**MASARYK UNIVERSITY**

**Faculty of Medicine**

**Department of Biology**

# Comparative studies on *Treponema pallidum*: comparative genomics of yaws agents and phenomics of BAC clones of syphilis agent

**Mgr. Darina Čejková**

**Doctoral Thesis**

**Supervisor: doc. MUDr. David Šmajs, Ph.D.**

**Brno 2012**

# Bibliografická identifikace

| | |
|---|---|
| JMÉNO A PŘÍJMENÍ: | Darina Čejková |
| NÁZEV DISERTAČNÍ PRÁCE: | Komparativní studie *Treponema pallidum*: komparativní genomika původce yaws a fenomika klonů BAC knihovny původce syfilidy |
| STUDIJNÍ OBOR: | Lékařská biologie 5103V022 |
| ŠKOLITEL: | doc. MUDr. David Šmajs, Ph.D. |
| ROK OBHAJOBY: | 2012 |
| KLÍČOVÁ SLOVA: | *Treponema pallidum*, yaws, syfilis, komparativní genomika, *rrn* operon, fenotypový microarray, sekvencování nové generace |

# Bibliographic identification

| | |
|---|---|
| NAME AND SURNAME: | Darina Čejková |
| TITLE OF DOCTORAL THESIS: | Comparative studies on *Treponema pallidum*: comparative genomics of yaws agents and phenomics of BAC clones of syphilis agent |
| STUDY PROGRAMME: | Medical Biology 5103V022 |
| SUPERVISOR: | doc. MUDr. David Šmajs, Ph.D. |
| DEFENDED IN: | 2012 |
| KEYWORDS: | *Treponema pallidum*, yaws, syfilis, comparative genomics, *rrn* operon, phenotype microarray, next-generation sequencing |

My work could not have been carried out without contribution from other people. I would like to thank my supervisor, doc. MUDr. David Šmajs, Ph.D. for his advices, ideas, support and encouragement while working in his laboratory and abroad.

I am grateful to Professor George Weinstock, Ph.D. from The Genome Institute at Washington University, St. Louis, Missouri, USA for the environment capable to improve my knowledge, skills and experience during my stay in his laboratory.

Many thanks belong to my colleagues who share the interest in comparative genomics of pathogenic treponemes, namely Dr. Petra Matějková, Dr. Michal Strouhal, Dr. Lei Chen, Mgr. Lenka Mikalová, and especially Mgr. Marie Zobaníková for overnight discussions.

# CONTENT:

# ABSTRACT

The species *Treponema pallidum* includes several obligate human pathogens with similar genome sizes and a high degree of sequence similarity. The subspecies *pallidum* (TPA) is an agent of sexually transmitted syphilis, the subspecies *pertenue* (TPE) causes non-venereal tropical disease yaws, and the subspecies *endemicum* (TEN) causes non-venereal endemic syphilis. In addition to human pathogens, the simian Fribourg-Blanc strain, clustering within the TPE group, causes a yaws-like disease in baboons. Other closely related treponemes include *T. paraluiscuniculi* and *T. paraluisleporis*, the agents of rabbit and hare venereal disease, respectively. All these organisms cannot be continuously cultivated *in vitro* and they are considered as non-cultivable pathogenic treponemes.

The diseases caused by these treponemes are quite distinct, although the pathogens are microscopically and serologically indistinguishable, being >99.5% identical based on the whole genome fingerprinting (WGF) results. These results showed that no major gene deletions, insertions or rearrangement are present among pathogenic treponemes. Relatively subtle genetic differences are thus responsible for different pathogenicity and host range.

To precisely define genetic differences between TPE and other subspecies, high quality whole genome sequences of three TPE strains (Samoa D, CDC-2, Gauthier) were determined by a combination of several next-generation sequencing techniques, including Comparative Genome Sequencing, 454 pyrosequencing and Illumina techniques. Paralogous regions, ambiguous bases and gaps were resolved using dideoxy-terminator sequencing. To verify the final genome assemblies, the WGF was compared to the *in silico* restriction enzyme analysis of each sequenced TPE genome. No discrepancies between *in silico* and experimental restriction site analyses of TPE genomes were found, thus estimating a sequencing error rate to be lower than $10^{-4}$.

The TPE Samoa D genome, comprising 1,139,330 bp, was sequenced by all three next-generation sequencing techniques mentioned above, while CDC-2 (1,139,744 bp) and Gauthier (1,139,441 bp) genomes were determined by 454 and Illumina approaches. To determine the Gauthier genome sequence, a pooled segment genomic sequencing method needed to be developed. Briefly, chromosomal DNA was amplified in overlapping segments and equimolar PCR products were pooled and sequenced.

The TPE Samoa D genome was annotated *de novo* using Glimmer3, FgenesB and GeneMark software. For genes with unknown function, a gene size limit of 150 bp was

applied. The genome consists of 1125 genes, including 54 untranslated genes (coding for rRNAs, tRNAs and other non-coding RNAs). The coding regions comprise 95.36% of the genome length. In total, 3 pseudogenes, 648 genes encoding proteins with predicted function, 141 genes encoding conserved hypothetical proteins, 129 genes encoding treponemal conserved hypothetical proteins, and 147 genes encoding hypothetical proteins were annotated.

The TPE Samoa D gene annotation was used as a scaffold for the TPE CDC-2 and Gauthier, and TPA DAL-1 annotation. The same number of genes was preserved in both CDC-2 and Gauthier TPE strains. Two genes were re-annotated in the CDC-2 genome due to frameshuft mutation caused by different number of homopolymers within homopolymeric strings. In total, 10 genes were re-annotated in the Gauthier genome, 9 genes due to grameshift mutation (6 within homopolymeric tract) and one gene with a mutation in the stopcodon.

In total, 16 genes needed to be re-annotated in the TPA DAL-1 strain, including 14 genes carrying a frameshift mutation (7 within homopolymeric tract) and 2 genes with a cluster of multiple mutations. As a consequence, 4 Samoa D orthologues were not annotated, 5 genes were considered pseudogenes, 2 gene loci were annotated as 4 genes (each locus was split into 2 genes) and 5 genes were truncated or elongated based on manual prediction.

The major intra- and inter-strain variability among TPE strains was observed within *tpr* (*Treponema pallidum* repeat) genes, genes coding for putative virulence factors. Whereas *tpr*K (TPE_0897) gene showed the highest intra-strain heterogeneity, the *tpr*D (TPE_0131) gene showed the highest heterogeneity between strains. In addition to high variability observed in *tpr* genes, different numbers of nucleotides within several homopolymeric strings were revealed between and within TPE strains. Indels within homopolymeric strings can result in frameshift mutation or different gene expression.

Additional frameshift mutations were observed only in the Gauthier strain, affecting three genes coding for hypothetical (TPEGAU_0856a) or treponemal conserved hypothetical proteins (TPEGAU_0629, TPEGAU_0858).

Among all examined TPE strains, three genes with variable number of tandem repeats do not disrupt an open reading frame. The number of 60-bp tandem repeat in the *arp* gene (TPE_0433 encoding acidic repeat protein) varies between 4 and 12, the number of 24-bp tandem repeats in TPE_0470 (conserved hypothetical protein) is between 12 and 37, and the

number of 9-bp tandem repeats in TPE_0967 (treponemal conserved hypothetical protein) alternates between two and four.

Two other indels, preserving the open reading frames, were identified in TPE_0067 (cell division protein) and TPE_0136 (outer membrane protein). Interestingly, the deleted loci were terminated by direct repeats. Moreover, TPE_0136 harbors additional single nucleotide mutations. Within TPE strains, three genes were found to be under positive selection, including *tp*92 (TPE_0326, outer membrane protein), *mcp*2 (TPE_0488, methyl-accepting chemotaxis protein) and TPE_0548 (treponemal conserved hypothetical membrane protein). These genes likely encode virulence factors or antigens.

The identified changes that are specific to individual strains represent suitable targets for molecular diagnostics and epidemiologic typing.

The *rrn* operons were examined in 20 strains of non-cultivable pathogenic treponemes, including 11 strains of TPA, 5 strains of TPE, 2 strains of TEN, a simian Fribourg-Blanc and a rabbit *T. paraluiscuniculi* strain. Every examined strain carries 2 *rrn* operons, each with a different 16S-23S ribosomal intergenic spacer, containing a gene encoding either tRNA-Ala or tRNA-Ile. With the exception of genes coding for tRNAs, only a deletion upstream of the 16S rDNA and additional 17 heterogenous nucleotide positions were found within the *rrn* operons among treponemal strains. The sequence of the *rrn* operons reflects the classification, while different *rrn* spacer patterns (Ile/Ala and Ala/Ile) appeared to be randomly distributed across the species/subspecies classification, time, and geographical source of the treponemal strains. Moreover, 16S-23S ribosomal intergenic spacers were determined for 30 clinical samples (of the Czech Republic origin) belonging to 5 different genotypes. All clinical samples showed the Ile/Ala pattern. The reciprocal translocation of genes coding for tRNA is likely mediated by a *rec*BCD-like recombination pathway.

A minimal set (19 clones) of the BAC library of TPA Nichols DNA was tested under 1342 different phenotype conditions. Overall, 190 substrates were tested as sole carbon sources and 192 pH and osmotic sensitivity and 960 antibiotic resistance assays were performed. An increased resistance to lincomycin and cephalosporins was observed in the DSTP001 clone, resistance to cinoxacin and 2,4-diamino-6,7-diisopropyl-pteridine in the DSTP334 clone, and resistance to guanazole and D,L-serine hydroxamate in the DSTP094 clone when compared to other clones and the host *E. coli* strain. The treponemal genes causing the higher resistance of these BAC clones need to be further examined.

# ABSTRAKT

Druh *Treponema pallidum* zahrnuje obligátní lidské patogeny, jež mají podobnou velikost genomu a vysokou sekvenční homologii. Poddruh *pallidum* (TPA) je původcem sexuálně přenosné syfilis, poddruh *pertenue* (TPE) způsobuje nevenerické tropické onemocnění yaws a poddruh *endemicum* (TEN) způsobuje nevenerickou endemickou syfilis. Opičí izolát Fribourg-Blanc, který řadíme k TPE kmenům, způsobuje onemocnění podobné yaws u paviánů. Mezi další blízce příbuzná treponemata patří *T. paraluiscuniculi* a *T. paraluisleporis*, původce králičí a zaječí venerické syfilis. Všechna výše zmíněná treponemata nelze kultivovat v podmínkách *in vitro* a jsou považována za (nekultivovatelná) patogenní treponemata.

Ačkoliv jednotlivá patogenní treponemata nelze od sebe morfologicky ani sérologicky rozeznat, klinická manifestace jednotlivých onemocnění je značně odlišná. Metoda whole genome fingerprinting (WGF) potvrdila více než 99,5% sekvenční identitu mezi poddruhy TPA a TPE. Žádné rozsáhlé inzerce, delece či chromozomové přestavby nebyly touto metodou objeveny. Lze tedy předpokládat, že rozdílná patogenita je způsobena pouze malými genetickými změnami.

Aby bylo možné identifikovat tyto malé genetické změny mezi TPE a dalšími poddruhy, tři zástupci poddruhu TPE (Samoa D, CDC-2 a Gauthier) byly podrobeny vysoce kvalitní celogenomové sekvenaci. Výsledné genomové sekvence byly získány s použitím alespoň dvou sekvenačních technik „nové generace“ (next-generation sequencing), technikami komparativního genomového sekvencování, 454 pyrosekvencování nebo sekvencování metodou Illumina. Paralogní oblasti, nejednoznačně určené nukleotidy a mezery mezi kontigy byly sekvencovány dideoxyterminátorovou metodou. Výsledné celogenomové sekvence byly ověřeny metodou WGF. Experimentálně získaný restrikční profil byl porovnán s restrikčním profilem získaným analýzou výsledné sekvence *in silico*. Tento postup neodhalil žádný rozdíl. Proto je možné předpokládat, že frekvence sekvenační chyby je nanejvýše $10^{-4}$.

Genom kmene TPE Samoa D, který obsahuje 1139330 bp, byl sekvencován všemi třemi sekvenačními metodami nové generace, kdežto kmeny CDC-2 (1139744 bp) a Gauthier (1139441 bp) byly sekvencovány metodou 454 a Illumina. Aby bylo možné určit kompletní sekvenci kmene Gauthier, musela býti vyvinuta metoda Pooled Segment Genomic Sequencing. Chromozomální DNA kmene Gauthier byla amplifikována v překrývajících se oblastech. Vzniklé PCR produkty byly ekvimolárně smíchány a sekvencovány.

Genom kmene TPE Samoa D byl anotován *de novo* za použití softwarů Glimmer3, FgenesB a GeneMark. Pro geny kódující proteiny s nezmámou funkcí byla stanovena minimální hranice 150 bp. Genome kmene Samoa D se sestává z 1125 genů, z nichž 54 je nepřekládáno (tyto geny kódují tRNA, rRNA a nekódující RNA). Kódující oblasti zaujímají 95,36 % délky genomu. Celkově byly anotovány 3 pseudogeny, 648 genů kódujících proteiny se známou funkcí, 141 genů kódujících konzervovavé hypotetické proteiny o neznámé funkci, 129 genů kódujících treponemální konzervovavé hypotetické proteiny a 147 genů kódujících hypotetické proteiny.

Anotace kmene Samoa D byla použita jako předloha pro anotaci kmenů TPE CDC-2, Gauthier a TPA DAL-1. U obou dalších TPE kmenů byl zachován stejný počet anotovaných genů. Odlišný počet nukleotidů v homopolymerních repeticích způsobil v genomu CDC-2 posunovou mutaci ve 2 genech, které byly reanotovány. V genomu Gauthier bylo reanotováno celkem 10 genů, včetně 9 genů s mutací způsobující posun čtecího rámce (6 z nich se nacházelo v oblasti homopolymerních repetic) a mutaci ve stop kodonu.

Celkem 16 genů bylo reanotováno u kmene TPA DAL-1, včetně 14 genů s mutací způsobující posun čtecího rámce (sedm z nich se nacházelo v oblasti homopolymerních repetic), 2 genů se shluky mnohočetných změn, 2 genových lokusů anotovaných jako 4 geny (oba lokusy byly anotovány jako 2 geny) a 5 genů, které byly zkráceny či prodlouženy v závislosti na manuální anotaci.

Nejvyšší vnitrokmenová a mezikmenová variabilita mezi zástupci poddruhu TPE byla pozorována v oblasti *tpr* (*Treponema pallidum* repeat) genů, jež jsou považovány za geny kódující virulenční faktory. Největší vnitrokmenová heterogenita byla prokázána v genu *tpr*K (TPE_0897), naopak gen *tpr*D (TPE_0131) vykazoval největší variabilitu mezi kmeny. Vysoká mezikmenová i vnitrokmenová variabilita byla sledována nejenom uvnitř *tpr* genů, ale také v několika oblastech s variabilním počtem nukleotidů uvnitř homopolymerní repetice. Nalelzené indely v těchto oblastech mohou mít za následek posunovou mutaci či rozdílnou regulaci genové exprese.

Pouze v kmeni Gauthier byly objeveny další mutace, které způsobují posun čtecího rámce u celkově tří genů. Gen TPEGAU_0856a kóduje hypotetický protein a geny TPEGAU_0629 a TPEGAU_0858 kódují treponemální konzervované hypotetické proteiny.

U všech vyšetřených TPE kmenů byl popsán rozdílný počet tandemových repetic, které nezpůsobují posun čtecího rámce. Počet 60bp tandemových repetic genu *arp* (TPE_0433 kódujícího acidic repeat protein) kolísá mezi 4 a 12, počet 24bp tandemových repetic genu TPE_0470 (konzervovaný hypotetický protein) se nachází v rozmezí 12 a 37

a počet 9bp tandemových repetic genu TPE_0967 (treponemální konzervovaný hypotetický protein) alternuje mezi dvěmi a čtyřmi repeticemi.

Dva další indely, které nezapříčinily posun čtecího rámce, byly identifikovány v genech TPE_0067 (protein, který se účastní buněčného dělení) a TPE_0136 (protein vnější membrány). Oba deletované úseky byly ohraničeny přímými repeticemi. Navíc gen TPE_0136 nese další jednonukleotidové záměny. V rámci poddruhu TPE byly nalezeny tři geny pod pozitivním selekčním tlakem. Jsou to: *tp*92 (TPE_0326, protein vnější membrány, povrchový antigen Tp92), gen *mcp*2 (TPE_0488, protein, který se účastní chemotaxe) a TPE_0548 (treponemální konzervovaný hypotetický membránový protein). Je tedy možné, že tyto geny kódují virulenční faktory poddruhu TPE.

Některé z identifikovaných změn mezi TPE kmeny se mohou stát vhodnými kandidáty pro molekulární diagnostiku a epidemiologickou typizaci.

Operony s geny kódující ribozomální RNA (*rrn*) byly vyšetřeny ve 20 kmenech nekultivovatelných patogenních treponemat, včetně 11 kmenů poddruhu TPA, 5 kmenů poddruhu TPE, 2 kmenů poddruhu TEN, opičího izobátu Fribourg-Blamc a králičího zástupce druhu *T. paraluiscuniculi*. Každý vyšetřený kmen obsahoval 2 *rrn* operony, které se lišily v oblasti 16S-23S ribozomálního intergenového mezerníku. Tento mezerník obsahuje jeden gen kódující tRNA, a to buď gen pro alanin-tRNA nebo gen kódující izoleucin-tRNA. Oba geny pro tRNA jsou vždy přítomny v genomu, mění se tedy jejich lokalizace. Pokud pomineme geny kódující odlišné tRNA produkty, pouze jedna delece předcházející genu pro 16S rRNA a dalších 17 heterologních nukleotidů byly nalezeny v *rrn* operonech treponemálních kmenů. Struktura konzervovaných a heterologních oblastí *rrn* operonu odpovídala klasifikaci treponemálních kmenů. Nicméně odlišná lokalizace genů pro tRNA (vzorec Ile/Ala nebo Ala/Ile) se zdá býti náhodná navzdory druhové a poddruhové klasifikaci, času a místa původu izolátů. Oblast 16S-23S ribozomálního intergenového mezerníku byla dále ještě vyšetřena u 30 klinických izolátů (pocházejících z České republiky), které náležely do pěti odlišných genotypů. Všechny klinické izoláty obsahovaly stejný vzorec (Ile/Ala), kdy gen pro izoleucin-tRNA byl lokalizován v *rrn*1 a gen pro alalnin-tRNA byl umístěn v *rrn*2 operonu. Reciproká translokace *rrn* operonů je pravděpodobně zajištěna rekombinačním mechanizmem, který je podobný mechanizmu dráhy *rec*BCD.

