.

| | | **Coverage** | | |
|---|---|---|---|---|
| **Species** | **% Mapped** | **Mean** | **Median** | **Standard deviation** |
| **Self-mapped** | | | | |
| *L. panamensis*PSC-1 | 85.88 | 157.40 | 143 | 408.16 |
| *L. naiffi* CL223 | 98.69 | 42.58 | 36 | 77.92 |
| *L. guyanensis* CL085 | 84.08 | 65.61 | 56 | 124.77 |
| *L. peruviana* PAB-4377 | 96.02 | 83.89 | 75 | 613.92 |
| *L. peruviana* LEM1537 | 97.5 | 51.55 | 42 | 95.48 |
| *L. braziliensis* M2904 | 99 | 77.98 | 74 | 105.80 |
| *L. braziliensis* M2904 control | 99.15 | 82.76 | 75 | 100.87 |
| **Mapped to *L. braziliensis* M2904** | | | | |
| *L. guyanensis* CL085 | 81.64 | 60.62 | 54 | 97.42 |
| *L. guyanensis* M4147 | 90.89 | 32.73 | 29 | 226.48 |
| *L. panamensis* PSC-1 | 90.01 | 155.91 | 141 | 1324.16 |
| *L. panamensis* WR120 | 92.28 | 24.49 | 22 | 171.17 |
| *L. shawi* M8408 | 86.84 | 25.51 | 22 | 319.81 |
| *L. naiffi* CL223 | 98.49 | 39.45 | 35 | 54.94 |
| *L. naiffi* M5533 | 87.61 | 47.76 | 43 | 420.74 |
| *L. lainsoni* M6426 | 78.59 | 19.87 | 17 | 312.06 |
| *L. peruviana* PAB-4377 | 95.26 | 85.45 | 76 | 855.65 |
| *L. peruviana* LEM1537 | 98.66 | 55.29 | 47 | 86.21 |
| **Mapped to *L. panamensis* PSC-1** | | | | |
| *L. panamensis* WR120 | 89.06 | 24.97 | 23 | 39.46 |

**Table B6:** Percentage reads mapped and coverage statistics for each mapping. All statistics were calculated from duplicate removed BAM file.

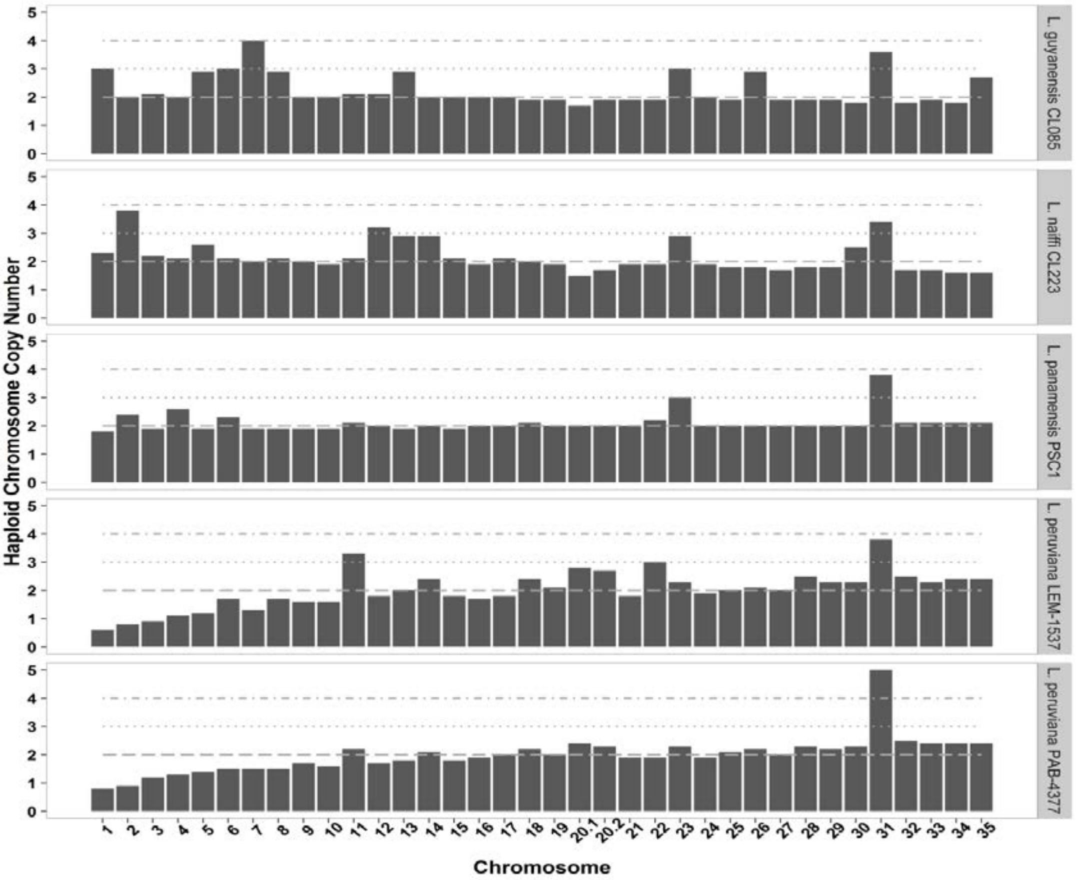


**Figure B14:** Chromosome copy number predicted using read-depth coverage of reads mapped to *L. braziliensis* M2904 for *L. guyanensis* CL085, *L. naiffi* CL223, *L. panamensis* PSC-1, *L. peruviana* LEM-1537 and *L. peruviana* PAB-4377.

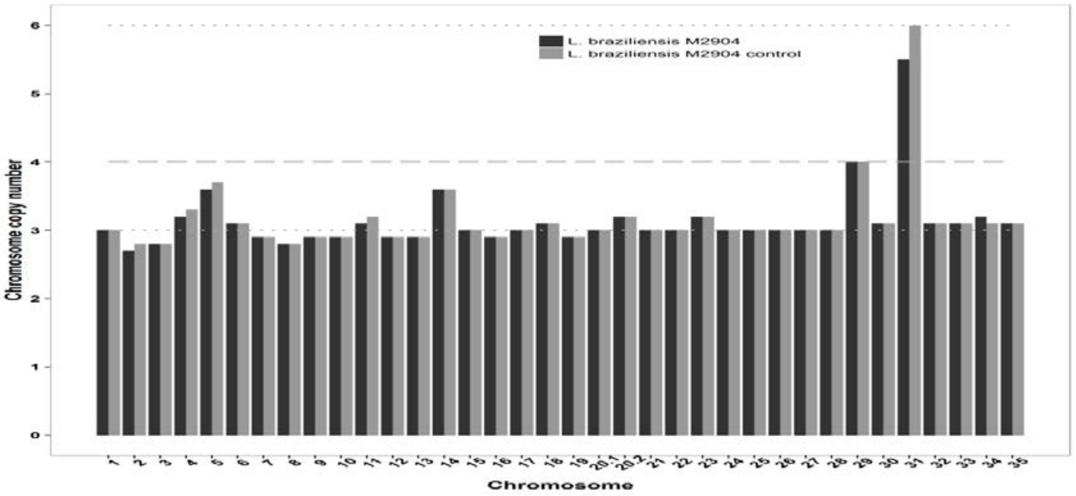


**Figure B15:** Chromosome copy number of the control *L. braziliensis* M2904 assembly which replicates the results of [177] for *L. braziliensis* M2904.

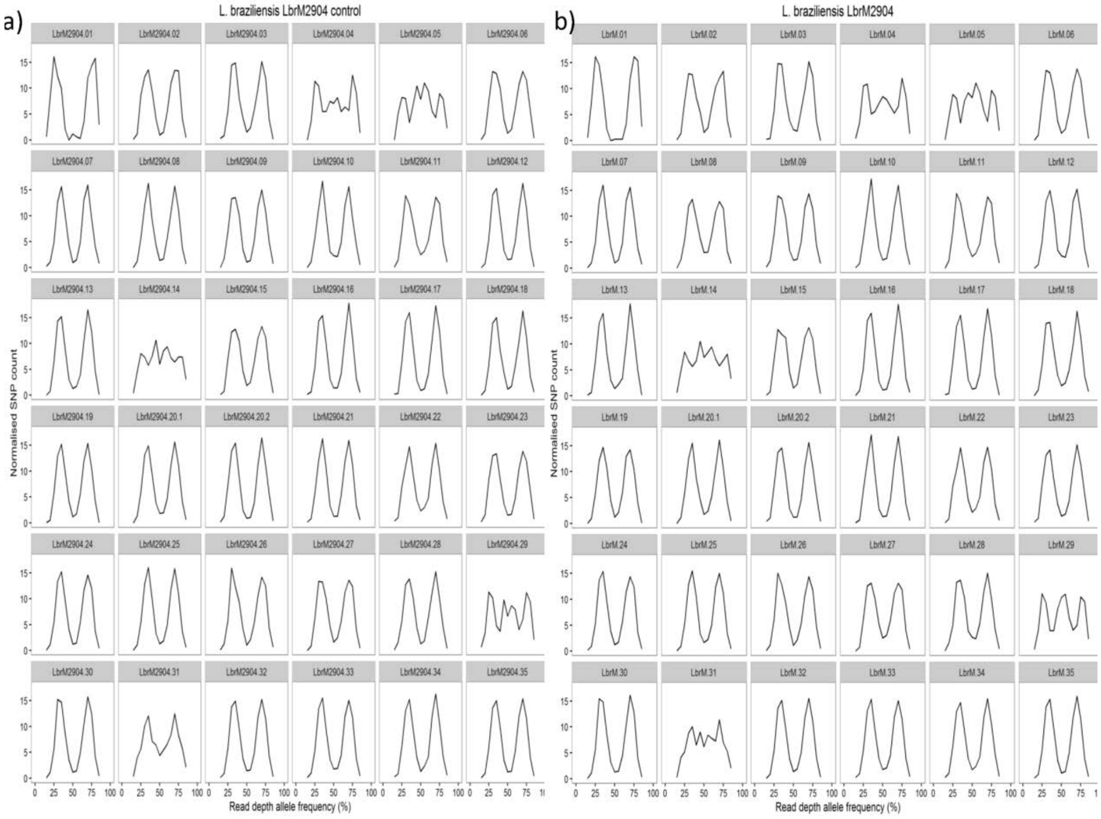


**Figure B16:** Read depth allele frequency distributions (RDAF) of heterozygous SNPs called from self-mapped reads for a) each chromosome of *L. braziliensis* M2904 control genome compared with b) each chromosome of the reference *L. braziliensis* M2904 genome showing that both produce the same distributions for every chromosome.

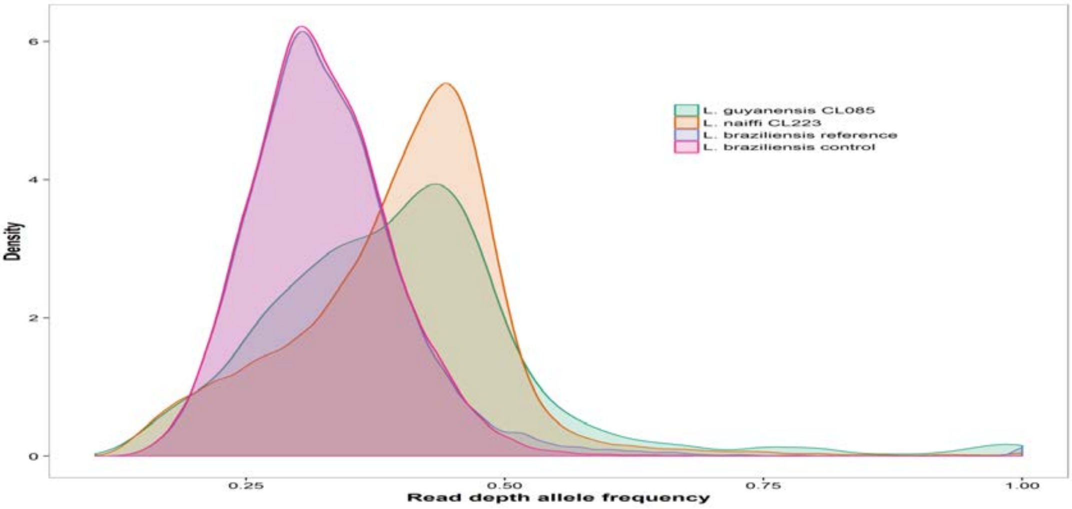


**Figure B17:** Read depth allele frequencies of self-mapped SNPs for *L. naiffi*, *L. guyanensis* and the *L. braziliensis* reference and control genomes showing that *L. naiffi* and *L. guyanensis* are mainly disomic and *L. braziliensis* is predominately trisomic.

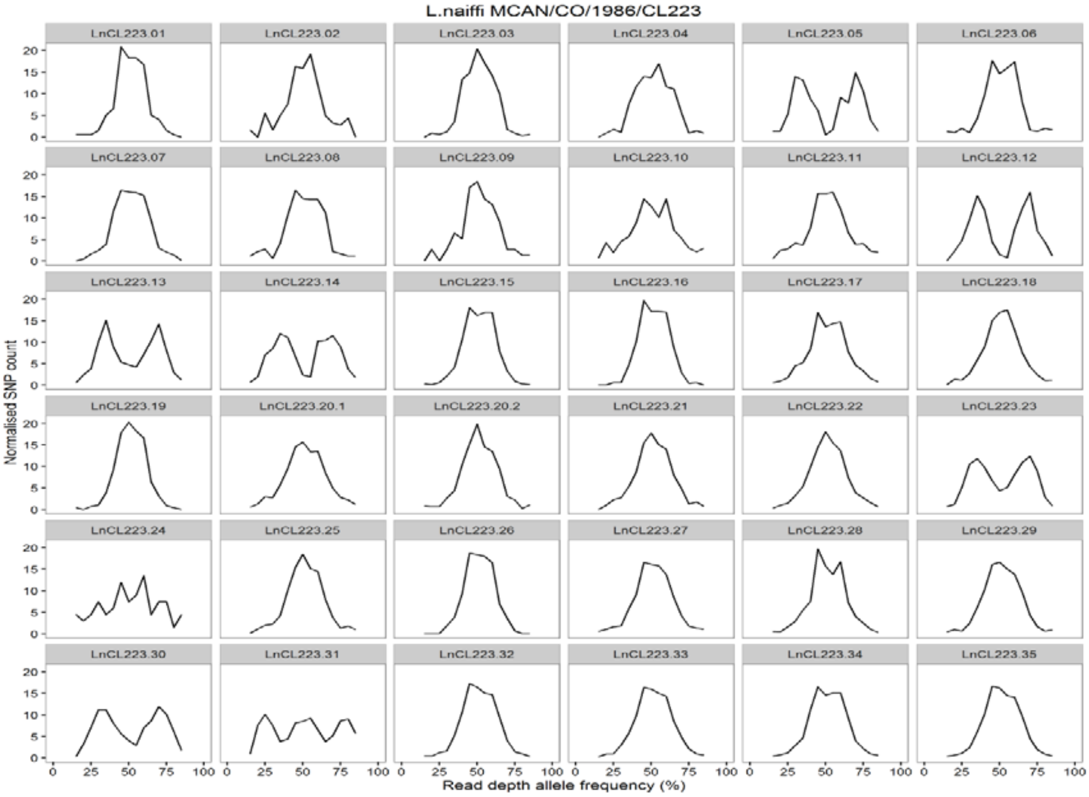


**Figure B18:** Read depth allele frequency distributions (RDAF) of each chromosome of *L. naiffi* CL223 determined using heterozygous SNPs called from self-mapped reads.

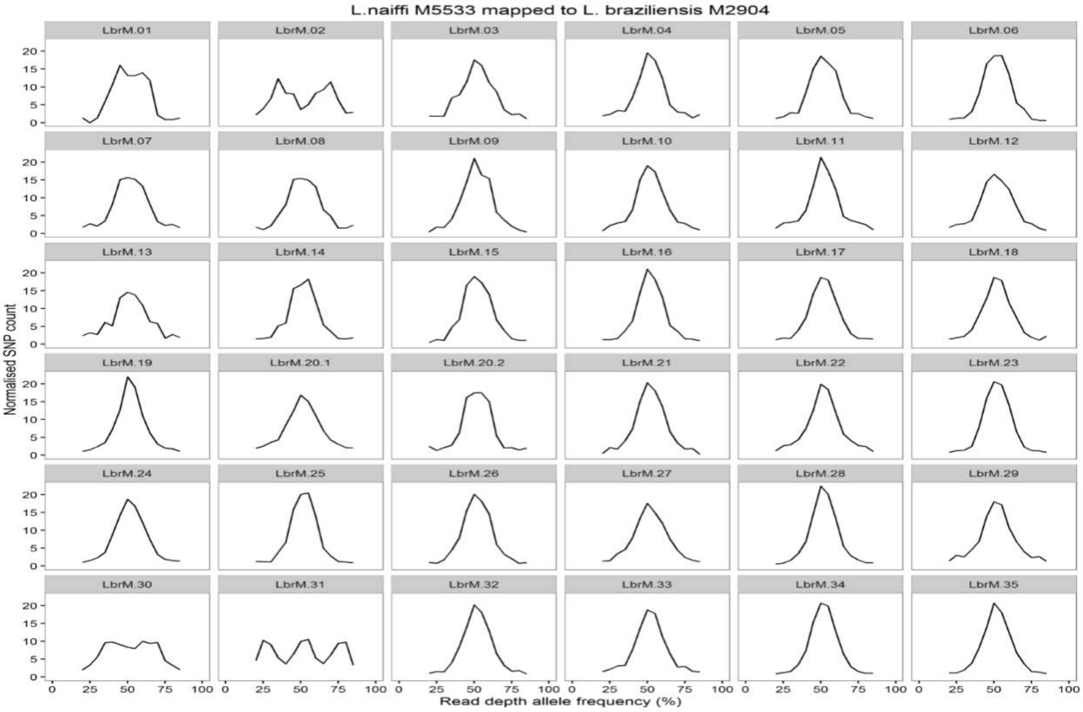


**Figure B19:** Read depth allele frequency distributions (RDAF) of each chromosome of *L. naiffi* M5533 determined using heterozygous SNPs from reads mapped to *L. braziliensis* M2904.

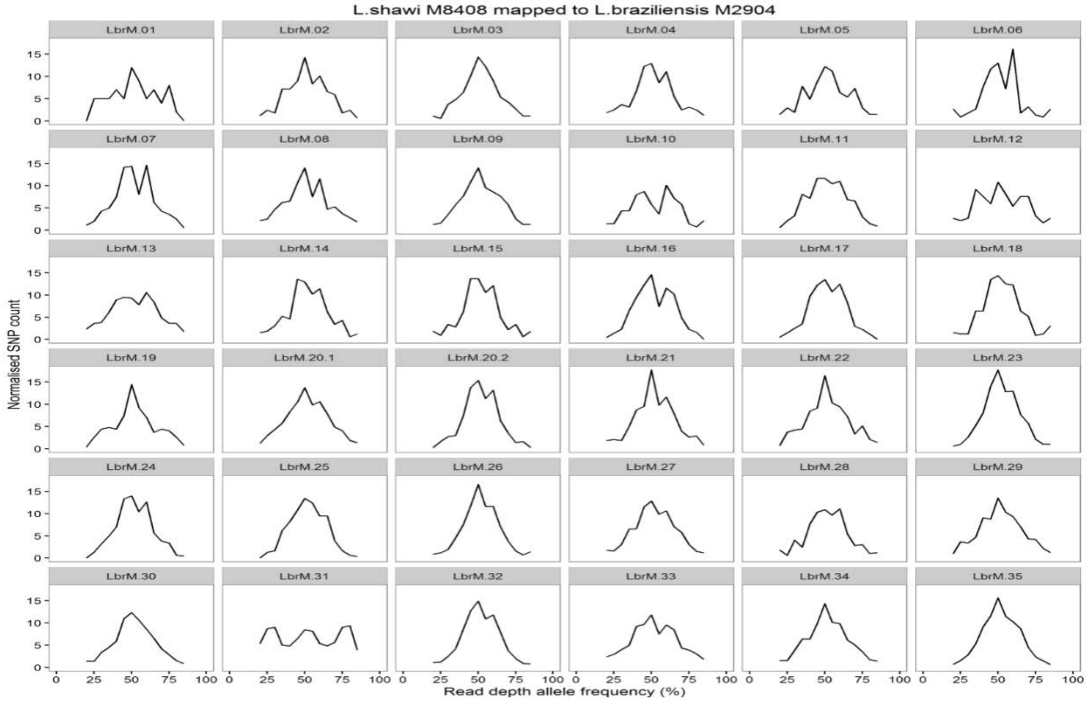


**Figure B20:** Read depth allele frequency distributions (RDAF) of each chromosome of *L. shawi* M8408 determined using heterozygous SNPs from reads mapped to *L. braziliensis* M2904.

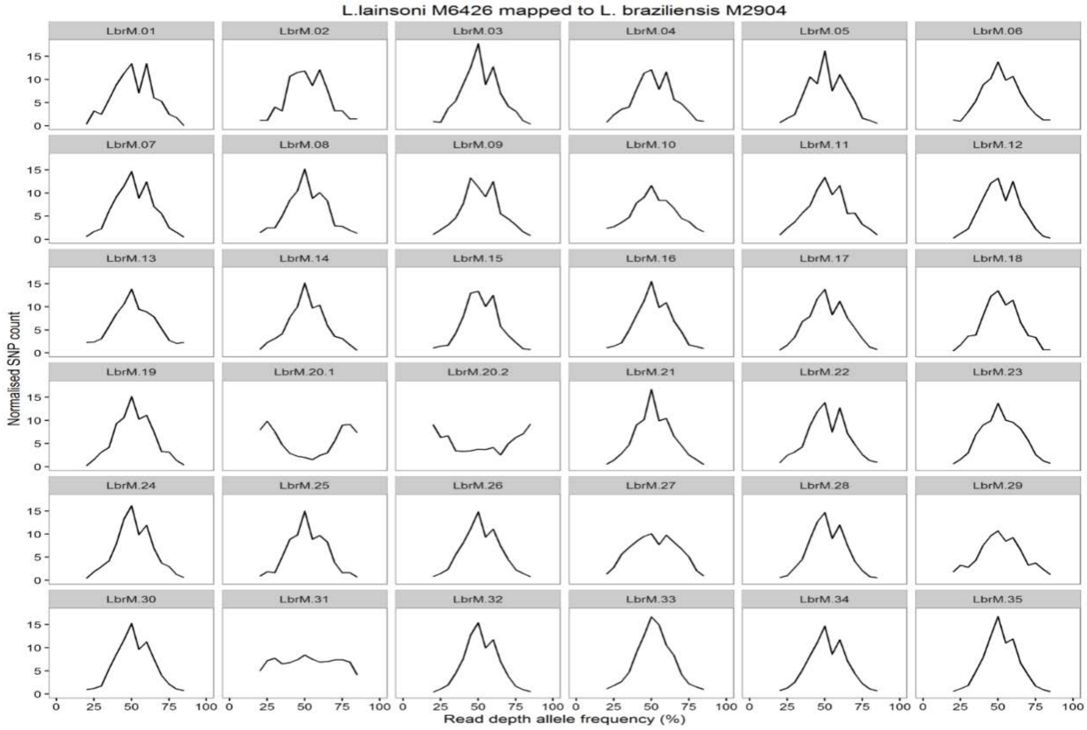


**Figure B21:** Read depth allele frequency distributions (RDAF) of each chromosome of *L. lainsoni* M6426 determined using heterozygous SNPs from reads mapped to *L. braziliensis* M2904.

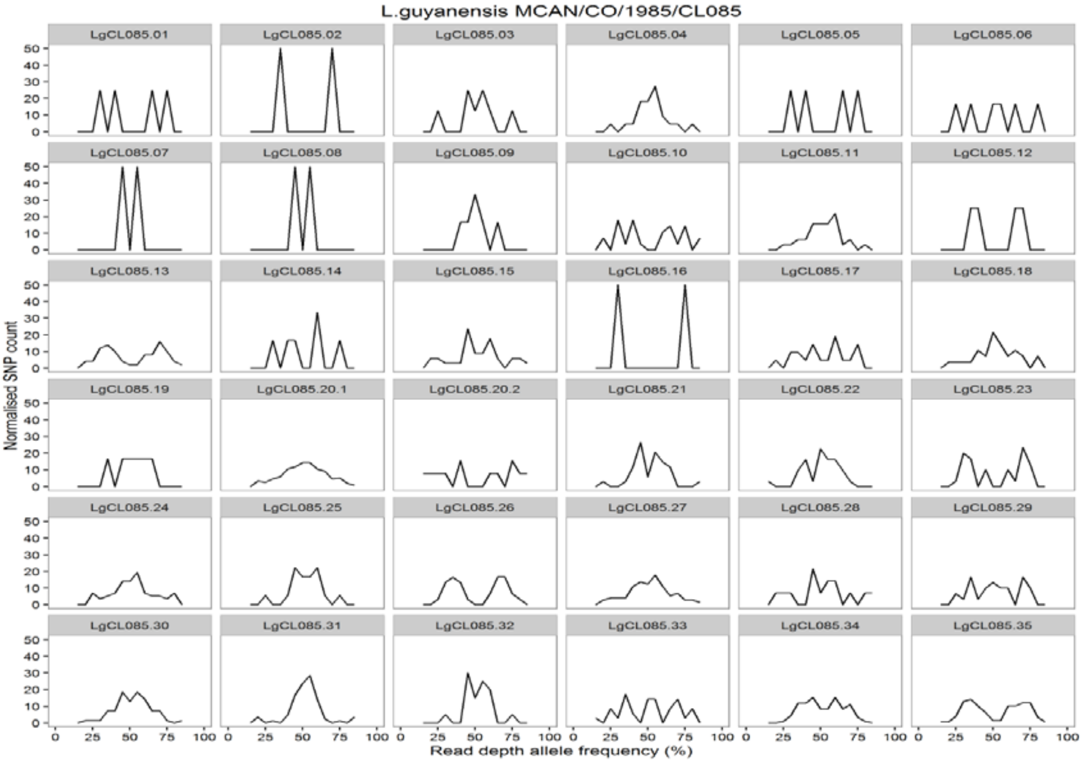


**Figure B22:** Read depth allele frequency distributions (RDAF) of each chromosome of *L. guyanensis* CL085 determined using heterozygous SNPs from self-mapped reads. Most chromosomes have uninformative RDAF plots here due to low number of heterozygous SNPs.

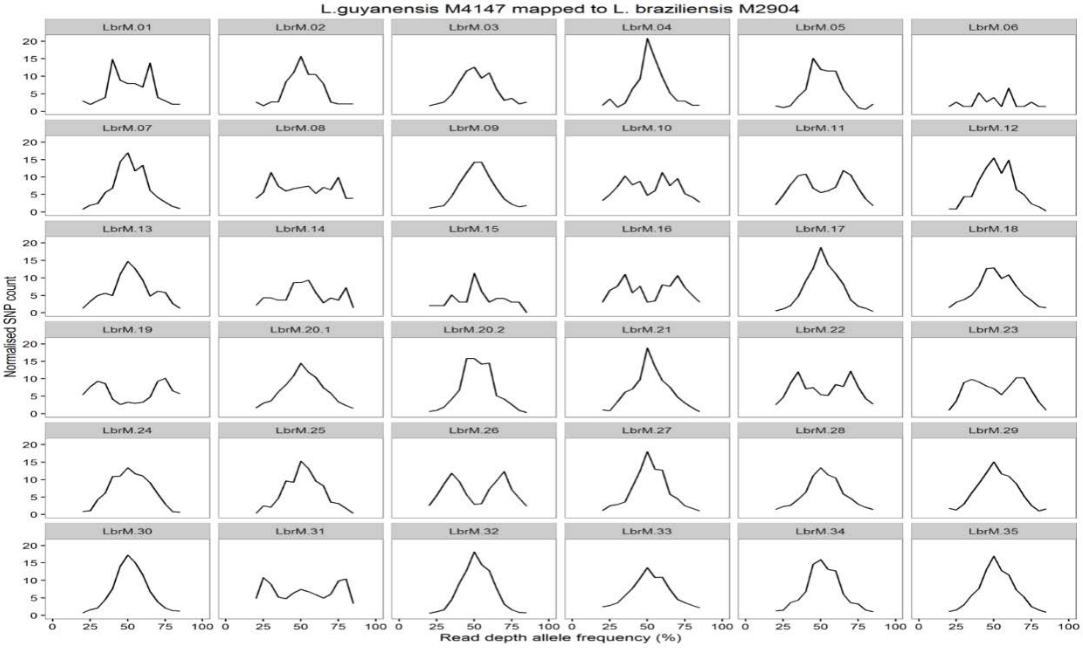


**Figure B23:** Read depth allele frequency distributions (RDAF) of each chromosome of *L. guyanensis* M4147 determined using heterozygous SNPs from reads mapped to *L. braziliensis* M2904.

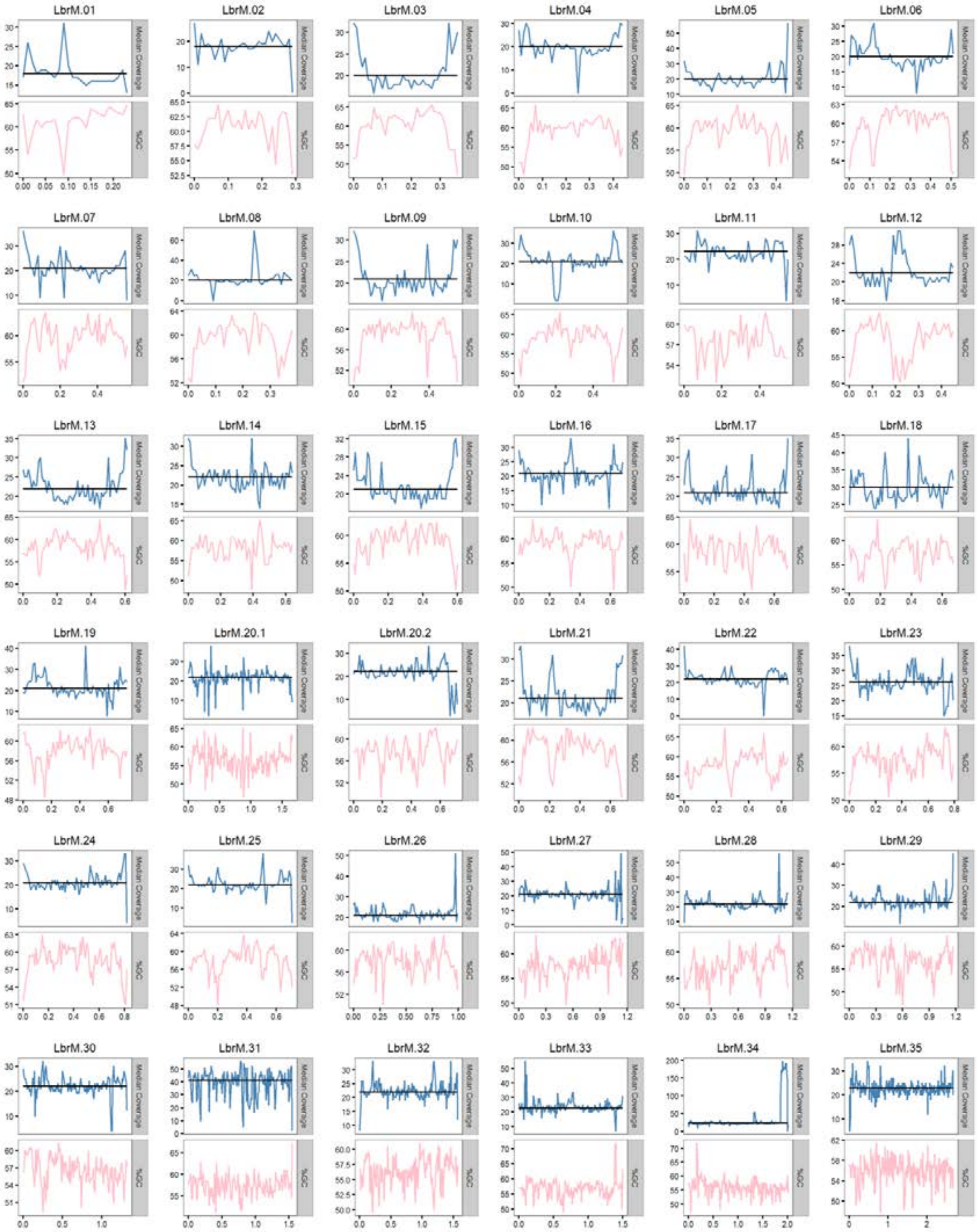


**Figure B24:** Median coverage (blue) in 10 kb blocks across each chromosome for *L. shawi* M8408 reads mapped to *L. braziliensis* M2904. The black horizontal line in each plot denotes the median chromosomal coverage and the pink line indicates %GC content measured in 10 kb blocks. Note the increase in coverage at the 3’ end of chromosome 34 (second last plot on the bottom row) indicating amplification of inverted repeats at that locus to form a linear minichromosome.

**Table B7:** Copy number of 10 kb non-overlapping loci in each genome. Gray shading indicates the loci that a minichromosome is amplified from on *L. panamensis* PSC-1 and *L. shawi* M8408. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

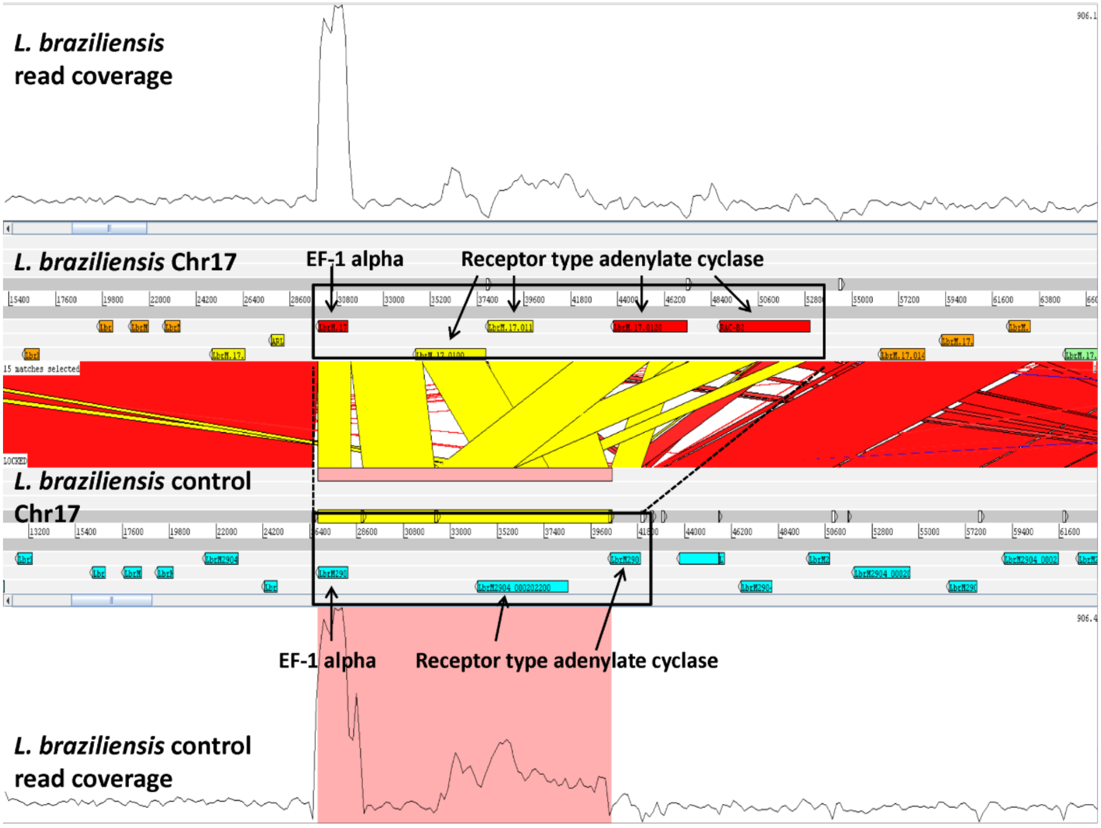


**Figure B25:** Amplified locus in the *L. braziliensis* M2904 control genome (bottom half) that is completely assembled in the original *L. braziliensis* M2904 genome with the exception of the elongation factor 1-alpha gene (EF-1 alpha) which is not fully assembled in either genome due to its very high copy number.

**Table B8:** Amplifications of loci > 10 kb on *L. guyanensis* CL085 and *L. naiffi* CL223. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B9:** Orthologous groups (OGs) in *Leishmania*. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

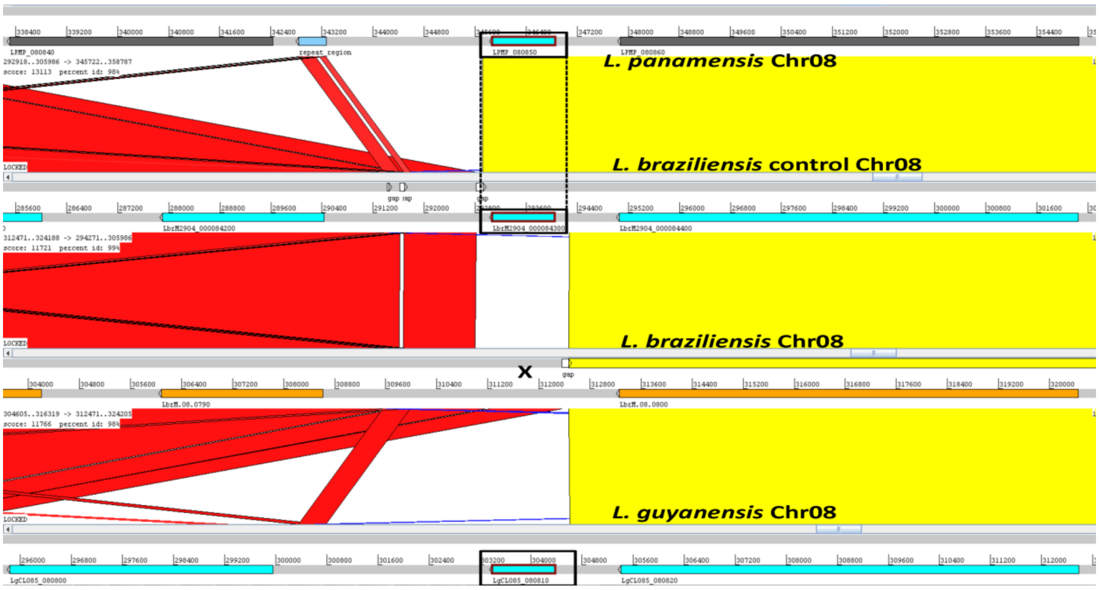


**Figure B26:** Incomplete assembly of a locus on chromosome 8 of *L. braziliensis* M2904, marked with an 'X' here, that is present on the control *L. braziliensis genome*, the *L. panamensis* PSC-1 genome (LPMP_080850 on *L. panamensis* Chr 8), *L. guyanensis* CL085 (LgCL085_080810) and *L. naiffi* (not shown here).