Nejmenší sada (19 klonů) BAC knihovny kmene TPA Nichols byla podrobena testům zkoumajícím rozličné fenotypové projevy. Celkově bylo provedeno 1342 paralelních testů; 190 substrátů bylo testováno jako jediných zdroj uhlíku, 192 testů sledovalo citlivost bakterií k různému pH a osmotického prostředí a 960 testů prověřovalo rezistenci k antibiotikům. Tři

klony BAC knihovny vykazovaly zvýšenou rezistenci. Klon DSTP001 byl více rezistentní k lincomycinu a cefalosporinům, klon DSTP334 vykazoval zvýšenou rezistenci k antibiotikům cinoxacin a 2,4-diamino-6,7-diisopropyl-pteridin, klon DSTP094 byl více rezistentní k antibiotikům guanazol a D,L-serin hydroxamát. Konkrétní treponemální geny, které zvyšují odolnost hostitelské *Escherichia coli* k výše zmíněným antibiotikům, je nutné určit.

.

# 1. INTRODUCTION

The genus *Treponema* includes host-associated, anaerobic organisms belonging to spirochetes (Domain *Bacteria*, Phylum *Spirochaetes*, Class *Spirochaetes*, Order *Spirochaetales*, Family *Spirochaetaceae*) (Paster & Dewhirst, 2000). The spirochetes share a unique spiral cell architecture with periplasmic flagella, and comprise both pathogenic and non-pathogenic strains.

Within the mammalian habitat, the treponemes can be devided into (i) oral treponemes (usually commensal organisms; although high incidence of oral treponemes was associated with periodontal disease and endodontic infection), (ii) skin-associated treponemes (commensal), (iii) digital dermatitis treponemes (causing digital dermatitis in dairy), (iv) intestinal treponemes (commensal), (v) a group of treponemes closely related to *T. pallidum*, the agent of syphilis (Norris *et al.*, 2006). Many treponemes contribute to commensal flora of the host organism and can be cultivated *in vitro*. However, *T. pallidum* is an obligate human pathogen which cannot be continuously cultivated under *in vitro* conditions.

## 1.1. Closely related non-cultivable pathogenic treponemes

The species *Treponema pallidum* comprises several human pathogens including subspecies *pallidum* (TPA), causing venereal syphilis, subspecies *pertenue* (TPE), causing tropical disease yaws, and subspecies *endemicum* (TEN), causing endemic syphilis. Whereas syphilis is venereal disease, yaws and endemic syphilis are non-venereal, so-called endemic treponematoses. The following short disease characteristics were taken from reviews of Antal *et al.* (2002), Norris *et al.* (2001; 2006), Kestelyn (2010) and Farnsworth and Rosen (2006).

Syphilis is a multistage systemic infection caused by TPA with worldwide distribution. Only several dozen (Magnuson *et al.*, 1956) treponemes are needed to cause disease. Usually, treponemes are inoculated onto mucosal or skin tissue during sexual intercourse, thus adolescents and adults are primary carriers. The course of the disease was studied in detail by The Oslo study (Clark & Danbolt, 1955; Harrison, 1956) and the Tuskegee study (White, 2000), dividing syphilis into primary, secondary, latent and tertiary stages. If untreated, more than 30% of patients can progress to the tertiary (gummatous,

cardio- and/or neurosyphilitic) stage. Moreover, syphilis can be transmitted vertically to the fetus via transplacental invasion or amniotic fluid infecton (Wendel *et al.*, 1991).

Yaws is the most prevalent non-venereal treponematosis. The agent, TPE, is transmitted by direct skin contact through open lesions, bites and excoriations. This chronic and relapsing disease leads to skin ulceration, cartilage, joints and bones destruction. If untreated, 10% of patients can progress to the tertiary stage with deformities of skin and subcutaneous tissues. TPE is supposed to be less virulent than TPA. The reservoir is formed by children (< 15 years) in humid and tropical areas of South America, Africa, the Caribbean islands and Indonesia. The congenital mode of transmission is unlikely but several cases were reported (Roman & Roman, 1986).

Endemic syphilis affects children, adolescents and adults in arid areas (Africa, Middle East). The agent, TEN, is less virulent than TPA and TPE. Oral contact (or common feeding utensils) is the main mode of transmission; congenital transmission is rare but plausible (Akrawi, 1949). The primary stage is usually asymptomatic. If untreated, malformation of the nasopharynx, skin and bones can occur.

The World Health Organization (WHO) worldwide estimates about 12 million new cases of syphilis per year (WHO, 2001) with 4 million new cases in Sub-Saharan Africa, 4 million in South and Southeast Asia, abd 3 million in the Caribbean and Latin America. In total, 460,000 cases of yaws were estimated worldwide (Asiedu *et al.*, 2008), 400,000 from West and Central Africa and 50,000 from Indonesia and Pacific islands. Low worldwide prevalence was estimated for endemic syphilis (WHO, 1998).

Displaying distinct clinical manifestations, the agents were originally classified as three species (*T. pallidum*, *T. pertenue*, *T. endemicum*). However, Miao and Fieldsteel (1980) revealed > 95% DNA homology between individual agents with a DNA-DNA cross-hybridization study, resulting in re-classification of the agents as subspecies (Smibert, 1984). Moreover, the TPA, TPE and TEN subspecies are morphologically (Engelkens *et al.*, 1991; Hovind-Hougen *et al.*, 1976; Ovcinnikov & Delektorskij, 1971) and serologically identical (Baker-Zander & Lukehart, 1983; Noordhoek *et al.*, 1990).

Clinical manifestation and epidemiological characteristics are the primary factors used to diagnose individual treponematosis (Antal *et al.*, 2002; Clyne & Jerrard, 2000). To confirm treponemal infection, tests are used that are based on morphology (dark-field microscopy) and both non-treponemal-specific serology (Rapid Plasma Reagin, Venereal Disease Research Laboratory) and treponemal-specific serology (*T. pallidum*

Hemagglutination, Fluorescent Treponema Antibodies-Absorbed, *T. pallidum* Particle Agglutination) (Ho & Lukehart, 2011). Indeed, many cases of treponematoses may be mis-diagnosed, due to human mobility (Vabres, 2011), overlapped distribution of syphilis and endemic treponematoses in some areas (Akrawi, 1949; Antal *et al.*, 2002; Wilson, 1973), poor expertise of health care providers (Farnsworth & Rosen, 2006) or an unexpected clinical manifestation or means of transmission (Mitja *et al.*, 2011; Roman & Roman, 1986; Turner & Hollander, 1957).

All three *T. pallidum* subspecies are obligate human pathogens and no animal (non-human) reservoir exists. Thus far, attempts to continously cultivate these organisms *in vitro* have been unsuccessful (Cox & Radolf, 2006). Prolonged cultivation of TPA strains was achieved in tissue culture when attached to cotton-tail rabbit epithelial (Sf1Ep) cells (Fieldsteel *et al.*, 1981; Norris, 1982) or RAB-9 rabbit fibroblasts (Cox, 1994). On the contrary, TPE and TEN isolates failed to grow in tissue culture (Cox, 1994; Cox *et al.*, 1984).

To maintain the viability of *T. pallidum* laboratory samples, the New Zealand white rabbit is the animal of choice for all three subspecies (Norris *et al.*, 2006). Rabbits are inoculated intradermally or intratesticularly. Rabbits inoculated with TPA strains develop lesions similar to those in human syphilis (Schell, 1983).

Rabbits are natural hosts for the closely related species *T. paraluiscuniculi* (Baker-Zander & Lukehart, 1984; Centurion-Lara *et al.*, 1997; Šmajs *et al.*, 2011), the agent of venereal rabbit syphilis (Bayon, 1913). The causative agent of human pinta, *T. carateum,* and the unclassified simian isolate Fribourg-Blanc, are other closely related non-cultivable pathogenic treponemes. *T. carateum* has probably been eradicated (Falabella, 1994), and no laboratory isolate exists. Recently, it was also shown that hare isolate *T. paraluisleporis* (Lumeij, 2010; Lumeij *et al.*, 1994), the agent of hare venereal syphilis, is similar to *T. paraluiscuniculi* based on 16S rDNA sequencing and is also able to be propagated in rabbits (Šmajs, unpublished results).

## 1.2. *Treponema pallidum* ssp. *pertenue* and neglected tropical disease yaws

By the early 1950s, when WHO established The Global Yaws Control Programme (Rinaldi, 2008), the estimated prevalence of yaws was about 50 to 100 million cases worldwide. During 1952-1964 WHO and the United Nations Children's Emergency Fund

treated more than 300 million people in 46 countries by a single intramuscular injection of long-acting penicillin (Guthe, 1960). By 1964, the global prevalence had dropped to 2.5 million (Antal & Causse, 1985). However, the disease was not eradicated entirely. The reservoir remained in rural areas with poor hygiene, housing and health care conditions (Walker & Hay, 2000). The disease "at the end of the road" started to re-emerge since the end of the 1970s in Africa, Asia, South America and the Pacific islands, which prompted WHO's Trepomenatoses Control Programme (Amin *et al.*, 2010; Asiedu *et al.*, 2008). The yaws control program was integrated into primary health care in developed cities but not in rural areas (Meheus, 1985). Recently, yaws reservoirs include western and central Africa, Southeast Asia and Pacific islands with about a half-million infected people (Capuano & Ozaki, 2011; Fegan *et al.*, 2010; Gerstl *et al.*, 2009; Mitja *et al.*, 2011). New hopes to get rid of yaws have emerged with a new planned eradication campaign (Amin *et al.*, 2010; Asiedu *et al.*, 2008; Maurice, 2012; Rinaldi, 2008).

The etiologic agent of yaws, *Treponema pallidum* ssp. *pertenue* (TPE) was discovered in 1905 (Castellani, 1905) when Dr. Schaudinn, co-discover of the syphilis spirochete (Schaudinn & Hoffman, 1905a; Schaudinn & Hoffman, 1905b), confirmed the presence of "*Spirochaeta pallida*" in patients with yaws. Although TPE is typically considered to be strictly a human pathogen, Castellani (1907) inoculated monkeys (*Macacus sp.*) with yaws spirochetes shortly after his discovery. Additionally, chimpanzees, rabbits, guinea pigs and especially hamsters are susceptible for TPE propagation (Norris *et al.*, 2006; Turner & Hollander, 1957). For comparison, TPA strains do not develop skin infection in hamsters but they can be easily propagated in rabbits.

The geographic distribution of yaws is similar to the distribution of free-ranging monkeys and apes. In the 1960s, Fribourg-Blanc *et al.* searched for any treponemal disease in free-ranging monkeys. While no infected monkey was found in Kenya and Cambodia, 65% of baboons in the Republic of Guinea showed a high serological reaction against TPA Nichols strain (Fribourg-Blanc *et al.*, 1963; Fribourg-Blanc *et al.*, 1966). The captured baboons did not show any pathological lesions, but treponemes similar to TPA or TPE were found in their popliteal lymph nodes (Fribourg-Blanc & Mollaret, 1969). The isolates, when inoculated into hamsters, caused skin lesions. Only one strain, later called Fribourg-Blanc, was successfully passaged and became the type strain for non-human pathogenic treponemes. Based on morphology, serological cross-reactivity, susceptibility to grow in hamster and geographical

distribution, it was supposed that Fribourg-Blanc and TPE strains are identical organisms. Other studies confirmed morphological and serological relatedness to pathogenic human treponemes, including resistance to *in vitro* cultivation and active yaws-like treponematosis in human beings (Baker-Zander & Lukehart, 1984; Engelkens *et al.*, 1991; Hovind-Hougen *et al.*, 1976; Ovcinnikov & Delektorskij, 1970; Sepetjian *et al.*, 1969; Smith *et al.*, 1971; Turner & Hollander, 1957).

The hunt for simian treponematoses revealed a high-titer of anti-treponemal antibodies among chimpanzees and gorillas (Fribourg-Blanc & Mollaret, 1969). Another study reported simian treponematoses in chimpanzees (20% prevalence) and gorillas (14%) killed in the Belgian Congo (Democratic Republic of the Congo), French Congo (Republic of the Congo) and Cameroon at the beginning of the 21$^{st}$ century (Lovell *et al.*, 2000). More recent studies have shown simian treponematoses in Tanzania (Knauf *et al.*, 2011) and in the Republic of Congo (Levrero *et al.*, 2007). About 17% of examined wild gorilla populations in the Republic of Congo developed lesions similar to those from yaws ones. Interestingly, lesions were more prevalent in males, especially in unmated adult males.

Overall, screening for simian treponematoses showed a similar prevalence among wild monkey populations in central and western Africa, in the region overlapping with the last human yaws infections. It is an open question wheather monkeys are a natural reservoir of yaws and if it is possible to transmit treponematosis between human and simian hosts (Capuano & Ozaki, 2011).

Most of my study focused on whole genome sequencing and comparison of three TPE strains. The Samoa D strain was isolated in 1953 in Apia, Western Samoa (Samoa) from a seven-month boy with typical lesions of yaws (Turner & Hollander, 1957). The CDC-2 strain was isolated in 1980 in Akorabo (Ghana) from a patient with skin lesions typical for yaws (Liska *et al.*, 1982). The Gauthier strain was isolated in 1960 in Brazzaville, Republic of Congo (Gastinel *et al.*, 1963).

### 1.3. Genomic and pre-genomic era of *T. pallidum*

For a long time, the pathogenic treponomes were mysterious organisms. Little was known about their biology and unsuccessful attempts of continuous *in vitro* cultivation led researchers to a catch-22: how could treponemes can be cultivated without knowledge about

them and how could knowledge about treponemes be gained without cultivation (Cox & Radolf, 2006). Small progress was made with the application of electron microscopy, protein electrophoresis, biochemical methods and molecular biology (Norris *et al.*, 2001). However, the research on pathogenic treponemes is still limited because genetic manipulation is impossible.

The whole genome sequencing project of TPA Nichols strain was established due to a lack of knowledge and increased global prevalence (WHO, 2001). The TPA Nichols strain was isolated in 1912 in Washington, D.C., USA, from cerebrospinal fluid of a patient with neurosyphilis (Nichols & Hough, 1913). Since then, the Nichols strain has become the type strain for many studies. Despite years of propagation in the laboratory, this strain is still virulent for humans (Fitzgerald *et al.*, 1976; Magnuson *et al.*, 1956).

The genome of TPA Nichols is composed of a single circular chromosome of 1,138,011 Mb (Fraser *et al.*, 1998). The average guanosine and cytosine content (GC content) is 52.8%. Although several studies observed a small plasmid DNA in treponemal samples (Norgard & Miller, 1981; Piruzian, 1989), this observation has not been confirmed. The first assessment of Nichols' size was based on DNA-DNA hybridization (Miao & Fieldsteel, 1978). Miao and Fieldsteel predicted a genome of 9.05 x $10^9$ Da (13.7 Mb) with GC content varying between 52.4 and 53.7%. The genome size was clearly overestimated; however, GC content was in agreement with the subsequent sequencing project. An additional study to predict a genome size used pulsed field gel electrophoresis (Walker *et al.*, 1991). Genomic DNA was fragmented by three restriction endonucleases, *Not*I, *Sfi*I and *Srf*I. The sum of all fragment sizes predicted a genome as small as 0.9 Mb. More precise estimation, ranging between 1.03 and 1.08 Mb, was determined later by physical mapping (Walker *et al.*, 1995).

Eventually, the shot-gun dideoxy-terminator (DDT) sequencing approach was used for a Nichols genome project (Fraser *et al.*, 1998). In total, 1041 open reading frames (ORFs) were predicted by genome annotation tools, with an average gene length of 1023 bp. Gene annotation determined 577 (55.4%) ORFs with predicted protein function, 177 (17.0%) ORFs coding for conserved hypothetical proteins (genes of unknown function but conserved across genera and higher taxonomic levels), and 287 (27.6%) ORFs encoding hypothetical proteins (genes of unknown function). Later, additional five genes were annotated, four ORFs with predicted functions (TP0206a, TP0250a, TP0451a, TP0950a) and an ORF coding for a hypothetical protein (TP0949a). On the contrary, two other ORFs (TP0635, TP0950), both coding for hypothetical proteins, were removed from the updated Nichols annotation (AE000520.1).

ORFs account for 92.9% of the TPA Nichols chromosomal DNA. Only 476 homologues were found between the TPA Nichols genome and the genome of *Borrelia burgdorferi* (strain B31), the Lyme disease spirochete: 76% of those are genes with predicted function (Fraser *et al.*, 1997). When a large number of genes with unknown function are indicated within a bacterial genome, it is plausible that bacterial genomes have been highly over-annotated (Ussery & Hallin, 2004).

With the exception of very small symbiotic bacteria (McCutcheon & Moran, 2012), the genome of *Treponema pallidum* possesses one of the smallest known bacterial genomes. Based on the idea of reductive genome evolution, a close fellowship of TPA with its human host was postulated (Norris *et al.*, 2001). Surprisingly, although the TPA genome lacks a restriction-modification system to defend its genome against foreign infectious elements, no foreign infectious DNA, such as transposons, bacteriophages or extrachromosomal DNA fragments, has been found in the TPA genome. Moreover, no horizontal gene transfer has been described.

Genes involved in glycolysis, DNA replication, DNA repair and gene expression are present, but the TPA genome does not posses genes involved in the Krebs cycle, oxidative phosphorylation pathways, or in synthesis of enzyme co-factors, fatty acids and most amino acids. The primary carbon source is glucose: alternatively, mannose and maltose can be utilized. A transport system is needed to compesate for such limited biosynthetic pathways. About 5% of the TPA genome codes for transporters. TPA has a limited tolerance for heat and oxygen stress (Cox, 1994; Cox *et al.*, 1990) and it lacks genes coding for superoxide dismutase, catalase or peroxidase, as well as a sigma-32 gene regulating gene expression of heat shock proteins.

TPA has a common spirochete morphology, a spiral architecture with periplasmic endoflagella. The rotation of endoflagella around the cytoplasmatic cylinder causes the snake-like movement of the treponeme along its longitudinal axis. Again, about 5% of the TPA genome is devoted to motility and chemotaxis. The outer membrane (OM) of TPA does not contain lipopolysaccharides. When compared to other spirochetes or gram-negative bacteria, a paucity of integral membrane proteins was observed (Bourell *et al.*, 1994; Radolf *et al.*, 1989; Walker *et al.*, 1989). The lack of antigens on their surface protects the treponemes against the host immune response; thus, TPA has been termed a "stealth" pathogen. TPA lives in multiple tissues and is able to adhere to various eukaryotic cells via binding to extra-cellular matrix (ECM) compounds, such as laminin, fibronectin and albumin (Fitzgerald *et al.*, 1975; Peterson *et al.*, 1983). Abundant lipoproteins and nedoflagella are located in the

periplasmic layer of the TPA of cell membrane. Several studies have confirmed that components of the periplasmic layer are highly immunogenic (Blanco *et al.*, 1990; Chamberlain *et al.*, 1989).

When considering virulence factors in TPA, the proteins involved in chemotaxis and motility, outer membrane proteins, lipoproteins and ECM-binding proteins are of interest. The genome sequencing project revealed a 12-member paralogous family of *tpr* (*Treponema pallidum* repeat) genes, marked as *tpr*A to *tpr*L. The *tpr* genes are orthologous to a major sheath protein (MSP) of *T. denticola*, an oral spirochete. The MSP is highly immunogenic and is supposed to be a pore-forming adhesin on the outer membrane surface. A similar porin function has been predicted for Tpr proteins. In addition to the *tpr* genes, other 5 genes, encoding a putative hemolysins, have been proposed as virulence factors. All five genes were cloned into *E. coli*, but no hemolytic activity on sheep blood agar plates was observed (Šmajs *et al.*, 2002).

Since the discovery of the syphilis and yaws treponemes, several studies have been conducted in order to determine differences beyond the different clinical manifestations. Neither electron microscopy nor biochemical tests were successful in answering this question (Baker-Zander & Lukehart, 1983; Engelkens *et al.*, 1991; Hovind-Hougen *et al.*, 1976; Noordhoek *et al.*, 1990; Ovcinnikov & Delektorskij, 1971). Although DNA-DNA cross-hybridization studies (Miao & Fieldsteel, 1980) and protein two-dimensional gel electrophoresis showed subtle differences (Thornburg & Baseman, 1983), these tools were not able to specify differences in greater detail.

Reasonable costs for analysis of restriction fragment length polymorphism, PCR, and DNA sequencing paved new avenues to find differences between TPA and TPE isolates. Noordhoek *et al.* 1989) identified a single base difference between the TPA Nichols strain and the TPE CDC-2575 strain in the *tp*F1 (TP1038) gene at position 123. Centurion-Lara *et al.*(1998) found a single nucleotide change in the *Eco*47III restriction site within the 5' flanking region of the *tpp*15 gene (TP0171) coding for a 15-kDa lipoprotein. Cameron *et al.* (1999) described a single base difference at position 579 in the *gpd* gene (TP0257) coding for glycerophosphodiester phosphodiesterase. The same group later discovered differences in several nucleotide positions in the *tp*92 gene (TP0326), gene coding for a 92-kDa antigen (Cameron *et al.*, 2000). Except for the TP0326 gene, only single nucleotide polymorphisms were found in genes between the TPA and TPE strains iduring the pre-genomic and early genomic era.

### 1.4. Post-genomic era of *T. pallidum*

The complete genome sequence of TPA Nichols opened the door for global-scale research on treponemes. In 2002, Šmajs *et al.* (2002) constructed a treponemal BAC library in *E. coli* strain DH10B. TPA Nichols genomic DNA was fragmented and cloned into the pBeloBAC11 vector in order to perform gene function studies via heterologous expression. Only 13 (out of 1039, 0.25%) genes were unable to be cloned into the library. A minimal set of 19 library clones covers the remaining 1026 Nichols genes at full length.

To further characterize genes, especially hypothetical ones, the transcriptional and proteomic profiles of Nichols strain cultivated in rabbit testes were examined (Šmajs *et al.*, 2005). The most highly expressed genes were genes encoding flagellins and lipoproteins, followed by genes coding for ribosomal and chaperone proteins. Out of 19 abundant proteins, 13 corresponding genes showed high transcriptional activity.

In 2003, a Nichols proteome array was constructed (McKevitt *et al.*, 2003). The proteome array consisted of 991 *E. coli* clones, each harboring an individual TPA ORF. In order to identify TPA antigens, rabbit (McKevitt *et al.*, 2005) and human (Brinkman *et al.*, 2006) antibody reactivities with the proteome array were examined. The rabbit and human immunoproteomes consist mainly of nine and ten proteins, respectively.

In 2008, Titz *et al.* (2008) used plasmid DNA of the proteome array to characterize a protein-protein interaction network. Using an yeast two-hybrid system, he identified 991 high-confidence interactions between 576 Nichols proteins. Moreover, he predicted novel functions for 18 genes, especially those involved in DNA metabolism.

In 2010, McGill *et al.* (2010) studied the treponemal proteome, and the rabbit and human immuno-proteome. Contrary to previous studies based on heterologous expression, the Nichols strain used in this study was isolated from rabbit testes. The study identified 88 polypeptides and revealed 15 highly immunogenic proteins, including three hypothetical ones.

In 2006, a novel computational tool, SpLip, was developed to identify spirochaetal lipoproteins within a genome sequence (Setubal *et al.*, 2006). The SpLip software identifies 58 genes coding for lipoproteins in the TPA Nichols genome.

Comparative genomic studies can reveal additional information about treponemes. Comparative genomics of closely related organisms, e.g. TPA and TPE, can identify genes that cause distinct pathogenicity or tropisms to human tissues and differential ability to grow in various animal models (Šmajs *et al.*, 2012). On the other hand, intra-subspecies genome comparison may identify heterogenous genes which have no effect on type of disease but may

outline genetic differences between distinct strain phenotypes, such as the pattern of virulence. These heterogenous genes may be useful tools for epidemiological typing studies.

In addition to the TPA Nichols strain, the genomes of TPA, TPE, and *T. paraluiscuniculi* (TPc) strains were completed. The complete genome sequence of the TPA SS14 (Matějková *et al.*, 2008) strain was determined by a comparative genome sequencing method (CGS) (Albert *et al.*, 2005). CGS identified heterogenous loci between two genomic DNAs which were separately hybridized to the chip. The chip, consisting of an array of oligonucleotides, was designed based on the known Nichols genome sequence and Nichols genomic DNA (gDNA) was used as the reference hybridization. The SS14 strain was isolated in 1977 in Atlanta, GA, USA from a patient with secondary syphilis (Stamm *et al.*, 1983). The genome consists of 1,139,407 bp. In total, 327 single nucleotide changes, 14 deletions and 18 insertions were found when compared to the Nichols strain (AE000520.1). Unfortunatelly, several sequencing errors have been found in the Nichols genome (Giacani *et al.*, 2012a), so the genome sequence of the SS14 strain is likely to carry the same errors, since it used the Nichols sequence as a reference. To avoid a similar propagation of sequencing errors, new sequencing projects have been based on other next-generation methods (Pětrošová, unpublished results).

The TPA genome of a Chicago strain (Giacani *et al.*, 2010a) was determined using the Illumina sequencing method (Bennett, 2004). The Chicago strain was isolated in 1951 from a patient with primary syphilis in Chicago, IL, USA (Turner & Hollander, 1957). The complete genome consists of 1,139,281 bp. Based on comparison with the Nichols strain (AE000520.1), 44 single nucleotide changes, 103 small (≤ 3 bp) insertions/deletions (indels) and one large insertion were described. However, only 20 single nucleotide changes and 2 indels were confirmed (Giacani *et al.*, 2012a) based on re-appraisal considering the sequencing errors in AE000520.1.

The complete genome of *T. paraluiscuniculi*, Cuniculi A strain, showed further genome reduction (Šmajs *et al.*, 2011) compared to TPA strains. The genome (1,133,390 bp) contained 51 pseudogenes (61.9% of unknown function) and 33 genes with multiple sequence changes. Proteins involved in virulence, gene regulation and DNA metabolism (DNA repair and recombination) were among the most affected. Genome reduction was postulated to be an adaptative process in response to infection of the rabbit.

The complete genome sequence of the TPA DAL-1 strain (Zobaníková *et al.*, 2012) was independently sequenced by both the 454 (Margulies *et al.*, 2005) and Illumina methods and gaps were closed by DDT sequencing. The DAL-1 strain was isolated from the amniotic

fluid of a pregnant woman with secondary syphilis in Dallas, TX, USA in 1991 (Wendel *et al.*, 1991). The complete genome (1,139,971 bp) differs from the re-sequenced Nichols strain by 54 single nucleotide changes, 9 insertions and 9 deletions.

The whole genome sequences of TPE Samoa D, CDC-2 and Gauthier genomes were compared to the re-sequenced TPA Nichols (Pětrošová, unpublished results), re-sequenced SS14 (Pětrošová, unpublished results), Chicago and DAL-1 strains (Čejková *et al.*, 2012). Overall, only 0.2% difference at DNA level was found between TPA and TPE subspecies. The whole genome nucleotide difference was 4.7 - 4.8 times higher between subspecies than within them. In total, 17 genes, consistently divergent between TPA and TPE strains, were found to be under positive selection pressure. Moreover, 11 of them were predicted to be membrane or exported proteins, suggesting their participation in virulence.

Recently, the whole genome sequence of the TPA Mexico A strain was published (Pětrošová *et al.*, 2012). The Illumina high-throughput sequencing platform was used. The gaps were closed by DDT sequencing. The strain was isolated from a patient with primary syphilis in Mexico City, Mexico in 1953 (Turner & Hollander, 1957). The complete genome consists of 1,140,038 bp and differs from Nichols (AE000520.1) by 438 nucleotide replacements, 94 insertions and 38 deletions, and differs from the Chicago strain by 419 substitutions, 18 insertions and 20 deletions. The strain is more closely related to SS14 than it is to the Nichols, DAL-1 and Chicago strains. Although 175 substitutions, 85 insertions and 28 deletions were proposed, the SS14 sequence used in the comparison contains sequencing errors. Surprisingly, two Mexico A genes, TPAMA_0326 (*tp*92, outer membrane protein) and TPAMA_0488 (*mcp*2, methyl-accepting chemotaxis protein) showed a mosaic of TPA and TPE nucleotide block patterns. Considering the mosaic pattern of these genes and place of origin, a recombination event between both subspecies was postulated.

In addition to sequencing projects, physical mapping (whole genome fingerprinting, WGF) was performed for TPA (Nichols, SS14, Mexico A, DAL-1), TPE (CDC-2, Gauthier, Samoa D), a simian isolate (Fribourg-Blanc) and *T. paraluiscuniculi* (Cuniculi A) treponemes (Mikalová *et al.*, 2010; Strouhal *et al.*, 2007).

Several typing systems for treponematoses have been developed since the TPA Nichols complete genome was completed. In 1998, Pillay *et al.* (1998) established the first typing system based on restriction fragment length polymorphism (RFLP) of *tpr*E (TP0313), *tpr*G (TP0316) and *tpr*J (TP0621) genes and the variable number of 60-bp tandem repeats in the *arp* (TP0433-0434; acidic repeat protein). While both ends of the *arp* gene are conserved

among strains and subspecies, the number of repeats and the repeat motif vary between strains (Liu *et al.*, 2007; Pillay *et al.*, 1998). Only repeat motif (II) was observed in non-venereal trepanomatoses, whereas venereal treponematoses (both human and rabbit) contain several distinct motifs (Harper *et al.*, 2008a). Moreover, the gene is under positive selection in venereal strains but not in TPE strains (Čejková *et al.*, 2012; Šmajs *et al.*, 2011).

Another typing system to differentiate treponemal isolates was employed in 2006 (Centurion-Lara *et al.*, 2006). To previously known difference in restriction site within the 5' flanking region of the *tpp* gene (TP0171), Centurion-Lara *et al.* added RFLPs of the *tpr*I (TP0620) and *tpr*C (TP0117) genes.

Interestingly, in the same year, the TP0136 and TP0548 heterogenous genes were used for typing (Flasarová *et al.*, 2006). TP0136 was postulated to be an outer surface lipoprotein antigen with binding capacity to fibronectin (Brinkman *et al.*, 2008). In a subsequent study, only several intact treponemes reacted with anti-TP0136 antisera, whereas all examined disrupted treponemes showed positive reaction, suggesting a periplasmic localization of the protein encoded by TP0136 (Cox *et al.*, 2010). TP0548 gene codes for hypothetical protein with putative outer membrane localization (Cox *et al.*, 2010).

A mutation A2058G in both 23S rDNA genes was observed in the SS14 strain (Stamm & Bergen, 2000b), resulting in resistance to macrolide antibiotics (Lukehart *et al.*, 2004; Mitchell *et al.*, 2006). Prevalence of macrolide resistance in treponemes is increasing among syphilis patients (Katz & Klausner, 2008; Mitchell *et al.*, 2006), and only a few two tests are available to detect the mutation (Lukehart *et al.*, 2004; Pandori *et al.*, 2007). A novel mutation, A2059G, has been found recently in both 23S rDNA genes. Again, the mutation resulted in resistance to macrolide antibiotics (Matějková *et al.*, 2009). The authors developed a new test based on the RFLP of the 23S rDNA gene to detect both mutations.

Most epidemiological studies were performed only for the syphilis spirochete (Flasarová *et al.*, 2012; Marra *et al.*, 2010; Martin *et al.*, 2009; Molepo *et al.*, 2007; Pillay *et al.*, 2002; Sutton *et al.*, 2001). Recently, the treponemal typing system confirmed endemic syphilis in a child patient (Fanella *et al.*, 2012). A one-year old girl was born in Canada, while her siblings and parents came from the Republic of Senegal. Broader examination of the family revealed other siblings positive for endemic syphilis. It was concluded that direct close contact among family members had spread the disease. The typing system was also used in the case of a 10-year pygmy boy from the Republic of Congo (Pillay *et al.*, 2011). The boy, originally suspected of having monkeypox, was tested positive for yaws. This case claimed to be the first yaws report to use the molecular biology method for testing. In addition to human

samples, the simian samples collected from wild baboons in the Republic of Congo were also tested by molecular biology techniques (Knauf *et al.*, 2011). The simian isolates clustered with TPE strains. The same pattern was also found in another simian isolate, Fribourg-Blanc (Harper *et al.*, 2008b), via detecting 21 loci including 4 that were used in the study of Congo baboons. Moreover, an ongoing sequencing project of the Fribourg-Blanc strain did not find any major significant differences between simian and yaws spirochetes (Zobaníková, personal communication).

Molecular typing identified two strains as TPA that were originally classified as TPE. The Haiti B strain was isolated in 1951 in Côtes-de-Fer, Haiti from an 11-year boy with typical yaws lesions (Turner & Hollander, 1957). The strain Madras was isolated in 1954 in Madras, India. Unfortunately, no detailed information about the origin of the Madras strain was published. Both strains grew better in hamsters than in rabbits. Because of strain origins (location, child patient) and the ability to grow in hamsters, the evidence suggested these were TPE strains. However, further analysis indicated that the Haiti B strain clustered within TPA strains (Cameron *et al.*, 1999; Cameron *et al.*, 2000; Centurion-Lara *et al.*, 1998). Enhanced screening tools confirmed that Haiti B and Madras strains belonged to the TPA branch (Harper *et al.*, 2008a; Harper *et al.*, 2008b). It would be very interesting to confirm once again that these two TPA strains carry features of non-venereal strains, especially if it is known that mis-labeling of treponeme strains had occurred in the laboratory in the past (when TPA SS14 strains was labeled as TPE Gauthier).

## 1.5. Focusing on *T. pallidum* gene function

In general, bacterial outer membrane proteins mediate an interaction between bacteria and host environment and they are considered as virulence factors of pathogenic bacteria. *T. pallidum* possess a unique cell wall with a paucity of membrane proteins (Walker *et al.*, 1989), resulting in poor antigenity (Cox *et al.*, 1992). The quest for *T. pallidum* outer membrane proteins is therefore very worthy for understanding the pathogenesis of *T. pallidum* and for vaccine development.

Recent studies have identified several treponemal proteins as candidates for outer membrane proteins (Cameron, 2006), including transporter proteins (encoded by TP0163), lipoproteins (TP0257, TP0453), extra-cellular matrix (ECM)-binding proteins (TP0155,

TP0483, TP0751), Tpr proteins (TP0316, TP0620, TP0897), Tp92 antigen (TP0326), outer membrane protein (TP0136) and Arp protein (TP0433-0434).

TP0163 encodes a periplasmic binding protein of the ABC transporter system with affinity to $Zn^{2+}$ and $Mn^{2+}$ ions (Desrosiers *et al.*, 2007). Strong serological reactions against recombinant Gpd (TP0257, glycerophosphodiester phosphodiesterase), TP0453 (membrane hypothetical protein) and Tp92 (outer membrane protein) proteins were observed using human sera of patients with primary syphilis (Van Voorhis *et al.*, 2003). While lipoprotein TP0453 is periplasmic, the non-lipidated variant is attached to the membrane (Hazlett *et al.*, 2005), suggesting a porter function during outer membrane biogenesis (Luthra *et al.*, 2011). The Tp92 protein was confirmed to be exposed on the surface (Desrosiers *et al.*, 2011) and involved in the virulence mechanism of TPA (Jun *et al.*, 2008). A novel outer membrane candidate (TP0865, hypothetical outer membrane protein) was predicted using a combination of several computational tools (Cox *et al.*, 2010). According to this study, the TP0865 protein received a similar score for the outer membrane protein prediction as Tp92 antigen.

Laminin and fibronecting-binding abilities have been proposed for TP0751 and TP0136 proteins. TP0751 is a $Zn^{2+}$-dependent protease, expressed during syphilis infection (Cameron, 2003; Houston *et al.*, 2011). Attachment to laminin was confirmed *in vivo* using a heterologous expression system in *T. phagedenis*, the treponeme which does not bind to laminin (Cameron *et al.*, 2008). TP0155 and TP0483 also bind fibronectin (Cameron *et al.*, 2004). TP0155 (M23B subfamily peptidase) contains a LysM domain with a peptidoglycan-binding affinity. It is of interest, that *T. denticola* contains 4 homologous proteins of TP0155, and one og them, homologous protein TDE2318 is exposed on the outer membrane (Bamford *et al.*, 2010). The recombinant TP0483 (hypothetical protein) protein has also been shown to bind to fibronectin (Dickerson *et al.*, 2012).