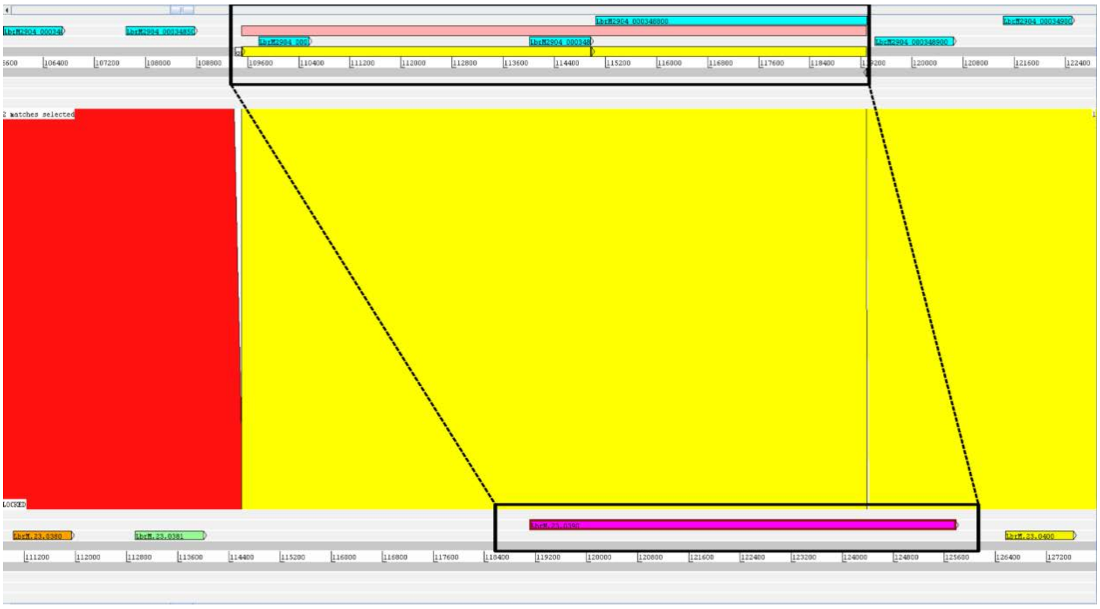


**Figure B27:** Incorrect models of DCL1 gene on chromosome 23 of the *L. braziliensis* M2904 control genome (top) due to gaps. Only one gene model is present at the homologous locus on chromosome 23 of *L. braziliensis* M2904

**Table B10:** Genes in orthologous groups (OGs) that are in the *L. braziliensis* M2904 reference annotation but not the control *L. braziliensis* M2904 genome annotation. The *L. braziliensis* reftest column indicates the number of genes in each OG annotated on the *L. braziliensis* M2904 reference genome by the Companion pipeline in a test run. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

| OG | | Number of assembled genes in OG | | | Assembled genes in OG | OG haploid copy number |
|---|---|---|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. braziliensis control*** | ***L. braziliensis*** | ***L. braziliensis reftest*** | ***L. braziliensis control*** | ***L braziliensis* control** |
| OG5_126627 | hypothetical protein | 3 | 0 | 3 | LbrM2904_000804300,LbrM2904_000616100, LbrM2904_000337500 | 3.91 |
| OG5_127141 | serine/threonine protein phosphatase, putative | 1 | 0 | 0 | LbrM2904_000402700 | 1.21 |
| OG5_127356 | signal peptidase type I, putative, serine peptidase, Clan SF, Family S26A | 1 | 0 | 0 | LbrM2904_000798200 | 0.85 |
| OG5_127596 | aminophospholipid translocase, putative , | 1 | 0 | 0 | LbrM2904_000278200 | 0.81 |
| OG5_127804 | ATP-dependent DEAD-box RNA helicase, putative , putative | 1 | 0 | 0 | LbrM2904_000671100 | 0.95 |
| OG5_128007 | hypothetical protein, conserved , | 2 | 0 | 0 | LbrM2904_000131200,LbrM2904_000131300 | 1.71 |
| OG5_128035 | translation initiation factor eif-2b beta subunit, putative,eIF-2B GDP-GTP exchange factor, putative | 1 | 0 | 0 | LbrM2904_000797200 | 1.06 |
| OG5_128043 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000088400 | 1.04 |
| OG5_128097 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000333100 | 1.1 |
| OG5_128550 | surface antigen-like protein | 1 | 0 | 0 | LbrM2904_000799500 | 0.95 |
| OG5_128609 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000072800 | 0.96 |
| OG5_128636 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000084300 | 1.08 |
| OG5_128693 | ribosomal RNA processing protein, putative | 1 | 0 | 0 | LbrM2904_000799900 | 1.05 |
| OG5_128713 | serine acetyltransferase , serine acetyltransferase, putative | 1 | 0 | 0 | LbrM2904_000275100 | 0.85 |
| OG5_129552 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000308600 | 1.18 |
| OG5_129928 | hypothetical | 1 | 0 | 1 | LbrM2904_000 | 0.92 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | protein, conserved | | | | 692400 | |
| OG5_132126 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000198500 | 0.83 |
| OG5_132827 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000008000 | 1.06 |
| OG5_134066 | tuzin-like protein | 1 | 0 | 3 | LbrM2904_000274300 | 5.67 |
| OG5_142321 | metalloprotease-like protein,peptide deformylase, putative (EMBL:AY353252) | 1 | 0 | 0 | LbrM2904_000814200 | 0.96 |
| OG5_149604 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000011500 | 1.17 |
| OG5_149914 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000754200 | 0.97 |
| OG5_151904 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000814400 | 0.87 |
| OG5_151905 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000109200 | 1.06 |
| OG5_151907 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000444600 | 0.88 |
| OG5_151908 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000461300 | 0.99 |
| OG5_151909 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000788200 | 0.85 |
| OG5_152437 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000666700 | 0.97 |
| OG5_154739 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000063800 | 1.05 |
| OG5_154740 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000354400 | 0.9 |
| OG5_154741 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000435200 | 1.03 |
| OG5_155868 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000190500 | 1.01 |
| OG5_158384 | hypothetical protein, conserved,inositol 5-phosphatase-like protein | 1 | 0 | 0 | LbrM2904_000796400 | 1 |
| OG5_161893 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000542500 | 0.78 |
| OG5_162271 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000654500 | 0.81 |
| OG5_162319 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000391200 | 1.21 |
| OG5_162394 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000697600 | 0.99 |
| OG5_164625 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000413900 | 0.9 |

| | | | | | | |
|---|---|---|---|---|---|---|
| OG5_164748 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000087800 | 0.84 |
| OG5_167675 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000667500 | 4.26 |
| OG5_167687 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000786200 | 0.93 |
| OG5_171326 | surface antigen-like protein | 1 | 0 | 0 | LbrM2904_000816800 | 0.84 |
| OG5_173551 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000497700 | 0.94 |
| OG5_174131 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000769300 | 0.82 |
| OG5_185212 | hypothetical protein ,T. brucei spp.-specific protein | 1 | 0 | 1 | LbrM2904_000201400 | 1.01 |
| OG5_200862 | hypothetical protein, conserved | 1 | 0 | 0 | LbrM2904_000173600 | 0.94 |
| OG5_204172 | hypothetical protein | 2 | 0 | 0 | LbrM2904_000805200,LbrM2904_000805300 | 2.12 |
| OG5_204174 | hypothetical protein | 2 | 0 | 2 | LbrM2904_000797500,LbrM2904_000797600 | 1.87 |
| OG5_204183 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000798100 | 0.83 |
| OG5_204184 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000082900 | 1.03 |
| OG5_204189 | hypothetical protein | 2 | 0 | 4 | LbrM2904_000111100,LbrM2904_000111400 | 5.51 |
| OG5_204190 | hypothetical protein, conserved, | 2 | 0 | 1 | LbrM2904_000112500,LbrM2904_000112600 | 1.81 |
| OG5_204193 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000137800 | 0.8 |
| OG5_204214 | hypothetical protein | 1 | 0 | 0 | LbrM2904_000792800 | 0.92 |
| OG5_204230 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000541500 | 0.92 |
| OG5_204253 | hypothetical protein | 2 | 0 | 2 | LbrM2904_000674800,LbrM2904_000674900 | 1.95 |
| OG5_206778 | selenoprotein, putative | 1 | 0 | 1 | LbrM2904_000782600 | 0.99 |
| OG5_236274 | multi drug resistance protein-like ,hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000379800 | 1.26 |
| OG5_236303 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000761400 | 0.99 |
| OG5_247477 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000808600 | 1.03 |
| OG5_247657 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000431500 | 1.29 |
| OG5_248246 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000458300 | 1.46 |

**Table B11:** Genes in orthologous groups (OGs) that are annotated on the control *L. braziliensis* M2904 genome but not the published *L. braziliensis* M2904 genome. The '*L. braziliensis* reftest'

column indicates the number of genes in each OG annotated on the *L. braziliensis* M2904 reference genome by the Companion pipeline in a test run.

| **OG** | | **Number of assembled genes in OG** | | | **Assembled genes in OG** | |
|---|---|---|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. braziliensis control*** | ***L. braziliensis*** | ***L. braziliensis reftest*** | ***L. braziliensis control*** | ***L. braziliensis*** |
| OG5_127444 | phosphatidylinositol 3-kinase 2, putative | 0 | 1 | 0 | | LbrM.14.0020 |
| OG5_127707 | argininosuccinate synthase, putative | 0 | 1 | 0 | | LbrM.23.0290 |
| OG5_128257 | ubiquitin-fusion protein | 0 | 3 | 0 | | LbrM.31.2110,LbrM.31.2130, LbrM.31.2290 |
| OG5_128938 | transporter, putative,major facilitator superfamily protein (MFS), putative | 0 | 1 | 0 | | LbrM.18.0050 |
| OG5_129071 | polyprenyl synthase, putative | 0 | 1 | 0 | | LbrM.19.0530 |
| OG5_129874 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.28.0160 |
| OG5_130907 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.32.0350 |
| OG5_131213 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.20.0631 |
| OG5_131275 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.35.4290 |
| OG5_132234 | PFPI/DJ-1-like protein, putative,cysteine peptidase, Clan PC(C), family C56, putative | 0 | 1 | 0 | | LbrM.34.3890 |
| OG5_132308 | phosphoinositide phosphatase, putative;with=GeneDB:LmjF22.0250 ,phosphoinositide phosphatase, putative; | 0 | 1 | 0 | | LbrM.22.0240 |
| OG5_133951 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.35.3882 |
| OG5_134187 | hypothetical protein | 0 | 1 | 0 | | LbrM.29.0740 |
| OG5_135460 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000326700 | LbrM.21.1560 |
| OG5_139861 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000817800 | LbrM.31.3710 |
| OG5_142178 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.19.0880 |
| OG5_144320 | hypothetical protein | 0 | 1 | 0 | | LbrM.18.1720 |
| OG5_144830 | inosine-adenosine-guanosine-nucleoside hydrolase, putative | 1 | 1 | 0 | LbrM2904_000522200 | LbrM.29.2850 |
| OG5_145867 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000081600 | LbrM.08.0640 |
| OG5_145983 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000377400 | LbrM.24.1160 |
| OG5_146098 | hypothetical protein, conserved ,hypothetical protein, conserved ,hypothetical protein, conserved ,hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000727300 | LbrM.35.0890 |
| OG5_146267 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.26.1000 |
| OG5_146874 | hypothetical protein, | 0 | 1 | 0 | | LbrM.32.2070 |

| | | | | | | |
|---|---|---|---|---|---|---|
| OG5_148233 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000014300 | LbrM.02.0190 |
| OG5_148319 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.09.1060 |
| OG5_148512 | hypothetical protein, conserved , | 0 | 1 | 0 | | LbrM.19.1601 |
| OG5_148600 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000326800 | LbrM.21.1550 |
| OG5_148779 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.27.1430 |
| OG5_148809 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.28.0440 |
| OG5_149001 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000692100 | LbrM.34.2561 |
| OG5_150188 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.35.4850 |
| OG5_151425 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000014400 | LbrM.02.0200 |
| OG5_151769 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.30.1300 |
| OG5_154335 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.20.3561 |
| OG5_154545 | arginine N-methyltransferase, putative | 1 | 1 | 0 | LbrM2904_000023500 | LbrM.15.1340 |
| OG5_154597 | protein kinase-like protein | 1 | 1 | 0 | LbrM2904_000810500 | LbrM.27.2860 |
| OG5_154637 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000522400 | LbrM.29.2860 |
| OG5_156056 | hypothetical protein | 0 | 1 | 0 | | LbrM.27.2850 |
| OG5_156080 | hypothetical protein | 1 | 1 | 0 | LbrM2904_000796600 | LbrM.03.0950 |
| OG5_158620 | hypothetical protein ,putative Pfkb family sugar kinase ,putative | 0 | 1 | 0 | | LbrM.21.0880 |
| OG5_160885 | proteophosphoglycan ppg1 | 0 | 1 | 0 | | LbrM.14.1690 |
| OG5_163246 | long-chain-fatty-acid-CoA ligase, putative | 1 | 1 | 0 | LbrM2904_000807000 | LbrM.01.0530 |
| OG5_164139 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.28.3240 |
| OG5_166072 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000114700 | LbrM.10.1370 |
| OG5_166750 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.30.0931 |
| OG5_166775 | hypothetical protein | 0 | 1 | 0 | | LbrM.34.4700 |
| OG5_169451 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.20.0610 |
| OG5_170968 | ribosomal protein L32-like protein | 0 | 1 | 0 | | LbrM.34.1870 |
| OG5_171467 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000803400 | LbrM.02.0730 |
| OG5_175808 | CYC2-like cyclin, putative | 1 | 1 | 0 | LbrM2904_000363800 | LbrM.05.0890 |
| OG5_177088 | RNA binding protein, putative | 1 | 1 | 0 | LbrM2904_000812500 | LbrM.29.2920 |
| OG5_180610 | ATPase beta subunit, putative | 0 | 1 | 0 | | LbrM.25.2500 |
| OG5_180644 | ribosomal protein L3, putative | 0 | 1 | 0 | | LbrM.32.3410 |
| OG5_182276 | elongation factor 2 | 0 | 1 | 0 | | LbrM.35.0250 |
| OG5_183345 | carboxypeptidase, putative, metallo-peptidase, Clan MA(E), family 32 | 0 | 2 | 0 | | LbrM.27.1350,LbrM.13.1580 |

| | | | | | | |
|---|---|---|---|---|---|---|
| OG5_18 3372 | hypothetical protein | 0 | 1 | 0 | | LbrM.15.0141 |
| OG5_18 3568 | hypothetical protein | 0 | 1 | 0 | | LbrM.21.0510 |
| OG5_18 3582 | centromere/microtubule binding protein cbf5, putative | 0 | 1 | 0 | | LbrM.21.1980 |
| OG5_18 3761 | ,hypothetical protein, conserved (pseudogene) ,hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.28.0430 |
| OG5_18 3825 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.29.1931 |
| OG5_18 3841 | hypothetical protein | 1 | 1 | 0 | LbrM2904_000 796500 | LbrM.29.2930 |
| OG5_18 3919 | hypothetical protein | 0 | 1 | 0 | | LbrM.31.2341 |
| OG5_18 3936 | hypothetical protein, | 0 | 1 | 0 | | LbrM.31.3021 |
| OG5_18 4000 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.32.3711 |
| OG5_18 4006 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.32.3930 |
| OG5_18 4110 | hypothetical protein, conserved | 1 | 1 | 0 | LbrM2904_000 702600 | LbrM.34.3620 |
| OG5_18 4122 | hypothetical protein, | 0 | 1 | 0 | | LbrM.35.0690 |
| OG5_20 4134 | hypothetical protein | 1 | 1 | 0 | LbrM2904_000 034200 | LbrM.04.0690 |
| OG5_20 4145 | hypothetical protein, | 0 | 1 | 0 | | LbrM.14.1090 |
| OG5_20 4162 | hypothetical protein | 0 | 1 | 0 | | LbrM.32.1680 |
| OG5_20 4163 | hypothetical protein, conserved (pseudogene) | 1 | 1 | 0 | LbrM2904_000 621700 | LbrM.32.3280 |
| OG5_20 4166 | hypothetical protein , | 0 | 1 | 0 | | LbrM.35.0410 |
| OG5_20 4168 | beta-adaptin protein, putative | 0 | 1 | 0 | | LbrM.35.5870 |
| OG5_20 8779 | 60S ribosomal protein L19, putative | 0 | 1 | 0 | | LbrM.06.0420 |
| OG5_21 3325 | serine/threonine protein phosphatase, putative | 0 | 1 | 0 | | LbrM.25.1290 |
| OG5_22 2491 | zinc-finger protein ZPR1, putative | 0 | 1 | 0 | | LbrM.17.1560 |
| OG5_23 6236 | heat shock 70-related protein 1, mitochondrial precursor, putative | 0 | 1 | 0 | | LbrM.30.2450 |
| OG5_23 6242 | hypothetical protein, conserved | 0 | 1 | 0 | | LbrM.32.2640 |

**Table B12:** Genes in orthologous groups (OGs) that are on the published *L. braziliensis* M2904 genome annotation but not the Companion annotation for the same genome

| **OG** | | **Number of assembled genes in OG** | | | **Assembled genes in OG** | | **OG haploid copy number** |
|---|---|---|---|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. braziliensis control*** | ***L. braziliensis*** | ***L. braziliensis reftest*** | ***L. braziliensis control*** | ***L. braziliensis reftest*** | ***Lbraziliensis control*** |
| OG5_126627 | hypothetical protein | 3 | 0 | 3 | LbrM2904_000804300,LbrM2904_000616100,LbrM2904_000337500 | LbrM_000683800,LbrM_000381500,LbrM_000382700 | 3.91 |
| OG5_129552 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000308600 | LbrM_000351100 | 1.18 |
| OG5_129928 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000692400 | LbrM_000765000 | 0.92 |
| OG5_130626 | epsilon-adaptin, putative, AP-1/4 adapter complex gamma/epsilon subunit, putative | 0 | 0 | 1 | | LbrM_000595300 | 0 |
| OG5_134066 | tuzin-like protein | 1 | 0 | 3 | LbrM2904_000274300 | LbrM_000300300,LbrM_000304900,LbrM_000334700 | 5.67 |
| OG5_149604 | hypothetical protein, conserved , | 1 | 0 | 1 | LbrM2904_000011500 | LbrM_000030700 | 1.17 |
| OG5_149914 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000754200 | LbrM_000830500 | 0.97 |
| OG5_152437 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000666700 | LbrM_000737500 | 0.97 |
| OG5_154739 | hypothetical protein, conserved conserved | 1 | 0 | 1 | LbrM2904_000063800 | LbrM_000087600 | 1.05 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| OG5_162271 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000654500 | LbrM_000724900 | 0.81 |
| OG5_162319 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000391200 | LbrM_000439000 | 1.21 |
| OG5_162394 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000697600 | LbrM_000770200 | 0.99 |
| OG5_164954 | hypothetical protein, conserved | 0 | 0 | 1 | | LbrM_000635300 | 0 |
| OG5_167675 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000667500 | LbrM_000716400 | 4.26 |
| OG5_173551 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000497700 | LbrM_000551900 | 0.94 |
| OG5_174131 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000769300 | LbrM_000845500 | 0.82 |
| OG5_185212 | hypothetical protein,T. brucei spp.-specific protein | 1 | 0 | 1 | LbrM2904_000201400 | LbrM_000235700 | 1.01 |
| OG5_204174 | hypothetical protein | 2 | 0 | 2 | LbrM2904_000797500,LbrM2904_000797600 | LbrM_000020700,LbrM_000020800 | 1.87 |
| OG5_204184 | hypothetical protein, | 1 | 0 | 1 | LbrM2904_000082900 | LbrM_000107800 | 1.03 |
| OG5_204189 | hypothetical protein | 2 | 0 | 4 | LbrM2904_000111100,LbrM2904_000111400 | LbrM_000139400,LbrM_000139500,LbrM_000139700,LbrM_000139800 | 5.51 |
| OG5_204190 | hypothetical protein | 2 | 0 | 1 | LbrM2904_000112500,LbrM2904_000112600 | LbrM_000140900 | 1.81 |
| OG5_204230 | hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000541500 | LbrM_000598700 | 0.92 |
| OG5_204253 | hypothetical protein | 2 | 0 | 2 | LbrM2904_000674800,LbrM2904_000674900 | LbrM_000746400,LbrM_000746500 | 1.95 |
| OG5_20677 | selenoprotei | 1 | 0 | 1 | LbrM2904_000782600 | LbrM_000860400 | 0.99 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| 8 | n, putative, hypothetical protein |  |  |  |  |  |  |
| OG5_236274 | multidrug resistance protein-like ,hypothetical protein, conserved | 1 | 0 | 1 | LbrM2904_000379800 | LbrM_000427200 | 1.26 |
| OG5_236303 | hypothetical protein , | 1 | 0 | 1 | LbrM2904_000761400 | LbrM_000838300 | 0.99 |
| OG5_247477 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000808600 | LbrM_000011600 | 1.03 |
| OG5_247657 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000431500 | LbrM_000480800 | 1.29 |
| OG5_248246 | hypothetical protein | 1 | 0 | 1 | LbrM2904_000458300 | LbrM_000511100 | 1.46 |

**Table B13:** Genes in orthologous groups (OGs) that are annotated on *L. braziliensis* M2904 by Companion but not on the published *L. braziliensis* M2904 annotation.

**Not in *Leishmania* or *Trypanosoma***

| **OG** | | **Number of assembled genes in OG** | **Assembled genes in OG** | **Species in OG** |
|---|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. naiffi*** | ***L. naiffi*** | **Species names** |
| OG5_134703 | methylenetetrahydrofolate reductase putative, hypothetical protein | 1 | LnCL223_352070 | *Aspergillus fumigatus Af293, Debaryomyces hansenii CBS767, Emericella nidulans, Aspergillus oryzae RIB40, Coccidioides immitis RS, Candida albicans, Candida glabrata CBS 138, Cryptococcus neoformans var. grubii H99, Cryptococcus bacillisporus, Eremothecium gossypii, Geobacter sulfurreducens PCA, Gibberella zeae PH-1, Kluyveromyces lactis NRRL Y-1140, Saccharomyces cerevisiae S288c, Schizosaccharomyces pombe, Scheffersomyces stipitis CBS 6054, Synechococcus sp. WH 8102, Thermotoga maritima MSB8, Thermoplasma volcanium GSS1, Neurospora crassa OR74A, Laccaria bicolor S238N-H82, Phanerochaete chrysosporium, Yarrowia lipolytica CLIB122, Yersinia pestis CO92* |
| OG5_212197 | hypothetical protein, Transferase family, putative | 1 | LnCL223_341350 | *Neurospora crassa OR74A, Physcomitrella patens subsp. patens* |

**In *Trypanosoma***

| **OG** | | **Number of assembled genes in OG** | | | | **Assembled genes in OG** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Orthologous Group ID** | **Description** | ***L. naiffi*** | ***T. brucei gambiense*** | ***T.brucei*** | ***T.vivax*** | ***L. naiffi*** | ***T. brucei gambiense*** | ***T.brucei*** | ***T.vivax*** |
| OG5_167811 | hypothetical protein, conserved in T. vivax ,hypothetical protein, conserved in T. vivax ,hypothetical protein, conserved in T. vivax ,hypothetical protein, conserved in T. vivax, (fragment) ,hypothetical protein, conserved in T. vivax ,hypothetical protein, conserved | 1 | 0 | 0 | 6 | LnCL223_292920 | | | TvY486_0200060,TvY486_0042930,TvY486_1008220,TvY486_0803125,TvY486_0043250,TvY486_1101880 |
| OG5_206550 | hypothetical protein ,hypothetical protein | 1 | 1 | 2 | 0 | LnCL223_312570 | Tbg972.1.1410 | Tb427.01.2420,Tb927.1.2420 | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ,T. brucei spp.-specific protein,hypothetical protein | | | | | | | | |

**Other**

| **Orthologous Group ID** | **Description** | ***L. naiffi*** | ***L. braziliensis control*** | ***L. braziliensis*** | ***L. braziliensis reftest*** | ***L. naiffi*** | ***L. braziliensis control*** | ***L. braziliensis reftest*** | **Species in OG** |
|---|---|---|---|---|---|---|---|---|---|
| OG5_129552 | Putative methyltransferase/Methyltransferase domain containing protein | 1 | 1 | 0 | 1 | LnCL223_2021430 | LbrM2904_000308600 | LbrM_000351100 | *Aedes aegypti, Aspergillus fumigatus Af293, Anopheles gambiae str. PEST, Apis mellifera, Emericella nidulans, Aspergillus oryzae RIB40, Acyrthosiphon pisum, Arabidopsis thaliana, Brugia malayi, Bombyx mori, Candida albicans, Caenorhabditis briggsae AF16, Caenorhabditis elegans, Candida glabrata CBS 138, Cryptosporidium hominis TU502, Coccidioides immitis RS, Ciona intestinalis, Canis lupus familiaris, Cryptosporidium muris RN66, Cryptococcus neoformans var. grubii H99, Cryptococcus bacillisporus, Cryptosporidium parvum Iowa II, Culex pipiens, Coccidioides posadasii RMSCC 3488, Chlamydomonas reinhardtii, Dictyostelium discoideum AX4, Debaryomyces hansenii CBS767, Drosophila melanogaster, Danio rerio, Equus caballus, Eremothecium gossypii, Gallus gallus, Gibberella zeae PH-1, Homo sapiens, Ixodes scapularis, Kluyveromyces lactis NRRL Y-1140, Laccaria bicolor S238N-H82,Monosiga brevicollis MX1, Monodelphis domestica, Micromonas sp. RCC299, Macaca mulatta, Mus musculus, Neurospora crassa OR74A, Oryza sativa Japonica Group, Ostreococcus tauri, Phanerochaete chrysosporium, Physcomitrella patens subsp. patens, Phytophthora ramorum, Scheffersomyces stipitis CBS 6054, Pan troglodytes, Ricinus communis, Rattus norvegicus, Saccharomyces cerevisiae S288c, Schizosaccharomyces pombe, Trichoplax adhaerens, Tetraodon nigroviridis, Takifugu rubripes, Yarrowia lipolytica CLIB122* |

**Table B14:** Genes in orthologous groups (OGs) that are only found on *L. naiffi* CL223 when compared with other *Leishmania* species

**Table B15:** Genes in orthologous groups (OGs) that are only found on *L. guyanensis* CL085. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

| **OG** | | **Number of assembled genes in OG** | | | | | | | | | | | **Assembled genes in OG** | | | | | **OG haploid copy number** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ***Leishmania Viannia*** | | | | | | ***Leishmania Leishmania*** | | | ***Sauroleishmania*** | | ***Leishmania Viannia*** | | | | ***Sauroleishmania*** | ***Leishmania Viannia*** | | |
| **Orthologous Group ID** | **Description** | ***L. guyanensis*** | ***L. naiffi*** | ***L. braziliensis control*** | ***L. braziliensis*** | ***L. braziliensis reftest*** | ***L. panamensis*** | ***L. major*** | ***L. infantum*** | ***L. mexicana*** | ***L. adleri*** | ***L. tarentolae*** | ***L. guyanensis*** | ***L. naiffi*** | ***L. braziliensis control*** | ***L. braziliensis reftest*** | ***L. adleri*** | ***L. guyanensis*** | ***L. naiffi*** | ***Lbraziliensis control*** |
| OG5_185212 | hypothetical protein,T. brucei spp.-specific protein | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | LgCL085_170040 | LnCL223_170040 | LbrM2904_000201400 | LbrM_000235700 | LaHO174_170020 | 0.98 | 1.1 | 1.01 |
| OG5_206778 | selenoprotein, putative,hypothetical protein | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | LgCL085_356500 | LnCL223_356340 | LbrM2904_000782600 | LbrM_000860400 | LaHO174_3623670 | 0.93 | 1.17 | 0.99 |

**Table B16:** Genes previously unique to *L. adleri* but also in *L. naiffi* CL223 and *L. guyanensis* CL085

**Table B17**: Genes exclusive to *Viannia* genomes. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B18:** *L. guyanensis* CL085 orthologous groups with >2 the number of haploid copies of genes compared with assembled copies. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B19:** *L. naiffi* CL223 orthologous groups with >2 the number of haploid copies of genes compared with assembled copies. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B20:** *L. braziliensis* M2904 orthologous groups with >2 the number of haploid copies of genes compared with assembled copies. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B21:** Genes involved in the RNAi pathway. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B22:** *L. naiffi* CL223 gene arrays. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B23:** *L. guyanensis* CL085 gene arrays. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

**Table B24:** *L. braziliensis* M2904 control gene arrays. This table is online at https://figshare.com/s/84575f55f47386a2d4e7

# Appendix C –Identification of genomic and transcriptomic changes associated with drug resistance in Methicillin Resistant *Staphylococcus aureus* including a large tandem amplification of the SCC*mec*V element

*Some tables in this section were too large to include and so are deposited on Figshare at https://figshare.com/s/c0370d2800fe73a007e1. The legends of each of these tables are included in this appendix as well as an indication that the particular table can be found online.*

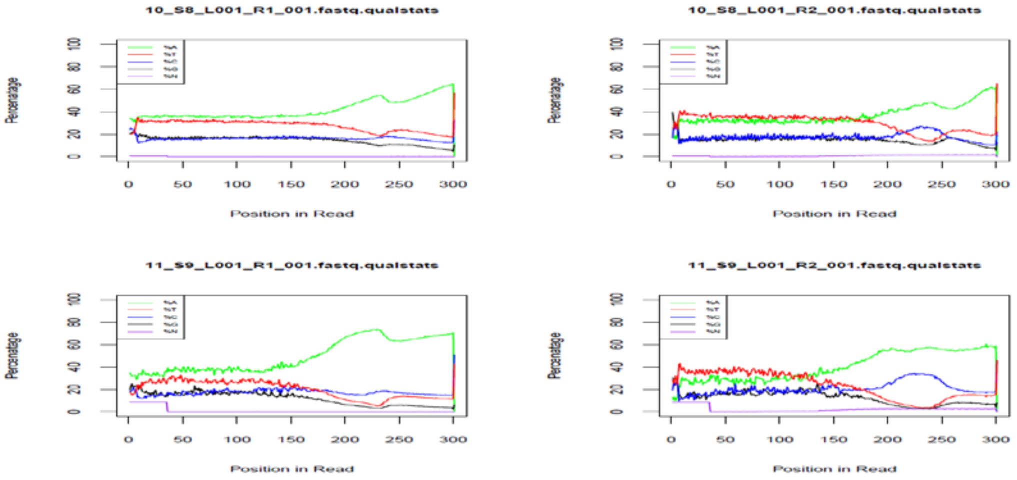


**Figure C1:** Nucleotide distribution plots of RNASeq reads from samples 10_S8 and 11_S9 in the low dose experiment showing an abnormal distribution of bases after approximately 150 bp.

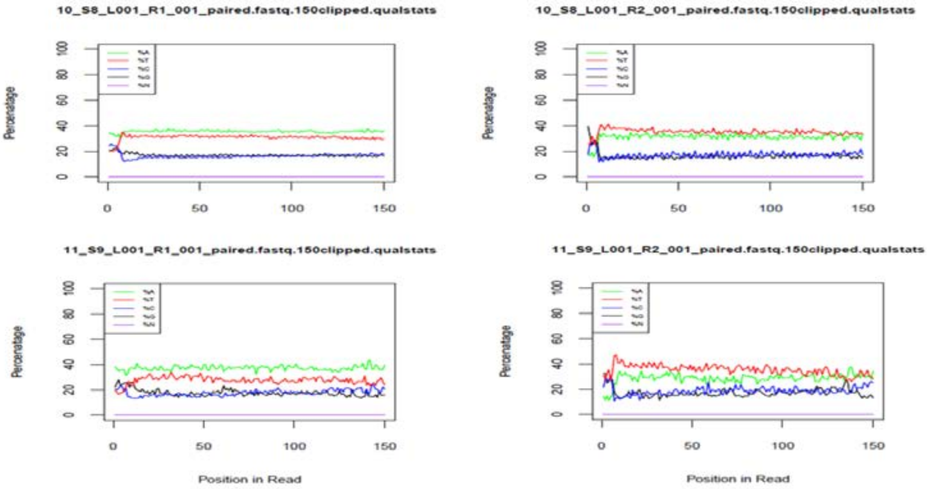


**Figure C2:** Nucleotide distribution plots of RNASeq reads from samples 10_S8 and 11_S9 in the low dose experiment showing improved distributions after removal of the last 150 bp of all reads.

|  | Unprocessed Reads |  |  | QC Processed Reads |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | R1 | R2 | Total | R1 paired | R2 paired | Total paired left | R1 unpaired | R2 unpaired | Unassigned unpaired | Total paired + unpaired | % Total remaining reds | Reads discarded |
| **1a_S1** | 798,437 | 798,437 | 1,596,874 | 620,803 | 620,803 | 1,241,606 | 22,336 | 107,284 | 15,343 | 1,386,569 | 86.83 | 210,305 |
| **1b_S2** | 957,761 | 957,761 | 1,915,522 | 782,581 | 782,581 | 1,565,162 | 29,323 | 89,394 | 13,743 | 1,697,622 | 88.62 | 217,900 |
| **1c_S3** | 1,198,086 | 1,198,086 | 2,396,172 | 965,680 | 965,680 | 1,931,360 | 83,700 | 62,111 | 14,039 | 2,091,210 | 87.27 | 304,962 |
| **2a_S4** | 567,673 | 567,673 | 1,135,346 | 426,898 | 426,898 | 853,796 | 9,552 | 100,961 | 13,141 | 977,450 | 86.09 | 157,896 |
| **2b_S5** | 1,040,838 | 1,040,838 | 2,081,676 | 834,132 | 834,132 | 1,668,264 | 24,127 | 125,073 | 17,163 | 1,834,627 | 88.13 | 247,049 |
| **2c_S6** | 969,738 | 969,738 | 1,939,476 | 758,126 | 758,126 | 1,516,252 | 33,283 | 117,057 | 16,903 | 1,683,495 | 86.8 | 255,981 |
| **3a_S7** | 1,627,376 | 1,627,376 | 3,254,752 | 1,336,947 | 1,336,947 | 2,673,894 | 134,123 | 48,611 | 15,101 | 2,871,729 | 88.23 | 383,023 |
| **3b_S8** | 1,382,523 | 1,382,523 | 2,765,046 | 1,129,113 | 1,129,113 | 2,258,226 | 121,911 | 36,736 | 12,764 | 2,429,637 | 87.87 | 335,409 |
| **3c_S9** | 1,071,442 | 1,071,442 | 2,142,884 | 831,232 | 831,232 | 1,662,464 | 144,804 | 18,869 | 10,852 | 1,836,989 | 85.73 | 305,895 |
| **4a_S10** | 1,117,642 | 1,117,642 | 2,235,284 | 888,300 | 888,300 | 1,776,600 | 128,420 | 23,778 | 10,940 | 1,939,738 | 86.78 | 295,546 |
| **4b_S11** | 1,750,255 | 1,750,255 | 3,500,510 | 1,460,589 | 1,460,589 | 2,921,178 | 100,787 | 58,008 | 14,469 | 3,094,442 | 88.4 | 406,068 |
| **4c_S12** | 999,344 | 999,344 | 1,998,688 | 757,685 | 757,685 | 1,515,370 | 139,824 | 20,096 | 11,020 | 1,686,310 | 84.37 | 312,378 |
| **5a_S13** | 1,262,246 | 1,262,246 | 2,524,492 | 976,625 | 976,625 | 1,953,250 | 158,882 | 28,250 | 13,287 | 2,153,669 | 85.31 | 370,823 |
| **5b_S14** | 1,368,276 | 1,368,276 | 2,736,552 | 1,125,802 | 1,125,802 | 2,251,604 | 123,428 | 33,406 | 12,343 | 2,420,781 | 88.46 | 315,771 |
| **5c_S15** | 1,061,449 | 1,061,449 | 2,122,898 | 795,533 | 795,533 | 1,591,066 | 166,599 | 17,114 | 11,501 | 1,786,280 | 84.14 | 336,618 |
| **6a_S16** | 1,392,942 | 1,392,942 | 2,785,884 | 1,134,845 | 1,134,845 | 2,269,690 | 128,496 | 37,808 | 13,068 | 2,449,062 | 87.91 | 336,822 |
| **6b_S17** | 2,126,836 | 2,126,836 | 4,253,672 | 1,737,667 | 1,737,667 | 3,475,334 | 185,098 | 62,347 | 20,785 | 3,743,564 | 88.01 | 510,108 |
| **6c_S18** | 1,052,278 | 1,052,278 | 2,104,556 | 838,439 | 838,439 | 1,676,878 | 112,384 | 25,747 | 10,518 | 1,825,527 | 86.74 | 279,029 |

**Table C1:** Quality control statistics of DNA samples in the low dose experiment. The QC processed read numbers are the final reads used for all analysis and are reads that have been trimmed, had 50 bases clipped from their 5’ ends and error corrected.  R1 denotes the first (forward) read of a pair and R2 denotes the second (reverse) read of a pair.

|  | Unprocessed Reads |  |  | QC Processed Reads |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | R1 | R2 | Total | R1 paired | R2 paired | Total paired remaining | % of Paired reads left | R1 unpaired | R2 unpaired | Total paired + unpaired | %Total Reads remaining | Reads discarded* |
| 1_S1 | 2,179,206 | 2,179,206 | 4,358,412 | 1,731,794 | 1,731,794 | 3,463,588 | 79 | 261,414 | 50,469 | 3,775,471 | 87 | 582,941 |
| 2_S2 | 2,198,439 | 2,198,439 | 4,396,878 | 1,684,939 | 1,684,939 | 3,369,878 | 77 | 136,221 | 77,676 | 3,583,775 | 82 | 813,103 |
| 4_S3 | 2,545,365 | 2,545,365 | 5,090,730 | 1,946,682 | 1,946,682 | 3,893,364 | 76 | 353,505 | 60,396 | 4,307,265 | 85 | 783,465 |
| 6_S4 | 2,596,020 | 2,596,020 | 5,192,040 | 2,122,028 | 2,122,028 | 4,244,056 | 82 | 222,915 | 73,578 | 4,540,549 | 87 | 651,491 |
| 7_S5 | 2,576,335 | 2,576,335 | 5,152,670 | 1,942,782 | 1,942,782 | 3,885,564 | 75 | 429,000 | 47,772 | 4,362,336 | 85 | 790,334 |
| 8_S6 | 2,952,649 | 2,952,649 | 5,905,298 | 2,386,503 | 2,386,503 | 4,773,006 | 81 | 308,229 | 70,791 | 5,152,026 | 87 | 753,272 |
| 9_S7 | 2,778,065 | 2,778,065 | 5,556,130 | 2,086,012 | 2,086,012 | 4,172,024 | 75 | 467,314 | 50,898 | 4,690,236 | 84 | 865,894 |
| 10_S8 | 2,346,668 | 2,346,668 | 4,693,336 | 1,851,954 | 1,851,954 | 3,703,908 | 79 | 294,503 | 54,367 | 4,052,778 | 86 | 640,558 |
| 11_S9 | 2,338,505 | 2,338,505 | 4,677,010 | 1,790,716 | 1,790,716 | 3,581,432 | 77 | 145,744 | 79,046 | 3,806,222 | 81 | 870,788 |
| 12_S10 | 2,331,266 | 2,331,266 | 4,662,532 | 1,812,485 | 1,812,485 | 3,624,970 | 78 | 144,220 | 83,645 | 3,852,835 | 83 | 809,697 |
| 13_S11 | 2,364,611 | 2,364,611 | 4,729,222 | 1,896,075 | 1,896,075 | 3,792,150 | 80 | 218,416 | 55,366 | 4,065,932 | 86 | 663,290 |
| 14_S12 | 2,275,443 | 2,275,443 | 4,550,886 | 1,863,222 | 1,863,222 | 3,726,444 | 82 | 202,782 | 202,782 | 4,132,008 | 91 | 418,878 |
| 15_S13 | 2,360,355 | 2,360,355 | 4,720,710 | 1,932,354 | 1,932,354 | 3,864,708 | 82 | 181,161 | 82,535 | 4,128,404 | 87 | 592,306 |
| 16_S14 | 2,512,428 | 2,512,428 | 5,024,856 | 2,066,636 | 2,066,636 | 4,133,272 | 82 | 174,527 | 81,874 | 4,389,673 | 87 | 635,183 |