Since the genome of TPA Nichols was determined, research has focused on gene identification, with *tpr* genes and genes coding for membrane proteins are of primary interest. The *tpr* genes were subdivided into three subfamilies based on protein sequence homology (Centurion-Lara *et al.*, 1999), into Subfamily I (*tpr*C, *tpr*D, *tpr*F, *tpr*I), Subfamily II (*tpr*E, *tpr*G, *tpr*J), and Subfamily III (*tpr*A, *tpr*B, *tpr*H, *tpr*K, *tpr*L). Within a strain, Tpr proteins of subfamily I and II show conserved N- and C- termini while central regions vary in length and sequence. Only limited homology is observed among Tpr proteins of Subfamily III. Tpr proteins (especially TprK) are targets for cellular and humoral immune response (Giacani *et al.*, 2007b; Leader *et al.*, 2003; Sun *et al.*, 2004).

Heterogenous sequence variation (Centurion-Lara *et al.*, 2000a; Centurion-Lara *et al.*, 2000b; Centurion-Lara *et al.*, 2006; Giacani *et al.*, 2005a; Sun *et al.*, 2004) were observed and recombination events (Gray *et al.*, 2006) were suggested among homologous proteins of different isolates. Moreover, an intra-strain heterogeneity of the TprK protein was detected during rabbit (LaFond *et al.*, 2006) and human (LaFond *et al.*, 2003) infection. It was postulated that the mechanism of gene conversion created heterogeneity of the *tpr*K gene (Centurion-Lara *et al.*, 2004). Antigenic variation of the TprK protein can help TPA to escape from host immune pressure and to persist within a host. Heterogeneity was also described in *T. denticola* (Gaibani *et al.*, 2010). A Tpr homologue, MSP, was also found to be heterogenous among clinical isolates. As mentioned earlier, MSP is an outer membrane oligomeric protein with porin activity (Fenno *et al.*, 1996).

The TprC/D protein was predicted to be a trimeric porin localized in the outer membrane (Anand *et al.*, 2012), although the exact localization of Tpr proteins is still debated (Cox *et al.*, 2010; Giacani *et al.*, 2005b). Gene expression, gene regulation and phase variation of *tpr* genes has also been studied (Giacani *et al.*, 2007a; Giacani *et al.*, 2007b; Giacani *et al.*, 2009).

In addition, the structure of TPA cell wall was depicted using high-resolution cryo-electron tomography (Liu *et al.*, 2010). Many proteins are periplasmic. If attached to the outer membrane, they are anchored from the periplasmic space and do not span the outer leaflet.

## 1.6. *rrn* operons in treponemes

The ribosomes are ribonucleoproteins composed of rRNAs and ribosomal proteins. Because ribosomes are responsible for protein production in all organisms, rDNA genes as well as genes coding for ribosomal proteins tend to be conserved. The prokaryotic ribosome consists of the 30S and the 50S subunit. The 30S subunit contains 16S rRNA and about 20 ribosomal proteins, while the 50S subunit comprises 23S and 5S rRNAs and about 30 ribosomal proteins. The rDNA genes are co-localized in the ribosomal RNA (*rrn*) operons. The typical bacterial *rrn* operon consists of 16S-23S-5S rDNA genes. In addition, the *rrn* operons may contain genes coding for tRNA and regulatory regions. The *rrn* operons are highly transcribed in bacteria (Condon *et al.*, 1992), especially during the log growth phase. It is generally believed that bacteria with a short generation time have multiple *rrn* operons in the genome. Multiple copies of 16S and 23S rDNA genes in an organism are almost identical

(Pei *et al.*, 2009; Pei *et al.*, 2010), suggesting homogenization of rDNA genes through homologous recombination (Liao, 2000). The 16S and 23S rDNA genes are widely used in bacterial phylogenetic studies, while the 5S rDNA genes are too short to be useful.

In addition to the rDNA genes, the *rrn* operons contain intergenic spacer regions (ISR). The ISRs are not involved in ribosomal function, thus they are not under functional constraints resulting in higher heterogeneity among bacteria (de Vries *et al.*, 2006; Gurtler, 1999). The 16S-23S ISR, containing gene(s) encoding tRNA, is of great interest. The 16S-23S ISRs vary in length, composition of tRNA encoding genes and intragenomic nucleotide diversity (Stewart & Cavanaugh, 2007) and have been used for bacterial identification and molecular typing (Indra *et al.*, 2010; Sadeghifard *et al.*, 2006) and evolutionary studies.

Two *rrn* operons were observed in pathogenic treponemes (Fukunaga *et al.*, 1992) composed of 16S-23S-5S rDNA genes. The 16S-23S ISRs of TPA Nichols (Fraser *et al.*, 1998) contain tRNA-Ile (tRNA-Ile-1; TP_t12) and tRNA-Ala (tRNA-Ala-3; TP_t15) genes within *rrn*1 and *rrn*2 operons, respectively.

Stamm *et al.* (2002) used the sequences of 16S-23S ISRs for molecular typing of dermatitis-associated treponemes in cattle. Cattle dermatitis-associated treponemes are divided into three phylotypes which cluster within the group of human saprophytic treponemes (*T. denticola, T. phagedenis, T. vincentii*). Human saprophytic treponemes are distinct from pathogenic species in several ways. The saprophytic treponemes can be cultured *in vitro.* The genomes of saprophytic treponemes are capable of the acquisition of foreign DNA through horizontal transfer, whereas the genomes of pathogenic treponemes are not. The complete genome of *T. denticola* (Seshadri *et al.*, 2004) is more than twice the size of the complete genome of the TPA Nichols strain (Fraser *et al.*, 1998). *T. denticola* and *T. phagedenis* contain two copies of *rrn* operon while *T. vincentii* contains only one copy.

### 1.7. High-throughput sequencing techniques

The first *de novo* whole genome bacterial sequences were determined by shot-gun sequencing and DNA gap-closure techniques (Fleischmann *et al.*, 1995; Fraser *et al.*, 1997). Genomic DNA was fragmented and cloned into plasmid vectors and sequenced by a paired-ends dideoxy-terminator (DDT) approach (Edwards *et al.*, 1990). Reads were assembled into contigs and gaps between contigs were closed by a primer walking method. To determine

whole genome of eukaryotic cells was even more complicated, requiring cloning into BAC vectors and physical mapping (Mardis, 2008a). Overall, the coverage was low (6x) and error rate was relatively high ($10^{-4}$, Pětrošová, unpublished results).

Since the original sequencing of the TPA Nichols genome, several next-generation sequencing (NGS) technologies have been developed. A short description of 454 pyrosequencing and Illumina sequencing strategies will be provided because both methods were used to accomplish whole genome sequencing of TPE strains.

454 pyrosequencing (Roche) is a massively-parallel sequencing-by-synthesis system (Margulies *et al.*, 2005). Adapter-ligated DNA fragments are attached to DNA-capture beads and each DNA:bead complex is isolated into individual water-in-oil micelles. Each water-in-oil emulsion contains components of PCR reaction, used for the amplification of DNA bound to the beads. Finally, each DNA-bound bead is placed into a single well on a PicoTiterPlate, a fiber optic chip, providing a fixed location for the next steps. A mix of enzymes such as polymerase, sulfurase, and luciferase is also packed into the well. The PicoTiterPlate is then placed into the instrument (454 FLX, 454 FLX Titanium) for sequencing. At this stage, the four nucleotides (T, A, G, C) are washed in series over the PicoTiterPlate. During the nucleotide flow, 100,000 beads, each with million copies of DNA, are sequenced in parallel. When a nucleotide is complementary to the template strand present in a well, the polymerase extends the existing DNA strand by adding the nucleotide(s). Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal which is recorded by the CCD camera in the instrument. Before the next cycle begins, unincorporated nucleotides are washed away. Today, up to 450-bp fragments can be sequenced.

Illumina sequencing is also a massively-parallel sequencing-by-synthesis system (Bennett *et al.*, 2004). Adapter-ligated DNA fragments are attached to the surface of a flow cell. The surface of the flow cell is covered by oligonucleotides that are complementary to the adapters. In the Cluster Station, bridge amplification of a single molecule is performed by DNA polymerase directly on the surface of the flow cell. The amplification step results into millions of copies, clusters, of the original DNA fragment which is subject to the sequencing step in the Illumina instrument (Genome Analyzer, HiSeq, MiSeq). At this stage, all four fluorescently labeled nucleotides (T, A, G. C) are washed simultaneously over the flow cell, along with DNA polymerase and primers. After laser excitation, fluorescence from each cluster is emitted and recorded. To prevent incorporation of multiple nucleotides during the sequencing cycle, labeled nucleotides are blocked at the 3'-OH group. Before the next cycle starts, unincorporated nucleotides and DNA polymerase are washed away, followed by

removal of the fluorophore and unblocking of the 3'-OH termini of incorporated nucleotides (Mardis, 2008a). The next cycles (runs up to 150 cycles are plausible today) are performed in a similar way to the first one. However, there is no need to add new primers.

Other NGS sequencing methods that have been developed and used: SOLiD (sequencing by ligation), Helicos Heliscope, Ion Torrent Life Technologies and Pacific Biosciences SMRT (real-time sequencing). Details about these methods can be found elsewhere (Glenn, 2011; Gupta, 2008; Mardis, 2008a).

NGS can be used not only for whole genome sequencing (including ancient samples), but also for other genomic studies such as metagenomics, genotyping and even immune escape analysis. For example, the Human Microbiome Project (http://nihroadmap.nih.gov/hmp) has several aims: to sequence about 3000 bacterial genomes, to genotype (by 16S rDNA sequencing) bacteria living in human healthy volunteers from 15 (18 in females) body sites, and lastly to screen the metagenomes in these body site samples. All steps are performed either by 454 or Illumina sequencing (Weinstock, 2012). Moreover, NGS can be in applied in transcriptomic studies (RNA-seq). RNA-seq has some advantages when compared to earlier approaches. It is more sensitive than hybridization techniques, no reference sequenced genome is needed, and expression of mRNA and even non-coding RNA can be studied. The chromatin immunoprecipitation and sequencing (ChIP-seq) method is the last major application suitable for NGS. Using this method, DNA-protein interaction and epigenetic regulation (histone modification) can be examined (Luciani *et al.*, 2012; Mardis, 2008a, b).

## 1.8. High-throughput phenotyping technique - Phenotype MicroArrays

Phenotype is a function of genotype and environmental conditions. Bacteria occupy almost all environmental niches on the globe with a huge adaptation ability to the different conditions. Bacteria adapt to inconvenient conditions such as growth competition, toxic chemicals, temperature, pressure, electromagnetic radiation and desiccation (Bochner, 2009).

Recent bacterial taxonomy continues to be based upon global phenotypic profiles, despite all the genotypic information that sequencing data have provided. Phenotype description is more subjective than objective. On the other hand, DNA sequence typing is usually based only on one gene, 16S rDNA. A combination of genotyping and phenotyping

methods should be used in taxonomy, epidemiology and diagnostics because they provide complementary information and each method alleviates the disadvantages of the other.

To study a whole set of microbial phenotypes, many assays or experiments were needed. In 2001, the BiOLOG company (Bochner *et al.*, 2001) designed a high-throughput Phenotype MicroArray (PM) technique which permits the study of almost 1920 different phenotypes at one time during a single incubation. Twenty 96-well microplates (PM01 through PM20) represent the whole PM set for bacterial organisms, comprising 2 plates of sole carbon sources (PM01 – PM02), 4 plates of sole nitrogen sources (PM03, PM06 – PM08), a plate combined of sulphur and phosphorus sole sources (PM04), a plate testing biosynthetic pathways (PM05), a plate of ions and osmolytes (PM09), a plate testing pH effects (PM10) and 10 plates of toxins and antibiotics (PM11 through PM20) (Bochner, 2009).

To increase sensitivity and expand the applications of the assay, the PMs measure cell respiration instead of biomass amount (Bochner, 2009). If the substrate in a well is utilized, the organism can survive and respire, even if the bacteria may not proliferate. During respiration, the production of NADH molecules is elevated and NADH molecules provide electrons throughout the electron transport chain. On the membrane interface, electrons irreversibly reduce (colorless) tetrazolium violet, which has been added to the inoculum, to insoluble purple formazan, enabling a colorimetric assay (Bochner & Savageau, 1977; Tachon *et al.*, 2009).

Within the incubation chamber of the OmniLog® instrument, the PM colorimetric reporter system (Sturino *et al.*, 2010) allows capture and recording of color quantities every 15 minutes from up to 48 parallel running plates. The colorimetric data can be displayed in the form of kinetic graphs and analyzed by the OmniLog® software package. Different package tools also enable to compare kinetic data between several replicates (File Management/Kinetic software) and compare averaged replicates between two tested samples (Parametric software). Although more sophisticated statistical approaches are possible (Sturino *et al.*, 2010), several important studies have been analyzed by the provided software (Chaudhuri *et al.*, 2010; Ihssen & Egli, 2005; Wang *et al.*, 2010; Xue *et al.*, 2011).

A global cellular phenotype (phenomic) profile can reveal some important information and improve our knowledge about the biology of the studied microorganisms. High-throughput phenomic techniques can be used to empirically investigate metabolic pathways that are predicted from whole genome annotation (Chaudhuri *et al.*, 2010; Lim *et al.*, 2010). They can be used to discover new gene or even pseudogene functions (Soo *et al.*, 2011; Wang

*et al.*, 2010), reveal novel ways of regulating bacterial gene expression (Ihssen & Egli, 2005), or even study microbial communities (Pruss *et al.*, 2010; van Heerden *et al.*, 2002).

# 2. THE AIMS OF THIS THESIS

- To determine the complete genome sequence of *T. pallidum* ssp. *pertenue* CDC-2 strain
- To determine the complete genome sequence of *T. pallidum* ssp. *pertenue* Gauthier strain - in cooperation with Dr. Michal Strouhal and Marie Zobaníková
- To perform *de novo* gene annotation on the *T. pallidum* ssp. *pertenue* Samoa D genome and to apply novel gene prediction to the genomes of on *T. pallidum* ssp. *pertenue* CDC-2 and Gauthier, and *T. pallidum* ssp. *pallidum* DAL-1
- To assess intra- and inter-strain variability within and between *T. pallidum* ssp. *pertenue* CDC-2, Gauthier and Samoa D strains
- To analyse *rrn* operons in pathogenic treponemes with regard to reciprocal translocation
- To compare metabolic profiles of different *T. pallidum* ssp. *pallidum* BAC library clones

# 3. MATERIALS AND METHODS

**Bacterial strains.** In total, 20 strains of the *Treponema* genus (Table 1) used in this study comprise a baboon isolate (unclassified *T. pallidum* strain Fribourg-Blanc), a rabbit syphilis strain (*T. paraluiscuniculi*, TPc) and 18 human strains (including 11 strains of *T. pallidum* ssp. *pallidum*, TPA; 5 of *T. pallidum* ssp. *pertenue*, TPE; and 2 of *T. pallidum* ssp. *endemicum*, TEN).

For phenotype profile screening, the BAC library (containing 19 clones) of TPA Nichols DNA (Šmajs *et al.*, 2002) was examined. The host strain (*E. coli* DH10B) and strain carrying an empty vector (*E. coli* DH10B pBeloBAC11) were also used in comparative phenomic experiment.

**Culture media.** Rich broth LB medium consisting of yeast extract (Hi-Media, Mumbai, India) 5 $gl^{-1}$, tryptone (Hi-Media) 10 $gl^{-1}$, and sodium chloride 10 $gl^{-1}$ in water was used throughout the study. For selection and maintenance of plasmids, 100 μg of ampicillin per ml of liquid medium or per ml of 1.5% (w/v) LB agar, was added. If not indicated, bacteria were grown at 37°C overnight.

**Isolation of treponemal DNA.** TPA Nichols and SS14, TPE Samoa D, CDC-2 and Gauthier, and TPc Cuniculi A chromosomal DNAs were prepared as previously described by Fraser *et al.* (1998), by extracting DNA from experimentally infected rabbits. Treponemes were purified by Hypaque gradient centrifugation (Baseman *et al.*, 1974). Because high input of DNA was required for the sequencing approach, whole genome amplification (WGA, REPLI-g Midi Kit, Qiagen, Hilden, Germany) was performed for TPA Nichols, TPE CDC-2 and Gauthier DNA according to the manufacturer's instructions. In addition, non-WGA DNAs from TPA Nichols and SS14, TPE Samoa D and CDC-2 and TPc Cuniculi A were used. The Philadelphia 1, Philadelphia 2, DAL-1, Mexico A, Bal 73-1, Grady, MN-3, Madras, Haiti B (TPA), CDC-1, CDC-2, Gauthier and Samoa F (TPE), Bosnia A and Iraq B (TEN), and Fribourg-Blanc (a simian *T. pallidum*) strains were obtained as rabbit testicular tissues containing treponemal cells. After the samples were briefly centrifuged at 100 g for 5 min, the supernatant containing treponemal DNA was amplified using the REPLI-g kit.

Treponemal strains and DNAs were kindly provided by Dr. David Cox (Centers for Disease Control and Prevention, Atlanta, GA, USA), Dr. Steven J. Norris (The University of

**Table 1.** Treponemal strains used in this study.

| Strain name | *Treponema* (sub)species | Place of isolation | Date of isolation | Reference | The source of the material |
|---|---|---|---|---|---|
| Bal 73-1 | TPA | Baltimore, USA | 1968 | (Hardy *et al.*, 1970) | David L. Cox |
| Bosnia A | TEN | Bosnia | 1950 | (Turner & Hollander, 1957) | Sylvia M. Bruisten |
| CDC-1 | TPE | Dersuso, Ghana | 1980 | (Liska *et al.*, 1982) | David L. Cox |
| CDC-2 | TPE | Akorabo, Ghana | 1980 | (Liska *et al.*, 1982) | Steven J. Norris |
| Cuniculi A | *paraluiscuniculi* | ? | pre-1957 | (Turner & Hollander, 1957) | Steven J. Norris |
| DAL-1 | TPA | Dallas, USA | 1991 | (Wendel *et al.*, 1991) | David L. Cox |
| Fribourg-Blanc | simian isolate | Guinea | 1966 | (Fribourg-Blanc & Mollaret, 1969) | David L. Cox |
| Gauthier | TPE | Brazzaville, Congo | 1960 | (Gastinel *et al.*, 1963) | Steven J. Norris |
| Grady | TPA | Atlanta, USA | 1980s | ? | David L. Cox |
| Haiti B | TPA | Côtes-de-Fer, Haiti | 1951 | (Turner & Hollander, 1957) | David L. Cox |
| Iraq B | TEN | Iraq | 1951 | (Turner & Hollander, 1957) | Kristin N. Harper |
| Madras | TPA | Madras, India | 1954 | Laboratory notebook of Rob George, CDC | David L. Cox |
| Mexico A | TPA | Mexico City, Mexico | 1953 | (Turner & Hollander, 1957) | David L. Cox |
| MN-3 | TPA | Minnesota, USA | ? | ? | David L. Cox |
| Nichols | TPA | Washington, D.C., USA | 1912 | (Nichols & Hough, 1913) | Steven J. Norris |
| Philadelphia 1 | TPA | Philadelphia, USA | 1988 | ? | David L. Cox |
| Philadelphia 2 | TPA | Philadelphia, USA | ? | ? | David L. Cox |
| Samoa D | TPE | Apia, Samoa | 1953 | (Turner & Hollander, 1957) | Steven J. Norris |
| Samoa F | TPE | Apia, Samoa | 1953 | (Turner & Hollander, 1957) | Steven J. Norris |
| SS14 | TPA | Atlanta, USA | 1977 | (Stamm *et al.*, 1983) | Steven J. Norris |

?, indicates missing data

Texas Health Science Center at Houston, Houston, TX, USA), Dr. Kristin N. Harper (Emory University, Atlanta, GA, USA) and Dr. Sylvia M. Bruisten (Public Health Laboratory, Amsterdam, NL). I also thank other members of Laboratory of bacterial genetics and

genomics (Department of Biology, Faculty of Medicine, Masaryk University) for genomic DNA extracted from rabbit testicular tissues.

**DNA sequencing and assembly of the TPE CDC-2 genome**. Whole genome DNA sequencing was performed by combined approach using 454 pyrosequencing, Illumina approach and dideoxy-terminator (DDT) sequencing technologies. The amplified genomic DNA was sequenced by the 454 technique using the GS20 apparatus (454 Life Sciences Corporation, Branford, CT, USA) to average depth coverage 36x. 454 reads with an average length of 101 bp were assembled by Newbler assembler into 74 contigs covering 98.6% of the reference TPA Nichols genome (AE000520.1). Remaining gaps were closed by DDT approach. To improve accuracy of the complete sequence and increase coverage, an Illumina (Illumina, San Diego, CA, USA) sequencing technology was also applied using the Genome Analyzer (GA) sequencing apparatus. Reads of 36 bp (depth coverage 74x) were assembled by Velvet assembler (Zerbino & Birney, 2008) into 647 contigs. All ambiguous positions between 454 and Illumina contigs were resolved by DDT method. 454 and Illumina sequencing steps and assemblies were performed at Human Genome Sequencing Center (HGSC, Baylor College of Medicine, Houston, TX, USA).

**DNA sequencing and assembly of the TPE Gauthier genome.** Again, the amplified genomic DNA of TPE Gauthier was sequenced by the 454 (26x coverage, 101 bp average read length, 1508 contigs, 92.6% coverage of AE000520.1 genome), and Illumina methods (65x coverage, 36 bp average read length, 902 contigs). Both steps were performed at HGSC on GS20 and GA platforms.

Due to substantial contamination of rabbit DNA (data not shown) and too many gaps in the final genome assembly, chromosomal Gauthier DNA was amplified in 134 overlapping *Treponema pallidum* intervals (TPI) (Strouhal *et al.*, 2007) by GeneAmp® XL PCR kit (Applied Biosystems, Forster City, CA, USA). To avoid misassembly of sequentially related genes, equimolar XL PCR products were combined into four pools sequenced as four different samples on a 454 platform. 454 data were as follows: coverage 54x, average read length 233 bp, 87 contigs assembled by Newbler, 98.5% length coverage of TPA Nichols. Only one pool of all 134 XL PCR products was used for Illumina sequencing. Illumina pair-end data represented 35-bp reads, coverage 206x, assembled in 183 contigs by Velvet. All gaps in the complete sequence were filled in using the DDT approach. My colleague, Dr. Michal Strouhal prepared the XL PCR products and pools for sequencing. More detailed

information about this procedure can be found in his Doctoral thesis (Strouhal, 2010). 454 and Illumina sequencing techniques were performed at The Genome Institute (TGI, Washington University in St. Louis, Saint Louis, MO, USA) on GS FLX Titanium and GAIIx platforms, respectively.

**Gap Closure using DDT sequencing.** In order to close draft genome sequences, the LASERGENE program package (DNASTAR, Madison, WI, USA) was used to align 454 and/or Illumina contigs to reference genomes. This strategy could be applied because it is known that gene order in pathogenic treponemes is syntenic with the exception of paralogous regions (Mikalová *et al.*, 2010).

Regions containing gaps betweeen contigs were extended outwards in both direction to design primers using the Primer3 software (Rozen & Skaletsky, 2000). The regions were amplified using *Taq* DNA polymerase (New England BioLabs, Franfurkt am Main, Germany) or GeneAmp® XL PCR Kit according to the manufacturer's instructions. An extra-large PCR (XL PCR) comprised two reagent solutions: the lower mix was composed of 6.6 μL of PCR grade water, 6.0 μL of 3.3X XL Buffer II, 4.0 μL of 10 nM dNTP Blend, 0.5 μl of primer F (100 nmol/L), 0.5 μL of primer R (100 nmol/L), 2.4 μL $Mg(OAc)_2$ Solution; while the upper mix consisted of 19.0 μL of PCR grade water, 9.0 μL of 3.3X XL Buffer II, 1.0 μL of r*Tth* DNA Polymerase XL and 1.0 μL of DNA template. The XL PCR reaction was performed in GeneAmp PCR system 9700 (Applied Biosystems, Foster City, CA, USA). Lower mix was initially denaturated at 80 °C (60 s). After the upper solution was added, the amplification steps followed by 16 cycles (94 °C for 15 s; 65 °C for 10 min), then continued by additional 12 cycles (94 °C for 15 s; 67 °C for 10 min with an increment of 15s for each additional cycle) leading to final extension step at 67 °C for 10 min.

The resulting PCR (or XL PCR) products were purified by the QIAquick PCR Purification Kit (QIAGEN, Valencia, CA, USA) or ExoSAP-IT kit (GE Healthcare, Chalfont St. Giles, UK) and undergone DDT sequencing with the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA). The sequencing was performed using the original amplification primers or alternatively, additional internal oligonucleotide sequencing primers were designed by the Primer3 software. Throughout the thesis, DDT sequencing with internal primers is considered as primer walking method. All DDT sequencing reads were assembled and analyzed in LASERGENE software.

Found nucleotide changes between 454 and Illumina data were examined using the same steps for PCR, purification and DDT sequencing. Moreover, the same procedure was used to confirm all differences between TPE CDC-2 and TPE Samoa D genomes.

**DDT sequencing of paralogous regions in the TPE CDC-2 genome**. To prevent misassembly of paralogous sequences, the corresponding regions were XL PCR amplified. Paralogous loci included rDNA loci and several TPI regions in the vicinity of *tpr* (*Treponema pallidum* repeat) genes including TPI-12 (containing *tpr*D gene), TPI-25A (*tpr*E), TPI-25B (*tpr*F, *tpr*G), TPI-48 (*tpr*I, *tpr*J), and *tpr*K and *tpr*L genes (Mikalová *et al.*, 2010).

Whereas rDNA loci were directly sequenced from XL PCR products using primer walking approach, loci containing *tpr* genes were cloned into pCR®2.1 TOPO® vector (TOPO® TA Cloning® Kit, Invitrogen, Carlsbad, CA, USA). Usually, a colony from each cloned locus was sequenced using universal primers provided in the kit. If needed, additional internal primers were designed for primer walking approach. However, due to high heterogeneity observed in the *tpr*K gene (Centurion-Lara *et al.*, 2000a), 24 colonies were examined.

To increase coverage and investigate intra-strain heterogeneity, small insert libraries (SMIL) from TPI-11, TPI-12, TPI-25A, TPI-25B, TPI-48 and *tpr*L gene loci were prepared and sequenced as described previously (Andersson *et al.*, 1996; Matějková *et al.*, 2008). In brief, purified XL PCR products were mechanically sheared to length fragments of 1000-1500 bp using hydrodynamic forces in the Hydroshear instrument (Genomic Solution, Holliston, MA, USA). The fragments were blunt-ended, followed by ligation of samples to facilitate cloning into pUC18 vector. Transformation into competent cells under heat shock stress resulted in SMIL construction. For each TPI region, 96 recombinant plasmids were sequenced by pUC universal primers (pUC-545 and pUC-780).

**DDT sequencing of regions containing repetitive sequences.** Repetitive sequences of 60 and 24 bp are present in the *arp* (TPI-32B) and TP0470 (TPI-34) genes, respectively. The genes were XL PCR amplified, cloned into pCR®2.1 TOPO® vectors and sequenced by universal plasmid primers.

**DDT sequencing of regions carrying homopolymeric tracts.** If high quality Illumina data collapsed in the vicinity of tracts composed of nucleotide homopolymers, the surrounding regions were amplified and XL PCR products were used as templates for DDT

sequencing as described above. When the sequencing results did not resolve the consensus sequence, the XL PCR product was cloned into pCR®2.1 TOPO® vector. Trying to estimate the consensus sequences (and decrease the influence of population heterogeneity or DNA polymerase errors), the plasmid DNA was amplified directly from a colony using TempliPhi Amplification kit (Amersham Biosciences, Piscataway, NJ, USA) which contains a Φ29 DNA polymerase of high fidelity. Thus, 4 to 8 colonies from each problematic region were amplified and DDT sequenced.

**Detection of intra-strain heterogeneity within TPE genomes.** Next-generation data were used to estimate intra-strain heterogeneity within TPE Samoa D, CDC-2 and Gauthier strains. Initially, individual strain specific Illumina reads were aligned to a final genome using Burrows-Wheeler Aligner (Li & Durbin, 2009). Then, SAMtools package (Li *et al.*, 2009) was applied to determinate nucleotide changes, e.g. intra-strain heterogeneity, using the default settings.

Moreover, TPE CDC-2 genome heterogeneity was estimated based on sequencing results from variable SMIL clones and clones harboring the *tpr*K gene.

**Gene identification and annotation.** A semiautomatic bacterial genome annotation was designed at HGSC and applied for TPE Samoa D strain gene identification. FgenesB (http://linux1.softberry.com), GeneMark (Lukashin & Borodovsky, 1998) and Glimmer (Delcher *et al.*, 1999) software were used independently for prediction of open reading frames (ORFs) for the TPE Samoa D genome. Visualization of gene prediction was performed using the Genboree system (http://www.genboree.org) and the CONAN database (Highlander *et al.*, 2007; McLeod *et al.*, 2004). To predict genes coding for tRNA, rRNA and non-coding RNAs, tRNAscan (Lowe & Eddy, 1997), RNAmmer (Lagesen *et al.*, 2007) and Rfam (Gardner *et al.*, 2009) were used. This automatic process was conducted at HGSC.

For each predicted ORF, DNA comparisons were performed using BLASTN and BLASTX algorithms, and protein sequences were analyzed by BLASTP versus the non-redundant (nr) database at NCBI (Sayers *et al.*, 2009). When appropriate, other predictive tools were used as described previously (Gioia *et al.*, 2007). Briefly, conserved domains were predicted by Conserved Domain Database (Sayers *et al.*, 2009) and InterProScan (Zdobnov & Apweiler, 2001). Proposed enzymes were subjected to the ExPASy ENZYME (Gasteiger *et al.*, 2003) analysis. Predicted peptidases and transporters were analyzed by MEROPs (Rawlings *et al.*, 2008) peptidase database and Transporter Classification Database (Saier *et*

*al.*, 2006), respectively. Protein putative localization within a cell was determined using PSORTb (Gardy *et al.*, 2005), SignalP (Bendtsen *et al.*, 2004) and LipoP (Juncker *et al.*, 2003) software.

For proteins of unknown functions, the gene minimal size limit of 150 bp was set. For the TPE CDC-2 and Gauthier and TPA DAL-1 genome annotations, predicted gene coordinates from Samoa D genome were adapted and recalculated using Consed finishing package (Green, 1993). If needed (e.g. in ORFs carrying a frameshift or a non-sense mutation), genes were newly manually predicted using the same tools as in semiautomatic approach. In most cases, the original locus tag numbering of annotated TPA Nichols genes was preserved in TPA DAL-1 and TPE orthologs.

Newly predicted ORFs were named based on their proximity to their preceding TPA Nichols gene with a letter suffix (e.g. TPESAMD_0001a is a newly annotated ORF located downstream of TPESAMD_0001 which is an orthologue of TP0001 gene).

**Submission into GenBank database.** The final gene annotation spreadsheet for each genome was converted into .tbl format using the awk programming language. The instructions of Bacterial Genome Submission Guide (provided by NCBI) were followed to create an .sqn file, a genome annotation format suitable for direct submission.

The genomes of TPE Samoa D, CDC-2, Gauthier, and TPA DAL-1 have been deposited in the GenBank under the accession numbers CP002374, CP002375, CP002376 and CP003115, respectively.

**Comparative genome analysis.** In order to define variability among TPE Samoa D, CDC-2 and Gauthier strains at the DNA level, two different methods were used to compare whole genome sequences. The whole genome alignment of all genomes was performed by SeqMan software from LASERGENE program package with manual corrections. In addition, a pairwise comparative software Cross-match (Consed program package) was applied to investigate differences in all three possible genome pairs (Samoa D – CDC-2; Gauthier – CDC-2; Samoa D – Gauthier).

**Detection of recombination events between genomes.** Rrecombinant events must have been tested prior the estimation of selection type. A whole genome alignment of TPE strains was analyzed by the Recombination Detection Program package (version RDP3) (Martin *et al.*, 2010). Four methods, including RDP, GENECONV (Sawyer, 1989), Maximum

Chi-squared (Smith, 1992) and Chimaera (Posada & Crandall, 2001), implemented in the RDP3 package, were applied using non-default settings including maximum p-value of 0.01 and a Bonferroni correction. For the RDP method, a window size of 10 nt was used. For the MaxChi method, the number of variable sites per window was set to 30. Recombinant regions predicted by at least three methods were considered significant.

**Determination of genes under selection among TPE strains.** To determine genes under positive selection, TPE orthologues coding for protein with > 1 amino acid replacements have been chosen. The number of synonymous substitutions per a synonymous site (Ks), the number of nonsynonymous substitutions per a nonsynonymous site (Ka), and the codon-based test for estimation of selection type was calculated using the Kumar model (Nei & Kumar, 2000) in the MEGA4 software (Tamura *et al.*, 2007).

A highly variable *tpr*K gene, genes containing frameshift mutation and genes prone to recombination events were omitted from this analysis.

**PCR amplification and DNA sequencing of *rrn* operons.** Primers RNA1F (5'-GTGTGTGAGTCTGGCAGGAA-3') and RNA1R (5'-TTATTGCTGTGCGCATCTTC-3'); and RNA2F (5'-ACAAGTGAGCGAAGCGTTTT-3') and RNA2R (5'-CCAAGAGAGCTACCCGTCTG-3') were used for the amplification of *rrn* operons from treponemal strains; these primer-pairs produced XL-PCR products of 5.85 and 5.92 kb, respectively.

For more details regarding amplification, purification, DNA sequencing and assembly approaches please see the above sections.

**Phylogenetic analysis of *rrn* operons.** In addition to the *rrn* operons investigated in 20 strains (Table 1), the *rrn* operons of the TPA Chicago (CP001752.1) (Giacani *et al.*, 2010a) were included in the evolutionary analysis. Concatenated sequences of *rrn*1 and *rrn*2 operons (Table S1) were used for the construction of evolutionary trees using neighbor-joining method (Saitou & Nei, 1987) in MEGA4 software. The bootstrap consensus trees were determined from 1000 bootstrap re-samplings. Branches with less than 50% bootstrap support were collapsed.

**Nucleotide sequence accession numbers.** The nucleotide sequences of *rrn* operons reported in this study have been deposited in GenBank under the accession numbers JX120527-JX120565.

**Comparative phenomics applied on BAC library of TPA Nichols.** BAC library clones carrying TPA Nichols DNA were screened for phenotype profile using Phenotype MicroArray plates (BiOLOG, Hayward, CA, USA). The experiments were performed as described by Bochner *et al.* (2001). Initially, individual clones from BAC library, including negative control with an empty vector were pre-grown overnight at 37°C on LB agar plates and at least once again re-streaked onto a sterile plate. Thereafter, colonies were transferred into BiOLOG inoculating fluid to compile a suspension of 85% transmittance. The suspensions were added into BiOLOG defined minimal medium (carbon-free) and nutrient-rich medium enriched with inhibitors and other chemicals to test for carbon sources (PM01 and PM02 plates), sensitivity (PM09 and PM10 plates) and resistance (PM11 through PM20 plates). Altogether, each BAC library clone was tested for 1342 phenotypes. The cell-inoculated suspensions were added to appropriate phenotype microarray plates and grown at 33°C for 24 hours using OmniLog instrument. During the growth, OmniLog reader measured the amount of color production (and thus corresponding respiration) and recorded data every 15 minutes.

Every library clone (in total 19 clones) and negative control (*E. coli* DH10B pBeloBAC11) were examined in four replicates. Strain-specific data from all replicates were analyzed using File Management/Kinetic module (OL_PM_FM/Kin, BiOLOG) which calculates the mean signal at any measured time point. The obtained averaged data for each strain were further analyzed via parametric tests. Parametric tests between negative control and an examined BAC library clone were performed using Parametric software (OL_PM_Par, BiOLOG).

Moreover, the host strain (*E. coli* DH10B) was examined in the same way and phenotype profile was compared to a strain served as negative control (*E. coli* DH10B pBeloBAC11).

# 4. RESULTS

## 4.1 Whole genome sequencing of the *Treponema pallidum* ssp. *pertenue* CDC-2 genome

The whole genome sequencing project of *Treponema pallidum* ssp. *pertenue* (TPE) CDC-2 started on the 454 pyrosequencing platform GS20, resulting in 165 contigs (N50 of 52 kb). The sequencing process including assembly was performed at the Human Genome Sequencing Center (HGSC, Baylor College of Medicine, Houston, TX, USA).