**Table C2:** Quality control statistics of RNA samples in the low dose experiment. The QC processed read numbers are the final reads used for all analysis and are reads that have been trimmed and had 150 bp removed from their 3’ ends. R1 denotes the first (forward) read of a pair and R2 denotes the second (reverse) read of a pair.

|  | Unprocessed Reads |  |  | QC Processed Reads |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sample** | **R1** | **R2** | **Total** | **R1 paired** | **R2 paired** | **Total paired left** | **R1 unpaired** | **R2 unpaired** | **Unassigned unpaired** | **Total paired + unpaired** | **% Total remaining reds** | **Reads discarded** |
| **1A_S1** | 792,746 | 792,746 | 1,585,492 | 580,517 | 580,517 | 1,161,034 | 35,056 | 129,143 | 0 | 1,325,233 | 83.58 | 260,259 |
| **2A_S2** | 1,196,498 | 1,196,498 | 2,392,996 | 926,864 | 926,864 | 1,853,728 | 62,822 | 131,419 | 0 | 2,047,969 | 85.58 | 345,027 |
| **3A_S3** | 1,001,489 | 1,001,489 | 2,002,978 | 781,061 | 781,061 | 1,562,122 | 81,618 | 65,493 | 0 | 1,709,233 | 85.33 | 293,745 |
| **4A_S4** | 507,155 | 507,155 | 1,014,310 | 345,421 | 345,421 | 690,842 | 18,799 | 113,992 | 0 | 823,633 | 81.20 | 190,677 |
| **5A_S5** | 1,130,129 | 1,130,129 | 2,260,258 | 859,758 | 859,758 | 1,719,516 | 53,032 | 151,623 | 1 | 1,924,172 | 85.13 | 336,086 |
| **6A_S6** | 938,021 | 938,021 | 1,876,042 | 702,281 | 702,281 | 1,404,562 | 43,773 | 135,375 | 0 | 1,583,710 | 84.42 | 292,332 |
| **7A_S7** | 1,948,796 | 1,948,796 | 3,897,592 | 1,540,946 | 1,540,946 | 3,081,892 | 216,814 | 60,945 | 0 | 3,359,651 | 86.20 | 537,941 |
| **8A_S8** | 1,965,346 | 1,965,346 | 3,930,692 | 1,552,719 | 1,552,719 | 3,105,438 | 237,457 | 49,670 | 1 | 3,392,565 | 86.31 | 538,127 |
| **9A_S9** | 1,846,212 | 1,846,212 | 3,692,424 | 1,382,010 | 1,382,010 | 2,764,020 | 308,584 | 27,812 | 0 | 3,100,416 | 83.97 | 592,008 |
| **10A_S10** | 1,772,586 | 1,772,586 | 3,545,172 | 1,364,945 | 1,364,945 | 2,729,890 | 251,285 | 38,605 | 0 | 3,019,780 | 85.18 | 525,392 |
| **11A_S11** | 1,494,639 | 1,494,639 | 2,989,278 | 1,093,319 | 1,093,319 | 2,186,638 | 266,496 | 24,220 | 0 | 2,477,354 | 82.87 | 511,924 |
| **12A_S12** | 1,737,719 | 1,737,719 | 3,475,438 | 1,376,634 | 1,376,634 | 2,753,268 | 194,549 | 52,442 | 0 | 3,000,259 | 86.33 | 475,179 |
| **13A_13** | 1,509,633 | 1,509,633 | 3,019,266 | 1,106,960 | 1,106,960 | 2,213,920 | 255,989 | 27,441 | 0 | 2,497,350 | 82.71 | 521,916 |
| **14A_14** | 1,333,727 | 1,333,727 | 2,667,454 | 983,919 | 983,919 | 1,967,838 | 227,600 | 21,679 | 1 | 2,217,117 | 83.12 | 450,337 |
| **15A_S15** | 1,463,644 | 1,463,644 | 2,927,288 | 1,052,992 | 1,052,992 | 2,105,984 | 282,898 | 17,142 | 0 | 2,406,024 | 82.19 | 521,264 |
| **B1_S16** | 1,366,973 | 1,366,973 | 2,733,946 | 1,012,876 | 1,012,876 | 2,025,752 | 221,607 | 27,018 | 1 | 2,274,377 | 83.19 | 459,569 |

**Table C3:** Quality control statistics of DNA samples in the bioreactor experiment. The QC processed reads numbers are the final reads used for all analysis and are reads that have been trimmed and, had 20 bp clipped from their 5’ end s and were error corrected. R1 denotes the first (forward) read of a pair and R2 denotes the second (reverse) read of a pair.

| | Insert size | | Read lengths | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample | Mean | Standard deviation | Mean | Median | Minimum | Maximum | Standard deviation | Total reads |
| 1a_S1 | 225.80 | 118.56 | 156.94 | 162 | 21 | 281 | 68.76 | 1,386,569 |
| 1b_S2 | 197.98 | 118.22 | 152.47 | 154 | 21 | 281 | 69.99 | 1,697,622 |
| 1c_S3 | 203.44 | 122.41 | 159 | 161 | 21 | 281 | 73.07 | 2,091,210 |
| 2a_S4 | 225.51 | 113.64 | 155.47 | 160 | 21 | 281 | 67.22 | 977,450 |
| 2b_S5 | 178.21 | 118.96 | 141.59 | 138 | 21 | 281 | 70.12 | 1,834,627 |
| 2c_S6 | 204.55 | 121.61 | 149.75 | 151 | 21 | 281 | 69.62 | 1,683,495 |
| 3a_S7 | 182.00 | 119.20 | 155.35 | 153 | 21 | 281 | 74.57 | 2,871,729 |
| 3b_S8 | 177.50 | 121.28 | 153.4 | 150 | 21 | 281 | 76.72 | 2,429,637 |
| 3c_S9 | 256.20 | 112.32 | 190.5 | 203 | 21 | 281 | 71.83 | 1,836,989 |
| 4a_S10 | 277.67 | 111.98 | 195.91 | 212 | 21 | 281 | 71.39 | 1,939,738 |
| 4b_S11 | 139.86 | 113.20 | 132.15 | 119 | 21 | 281 | 76.77 | 3,094,442 |
| 4c_S12 | 298.54 | 113.92 | 194.07 | 212 | 21 | 281 | 73.64 | 1,686,310 |
| 5a_S13 | 286.70 | 113.41 | 192.96 | 210 | 21 | 281 | 72.63 | 2,153,669 |
| 5b_S14 | 224.36 | 112.30 | 179.23 | 188 | 21 | 281 | 70.59 | 2,420,781 |
| 5c_S15 | 298.99 | 110.48 | 197.68 | 215 | 21 | 281 | 73.36 | 1,786,280 |
| 6a_S16 | 234.72 | 112.85 | 182.35 | 191 | 21 | 281 | 69.84 | 2,449,062 |
| 6b_S17 | 242.66 | 111.06 | 183.81 | 195 | 21 | 281 | 69.63 | 3,743,564 |
| 6c_S18 | 240.85 | 112.80 | 183.18 | 193 | 21 | 281 | 70.35 | 1,825,527 |

**Table C4:** Insert sizes and read lengths of DNA reads in the low dose experiment.

| | Insert size | | Read lengths | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample | Mean | Standard deviation | Mean | Median | Minimum | Maximum | Standard deviation | Total reads |
| 1_S1 | 125.58 | 55.62 | 111.73 | 113 | 35 | 150 | 33.84 | 3,775,471 |
| 2_S2 | 81.67 | 28.03 | 80.27 | 74 | 35 | 150 | 27.32 | 3,583,775 |
| 4_S3 | 128.75 | 55.98 | 113.49 | 117 | 35 | 150 | 34.05 | 4,307,265 |
| 6_S4 | 109.95 | 39.65 | 104.3 | 102 | 35 | 150 | 33.82 | 4,540,549 |
| 7_S5 | 144.46 | 60.91 | 122.68 | 136 | 35 | 150 | 31.79 | 4,362,336 |
| 8_S6 | 118.34 | 43.43 | 110.52 | 112 | 35 | 150 | 32.96 | 5,152,026 |
| 9_S7 | 152.24 | 77.07 | 121.44 | 136 | 35 | 150 | 33.06 | 4,690,236 |
| 10_S8 | 126.54 | 53.96 | 113.47 | 118 | 35 | 150 | 34.02 | 4,052,778 |
| 11_S9 | 87.81 | 31.57 | 85.41 | 80 | 35 | 150 | 30.28 | 3,806,222 |
| 12_S10 | 86.23 | 30.95 | 84.01 | 79 | 35 | 150 | 29.46 | 3,852,835 |
| 13_S11 | 103.96 | 36.74 | 99.25 | 96 | 35 | 150 | 32.09 | 4,065,932 |
| 14_S12 | 114.06 | 39.39 | 106.4 | 106 | 35 | 150 | 32.58 | 3,992,627 |
| 15_S13 | 95.16 | 31.85 | 91.31 | 88 | 35 | 150 | 29.88 | 4,128,404 |
| 16_S14 | 86.88 | 32.06 | 84.97 | 79 | 35 | 150 | 29.37 | 4,389,673 |

**Table C5:** Insert sizes and read lengths of RNA reads in the low dose experiment.

|  | Insert size |  | Read lengths |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| **Sample** | **Mean** | **Standard deviation** | **Mean** | **Median** | **Minimum** | **Maximum** | **Standard deviation** | **Total reads** |
| **1A_S1** | 223.87 | 109.36 | 169.97 | 166 | 36 | 301 | 78.49 | 1,325,233 |
| **2A_S2** | 206.17 | 108.12 | 164.44 | 157 | 36 | 301 | 78.55 | 2,047,969 |
| **3A_S3** | 197.92 | 103.64 | 164.38 | 156 | 36 | 301 | 78.81 | 1,709,233 |
| **4A_S4** | 236.30 | 111.15 | 171.62 | 170 | 36 | 301 | 79.87 | 823,633 |
| **5A_S5** | 216.28 | 111.96 | 168.67 | 163 | 36 | 301 | 79.99 | 1,924,172 |
| **6A_S6** | 210.28 | 110.66 | 163.39 | 156 | 36 | 301 | 79.05 | 1,583,710 |
| **7A_S7** | 186.65 | 103.08 | 161.5 | 150 | 36 | 301 | 80.16 | 3,359,651 |
| **8A_S8** | 194.61 | 103.07 | 167.83 | 159 | 36 | 301 | 79.84 | 3,392,566 |
| **9A_S9** | 241.12 | 104.07 | 198.26 | 204 | 36 | 301 | 78.97 | 3,100,416 |
| **10A_S10** | 238.41 | 105.52 | 196.08 | 202 | 36 | 301 | 79.52 | 3,019,780 |
| **11A_S11** | 264.23 | 110.35 | 209.65 | 225 | 36 | 301 | 80.63 | 2,477,354 |
| **12A_S12** | 186.94 | 101.01 | 161.45 | 150 | 36 | 301 | 78.85 | 3,000,259 |
| **13A_13** | 286.73 | 104.29 | 216.65 | 239 | 36 | 301 | 80.30 | 2,497,350 |
| **14A_14** | 283.29 | 103.54 | 216.39 | 238 | 36 | 301 | 79.60 | 2,217,118 |
| **15A_S15** | 283.79 | 103.99 | 217.84 | 240 | 36 | 301 | 80.01 | 2,406,024 |
| **B1_S16** | 288.22 | 104.51 | 217.92 | 241 | 36 | 301 | 80.02 | 2,274,378 |

**Table C6:** Insert sizes and read lengths of DNA reads in the bioreactor experiment

| Sample | Mean | Median | Minimum | Maximum | Standard Deviation |
|---|---|---|---|---|---|
| **1a_S1** | 36.31 | 35 | 0 | 175 | 14.09 |
| **1b_S2** | 44.76 | 44 | 0 | 213 | 16.19 |
| **1c_S3** | 71.16 | 69 | 0 | 362 | 27.51 |
| **2a_S4** | 35.32 | 34 | 0 | 172 | 12.16 |
| **2b_S5** | 67.02 | 66 | 0 | 282 | 20.96 |
| **2c_S6** | 80.75 | 78 | 0 | 393 | 25.56 |
| **3a_S7** | 138.97 | 136 | 0 | 613 | 49.40 |
| **3b_S8** | 109.08 | 105 | 0 | 702 | 38.93 |
| **3c_S9** | 108.81 | 106 | 0 | 625 | 33.07 |
| **4a_S10** | 75.73 | 74 | 0 | 399 | 26.01 |
| **4b_S11** | 94.09 | 91 | 0 | 405 | 33.76 |
| **4c_S12** | 85.92 | 85 | 0 | 351 | 28.90 |
| **5a_S13** | 112.21 | 109 | 0 | 524 | 40.81 |
| **5b_S14** | 118.27 | 113 | 0 | 573 | 43.45 |
| **5c_S15** | 88.60 | 86 | 0 | 491 | 40.49 |
| **6a_S16** | 142.77 | 137 | 0 | 707 | 53.99 |
| **6b_S17** | 194.04 | 185 | 0 | 1343 | 74.00 |
| **6c_S18** | 97.38 | 95 | 0 | 471 | 31.40 |

**Table C0.7:** Coverage of reads mapped to the chromosome (accession NC_007793) in the low dose experiment

| Sample | Mean | Median | Minimum | Maximum | Standard Deviation |
|---|---|---|---|---|---|
| **1A_S1** | 73.46 | 73 | 0 | 354 | 26.16 |
| **2A_S2** | 108.18 | 107 | 0 | 528 | 43.97 |
| **3A_S3** | 95.14 | 95 | 0 | 499 | 32.82 |
| **4A_S4** | 47.15 | 46 | 0 | 223 | 17.23 |
| **5A_S5** | 106.09 | 106 | 0 | 452 | 37.21 |
| **6A_S6** | 85.77 | 85 | 0 | 395 | 32.02 |
| **7A_S7** | 183.84 | 184 | 0 | 1048 | 60.18 |
| **8A_S8** | 191.28 | 175 | 0 | 4566 | 214.95 |
| **9A_S9** | 197.19 | 198 | 0 | 802 | 78.30 |
| **10A_S10** | 190.60 | 190 | 0 | 1051 | 65.54 |
| **11A_S11** | 172.03 | 169 | 0 | 1086 | 48.98 |
| **12A_S12** | 149.62 | 142 | 0 | 1615 | 71.22 |
| **13A_13** | 179.78 | 176 | 0 | 1033 | 90.17 |
| **14A_14** | 161.89 | 158 | 0 | 997 | 84.75 |
| **15A_S15** | 176.38 | 176 | 0 | 845 | 80.60 |
| **B1_S16** | 157.73 | 155 | 0 | 823 | 64.55 |

**Table C8:** Coverage of reads mapped to the chromosome (accession NC_007793) in the bioreactor experiment

| Accession | Details | Read number |
|---|---|---|
| ERR580965 | Ion torrent 200bp WGS single end | 3,255,922 |
| ERR580966 | Ion torrent 400bp WGS single end | 5,900,237 |
| ERR580967 | MiSeq 2X 150bp WGS paired end | 1,562,528 (3,125,056 total) |
| ERR580968 | MiSeq 2X 250bp WGS paired end | 1,221,738 (2,443,476 total) |

**Table C9:** Accession numbers and details of reads from [520].

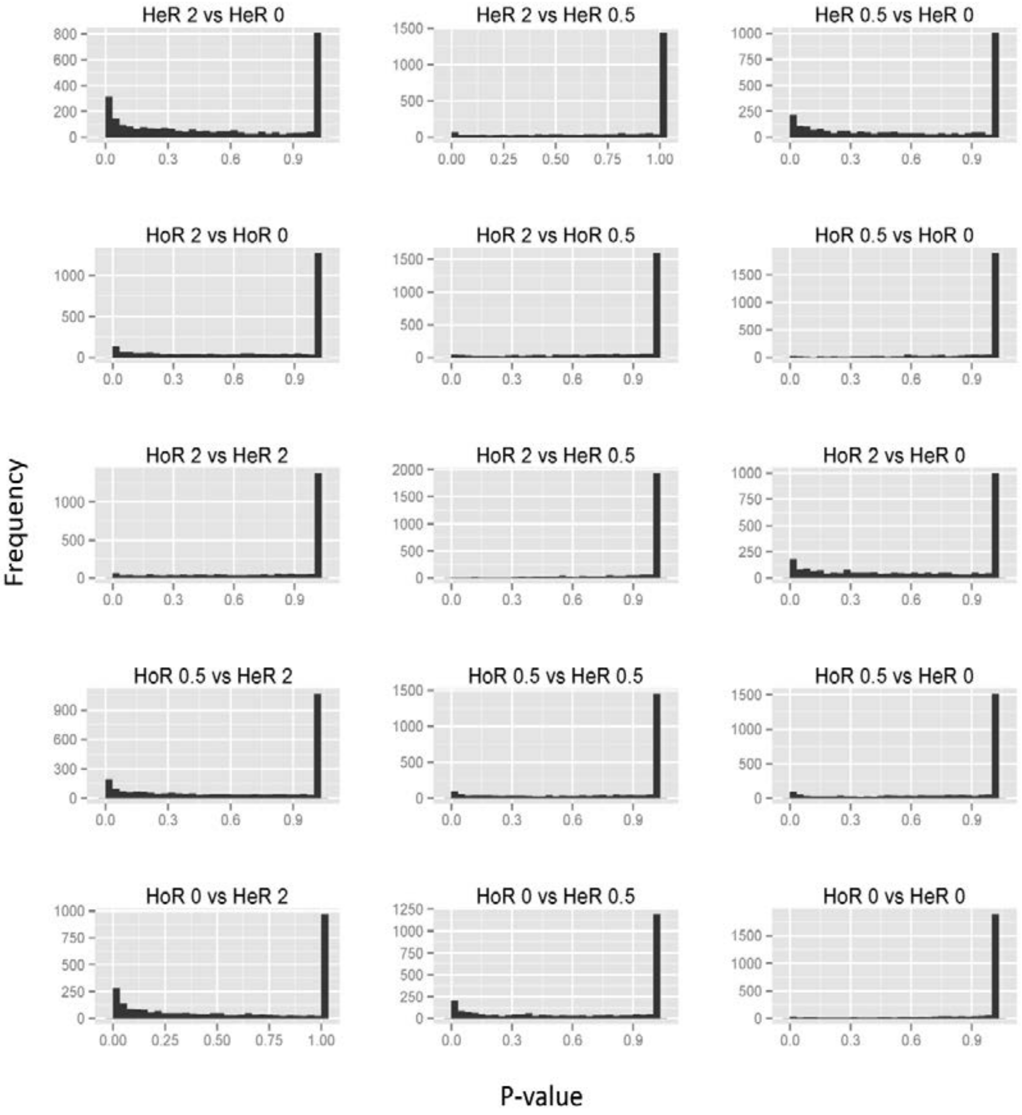


**Figure C3:** Histogram of p-values for each pairwise differential expression test in the low dose experiment

| **Gene symbol** | **Locus tag** | **Product** | **Mutation** | **Codon Change** | **Amino Acid Change/Nucleotide Change** | **Caller** | **SVTYPE** | **SVLEN** | **START** | **END** | **REF** | **ALT** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 317 | - | A | G |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 518090 | - | T | C |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 944692 | - | A | **T** |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1352167 | - | T | A |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1355744 | - | T | A |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1542521 | - | G | **A** |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1961241 | - | G | C |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1961354 | - | C | T |
| *serS* | SAUSA300_0009 | seryl-tRNA synthetase | Missense | agC/agA | p.Ser16Arg/c.48C>A | both | SNP | - | 12860 | - | C | A |
| - | SAUSA300_0375 | putative phosphoglycerate mutase family protein | Missense | gaA/gaT | p.Glu106Asp/c.318A>T | both | SNP | - | 425833 | - | A | T |
| - | SAUSA300_1232 | catalase | Missense | gCa/gTa | p.Ala173Val/c.518C>T | both | SNP | - | 1350514 | - | C | T |
| - | SAUSA300_1396 | phiSLT ORF151-like protein major tail protein | Missense | Aat/Tat | p.Asn29Tyr/c.85A>T | both | SNP | - | 1564562 | - | T | A |
| *splF* | SAUSA300_1753 | serine protease SplF | Missense | gGa/gCa | p.Gly11Ala/c.32G>C | both | SNP | - | 1939443 | - | C | G |
| - | SAUSA300_2324 | PTS system sucrose-specific IIBC component | Missense | tTa/tCa | p.Leu293Ser/c.878T>C | both | SNP | - | 2498396 | - | A | G |
| *narH* | SAUSA300_2342 | respiratory nitrate reductase, beta subunit | Missense | gCa/gTa | p.Ala178Val/c.533C>T | both | SNP | - | 2516156 | - | G | A |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | SAUSA300_0026 | rRNA large subunit methyltransferase | Synonymous | gcT/gcA | p.Ala156Ala/c.468T>A | both | SNP | - | 34179 | - | T | A |
| *sdrC* | SAUSA300_0546 | sdrC protein | Synonymous | tcT/tcA | p.Ser724Ser/c.2172T>A | both | SNP | - | 613318 | - | T | A |
| - | SAUSA300_1236 | hypothetical protein | Synonymous | taC/taT | p.Tyr81Tyr/c.243C>T | both | SNP | - | 1354028 | - | C | T |
| - | Intergenic | - | Intergenic | - | - | both | DEL | -1 | 517960 | 517961 | CT | C |
| - | Intergenic | - | Intergenic | - | - | both | DEL | -1 | 593858 | 593859 | AT | A |
| - | Intergenic | - | Intergenic | - | - | nucmer | DEL | -1 | 973710 | 973711 | AT | T |
| - | Intergenic | - | Intergenic | - | - | pindel | DEL | -121 | 2295770 | 2295891 | AGTCAAGCGCTCGCATACTGCTTTATTTTCAAAAAATCAAATGCTCATTTACAAAAGTAAACTCCGCTTTAATTTTTCTTAATGCATTGTCTGACAATCGCTTTCTTTAAAAAGAATAGATT | A |
| - | SAUSA300_0407 | superantigen-like protein | Frameshift | aatacagct/ | p.Asn108fs/c.324_328delTACAG | both | DEL | -5 | 460961 | 460966 | ATACAG | A |
| - | SAUSA300_0567 | hypothetical protein | Frameshift | ggc/ | p.Gly123fs/c.368delG | both | DEL | -1 | 641665 | 641666 | GC | G |
| - | SAUSA300_1810 | IS1181, transposase | Frameshift | aag/ | p.Lys440fs/c.1319delA | manual | DEL | -1 | 1994918 | 1994919 | AA | A |
| - | SAUSA300_1993 | Pfkb family kinase | Frameshift | act/actG | p.Thr255_Gly256fs/c.765_766insG | nucmer | INS | 1 | 2150261 | 2150261 | C | CC |

**Table C10:** Genetic background of USA300 used in the low dose experiment. Substitution notation is in Human Genome Variation Society (HGVS) format (http://www.hgvs.org/mutnomen/standards.html). The table is ordered by ‘Effect’, then ‘SVTYPE’ and then ‘Position’. The ‘Caller’ column indicates whether the variant was detected by Samtools, Nucmer, Pindel or or manually (‘Manual’ using the BAM file in IGV). In the case of SNPs, ‘both’ means the SNP was called by Nucmer and by samtools and in the case of indels it means the SNP was called by both Nucmer and Pindel. The amino acid changed in each codon is in capital letters. Two intergenic SNPs at positions 944,692 and 1,542,521 bp on the chromosome had the read depth allele frequency of Sample 1a_S1 (HeR0) added from the RNA (the frequency was one indicating that these SNPs were also fixed in Sample 1a_S1).

| Gene symbol | Locus tag | Product | Mutation | Codon Change | Amino Acid Change/Nucleotide Change | Caller | Position on chromosome | REF | ALT |
|---|---|---|---|---|---|---|---|---|---|
| - | Intergenic | - | Intergenic | - | - | samtools | 512748 | A | G |
| - | Intergenic | - | Intergenic | - | - | both | 514487 | C | T |
| - | Intergenic | - | Intergenic | - | - | samtools | 1997102 | A | T |
| - | Intergenic | - | Intergenic | - | - | samtools | 1997489 | G | C |
| - | Intergenic | - | Intergenic | - | - | samtools | 2699115 | G | T |
| *sdrE* | SAUSA300_0548 | sdrE protein | Synonymous | tcT/tcA | p.Ser1077Ser/c.3231T>A | both | 622126 | T | A |
| - | Intergenic | - | Intergenic | - | - | samtools | 1997484 | A | T |

**Table C11:** Seven SNPs that were polymorphic in all eighteen DNA samples in the low dose experiment. Substitution notation is in Human Genome Variation Society (HGVS) format (http://www.hgvs.org/mutnomen/standards.html). The amino acid changed in each codon is in capital letters. In the Caller column, 'both' means that the SNP was called by both nucmer and samtools.

**Table C12:** SNPs called in at least one sample in the low dose experiment. Table is online at https://figshare.com/s/c0370d2800fe73a007e1

**Table C13:** Log2 fold change and adjusted p-value produced by DeSeq2 for genes that were differentially expressed in at least one comparisons with HeR0. So HeR0.5 columns indicates HeR0.5 vs HeR0, HeR2 indicates HeR2 vs HeR0, HoR0 indicates HoR0 vs HeR0, HoR0.5 indicates HoR0.5 vs HeR0 and HoR2 indicates HoR2 vs HeR0. Gray shading in cells containing numbers indicates that the gene was differentially expressed in that comparison and shading in cells with gene names indicates that genes were adjacent to each other on the chromosome. Table is online at https://figshare.com/s/c0370d2800fe73a007e1

***Down-regulated genes***

| Locus Tag | Official Gene Symbol | Product | log2FoldChange | padj |
|---|---|---|---|---|
| SAUSA300_0142 | *phnE* | phosphonate ABC transporter permease | -2.908 | 0.071375 |
| SAUSA300_0895 | *oppB* | oligopeptide ABC transporter permease | -2.891 | 0.074495 |
| SAUSA300_0966 | *purE* | phosphoribosylaminoimidazole carboxylase catalytic subunit | -2.621 | 0.099783 |
| SAUSA300_0967 | *purK* | phosphoribosylaminoimidazole carboxylase ATPase subunit | -2.889 | 0.045017 |
| SAUSA300_0970 | *purQ* | phosphoribosylformylglycinamidine synthase I | -3.396 | 0.016586 |
| SAUSA300_0971 | *purL* | phosphoribosylformylglycinamidine synthase II | -3.634 | 0.009576 |
| SAUSA300_0972 | *purF* | amidophosphoribosyltransferase | -3.531 | 0.010269 |
| SAUSA300_0973 | *purM* | phosphoribosylaminoimidazole synthetase | -3.703 | 0.009576 |
| SAUSA300_0974 | *purN* | phosphoribosylglycinamide formyltransferase | -3.239 | 0.033011 |
| SAUSA300_0975 | *purH* | bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase | -3.235 | 0.016586 |
| SAUSA300_0976 | *purD* | phosphoribosylamine--glycine ligase | -3.297 | 0.016586 |
| SAUSA300_1053 | *flr* | FPRL1 inhibitory protein | -2.399 | 0.096997 |
| SAUSA300_1280 | *pstB* | phosphate transporter ATP-binding protein | -3.009 | 0.056377 |
| SAUSA300_1281 | *pstA* | phosphate ABC transporter permease PstA | -2.901 | 0.074495 |
| SAUSA300_1282 | *pstC* | phosphate ABC transporter permease PstC | -2.996 | 0.081148 |
| SAUSA300_1487 | - | replication initiation factor family protein | -3.097 | 0.051069 |
| SAUSA300_1919 | *scn* | staphylococcal complement inhibitor | -2.820 | 0.00025 |
| SAUSA300_2561 | *phoB* | alkaline phosphatase | -3.609 | 0.012108 |

***Up-regulated genes***

| Locus Tag | Official Gene Symbol | Product | log2FoldChange | padj |
|---|---|---|---|---|
| SAUSA300_0113 | *spa* | immunoglobulin G binding protein A | 4.425 | 1.71E-07 |
| SAUSA300_0278 | - | hypothetical protein | 2.852 | 0.045017 |
| SAUSA300_0 | *nusG* | transcription antitermination protein | 2.072 | 0.034265 |

| | | | | |
|---|---|---|---|---|
| 521 | | | | |
| SAUSA300_0684 | *fruB* | fructose 1-phosphate kinase | 2.771 | 0.094768 |
| SAUSA300_0964 | - | Chitinase-related protein | 2.261 | 0.056377 |
| SAUSA300_0993 | *pdhA* | pyruvate dehydrogenase E1 component, alpha subunit | 2.053 | 0.051069 |
| SAUSA300_1074 | *ftsL* | cell division protein | 2.569 | 0.081148 |
| SAUSA300_1330 | *ilvA* | threonine dehydratase | 3.234 | 0.012108 |
| SAUSA300_1622 | *tig* | trigger factor | 2.246 | 0.056377 |
| SAUSA300_1629 | *thrS* | threonyl-tRNA synthetase | 2.307 | 0.056377 |
| SAUSA300_1708 | *rot* | repressor of toxins | 2.175 | 0.058378 |
| SAUSA300_1982 | *groEL* | chaperonin GroEL | 2.157 | 0.07863 |
| SAUSA300_1983 | *groES* | co-chaperonin GroES | 2.609 | 0.016586 |
| SAUSA300_2150 | *lacE* | PTS system, lactose-specific IIBC component | 2.691 | 0.051069 |
| SAUSA300_2151 | *lacF* | PTS system, lactose-specific IIA component | 3.105 | 0.016427 |

**Table C14:** Log2 fold change and adjusted p-values of differentially expressed genes in the HoR2 vs HoR0 comparison

| ***Down-regulated genes*** | | | | |
|---|---|---|---|---|
| **Locus Tag** | **Official Gene Symbol** | **Product** | **log2FoldChange** | **padj** |
| SAUSA300_0409 | - | 50S ribosomal protein L25/general stress protein Ctc | -2.652 | 0.0493856 |
| SAUSA300_0971 | *purL* | phosphoribosylformylglycinamidine synthase II | -3.274 | 0.0021142 |
| SAUSA300_0972 | *purF* | amidophosphoribosyltransferase | -3.344 | 0.0021142 |
| SAUSA300_0973 | *purM* | phosphoribosylaminoimidazole synthetase | -3.352 | 0.0021142 |
| SAUSA300_0974 | *purN* | phosphoribosylglycinamide formyltransferase | -3.290 | 0.0021622 |
| SAUSA300_0975 | *purH* | bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase | -3.169 | 0.0021142 |
| SAUSA300_0976 | *purD* | phosphoribosylamine--glycine ligase | -3.177 | 0.0021142 |
| SAUSA300_1053 | - | formyl peptide receptor-like 1 inhibitory protein | -2.482 | 0.0043709 |
| SAUSA300_1055 | *efb* | fibrinogen-binding protein | -2.349 | 0.0071473 |

|  |  |  |  |  |
|---|---|---|---|---|
| SAUSA300_1056 | - | hypothetical protein | -2.434 | 0.0045343 |
| SAUSA300_1919 | *scn* | staphylococcal complement inhibitor | -2.854 | 1.81E-08 |
| ***Up-regulated genes*** |  |  |  |  |
| **Locus Tag** | **Official Gene Symbol** | **Product** | **log2FoldChange** | **padj** |
| SAUSA300_0113 | *spa* | immunoglobulin G binding protein A | 2.730 | 0.0021142 |
| SAUSA300_0225 | - | putative acyl-CoA acetyltransferase FadA | 3.204 | 0.0036288 |
| SAUSA300_0226 | - | 3-hydroxyacyl-CoA dehydrogenase | 3.173 | 0.0109714 |
| SAUSA300_0278 | - | ESAT-6-like protein | 2.656 | 0.0064794 |
| SAUSA300_0703 | *ltaS* | Lipoteichoic acid synthase | 1.706 | 0.0137291 |
| SAUSA300_0891 | *oppA* | oligopeptide ABC transporter substrate-binding protein | 2.155 | 0.0920496 |
| SAUSA300_0964 | - | Chitinase-related protein | 1.822 | 0.0684679 |
| SAUSA300_1606 | - | hypothetical protein | 2.172 | 0.0822155 |
| SAUSA300_1629 | *thrS* | threonyl-tRNA synthetase | 2.133 | 0.0069931 |
| SAUSA300_1982 | - | phi77 ORF002-like protein, phage minor structural protein | 1.798 | 0.0569837 |
| SAUSA300_1983 | - | phi77 ORF004-like protein phage tail component | 2.071 | 0.0165086 |
| SAUSA300_2059 | *atpG* | F0F1 ATP synthase subunit gamma | 1.808 | 0.0429941 |
| SAUSA300_2195 | *rpsQ* | 30S ribosomal protein S17 | 2.030 | 0.0822155 |
| SAUSA300_2537 | - | L-lactate dehydrogenase | 2.473 | 0.0021142 |

**Table C15:** Log2 fold change and adjusted p-values of differentially expressed genes in the HoR0.5 vs HoR0 comparison

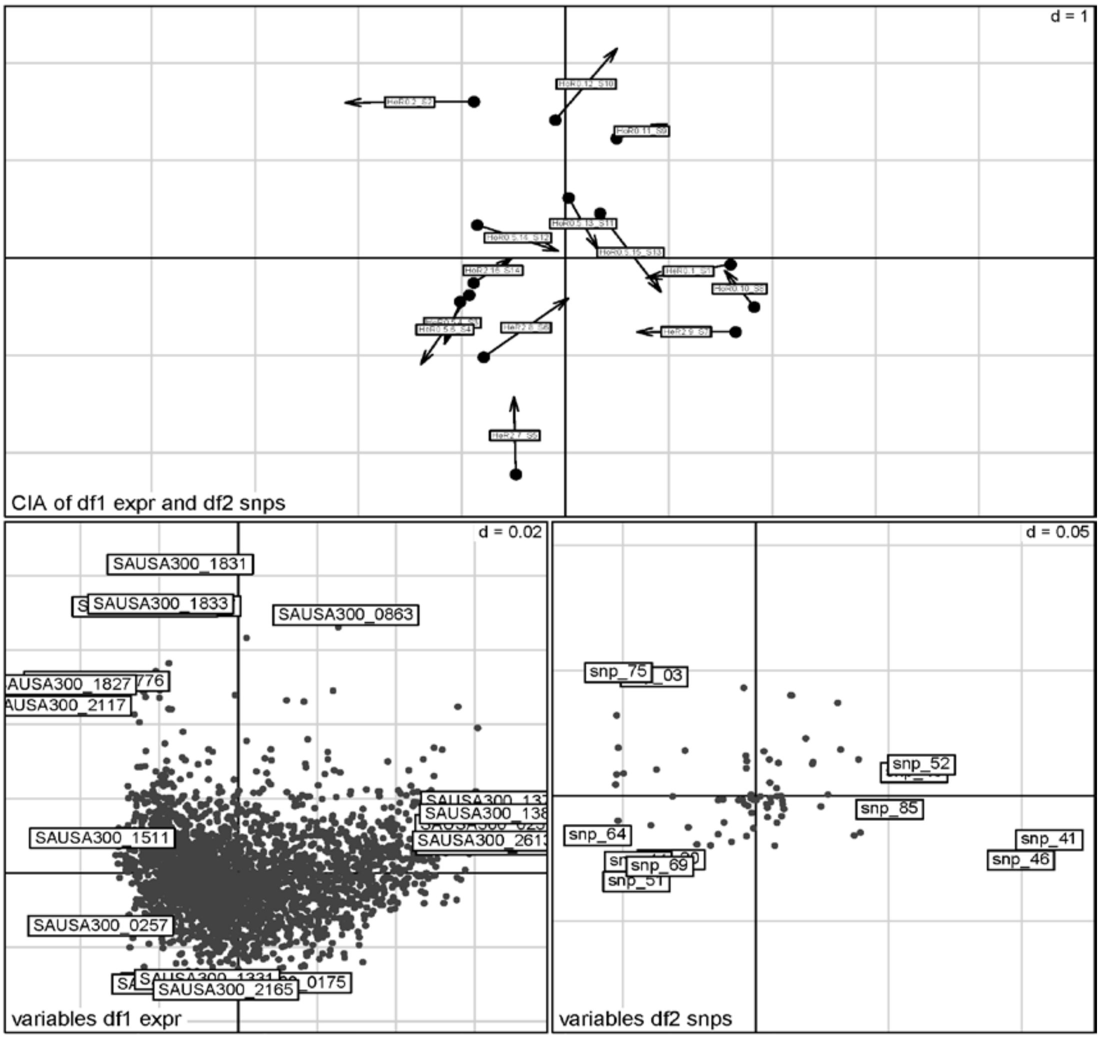


**Figure C4:** Co-inertia of the full dataset of 95 SNPs (excluding one on a plasmid) and expression of 2,367 genes. The arrows in the top part of the figure indicate the degree of relationship between the expression data and the SNP data in each sample. Long arrows signify low concordance between the datasets. The bottom left plot represents the genes. Genes that contributed most to the similarity between the SNP and expression datasets are the furthest from the centre. The bottom right plot shows the SNP samples. SNPs and genes which are changed in the same direction away from the origin are changing in a similar way. The insert plot shows the results of the RV value Monte Carlo permutation test with the RV value of the co-inertia analysis shown as the line with the diamond on top and the frequency of RV values calculated by the permutation test shown in the histogram.