Since the chromosome of pathogenic treponemes preserved gene organization and content (Mikalová *et al.*, 2010; Strouhal *et al.*, 2007), the assembled contigs were aligned to the reference genome, *Treponema pallidum* ssp. *pallidum* (TPA) Nichols (AE000520.1), using SeqMan assembly software (from LASERGENE suite). Out of total 165 contigs, only 74 ones hit the reference genome.

To determine the whole genome sequence of the TPE CDC-2 strain (composed of a circular chromosome), a total number of 41 gaps needed to be closed by DDT sequencing. In addition, two genes with variable number of tandem repeats in different strains, orthologues to *arp* and TP0470 genes, were also chosen for DDT sequencing (Fraser *et al.*, 1998; Harper *et al.*, 2008a; Liu *et al.*, 2007; Mikalová *et al.*, 2010; Pillay *et al.*, 1998). A combination of several methods (Table 2) was used to accomplish this project. The total size of the missing genome sequence was estimated to be approximately 20 kb (1.8% of TPA Nichols genome

To ascertain sequences within each operon, *rrn* operons and their flanking regions were amplified using XL PCR and DDT sequenced by primer walking technique. To our surprise, tRNA-Ala gene was located in *rrn*1 and tRNA-Ile in *rrn*2 operon in TPE CDC-2 genome whereas other treponemal genomes sequenced at that time (Fraser *et al.*, 1998; Matějková *et al.*, 2008; Šmajs *et al.*, 2011) revealed tRNA-Ile located in *rrn*1 and tRNA-Ala in *rrn*2 operon.

In the TPE CDC-2 genome, four tandem repeats in the *arp* gene were found which is in concordance with the study of Liu *et al.* (2007). The number of tandem repeats in the TP0470 gene was not revealed directly from DDT sequencing. Current DDT sequencing enables to sequence up to 800 bp, thus region containing almost three dozens of repeats of 24 bp is beyond the limit. Nevertheless, DNA sequence was determined from both ends of a cloned XL PCR product. The exact number of repeats, 37, was estimated on the agarose gel.

**Table 2.** Regions in TPE CDC-2 genome chosen for DDT sequencing based on 454 pyrosequencing results.

| Coordinates in TPA Nichols genome (AE000520.1) | Locus tag (AE000520.1)/ Intergenic region (IGR) | Reason for finishing | Resolved by | Note |
|---|---|---|---|---|
| 8798 | TP0009 | gap (0 bp) | DDT sequencing from PCR product | |
| 9272 | TP0009 | gap (0 bp) | DDT sequencing from PCR product | |
| 24481-24484 | IGR TP0021-0022 | gap (4 bp) | DDT sequencing from PCR product | |
| 127346 | TP0111 | gap (0 bp) | DDT sequencing from PCR product | |
| 136259-136867 | TP0116-0118 | gap (607 bp) / paralogous region | DDT sequencing (primer walking approach) from PCR product | TPI-11[a] |
| 148481-148529 | TP0126-0127 | gap / paralogous region | cloning and DDT sequencing (primer walking approach) | TPI-12[b] |
| 151226-153011 | TP0130-0132 | gap (1784 bp) / paralogous region | DDT sequencing (primer walking approach) from PCR product | TPI-12 |
| 154419-154931 | TP0134 | gap (13 bp) / paralogous region | DDT sequencing from PCR product | TPI-12 |
| 155309-155360 | TP0134 | gap (52 bp) / paralogous region | DDT sequencing from PCR product | TPI-12 |
| 157010-157017 | TP0136 | gap (8 bp) | DDT sequencing from PCR product | |
| 158251-158524 | TP0136-0138 | gap (274 bp) | DDT sequencing (primer walking approach) from PCR product | |
| 165860-165873 | TP0144-0145 | gap (14 bp) | DDT sequencing from PCR product | |
| 201751 | TP0185 | gap (0 bp) | DDT sequencing from PCR product | |
| 230514-230555 | TP0225-0226 | gap (42 bp) / paralogous region | DDT sequencing (primer walking approach) from XL PCR product | *rrn* operon[c] |
| 261220 | TP0248 | gap (0 bp) | DDT sequencing from PCR product | |
| 280201-283598 | TP0265-0267 | gap (3398 bp) / paralogous region | DDT sequencing (primer walking approach) from XL PCR product | *rrn* operon |
| 285823-285838 | TP0269 | gap (16 bp) | DDT sequencing from PCR product | |
| 327896-334776 | TP0312-0319 | gap (6881 bp) / paralogous region | SMIL | TPI-25[d] |
| 461207-461507 | TP0433-0434 | region with repeats | DDT sequencing from XL PCR product | |
| 497262-497711 | TP0470 | region with repeats | cloning and DDT sequencing (primer walking approach) | exact number of repeats was estimated from agarose gel |
| 513615-513625 | TP0483 | gap (11 bp) | DDT sequencing from PCR product | |
| 518922-518945 | TP0486 | gap (24 bp) | DDT sequencing from PCR product | |

| Coordinates in TPA Nichols genome (AE000520.1) | Locus tag (AE000520.1)/ Intergenic region (IGR) | Reason for finishing | Resolved by | Note |
|---|---|---|---|---|
| 624343-624379 | TP0575 | gap (37 bp) | DDT sequencing from PCR product | |
| 624566-624585 | TP0575 | gap (20 bp) | DDT sequencing from PCR product | |
| 629889 | TP0579 | gap (0 bp) | DDT sequencing from PCR product | |
| 659785-659787 | TP0609 | gap (3 bp) | DDT sequencing from PCR product | |
| 670217-675757 | TP0618-0622 | gap (5541 bp) / paralogous region | SMIL | TPI-48[e] |
| 675906-675927 | TP0622 | gap (20 bp) | DDT sequencing from PCR product | TPI-48 |
| 698042 | TP0639 | gap (0 bp) | DDT sequencing from PCR product | |
| 698179 | TP0639 | gap (0 bp) | DDT sequencing from PCR product | |
| 835434-835619 | TP0769 | gap (186 bp) | DDT sequencing from PCR product | |
| 858811 | TP0792 | gap (0 bp) | DDT sequencing from PCR product | |
| 859004 | IGR TP0792-0793 | gap (0 bp) | DDT sequencing from PCR product | |
| 871839-871840 | TP0803 | gap (2 bp) | DDT sequencing from PCR product | |
| 944713-944717 | TP0865 | gap (5 bp) | DDT sequencing from PCR product | |
| 974330 | IGR TP0985-0986 | gap (0 bp) | DDT sequencing from PCR product | |
| 974764-974829 | TP0897 | gap (66 bp) / paralogous region | cloning and DDT sequencing | *tpr*K gene[f] |
| 976298 | TP0898 | gap (0 bp) | DDT sequencing from PCR product | |
| 1047221 | TP0965 | gap (0 bp) | DDT sequencing from PCR product | |
| 1047475-1047476 | TP0965 | gap (2 bp) | DDT sequencing from PCR product | |
| 1060426-1060452 | TP0976 | gap (27 bp) | DDT sequencing from PCR product | |
| 1071433-1071614 | TP0986-0987 | gap (182 bp) | DDT sequencing from PCR product | |
| 1093549 | TP1005 | gap (0 bp) | DDT sequencing from PCR product | |

[a] Small insert library (SMIL) from XL PCR product amplifying 133144-145876 region (TPI-11) was constructed and sequenced.

[b] XL PCR product of 145858-155696 region (TPI-12) was cloned and sequenced by primer walking approach. In addition, SMIL was constructed and sequenced.

[c] XL PCR products from *rrn*1 (229575-235425) and *rrn*2 (278035-283955) regions were sequenced by primer walking technique.

[d] XL PCR products from 325002-330765 (TPI-25A) and 330289-335856 (TPI-25B) regions were cloned and sequenced by primer walking approach. In addition, SMILs were constructed and sequenced.

[e] XL PCR product of 667268-678727 region (TPI-48) was cloned and sequenced by primer walking approach. In addition, SMIL was constructed and sequenced.

[f] In total, 24 colonies from XL PCR product (973468-976794) were sequenced by primer walking technique.

An XL PCR product of the TP0470 homologue was compared to PCR products of known sizes.

The main disadvantage of the 454 technique was high sequencing error rate in the vicinity of homopolymers (Huse *et al.*, 2007) and the likelihood of an error increased with the length. To avoid such high error rate, an Illumina sequencing technology was also applied, resulting in an additional 74x depth coverage and 647 contigs. Illumina sequencing including assembly was performed at HGSC. Based on Illumina results, 20 insertions and 104 deletions were adopted in the CDC-2 genome. In addition to identified indels, only five ambiguous single nucleotide changes (single nucleotide polymorphism, SNP) were found between 454 and Illumina contigs (Table 3). Sequencing data corresponding to these five positions confirmed higher accuracy of Illumina results.

However, Illumina sequencing did not resolve all homopolymeric consensi due to gaps in additional 14 regions (Table 3). The results confirmed intra-strain variability regarding the numbers of homopolymers in a tract and thus median value from each region was chosen for the final complete genome of TPE CDC-2 strain.

In summary, complete genome of TPE CDC-2 was assembled from 454 and Illumina sequencing contigs. Following assembly, the genome was further improved by manual finishing. Different optimization steps of DDT sequencing had to be developed to close gaps, and to resolve heterogeneity in homopolymers and paralogous sequences.

To verify the final version of the CDC-2 genome, TPE CDC-2 and Samoa D (Čejková *et al.*, 2012) genomes were compared. The TPE Samoa D genome was independently sequenced by three next-generation techniques and finished by my colleague Marie Zobaníková. All nucleotide discrepancies were confirmed in the CDC-2 genome by DDT approach. In addition, whole genome fingerprinting (WGF) (Mikalová *et al.*, 2010; Strouhal *et al.*, 2007) method was applied on CDC-2 genome. Whole genome restriction fragments were compared *in silico* to final genome version of CDC-2 strain. This analysis was provided by Lenka Mikalová. Overall, only one resticton site (*Sph* I site, in the TPI21-A) was missing in the final version (Čejková *et al.*, 2012). When screening raw sequencing data, neither 454 nor Illumina reads showed this cleavage site. Shorter incubation time with *Sph* I enzyme did not result in cleavage of DNA in this site. It is likely that a non-specific cleavage occured within two-hour incubation. Thus sequencing error rate of $10^{-4}$ or lower was predicted based on experimentally confirmed restriction profile (Šmajs *et al.*, 2011).

**Table 3.** Regions in TPE CDC-2 genome demanding to be sequenced by DDT method after 454, gap closing and Illumina approaches were used.

| Coordinates in TPA Nichols genome (AE000520.1) | Locus tag (AE000520.1) | Reason for finishing | Resolved by |
|---|---|---|---|
| 38259 | IGR[a] TP0029-0030 | SNP[b] between 454 and Illumina | DDT sequencing from PCR product |
| 49355-49634 | TP0040 | homopolymeric tract[c] | DDT sequencing from XL PCR product[d] |
| 94710-94718 | IGR TP0084-0085 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing[e] |
| 122219-122228 | IGR TP0107-0108 | homopolymeric tract | DDT sequencing from XL PCR product |
| 140941-140949 | IGR TP0121-0122 | homopolymeric tract | DDT sequencing from XL PCR product |
| 198527-198535 | TP0180 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 207435-207442 | IGR TP0192-0193 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 372050-372062 | TP0347 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 382017 | TP0358 | SNP between 454 and Illumina | DDT sequencing from PCR product |
| 382018 | TP0358 | SNP between 454 and Illumina | DDT sequencing from PCR product |
| 405215-405224 | IGR TP0379-0380 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 407928-407970 | TP0381 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 490873-490878 | IGR TP0460-0461 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 669837-669846 | TP0618 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 757758 | TP0688 | SNP between 454 and Illumina | DDT sequencing from PCR product |
| 757759 | TP0688 | SNP between 454 and Illumina | DDT sequencing from PCR product |
| 936399-936409 | TP0859 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 944081-944086 | TP0865 | homopolymeric tract | DDT sequencing from XL PCR product |
| 1053934-1053942 | TP0969 | homopolymeric tract | DDT sequencing from XL PCR product |

[a] Intergenic region

[b] Discrepancy / Single nucleotide polymorfism

[c] Illumina sequencing results collapsed in homopolymeric tracts and/or vicinity.

[d] The number of nucleotides in a tract was determined by DDT sequencing from XL PCR product.

[e] Up to 8 colonies were cloned and sequenced using TempliPhi Amplification kit and DDT sequencing.

The complete genome sequence of TPE CDC-2 strain consists of 1,139,744 bp with GC content of 52.8%. The genome was annotated and compared with other TPE complete genomes. More details will be provided below.

For more details see Attachment-2.doc.

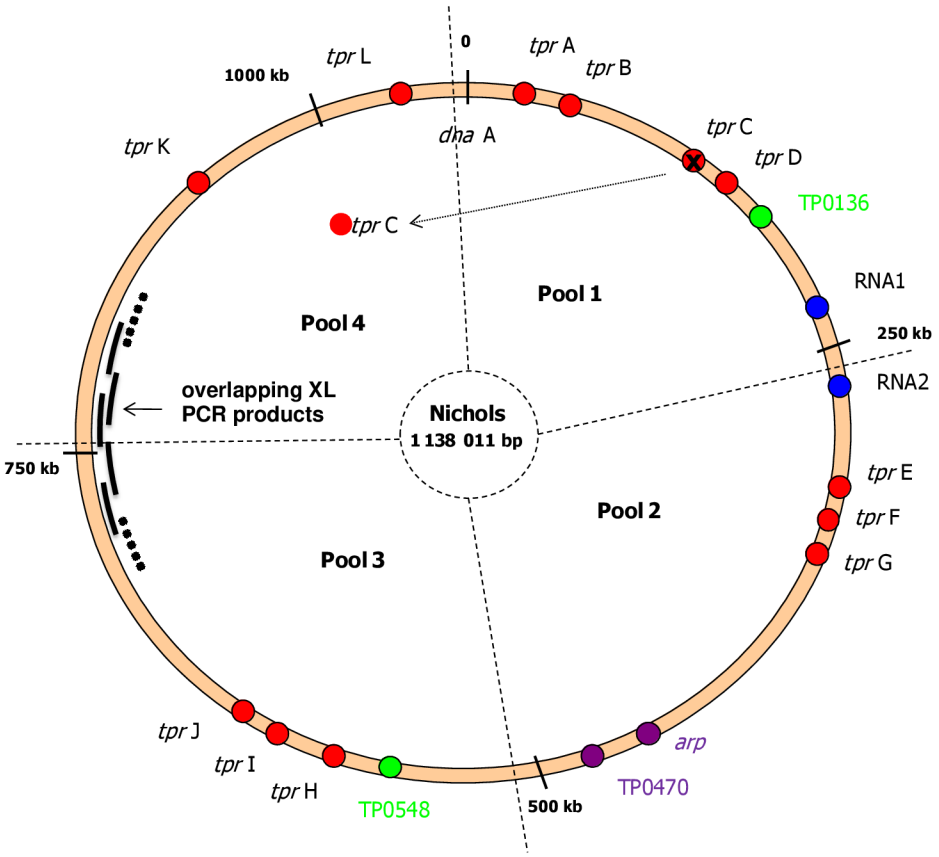
## 4.2 Whole genome sequencing of the *Treponema pallidum* ssp. *pertenue* Gauthier genome

The first attempt to determine whole genome sequence of the TPE Gauthier strain started at the HGSC by 454 pyrosequencing, followed by Illumina sequencing. Both methods resulted in a high depth coverage; however, high contamination with rabbit DNA was observed. Moreover, even if results were combined from both sequencing runs, almost 400 contigs were scattered throughout the genome. Closing gaps and finishing the complete genome sequence would have been too laborious and expensive.

The ongoing whole genome fingerprinting project (Mikalová *et al.*, 2010) suggested an idea of pooled segment genome sequencing (PSGS) approach applied on TPE Gauthier genome. Samples were prepared by Dr. Michal Strouhal and the methodology was extensively described in his Doctoral Thesis (Strouhal, 2010). Briefly, genomic DNA was amplified in 134 overlapping XL PCR products (Figure 1). Four pools (Pool 1 - Pool 4) of equimolar XL PCR products were combined and labeled with 4 different adapters (multiplex identifiers, MID) to ascertain the precise assembly of paralogous regions. Paralogous regions in treponemal genomes comprise *rrn* operons (marked RNA1 and RNA2 in Figure 1) and regions in the vicinity of *tpr* genes (marked *tpr*A through *tpr*L). In order to separate *tpr*C and *tpr*D sequencing data from each other, XL PCR product of TPI-11 containing *tpr*C gene was merged with Pool 4. Both genes belong to highly heterogenous Subfamily I (Centurion-Lara *et al.*, 2000b; Gray *et al.*, 2006; Sun *et al.*, 2004) and are localized in close proximity in treponemal genomes. XL PCR products of chromosomal Gauthier DNA were sequenced as four pools by 454 and as one pool by Illumina sequencing at The Genome Institute (TGI, Washington University in St. Louis, St. Louis, MO, USA). While Illumina contigs (183) were assembled at TGI, the author of this Doctoral Thesis analyzed 454 data.

Altogether, 454 sequencing reads from Pool 1 were assembled into 6 contigs (59x depth coverage), from Pool 2 into 8 contigs (63x), from Pool 3 into 11 contigs (67x) and from Pool 4 into 19 contigs (22x). In total, contigs from all Pools corresponded to 1,120,447 bp of the TPE Gauthier genome. Overall, 95.15% 454 reads were used for assembly.

**Figure 1.** Preparation of TPE Gauthier DNA for pooled segment genome sequencing. Figure is reprinted with courtesy of Dr. Michal Strouhal (Strouhal, 2010)




Based on the experience with TPE Samoa D and TPE CDC-2 finishing processes, several TPI regions were directly XL PCR amplified, cloned and DDT sequenced, including TPI-11, TPI-12, TPI-25A, TPI-25B, and TPI-48, and tandem repeats in the *arp* gene and TP0470 orhologue gene. The TP0136 orthologue locus showed a gap in both 454 and Illumina sequence. In addition, the *tpr*K gene was also XL PCR amplified and cloned, and 12 colonies were chosen for DDT sequencing. The number of tandem repeats (60-bp repeats) in the the *arp* gene was not revealed directly from DDT sequencing. Similar to the number of tandem repeats estimated in the TP0470 gene orthologue in TPE CDC-2 genome, the exact number of repeats within the TPE Gauthier *arp* gene, 26, was estimated based on comparison of the PCR products of known sizes on the agarose gel.

A combination of 454, Illumina and the above mentioned DDT sequencing covered a total of 1,139,296 bp. Several remaining gaps and problematic regions needed to be resolved (Table 4) via additional DDT sequencing.

**Table 4.** Regions in TPE Gauthier genome necessary to be sequenced by DDT method after 454 and Illumina sequencing approaches were used.

| Coordinates in TPA Nichols genome (AE000520.1) | Locus tag (AE000520.1) | Reason for finishing | Resolved by |
|---|---|---|---|
| 12476-12485 | TP0012 | homopolymeric tract[a] | cloning, TempliPhi kit and DDT sequencing[b] |
| 49355-49364 | TP0040 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 122219-122228 | IGR[c] TP0107-0108 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 140941-140949 | IGR TP0121-0122 | homopolymeric tract | DDT sequencing from XL PCR product |
| 170847-170863 | TP0147 | homopolymeric tract | DDT sequencing from XL PCR product |
| 198527-198535 | TP0180 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 294249-294258 | TP0279 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 352316-352324 | TP0330 | gap (9 bp) | DDT sequencing from PCR product |
| 372050-372062 | TP0347 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 372788-372796 | TP0348 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 373925-373942 | TP0348,TP0349 | homopolymeric tract and partial repeat | DDT sequencing from XL PCR product |
| 407928-407970 | TP0381 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 452000-452004 | TP0424 | gap (5 bp) | DDT sequencing from PCR product |
| 509476-509484 | TP0479 | homopolymeric tract | DDT sequencing from XL PCR product |
| 532939-532948 | TP0498 | homopolymeric tract | DDT sequencing from XL PCR product |
| 757632-757641 | TP0690 | homopolymeric tract | DDT sequencing from XL PCR product |
| 835459-835525 | TP0769 | gap (67 bp) | DDT sequencing from PCR product |
| 934768-934785 | TP0858 | gap (18 bp) | DDT sequencing from PCR product |
| 936399-936409 | TP0859 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 944081-944097 | TP0865 | homopolymeric tract | cloning, TempliPhi kit and DDT sequencing |
| 948434-948446 | TP0870 | gap (13 bp) | DDT sequencing from PCR product |
| 977778-977795 | TP0898 | gap (18 bp) | DDT sequencing from PCR product |
| 1050288-1050310 | TP0967 | partial repeat | DDT sequencing from XL PCR product |
| 1053934-1053942 | TP0969 | homopolymeric tract | DDT sequencing from XL PCR product |
| 1070550-1070557 | IGR TP0985-0986 | homopolymeric tract | DDT sequencing from XL PCR product |

[a] Illumina sequencing results collapsed in homopolymeric tracts and/or vicinity

[b] Up to 8 colonies were cloned and sequenced using TempliPhi Amplification kit and DDT sequencing.

[c] Intergenic region

The complete Gauthier genome was compared to contigs obtained from the first 454 and Illumina sequencing approaches. Genomic DNA used in these sequencing steps were not amplified by XL PCR, thus the comparison can reveal any nucleotide change introduced by XL PCR. The most common error during XL PCR can occur by adopting designed primer sequences into relevant genome positions. The contigs from non XL PCR amplified genomic DNA covered all positions corresponding to 134 primer-pairs designed for XL PCR. No discrepancy was found at these loci. In addition, only 7 discrepancies were found throughout the whole genome comparison, and all 7 discrepancies were found in homopolymeric tracts already examined in the finishing step.

A whole genome fingerprinting evaluation was also applied on TPE Gauthier genome. No discrepancy between restriction profile and *in silico* analysis was found. The sequencing error rate of $10^{-4}$ or lower was also presumed.

The complete genome sequence of TPE Gauthier strain consists of 1,139,417 bp with GC content of 52.8%. The genome was annotated and compared with other complete TPE genomes. More details will be provided below.

For more details see Attachment-2.doc.

### 4.3. Gene annotation of the TPE Samoa D genome

TPE Samoa D genome, completed by Marie Zobaníková, was annotated *de novo*. Since the time of the TPA Nichols genome annotation (Fraser *et al.*, 1998) the algorithms of predictive tools have been improved enormously. Moreover, dozens of sequencing errors were identified in the TPA Nichols sequence (AE000520.1) resulting in incorrect annotation as many as 10% genes (Giacani *et al.*, 2012a; Šmajs *et al.*, 2011).

A semi-automatic bacterial genome annotation was applied for the TPE Samoa D strain gene identification at HGSC. The automated process was involved in prediction of open reading frames (ORFs), tRNA and rRNA encoding genes and other non-coding RNAs. To further characterize ORFs, manual steps were involved in genome annotation. For each predicted ORF, DNA and protein sequences were compared to non-redundant (nr) database. Moreover, start codon, protein localization within a cell and conserved domains were predicted. Peptidases and transporters were compared with MEROPs peptidase database and Transporter Classification Database, respectively. The manual gene annotation of every ORF was provided by at least two researchers and ambiguous gene predictions were reconciled.

Whereas other members of the team annotated only part of the TPE Samoa D genome, the author of this Doctoral Thesis annotated all ORFs. Moreover, the author converted data into suitable formats and submitted gene annotation files into the GenBank.

TPESAMD locus tag was registered for the TPE Samoa D annotated genes. Genes were called using the locus tag, followed by an underscore and number (e.g. TPESAMD_0001). A minimal length limit of 150 bp was set for all ORFs of unknown function with no hit to other sequences. An overview of the Samoa D genome annotation is depicted in Attachment 2 (TPESAMD.tbl file).

In the TPE Samoa D genome, 1125 genes were annotated, composed of 1068 ORFs, 3 pseudogenes, 6 genes encoding rRNA, 45 genes encoding tRNA genes, and additional three genes coding for non-coding RNA. In total, 600 genes were coded on a leading strand while 525 on a lagging DNA strand. The average and median gene lengths were calculated to 980 bp and 831 bp, respectively. The intergenic regions (IGR) comprised 52.844 kb and represented 4.64% of the total genome length. Among predicted ORFs, there were 651 genes encoding proteins with predicted function (60.96% of ORFs), 141 genes encoding conserved hypothetical proteins (CHP, 13.20%), 129 genes encoding treponemal conserved hypothetical proteins (TCHP, 12.08%), and 147 genes encoding hypothetical proteins (HP, 13.76%). The group of genes with predicted function (with average/median gene length of 843/657 bp) was further subdivided into genes involved in cell processes (63 genes; 5.90% of ORFs), cell structure (62; 5.81%), DNA metabolism (51; 4.78%), general metabolism (160; 14.98%), regulation (35; 3.37%), transcription (16; 1.50%), translation (121; 11.33%), transport ( 112; 10.49%) and virulence (31; 2.90%).

During annotation steps, the DNA sequences of *de novo* predicted TPE Samoa D ORFs was compared to nr database. Usually, the highest hit belonged to TPA Nichols gene orthologues (AE000520.1). To our surprise, 50 TPA Nichols genes were fused into 24 TPE Samoa D orthologues (Table 5). While 22 fused Samoa D genes were composed of two putative adjacent Nichols genes, the additional two Samoa D fused genes consisted of three putative neighboring Nichols genes. With the exception of 2 gene fusions (TPESAMD_0314, fusion of TP0314 and TP0315; TPESAMD_0859, fusion of TP0859 and TP0860), all fusions were also present in the re-sequenced Nichols genome (Pětrošová, personal communication).

On the other hand, an orthologous locus to TP0127 carries a frameshift mutation in the TPE Samoa D genome, resulting into prediction of two novel ORFs, TPESAMD_0127a and TPESAMD_0127b.

**Table 5.** Gene fusions – differences found between TPE Samoa D and TPA Nichols (AE000520.1) genomes during annotation process.

| Annotated TPA Nichols genes (AE000520.1)[a] | Annotated TPE Samoa D genes (CP002374.1) | Gene function in Samoa D genome |
|---|---|---|
| TP0006, TP0007, TP0008 | TPESAMD_0006 | probable lipoprotein |
| TP0013, TP0014 | TPESAMD_0013 | probable lipoprotein |
| TP0018, TP0019 | TPESAMD_0018 | transcription elongation factor GreA |
| TP0127 | TPESAMD_0127a, TPESAMD_0127b | hypothetical proteins |
| TP0172, TP0173 | TPESAMD_0172 | GGDEF domain protein |
| TP0174, TP0175, TP0176 | TPESAMD_0174 | conserved hypothetical protein |
| TP0284, TP0285 | TPESAMD_0284 | Fe-S oxidoreductase domain protein |
| TP0286, TP0287 | TPESAMD_0286 | conserved hypothetical protein |
| TP0288, TP0289 | TPESAMD_0288 | cytidylyltransferase domain protein |
| TP0299, TP0300 | TPESAMD_0300 | sugar ABC superfamily ATP binding cassette transporter, ABC protein |
| TP0314, TP0315 | TPESAMD_0314 | hypothetical protein |
| TP0324, TP0325 | TPESAMD_0324 | hypothetical outer membrane protein |
| TP0377, TP0378 | TPESAMD_0377 | flagellar basal body-associated protein FliL |
| TP0419, TP0420 | TPESAMD_0419 | acid phosphatase |
| TP0433, TP0434 | TPESAMD_0433 | acidic repeat protein |
| TP0462, TP0463 | TPESAMD_0462 | conserved hypothetical protein |
| TP0468, TP0469 | TPESAMD_0468 | hypothetical protein |
| TP0481, TP0482 | TPESAMD_0481 | hypothetical protein |
| TP0587, TP0588 | TPESAMD_0587 | DNA-directed DNA polymerase III delta subunit |
| TP0597, TP0598 | TPESAMD_0597 | hypothetical protein |
| TP0702, TP0703 | TPESAMD_0702 | M23B subfamily peptidase |
| TP0781, TP0782 | TPESAMD_0781 | hypothetical protein |
| TP0859, TP0860 | TPESAMD_0859 | hypothetical protein |
| TP0899, TP0900 | TPESAMD_0899 | AddB protein |
| TP0928, TP0929 | TPESAMD_0928 | hypothetical protein |

[a] Re-sequencing of selected TPA Nichols genes (unpublished results) resulted in similar fusions (with the exceptions of TP0314, TP0315; and TP0859, TP0860) as in the TPE Samoa D genome.

Interestingly, the TPA Chicago strain, sequenced and annotated independently by another group (Giacani *et al.*, 2010a; Giacani *et al.*, 2012a), revealed the same fused genes as re-sequenced TPA Nichols strain, thus confirmed suggested sequencing errors in the Nichols first sequencing project (AE000520.1).

In addition, TPE Samoa D annotated ORFs were compared with genome annotation of TPA Nichols (AE000520.1). A set of 95 genes were newly predicted (Table S2) in the TPE Samoa D genome. Newly predicted genes were mostly identified as HP, TCHP, CHP or hypothetical membrane proteins (HMP). Only one gene TPESAMD_0409a was assigned to preprotein translocase YajC (transport function).

On the contrary, a set of 40 Nichols genes were not annotated in the Samoa D genome (Table S3). All genes but one were defined as genes coding for hypothetical proteins. Product of the TPA Nichols gene TP0590 was named ribosomal protein L36. However, no orthologues were found in nr database, nor any domain was predicted. Thus, the TP0590 gene was also considered as gene encoding hypothetical protein. Because the 150 bp gene limit was applied for hypothetical proteins, 32 out 40 genes were excluded from the Samoa D genome annotation. Additional 8 genes in TPA genome annotation were surrogated by other genes in Samoa D, because a different algorithm for gene prediction was applied.

Based on improved algorithm in the prediction tools of conserved domains and cell localization, and increased amount data in the database, 107 genes in the Samoa D genome were renamed when compared to the Nichols version (Table S4). A novel gene function was assigned to 9 genes previously defined as hypothetical and 71 genes previously defined as conserved hypothetical genes. For nine genes with predicted function, an additional function was predicted, resulting in bifunctional proteins. Remaining 18 genes were renamed according to the standardized gene naming at the HGSC (Gioia *et al.*, 2007; Highlander *et al.*, 2007).

In total, 25 renamed genes were assigned to transport function, 45 genes to general metabolism, 12 genes to translation, 9 genes to cell processes, 2 genes to virulence, 3 genes to DNA metabolism, 6 genes to cell structure, 4 genes to regulation and one renamed gene to a group of unknown proteins.

In summary, *de novo* gene annotation process improved gene predicition, especially for those genes carrying a frameshift mutation in the TPA Nichols genome due to sequencing errors. Moreover, novel gene functions were assigned to 107 (10%) annotated ORFs when compared to Nichols gene naming. Genome annotation of the TPE Samoa D genome found only one novel ORF with assigned gene function when compared to TPA Nichols annotation (AE000520.1). On contrary, no ORF with assigned gene function was missing.

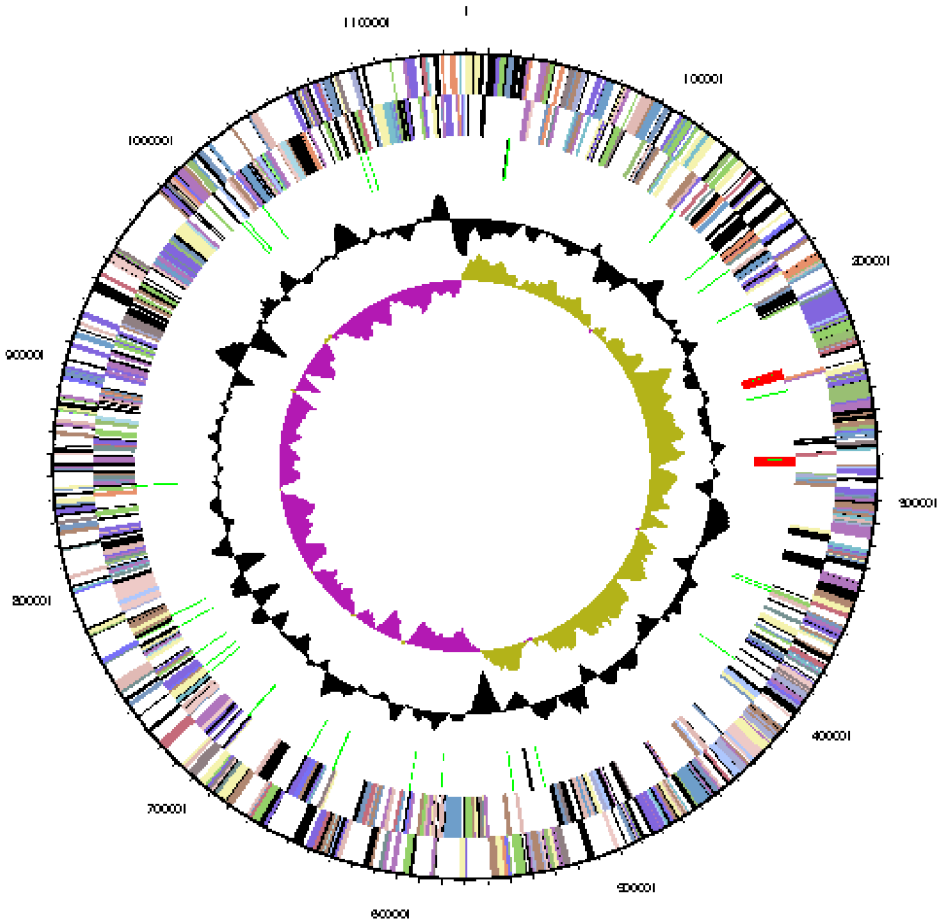
### 4.4. Gene annotation of the TPE CDC-2, TPE Gauthier and TPA DAL-1 genomes

TPA DAL-1 genome was sequenced by 454 pyrosequencing and Illumina methods at the HGSC. The finishing steps were performed by Pavol Mikolka and Marie Zobaníková. TPECDC2, TPEGAU and TPADAL locus tags were assigned for ORFs annotated in CDC-2, Gauthier and DAL-1 genomes, respectively. TPE Samoa D gene annotation served as a reference for annotation of TPE CDC-2, TPE Gauthier and TPA DAL-1 genomes. Predicted

gene coordinates from the Samoa D genome were recalculated based on script using pairwise comparative software Cross_match. The script may screen for any discrepancy between pairs of DNA sequence, including a frameshift mutation, a premature stop codon and a nucleotide substitution in start and stop codons.

In congruence with the WGF results (Mikalová *et al.*, 2010), preserved gene organization were found in all four treponemal genomes investigated. Only small genome rearrangements were found between paralogous families, regions comprising *tpr* genes and *rrn* operons.

**Figure 2.** Circular representation of the TPE CDC-2 chromosome. From outside to the center: Genes on a leading strand (colored by COG categories), Genes on a lagging strand (colored by COG categories), RNA genes (tRNA encoding genes are shown in green, rDNAs in red, other RNAs on black), GC content, GC skew. Downloaded from https://img.jgi.doe.gov/, DOE Joint Genome Institute Integrated microbial genomics expert review.



Out of 1125 genes annotated in TPE Samoa D, 1123 genes were also predicted in the CDC-2 genome while gene orthologues of TPESAMD_0126 and TPESAMD_0924a needed further examination (Table 6). Different number of homopolymeric guanosine tracts in the TPESAMD_0126 (treponemal conserved hypothetical membrane protein, TCHMP, 1908 bp) orthologous region caused a frameshift mutation in CDC-2. While 10 guanosine nucleotides are present in Samoa D, only 9 nucleotides are located in CDC-2, resulting in prediction of two genes, TPECDC2_0126 (HP, 672 bp) and TPECDC2_0126a (HP, 651 bp). The size of TPECDC2_0126 gene is identical to 223AA protein prediction based on promoter sequence analysis (Giacani *et al.*, 2012b). Nine guanosine nucleotides were found also in TPE Gauthier and TPA DAL-1 orthologous region and in other genomes (Giacani *et al.*, 2012b).

The TPESAMD_0924a (HP, 234 bp) orthologous sequence in the CDC-2 genome contains also a frameshift mutation due to variable number of nucleotides in the homopolymeric tract. The predicted homologous gene was shorter than 150 bp, and therefore was not annotated.