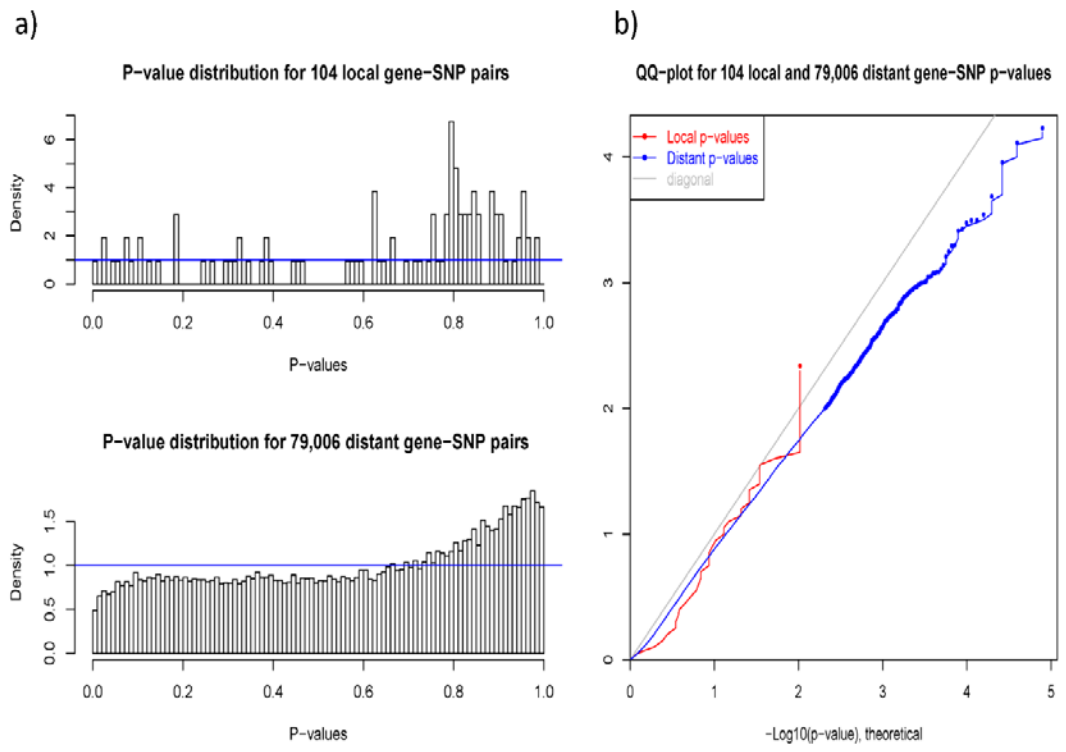


**Figure C5:** EQTL results. a)distribution of p-values for the local and distant SNP gene pairs.. b) QQ-plot for the local and distant gene pairs.

| Comparison (DNA vs RNA) | RV value | Simulated P value |
|---|---|---|
| **All samples** | 0.33 | 0.86 |
| **All HeR samples** | 0.48 | 0.78 |
| **All HoR samples** | 0.44 | 0.86 |
| **HeR0 samples** | 1 | 1 |
| **HeR0.5 samples** | 1 | 1 |
| **HeR2 samples** | 0.68 | 0.64 |
| **HoR0 samples** | 0.87 | 0.15 |
| **HoR0.5 samples** | 0.9 | 0.13 |
| ***rpoB* SNP samples** | 0.62 | 0.71 |

**Table C16:** RV value, simulated p-value based on 100 replicates of co-inertia analysis using a monte carlo permutation test. Significance threshold is 5%. HoR2 is not included here due to only having one RNASeq sample. '*rpoB* SNP samples' means the four samples (one HeR0.5 and three HeR2 samples) that have either the fixed or polymorphic *rpoB* SNP.

| Gene symbol | Locus tag | Product | Effect | Codon Change | Amino Acid Change/Nucleotide Change | Caller | SVTYPE | SVLEN | START | END | REF | ALT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 317 | - | A | G |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 518090 | - | T | C |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 556291 | - | T | C |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 944692 | - | A | T |
| - | Intergenic | - | I Intergeni | - | - | both | SNP | - | 1352167 | - | T | A |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1355744 | - | T | A |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1542521 | - | G | A |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1961241 | - | G | C |
| - | Intergenic | - | Intergenic | - | - | both | SNP | - | 1961354 | - | C | T |
| *serS* | SAUSA300_0009 | seryl-tRNA synthetase | Missense | agC/agA | p.Ser16Arg/c.48C>A | nucmer | SNP | - | 12860 | - | C | A |
| - | SAUSA300_0375 | putative phosphoglycerate mutase family protein | Missense | gaA/gaT | p.Glu106Asp/c.318A>T | nucmer | SNP | - | 425833 | - | A | T |
| - | SAUSA300_1232 | catalase | Missense | gCa/gTa | p.Ala173Val/c.518C>T | both | SNP | - | 1350514 | - | C | T |
| - | SAUSA300_1396 | phiSLT ORF151-like | Missense | Aat/Tat | p.Asn29Tyr/c.85A>T | both | SNP | - | 1564562 | - | T | A |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | protein major tail protein | | | | | | | | | | |
| *splF* | SAUSA300_1753 | serine protease SplF | Missense | gGa/gCa | p.Gly11Ala/c.32G>C | both | SNP | - | 1939443 | - | C | G |
| - | SAUSA300_2324 | PTS system sucrose-specific IIBC component | Missense | tTa/tCa | p.Leu293Ser/c.878T>C | both | SNP | - | 2498396 | - | A | G |
| *narH* | SAUSA300_2342 | respiratory nitrate reductase, beta subunit | Missense | gCa/gTa | p.Ala178Val/c.533C>T | both | SNP | - | 2516156 | - | G | A |
| - | SAUSA300_1236 | hypothetical protein | Synonymous | taC/taT | p.Tyr81Tyr/c.243C>T | both | SNP | - | 1354028 | - | C | T |
| - | Intergenic | - | Intergenic | - | - | both | DEL | -1 | 517960 | 517961 | CT | T |
| - | Intergenic | - | Intergenic | - | - | both | DEL | -1 | 593858 | 593859 | AT | T |
| - | Intergenic | - | Intergenic | - | - | nucmer | DEL | -1 | 973710 | 973711 | AA | A |
| - | Intergenic | - | Intergenic | - | - | pindel | DEL | -121 | 2295770 | 2295891 | AGTCAAGCGCTCGCAT<br>ACTGCTTTA<br>TTTTCAAAAAATCAAAT<br>GCTCATTTAC<br>AAAAGTAAAC<br>TCCGCTTTAATTTTTCTTA<br>ATGCAT | A |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | TGTCTGACAATCGCTTTCTTTAAAAAGAATAGATT | |
| - | SAUSA300_0407 | superantigen-like protein | Frameshift | aatacagct/ | p.Asn108fs/c.324_328delTACAG | both | DEL | -5 | 460961 | 460966 | ATACAG | A |
| - | SAUSA300_0567 | hypothetical protein | Frameshift | ggc/ | p.Gly123fs/c.368delG | both | DEL | -1 | 641665 | 641666 | GC | C |
| - | SAUSA300_1810 | IS1181, transposase | Frameshift | aag/ | p.Lys440fs/c.1319delA | nucmer | DEL | -1 | 1994918 | 1994919 | AA | A |
| - | SAUSA300_1993 | Pfkb family kinase | Frameshift | act/actG | p.Thr255_Gly256fs/c.765_766insG | nucmer | INS | 1 | 2150261 | 2150261 | C | CC |

**Table C17**: The genetic background of the USA300 sample in the bioreactor experiment. Variants were called using the USA300_FPR3757chromosome (NC_007793). Substitution notation is in HGVS format (http://www.hgvs.org/mutnomen/standards.html). The table is ordered by 'Effect', then 'SVTYPE' and then 'Position'. The 'Caller' column indicates whether the SNP was called by samtools or nucmer originally or both. The amino acid changed in each codon is in capital letters.

| Type | Caller | Reference | Position | REF | ALT |
|---|---|---|---|---|---|
| Intergenic | samtools | NC_007793 | 512748 | A | G |
| Intergenic | both | NC_007793 | 514487 | C | T |
| Intergenic | both | NC_007793 | 558118 | T | C |
| Intergenic | both | NC_007793 | 1084103 | T | G |
| Intergenic | both | NC_007793 | 1997102 | A | T |
| Intergenic | nucmer | NC_007793 | 1997387 | A | G |
| Intergenic | both | NC_007793 | 1997484 | A | T |
| Intergenic | both | NC_007793 | 1997489 | G | C |

**Table C18:** SNPs that were polymorphic in all samples of the bioreactor experiment

**Table C19:** 84 SNPs were present in at least one sample of the 16 DNA samples in the bioreactor dose experiment. Values in cells are the read depth allele frequencies (RDAF) of the alternative allele of each SNP with the exception of zero values which means no SNP was found at that particular locus in that sample. Table is online at https://figshare.com/s/c0370d2800fe73a007e1

| - | **Gene symbol** | **-** | ***clfA*** | ***clfB*** |
|---|---|---|---|---|
| | **Locus tag** | SAUSA300_pUSA030002 | SAUSA300_0772 | SAUSA300_2565 |
| | **Product** | IS431mec-like transposase | clumping factor A | clumping factor B |
| | **Mutation** | Synonymous | Synonymous | Synonymous |
| | **Codon Change** | cgT/cgC | tcG/tcA | tcG/tcA |
| | **Amino Acid Effect** | R204R | S627S | S812S |
| | **DNA effect** | 612T>C | 1881G>A | 2436G>A |
| | **Caller** | nucmer | nucmer | nucmer |
| | **Position** | 1790 | 861,028 | 2,774,604 |
| | **REF** | T | G | C |
| | **ALT** | C | A | T |
| **HoR0** | **8A_S8** | 0.98 | 0 | 0 |
| **HoR0** | **9A_S9** | 0 | 0.91 | 0.99 |
| **HoR0** | **10A_S10** | 0 | 0 | 0.97 |

**Table C20:** Fixed synonymous mutations at three genes in the bioreactor samples. The IS431mec-like transposase gene is on the pUSA03 plasmid**.**

| **Gene** | ***gdpP*** | **-** | **intergenic & *apt*** | **-** |
|---|---|---|---|---|
| **Locus tag** | SAUSA300_0014 | Intergenic-upstream of *sarA* (sausa300_0605) | Intergenic & SAUSA300_1591 | SAUSA300_1968 |
| **Product** | GGDEF domain protein containing phosphodiesterase activity | - | adenine phosphoribosyltransferase | putative phage transcriptional regulator |
| **Mutation** | Frameshift | Intergenic | stop lost and inframe deletion (most of deletion is intergenic, some at end of *apt* gene) | Frameshift |
| **Codon Change** | atg/ | - | Ggcggtatcgtagtaggtattgcatttataattgaattgaaatat<br>ttaaatggtattgaaaaaattaaagattacg<br>atgttatgagtttaatctcatacgacgaataa/ | ttt/ |
| **Amino Acid Change/Nucleotide Change** | p.Met458fs/c.1372_1373delAT | - | p.Gly138_Ter173del/c.412_814delGGCGG<br>TATCGTAGTAGGTA<br>TTGCATTTATA<br>ATTGAATTGAAATATTTAAATGGTATTGAAAAAATT<br>AAAGATTACGATGTTAT<br>GAGTTT<br>AATCTCATACGAC<br>GAATAATAAATAATATAATTTTAT<br>CAAATGAAATCCTTCATCAAATGTATAAGAACC<br>AATGACTTAATTAAAA<br>AAGTTGTTTA<br>AGTTTT | p.Phe36fs/c.106delT |

| | | | | |
|---|---|---|---|---|
| | | | CTTAACATGAGATGTTAGGATTTTTTATTTACTGAAAATGTTAGATG<br>ATTGAGCATTATA<br>CCTTAATAA<br>CATCGTTTATTTATTTCATAAATTGTAGTATCATAGAACTAATAT<br>TTAAAAAATGA<br>AACAGTAGATTTAGGTCGAATT<br>TTTGTAAAAGTTTTAAAAGTAGGAATAGTATACAAATTAAAC<br>TCGCTCAAGTAAAATTAATATTA | |
| **Amino Acid Length** | 655 | - | 172 | 149 |
| **Caller** | nucmer | pindel | pindel | nucmer |
| **SVTYPE** | DEL | DEL | DEL | DEL |
| **SVLEN** | -2 | -125 | -403 | -1 |
| **START** | 19715 | 678894 | 1743267 | 2123182 |
| **END** | 19717 | 679019 | 1743670 | 2123183 |
| **REF** | TAT | CTATTTGATGCATCTT<br>GCTCGATAC<br>ATTTG<br>CCCGATAAT<br>ATATTGAT | GTAATATTAATT<br>TTACTTGAGCGAGTTTAATTT<br>GTATACTATTCCTACTT<br>TTAAAACT<br>TTTACAAAAATTCGACCTAAATCTACT | AA |

| | | | | |
|---|---|---|---|---|
| | | ATCTAAT<br>CTTTATTTAT<br>TATAGATA<br>TGTTAGTCATAA<br>TTTTGCATTA<br>AATAAGTTTT<br>ATTAA<br>ATATATTTA<br>ATGCTCTA | GTTTCATTTTTTAAATATTAGTTCTATGATACTACAAT<br>TTATGAAATAAATAAACGATG<br>TTATTAAGGTATAATGCTCA<br>ATCATCTAACATT<br>TTCAGTAAATAAAAAATCCTAACATCTCA<br>TGTTAAGAAAACTTAAACAACTTTTTTAATTAAGTCA<br>TTGGTTCTTATACATTTGATGAAGGATTTC<br>ATTTGATAAAATTATATTATT<br>TATTATTCGTCGTATGAGATTAAACTCATAA<br>CATCGTAATCTTTAATTTTT<br>TCAATACCATTTAAATATTTCAATTCAATTATAAATGCA<br>ATACCTACTACGATACCGCC | |
| **ALT** | T | C | G | A |
| **Sample(s)** | B1_S16 | 8A_S8 | 2A_S2,3A_S3,4A_S4,5A_S5,6A_S6,9A_S9,<br>11A_S11,13A_S13,14A_S14,15A_S15 | 8A_S8 |

**Table C21:** Fixed deletions in at least one sample in the bioreactor experiment. SVLEN indicates the length of the deletion e.g. -1 denotes a one base pair deletion.

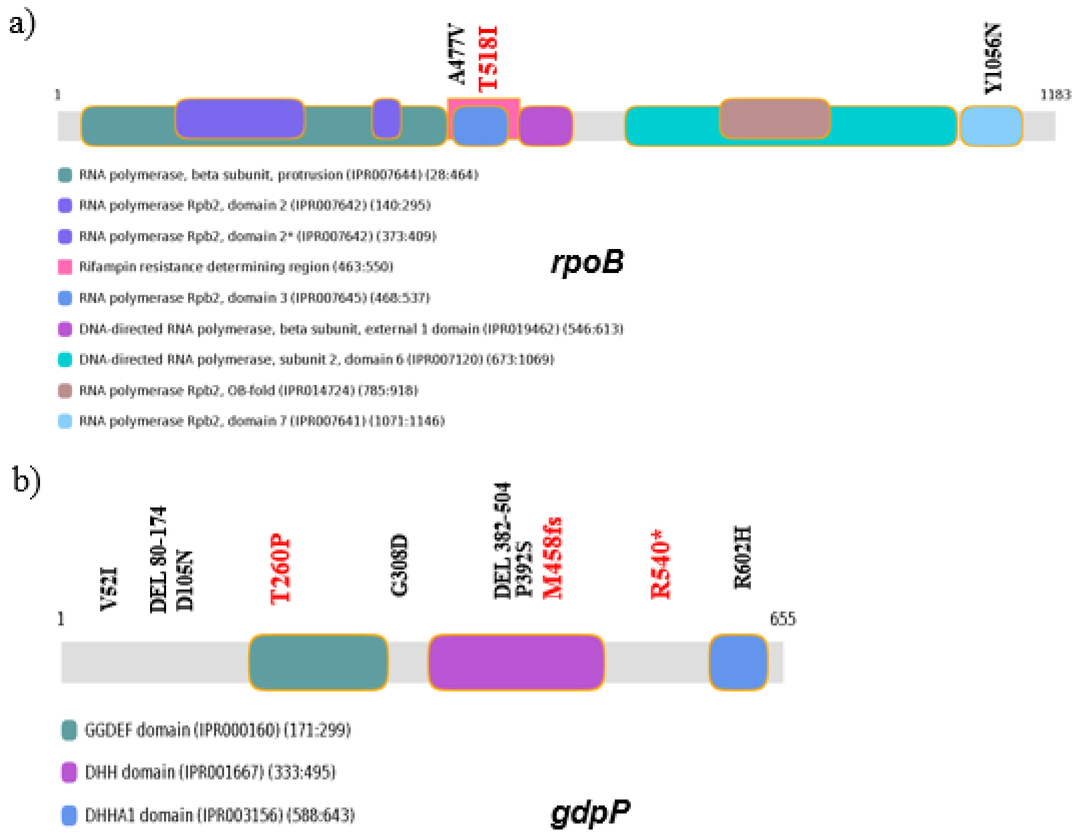

a)
A477V
T518I
Y1056N
1
1183
RNA polymerase, beta subunit, protrusion (IPR007644) (28:464)
RNA polymerase Rpb2, domain 2 (IPR007642) (140:295)
RNA polymerase Rpb2, domain 2* (IPR007642) (373:409)
Rifampin resistance determining region (463:550)
RNA polymerase Rpb2, domain 3 (IPR007645) (468:537)
DNA-directed RNA polymerase, beta subunit, external 1 domain (IPR019462) (546:613)
DNA-directed RNA polymerase, subunit 2, domain 6 (IPR007120) (673:1069)
RNA polymerase Rpb2, OB-fold (IPR014724) (785:918)
RNA polymerase Rpb2, domain 7 (IPR007641) (1071:1146)
rpoB
b)
V52I
DEL 80-174
D105N
T260P
G308D
DEL 382-504
P392S
M458fs
R540*
R602H
1
655
GGDEF domain (IPR000160) (171:299)
DHH domain (IPR001667) (333:495)
DHHA1 domain (IPR003156) (588:643)
gdpP

**Figure C6:** Mutations in *gdpP* and *rpoB* gene domains

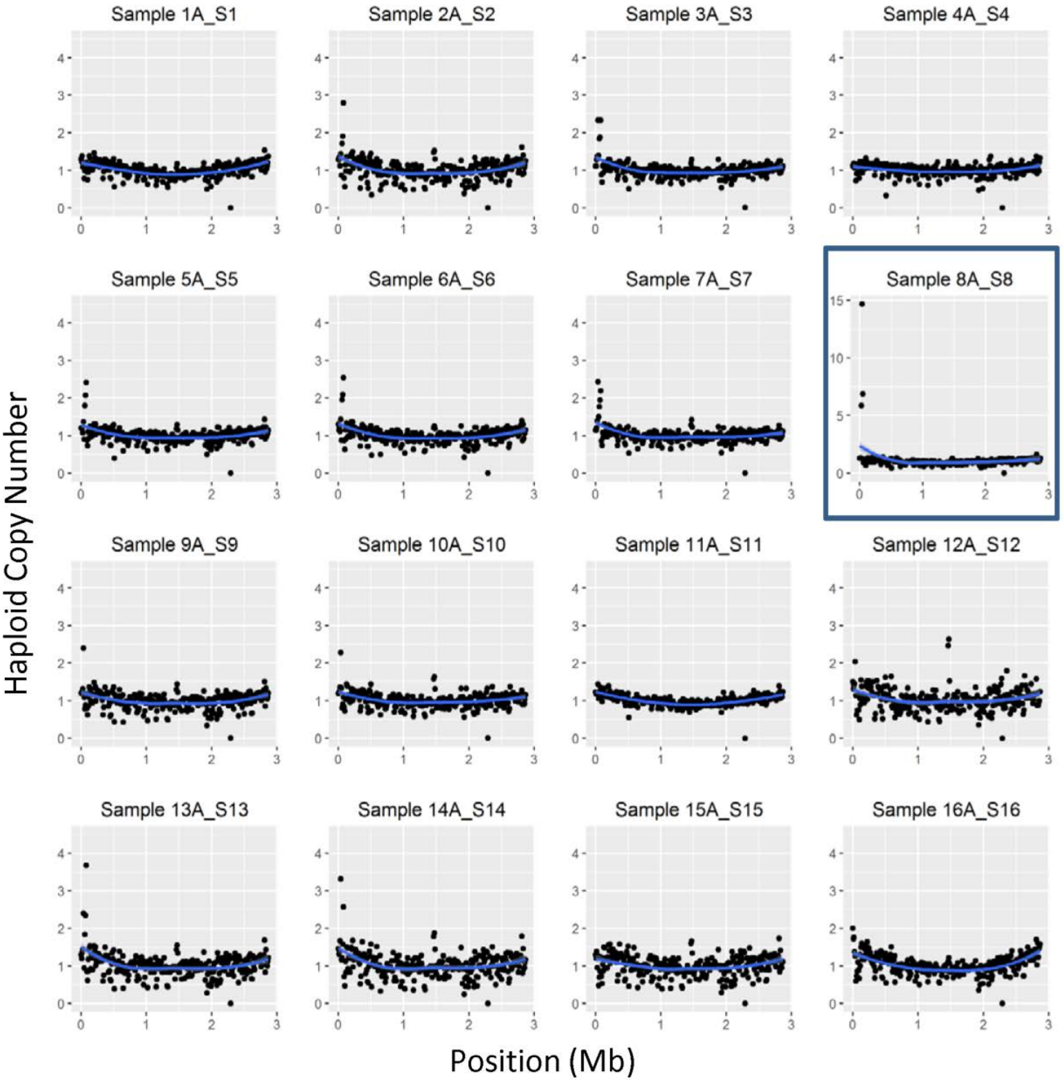


**Figure C7:** Copy number of 10 kb non-overlapping windows across the chromosome of each sample in the bioreactor experiment. All plots have maximum copy number of four with the exception of sample 8A_S8 (inside blue square) which has a maximum copy number of 15 for one 10 kb section of the SCC*mec*IV element (which is represented by the three dots with increased copy number at the 5' edge of the sample). The blue lines are lowess smoothers applied to the data points (black) with the grey area surrounding each blue line showing the standard error.

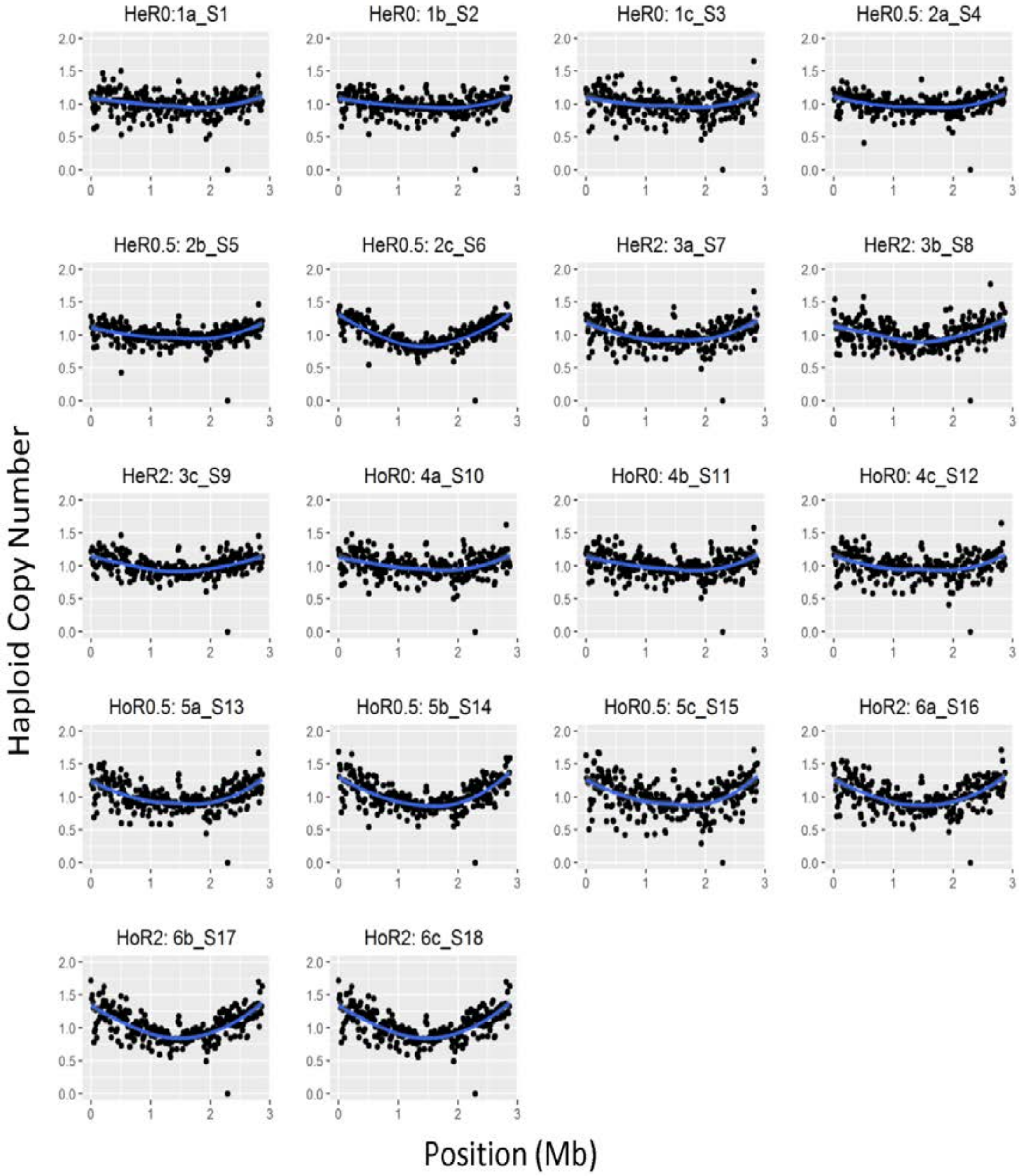


**Figure C8:** Copy number of 10 kb non-overlapping windows across the chromosome of each sample in the low dose experiment. The blue lines are lowess smoothers applied to the data points (black) with the grey area surrounding each blue line showing the standard error.

| Sample | Interval Start | Interval End | Interval median | Chromosomal median | Rounded Copy Number | Annotation |
|---|---|---|---|---|---|---|
| 2A_S2 | 80001 | 90000 | 282 | 101 | 3 | part of ACME |
| 3A_S3 | 40001 | 50000 | 212 | 91 | 2 | part of SCC*mec*IV |
| 3A_S3 | 80001 | 90000 | 212 | 91 | 2 | part of ACME |
| 5A_S5 | 70001 | 80000 | 203 | 98 | 2 | part of ACME |
| 5A_S5 | 80001 | 90000 | 235 | 98 | 2 | part of ACME |
| 6A_S6 | 70001 | 80000 | 167 | 80 | 2 | part of ACME |
| 6A_S6 | 80001 | 90000 | 203 | 80 | 3 | part of ACME |
| 7A_S7 | 40001 | 50000 | 422 | 174 | 2 | part of SCC*mec*IV |
| 7A_S7 | 80001 | 90000 | 381 | 174 | 2 | part of ACME |
| 8A_S8 | 30001 | 40000 | 968 | 165 | 6 | full SCC*mec*IV |
| 8A_S8 | 40001 | 50000 | 2422.5 | 165 | 15 | |
| 8A_S8 | 50001 | 60000 | 1132 | 165 | 7 | |
| 9A_S9 | 40001 | 50000 | 438 | 183 | 2 | part of SCC*mec*IV |
| 10A_S10 | 40001 | 50000 | 396 | 174 | 2 | part of SCC*mec*IV |
| 12A_S12 | 40001 | 50000 | 276 | 135 | 2 | part of SCC*mec*IV |
| 12A_S12 | 1460001 | 1470000 | 332 | 135 | 2 | |
| 12A_S12 | 1470001 | 1480000 | 356 | 135 | 3 | |
| 13A_S13 | 40001 | 50000 | 378.5 | 158 | 2 | part of SCC*mec*IV |
| 13A_S13 | 70001 | 80000 | 381.5 | 158 | 2 | part of ACME |
| 13A_S13 | 80001 | 90000 | 581 | 158 | 4 | part of ACME |
| 14A_S14 | 40001 | 50000 | 469.5 | 142 | 3 | part of SCC*mec*IV |
| 14A_S14 | 80001 | 90000 | 365 | 142 | 3 | part of ACME |
| B1_S16 | 1 | 10000 | 273 | 136 | 2 | Near Origin of replication |

**Table C22:** Copy number and annotation results from an examination of read depth coverage in 10 kb non-overlapping windows across the chromosome of each sample in the bioreactor experiment. Grey shading delineates either adjacent 10 kb loci or the next sample in the table.

| Sample | Interval Start | Interval End | Interval median | Chromosomal median | Rounded copy number | Annotation |
|---|---|---|---|---|---|---|
| 8A_S8 | 25001 | 50000 | 1238 | 165 | 8 | SCC*mec*IV |
| 12A_S12 | 1450001 | 1475000 | 289 | 135 | 2 | |

**Table C23:** Results of examining read depth coverage 25 kb non-overlapping windows across the chromosome of each sample in the bioreactor experiment.

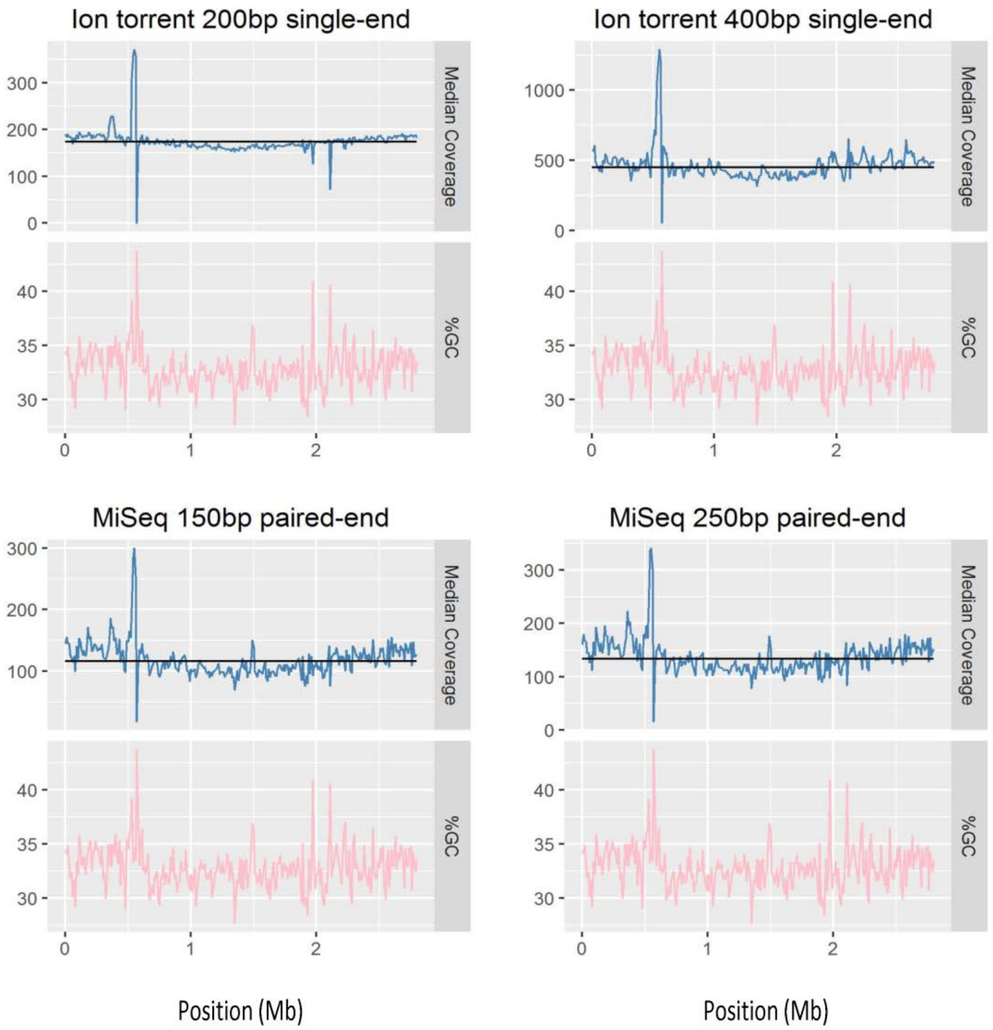


**Figure C9:** An amplification of a 40 kb locus (approximately 530,000 - 570,000 bp) on *S. aureus* COL visible in each of four sets of reads (sequenced from the same cell culture) mapped to the genome. The amplification had two to three copies.

# Bibliography

[1] Bloomfield M 1897 Hymns of the Atharva-veda *Sacred Books of the East* ed M FM (Oxford: Clarendon Press) p Hymns 1.23, 1.24, 11.31

[2] Galperin M Y and Koonin E V 1999 Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol.* **10** 571–8

[3] Ventola C L 2015 The antibiotic resistance crisis: part 1: causes and threats. *P T A peer-reviewed J. Formul. Manag.* **40** 277–83

[4] Bos K I, Schuenemann V J, Golding G B, Burbano H a, Waglechner N, Coombes B K, McPhee J B, DeWitte S N, Meyer M, Schmedes S, Wood J, Earn D J D, Herring D A, Bauer P, Poinar H N and Krause J 2011 A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478** 506–10

[5] Simon-Lorière E, Faye O, Faye O, Koivogui L, Nfaly M, Keita S, Thiberge J-M, Diancourt L, Bouchier C, Vandenbogaert M, Caro V, Fall G, Buchmann J P, Matranga C, Sabeti P C, Manuguerra J-C, Holmes E C and Amadou A.Sall 2015 Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic *Nature* **524** 102–4

[6] Sanger F, Nicklen S and Coulson A R 1977 DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74** 5463–7

[7] Sanger F and Coulson A R 1975 A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94** 441–8

[8] Sanger F, Coulson a R, Friedmann T, Air G M, Barrell B G, Brown N L, Fiddes J C, Hutchison C a, Slocombe P M and Smith M 1978 The nucleotide sequence of bacteriophage phiX174. *J. Mol. Biol.* **125** 225–46

[9] Fleischmann R D, Adams M D, White O, Clayton R A, Kirkness E F, Kerlavage A R, Bult C J, Tomb J F, Dougherty B A, Merrick J M, Mckenney K, Sutton G, Fitzhugh W, Fields C, Gocayne J D, Scott J, Shirley R, Liu L I, Glodek A, Kelley J M, Weidman J F, Phillips C A, Spriggs T, Hedblom E, Cotton M D, Utterback T R, Hanna M C, Nguyen D T, Saudek D M, Brandon R C, Fine L D, Fritchman J L, Fuhrmann J L, Geoghagen N S M, Gnehm C L, Mcdonald L A, Small K V, Fraser C M, Smith H O and Venter J C 1995 Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd *Science (80-. ).* **269** 496–512

[10] Lander E S, Linton L M, Birren B, Nusbaum C, Zody M C, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov J P, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin J C, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston R H, Wilson R K, Hillier L W, McPherson J D, Marra M A, Mardis E R, Fulton L A, Chinwalla A T, Pepin K H, Gish W R, Chissoe S L, Wendl M C, Delehaunty K D, Miner T L, Delehaunty A, Kramer J B, Cook L L, Fulton R S, Johnson D L, Minx P J, Clifton S W, Hawkins T, Branscomb E, Predki P, Richardson P,

Wenning S, Slezak T, Doggett N, Cheng J-F, Olsen A, Lucas S, Elkin C, et al 2001 Initial sequencing and analysis of the human genome *Nature* **409** 860–921

[11] Venter J C, Adams M D, Myers E W, Li P W, Mural R J, Sutton G G, Smith H O, Yandell M, Evans C A, Holt R A, Gocayne J D, Amanatides P, Ballew R M, Huson D H, Wortman J R, Zhang Q, Kodira C D, Zheng X H, Chen L, Skupski M, Subramanian G, Thomas P D, Zhang J, Gabor Miklos G L, Nelson C, Broder S, Clark A G, Nadeau J, McKusick V A, Zinder N, Levine A J, Roberts R J, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian A E, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman T J, Higgins M E, Ji R R, Ke Z, Ketchum K A, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov G V, Milshina N, Moore H M, Naik A K, Narayan V A, Neelam B, Nusskern D, Rusch D B, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, et al 2001 The sequence of the human genome. *Science* **291** 1304–51

[12] Mitra R D, Shendure J, Olejnik J, Krzymanska-Olejnik E and Church G M 2003 Fluorescent in situ sequencing on polymerase colonies *Anal. Biochem.* **320** 55–65

[13] Turcatti G, Romieu A, Fedurco M and Tairi A P 2008 A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis *Nucleic Acids Res.* **36** e25

[14] Shendure J, Porreca G J, Reppas N B, Lin X, McCutcheon J P, Rosenbaum A M, Wang M D, Zhang K, Mitra R D and Church G M 2005 Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309** 1728–32

[15] Zhao W, He X, Hoadley K A, Parker J S, Hayes D N and Perou C M 2014 Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling *BMC Genomics* **15** 419

[16] Jain M, Fiddes I T, Miga K H, Olsen H E, Paten B and Akeson M 2015 Improved data analysis for the MinION nanopore sequencer *Nat. Methods* **12** 351–6

[17] Bayley H 2006 Sequencing single molecules of DNA *Curr. Opin. Chem. Biol.* **10** 628–37

[18] Deamer D, Akeson M and Branton D 2016 Three decades of nanopore sequencing *Nat. Biotechnol.* **34** 518–24

[19] Laver T, Harrison J, O'Neill P A, Moore K, Farbos A, Paszkiewicz K and Studholme D J 2015 Assessing the performance of the Oxford Nanopore Technologies MinION *Biomol. Detect. Quantif.* **3** 1–8

[20] Ewing B and Green P 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities *Genome Res.* **8** 186–94

[21] Kircher M, Heyn P and Kelso J 2011 Addressing challenges in the production and analysis of illumina sequencing data *BMC Genomics* **12** 382

[22] Kircher M, Stenzel U and Kelso J 2009 Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* **10** R83

[23] Bolger A M, Lohse M and Usadel B 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30** 2114-20

[24] Del Fabbro C, Scalabrin S, Morgante M and Giorgi F M 2013 An extensive evaluation of read trimming effects on illumina NGS data analysis *PLoS One* **8**

[25] Kelley D R, Schatz M C and Salzberg S L 2010 Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11** R116

[26] Nikolenko S I, Korobeynikov A I and Alekseyev M A 2013 BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14 Suppl 1** S7

[27] Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov A S, Lesin V M, Nikolenko S I, Pham S, Prjibelski A D, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev M a. and Pevzner P a. 2012 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing *J. Comput. Biol.* **19** 455–77

[28] Magoč T and Salzberg S L 2011 FLASH: Fast length adjustment of short reads to improve genome assemblies *Bioinformatics* **27** 2957–63

[29] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden T L 2009 BLAST+: architecture and applications. *BMC Bioinformatics* **10** 421

[30] Kumar S, Jones M, Koutsovoulos G, Clarke M and Blaxter M 2013 Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **4** 237

[31] Denisov G, Walenz B, Halpern A L, Miller J, Axelrod N, Levy S and Sutton G 2008 Consensus generation and variant detection by Celera Assembler *Bioinformatics* **24** 1035–40

[32] Flicek P and Birney E 2009 Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* **6** S6–12

[33] Coughlan S, Barreira S, Seoighe C and Downing T 2014 Genome-wide Variation Discovery using Sequence Assembly, Mapping and Population Wide Analysis. In Bishop, O.Z. ed. *Bioinforma. data Anal. Microbiol. Caister Acad. Pres* 51-80

[34] Zerbino D R and Birney E 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18** 821–9

[35] Butler J, MacCallum I, Kleber M, Shlyakhter I A, Belmonte M K, Lander E S, Nusbaum C and Jaffe D B 2008 ALLPATHS: De novo assembly of whole-genome shotgun microreads *Genome Res.* **18** 810–20

[36] Simpson J T, Wong K, Jackman S D, Schein J E, Jones S J M and Birol I 2009 ABySS: A parallel assembler for short read sequence data *Genome Res.* **19** 1117–23

[37] Compeau P E C, Pevzner P A and Tesler G 2011 How to apply de Bruijn graphs to genome assembly *Nat. Biotechnol.* **29** 987–91

[38] Berger B, Peng J and Singh M 2013 Computational solutions for omics data *Nat. Rev. Genet.* **14** 333–46

[39] Hunt M, Newbold C, Berriman M and Otto T D 2014 A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.* **15** R42

[40] Boetzer M, Henkel C V, Jansen H J, Butler D and Pirovano W 2011 Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27** 578–9

[41] Langmead B 2010 Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinform.* John Wiley & Sons **32** 11.7.1-14

[42] Antipov D, Korobeynikov A, McLean J S and Pevzner P A 2016 HybridSPAdes: An algorithm for hybrid assembly of short and long reads *Bioinformatics* **32** 1009–15

[43] Deshpande V, Fung E D K, Pham S and Bafna V 2013 Cerulean: A hybrid assembly using high throughput short and long reads *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8126** LNBI 349–63

[44] Ye C, Hill C, Wu S, Ruan J, Zhanshan and Ma 2016 DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies *Scientific Reports* **6**

[45] English A C, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny D M, Reid J G, Worley K C and Gibbs R A 2012 Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology *PLoS One* **7**

[46] Bashir A, Klammer A a, Robins W P, Chin C-S, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdovino M, Mollova E, Luong K, Lin S, LaMay B, Joshi A, Rowe L, Frace M, Tarr C L, Turnsek M, Davis B M, Kasarskis A, Mekalanos J J, Waldor M K and Schadt E E 2012 A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30** 701–7

[47] Pop M 2009 Genome assembly reborn: Recent computational challenges *Brief. Bioinform.* **10** 354–66

[48] Goldberg S M D, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz S A, Lauro F M, Li K, Rogers Y-H, Strausberg R, Sutton G, Tallon L, Thomas T, Venter E, Frazier M and Venter J C 2006 A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes *Proc. Natl. Acad. Sci.* **103** 11240–5

[49] Swain M T, Tsai I J, Assefa S A, Newbold C, Berriman M and Otto T D 2012 A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat. Protoc.* **7** 1260–84

[50] Tsai I J, Otto T D and Berriman M 2010 Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11** R41

[51] Boetzer M and Pirovano W 2012 Toward almost closed genomes with GapFiller. *Genome Biol.* **13** R56

[52] Otto T D, Sanders M, Berriman M and Newbold C 2010 Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26** 1704–7

[53] Ning Z, Cox a J and Mullikin J C 2001 SSAHA: a fast search method for large DNA databases. *Genome Res.* **11** 1725–9

[54] Manske H M and Kwiatkowski D P 2009 SNP-o-matic. *Bioinformatics* **25** 2434–5

[55] Carver T J, Rutherford K M, Berriman M, Rajandream M-A, Barrell B G and Parkhill J 2005 ACT: the Artemis Comparison Tool. *Bioinformatics* **21** 3422–3

[56] Gurevich A, Saveliev V, Vyahhi N and Tesler G 2013 QUAST: Quality assessment tool for genome assemblies *Bioinformatics* **29** 1072–5

[57] Kurtz S, Phillippy A, Delcher A L, Smoot M, Shumway M, Antonescu C and Salzberg S L 2004 Versatile and open software for comparing large genomes. *Genome Biol.* **5** R12

[58] Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M and Otto T D 2013 REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14** R47

[59] Galardini M, Biondi E G, Bazzicalupo M and Mengoni A 2011 CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol. Med.* **6** 11

[60] Rissman A I, Mau B, Biehl B S, Darling A E, Glasner J D and Perna N T 2009 Reordering contigs of draft genomes using the Mauve Aligner *Bioinformatics* **25** 2071–3

[61] Lu C L, Chen K-T, Huang S-Y and Chiu H-T 2014 CAR: contig assembly of prokaryotic draft genomes using rearrangements. *BMC Bioinformatics* **15** 381

[62] van Hijum S A F T, Zomer A L, Kuipers O P and Kok J 2005 Projector 2: Contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies *Nucleic Acids Res.* **33**

[63] Assefa S, Keane T M, Otto T D, Newbold C and Berriman M 2009 ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25** 1968–9

[64] Benjamini Y and Speed T P 2012 Summarizing and correcting the GC content bias in high-throughput sequencing *Nucleic Acids Res.* **40**

[65] Chen Y C, Liu T, Yu C H, Chiang T Y and Hwang C C 2013 Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly *PLoS One* **8**

[66] Otto T D, Dillon G P, Degrave W S and Berriman M 2011 RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39** e57

[67] Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M and Otto T D 2016 *Companion* : a web server for annotation and analysis of parasite genomes *Nucleic Acids Res.* **44** W29–34

[68] Seemann T 2014 Prokka: Rapid prokaryotic genome annotation *Bioinformatics* **30** 2068–9

[69] Overbeek R, Olson R, Pusch G D, Olsen G J, Davis J J, Disz T, Edwards R A, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam A R, Xia F and Stevens R 2014 The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) *Nucleic Acids Res.* **42**

[70] Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang

H, Wang J and Wang J 2010 De novo assembly of human genomes with massively parallel short read sequencing *Genome Res.* **20** 265–72

[71] Alkan C, Sajjadian S and Eichler E E 2010 Limitations of next-generation genome sequence assembly *Nat. Methods* **8** 61–5

[72] Baker M 2012 De novo genome assembly: what every biologist should know *Nat. Methods* **9** 333–7

[73] Green P 2002 Whole-genome disassembly. *Proc. Natl. Acad. Sci. U. S. A.* **99** 4143–4

[74] Chain P S G, Grafham D V, Fulton R S, Fitzgerald M G, Hostetler J, Muzny D, Ali J, Birren B, Bruce D C, Buhay C, Cole J R, Ding Y, Dugan S, Field D, Garrity G M, Gibbs R, Graves T, Han C S, Harrison S H, Highlander S, Hugenholtz P, Khouri H M, Kodira C D, Kolker E, Kyrpides N C, Lang D, Lapidus A, Malfatti S A, Markowitz V, Metha T, Nelson K E, Parkhill J, Pitluck S, Qin X, Read T D, Schmutz J, Sozhamannan S, Sterk P, Strausberg R L, Sutton G, Thomson N R, Tiedje J M, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium and Detter J C 2009 Genomics. Genome project standards in a new era of sequencing. *Science* **326** 236–7

[75] Simão F A, Waterhouse R M, Ioannidis P, Kriventseva E V. and Zdobnov E M 2015 BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs *Bioinformatics* **31** 3210–2

[76] Burrows M and Wheeler D 1994 A block-sorting lossless data compression algorithm *Algorithm, Data Compression* 18

[77] Li H, Ruan J and Durbin R 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18** 1851–8

[78] Li R, Li Y, Kristiansen K and Wang J 2008 SOAP: Short oligonucleotide alignment program *Bioinformatics* **24** 713–4

[79] Li H and Durbin R 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26** 589–95

[80] Langmead B and Salzberg S L 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9** 357–9

[81] Li H and Durbin R 2009 Fast and accurate short read alignment with Burrows-Wheeler transform *Bioinformatics* **25** 1754–60

[82] Langmead B, Trapnell C, Pop M and Salzberg S L 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10** R25

[83] Tattini L, D'Aurizio R and Magi A 2015 Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front. Bioeng. Biotechnol.* **3** 92

[84] Pang A W, MacDonald J R, Pinto D, Wei J, Rafiq M a, Conrad D F, Park H, Hurles M E, Lee C, Venter J C, Kirkness E F, Levy S, Feuk L and Scherer S W 2010 Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11** R52

[85] Brand H, Collins R L, Hanscom C, Rosenfeld J A, Pillalamarri V, Stone M R, Kelley F, Mason T, Margolin L, Eggert S, Mitchell E, Hodge J C, Gusella J F, Sanders S J and Talkowski M E 2015 Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation *Am. J. Hum. Genet.* **97** 170–6

[86] Brandler W M, Antaki D, Gujral M, Noor A, Rosanio G, Chapman T R, Barrera D J, Lin G N, Malhotra D, Watts A C, Wong L C, Estabillo J A, Gadomski T E, Hong O, Fajardo K V F, Bhandari A, Owen R, Baughn M, Yuan J, Solomon T, Moyzis A G, Maile M S, Sanders S J, Reiner G E, Vaux K K, Strom C M, Zhang K, Muotri A R, Akshoomoff N, Leal S M, Pierce K, Courchesne E, Iakoucheva L M, Corsello C and Sebat J 2016 Frequency and Complexity of de Novo Structural Mutation in Autism *Am. J. Hum. Genet.* **98** 667–79

[87] Frost L S, Leplae R, Summers A O and Toussaint A 2005 Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3** 722–32

[88] Lin K, Bonnema G, Sanchez-Perez G and De Ridder D 2014 Making the difference: Integrating structural variation detection tools *Brief. Bioinform.* **16** 852–64

[89] Zhao M, Wang Q Q, Wang Q Q, Jia P and Zhao Z 2013 Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 (Suppl 11)** S1

[90] Robinson J T, Thorvaldsdóttir H, Winckler W, Guttman M, Lander E S, Getz G and Mesirov J P 2011 Integrative genomics viewer. *Nat. Biotechnol.* **29** 24–6

[91] Medvedev P, Stanciu M and Brudno M 2009 Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6** S13-20

[92] Ye K, Schulz M H, Long Q, Apweiler R and Ning Z 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25** 2865–71

[93] Rausch T, Zichner T, Schlattl A, Stütz A M, Benes V and Korbel J O 2012 DELLY: Structural variant discovery by integrated paired-end and split-read analysis *Bioinformatics* **28**

[94] Sindi S S, Onal S, Peng L C, Wu H-T and Raphael B J 2012 An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.* **13** R22

[95] Wong K, Keane T M, Stalker J and Adams D J 2010 Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* **11** R128

[96] Lam H Y K, Pan C, Clark M J, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein M B, Kidd J M, Bustamante C D and Snyder M 2012 Detecting and annotating genetic variations using the HugeSeq pipeline *Nat. Biotechnol.* **30** 226–9

[97] Delcher A L, Phillippy A, Carlton J and Salzberg S L 2002 Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30** 2478–83

[98] McVean G A, Altshuler (Co-Chair) D M, Durbin (Co-Chair) R M, Abecasis G R, Bentley D R, Chakravarti A, Clark A G, Donnelly P, Eichler E E, Flicek P, Gabriel S B, Gibbs R A, Green E D, Hurles M E, Knoppers B M, Korbel J O, Lander E S, Lee C, Lehrach H, Mardis E R, Marth G T, McVean G A, Nickerson D A, Schmidt J P, Sherry S T, Wang J, Wilson R K,

Gibbs (Principal Investigator) R A, Dinh H, Kovar C, Lee S, Lewis L, Muzny D, Reid J, Wang M, Wang (Principal Investigator) J, Fang X, Guo X, Jian M, Jiang H, Jin X, Li G, Li J, Li Y, Li Z, Liu X, Lu Y, Ma X, Su Z, Tai S, Tang M, Wang B, Wang G, Wu H, Wu R, Yin Y, Zhang W, Zhao J, Zhao M, Zheng X, Zhou Y, Lander (Principal Investigator) E S, Altshuler D M, Gabriel (Co-Chair) S B, Gupta N, Flicek (Principal Investigator) P, Clarke L, Leinonen R, Smith R E, Zheng-Bradley X, Bentley (Principal Investigator) D R, Grocock R, Humphray S, James T, Kingsbury Z, Lehrach (Principal Investigator) H, Sudbrak (Project Leader) R, Albrecht M W, Amstislavskiy V S, Borodina T A, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo M-L, Sherry (Principal Investigator) S T, McVean (Principal Investigator) G A, Mardis (Co-Principal Investigator) (Co-Chair) E R, Wilson (Co-Principal Investigator) R K, Fulton L, Fulton R, Weinstock G M, Durbin (Principal Investigator) R M, Balasubramaniam S, Burton J, Danecek P, Keane T M, Kolb-Kokocinski A, et al 2012 An integrated map of genetic variation from 1,092 human genomes *Nature* **491** 56–65

[99] Weisenfeld N I, Yin S, Sharpe T, Lau B, Hegarty R, Holmes L, Sogoloff B, Tabbaa D, Williams L, Russ C, Nusbaum C, Lander E S, MacCallum I and Jaffe D B 2014 Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46** 1350–5

[100] Han E, Sinsheimer J S and Novembre J 2014 Characterizing bias in population genetic inferences from low-coverage sequencing data *Mol. Biol. Evol.* **31** 723–35

[101] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo M A 2010 The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data *Genome Res.* **20** 1297–303

[102] Li H 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27** 2987–93

[103] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R 2009 The Sequence Alignment/Map format and SAMtools *Bioinformatics* **25** 2078–9

[104] Garrison E and Marth G 2012 Haplotype-based variant detection from short-read sequencing *arXiv Prepr. arXiv1207.3907*

[105] Clevenger J, Chavarro C, Pearl S A, Ozias-Akins P and Jackson S A 2015 Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations *Mol. Plant* **8** 831–46

[106] Li H 2014 Toward better understanding of artifacts in variant calling from high-coverage samples *Bioinformatics* **30** 2843–51

[107] Trapnell C, Pachter L and Salzberg S L 2009 TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25** 1105–11

[108] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D R, Pimentel H, Salzberg S L, Rinn J L and Pachter L 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7** 562–78

[109] Schulz M H, Zerbino D R, Vingron M and Birney E 2012 Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels *Bioinformatics* **28** 1086–92

[110] Grabherr M G, Haas B J, Yassour M, Levin J Z, Thompson D A, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren B W, Nusbaum C, Lindblad-Toh K, Friedman N and Regev A 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29** 644–52

[111] Anders S, Pyl P T and Huber W 2015 HTSeq-A Python framework to work with high-throughput sequencing data *Bioinformatics* **31** 166–9

[112] Trapnell C, Hendrickson D G, Sauvageau M, Goff L, Rinn J L and Pachter L 2013 Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31** 46–53

[113] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak M W, Gaffney D J, Elo L L, Zhang X and Mortazavi A 2016 A survey of best practices for RNA-seq data analysis *Genome Biol.* **17** 13

[114] Love M I, Huber W and Anders S 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome Biol.* **15** 550

[115] Robinson M D, McCarthy D J and Smyth G K 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–40

[116] Li J, Witten D M, Johnstone I M and Tibshirani R 2012 Normalization, testing, and false discovery rate estimation for RNA-sequencing data *Biostatistics* **13** 523–38

[117] Benjamini Y and Hochberg Y 1995 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc. B* **57** 289–300

[118] Leek J T, Scharpf R B, Bravo H C, Simcha D, Langmead B, Johnson W E, Geman D, Baggerly K and Irizarry R a 2010 Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–9

[119] Leek J T 2014 svaseq: removing batch effects and other unwanted noise from sequencing data *Nucleic Acids Res.* **42** e161

[120] Stegle O, Parts L, Piipari M, Winn J and Durbin R 2012 Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7** 500–7

[121] Li S, Labaj P P, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu P-Y, Wang M, Wang C, Thierry-Mieg D, Thierry-Mieg J, Kreil D P and Mason C E 2014 Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32** 888–95

[122] Simpson A G B, Stevens J R and Lukeš J 2006 The evolution and diversity of kinetoplastid flagellates *Trends Parasitol.* **22** 168–74

[123] Akhoundi M, Kuhls K, Cannet A, Votýpka J, Marty P, Delaunay P and Sereno D 2016 A Historical Overview of the Classification, Evolution, and Dispersion of *Leishmania* Parasites and Sandflies ed A-L Bañuls *PLoS Negl. Trop. Dis.* **10** e0004349

[124] Rose K, Curtis J, Baldwin T, Mathis A, Kumar B, Sakthianandeswaren A, Spurck T, Low

Choy J and Handman E 2004 Cutaneous leishmaniasis in red kangaroos: Isolation and characterisation of the causative organisms *Int. J. Parasitol.* **34** 655–64

[125] WHO technical report series 2010 Control of the leishmaniasis: report of a meeting of the WHO Expert Committee on the Control of Leishmaniases, Geneva, 22-26 March 2010. *World Health Organ. Tech. Rep. Ser.* **949** 202

[126] Alvar J, Vélez I D, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, de Boer M, Boer M den, Team the W L C, Hotez P, Molyneux D, Fenwick A, Ottesen E, Sachs S E, Hotez P, Remme J, Buss P, Alleyne G, Morel C, Alvar J, Yactayo S, Bern C, Bern C, Maguire J, Alvar J, Cattand P, Desjeux P, Guzman M, Jannin J, Kroeger A, Mathers C, Ezzati M, Lopez A, World B, AbouZahr C, Vaughan J, King C, Bertino A, King C, Dickman K, Tisch D, Conteh L, Engels T, Molyneux D, Desjeux P, Gupta P Sen, Bern C, Hightower A, Chowdhury R, Ali M, Amann J, Collin S, Davidson R, Ritmeijer K, Keus K, Melaku Y, Rey L, Martins C, Ribeiro H, Lima A, Zijlstra E, el-Hassan A, Ismael A, Ghalib H, Desjeux P, Copeland H, Arana B, Navin T, Maia-Elkhoury A, Carmo E, Sousa-Gomes M, Mota E, Mosleh I, Geith E, Natsheh L, Abdul-Dayem M, Abotteen N, Singh S, Reddy D, Rai M, Sundar S, Singh V, Ranjan A, Topno R, Verma R, Siddique N, Yadon Z, Quigley M, Davies C, Rodrigues L, Segura E, Reithinger R, Mohsen M, Aadil K, Sidiqi M, Erasmus P, Ashford R, et al 2012 Leishmaniasis worldwide and global estimates of its incidence **7**, ed M Kirk (Public Library of Science)

[127] Cupolillo E, Medina-Acosta E, Noyes H, Momen H and Grimaldi G 2000 A revised classification for *Leishmania* and *Endotrypanum*. *Parasitol. Today* **16** 142–4

[128] Schönian G, Mauricio I and Cupolillo E 2010 Is it time to revise the nomenclature of *Leishmania? Trends Parasitol.* **26** 466-9

[129] Gramiccia M and Gradoni L 2005 The current status of zoonotic leishmaniases and approaches to disease control. *Int. J. Parasitol.* **35** 1169–80

[130] Silveira F T, Ishikawa E A Y, De Souza A A A and Lainson R 2002 An outbreak of cutaneous leishmaniasis among soldiers in Belém, Pará State, Brazil, caused by *Leishmania (Viannia) lindenbergi* n. sp. A new leishmanial parasite of man in the Amazon region. *Parasite* **9** 43–50

[131] Kato H, Calvopiña M, Criollo H and Hashiguchi Y 2013 First human cases of *Leishmania (Viannia) naiffi* infection in Ecuador and identification of its suspected vector species. *Acta Trop.* **128** 710–3

[132] Braga R R, Lainson R, Ishikawa E A Y and Shaw J J 2003 *Leishmania (Viannia) utingensis* n. sp., a parasite from the sandfly *Lutzomyia (Viannamyia) tuberculata* in Amazonian Brazil. *Parasite* **10** 111–8

[133] Manson-Bahr P E and Heisch R B 1961 Transient infection of man with a *Leishmania (L. adleri)* of lizards. *Ann. Trop. Med. Parasitol.* **55** 381–2

[134] Yang B-B, Chen D-L, Chen J-P, Liao L, Hu X-S and Xu J-N 2013 Analysis of kinetoplast cytochrome b gene of 16 *Leishmania* isolates from different foci of China: different species of *Leishmania* in China and their phylogenetic inference. *Parasit. Vectors* **6** 32

[135] Novo S P C, Leles D, Bianucci R and Araujo A 2015 *Leishmania tarentolae* molecular signatures in a 300 hundred-years-old human Brazilian mummy. *Parasit. Vectors* **8** 72

[136] Real F, Vidal R O, Carazzolle M F, Mondego J M C, Costa G G L, Herai R H, Würtele M, de Carvalho L M, Carmona e Ferreira R, Mortara R A, Barbiéri C L, Mieczkowski P, da Silveira J F, Briones M R D S, Pereira G A G, Bahia D, E Ferreira R C, Mortara R A, Barbiéri C L, Mieczkowski P, da Silveira J F, Briones M R D S, Pereira G A G and Bahia D 2013 The Genome Sequence of *Leishmania (Leishmania) amazonensis*: Functional Annotation and Extended Analysis of Gene Models. *DNA Res.* **20** 1–15

[137] Kwakye-Nuako G, Mosore M-T, Duplessis C, Bates M D, Puplampu N, Mensah-Attipoe I, Desewu K, Afegbe G, Asmah R H, Jamjoom M B, Ayeh-Kumi P F, Boakye D A and Bates P A 2015 First isolation of a new species of *Leishmania* responsible for human cutaneous leishmaniasis in Ghana and classification in the *Leishmania enriettii* complex. *Int. J. Parasitol.* **45** 679-84

[138] Rohoušová I, Hostomská J, Vlková M, Kobets T, Lipoldová M and Volf P 2011 The protective effect against *Leishmania* infection conferred by sand fly bites is limited to short-term exposure *Int. J. Parasitol.* **41** 481–5

[139] Marovich M A, Lira R, Shepard M, Fuchs G H, Kreutzer R D, Nutman T B, Neva F a, Kruetzer R, Nutman T B and Neva F a 2001 Leishmaniasis recidivans recurrence after 43 years: a clinical and immunologic report after successful treatment *Clin. Infect. Dis.* **33** 1076–9

[140] Wortmann G W, Aronson N E, Miller R S, Blazes D and Oster C N 2000 Cutaneous leishmaniasis following local trauma: a clinical pearl. *Clin. Infect. Dis.* **31** 199–201

[141] Endris M, Takele Y, Woldeyohannes D, Tiruneh M, Mohammed R, Moges F, Lynen L, Jacobs J, Van Griensven J and Diro E 2014 Bacterial sepsis in patients with visceral leishmaniasis in Northwest Ethiopia *Biomed Res. Int.* **2014**

[142] Druzian A F, de Souza A S, de Campos D N, Croda J, Higa M G, Dorval M E C, Pompilio M A, de Oliveira P A and Paniago A M M 2015 Risk Factors for Death from Visceral Leishmaniasis in an Urban Area of Brazil *PLoS Negl. Trop. Dis.* **9**

[143] Sundar S, Thakur B B, Tandon A K, Agrawal N R, Mishra C P, Mahapatra T M and Singh V P 1994 Clinicoepidemiological study of drug resistance in Indian kala-azar. *BMJ* **308** 307

[144] Sundar S, Sinha P R, Agrawal N K, Srivastava R, Rainey P M, Berman J D, Murray H W and Singh V P 1998 A cluster of cases of severe cardiotoxicity among kala-azar patients treated with a high-osmolarity lot of sodium antimony gluconate *Am. J. Trop. Med. Hyg.* **59** 139–43

[145] Croft S L, Sundar S and Fairlamb A H 2006 Drug resistance in leishmaniasis *Clin. Microbiol. Rev.* **19** 111–26

[146] McCall L I, Zhang W W, Ranasinghe S and Matlashewski G 2013 Leishmanization revisited: Immunization with a naturally attenuated cutaneous *Leishmania donovani* isolate from Sri Lanka protects against visceral leishmaniasis *Vaccine* **31** 1420–5

[147] Khamesipour A, Dowlati Y, Asilian A, Hashemi-Fesharki R, Javadi A, Noazin S and Modabber F 2005 Leishmanization: Use of an old method for evaluation of candidate vaccines against leishmaniasis *Vaccine* **23** 3642–8

[148] Moradin N and Descoteaux A 2012 *Leishmania* promastigotes: building a safe niche within macrophages. *Front. Cell. Infect. Microbiol.* **2** 121

[149] McConville M J, de Souza D, Saunders E, Likic V A and Naderer T 2007 Living in a phagolysosome; metabolism of *Leishmania amastigotes Trends Parasitol.* **23** 368–75

[150] Bates P a and Rogers M E 2004 New insights into the developmental biology and transmission mechanisms of *Leishmania*. *Curr. Mol. Med.* **4** 601–9

[151] Bates P A 2007 Transmission of *Leishmania* metacyclic promastigotes by phlebotomine sand flies *Int. J. Parasitol.* **37** 1097–106

[152] Dostálová A and Volf P 2012 *Leishmania* development in sand flies: parasite-vector interactions overview. *Parasit. Vectors* **5** 276

[153] Pimenta P, Turco S, McConville M, Lawyer P, Perkins P and Sacks D 1992 Stage-specific adhesion of *Leishmania* promastigotes to the sandfly midgut *Science (80-. ).* **256** 1812–5

[154] Rogers M E, Ilg T, Nikolaev A V, Ferguson M A J and Bates P A 2004 Transmission of cutaneous leishmaniasis by sand flies is enhanced by regurgitation of fPPG. *Nature* **430** 463–7

[155] McConville M J, Turco S J, Ferguson M A J and Sacks D L J 1992 Developmental modification of lipophosphoglycan during the differentiation of *Leishmania major* promastigotes to an infectious stage. *EMBO J.* **8** 396

[156] Myskova J, Svobodova M, Beverley S M and Volf P 2007 A lipophosphoglycan-independent development of *Leishmania* in permissive sand flies *Microbes Infect.* **9** 317–24

[157] Volf P, Hajmova M, Sadlova J and Votypka J 2004 Blocked stomodeal valve of the insect vector: Similar mechanism of transmission in two trypanosomatid models *Int. J. Parasitol.* **34** 1221–7

[158] Andrade B B, De Oliveira C I, Brodskyn C I, Barral A and Barral-Netto M 2007 Role of sand fly saliva in human and experimental leishmaniasis: Current insights *Scand. J. Immunol.* **66** 122–7

[159] Rogers M E, Chance M L and Bates P a 2002 The role of promastigote secretory gel in the origin and transmission of the infective stage of *Leishmania mexicana* by the sandfly *Lutzomyia longipalpis Parasitology* **124** 495–507

[160] Maroli M, Feliciangeli M D, Bichaud L, Charrel R N and Gradoni L 2013 Phlebotomine sandflies and the spreading of leishmaniases and other diseases of public health concern. *Med. Vet. Entomol.* **27** 123–47

[161] Kaye P and Scott P 2011 Leishmaniasis: complexity at the host-pathogen interface. *Nat. Rev. Microbiol.* **9** 604–15

[162] Maroli M, Feliciangeli M D, Bichaud L, Charrel R N and Gradoni L 2013 Phlebotomine sandflies and the spreading of leishmaniases and other diseases of public health concern *Med. Vet. Entomol.* **27** 123–47

[163] Mukherjee S, Hassan M Q, Ghosh A, Ghosh K N, Bhattacharya A and Adhya S 1997 Short report: *Leishmania* DNA in Phlebotomus and Sergentomyia species during a kala-azar epidemic. *Am. J. Trop. Med. Hyg.* **57** 423–5

[164] Dougall A M, Alexander B, Holt D C, Harris T, Sultan A H, Bates P A, Rose K and Walton S F 2011 Evidence incriminating midges (Diptera: Ceratopogonidae) as potential vectors of *Leishmania* in Australia *Int. J. Parasitol.* **41** 571–9

[165] Dantas-Torres F 2011 Ticks as vectors of *Leishmania* parasites. *Trends Parasitol.* **27** 155–9

[166] Maroli M, M.D. F, Bichaud L, Charrel R N and Gradoni L 2013 Phlebotomine sandflies and the spreading of leishmaniases and other diseases of public health concern *Med. Vet. Entomol.* **27** 123–47

[167] Killick-Kendrick R 1999 The biology and control of Phlebotomine sand flies *Clin. Dermatol.* **17** 279–89

[168] Roque A L R and Jansen A M 2014 Wild and synanthropic reservoirs of *Leishmania* species in the Americas *Int. J. Parasitol. Parasites Wildl.* **3** 251–62

[169] Santiago M E B, Vasconcelos R O, Fattori K R, Munari D P, Michelin A de F and Lima V M F 2007 An investigation of *Leishmania* spp. in Didelphis spp. from urban and peri-urban areas in Bauru (São Paulo, Brazil) *Vet. Parasitol.* **150** 283–90

[170] Podaliri Vulpiani M, Iannetti L, Paganico D, Iannino F and Ferri N 2011 Methods of Control of the *Leishmania infantum* Dog Reservoir: State of the Art. *Vet. Med. Int.* **2011** 215964

[171] Baneth G, Zivotofsky D, Nachum-Biala Y, Yasur-Landau D and Botero A-M 2014 Mucocutaneous *Leishmania tropica* infection in a dog from a human cutaneous leishmaniasis focus. *Parasit. Vectors* **7** 118

[172] Dantas-Torres F 2009 Canine leishmaniosis in South America. *Parasit. Vectors* **2** S1

[173] Dantas-Torres F, Solano-Gallego L, Baneth G, Ribeiro V M, de Paiva-Cavalcanti M and Otranto D 2012 Canine leishmaniosis in the Old and New Worlds: Unveiled similarities and differences *Trends Parasitol.* **28** 531–8

[174] Madeira M F, Figueiredo F B, Pinto A G S, Nascimento L D, Furtado M, Mouta-Confort E, de Paula C C, Bogio A, Gomes M C A, Bessa A M S and Passos S R L 2009 Parasitological diagnosis of canine visceral leishmaniasis: Is intact skin a good target? *Res. Vet. Sci.* **87** 260–2

[175] Queiroz P V S, Monteiro G R G, Macedo V P S, Rocha M a C, Batista L M M, Queiroz J W, Jerônimo S M B and Ximenes M F F M 2009 Canine visceral leishmaniasis in urban and rural areas of Northeast Brazil. *Res. Vet. Sci.* **86** 267–73

[176] Dantas-Torres F, de Brito M E F and Brandão-Filho S P 2006 Seroepidemiological survey on canine leishmaniasis among dogs from an urban area of Brazil. *Vet. Parasitol.* **140** 54–60

[177] Rogers M B, Hilley J D, Dickens N J, Wilkes J, Bates P a, Depledge D P, Harris D, Her Y, Herzyk P, Imamura H, Otto T D, Sanders M, Seeger K, Dujardin J-C, Berriman M, Smith D F, Hertz-Fowler C and Mottram J C 2011 Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. *Genome Res.* **21** 2129–42

[178] Peacock C S, Seeger K, Harris D, Murphy L, Ruiz J C, Quail M a, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream M-A, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S,

Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf S L, Faulconbridge A, Jeffares D, Depledge D P, Oyola S O, Hilley J D, Brito L O, Tosi L R O, Barrell B, Cruz A K, Mottram J C, Smith D F and Berriman M 2007 Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet.* **39** 839–47

[179] Harkins K M, Schwartz R S, Cartwright R A and Stone A C 2016 Phylogenomic reconstruction supports supercontinent origins for *Leishmania*. *Infect. Genet. Evol.* **38** 101–9

[180] Myler P J, Sisk E, McDonagh P D, Martinez-Calvillo S, Schnaufer a, Sunkin S M, Yan S, Madhubala R, Ivens a and Stuart K 2000 Genomic organization and gene function in *Leishmania*. *Biochem. Soc. Trans.* **28** 527–31

[181] Britto C, Ravel C, Bastien P, Blaineau C, Pagès M, Dedet J-P P and Wincker P 1998 Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes *Gene* **222** 107–17

[182] Deschamps P, Lara E, Marande W, Lopez-Garcia P, Ekelund F and Moreira D 2011 Phylogenomic Analysis of Kinetoplastids Supports That Trypanosomatids Arose from within Bodonids *Mol. Biol. Evol.* **28** 53–8

[183] Opperdoes F R, Butenko A, Flegontov P, Yurchenko V and Lukeš J 2016 Comparative Metabolism of Free-living *Bodo saltans* and Parasitic Trypanosomatids *J. Eukaryot. Microbiol.*

[184] Jackson A P, Otto T D, Aslett M, Armstrong S D, Bringaud F, Schlacht A, Hartley C, Sanders M, Wastling J M, Dacks J B, Acosta-Serrano A, Field M C, Ginger M L and Berriman M 2016 Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism *Curr. Biol.* **26** 161–72

[185] Fiebig M, Kelly S and Gluenz E 2015 Comparative Life Cycle Transcriptomics Revises *Leishmania mexicana* Genome Annotation and Links a Chromosome Duplication with Parasitism of Vertebrates *PLoS Pathog.* **11**

[186] Downing T, Imamura H, Decuypere S, Clark T G, Coombs G H, Cotton J A, Hilley J D, de Doncker S, Maes I, Mottram J C, Quail M A, Rijal S, Sanders M, Schönian G, Stark O, Sundar S, Vanaerschot M, Hertz-Fowler C, Dujardin J-C and Berriman M 2011 Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* **21** 2143–56

[187] Imamura H, Downing T, van den Broeck F, Sanders M J, Rijal S, Sundar S, Mannaert A, Vanaerschot M, Berg M, de Muylder G, Dumetz F, Cuypers B, Maes I, Domagalska M, Decuypere S, Rai K, Uranw S, Bhattarai N R, Khanal B, Prajapati V K, Sharma S, Stark O, Sch??nian G, de Koning H P, Settimo L, Vanhollebeke B, Roy S, Ostyn B, Boelaert M, Maes L, Berriman M, Dujardin J C and Cotton J A 2016 Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent *Elife* **5**

[188] Llanes A, Restrepo C M, Del Vecchio G, Anguizola F J and Lleonart R 2015 The genome of *Leishmania panamensis*: insights into genomics of the *L. (Viannia)* subgenus. *Sci. Rep.* **5** 8550

[189] Valdivia H O, Reis-Cunha J L, Rodrigues-Luiz G F, Baptista R P, Baldeviano G C, Gerbasi R V, Dobson D E, Pratlong F, Bastien P, Lescano A G, Beverley S M and Bartholomeu D C 2015 Comparative genomic analysis of *Leishmania (Viannia) peruviana* and *Leishmania (Viannia) braziliensis*. *BMC Genomics* **16** 715

[190] Raymond F, Boisvert S, Roy G, Ritt J-F, Légaré D, Isnard A, Stanke M, Olivier M, Tremblay M J, Papadopoulou B, Ouellette M and Corbeil J 2012 Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species. *Nucleic Acids Res.* **40** 1131–47

[191] Ivens A C, Peacock C S, Worthey E A, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream M-A, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley S M, Bianchettin G, Borzym K, Bothe G, Bruschi C V, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson R M R, Cronin A, Cruz A K, Davies R M, De Gaudenzi J, Dobson D E, Duesterhoeft A, Fazelina G, Fosker N, Frasch A C, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram J C, Müller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'neil S, Pentony M, Pohl T M, Price C, Purnelle B, Quail M A, Rabbinowitsch E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz J C, Rutter S, Saunders D, Schäfer M, Schein J, Schwartz D C, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G, Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith D F, Blackwell J M, et al 2005 The genome of the kinetoplastid parasite, *Leishmania major*. *Science* **309** 436–42

[192] Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K and Myler P J 2003 Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. *Mol. Cell* **11** 1291–9

[193] Myler P J, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, Swartzell S, Westlake T, Bastien P, Fu G, Ivens a and Stuart K 1999 *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. U. S. A.* **96** 2902–6

[194] Liang X H, Haritan A, Uliel S and Michaeli S 2003 trans and cis splicing in trypanosomatids: Mechanism, factors, and regulation *Eukaryot. Cell* **2** 830–40

[195] Worthey E A, Martinez-Calvillo S, Schnaufer A, Aggarwal G, Cawthra J, Fazelinia G, Fong C, Fu G, Hassebrock M, Hixson G, Ivens A C, Kiser P, Marsolini F, Rickell E, Salavati R, Sisk E, Sunkin S M, Stuart K and Myler P J 2003 *Leishmania major* chromosome 3 contains two long convergent polycistronic gene clusters separated by a tRNA gene *Nucleic Acids Res.* **31** 4201–10

[196] Lee M G and Van der Ploeg L H 1997 Transcription of protein-coding genes in trypanosomes by RNA polymerase I. *Annu. Rev. Microbiol.* **51** 463–89

[197] Clayton C and Shapira M 2007 Post-transcriptional regulation of gene expression in trypanosomes and leishmanias *Mol. Biochem. Parasitol.* **156** 93–101

[198] Siegel T N, Hekstra D R, Wang X, Dewell S and Cross G A M 2010 Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites *Nucleic Acids Res.* **38** 4946–57

[199] Thomas S, Green A, Sturm N R, Campbell D A and Myler P J 2009 Histone acetylations mark origins of polycistronic transcription in *Leishmania major. BMC Genomics* **10** 152

[200] Martínez-Calvillo S, Nguyen D, Stuart K and Myler P J 2004 Transcription initiation and termination on *Leishmania major* chromosome 3 *Eukaryot. Cell* **3** 506–17

[201] van Luenen H G a M, Farris C, Jan S, Genest P-A, Tripathi P, Velds A, Kerkhoven R M, Nieuwland M, Haydock A, Ramasamy G, Vainio S, Heidebrecht T, Perrakis A, Pagie L, van Steensel B, Myler P J and Borst P 2012 Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. *Cell* **150** 909–21

[202] Murray A, Fu C, Habibi G and McMaster W R 2007 Regions in the 3′ untranslated region confer stage-specific expression to the *Leishmania mexicana* a600-4 gene *Mol. Biochem. Parasitol.* **153** 125–32

[203] Haile S and Papadopoulou B 2007 Developmental regulation of gene expression in trypanosomatid parasitic protozoa *Curr. Opin. Microbiol.* **10** 569–77

[204] Depledge D P, Evans K J, Ivens A C, Aziz N, Maroof A, Kaye P M and Smith D F 2009 Comparative expression profiling of *Leishmania*: modulation in gene expression between species and in different host genetic backgrounds. *PLoS Negl. Trop. Dis.* **3** e476

[205] Rochette A, Raymond F, Ubeda J-M, Smith M, Messier N, Boisvert S, Rigault P, Corbeil J, Ouellette M, Papadopoulou B, Murray H, Berman J, Davies C, Saravia N, Sacks D, Kamhawi S, Zilberstein D, Shapira M, Garlapati S, Dahan E, Shapira M, Barak E, Amin-Spector S, Gerliak E, Goyard S, Holland N, Zilberstein D, MacFarlane J, Blaxter M, Bishop R, Miles M, Kelly J, Glaser T, Moody S, Handman E, Bacic A, Spithill T, Turco S, Descoteaux A, McConville M, Blackwell J, Charest H, Matlashewski G, Argaman M, Aly R, Shapira M, Hubel A, Krobitsch S, Horauf A, Clos J, Burchmore R, Landfear S, Rochette A, McNicoll F, Girard J, Breton M, Leblanc E, Bergeron M, Papadopoulou B, Handman E, Osborn A, Symons F, Driel R van, Cappai R, Barr S, Gedamu L, Moore L, Santrich C, LeBowitz J, Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream M, Adlem E, Aert R, Peacock C, Seeger K, Harris D, Murphy L, Ruiz J, Quail M, Peters N, Adlem E, Tivey A, Aslett M, Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P, Clayton C, Shapira M, Haile S, et al 2008 Genome-wide gene expression profiling analysis of *Leishmania major* and *Leishmania infantum* developmental stages reveals substantial differences between the two species *BMC Genomics* **9** 255

[206] Mukherjee A, Langston L D and Ouellette M 2011 Intrachromosomal tandem duplication and repeat expansion during attempts to inactivate the subtelomeric essential gene GSH1 in *Leishmania Nucleic Acids Res.* **39** 7499–511

[207] Ritt J-F, Raymond F, Leprohon P, Légaré D, Corbeil J and Ouellette M 2013 Gene Amplification and Point Mutations in Pyrimidine Metabolic Genes in 5-Fluorouracil Resistant *Leishmania infantum PLoS Negl. Trop. Dis.* **7** e2564

[208] Pérez-Victoria F J, Gamarro F, Ouellette M and Castanys S 2003 Functional cloning of the miltefosine transporter: A novel p-type phospholipid translocase from leishmania involved in drug resistance *J. Biol. Chem.* **278** 49965–71

[209] Leprohon P, Fernandez-Prada C, Gazanion É, Monte-Neto R and Ouellette M 2015 Drug resistance analysis by next generation sequencing in *Leishmania*. *Int. J. Parasitol. Drugs drug Resist.* **5** 26–35

[210] Sterkers Y, Lachaud L, Crobu L, Bastien P and Pagès M 2011 FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell. Microbiol.* **13** 274–83

[211] Zhang W W, Ramasamy G, McCall L-I, Haydock A, Ranasinghe S, Abeygunasekara P, Sirimanna G, Wickremasinghe R, Myler P and Matlashewski G 2014 Genetic analysis of

*Leishmania donovani* tropism using a naturally attenuated cutaneous strain. *PLoS Pathog.* **10** e1004244

[212] Leprohon P, Légaré D, Raymond F, Madore E, Hardiman G, Corbeil J and Ouellette M 2009 Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. *Nucleic Acids Res.* **37** 1387–99

[213] Ubeda J-M, Légaré D, Raymond F, Ouameur A A, Boisvert S, Rigault P, Corbeil J, Tremblay M J, Olivier M, Papadopoulou B and Ouellette M 2008 Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy. *Genome Biol.* **9** R115

[214] Beverley S M and Coburn C M 1990 Recurrent de novo appearance of small linear DNAs in *Leishmania major* and relationship to extra-chromosomal DNAs in other species *Mol. Biochem. Parasitol.* **42** 133–41

[215] Tripp C A, Myler P J and Stuart K 1991 A DNA sequence (LD1) which occurs in several genomic organizations in *Leishmania Mol. Biochem. Parasitol.* **47** 151–60

[216] Navarro M, Liu J, Muthui D, Ortiz G, Segovia M and Hamers R 1994 Inverted repeat structure and homologous sequences in the LD1 amplicons of *Leishmania* spp. *Mol. Biochem. Parasitol.* **68** 69–80

[217] Ubeda J-M, Raymond F, Mukherjee A, Plourde M, Gingras H, Roy G, Lapointe A, Leprohon P, Papadopoulou B, Corbeil J and Ouellette M 2014 Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite *Leishmania. PLoS Biol.* **12** e1001868

[218] Laffitte M-C N, Leprohon P, Papadopoulou B, Ouellette M, Laffitte M-C N, Leprohon P, Papadopoulou B and Ouellette M 2016 Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance *F1000Research* **5** 2350

[219] Rougeron V, Bañuls A L, Carme B, Simon S, Couppié P, Nacher M, Hide M and De Meeûs T 2011 Reproductive strategies and population structure in *Leishmania*: Substantial amount of sex in *Leishmania Viannia guyanensis Mol. Ecol.* **20** 3116–27

[220] Kuhls K, Cupolillo E, Silva S O, Schweynoch C, Boité M C, Mello M N, Mauricio I, Miles M, Wirth T and Schönian G 2013 Population structure and evidence for both clonality and recombination among Brazilian strains of the subgenus *Leishmania (Viannia). PLoS Negl. Trop. Dis.* **7** e2490

[221] Kelly J M, Law J M, Chapman C J, Van Eys G J and Evans D A 1991 Evidence of genetic recombination in *Leishmania*. *Mol. Biochem. Parasitol.* **46** 253–63

[222] Akopyants N S, Kimblin N, Secundino N, Patrick R, Peters N, Lawyer P, Dobson D E, Beverley S M and Sacks D L 2009 Demonstration of Genetic Exchange During Cyclical Development of *Leishmania* in the Sand Fly Vector *Science* **324** 265–8

[223] Odiwuor S, De Doncker S, Maes I, Dujardin J-C C and Van der Auwera G 2011 Natural *Leishmania donovani*/*Leishmania aethiopica* hybrids identified from Ethiopia. *Infect. Genet. Evol.* **11** 2113–8

[224] Lanotte G and Rioux J A 1990 [Cell fusion in *Leishmania* (Kinetoplastida, Trypanosomatidae)] *C R Acad Sci III* **310** 285–8

[225] Sterkers Y, Crobu L, Lachaud L, Pagès M and Bastien P 2014 Parasexuality and mosaic aneuploidy in *Leishmania*: alternative genetics *Trends Parasitol.* **9** 429-35

[226] Fitzgerald J R and Holden M T G 2016 Genomics of Natural Populations of *Staphylococcus aureus Annu. Rev. Microbiol.* **70** 459-78

[227] Foster T 1996 *Staphylococcus* Baron S. ed University of Texas Medical Branch at Galveston

[228] Freney J, Kloos W E, Hajek V, Webster J A, Bes M, Brun Y and Vernozy-Rozand C 1999 Recommended minimal standards for description of new staphylococcal species. Subcommittee on the taxonomy of staphylococci and streptococci of the International Committee on Systematic Bacteriology *Int.J.Syst.Bacteriol.* **49** 489–502

[229] Sasaki T, Tsubakishita S, Tanaka Y, Sakusabe A, Ohtsuka M, Hirotaki S, Kawakami T, Fukata T and Hiramatsu K 2010 Multiplex-PCR method for species identification of coagulase-positive staphylococci *J. Clin. Microbiol.* **48** 765–9

[230] Harris L G, Foster S J and Richards R G 2002 An introduction to *Staphylococcus aureus*, and techniques for identifying and quantifying *S. aureus* adhesisn in relation to adhesion to biomaterials: Review *Eur. Cells Mater.* **4** 39–60

[231] Wertheim H F, Melles D C, Vos M C, van Leeuwen W, van Belkum A, Verbrugh H a and Nouwen J L 2005 The role of nasal carriage in *Staphylococcus aureus* infections *Lancet Infect. Dis.* **5** 751–62

[232] Armstrong-Esther C A 1976 Carriage patterns of *Staphylococcus aureus* in a healthy non-hospital population of adults and children. *Ann. Hum. Biol.* **3** 221–7

[233] Peacock S J, Justice A, Griffiths D, de Silva G D I, Kantzanou M N, Crook D, Sleeman K and Day N P J 2003 Determinants of acquisition and carriage of *Staphylococcus aureus* in infancy. *J. Clin. Microbiol.* **41** 5718–25

[234] Myles I A and Datta S K 2012 *Staphylococcus aureus*: an introduction. *Semin. Immunopathol.* **34** 181–4

[235] Spellberg B, Guidos R, Gilbert D, Bradley J, Boucher H W, Scheld W M, Bartlett J G and Edwards Jr. J 2008 The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America *Clin.Infect.Dis.* **46** 155–64

[236] Sengupta S, Chattopadhyay M K and Grossart H-P 2013 The multifaceted roles of antibiotics and antibiotic resistance in nature *Front. Microbiol.* **4**

[237] Lowy F D 2003 Antimicrobial resistance: The example of *Staphylococcus aureus J. Clin. Invest.* **111** 1265–73

[238] Peacock S J and Paterson G K 2015 Mechanisms of Methicillin Resistance in *Staphylococcus aureus Annu. Rev. Biochem.* **84** 577–601

[239] Jevons M P 1961 ‘Celbenin’ - resistant *Staphylococci Br. Med. J.* **1** 124

[240] Griffiths C, Lamagni T L, Crowcroft N S, Duckworth G and Rooney C 2004 Trends in MRSA in England and Wales: analysis of morbidity and mortality data for 1993-2002. *Health Stat. Q.* 15–22

[241] Köck R, Becker K, Cookson B, van Gemert-Pijnen J E, Harbarth S, Kluytmans J, Mielke M, Peters G, Skov R L, Struelens M J, Tacconelli E, Navarro Torné A, Witte W and Friedrich A W 2010 Methicillin-resistant *Staphylococcus aureus* (MRSA): burden of disease and control challenges in Europe. *Euro Surveill.* **15** 19688

[242] Herold B C, Immergluck L C, Maranan M C, Lauderdale D S, Gaskin R E, Boyle-Vavra S, Leitch C D and Daum R S 1998 Community-acquired methicillin-resistant *Staphylococcus aureus* in children with no identified predisposing risk. *JAMA* **279** 593–8

[243] Rasigade J P, Laurent F, Lina G, Meugnier H, Bes M, Vandenesch F, Etienne J and Tristan A 2010 Global distribution and evolution of Panton-Valentine leukocidin-positive methicillin-susceptible *_Staphylococcus aureus_*, 1981-2007 *J.Infect.Dis.* **201** 1589–97

[244] Francis J S, Doherty M C, Lopatin U, Johnston C P, Sinha G, Ross T, Cai M, Hansel N N, Perl T, Ticehurst J R, Carroll K, Thomas D L, Nuermberger E and Bartlett J G 2005 Severe community-onset pneumonia in healthy adults caused by methicillin-resistant *Staphylococcus aureus* carrying the Panton-Valentine leukocidin genes *Clin. Infect. Dis.* **40** 100–7

[245] David M Z and Daum R S 2010 Community-associated methicillin-resistant *Staphylococcus aureus*: epidemiology and clinical consequences of an emerging epidemic. *Clin. Microbiol. Rev.* **23** 616–87

[246] Baba T, Takeuchi F, Kuroda M, Yuzawa H, Aoki K I, Oguchi A, Nagai Y, Iwama N, Asano K, Naimi T, Kuroda H, Cui L, Yamamoto K and Hiramatsu K 2002 Genome and virulence determinants of high virulence community-acquired MRSA *Lancet* **359** 1819–27

[247] Okuma K, Iwakawa K, Turnidge J D, Grubb W B, Bell J M, O'Brien F G, Coombs G W, Pearman J W, Tenover F C, Kapi M, Tiensasitorn C, Ito T and Hiramatsu K 2002 Dissemination of new methicillin-resistant *Staphylococcus aureus* clones in the community. *J. Clin. Microbiol.* **40** 4289–94

[248] Daum R S, Ito T, Hiramatsu K, Hussain F, Mongkolrattanothai K, Jamklang M and Boyle-Vavra S 2002 A novel methicillin-resistance cassette in community-acquired methicillin-resistant *Staphylococcus aureus* isolates of diverse genetic backgrounds *J.Infect.Dis.* **186** 1344–7

[249] Ma X X, Ito T, Tiensasitorn C, Jamklang M, Chongtrakool P, Boyle-Vavra S, Daum R S and Hiramatsu K 2002 Novel Type of Staphylococcal Cassette Chromosome mec Identified in Community-Acquired Methicillin-Resistant *Staphylococcus aureus* Strains *Antimicrob. Agents Chemother.* **46** 1147–52

[250] Karahan Z C, Tekeli A, Adaleti R, Koyuncu E, Dolapci I and Akan O A 2008 Investigation of Panton-Valentine Leukocidin Genes and SCC*mec* Types in Clinical *Staphylococcus aureus* Isolates from Turkey *Microb. Drug Resist.* **14** 203–10

[251] Buck J M, Como-Sabetti K, Harriman K H, Danila R N, Boxrud D J, Glennen A and Lynfield R 2005 Community-associated methicillin-resistant *Staphylococcus aureus*, Minnesota, 2000-2003 *Emerg.Infect.Dis.* **11** 1532–8

[252] Morrison M A, Hageman J C and Klevens R M 2006 Case definition for community-

associated methicillin-resistant *Staphylococcus aureus*. *J. Hosp. Infect.* **62** 241

[253] Miller L G, Perdreau-Remington F, Bayer A S, Diep B, Tan N, Bharadwa K, Tsui J, Perlroth J, Shay A, Tagudar G, Ibebuogu U and Spellberg B 2007 Clinical and epidemiologic characteristics cannot distinguish community-associated methicillin-resistant *Staphylococcus aureus* infection from methicillin-susceptible *S. aureus* infection: a prospective investigation. *Clin. Infect. Dis.* **44** 471–82

[254] McDougal L K, Fosheim G E, Nicholson A, Bulens S N, Limbago B M, Shearer J E S, Summers A O and Patel J B 2010 Emergence of resistance among USA300 methicillin-resistant *Staphylococcus aureus* isolates causing invasive disease in the United States *Antimicrob. Agents Chemother.* **54** 3804–11

[255] David M Z, Glikman D, Crawford S E, Peng J, King K J, Hostetler M A, Boyle-Vavra S and Daum R S 2008 What is community-associated methicillin-resistant *Staphylococcus aureus*? *J. Infect. Dis.* **197** 1235–43

[256] Popovich K J, Weinstein R A and Hota B 2008 Are community-associated methicillin-resistant *Staphylococcus aureus* (MRSA) strains replacing traditional nosocomial MRSA strains? *Clin. Infect. Dis.* **46** 787–94

[257] Jenkins T C, McCollister B D, Sharma R, McFann K K, Madinger N E, Barron M, Bessesen M, Price C S and Burman W J 2009 Epidemiology of Healthcare-Associated Bloodstream Infection Caused by USA300 Strains of Methicillin-Resistant *Staphylococcus aureus* in 3 Affiliated Hospitals • *Infect. Control Hosp. Epidemiol.* **30** 233–41

[258] Gonzalez B E, Rueda A M, Shelburne III S A, Musher D M, Hamill R J and Hulten K G 2006 Community-Associated Strains of Methicillin-Resistant *Staphylococcccus aureus* as the Cause of Healthcare-Associated Infection *Infect.Control Hosp.Epidemiol.* **27** 1051–6

[259] Seybold U, Kourbatova E V, Johnson J G, Halvosa S J, Wang Y F, King M D, Ray S M and Blumberg H M 2006 Emergence of Community-Associated Methicillin-Resistant *Staphylococcus aureus* USA300 Genotype as a Major Cause of Health Care Associated Blood Stream Infections *Clin. Infect. Dis.* **42** 647–56

[260] D’Agata E M C, Webb G F, Horn M A, Moellering R C and Ruan S 2009 Modeling the invasion of community-acquired methicillin-resistant *Staphylococcus aureus* into hospitals. *Clin. Infect. Dis.* **48** 274–84

[261] Epstein C R, Yam W C, Peiris J S M and Epstein R J 2009 Methicillin-resistant commensal staphylococci in healthy dogs as a potential zoonotic reservoir for community-acquired antibiotic resistance *Infect. Genet. Evol.* **9** 283–5

[262] Bloemendaal A L A, Brouwer E C and Fluit A C 2010 Methicillin resistance transfer from *Staphylocccus epidermidis* to methicillin-susceptible *Staphylococcus aureus* in a patient during antibiotic therapy *PLoS One* **5**

[263] Robinson D A and Enright M C 2003 Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus Antimicrob.Agents Chemother.* **47** 3926–34

[264] Grundmann H, Hori S, Enright M C, Webster C, Tami A, Feil E J and Pitt T 2002 Determining the genetic structure of the natural population of *Staphylococcus aureus*: a comparison of multilocus sequence typing with pulsed-field gel electrophoresis, randomly amplified polymorphic DNA analysis, and phage typing. *J. Clin. Microbiol.* **40** 4544–6

[265] Musser J M and Kapur V 1992 Clonal analysis of methicillin-resistant *Staphylococcus aureus* strains from intercontinental sources: association of the mec gene with divergent phylogenetic lineages implies dissemination by horizontal transfer and recombination. *J. Clin. Microbiol.* **30** 2058–63

[266] Musser J M, Schlievert P M, Chow A W, Ewan P, Kreiswirth B N, Rosdahl V T, Naidu A S, Witte W and Selander R K 1990 A single clone of *Staphylococcus aureus* causes the majority of cases of toxic shock syndrome. *Proc. Natl. Acad. Sci.* **87** 225–9

[267] Stefani S, Chung D R, Lindsay J A, Friedrich A W, Kearns A M, Westh H and Mackenzie F M 2012 Meticillin-resistant *Staphylococcus aureus* (MRSA): global epidemiology and harmonisation of typing methods. *Int. J. Antimicrob. Agents* **39** 273–82

[268] Deurenberg R H, Vink C, Kalenic S, Friedrich A W, Bruggeman C A and Stobberingh E E 2007 The molecular evolution of methicillin-resistant *Staphylococcus aureus Clin. Microbiol. Infect.* **13** 222–35

[269] Mediavilla J R, Chen L, Mathema B and Kreiswirth B N 2012 Global epidemiology of community-associated methicillin resistant *Staphylococcus aureus* (CA-MRSA) *Curr. Opin. Microbiol.* **15** 588–95

[270] Enright M C, Day N P J, Davies C E, Peacock S J and Spratt B G 2000 Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus J. Clin. Microbiol.* **38** 1008–15

[271] Feil E J, Cooper J E, Grundmann H, Robinson D A, Enright M C, Berendt T, Peacock S J, Smith J M, Murphy M, Spratt B G, Moore C E and Day N P J 2003 How clonal is *Staphylococcus aureus*? *J. Bacteriol.* **185** 3307–16

[272] Lindsay J A, Moore C E, Day N P, Peacock S J, Witney A A, Stabler R A, Husain S E, Butcher P D and Hinds J 2006 Microarrays Reveal that Each of the Ten Dominant Lineages of *Staphylococcus aureus* Has a Unique Combination of Surface-Associated and Regulatory Genes *J. Bacteriol.* **188** 669–76

[273] King J M, Kulhankova K, Stach C S, Vu B G and Salgado-Pabón W 2016 Phenotypes and Virulence among *Staphylococcus aureus* USA100, USA200, USA300, USA400, and USA600 Clonal Lineages. *mSphere* **1** e00071-16

[274] Planet P J, Narechania A, Chen L, Mathema B, Boundy S, Archer G and Kreiswirth B 2016 Architecture of a Species: Phylogenomics of *Staphylococcus aureus Trends Microbiol.***25** 153-166

[275] Aanensen D M, Feil E J, Holden M T G, Dordel J, Yeats C A, Fedosejev A, Goater R, Castillo-Ram??rez S, Corander J, Colijn C, Chlebowicz M A, Schouls L, Heck M, Pluister G, Ruimy R, Kahlmeter G, ??hman J, Matuschek E, Friedrich A W, Parkhill J, Bentley S D, Spratt B G, Grundmannj H, Krziwanek K, Stumvoll S, Koller W, Denis O, Struelens M, Nashev D, Budimir A, Kalenic S, Pieridou-Bagatzouni D, Jakubu V, Zemlickova H, Westh H, Larsen A R, Skov R, Laurent F, Ettienne J, Strommenger B, Witte W, Vourli S, Vatopoulos A, Vainio A, Vuopio-Varkila J, Fuzi M, Ungv??ri E, Murchan S, Rossney A, Miklasevics E, Balode A, Haraldsson G, Kristinsson K G, Monaco M, Pantosti A, Borg M, Van Santen-Verheuvel M, Huijsdens X, Marstein L, Jacobsen T, Simonsen G S, Airesde-Sousa M, De Lencastre H, Luczak-Kadlubowska A, Hryniewicz W, Straut M, Codita I, Perez-Vazquez M, Iglesias J O, Spik V C, Mueller-Premru M, Haeggman S, Olsson-Liljequist B, Ellington M and Kearns A 2016 Whole-genome sequencing for routine pathogen surveillance in public

health: A population snapshot of invasive *Staphylococcus aureus* in Europe *MBio* **7** e00444-16

[276] Méric G, Miragaia M, De Been M, Yahara K, Pascoe B, Mageiros L, Mikhail J, Harris L G, Wilkinson T S, Rolo J, Lamble S, Bray J E, Jolley K A, Hanage W P, Bowden R, Maiden M C J, Mack D, De Lencastre H, Feil E J, Corander J and Sheppard S K 2015 Ecological overlap and horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis Genome Biol. Evol.* **7** 1313–28

[277] Ito T, Kuwahara-Arai K, Katayama Y, Uehara Y, Han X, Kondo Y and Hiramatsu K 2014 Staphylococcal cassette chromosome *mec* (SCC*mec*) analysis of MRSA *Methods Mol. Biol.* **1085** 131–48

[278] Ito T, Katayama Y, Asada K, Mori N, Tsutsumimoto K, Tiensasitorn C and Hiramatsu K 2001 Structural comparison of three types of staphylococcal cassette chromosome *mec* integrated in the chromosome in methicillin-resistant *Staphylococcus aureus Antimicrob. Agents Chemother.* **45** 1323–36

[279] Ito T, Hiramatsu K, Oliveira D C, De Lencastre H, Zhang K, Westh H, O'Brien F, Giffard P M, Coleman D, Tenover F C, Boyle-Vavra S, Skov R L, Enright M C, Kreiswirth B, Kwan S K, Grundmann H, Laurent F, Sollid J E, Kearns A M, Goering R, John J F, Daum R and Soderquist B 2009 Classification of staphylococcal cassette chromosome *mec* (SCC*mec*): Guidelines for reporting novel SCC*mec* elements *Antimicrob. Agents Chemother.* **53** 4961–7

[280] Pinho M G, de Lencastre H and Tomasz A 2001 An acquired and a native penicillin-binding protein cooperate in building the cell wall of drug-resistant staphylococci *Proc.Natl.Acad.Sci.U.S.A* **98** 10886–91

[281] Katayama Y, Ito T and Hiramatsu K 2000 A new class of genetic element, staphylococcus cassette chromosome *mec*, encodes methicillin resistance in *Staphylococcus aureus Antimicrob. Agents Chemother.* **44** 1549–55

[282] Boundy S, Safo M K, Wang L, Musayev F N, O'Farrell H C, Rife J P and Archer G L 2013 Characterization of the *Staphylococcus aureus* rRNA methyltransferase encoded by *orfX*, the gene containing the staphylococcal chromosome Cassette *mec* (SCC*mec*) insertion site. *J. Biol. Chem.* **288** 132–40

[283] Noto M J and Archer G L 2006 A subset of *Staphylococcus aureus* strains harboring staphylococcal cassette chromosome *mec* (SCC*mec*) type IV is deficient in CcrAB-mediated SCC*mec* excision *Antimicrob. Agents Chemother.* **50** 2782–8

[284] Tsubakishita S, Kuwahara-Arai K, Sasaki T and Hiramatsu K 2010 The origin and molecular evolution of the determinant of methicillin-resistance in staphylococci *Antimicrob Agents Chemother* **54** 4352–9

[285] Katayama Y, Zhang H Z, Hong D and Chambers H F 2003 Jumping the barrier to beta-lactam resistance in *Staphylococcus aureus J Bacteriol* **185** 5465–72

[286] Noto M J, Kreiswirth B N, Monk A B and Archer G L 2008 Gene acquisition at the insertion site for SCC*mec*, the genomic island conferring methicillin resistance in *Staphylococcus aureus J. Bacteriol.* **190** 1276–83

[287] Waldron D E and Lindsay J A 2006 Sau1: A novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into *Staphylococcus aureus* and

between *S. aureus* isolates of different lineages *J. Bacteriol.* **188** 5578–85

[288] Tomasz A, Nachman S and Leaf H 1991 Stable classes of phenotypic expression in methicillin-resistant clinical isolates of staphylococci *Antimicrob. Agents Chemother.* **35** 124–9

[289] Wiegand I, Hilpert K and Hancock R E W 2008 Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **3** 163–75

[290] Berger-Bächi B and Rohrer S 2002 Factors influencing methicillin resistance in staphylococci *Arch. Microbiol.* **178** 165–71

[291] Hartman B J and Tomasz A 1986 Expression of methicillin resistance in heterogeneous strains of *Staphylococcus aureus Antimicrob. Agents Chemother.* **29** 85–92

[292] de Lencastre H, Figueiredo a M and Tomasz a 1993 Genetic control of population structure in heterogeneous strains of methicillin resistant *Staphylococcus aureus*. *Eur. J. Clin. Microbiol. Infect. Dis.* **12 Suppl 1** S13–8

[293] De Lencastre H, Chung M and Westh H 2000 Archaic strains of methicillin-resistant *Staphylococcus aureus*: Molecular and microbiological properties of isolates from the 1960s in Denmark *Microb. Drug Resist. Mech. Epidemiol. Dis.* **6** 1–10

[294] Ryffel C, Strässle A, Kayser F H and Berger-Bächi B 1994 Mechanisms of heteroresistance in methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **38** 724–8

[295] Kuwahara-Arai K, Kondo N, Hori S, Tateda-Suzuki E and Hiramatsu K 1996 Suppression of methicillin resistance in a *mecA*-containing pre-methicillin-resistant *Staphylococcus aureus* strain is caused by the *mecI*-mediated repression of PBP 2' production. *Antimicrob. Agents Chemother.* **40** 2680–5

[296] Hiramatsu K, Asada K, Suzuki E, Okonogi K and Yokota T 1992 Molecular cloning and nucleotide sequence determination of the regulator region of *mecA* gene in methicillin-resistant *Staphylococcus aureus* (MRSA). *FEBS Lett.* **298** 133–6

[297] Kondo N, Kuwahara-Arai K, Kuroda-Murakami H, Tateda-Suzuki E and Hiramatsu K 2001 Eagle-type methicillin resistance: New phenotype of high methicillin resistance under *mec* regulator gene control *Antimicrob. Agents Chemother.* **45** 815–24

[298] Aiba Y, Katayama Y, Hishinuma T, Murakami-Kuroda H, Cui L and Hiramatsu K 2013 Mutation of RNA polymerase β-subunit gene promotes heterogeneous-to-homogeneous conversion of β-lactam resistance in methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **57** 4861–71

[299] Chong Y P, Park S-J, Kim H S, Kim E S, Kim M-N, Park K-H, Kim S-H, Lee S-O, Choi S-H, Jeong J-Y, Woo J H and Kim Y S 2013 Persistent *Staphylococcus aureus* bacteremia: a prospective analysis of risk factors, outcomes, and microbiologic and genotypic characteristics of isolates. *Medicine (Baltimore).* **92** 98–108

[300] Cohen N R, Lobritz M A and Collins J J 2013 Microbial persistence and the road to drug resistance *Cell Host Microbe* **13** 632–42

[301] Brauner A, Fridman O, Gefen O and Balaban N Q 2016 Distinguishing between resistance, tolerance and persistence to antibiotic treatment. *Nat. Rev. Microbiol.* **14** 320–30

[302] Sandegren L and Andersson D I 2009 Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nat. Rev. Microbiol.* **7** 578–88

[303] Sonti R V. and Roth J R 1989 Role of gene duplications in the adaptation of *Salmonella typhimurium* to growth on limiting carbon sources *Genetics* **123** 19–28

[304] Gladman S L, Seemann T, Gao W, Stinear T P, Howden B P, Monk I R and Tobias N J 2015 Large tandem chromosome expansions facilitate niche adaptation during persistent infection with drug-resistant *Staphylococcus aureus Microb. Genomics* **1**

[305] Gill S R, Fouts D E, Archer G L, Mongodin E F, Deboy R T, Ravel J, Paulsen I T, Kolonay J F, Brinkac L, Beanan M, Dodson R J, Daugherty S C, Madupu R, Angiuoli S V, Durkin a S, Haft D H, Vamathevan J, Khouri H, Utterback T, Lee C, Dimitrov G, Jiang L, Qin H, Weidman J, Tran K, Kang K, Hance I R, Nelson K E and Fraser C M 2005 Insights on Evolution of Virulence and Resistance from the Complete Genome Analysis of an Early Methicillin-Resistant *Staphylococcus aureus* Strain and a Biofilm-Producing Methicillin-Resistant *Staphylococcus epidermidis* Strain *J. Bacteriol.* **187** 2426–38

[306] Lindsay J A and Holden M T G 2006 Understanding the rise of the superbug: investigation of the evolution and genomic variation of *Staphylococcus aureus Funct. Integr. Genomics* **6** 186–201

[307] Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama A, Mizutani-Ui Y, Takahashi N K, Sawano T, Inoue R, Kaito C, Sekimizu K, Hirakawa H, Kuhara S, Goto S, Yabuzaki J, Kanehisa M, Yamashita A, Oshima K, Furuya K, Yoshino C, Shiba T, Hattori M, Ogasawara N, Hayashi H and Hiramatsu K 2001 Whole genome sequencing of meticillin-resistant *Staphylococcus aureus Lancet* **357** 1225–40

[308] Holden M T G, Feil E J, Lindsay J a, Peacock S J, Day N P J, Enright M C, Foster T J, Moore C E, Hurst L, Atkin R, Barron A, Bason N, Bentley S D, Chillingworth C, Chillingworth T, Churcher C, Clark L, Corton C, Cronin A, Doggett J, Dowd L, Feltwell T, Hance Z, Harris B, Hauser H, Holroyd S, Jagels K, James K D, Lennard N, Line A, Mayes R, Moule S, Mungall K, Ormond D, Quail M a, Rabbinowitsch E, Rutherford K, Sanders M, Sharp S, Simmonds M, Stevens K, Whitehead S, Barrell B G, Spratt B G and Parkhill J 2004 Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc. Natl. Acad. Sci. U. S. A.* **101** 9786–91

[309] Diep B A, Gill S R, Chang R F, Phan T H, Chen J H, Davidson M G, Lin F, Lin J, Carleton H A, Mongodin E F, Sensabaugh G F and Perdreau-Remington F 2006 Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet* **367** 731–9

[310] Herron-Olson L, Fitzgerald J R, Musser J M and Kapur V 2007 Molecular correlates of host specialization in *Staphylococcus aureus PLoS One* **2** e1120

[311] Lindsay J A and Holden M T G 2004 *Staphylococcus aureus*: Superbug, super genome? *Trends Microbiol.* **12** 378–85

[312] Driebe E M, Sahl J W, Roe C, Bowers J R, Schupp J M, Gillece J D, Kelley E, Price L B, Pearson T R, Hepp C M, Brzoska P M, Cummings C A, Furtado M R, Andersen P S, Stegger

M, Engelthaler D M and Keim P S 2015 Using whole genome analysis to examine recombination across diverse sequence types of *Staphylococcus aureus PLoS One* **10** e0130955

[313] Jamrozy D M, Mohamed N, Anderson A S, Harris S R, Parkhill J, Tan C Y, Peacock S J and Holden M T G 2016 Pan-genomic perspective on the evolution of the *Staphylococcus aureus* USA300 epidemic *Microb. Genomics* **2**

[314] Holden M T G, Hsu L-Y, Kurt K, Weinert L a, Mather A E, Harris S R, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlicková H, Coombs G, Kearns A M, Hill R L R, Edgeworth J, Gould I, Gant V, Cooke J, Edwards G F, McAdam P R, Templeton K E, McCann A, Zhou Z, Castillo-Ramírez S, Feil E J, Hudson L O, Enright M C, Balloux F, Aanensen D M, Spratt B G, Fitzgerald J R, Parkhill J, Achtman M, Bentley S D and Nübel U 2013 A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res.* **23** 653–64

[315] Harris S R, Feil E J, Holden M T G, Quail M a, Nickerson E K, Chantratita N, Gardete S, Tavares A, Day N, Lindsay J A, Edgeworth J D, de Lencastre H, Parkhill J, Peacock S J and Bentley S D 2010 Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327** 469–74

[316] Köser C U, Holden M T, Ellington M J, Cartwright E J, Brown N M, Ogilvy-Stuart A L, Hsu L Y, Chewapreecha C, Croucher N J, Harris S R, Sanders M, Enright M C, Dougan G, Bentley S D, Parkhill J, Fraser L J, Betley J R, Schulz-Trieglaff O B, Smith G P and Peacock S J 2012 Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak *N.Engl.J.Med.* **366** 2267–75

[317] Thomas C M and Nielsen K M 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3** 711–21

[318] Morikawa K, Takemura A J, Inose Y, Tsai M, Nguyen Thi L T, Ohta T and Msadek T 2012 Expression of a Cryptic Secondary Sigma Factor Gene Unveils Natural Competence for DNA Transformation in *Staphylococcus aureus PLoS Pathog.* **8**

[319] Malachowa N and Deleo F R 2010 Mobile genetic elements of *Staphylococcus aureus Cell. Mol. Life Sci.* **67** 3057–71

[320] Frost L S, Leplae R, Summers A O and Toussaint A 2005 Mobile genetic elements: the agents of open source evolution *Nat.Rev.Microbiol.* **3** 722–32

[321] Goerke C, Pantucek R, Holtfreter S, Schulte B, Zink M, Grumann D, Bröker B M, Doskar J and Wolz C 2009 Diversity of prophages in dominant *Staphylococcus aureus* clonal lineages *J. Bacteriol.* **191** 3462–8

[322] Moore P C L and Lindsay J A 2001 Genetic variation among hospital isolates of methicillin-sensitive *Staphylococcus aureus*: Evidence for horizontal transfer of virulence genes *J. Clin. Microbiol.* **39** 2760–7

[323] Sumby P and Waldor M K 2003 Transcription of the toxin genes present within the Staphylococcal phage phiSa3ms is intimately linked with the phage's life cycle. *J. Bacteriol.* **185** 6841–51

[324] Coleman D, Knights J, Russell R, Shanley D, Birkbeck T H, Dougan G and Charles I 1991 Insertional inactivation of the *Staphylococcus aureus* beta-toxin by bacteriophage phi 13

occurs by site- and orientation-specific integration of the phi 13 genome. *Mol. Microbiol.* **5** 933–9

[325] Iandolo J J, Worrell V, Groicher K H, Qian Y, Tian R, Kenton S, Dorman A, Ji H, Lin S, Loh P, Qi S, Zhu H and Roe B a 2002 Comparative analysis of the genomes of the temperate bacteriophages phi 11, phi 12 and phi 13 of *Staphylococcus aureus* 8325. *Gene* **289** 109–18

[326] Yarwood J M, McCormick J K, Paustian M L, Orwin P M, Kapur V and Schlievert P M 2002 Characterization and expression analysis of *Staphylococcus aureus* pathogenicity island 3. Implications for the evolution of staphylococcal pathogenicity islands *J. Biol. Chem.* **277** 13138–47

[327] Alibayov B, Baba-Moussa L, Sina H, Zdeňková K and Demnerová K 2014 *Staphylococcus aureus* mobile genetic elements *Mol. Biol. Rep.* **41** 5005–18

[328] Sato'o Y, Omoe K, Ono H K, Nakane A and Hu D L 2013 A novel comprehensive analysis method for *Staphylococcus aureus* pathogenicity islands *Microbiol. Immunol.* **57** 91–9

[329] Novick R P and Subedi A 2007 The SaPIs: Mobile pathogenicity islands of *Staphylococcus Chem. Immunol. Allergy* **93** 42–57

[330] Úbeda C, Maiques E, Barry P, Matthews A, Tormo M Á, Lasa Í, Novick R P and Penadés J R 2008 SaPI mutations affecting replication and transfer and enabling autonomous replication in the absence of helper phage *Mol. Microbiol.* **67** 493–503

[331] Lindsay J A, Ruzin A, Ross H F, Kurepina N and Novick R P 1998 The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus Mol. Microbiol.* **29** 527–43

[332] Sievers F, Wilm A, Dineen D, Gibson T J, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson J D and Higgins D G 2011 Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7** 539

[333] Huson D H and Bryant D 2006 Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23** 254–67

[334] Croan D G, Morrison D a and Ellis J T 1997 Evolution of the genus *Leishmania* revealed by comparison of DNA and RNA polymerase gene sequences. *Mol. Biochem. Parasitol.* **89** 149–59

[335] Baneth G, Koutinas A F, Solano-Gallego L, Bourdeau P and Ferrer L 2008 Canine leishmaniosis – new concepts and insights on an expanding zoonosis: part one *Trends Parasitol.* **24** 324–30

[336] Noyes H A, Arana B A, Chance M L and Maingon R 1997 The *Leishmania hertigi* (Kinetoplastida; Trypanosomatidae) Complex and the Lizard *Leishmania*: Their Classification and Evidence for a Neotropical Origin of the *Leishmania-Endotrypanum* Clade *J. Eukaryot. Microbiol.* **44** 511–7

[337] Marcili A, Sperança M A, da Costa A P, Madeira M de F, Soares H S, Sanches C de O C C, Acosta I da C L, Girotto A, Minervino A H H, Horta M C, Shaw J J and Gennari S M 2014 Phylogenetic relationships of *Leishmania* species based on trypanosomatid barcode (SSU

rDNA) and gGAPDH genes: Taxonomic revision of *Leishmania (L.) infantum chagasi* in South America *Infect. Genet. Evol.* **25** 44–51

[338] Breton M, Tremblay M J, Ouellette M and Papadopoulou B 2005 Live nonpathogenic parasitic vector as a candidate vaccine against visceral leishmaniasis. *Infect. Immun.* **73** 6372–82

[339] Taylor V M, Muñoz D L, Cedeño D L, Vélez I D, Jones M A and Robledo S M 2010 *Leishmania tarentolae*: Utility as an in vitro model for screening of antileishmanial agents *Exp. Parasitol.* **126** 471–5

[340] Heisch R.B. 1958 On *leishmania adleri* sp. nov. from lacertid lizards (*Latastia* sp.) in Kenya. *Ann. Trop. Med. Parasitol.* **52** 68–71

[341] Adler S 1962 The behaviour of a lizard *Leishmania* in hamsters and baby mice *Rev. Inst. Med. Trop. Sao Paulo* **4** 61–4

[342] Haile T T and Lemma A 1977 Isolation of *Leishmania* parasites from *Arvicanthis* in Ethiopia. *Trans. R. Soc. Trop. Med. Hyg.* **71** 180–1

[343] Gebresilassie A, Abbasi I, Aklilu E, Yared S, Kirstein O D, Moncaz A, Tekie H, Balkew M, Warburg A, Hailu A and Gebre-Michael T 2015 Host-feeding preference of *Phlebotomus orientalis* (Diptera: Psychodidae) in an endemic focus of visceral leishmaniasis in northern Ethiopia. *Parasit. Vectors* **8** 270

[344] Zijlstra E E and el-Hassan A M 2001 Leishmaniasis in Sudan. Visceral leishmaniasis. *Trans. R. Soc. Trop. Med. Hyg.* **95 Suppl 1** S27-58

[345] Haile T and Anderson S D 2006 Visceral leishmaniasis in northern Ethiopia *East Afr. Med. J.* **83** 389–92

[346] Baleela R, Llewellyn M S, Fitzpatrick S, Kuhls K, Schönian G, Miles M A, Mauricio I L, Hassan M, Osman O, El-Raba'a F, Schallig H, Elnaiem D, Singh N, Mishra J, Singh R, Singh S, Zink A, Spigelman M, Schraut B, Greenblatt C, Nerlich A, Donoghue H, Dereure J, El-Safi S, Bucheton B, Boni M, Kheir M, Davoust B, Pratlong F, Feugier E, Lambert M, Dessein A, Dedet J, Seaman J, Mercer A, Sondorp E, Mueller Y, Nackers F, Ahmed K, Boelaert M, Djoumessi J, Eltigani R, Gorashi H, Hammam O, Ritmeijer K, Salih N, Worku D, Etard J, Chappuis F, Zijlstra E, el-Hassan A, Kuhls K, Mauricio I, Pratlong F, Presber W, Schonian G, Mauricio I, Howard M, Stothard J, Miles M, Mauricio I, Yeo M, Baghaei M, Doto D, Pratlong F, Zemanova E, Dedet J, Lukes J, Miles M, Zemanova E, Jirku M, Mauricio I, Horak A, Miles M, Lukes J, Kuhls K, Keilonat L, Ochsenreither S, Schaar M, Schweynoch C, Presber W, Schönian G, Alam M, Kuhls K, Schweynoch C, Sundar S, Rijal S, Shamsuzzaman A, Raju B, Salotra P, Dujardin J, Schönian G, Gelanew T, Kuhls K, Hurissa Z, Weldegebreal T, Hailu W, Kassahun A, et al 2014 *Leishmania donovani* populations in Eastern Sudan: temporal structuring and a link between human and canine transmission *Parasit. Vectors* **7** 496

[347] Maleki Ravasan N, Javadian E, Mohebali M, Dalimi Asl A, Sadraei J, Zarei Z and Oshaghi M A 2008 Natural infection of sand flies *Sergentomyia dentata* in Ardebil to Lizard *Leishmania Modares J. Med. Sci. Pathobiol.* **10** 65–73

[348] Hoogstraal H, Heyneman D, Dietlein D R, Brown H G, Jr R T, Van Peenen P, Saber A H and Rohrs L C 1963 Leishmaniasis in the Sudan Republic: epidemiological findings. *Bull. World Health Organ.* **28** 263–5

[349] Hotez P J, Woc-Colburn L and Bottazzi M E 2014 Neglected tropical diseases in Central America and Panama: Review of their prevalence, populations at risk and impact on regional development *Int. J. Parasitol.* **44** 597–603

[350] Mueller Y K, Nackers F, Ahmed K A, Boelaert M, Djoumessi J C, Eltigani R, Gorashi H A, Hammam O, Ritmeijer K, Salih N, Worku D, Etard J F and Chappuis F 2012 Burden of Visceral Leishmaniasis in Villages of Eastern Gedaref State, Sudan: An Exhaustive Cross-Sectional Survey *PLoS Negl. Trop. Dis.* **6** e1872

[351] Alvar J, Aparicio P, Aseffa A, Den Boer M, Cañavate C, Dedet J P, Gradoni L, Ter Horst R, López-Vélez R and Moreno J 2008 The relationship between leishmaniasis and AIDS: The second 10 years *Clin. Microbiol. Rev.* **21** 334–59

[352] el-Hassan A M and Zijlstra E E 2001 Leishmaniasis in Sudan. Cutaneous leishmaniasis. *Trans. R. Soc. Trop. Med. Hyg.* **95 Suppl 1** S1-17

[353] El Baidouri F, Diancourt L, Berry V, Chevenet F, Pratlong F, Marty P and Ravel C 2013 Genetic Structure and Evolution of the *Leishmania* Genus in Africa and Eurasia: What Does MLSA Tell Us. ed G Schönian *PLoS Negl. Trop. Dis.* **7** e2255

[354] Noyes H A, Arana B A, Chance M L and Maingon R 1997 The *Leishmania hertigi* (Kinetoplastida; *Trypanosomatidae*) Complex and the *Lizard Leishmania*: Their Classification and Evidence for a Neotropical Origin of the *Leishmania-Endotrypanum* Clade *J. Eukaryot. Microbiol.* **44** 511–7

[355] Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A and Barrell B 2000 Artemis: sequence visualization and annotation *Bioinformatics* **16** 944–5

[356] Quinlan A R and Hall I M 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26** 841–2

[357] Frith M C 2011 A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* **39** e23

[358] Danecek P, Auton A, Abecasis G, Albers C A, Banks E, DePristo M A, Handsaker R E, Lunter G, Marth G T, Sherry S T, McVean G and Durbin R 2011 The variant call format and VCFtools. *Bioinformatics* **27** 2156–8

[359] Young M D, Wakefield M J, Smyth G K and Oshlack A 2010 Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11** R14

[360] Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk B P, Carrington M, Depledge D P, Fischer S, Gajria B, Gao X, Gardner M J, Gingle A, Grant G, Harb O S, Heiges M, Hertz-Fowler C, Houston R, Innamorato F, Iodice J, Kissinger J C, Kraemer E, Li W, Logan F J, Miller J A, Mitra S, Myler P J, Nayak V, Pennington C, Phan I, Pinney D F, Ramasamy G, Rogers M B, Roos D S, Ross C, Sivam D, Smith D F, Srinivasamoorthy G, Stoeckert C J, Subramanian S, Thibodeau R, Tivey A, Treatman C, Velarde G and Wang H 2010 TriTrypDB: a functional genomic resource for the *Trypanosomatidae*. *Nucleic Acids Res.* **38** D457-62

[361] Li L, Stoeckert C J and Roos D S 2003 OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes *Genome Res.* **13** 2178–89

[362] Marques C A, Dickens N J, Paape D, Campbell S J, McCulloch R, Costa A, Hood I, Berger J, O'Donnell M, Langston L, Stillman B, Leonard A, Mechali M, Rienzi S, Lindstrom K, Mann T, Noble W, Raghuraman M, Brewer B, Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C, Thomas S, Green A, Sturm N, Campbell D, Myler P, Reynolds D, Cliffe L, Forstner K, Hon C, Siegel T, Sabatini R, Luenen H, Farris C, Jan S, Genest P, Tripathi P, Velds A, Tiengwe C, Marcello L, Farr H, Dickens N, Kelly S, Swiderski M, Muller C, Hawkins M, Retkute R, Malla S, Wilson R, Blythe M, Godoy P, Nogueira-Junior L, Paes L, Cornejo A, Martins R, Silber A, Tiengwe C, Marcello L, Farr H, Gadelha C, Burchmore R, Barry J, Renard-Guillet C, Kanoh Y, Shirahige K, Masai H, Rhind N, Gilbert D, Echeverry M, Bot C, Obado S, Taylor M, Kelly J, Lukes J, Skalicky T, Tyc J, Votypka J, Yurchenko V, Sayed N, Myler P, Blandin G, Berriman M, Crabtree J, Aggarwal G, Mannaert A, Downing T, Imamura H, Dujardin J, Rogers M, Hilley J, Dickens N, Wilkes J, Bates P, Depledge D, et al 2015 Genome-wide mapping reveals single-origin chromosome replication in *Leishmania*, a eukaryotic microbe *Genome Biol.* **16** 230

[363] Mannaert A, Downing T, Imamura H and Dujardin J-C C 2012 Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. *Trends Parasitol.* **28** 370–6

[364] Ropolo A S, Saura A, Carranza P G and Lujan H D 2005 Identification of variant-specific surface proteins in *Giardia muris* trophozoites. *Infect. Immun.* **73** 5208–11

[365] Jaskowska E, Butler C, Preston G and Kelly S 2015 *Phytomonas*: Trypanosomatids Adapted to Plant Environments *PLoS Pathog.* **11** 1–17

[366] Williams R A M, Woods K L, Juliano L, Mottram J C and Coombs G H 2009 Characterization of unusual families of ATG8-like proteins and ATG12 in the protozoan parasite *Leishmania major*. *Autophagy* **5** 159–72

[367] Eme L, Trilles A, Moreira D, Brochier-Armanet C, Thornton B, Toczyski D, Fang S, Weissman A, Pickart C, Schreiber A, Stengel F, Zhang Z, Enchev R, Kong E, Morris E, Robinson C, Fonseca P da, Barford D, Reed S, Manchado E, Eguren M, Malumbres M, Peters J, Yoon H, Feoktistova A, Wolfe B, Jennings J, Link A, Gould K, Hutchins J, Toyoda Y, Hegemann B, Poser I, Heriche J, Sykora M, Augsburg M, Hudecz O, Buschhorn B, Bulkescher J, Acquaviva C, Pines J, Matyskiela M, Rodrigo-Brenni M, Morgan D, Nasmyth K, Haering C, Onn I, Heidinger-Pauli J, Guacci V, Unal E, Koshland D, Oelschlaegel T, Schwickart M, Matos J, Bogdanova A, Camasses A, Havlis J, Shevchenko A, Zachariae W, Pesin J, Orr-Weaver T, Pimentel A, Venkatesh T, Capron A, Okresz L, Genschik P, Eloy N, Coppens F, Beemster G, Hemerly A, Ferreira P, Mde F L, Eloy N, Pegoraro C, Sagit R, Rojas C, Bretz T, Vargas L, Elofsson A, Oliveira A de, Hemerly A, Fulop K, Tarayre S, Kelemen Z, Horvath G, Kevei Z, Nikovics K, Bako L, Brown S, Kondorosi A, Kondorosi E, Kumar P, Wang C, Eisen J, Fraser C, Stechmann A, Cavalier-Smith T, Seidl M, et al 2011 The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes *BMC Evol. Biol.* **11** 265

[368] Listovsky T, Brandeis M and Zilberstein D 2011 *Leishmania* express a functional *Cdc20* homologue *Biochem Biophys Res Commun.* **408** 71-7

[369] Bessat M, Knudsen G, Burlingame A L, Wang C C, Fenn K, Matthews K, Hammarton T, McKean P, Hammarton T, Clark J, Douglas F, Boshart M, Mottram J, Kumar P, Wang C, Tu X, Wang C, Hammarton T, Monnerat S, Mottram J, Fededa J, Gerlich D, Glotzer M, Li Z, Lee J, Chu F, Burlingame A, Gunzl A, Ersfeld K, Gull K, Barford D, Musacchio A, Salmon E, Peters J, Pines J, Leuken R van, Clijsters L, Wolthuis R, Cohen-Fix O, Peters J, Kirschner M, Koshland D, Funabiki H, Yamano H, Kumada K, Nagao K, Hunt T, Stemmann O, Zou H, Gerber S, Gygi S, Kirschner M, Uhlmann F, Lottspeich F, Nasmyth K, Amon A, Amon A,

Irniger S, Nasmyth K, Kumar P, Wang C, Bessat M, Ersfeld K, Zachariae W, Shevchenko A, Andrews P, Ciosk R, Galova M, Hall M, Torres M, Schroeder G, Borchers C, Zachariae W, Shin T, Galova M, Obermaier B, Nasmyth K, Li Z, Wang C, Thornton B, Ng T, Matyskiela M, Carroll C, Morgan D, Mumberg D, Muller R, Funk M, Gietz R, Schiestl R, Gietz R, Woods R, Wirtz E, Leal S, Ochatt C, Cross G, Li Z, Wang C, Schimanski B, et al 2013 A Minimal Anaphase Promoting Complex/Cyclosome (APC/C) in *Trypanosoma brucei* ed Z Li *PLoS One* **8** e59258

[370] Mottram J C, Coombs G H and Alexander J 2004 Cysteine peptidases as virulence factors of *Leishmania Curr. Opin. Microbiol.* **7** 375–81

[371] Brotherton M C, Bourassa S, Leprohon P, Légaré D, Poirier G G, Droit A and Ouellette M 2013 Proteomic and genomic analyses of antimony resistant *Leishmania infantum* mutant *PLoS One* **8** e81899

[372] d'Avila-Levy C M, Marinho F A, Santos L O, Martins J L, Santos A L S and Branquinha M H 2006 Antileishmanial activity of MDL 28170, a potent calpain inhibitor *Int. J. Antimicrob. Agents* **28** 138–42

[373] Ouellette M, Drummelsmith J, El Fadili A, Kündig C, Richard D and Roy G 2002 Pterin transport and metabolism in *Leishmania* and related trypanosomatid parasites *Int. J. Parasitol.* **32** 385–98

[374] Vickers T J and Beverley S M 2011 Folate metabolic pathways in *Leishmania*. *Essays Biochem.* **51** 63–80

[375] Kaur J, Kumar P, Tyagi S, Pathak R, Batra S, Singh P and Singh N 2011 In silico screening, structure-activity relationship, and biologic evaluation of selective pteridine reductase inhibitors targeting visceral leishmaniasis. *Antimicrob. Agents Chemother.* **55** 659–66

[376] Dobson D E, Scholtes L D, Myler P J, Turco S J and Beverley S M 2006 Genomic organization and expression of the expanded *SCG/L/R* gene family of *Leishmania major*: Internal clusters and telomeric localization of *SCGs* mediating species-specific LPG modifications *Mol. Biochem. Parasitol.* **146** 231–41

[377] Dobson D E, Scholtes L D, Valdez K E, Sullivan D R, Mengeling B J, Cilmi S, Turco S J and Beverley S M 2003 Functional identification of galactosyltransferases (*SCGs*) required for species-specific modifications of the lipophosphoglycan adhesin controlling *Leishmania* major-sand fly interactions. *J. Biol. Chem.* **278** 15523–31

[378] Ramírez C A, Requena J M, Puerta C J, McKean P, Vaughan S, Gull K, Gull K, Murray H, Berman J, Davies C, Saravia N, Alvar J, Velez I, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, Boer M den, Fong D, Chang K, Coulson R, Connor V, Chen J, Ajioka J, Jackson A, Vaughan S, Gull K, Requena J, Jackson A, Vaughan S, Gull K, Landfear S, McMahon-Pratt D, Wirth D, Rogers M, Hilley J, Dickens N, Wilkes J, Bates P, Depledge D, Harris D, Her Y, Herzyk P, Imamura H, Otto T, Sanders M, Seeger K, Dujardin J, Berriman M, Smith D, Hertz-Fowler C, Mottram J, Peacock C, Seeger K, Harris D, Murphy L, Ruiz J, Quail M, Peters N, Adlem E, Tivey A, Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream M, Carver T, Norbertczak H, Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R, Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F, Baldauf S, Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream M, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, et al 2013 Alpha tubulin genes from *Leishmania braziliensis*: genomic organization, gene structure and insights on their expression *BMC Genomics* **14** 454

[379] Bhargava S R 2014 Emergence of Tubulin as a Vaccine against Parasitic Infections *Intellect. Prop. Rights Open Access* **2** 124

[380] Williams R A, Tetley L, Mottram J C and Coombs G H 2006 Cysteine peptidases CPA and CPB are vital for autophagy and differentiation in *Leishmania mexicana*. *Mol. Microbiol.* **61** 655–74

[381] Williams R A M, Smith T K, Cull B, Mottram J C, Coombs G H, Besteiro S, Williams R, Morrison L, Coombs G, Mottram J, Williams R, Tetley L, Mottram J, Coombs G, Yang Z, Klionsky D, Levine B, Mizushima N, Virgin H, Weidberg H, Shvets E, Elazar Z, Codogno P, Mehrpour M, Proikas-Cezanne T, Tanida I, Mizushima N, Kiyooka M, Ohsumi M, Ueno T, Shintani T, Mizushima N, Ogawa Y, Matsuura A, Noda T, Nair U, Cao Y, Xie Z, Klionsky D, Axe E, Walker S, Manifava M, Chandra P, Roderick H, Tooze S, Yoshimori T, Hailey D, Rambold A, Satpute-Krishnan P, Mitra K, Sougrat R, Rubinsztein D, Shpilka T, Elazar Z, Hanada T, Noda N, Satomi Y, Ichimura Y, Fujioka Y, Schweers R, Zhang J, Randall M, Loyd M, Li W, Pankiv S, Clausen T, Lamark T, Brech A, Bruun J, Nakatogawa H, Ichimura Y, Ohsumi Y, Rigden D, Herman M, Gillies S, Michels P, Herman M, Perez-Morga D, Schtickzelle N, Michels P, Williams R, Woods K, Juliano L, Mottram J, Coombs G, Mammucari C, Rizzuto R, Wang K, Klionsky D, Okamoto K, Kondo-Okamoto N, Stephenson L, Miller B, Ng A, Eisenberg J, Zhao Z, Zhang Y, Qi H, et al 2012 ATG5 Is Essential for ATG8-Dependent Autophagy and Mitochondrial Homeostasis in *Leishmania major PLoS Pathog.* **8** e1002695

[382] Dillon L A L, Suresh R, Okrah K, Corrada Bravo H, Mosser D M, El-Sayed N M, Duque G A, Descoteaux A, Nardy A F, Freire-de-Lima C, Morrot A, Jones B, Faron M, Rasmussen J, Fletcher J, Pittman K, Knoll L, Rikihisa Y, Sia J, Georgieva M, Rengarajan J, Simon S, Hilbi H, Alvar J, Vélez I, Bern C, Herrero M, Desjeux P, Cano J, Bogdan C, Donhauser N, Döring R, Röllinghoff M, Diefenbach A, Rittig M, Laufs H, Müller K, Fleischer J, Reiling N, Jahnke N, Jensenius J, Moll H, Flohé S, Röllinghoff M, Peters N, Egen J, Secundino N, Debrabant A, Kimblin N, Kamhawi S, Sarkar A, Aga E, Bussmeyer U, Bhattacharyya A, Möller S, Hellberg L, Ambit A, Woods K, Cull B, Coombs G, Mottram J, Beverley S, Turco S, Wheeler R, Gluenz E, Gull K, Etges R, Müller I, Reiner S, Locksley R, Scharton-Kersten T, Scott P, Bennett C, Misslitz A, Colledge L, Aebischer T, Blackburn C, Laskay T, Zandbergen G, Solbach W, Locksley R, Heinzel F, Fankhauser J, Nelson C, Sadick M, Zhang S, Kim C, Batra S, McKerrow J, Loke P, Kaye P, Scott P, Olivier M, Gregory D, Forget G, Sacks D, Sher A, Akopyants N, et al 2015 Simultaneous transcriptional profiling of *Leishmania major* and its murine macrophage host cell reveals insights into host-pathogen interactions *BMC Genomics* **16** 1108

[383] Casgrain P-A, Martel C, McMaster W R, Mottram J C, Olivier M, Descoteaux A, Matheoud D, Moradin N, Bellemare-Pelletier A, Shio M, Hong W, Olivier M, Moradin N, Descoteaux A, Desjardins M, Descoteaux A, Vinet A, Fukuda M, Turco S, Descoteaux A, Olivier M, Atayde V, Isnard A, Hassani K, Shio M, Duque G A, Descoteaux A, Duque G A, Fukuda M, Turco S, Stäger S, Descoteaux A, Courret N, Frehel C, Gouhier N, Pouchelet M, Prina E, Roux P, Alexander J, Vickerman K, Barbieri C, Brown K, Rabinovitch M, Real F, Pouchelet M, Rabinovitch M, Real F, Mortara R, Stow J, Manderson A, Murray R, Ndjamen B, Kang B, Hatsuzawa K, Kima P, Canton J, Ndjamen B, Hatsuzawa K, Kima P, Canton J, Kima P, Ilg T, Mottram J, Coombs G, Alexander J, Mottram J, Souza A, Hutchison J, Carter R, Frame M, Coombs G, Cameron P, McGachy A, Anderson M, Paul A, Coombs G, Mottram J, Denise H, McNeil K, Brooks D, Alexander J, Coombs G, Mottram J, Abu-Dayyeh I, Hassani K, Westra E, Mottram J, Olivier M, Leao S D S, Lang T, Prina E, Hellio R, Antoine J, Bahr V, Stierhof Y, Ilg T, Demar M, Quinten M, et al 2016 Cysteine Peptidase B Regulates *Leishmania mexicana* Virulence through the Modulation of GP63 Expression *PLOS Pathog.* **12** e1005658

[384] Yao C, Donelson J E and Wilson M E 2003 The major surface protease (MSP or GP63) of

*Leishmania* sp. Biosynthesis, regulation of expression, and function *Mol. Biochem. Parasitol.* **132** 1–16

[385] Kobe B and Kajava A V 2001 The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **11** 725–32

[386] Kedzierski L, Montgomery J, Bullen D, Curtis J, Gardiner E, Jimenez-Ruiz A and Handman E 2004 A leucine-rich repeat motif of *Leishmania* parasite surface antigen 2 binds to macrophages through the complement receptor 3. *J. Immunol.* **172** 4902–6

[387] Kolli B K, Kostal J, Zaborina O, Chakrabarty A M and Chang K-P 2008 *Leishmania*-released nucleoside diphosphate kinase prevents ATP-mediated cytolysis of macrophages. *Mol. Biochem. Parasitol.* **158** 163–75

[388] Landfear S M, Ullman B, Carter N S and Sanchez M A 2004 Nucleoside and nucleobase transporters in parasitic protozoa. *Eukaryot. Cell* **3** 245–54

[389] Almeida R, Norrish A, Levick M, Vetrie D, Freeman T, Vilo J, Ivens A, Lange U, Stober C, McCann S and Blackwell J M 2002 From genomes to vaccines: *Leishmania* as a model. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357** 5–11

[390] Holzer T R, McMaster W R and Forney J D 2006 Expression profiling by whole-genome interspecies microarray hybridization reveals differential gene expression in procyclic promastigotes, lesion-derived amastigotes, and axenic amastigotes in *Leishmania mexicana Mol. Biochem. Parasitol.* **146** 198–218

[391] Debrabant A, Lee N, Pogue G P, Dwyer D M and Nakhasi H L 2002 Expression of calreticulin P-domain results in impairment of secretory pathway in *Leishmania donovani* and reduced parasite survival in macrophages *Int. J. Parasitol.* **32** 1423–34

[392] McNicoll F, Drummelsmith J, Müller M, Madore E, Boilard N, Ouellette M and Papadopoulou B 2006 A combined proteomic and transcriptomic approach to the study of stage differentiation in *Leishmania infantum*. *Proteomics* **6** 3567–81

[393] El-Sayed N M, Myler P J, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey E a, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu D C, Haas B J, Tran A-N, Wortman J R, Alsmark U C M, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton J M, Cerqueira G C, Creasy T, Delcher A L, Djikeng A, Embley T M, Hauser C, Ivens A C, Kummerfeld S K, Pereira-Leal J B, Nilsson D, Peterson J, Salzberg S L, Shallom J, Silva J C, Sundaram J, Westenberger S, White O, Melville S E, Donelson J E, Andersson B, Stuart K D and Hall N 2005 Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309** 404–9

[394] Nandan D, Yi T, Lopez M, Lai C and Reiner N E 2002 *Leishmania* EF-1alpha Activates the Src Homology 2 Domain Containing Tyrosine Phosphatase SHP-1 Leading to Macrophage Deactivation *J. Biol. Chem.* **277** 50190–7

[395] Leifso K, Cohen-Freue G, Dogra N, Murray A and McMaster W R 2007 Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: the *Leishmania* genome is constitutively expressed. *Mol. Biochem. Parasitol.* **152** 35–46

[396] Shaw C D, Lonchamp J, Downing T, Imamura H, Freeman T M, Cotton J A, Sanders M, Blackburn G, Dujardin J C, Rijal S, Khanal B, Illingworth C J R, Coombs G H and Carter K C 2016 In vitro selection of miltefosine resistance in promastigotes of *Leishmania donovani*

from Nepal: Genomic and metabolomic characterization *Mol. Microbiol.* **99** 1134–48

[397] Rocha-Granados M C, Klingbeil M M 2016 *Leishmania* DNA Replication Timing: A Stochastic Event? *Trends Parasitol.* **32** 755-757

[398] Stanojcic S, Sollelis L, Kuk N, Crobu L, Balard Y, Schwob E, Bastien P, Pagès M and Sterkers Y 2016 Single-molecule analysis of DNA replication reveals novel features in the divergent eukaryotes *Leishmania* and *Trypanosoma brucei* versus mammalian cells. *Sci. Rep.* **6** 23142

[399] Simpson L and Shaw J 1989 RNA editing and the mitochondrial cryptogenes of kinetoplastid protozoa *Cell* **57** 355–66

[400] Rovai L, Tripp C, Stuart K and Simpson L 1992 Recurrent polymorphisms in small chromosomes of *Leishmania tarentolae* after nutrient stress or subcloning *Mol. Biochem. Parasitol.* **50** 115–25

[401] Iovannisci D M and Beverley S M 1989 Structural alterations of chromosome 2 in *Leishmania major* as evidence for diploidy, including spontaneous amplification of the mini-exon array. *Mol. Biochem. Parasitol.* **34** 177–88

[402] Reis-Cunha J L, Rodrigues-Luiz G F, Valdivia H O, Baptista R P, Mendes T A O, de Morais G L, Guedes R, Macedo A M, Bern C, Gilman R H, Lopez C T, Andersson B, Vasconcelos A T, Bartholomeu D C, Hotez P, Bottazzi M, Franco-Paredes C, Ault S, Periago M, Coura J, Dias J, Martins-Melo F, Alencar C, Ramos A, Heukelbach J, Vargas N, Pedroso A, Zingales B, Lewis M, Llewellyn M, Gaunt M, Yeo M, Carrasco H, Miles M, Minning T, Weatherly D, Flibotte S, Tarleton R, Ackermann A, Panunzi L, Cosentino R, Sanchez D, Aguero F, Panunzi L, Aguero F, Tibayrenc M, Kjellberg F, Ayala F, Tibayrenc M, Ayala F, Westenberger S, Barnabe C, Campbell D, Sturm N, Machado C, Ayala F, Sturm N, Vargas N, Westenberger S, Zingales B, Campbell D, Iskow R, Gokcumen O, Lee C, Martins C, Baptista C, Ienne S, Cerqueira G, Bartholomeu D, Zingales B, Clayton C, Martinez-Calvillo S, Vizuet-de-Rueda J, Florencio-Martinez L, Manning-Cela R, Figueroa-Angulo E, Pedroso A, Cupolillo E, Zingales B, Triana O, Ortiz S, Dujardin J, Solari A, Weatherly D, Boehlke C, Tarleton R, Branche C, Ochaya S, Aslund L, Andersson B, Tibayrenc M, Ward P, Moya A, Ayala F, Pablos L, Osuna A, Bartholomeu D, Paiva R, et al 2015 Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains *BMC Genomics* **16** 499

[403] Guan W, Cao D-P, Sun K, Xu J-N, Zhang J-R, Chen D-L and Chen J-P 2012 Phylogenic analysis of Chinese *Leishmania* isolates based on small subunit ribosomal RNA (SSU rRNA) and 7 spliced leader RNA (7SL RNA). *Acta Parasitol.* **57** 101–13

[404] Zhang J R, Guo X G, Liu J L, Zhou T H, Gong X, Chen D L and Chen J P 2016 Molecular detection, identification and phylogenetic inference of *Leishmania* spp. in some desert lizards from Northwest China by using internal transcribed spacer 1 (ITS1) sequences *Acta Trop.* **162** 83–94

[405] Kassahun A, Sadlova J, Benda P, Kostalova T, Warburg A, Hailu A, Baneth G, Volf P and Votypka J 2015 Natural infection of bats with *Leishmania* in Ethiopia. *Acta Trop.* **150** 166–70

[406] Leta S, Dao T H T, Mesele F and Alemayehu G 2014 Visceral Leishmaniasis in Ethiopia: An Evolving Disease *PLoS Negl. Trop. Dis.* **8**

[407] Boité M C, Mauricio I L, Miles M A and Cupolillo E 2012 New insights on taxonomy,

phylogeny and population genetics of *Leishmania (Viannia)* parasites based on multilocus sequence analysis. *PLoS Negl. Trop. Dis.* **6** e1888

[408] Dantas-Torres F and Brandão-Filho S P 2006 Visceral leishmaniasis in Brazil: Revisiting paradigms of epidemiology and control *Rev. Inst. Med. Trop. Sao Paulo* **48** 151–6

[409] Miró G, Gálvez R, Fraile C, Descalzo M A, Molina R, Curdi J L, Sánchez C A, Hernández J C, Peña A E, Martín-Sánchez J, Guilvard E, Acedo-Sánchez C, Wolf-Echeverri M, Sanchís-Marín M, Morillas-Márquez F, Rioux J, Guilvard E, Gállego J, Moreno G, Pratlong F, Portús M, Rispail P, Gállego M, Bastien P, Alvar J, Cañavate C, Molina R, Moreno J, Nieto J, Ayllon T, Tesouro M, Amusátegui I, Villaescusa A, Rodríguez-Franco F, Sainz A, Maia C, Nunes M, Campino L, Maia C, Nunes M, Cristovao J, Campino L, Martín-Sánchez J, Acedo C, Muñoz-Pérez M, Pesson B, Marchal O, Morillas-Márquez F, Solano-Gallego L, Rodríguez-Cortés A, Iniesta L, Quintana J, Pastor J, Espada Y, Portus M, Alberola J, Tabar M, Altet L, Francino O, Sánchez A, Ferrer L, Roura X, Maroli M, Pennisi M, Muccio T Di, Khoury C, Gradoni L, Gramiccia M, Silva S da, Rabelo P, Nde F G, Ribeiro R, Melo M, Ribeiro V, Michalick M, Naucke T, Menn B, Massberg D, Lorentz S, Amusátegui I, Sainz A, Aguirre E, Tesouro M, Morillas F, Rabasco F S, Ocaña J, Martín-Sánchez J, Ocana-Wihelmi J, Acedo C, Sanchís-Marín M, Gálvez R, Descalzo M, Miró G, Jiménez M, Martín O, Santos-Brandao F Dos, Guerrero I, Cubero E, et al 2011 Infectivity to *Phlebotomus perniciosus* of dogs naturally parasitized with *Leishmania infantum* after different treatments *Parasit. Vectors* **4** 52

[410] Borja L S, Sousa O M F de, Solcà M da S, Bastos L A, Bordoni M, Magalhães J T, Larangeira D F, Barrouin-Melo S M, Fraga D B M and Veras P S T 2016 Parasite load in the blood and skin of dogs naturally infected by *Leishmania infantum* is correlated with their capacity to infect sand fly vectors *Vet. Parasitol.* **229** 110–7

[411] Corredor A, Kreutzer R D, Tesh R B, Boshell J, Palau M T, Caceres E, Duque S, Pelaez D, Rodriguez G, Nichols S, Hernandez C A, Morales A, Young D G and Ferro de Carrasquilla C 1990 Distribution and etiology of leishmaniasis in Colombia *Am. J. Trop. Med. Hyg.* **42** 206–14

[412] World Health Organization 2010 Control of the leishmaniases. *World Health Organ. Tech. Rep. Ser.*

[413] Travi B L, Tabares C J and Cadena H 2006 *Leishmania (Viannia) braziliensis* infection in two Colombian dogs: a note on infectivity for sand flies and response to treatment. *Biomedica* **26 Suppl 1** 249–53

[414] Ramírez J D, Hernández C, León C M, Ayala M S, Flórez C, González C, Alvar J, Marcili A, Schönian G, Kuhls K, Mauricio I, Schönian G, Mauricio I, Cupolillo E, Rioux A, Grimaldi G, McMahon-Pratt D, Lukes J, Reithinger R, Alvar J, Yactayo S, Bern C, Bern C, Maguire J, Alvar J, Corredor A, Ovalle C, Porras L, Rey M, Ríos M, Camargo Y, Weigle M, Ferro C, Valderrama C, Young D, Corredor A, Montalvo A, Fraga J, Hassan S, Auwera G Van der, Foulet F, Marco J, Olson D, Etter A, Chaves M E, Arango M, Saravia N, Auwera G Van der, Dujardin J, Hernández C, Myint C K, Asato Y, Yamamoto Y, Asato Y, Boité M, Mauricio I, Miles M, Cupolillo E, Marlow M, Boité M, Ferreira G, Steindel M, Cupolillo E, Zemanova E, Mauricio I, Gelanew T, Rougeron V, Meeûs T De, Bañuls A, Ramírez J D, Llewellyn M, Feliciangeli M, Young D, Duncan M, Savani E, Lopez Y, Grimaldi G, Tesh R, Corrêa J, Eresh S, Bruijn M de, Mendoza-Leon J, Barker D, Silveira F, Ferro C, Cardenas E, Corredor D, Morales A, Munsterman L, Kreutzer D, Dutari L, Loaiza J, Rodriguez-Bonfante C, Bonfante-Garrido R, Grimaldi G, Momen H, Cupolillo E, Travi B, et al 2016 Taxonomy, diversity, temporal and geographical distribution of Cutaneous Leishmaniasis in Colombia: A retrospective study *Sci. Rep.* **6** 28266

[415] Ferro C, López M, Fuya P, Lugo L, Cordovez J M and González C 2015 Spatial Distribution of Sand Fly Vectors and Eco-Epidemiology of Cutaneous Leishmaniasis Transmission in Colombia. *PLoS One* **10** e0139391

[416] Ronet C, Beverley S M and Fasel N 2011 Muco-cutaneous leishmaniasis in the New World: The ultimate subversion *Virulence* **2** 547–52

[417] Ovalle C E, Porras L, Rey M, Ríos M and Camargo Y C 2006 Geographic distribution of *Leishmania* species isolated from patients at the National Institute of Dermatology Federico Lleras Acosta E.S.E., 1995-2005 *Biomédica* **26** 145–51

[418] Saravia N G, Weigle K, Navas C, Segura I, Valderrama L, Valencia A Z, Escorcia B and McMahon-Pratt D 2002 Heterogeneity, geographic distribution, and pathogenicity of serodemes of *Leishmania Viannia* in Colombia *Am. J. Trop. Med. Hyg.* **66** 738–44

[419] Vélez I D, Carrillo L M, López L, Rodríguez E and Robledo S M 2012 An epidemic outbreak of canine cutaneous leishmaniasis in colombia caused by *Leishmania braziliensis* and *Leishmania panamensis Am. J. Trop. Med. Hyg.* **86** 807–11

[420] Lainson R and Shaw J J 1989 *Leishmania (Viannia) naiffi* sp. n., a parasite of the armadillo, *Dasypus novemcinctus (L.)* in Amazonian Brazil. *Ann. Parasitol. Hum. Comp.* **64** 3–9

[421] Naiff R D, Freitas R A, Naiff M F, Arias J R, Barrett T V., Momen H and Grimaldi Júnior G 1991 Epidemiological and nosological aspects of *Leishmania naiffi* Lainson & Shaw, 1989. *Mem. Inst. Oswaldo Cruz* **86** 317–21

[422] Thomaz-Soccol V, Lanotte G, Rioux J A, Pratlong F, Martini-Dumas A and Serres E 1993 Monophyletic origin of the genus *Leishmania* Ross, 1903. *Ann. Parasitol. Hum. Comp.* **68** 107–8

[423] Pratlong F, Deniau M, Darie H, Eichenlaub S, Pröll S, Garrabe E, le Guyadec T and Dedet J P 2002 Human cutaneous leishmaniasis caused by *Leishmania naiffi* is wide-spread in South America. *Ann. Trop. Med. Parasitol.* **96** 781–5

[424] van Thiel P-P A M, Gool T Van, Kager P A and Bart A 2010 First cases of cutaneous leishmaniasis caused by *Leishmania (Viannia) naiffi* infection in Surinam. *Am. J. Trop. Med. Hyg.* **82** 588–90

[425] Fagundes-Silva G A, Sierra Romero G A, Cupolillo E, Gadelha Yamashita E P, Gomes-Silva A, De Oliveira Guerra J A and Da-Cruz A M 2015 *Leishmania (Viannia) naiffi*: Rare enough to be neglected? *Mem. Inst. Oswaldo Cruz* **110** 797–800

[426] Arias J R, Miles M A, Naiff R D, Povoa M M, de Freitas R A, Biancardi C B and Castellon E G 1985 Flagellate infections of Brazilian sand flies (Diptera: Psychodidae): isolation in vitro and biochemical identification of *Endotrypanum* and *Leishmania*. *Am. J. Trop. Med. Hyg.* **34** 1098–108

[427] Kato H, Gomez E A, Yamamoto Y, Calvopiña M, Guevara A G, Marco J D, Barroso P A, Iwata H and Hashiguchi Y 2008 Natural infection of *Lutzomyia tortura* with *Leishmania (Viannia) naiffi* in an Amazonian area of Ecuador. *Am. J. Trop. Med. Hyg.* **79** 438–40

[428] Azpurua J, De La Cruz D, Valderama A and Windsor D 2010 *Lutzomyia* Sand Fly Diversity and Rates of Infection by *Wolbachia* and an Exotic *Leishmania* Species on Barro Colorado

Island, Panama *PLoS Negl. Trop. Dis.* **4** e627

[429] Cássia-Pires R, Boité M C, D'Andrea P S, Herrera H M, Cupolillo E, Jansen A M and Roque A L R 2014 Distinct *Leishmania* Species Infecting Wild Caviomorph Rodents (Rodentia: Hystricognathi) from Brazil *PLoS Negl. Trop. Dis.* **8** e3389

[430] Van Der Snoek E M, Lammers A M, Kortbeek L M, Roelfsema J H, Bart A and Jaspers C A J J 2009 Spontaneous cure of American cutaneous leishmaniasis due to *Leishmania naiffi* in two Dutch infantry soldiers *Clin. Exp. Dermatol.* **34** e889–91

[431] Floch H 1954 *Leishmania tropica guyanensis* n.sp. agent de la leishmaniose tegumentarie de Guyanes et de l'Amerique Centralele *Arch Inst Pasteur La Guyane Française du Teritoire L'Inni* **15**

[432] Lainson R, Shaw J J and Povoa M 1981 The importance of edentates (sloths and anteaters) as primary reservoirs of *Leishmania braziliensis guyanensis*, causative agent of "pian-bois"; in north Brazil. *Trans. R. Soc. Trop. Med. Hyg.* **75** 611–2

[433] Corredor A, Gallego J F, Tesh R B, Morales A, Ferro De Carrasquilla C, Young D G, Kreutzer R D, Boshell J, Palau M T, Caceres E and Pelaez D 1989 Epidemiology of visceral leishmaniasis in Colombia *Am. J. Trop. Med. Hyg.* **40** 480–6

[434] Quinnell R J and Courtenay O 2009 Transmission, reservoir hosts and control of zoonotic visceral leishmaniasis. *Parasitology* **136** 1915–34

[435] Lainson R, Shaw J J, Ward R D, Ready P D and Naiff R D 1979 Leishmaniasis in brazil: XIII. Isolation of *leishmania* from armadillos (*dasypus novemcinctus*), and observations on the epidemiology of cutaneous leishmaniasis in north para' state *Trans. R. Soc. Trop. Med. Hyg.* **73** 239–42

[436] Ready P D, Lainson R, Shaw J J, Ward R D, Arias J R, Naiff R D, Arias J R, Naiff R D, Miles M A, de Souza A A, Gentile B, Le Pont F, Pajot F X, Besnard R, Lainson R, Lainson R, Shaw J J, Póvoa M, Lainson R, Shaw J J, Ready P D, Miles M A, Póvoa M, Lainson R, Shaw J J, Ward R D, Ready P D, Naiff R D, Lainson R, Ward R D, Shaw J J, Le Pont R, Pajot F X, Reguer R, Miles M A, de Souza A A, Povoa M, Ready P D, Arias J R, Freitas R A, Ready P D, Fraiha H, Lainson R, Shaw J J, Ready P D, Fraiha H, Lane R P, Arias J R, Pajot F-X, Ready P D, Lainson R, Shaw J J, Ward R D, Ward R D, Fraiha H, Wijers D J B and Linger R 1986 The ecology of *lutzomyia umbratilis* Ward &amp; Fraiha (Diptera: Psychodidae), the major vector to man of *Leishmania braziliensis guyanensis* in north-eastern Amazonian brazil *Bull. Entomol. Res.* **76** 21

[437] Rodríguez-Barraquer I, Góngora R, Prager M, Pacheco R, Montero L M, Navas A, Ferro C, Miranda M C and Saravia N G 2008 Etiologic agent of an epidemic of cutaneous leishmaniasis in Tolima, Colombia *Am. J. Trop. Med. Hyg.* **78** 276–82

[438] Balbino V Q, Marcondes C B, Alexander B, Luna L K S, Lucena M M M, Mendes A C S and Andrade P P 2001 First Report of *Lutzomyia (Nyssomyia) umbratilis* Ward & Frahia, 1977 outside of Amazonian Region, in Recife, State of Pernambuco, Brazil (Diptera: Psychodidae: Phlebotominae) *Mem. Inst. Oswaldo Cruz* **96** 315–7

[439] Young D G and Duncan M A 1994 Guide to the identification and geographic distribution of *Lutzomyia sand* flies in Mexico, the West Indies, Central and South America (Diptera: Psychodidae) Walter Reed Army Institute of Research Washingtin DC **54**

[440] Fouque F, Gaborit P, Issaly J, Carinci R, Gantier J-C, Ravel C and Dedet J-P 2007 Phlebotomine sand flies (Diptera: Psychodidae) associated with changing patterns in the transmission of the human cutaneous leishmaniasis in French Guiana. *Mem. Inst. Oswaldo Cruz* **102** 35–40

[441] Lainson R, Shaw J J, Ready P D, Miles M A and Póvoa M 1981 Leishmaniasis in Brazil: XVI. Isolation and identification of *Leishmania* species from sandflies, wild mammals and man in north Para State, with particular reference to *L. braziliensis guyanensis* causative agent of "pian-bois". *Trans. R. Soc. Trop. Med. Hyg.* **75** 530–6

[442] Van Der Meide W F, Jensema A J, Akrum R A E, Sabajo L O A, Lai A Fat R F M, Lambregts L, Schallig H D F H, Van Der Paardt M and Faber W R 2008 Epidemiology of cutaneous leishmaniasis in Suriname: A study performed in 2006 *Am. J. Trop. Med. Hyg.* **79** 192–7

[443] Garcia A L, Tellez T, Parrado R, Rojas E, Bermudez H and Dujardin J C 2007 Epidemiological monitoring of American tegumentary leishmaniasis: molecular characterization of a peridomestic transmission cycle in the Amazonian lowlands of Bolivia *Trans. R. Soc. Trop. Med. Hyg.* **101** 1208–13

[444] Rotureau B, Ravel C, Nacher M, Couppié P, Curtet I, Dedet J P and Carme B 2006 Molecular epidemiology of *Leishmania (Viannia) guyanensis* in French Guiana *J. Clin. Microbiol.* **44** 468–73

[445] Delgado O, Cupolillo E, Bonfante-Garrido R, Silva S, Belfort E, Grimaldi Jr G and Momen H 1997 Cutaneous Leishmaniasis in Venezuela Caused by Infection with a New Hybrid between *Leishmania (Viannia) braziliensis* and *L. (V.) guyanensis Mem. Inst. Oswaldo Cruz* **92** 581–2

[446] Bonfante-Garrido R, Meléndez E, Barroeta S, de Alejos M A, Momen H, Cupolillo E, McMahon-Pratt D and Grimaldi G 1992 Cutaneous leishmaniasis in western Venezuela caused by infection with *Leishmania venezuelensis* and *L. braziliensis* variants. *Trans. R. Soc. Trop. Med. Hyg.* **86** 141–8

[447] Jennings Y L, de Souza A A A, Ishikawa E A, Shaw J, Lainson R and Silveira F 2014 Phenotypic characterization of *Leishmania* spp. causing cutaneous leishmaniasis in the lower Amazon region, western Pará state, Brazil, reveals a putative hybrid parasite, *Leishmania (Viannia) guyanensis* × *Leishmania (Viannia) shawi shawi*. *Parasite* **21** 39

[448] Walsh J F, Molyneux D H and Birley M H 1993 Deforestation: effects on vector-borne disease. *Parasitology* **106 Suppl** S55–75

[449] Davies C R, Campbell-lendrum D, Reithinger R, Campbell-lendrum D, Feliciangeli D, Borges R and Rodriguez N 2000 The epidemiology and control of leishmaniasis in Andean countries Epidemiologia e controle da leishmaniose nos países andinos *Cad. Saude Pública, Rio Janeiro* **16** 925–50

[450] Valderrama-Ardila C, Alexander N, Ferro C, Cadena H, Marín D, Holford T R, Munstermann L E and Ocampo C B 2010 Environmental risk factors for the incidence of American cutaneous leishmaniasis in a sub-andean zone of Colombia (Chaparral, Tolima) *Am. J. Trop. Med. Hyg.* **82** 243–50

[451] Wincker P, Ravel C, Blaineau C, Pages M, Jauffret Y, Dedet J P and Bastien P 1996 The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human

pathogenic species. *Nucleic Acids Res.* **24** 1688–94

[452] Oddone R, Schweynoch C, Schönian G, De Sousa C D S, Cupolillo E, Espinosa D, Arevalo J, Noyes H, Mauricio I and Kuhls K 2009 Development of a multilocus microsatellite typing approach for discriminating strains of *Leishmania (Viannia)* species *J. Clin. Microbiol.* **47** 2818–25

[453] Lye L-F F, Owens K, Shi H, Murta S M F, Vieira A C, Turco S J, Tschudi C, Ullu E and Beverley S M 2010 Retention and Loss of RNA interference pathways in trypanosomatid protozoans *PLoS Pathog.* **6** e1001161

[454] Carver T, Harris S R, Berriman M, Parkhill J and McQuillan J A 2012 Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28** 464–9

[455] Smith D F, Peacock C S and Cruz A K 2007 Comparative genomics: from genotype to disease phenotype in the leishmaniases. *Int. J. Parasitol.* **37** 1173–86

[456] Barnes R L, Shi H, Kolev N G, Tschudi C and Ullu E 2012 Comparative genomics reveals two novel RNAi factors in *Trypanosoma brucei* and provides insight into the core machinery. *PLoS Pathog.* **8** e1002678

[457] Logan-Klumpler F J, De Silva N, Boehme U, Rogers M B, Velarde G, McQuillan J A, Carver T, Aslett M, Olsen C, Subramanian S, Phan I, Farris C, Mitra S, Ramasamy G, Wang H, Tivey A, Jackson A, Houston R, Parkhill J, Holden M, Harb O S, Brunk B P, Myler P J, Roos D, Carrington M, Smith D F, Hertz-Fowler C and Berriman M 2012 GeneDB-an annotation database for pathogens *Nucleic Acids Res.* **40** D98-108

[458] Smith M, Bringaud F and Papadopoulou B 2009 Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome. *BMC Genomics* **10** 240

[459] Bañuls A L, Jonquieres R, Guerrini F, Le Pont F, Barrera C, Espinel I, Guderian R, Echeverria R and Tibayrenc M 1999 Genetic analysis of *leishmania* parasites in Ecuador: are *Leishmania (Viannia) panamensis* and *Leishmania (V.) Guyanensis* distinct taxa? *Am. J. Trop. Med. Hyg.* **61** 838–45

[460] Segovia M and Ortiz G 1997 LDI amplifications in *Leishmania Parasitol. Today* **13** 342–8

[461] Fu G, Melville S, Brewster S, Warner J and Barker D C 1998 Analysis of the genomic organisation of a small chromosome of *Leishmania braziliensis* M2903 reveals two genes encoding GTP-binding proteins, one of which belongs to a new G-protein family and is an antigen *Gene* **210** 325–33

[462] Panagabko C, Morley S, Hernandez M, Cassolato P, Gordon H, Parsons R, Manor D and Atkinson J 2003 Ligand specificity in the CRAL-TRIO protein family *Biochemistry* **42** 6467–74

[463] Banerjee S, Basu S and Sarkar S 2010 Comparative genomics reveals selective distribution and domain organization of FYVE and PX domain proteins across eukaryotic lineages. *BMC Genomics* **11** 83

[464] Atayde V D, Shi H, Franklin J B, Carriero N, Notton T, Lye L F, Owens K, Beverley S M, Tschudi C and Ullu E 2013 The structure and repertoire of small interfering RNAs in

*Leishmania (Viannia) braziliensis* reveal diversification in the trypanosomatid RNAi pathway *Mol. Microbiol.* **87** 580–93

[465] Valdivia H O, Scholte L L S, Oliveira G, Gabaldón T and Bartholomeu D C 2015 The *Leishmania* metaphylome: a comprehensive survey of *Leishmania* protein phylogenetic relationships. *BMC Genomics* **16** 887

[466] Eisen J A, Coyne R S, Wu M, Wu D, Thiagarajan M, Wortman J R, Badger J H, Ren Q, Amedeo P, Jones K M, Tallon L J, Delcher A L, Salzberg S L, Silva J C, Haas B J, Majoros W H, Farzad M, Carlton J M, Smith R K, Garg J, Pearlman R E, Karrer K M, Sun L, Manning G, Elde N C, Turkewitz A P, Asai D J, Wilkes D E, Wang Y, Cai H, Collins K, Stewart B A, Lee S R, Wilamowska K, Weinberg Z, Ruzzo W L, Wloga D, Gaertig J, Frankel J, Tsao C C, Gorovsky M A, Keeling P J, Waller R F, Patron N J, Cherry J M, Stover N A, Krieger C J, Del Toro C, Ryder H F, Williamson S C, Barbeau R A, Hamilton E P and Orias E 2006 Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote *PLoS Biol.* **4** 1620–42

[467] Patrick K L, Shi H, Kolev N G, Ersfeld K, Tschudi C and Ullu E 2009 Distinct and overlapping roles for two Dicer-like proteins in the RNA interference pathways of the ancient eukaryote *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. U. S. A.* **106** 17933–8

[468] Shi H, Tschudi C and Ullu E 2006 An unusual Dicer-like1 protein fuels the RNA interference pathway in *Trypanosoma brucei*. *RNA* **12** 2063–72

[469] Steinkraus H B, Greer J M, Stephenson D C and Langer P J 1993 Sequence heterogeneity and polymorphic gene arrangements of the *Leishmania guyanensis gp63* genes *Mol. Biochem. Parasitol.* **62** 173–85

[470] Joshi P B, Sacks D L, Modi G and McMaster W R 1998 Targeted gene deletion of *Leishmania major* genes encoding developmental stage-specific leishmanolysin (GP63) *Mol. Microbiol.* **27** 519–30

[471] Joshi P B, Kelly B L, Kamhawi S, Sacks D L and McMaster W R 2002 Targeted gene deletion in *Leishmania major* identifies leishmanolysin (GP63) as a virulence factor. *Mol. Biochem. Parasitol.* **120** 33–40

[472] Olivier M, Atayde V D, Isnard A, Hassani K and Shio M T 2012 *Leishmania* virulence factors: focus on the metalloprotease GP63 *Microbes Infect.* **14** 1377–89

[473] Brittingham A, Morrison C J, McMaster W R, McGwire B S, Chang K P and Mosser D M 1995 Role of the *Leishmania* surface protease gp63 in complement fixation, cell adhesion, and resistance to complement-mediated lysis. *J. Immunol.* **155** 3102–11

[474] Hartley M A, Drexler S, Ronet C, Beverley S M and Fasel N 2014 The immunological, environmental, and phylogenetic perpetrators of metastatic leishmaniasis *Trends Parasitol.* **30** 412–22

[475] Acestor N, Masina S, Ives A, Walker J, Saravia N G and Fasel N 2006 Resistance to oxidative stress is associated with metastasis in mucocutaneous leishmaniasis. *J. Infect. Dis.* **194** 1160–7

[476] Jackson A P 2010 The evolution of amastin surface glycoproteins in trypanosomatid parasites. *Mol. Biol. Evol.* **27** 33–45

[477] Cupolillo E, Grimaldi Júnior G, Momen H and Beverley S M 1995 Intergenic region typing (IRT): a rapid molecular approach to the characterization and evolution of *Leishmania*. *Mol. Biochem. Parasitol.* **73** 145–55

[478] Berzunza-Cruz M, Cabrera N, Crippa-Rossi M, Sosa Cabrera T, Pérez-Montfort R and Becker I 2002 Polymorphism analysis of the internal transcribed spacer and small subunit of ribosomal RNA genes of *Leishmania mexicana*. *Parasitol. Res.* **88** 918–25

[479] Lainson R, Shaw J J, Silveira F T, Braga R R and Ishikawa E a 1990 Cutaneous leishmaniasis of man due to *Leishmania (Viannia) naiffi* Lainson and Shaw, 1989. *Ann. Parasitol. Hum. comparée* **65** 282–4

[480] Barreto M, Burbano M E and Barreto P 2000 *Lutzomyia* Sand Flies (Diptera: Psychodidae) from Middle and Lower Putumayo Department, Colombia, with New Records to the Country *Mem. Inst. Oswaldo Cruz* **95** 633–9

[481] Abba, A. M.; Superina M 2010 The 2009/2010 armadillo red list assessment *BioOne* **11** 135–84

[482] Ober H K, Degroote L W, McDonough C M, Mizell R F and Mankin R W 2011 Identification of an attractant for the nine-banded armadillo, *Dasypus novemcinctus Wildl. Soc. Bull.* **35** 421–9

[483] Santaella J, Ocampo C B, Saravia N G, Méndez F, Góngora R, Gomez M A, Munstermann L E and Quinnell R J 2011 *Leishmania (Viannia)* infection in the domestic dog in Chaparral, Colombia *Am. J. Trop. Med. Hyg.* **84** 674–80

[484] Vásquez-Trujillo A, González A E, Góngora A, Cabrera O, Santamaría E and Buitrago L S 2008 Identificación de *Leishmania (Viannia) guyanensis* en caninos, en zona rural del municipio de Villavicencio, Meta, Colombia *Orinoquia* **12** 173–81

[485] Ferro C, Marín D, Góngora R, Carrasquilla M C, Trujillo J E, Rueda N K, Marín J, Valderrama-Ardila C, Alexander N, Pérez M, Munstermann L E and Ocampo C B 2011 Phlebotomine vector ecology in the domestic transmission of American cutaneous leishmaniasis in Chaparral, Colombia. *Am. J. Trop. Med. Hyg.* **85** 847–56

[486] Pardo R H, Cabrera O L, Becerra J, Fuya P and Ferro C 2006 *Lutzomyia longiflocosa* as suspected vector of cutaneous leishmaniasis in a focus of cutaneous leishmaniasis on the sub-andean region of Tolima department, Colombia, and the knowledge on sandflies by the inhabitants *Biomédica* **26** 95–108

[487] Lakshmi B S, Wang R and Madhubala R 2014 *Leishmania* genome analysis and high-throughput immunological screening identifies tuzin as a novel vaccine candidate against visceral leishmaniasis *Vaccine* **32** 3816–22

[488] Iantorno S 2015 Genome Plasticity and Genetic Exchange in *Leishmania Tropica* , Ph.D. thesis, University of Cambridge

[489] Sampaio M C R, Barbosa A F, Este M G, Pirmez C, Bello A R and Traub-Csekö Y M 2009 A 245 kb mini-chromosome impacts on *Leishmania braziliensis* infection and survival *Biochem. Biophys. Res. Commun.* **382** 74–8

[490] Diep B A, Gill S R, Chang R F, Phan T H, Chen J H, Davidson M G, Lin F, Lin J, Carleton

H A, Mongodin E F, Sensabaugh G F and Perdreau-Remington F 2006 Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus Lancet* **367** 731–9

[491] Cingolani P, Platts A, Wang L L, Coon M, Nguyen T, Wang L, Land S J, Lu X and Ruden D M 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w 1118; iso-2; iso-3 *Fly (Austin).* **6** 80–92

[492] Tenover F C, McDougal L K, Goering R V, Killgore G, Projan S J, Patel J B and Dunman P M 2006 Characterization of a Strain of Community- Associated Methicillin-Resistant *Staphylococcus aureus* Widely Disseminated in the United States *J. Clin. Microbiol.* **44** 108–18

[493] Diep B A, Carleton H a, Chang R F, Sensabaugh G F and Perdreau-Remington F 2006 Roles of 34 virulence genes in the evolution of hospital- and community-associated strains of methicillin-resistant *Staphylococcus aureus*. *J. Infect. Dis.* **193** 1495–503

[494] Gilbert M, MacDonald J, Gregson D, Siushansian J, Zhang K, Elsayed S, Laupland K, Louie T, Hope K, Mulvey M, Gillespie J, Nielsen D, Wheeler V, Louie M, Honish A, Keays G and Conly J 2006 Outbreak in Alberta of community-acquired (USA300) methicillin-resistant *Staphylococcus aureus* in people with a history of drug use, homelessness or incarceration *CMAJ* **175** 149–54

[495] Witte W, Strommenger B, Cuny C, Heuck D and Nuebel U 2007 Methicillin-resistant *Staphylococcus aureus* containing the Panton-Valentine leucocidin gene in Germany in 2005 and 2006 *J. Antimicrob. Chemother.* **60** 1258–63

[496] Tineli M, Pontosti A, Lusardi C, Vimercati M, Monaco M 2007 First detected case of community-acquired methicillin-resistant *Staphylococcus aureus* skin and soft tissue infection in Italy *Euro Surveill.* **12** e070412.1.

[497] Centers for Disease Control and Prevention 2003 Methicillin-resistant *staphylococcus aureus* infections among competitive sports participants--Colorado, Indiana, Pennsylvania, and Los Angeles County, 2000-2003 *MMWR Morb.Mortal.Wkly.Rep.* **52** 793–5

[498] Miller L G, Eells S J, Taylor A R, David M Z, Ortiz N, Zychowski D, Kumar N, Cruz D, Boyle-Vavra S and Daum R S 2012 *Staphylococcus aureus* colonization among household contacts of patients with skin infections: Risk factors, strain discordance, and complex ecology *Clin. Infect. Dis.* **54** 1523–35

[499] Miller L G and Diep B A 2008 Colonization, fomites, and virulence: rethinking the pathogenesis of community-associated methicillin-resistant *Staphylococcus aureus* infection *Clin.Infect.Dis.* **46** 752–60

[500] Miller L G, Perdreau-Remington F, Rieg G, Mehdi S, Perlroth J, Bayer A S, Tang A W, Phung T O and Spellberg B 2005 Necrotizing fasciitis caused by community-associated methicillin-resistant *Staphylococcus aureus* in Los Angeles. *N. Engl. J. Med.* **352** 1445–53

[501] Gonzalez B E, Martinez-Aguilar G, Hulten K G, Hammerman W a, Coss-Bu J, Avalos-Mishaan A, Mason E O and Kaplan S L 2005 Severe Staphylococcal sepsis in adolescents in the era of community-acquired methicillin-resistant *Staphylococcus aureus*. *Pediatrics* **115** 642–8

[502] Kreisel K M, Stine O C, Johnson J K, Perencevich E N, Shardell M D, Lesse A J, Gordin F M, Climo M W and Roghmann M C 2011 USA300 methicillin-resistant *Staphylococcus aureus* bacteremia and the risk of severe sepsis: Is USA300 methicillin-resistant *Staphylococcus aureus* associated with more severe infections? *Diagn. Microbiol. Infect. Dis.* **70** 285–90

[503] Tenover F C, Tickler I A, Goering R V., Kreiswirth B N, Mediavilla J R and Persinga D H 2012 Characterization of nasal and blood culture isolates of methicillin-resistant *Staphylococcus aureus* from patients in United States hospitals *Antimicrob. Agents Chemother.* **56** 1324–30

[504] Diep B A, Stone G G, Basuino L, Graber C J, Miller A, des Etages S-A, Jones A, Palazzolo-Ballance A M, Perdreau-Remington F, Sensabaugh G F, DeLeo F R and Chambers H F 2008 The arginine catabolic mobile element and staphylococcal chromosomal cassette *mec* linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant *Staphylococcus aureus*. *J. Infect. Dis.* **197** 1523–30

[505] Mwangi M M, Kim C, Chung M, Tsai J, Vijayadamodar G, Benitez M, Jarvie T P, Du L and Tomasz A 2013 Whole-genome sequencing reveals a link between β-lactam resistance and synthetases of the alarmone (p)ppGpp in *Staphylococcus aureus*. *Microb. Drug Resist.* **19** 153–9

[506] Kim C, Mwangi M, Chung M, Milheirco C, De Lencastre H and Tomasz A 2013 The mechanism of heterogeneous beta-lactam resistance in MRSA: Key role of the stringent stress response *PLoS One* **8**

[507] Dordel J, Kim C, Chung M, Pardos de la Gándara M, Holden M T J, Parkhill J, de Lencastre H, Bentley S D and Tomasz A 2014 Novel determinants of antibiotic resistance: identification of mutated loci in highly methicillin-resistant subpopulations of methicillin-resistant *Staphylococcus aureus*. *MBio* **5** e01000

[508] Berger-Bächi B, Strässle A, Gustafson J E and Kayser F H 1992 Mapping and characterization of multiple chromosomal factors involved in methicillin resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **36** 1367–73

[509] de Lencastre H and Tomasz A 1994 Reassessment of the number of auxiliary genes essential for expression of high-level methicillin resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **38** 2590–8

[510] Cuirolo A, Plata K and Rosato A E 2009 Development of homogeneous expression of resistance in methicillin-resistant *Staphylococcus aureus* clinical strains is functionally associated with a beta-lactam-mediated SOS response. *J. Antimicrob. Chemother.* **64** 37–45

[511] Plata K B, Rosato R R and Rosato A E 2011 Fate of mutation rate depends on *agr* locus expression during oxacillin-mediated heterogeneous-homogeneous selection in methicillin-resistant *Staphylococcus aureus* clinical strains. *Antimicrob. Agents Chemother.* **55** 3176–86

[512] Rudkin J K, Edwards A M, Bowden M G, Brown E L, Pozzi C, Waters E M, Chan W C, Williams P, O'Gara J P and Massey R C 2012 Methicillin resistance reduces the virulence of healthcare-associated methicillin-resistant *staphylococcus aureus* by interfering with the *agr* quorum sensing system *J. Infect. Dis.* **205** 798–806

[513] Pozzi C, Waters E M, Rudkin J K, Schaeffer C R, Lohan A J, Tong P, Loftus B J, Pier G B, Fey P D, Massey R C and O'Gara J P 2012 Methicillin resistance alters the biofilm phenotype

and attenuates virulence in *Staphylococcus aureus* device-associated infections. *PLoS Pathog.* **8** e1002626

[514] Heyer G, Saba S, Adamo R, Rush W, Soong G, Cheung A and Prince A 2002 *Staphylococcus aureus agr* and *sarA* functions are required for invasive infection but not inflammatory responses in the lung *Infect. Immun.* **70** 127–33

[515] Morfeldt E, Janzou L, Arvidson S and Löfdahl S 1988 Cloning of a chromosomal locus (*exp*) which regulates the expression of several exoprotein genes in *Staphylococcus aureus MGG Mol. Gen. Genet.* **211** 435–40

[516] Watkins R R, David M Z and Salata R A 2012 Current concepts on the virulence mechanisms of meticillin-resistant *Staphylococcus aureus J. Med. Microbiol.* **61** 1179–93

[517] Collins J, Rudkin J, Recker M, Pozzi C, O'Gara J P and Massey R C 2010 Offsetting virulence and antibiotic resistance costs by MRSA. *ISME J.* **4** 577–84

[518] Sims D, Sudbery I, Ilott N E, Heger A and Ponting C P 2014 Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15** 121–32

[519] Ghoneim D H, Myers J R, Tuttle E and Paciorkowski A R 2014 Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res. Notes* **7** 864

[520] Jünemann S, Prior K, Albersmeier A, Albaum S, Kalinowski J, Goesmann A, Stoye J and Harmsen D 2014 GABenchToB: A genome assembly benchmark tuned on bacteria and benchtop sequencers *PLoS One* **9**

[521] Leek J T and Storey J D 2007 Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** 1724–35

[522] McCarthy D J and Smyth G K 2009 Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25** 765–71

[523] Cokelaer T, Pultz D, Harder L M, Serra-Musach J and Saez-Rodriguez J 2013 BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* **29** 3241–2

[524] Tatusov R L, Galperin M Y, Natale D A and Koonin E V 2000 The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28** 33–6

[525] Luo W and Brouwer C 2013 Pathview: An R/Bioconductor package for pathway-based data integration and visualization *Bioinformatics* **29** 1830–1

[526] Luo W, Friedman M S, Shedden K, Hankenson K D and Woolf P J 2009 GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10** 161

[527] Culhane A C, Perrière G and Higgins D G 2003 Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* **4** 59

[528] Culhane A C, Thioulouse J, Perrière G and Higgins D G 2005 MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics* **21** 2789–90

[529] Kroonenberg P M and Lombardo R 1999 Nonsymmetric Correspondence Analysis: A Tool for Analysing Contingency TablesWith a Dependence Structure *Multivariate Behav. Res.* **34** 367–96

[530] Shabalin A A 2012 Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28** 1353–8

[531] McCallum N, Stutzmann Meier P, Heusser R and Berger-Bächi B 2011 Mutational analyses of open reading frames within the *vraSR* operon and their roles in the cell wall stress response of *Staphylococcus aureus Antimicrob. Agents Chemother.* **55** 1391–402

[532] Utaida S, Dunman P M, Macapagal D, Murphy E, Projan S J, Singh V K, Jayaswal R K and Wilkinson B J 2003 Genome-wide transcriptional profiling of the response of *Staphylococcus aureus* to cell-wall-active antibiotics reveals a cell-wall-stress stimulon *Microbiology* **149** 2719–32

[533] Kuroda M, Kuroda H, Oshima T, Takeuchi F, Mori H and Hiramatsu K 2003 Two-component system *VraSR* positively modulates the regulation of cell-wall biosynthesis pathway in *Staphylococcus aureus Mol. Microbiol.* **49** 807–21

[534] Dunman P M, Murphy E, Haney S, Palacios D, Tucker-Kellogg G, Wu S, Brown E L, Zagursky R J, Shlaes D and Projan S J 2001 Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the *agr* and/or *sarA* loci *Journal of Bacteriology* **183** 7341–53

[535] Saïd-Salim B, Dunman P M, McAleese F M, Macapagal D, Murphy E, McNamara P J, Arvidson S, Foster T J, Projan S J and Kreiswirth B N 2003 Global regulation of *Staphylococcus aureus* genes by Rot. *J. Bacteriol.* **185** 610–9

[536] Geisinger E, Adhikari R P, Jin R, Ross H F and Novick R P 2006 Inhibition of *rot* translation by RNAIII, a key feature of *agr* function *Mol. Microbiol.* **61** 1038–48

[537] Boisset S, Geissmann T, Huntzinger E, Fechter P, Bendridi N, Possedko M, Chevalier C, Helfer A C, Benito Y, Jacquier A, Gaspin C, Vandenesch F and Romby P 2007 *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism *Genes Dev.* **21** 1353–66

[538] Peschel A, Otto M, Jack R W, Kalbacher H, Jung G and Götz F 1999 Inactivation of the *dlt* operon in *Staphylococcus aureus* confers sensitivity to defensins, protegrins, and other antimicrobial peptides. *J. Biol. Chem.* **274** 8405–10

[539] Brown S, Xia G, Luhachack L G, Campbell J, Meredith T C, Chen C, Winstel V, Gekeler C, Irazoqui J E, Peschel A and Walker S 2012 Methicillin resistance in *Staphylococcus aureus* requires glycosylated wall teichoic acids. *Proc. Natl. Acad. Sci. U. S. A.* **109** 18909–14

[540] Schirner K, Stone L K and Walker S 2011 ABC transporters required for export of wall teichoic acids do not discriminate between different main chain polymers *ACS Chemical Biology* **6** 407–12

[541] Brown S, Santa Maria J P and Walker S 2013 Wall Teichoic Acids of Gram-Positive Bacteria *Annu. Rev. Microbiol.* **67** 313–36

[542] Rooijakkers S H, Ruyken M, Roos a, Daha M R, Presanis J S, Sim R B, van Wamel W J,

van Kessel K P and van Strijp J a 2005 Immune evasion by a staphylococcal complement inhibitor that acts on C3 convertases *Nat.Immunol.* **6** 920–7

[543] Raeside C, Gaffé J, Deatherage D E, Tenaillon O, Briska A M, Ptashkin R N, Cruveiller S, Médigue C, Lenski R E, Barrick J E and Schneider D 2014 Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli MBio* **5** e01377-14

[544] Matthews P R and Stewart P R 1988 Amplification of a section of chromosomal DNA in methicillin-resistant *Staphylococcus aureus* following growth in high concentrations of methicillin. *J. Gen. Microbiol.* **134** 1455–64

[545] Wilson C M, Volkman S K, Thaithong S, Martin R K, Kyle D E, Milhous W K and Wirth D F 1993 Amplification of *pfmdr1* associated with mefloquine and halofantrine resistance in *Plasmodium falciparum* from Thailand *Mol. Biochem. Parasitol.* **57** 151–60

[546] Musher D M, Dowell M E, Shortridge V D, Flamm R K, Jorgensen J H, Le Magueres P and Krause K L 2002 Emergence of macrolide resistance during treatment of pneumococcal pneumonia. *N. Engl. J. Med.* **346** 630–1

[547] Sandegren L and Andersson D I 2009 Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nat. Rev. Microbiol.* **7** 578–88

[548] Albertson D G 2006 Gene amplification in cancer *Trends Genet.* **22** 447–55

[549] Normark E and Staffan T 1981 Recombination between short DNA homologies causes tandem duplication *Nature* **292** 269–71

[550] Edlund T, Grundstrom T and Normark S 1979 Isolation and characterization of DNA repetitions carrying the chromosomal beta-lactamase gene of *Escherichia coli* K-12 *Mol Gen Genet* **173** 115–25

[551] Higgins P G, Rosato A E, Seifert H, Archer G L and Wisplinghoff H 2009 Differential expression of *ccrA* in methicillin-resistant *Staphylococcus aureus* strains carrying staphylococcal cassette chromosome *mec* type II and IVa elements. *Antimicrob. Agents Chemother.* **53** 4556–8

[552] Stojanov M, Sakwinska O and Moreillon P 2013 Expression of scc*mec* cassette chromosome recombinases in methicillin-resistant *Staphylococcus aureus* and *Staphylococcus epidermidis J. Antimicrob. Chemother.* **68** 749–57

[553] Donnio P Y, Oliveira D C, Faria N A, Wilhelm N, Le Coustumier A and De Lencastre H 2005 Partial excision of the chromosomal cassette containing the methicillin resistance determinant results in methicillin-susceptible *Staphylococcus aureus J. Clin. Microbiol.* **43** 4191–3

[554] Adler M, Anjum M, Berg O G, Andersson D I and Sandegren L 2014 High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms *Mol. Biol. Evol.* **31** 1526–35

[555] Lovett S T, Drapkin P T, Sutera V A and Gluckman-Peskind T J 1993 A sister-strand exchange mechanism for *recA*-independent deletion of repeated DNA sequences in *Escherichia coli Genetics* **135** 631–42

[556] Galitski T and Roth J R 1997 Pathways for homologous recombination between chromosomal direct repeats in *Salmonella typhimurium Genetics* **146** 751–67

[557] Rao F, See R Y, Zhang D, Toh D C, Ji Q and Liang Z-X 2010 YybT is a signaling protein that contains a cyclic dinucleotide phosphodiesterase domain and a GGDEF domain with ATPase activity. *J. Biol. Chem.* **285** 473–82

[558] Wassmann P, Chan C, Paul R, Beck A, Heerklotz H, Jenal U and Schirmer T 2007 Structure of BeF3- -modified response regulator PleD: implications for diguanylate cyclase activation, catalysis, and feedback inhibition. *Structure* **15** 915–27

[559] Griffiths J M and O'Neill A J 2012 Loss of function of the GdpP protein leads to joint β-lactam/glycopeptide tolerance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **56** 579–81

[560] Durfee T, Hansen A-M, Zhi H, Blattner F R and Jin D J 2008 Transcription profiling of the stringent response in *Escherichia coli*. *J. Bacteriol.* **190** 1084–96

[561] Pereira S F F, Henriques A O, Pinho M G, de Lencastre H and Tomasz A 2007 Role of PBP1 in cell division of *Staphylococcus aureus*. *J. Bacteriol.* **189** 3525–31

[562] Wada A and Watanabe H 1998 Penicillin-binding protein 1 of *Staphylococcus aureus* is essential for growth. *J. Bacteriol.* **180** 2759–65

[563] Aubry-Damon H, Soussy C J and Courvalin P 1998 Characterization of mutations in the *rpoB* gene that confer rifampin resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **42** 2590–4

[564] Watanabe Y, Cui L, Katayama Y, Kozue K and Hiramatsu K 2011 Impact of *rpoB* mutations on reduced vancomycin susceptibility in *Staphylococcus aureus*. *J. Clin. Microbiol.* **49** 2680–4

[565] O'Neill A J, Huovinen T, Fishwick C W G and Chopra I 2006 Molecular genetic and structural modeling studies of *Staphylococcus aureus* RNA polymerase and the fitness of rifampin resistance genotypes in relation to clinical prevalence. *Antimicrob. Agents Chemother.* **50** 298–309

[566] Cui L, Isii T, Fukuda M, Ochiai T, Neoh H-M, Camargo I L B da C, Watanabe Y, Shoji M, Hishinuma T and Hiramatsu K 2010 An RpoB mutation confers dual heteroresistance to daptomycin and vancomycin in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **54** 5222–33

[567] Artsimovitch I, Patlan V, Sekine S, Vassylyeva M N, Hosaka T, Ochi K, Yokoyama S and Vassylyev D G 2004 Structural Basis for Transcription Regulation by Alarmone ppGpp *Cell* **117** 299–310

[568] Xu J, Tozawa Y, Lai C, Hayashi H and Ochi K 2002 A rifampicin resistance mutation in the *rpoB* gene confers ppGpp-independent antibiotic production in *Streptomyces coelicolor* A3(2) *Mol. Genet. Genomics* **268** 179–89

[569] Zhou Y N and Jin D J 1998 The *rpoB* mutants destabilizing initiation complexes at stringently controlled promoters behave like 'stringent' RNA polymerases in *Escherichia coli. Proc. Natl. Acad. Sci. U. S. A.* **95** 2908–13

[570] Rodríguez-Verdugo A, Gaut B S and Tenaillon O 2013 Evolution of *Escherichia coli* rifampicin resistance in an antibiotic-free environment during thermal stress. *BMC Evol. Biol.* **13** 50

[571] Dragosits M, Mozhayskiy V, Quinones-Soto S, Park J and Tagkopoulos I 2013 Evolutionary potential, cross-stress behavior and the genetic basis of acquired stress resistance in *Escherichia coli*. *Mol. Syst. Biol.* **9** 643

[572] Herring C D, Raghunathan A, Honisch C, Patel T, Applebee M K, Joyce A R, Albert T J, Blattner F R, van den Boom D, Cantor C R and Palsson B Ø 2006 Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* **38** 1406–12

[573] Barrick J E and Lenski R E 2013 Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14** 827–39

[574] Mei J-M, Nourbakhsh F, Ford C W and Holden D W 1997 Identification of *Staphylococcus aureus* virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis *Mol. Microbiol.* **26** 399–407

[575] Ge X, Kitten T, Chen Z, Lee S P, Munro C L and Xu P 2008 Identification of *Streptococcus sanguinis* genes required for biofilm formation and examination of their role in endocarditis virulence *Infect. Immun.* **76** 2551–9

[576] Yee R, Cui P, Shi W, Feng J and Zhang Y 2015 Genetic Screen Reveals the Role of Purine Metabolism in *Staphylococcus aureus* Persistence to Rifampicin *Antibiotics* **4** 627–42

[577] Chaudhuri R R, Allen A G, Owen P J, Shalom G, Stone K, Harrison M, Burgis T A, Lockyer M, Garcia-Lara J, Foster S J, Pleasance S J, Peters S E, Maskell D J and Charles I G 2009 Comprehensive identification of essential *Staphylococcus aureus* genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics* **10** 291

[578] Samant S, Lee H, Ghassemi M, Chen J, Cook J L, Mankin A S and Neyfakh A A 2008 Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathog.* **4** e37

[579] Leung K Y and Finlay B B 1991 Intracellular replication is essential for the virulence of *Salmonella typhimurium*. *Proc. Natl. Acad. Sci.* **88** 11470–4

[580] Buchmeier N A and Libby S J 1997 Dynamics of growth and death within a *Salmonella typhimurium* population during infection of macrophages *Can. J. Microbiol.* **43** 29–34

[581] Chakraburtty R, White J, Takano E and Bibb M J 1996 Cloning, characterization and disruption of a (p)ppGpp synthetase gene (*relA*) of *Streptomyces coelicolor* A3(2). *Mol. Microbiol.* **19** 357–68

[582] Huang Y-J, Tsai T-Y and Pan T-M 2007 Physiological response and protein expression under acid stress of *Escherichia coli* O157:H7 TWC01 isolated from Taiwan. *J Agric Food Chem* **55** 7182–91

[583] Hattangady D, Singh A, Muthaiyan A, Jayaswal R, Gustafson J, Ulanov A, Li Z, Wilkinson B and Pfeltz R 2015 Genomic, Transcriptomic and Metabolomic Studies of Two Well-Characterized, Laboratory-Derived Vancomycin-Intermediate *Staphylococcus aureus* Strains Derived from the Same Parent Strain *Antibiotics* **4** 76–112

[584] Roller B R K, Stoddard S F, Schmidt T M, Lauro F, Roller B R K, Schmidt T M, Pfeiffer T, Schuster S, Bonhoeffer S, Bachmann H, Stoddard S F, Smith B J, Hein R, Roller B R K, Schmidt T M, Kembel S W, Wu M, Eisen J A, Green J L, Angly F E, Klappenbach J A, Dunbar J M, Schmidt T M, Stevenson B S, Schmidt T M, Dethlefsen L, Schmidt T M, Vieira-Silva S, Rocha E P C, Giovannoni S J, Thrash J C, Temperton B, Eichorst S A, Kuske C R, Schmidt T M, Martiny A C, Treseder K, Pusch G, Condon C, Liveris D, Squires C, Schwartz I, Squires C L, Stouthamer A H, Fegatella F, Lim J, Kjelleberg S, Cavicchioli R, Kurland C G, Carini P, Strzelczyk E, Leniarska U, Morris J J, Lenski R E, Zinser E R, Raven J R, Andrews M, Quigg A, Taylor J R, Stocker R, Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M, Redmond M C, Valentine D L, Shrestha P M, Noll M, Liesack W, Nemergut D R, Young V B, Schmidt T M, Wieder W R, Bonan G B, Allison S D, Lee Z M, Schmidt T M, Eagon R, Conn H J, Datta S, Costantino N, Court D L, Gorlach K, Shingaki R, Morisaki H, Hattori T, Schut F, Schut F, Gottschal J C, Prins R A, Stevenson B S, Eichorst S A, Wertz J T, Schmidt T M, Breznak J A, Eichorst S A, Breznak J A, et al 2016 Exploiting rRNA operon copy number to investigate bacterial reproductive strategies *Nat. Microbiol.* **1** 16160

[585] Tamar S and Papadopoulou B 2001 A Telomere-mediated Chromosome Fragmentation Approach to Assess Mitotic Stability and Ploidy Alterations of *Leishmania* Chromosomes *J. Biol. Chem.* **276** 11662–73

[586] Perry J, Slater H R and Choo K H A 2004 Centric fission - Simple and complex mechanisms *Chromosom. Res.* **12** 627–40

[587] Cupolillo E, Grimaldi G and Momen H 1994 A general classification of new world *Leishmania* using numerical zymotaxonomy *Am. J. Trop. Med. Hyg.* **50** 296–311

[588] Trevisan D A C, Lonardoni M V C and Demarchi I G 2015 Diagnostic methods to cutaneous leishmaniasis detection in domestic dogs and cats *An. Bras. Dermatol.* **90** 868–72

[589] Ayllon T, Tesouro M A, Amusategui I, Villaescusa A, Rodriguez-Franco F and Sainz Á 2008 Serologic and molecular evaluation of *Leishmania infantum* in cats from central Spain *Annals of the New York Academy of Sciences* **1149** 361–4

[590] Dantas-Torres F 2009 Canine leishmaniosis in South America. *Parasit. Vectors* **2** 1–8

[591] Koutsovoulos G, Kumar S, Laetsch D, Stevens L, Daub J and Conlon C 2016 No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini Pnas* **113** 1–6

[592] Geiger T, Francois P, Liebeke M, Fraunholz M, Goerke C, Krismer B, Schrenzel J, Lalk M and Wolz C 2012 The Stringent Response of *Staphylococcus aureus* and Its Impact on Survival after Phagocytosis through the Induction of Intracellular PSMs Expression *PLoS Pathog.* **8** e1003016

[593] Schurch N J, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson G G, Owen-Hughes T, Blaxter M and Barton G J 2016 How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22** 839-51

[594] Seyednasrollah F, Laiho A and Elo L L 2013 Comparison of software packages for detecting differential expression in RNA-seq studies *Brief. Bioinform.* **16** 59–70

[595] Shao X, Kim J, Jeong H J and Levin B 2016 Antibiotic susceptibility of bacterial colonies: An assay and experiments with *Staphylococcus aureus*. *bioRxiv*

[596] Conlon B P, Rowe S E, Gandt A B, Nuxoll A S, Donegan N P, Zalis E A, Clair G, Adkins J N, Cheung A L and Lewis K 2016 Persister formation in *Staphylococcus aureus* is associated with ATP depletion. *Nat. Microbiol.* **1** 16051

[597] Church D M, Schneider V A, Steinberg K, Schatz M C, Quinlan A R, Chin C-S, Kitts P A, Aken B, Marth G T, Hoffman M M, Herrero J, Mendoza M L, Durbin R, Flicek P, Consortium I, Durbin R, Dunham I, Kundaje A, Aldred S, Collins P, Davis C, Doyle F, Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Collins F, Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Dennis M, Nuttle X, Sudmant P, Antonacci F, Graves T, Nefedov M, Watson C, Steinberg K, Huddleston J, Warren R, Malig M, Schein J, Church D, Schneider V, Graves T, Auger K, Cunningham F, Bouk N, Butler J, MacCallum I, Kleber M, Shlyakhter I, Belmonte M, Lander E, Zerbino D, Birney E, Schatz M, Delcher A, Salzberg S, Genovese G, Handsaker R, Li H, Altemose N, Lindgren A, Chambert K, Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Fritz M H-Y, Leinonen R, Cochrane G, Birney E, Danecek P, Auton A, Abecasis G, Albers C, Banks E, DePristo M, Raney B, Dreszer T, Barber G, Clawson H, Fujita P and Wang T 2015 Extending reference assembly models *Genome Biol.* **16** 13

[598] Pader V, Hakim S, Painter K L, Wigneshweraraj S, Clarke T B and Edwards A M 2016 *Staphylococcus aureus* inactivates daptomycin by releasing membrane phospholipids *Nat. Microbiol.* **2** 16194