In summary, TPE CDC-2 genome consists of 1125 predicted genes. The average and median gene lengths were calculated to 980 bp and 831 bp, respectively. The IGR comprised 52.963 kb and represented 4.65% of the total genome length. The average and median ORF lengths for genes with predicted function were calculated to 844 bp and 657 bp, respectively. Graphical circular representation of the TPE CDC-2 genome is depicted in Figure 2.

In total, 10 ORFs needed to be re-annotated in the TPE Gauthier genome (Table 7). In addition to TPESAMD_0126 and TPESAMD_0924a loci already discussed in the previous paragraph, loci comprising TPESAMD_0040, TPESAMD_0461, TPESAMD_0461a, TPESAMD_0479, TPESAMD_0622, TPESAMD_0629, TPESAMD_0856a, and TPESAMD_0858 needed further evaluation.

Additional three regions with variable number of nucleotides in homopolymeric tracts in the Gauthier genome resulted in a frameshift mutation of total 4 orthologous loci, because the same homopolymeric tract is present in TPESAMD_0461 (transcriptional regulator hypothetical protein, 360 bp) and TPESAMD_0461a (HP, 183 bp) orthologous genes. The different numbers of nucleotides in homopolymeric tracts were also included in TPESAMD_0040 (methyl-accepting chemotaxis protein, 2445 bp) and TPESAMD_0479 (HMP, 675 bp) orthologues.

**Table 6.** Samoa D gene loci manually re-annotated in CDC-2 genome due to orthologous changes affecting original open reading frame.

| TPE Samoa D gene (CP002374.1) | Gene length (bp) / orientation | Gene function | Type of change | TPECDC2 locus tags (length/strand) | Notes |
|---|---|---|---|---|---|
| TPESAMD_0126 | 1377 / minus | treponemal conserved hypothetical membrane protein | frameshift in homopolymeric tracts | _0126 (672 / -) and _0126a (651 / -) | 2 genes predicted due to frameshift |
| TPESAMD_0924a | 234 / minus | hypothetical protein | frameshift in homopolymeric tracts | | no orthologue predicted |

**Table 7.** Samoa D gene loci manually re-annotated in Gauthier genome due to orthologous changes affecting original open reading frame.

| TPE Samoa D gene (CP002374.1) | Gene length (bp) / orientation | Gene function | Type of change | TPEGAU locus tags (length/strand) | Notes |
|---|---|---|---|---|---|
| TPESAMD_0040 | 2445 / plus | probable methyl-accepting chemotaxis protein | frameshift in homopolymeric tracts | _0040 (2316 / +) | 5' truncation |
| TPESAMD_0126 | 1377 / minus | treponemal conserved hypothetical membrane protein | frameshift in homopolymeric tracts | _0126 (672 / -) and _0126a (651 / -) | 2 genes predicted due to frameshift |
| TPESAMD_0461 | 360 / plus | probable transcriptional regulator hypothetical protein | frameshift in homopolymeric tracts | _0461 (363 / +) | 5' truncation; 3' elongation |
| TPESAMD_0461a | 183 / minus | hypothetical protein | frameshift in homopolymeric tracts | _0461a (243 / -) | 3' elongation |
| TPESAMD_0479 | 675 / minus | hypothetical membrane protein | frameshift in homopolymeric tracts | _0675 (540 / -) | 5' truncation |

| TPE Samoa D gene (CP002374.1) | Gene length (bp) / orientation | Gene function | Type of change | TPEGAU locus tags (length/strand) | Notes |
|---|---|---|---|---|---|
| TPESAMD_0622 | 1782 / plus | treponemal conserved hypothetical membrane protein | read through the stopcodon | _0622 (1788 / +) | 3' elongation |
| TPESAMD_0629 | 813 / plus | treponemal conserved hypothetical protein | frameshift | _0629 (456 / +) | 5' truncation |
| TPESAMD_0856a | 1371 / minus | hypothetical protein | frameshift | _0856a (765 / -) | 5' truncation |
| TPESAMD_0858 | 1230 / plus | treponemal conserved hypothetical protein | frameshift | _0858 (1158 / +) | 5' elongation; 3' truncation |
| TPESAMD_0924a | 234 / minus | hypothetical protein | frameshift in homopolymeric tracts | | no orthologue predicted |

**Table 8.** Samoa D gene loci manually re-annotated in TPA DAL-1 genome due to orthologous changes affecting original open reading frame.

| TPE Samoa D gene (CP002374.1) | Gene length (bp) / orientation | Gene function | Type of change | TPADAL locus tags (length/strand) | Notes |
|---|---|---|---|---|---|
| TPESAMD_0009 | 1824 / minus | Tpr protein A | frameshift in homopolymeric tracts | _0009 (1825 / -) | pseudogene in DAL-1 |
| TPESAMD_0012 | 177 / plus | hypothetical protein | frameshift in homopolymeric tracts | | no orthologue predicted |
| TPESAMD_0040 | 2445 / plus | probable methyl-accepting chemotaxis protein | frameshift in homopolymeric tracts | _0040 (2451 / +) | 3' elongation |
| TPESAMD_0067 | 870 / plus | conserved hypothetical protein | frameshift | _0067 (852 / +) | pseudogene in DAL-1 |
| TPESAMD_0103 | 1932 / plus | ATP-dependent helicase RecQ | frameshift in homopolymeric tracts | _0103 (1824 / +) | 3' truncation |

| TPE Samoa D gene (CP002374.1) | Gene length (bp) / orientation | Gene function | Type of change | TPADAL locus tags (length/strand) | Notes |
|---|---|---|---|---|---|
| TPESAMD_0126 | 1377 / minus | treponemal conserved hypothetical membrane protein | frameshift in homopolymeric tracts | _0126 (672 / -) and _0126a (651 / -) | 2 genes predicted due to frameshift |
| TPESAMD_0126b | 408 / plus | hypothetical protein | multiple nucleotide changes | _0126d (324 / -) | no orthologue predicted |
| TPESAMD_0136 | 1413 / plus | treponemal conserved hypothetical outer membrane protein | frameshift | _0136 (1546 / +) | pseudogene in DAL-1 |
| TPESAMD_0179 | 1950 / minus | hypothetical protein | frameshift in homopolymeric tracts | _0179 (1884 / -) | 5' truncation |
| TPESAMD_0314 | 813 / minus | hypothetical protein | frameshift | _0314 (147 / -) and _0315 (648 / -) | fusion in TPE but not TPA genomes, 2 genes predicted |
| TPESAMD_0316 | 1830 / minus | Tpr protein F | frameshift | _0316 (1195 / -) | pseudogene in DAL-1 |
| TPESAMD_0548a | 198 / plus | hypothetical protein | multiple nucleotide changes | | no orthologue predicted |
| TPESAMD_0609 | 1572 / plus | asparagine--tRNA ligase | frameshift in homopolymeric tracts | _0609 (1568 / +) | pseudogene in DAL-1 |
| TPESAMD_0697 | 636 / minus | hypothetical membrane protein | frameshift | _0697 (609 / -) | 5' truncation |
| TPESAMD_911a | 276 / minus | hypothetical protein | frameshift | | no orthologue predicted |
| TPESAMD_1031 | 1671 / minus | Tpr protein L | frameshift | _1031 (1554 / -) | 5' truncation |

Furthermore, a deletion of 302 bp and 79 bp resulted in a frameshift mutation of TPESAMD_0629 (TCHP, 813 bp) and both TPESAMD_0856a (HP, 1371 bp) and TPESAMD_0858 (TCHP, 1230 bp) orthologous loci, respectively. The TPESAMD_0622 (TCHMP, 1782 bp) orthologue in TPE Gauthier harbors a mutation in the stop codon leading to 3' extension of the gene (Table 7).

In summary, TPE Gauthier genome consists of 1125 predicted genes. The average and median gene lengths were calculated to 979 bp and 831 bp, respectively. The IGR comprised 53.300 kb and represented 4.68% of total genome length. The average and median ORF lengths for genes with predicted function were calculated to 841 bp and 652.5 bp, respectively.

The complete genome sequence of TPA DAL-1 strain consists of 1,139,971 bp with GC content of 52.8%. The TPA DAL-1 strain carries in total 16 ORFs which demanded manual analysis and re-annotation (Table 8). All but two TPA DAL-1 orthologous loci bear a frameshift mutation, including 7 loci with frameshift mutation in homolopymeric tracts. Two additional orthologues, TPESAMD_0126b (HP, 408 bp) and TPESAMD_0548a (HP, 198 bp), carried multiple nucleotide changes. Whereas no orthologous ORF was predicted in the TPA DAL-1 genome at the TPESAMD_0548a locus, TPEDAL_0126d (HP, 324 bp) was predicted at the TPESAMD_0126b (Mikalová *et al.*, 2010). Moreover, orthologous ORFs were also not annotated for TPESAMD_0012 (HP, 177 bp) and TPESAMD_0911a (HP, 276 bp), due to length limit of newly predicted ORFs.

In the TPA DAL-1 genome, 5 pseudogenes were annotated for orthologous loci TPESAMD_0009 (encoding TprA protein, 1824 bp). TPESAMD_0067 (CHP, 870 bp), TPESAMD_0136 (treponemal conserved hypothetical outer membrane protein, TCHOMP, 1413 bp), TPESAMD_0316 (TprF protein, 1830 bp), and TPESAMD_0609 (asparagine--tRNA ligase, 1572 bp).

Remaining 5 ORFs were preserved in the TPA DAL-1 gene annotation, however, ORFs were either extended or truncated at 5' or 3' end (Table 8), based on manual annotation of the affected loci. The affected loci included orthologues of TPESAMD_0040, TPESAMD_0103 (ATP-dependent helicase RecQ, 1932 bp), TPESAMD_0179 (HP, 1950 bp), TPESAMD_0697 (HMP, 636 bp), and TPESAMD_1031 (TprL protein, 1671 bp).

In summary, the TPA DAL-1 genome consists of 1124 predicted genes, including 8 pseudogenes. The average and median gene lengths were calculated to 979 bp and 831 bp, respectively. The IGR comprised 53.785 kb and represented 4.72% of the total genome

length. The average and median ORF lengths for genes with predicted function were calculated to 982 bp and 834 bp, respectively.

The summarized genomic features of the TPE strains Samoa D, CDC-2 and Gauthier, and TPA strain DAL-1 are shown in Table 9.

**Table 9**. Summary of the genomic features of the Samoa D, CDC-2, Gauthier and DAL-1 strains.

| Genome parameter | TPE Samoa D | TPE CDC-2 | TPE Gauthier | TPA DAL-1 |
|---|---|---|---|---|
| Genome size | 1,139,330 bp | 1,139,744 bp | 1,139,441 bp | 1,139,971 bp |
| G+C content | 52.80% | 52.80% | 52.80% | 52.80% |
| No. of predicted genes | 1125 including 54 untranslated genes | 1125 including 54 untranslated genes | 1125 including 54 untranslated genes | 1124 including 54 untranslated genes |
| No. of fused genes | 25 (52 corresponding genes in the Nichols genomea) | 24 (50 corresponding genes in the Nichols genomea) | 24 (50 corresponding genes in the Nichols genomea) | 23 (48 corresponding genes in the Nichols genomea) |
| Sum of the intergenic region lengths (% of the genome length) | 52,844 bp (4.64%) | 52,963 bp (4.65%) | 53,300 bp (4.68%) | 53,785 bp (4.72%) |
| Average/median gene length | 980.3/831.0 bp | 980.4/831.0 bp | 979.3/831.0 bp | 979.0/831.0 bp |
| No. of genes encoded on plus/minus DNA strand | 600/525 | 600/525 | 600/525 | 598/526 |
| No. of genes coding for proteins with predicted function | 640 | 640 | 640 | 640 |
| No. of genes coding for treponemal conserved hypothetical proteins | 140 | 140 | 140 | 140 |
| No. of genes coding for conserved hypothetical proteins | 141 | 141 | 141 | 141 |
| No. of genes coding for hypothetical proteins | 147 | 147 | 147 | 146 |
| No. of annotated pseudogenes | 3 | 3 | 3 | 8 |
| No. of tRNA loci | 45 | 45 | 45 | 45 |
| No. of rRNA loci | 6 (2 operons) | 6 (2 operons) | 6 (2 operons) | 6 (2 operons) |
| No. of ncRNAs | 3 | 3 | 3 | 3 |

## 4.5. Intra-strain heterogeneity within the TPE Samoa D, CDC-2 and Gauthier strains

Prior to investigation of inter-strain differences, intra-strain heterogeneity within individual TPE genomes must have been determined. To accomplish it, the high-coverage Illumina reads of individual genomes were used. Applying Burrows-Wheeler alignment followed by SAMtools software (see Materials and Methods section) on Illumina raw data, only nucleotide replacements can be identified.

The analysis revealed 8, 15, and 2 positions showing intra-strain variability in the genomes of TPE CDC-2, Samoa D, and Gauthier, respectively. The most variable gene included *tpr*K gene; 5 variable positions were found in the CDC-2 (CDC-2 coordinates: 976582, 976873, 976876, 977057, 977101), 13 positions in the Samoa D (Samoa D coordinates: 976157, 976259, 976260, 796261, 976316, 976317, 976484, 976681, 976839, 977278, 977279, 977280, 977302), and 2 positions in the Gauthier (976268, 976401) genomes.

In addition to *tpr*K genes, the CDC-2 genome showed heterogeneity within TPECDC2_0313 (coordinates: 331372, 331379) and TPECDC2_0622 (678771); while Samoa D genome in TPESAMD_0110 (125082) and TPESAMD_0134 (155544) genes. It is of interest that all genes with heterogenous positions belong to paralogous families with the exception for TPESAMD_0110.

Other nucleotide change, an adenosine to guanosine transition, was found in the TPE CDC-2 genome at position 335053 (TPECDC2_0317).

An additional variability within TPE genomes was observed in homopolymeric tracts, mostly in guanosine and cytosine homopolymeric tracts alone or in combination with other homopolymeric tract (Table 3, Table 4, Table 10). Most homopolymeric tracts in TPE genomes were directly determined by Illumina sequencing. If Illumina, followed by DDT sequencing did not resolve the exact consensus number in homopolymeric tract, the region was considered to be intra-strain heterogenous.

In the CDC-2 genome, the heterogeneity was observed (Table 10) within intergenic region (IGR) TPECDC2_0084-0085 (11-13 cytosines in homopolymeric tracts, 11-13 C), TPECDC_0126 (9-10 C), TPECDC2_0179 (11-14 guanosine tract, 11-14 G), IGR TPECDC2_0192-0193 (9-11 G), TPECDC_0312a (9-12 G), IGR TPECDC2_0316-0317 (9-11 C), IGR TPECDC2_0317-0319 (11-13 C), TPECDC2_0347 (14-17 G), IGR TPECDC2_0379-0380 (11-12 C), IGR TPECDC2_0381-0383 (12-13 C), TPECDC2_0618

(9-12 C), IGR TPECDC2_0620-0621 (8-11 C), IGR TPECDC2_0621-0622 (11-12 C) and TPECDC2_0859 (13-21 G). In addition, a variable combination of 10-12 C and 9-10 G is located in the TPECDC2_0461 and TPECDC2_0461a genes.

**Table 10.** TPE loci showing intra-strain heterogeneity within nucleotide homopolymeric tracts. Intra-strain heterogeneity is highlighted.

| Coordinates (in TPE Samoa D genome; CP002374.1) | Affected gene / intergenic region (IGR) | CDC-2 | Gauthier | Samoa D[a] |
|---|---|---|---|---|
| 12479-12788 | TPESAMD_0012 | 10 G (Illumina)[b] | 10 - 11 G (TempliPhi)[c] | |
| 49360-49369 | TPESAMD_0040 | 10 G (DDT)[d] | 10 - 11 G (TempliPhi) | heterogeneity |
| 94409-94418 | IGR TPESAMD_0084-0085 | 11-13 C (TempliPhi) | 10 C (Illumina) | |
| 121917-121925 | IGR TPESAMD_0107-0108 | 9 C (DDT) | 10 - 11 C (TempliPhi) | heterogeneity |
| 140638-140646 | IGR TPESAMD_0121-0122 | 9 C (DDT) | 9 C (DDT) | heterogeneity |
| 148037-148046 | TPESAMD_0126 | 9 - 11 C (SMIL)[e] | 9 C (DDT) | |
| 199287-199296 | TPESAMD_0179 | 11 - 14 G (TempliPhi) | 16 - 17 G (TempliPhi) | |
| 208196-208203 | IGR TPESAMD_0192-0193 | 9 - 11 G (TempliPhi) | 9 G (Illumina) | |
| 328700-328710 | TPESAMD_0312a | 9 - 12 G (SMIL) | 11 G (DDT) | |
| 294985-294996 | TPESAMD_0279 | 12 G (Illumina) | 9 - 10 G (TempliPhi) | |
| 333729-333738 | IGR TPESAMD_0316-0317 | 9 - 11 C (SMIL) | 10 C (DDT) | |
| 336039-336048 | IGR TPESAMD_0317-0319 | 11 - 13 C (SMIL) | 11 C (DDT) | |
| 373403-373415 | TPESAMD_0347 | 14 - 17 G (TempliPhi) | 13 - 18 G (TempliPhi) | |
| 374141-374149 | IGR TPESAMD_0347-0348 | 9 G (Illumina) | 12 - 17 G (TempliPhi) | |
| 406570-406578 | IGR TPESAMD_0379-0380 | 11 - 12 C (TempliPhi) | 9 C (DDT) | |
| 409299-409311 | IGR TPESAMD_0381-0383 | 12 - 13 C (TempliPhi) | 13 - 17 C (TempliPhi) | heterogeneity |
| 492462-492501 | TPESAMD_0461 TPESAMD_0461a | 10 - 12 C + 9 - 10 G (TempliPhi) | 9 C + 9 G (Illumina) | |
| 671432-671440 | TPESAMD_0618 | 9 - 12 C (TempliPhi) | 9 C (DDT) | |
| 674508-674516 | IGR TPESAMD_0620-0621 | 8 - 11 C (SMIL) | 9 C (DDT) | |
| 676817-676827 | IGR TPESAMD_0621-0622 | 9 - 12 C (SMIL) | 11 C (DDT) | |
| 937991-938008 | TPESAMD_0859 | 13 - 21 G (TempliPhi) | 17 - 20 G (TempliPhi) | heterogeneity |
| 945685-945693 | TPESAMD_0865 | 9 C + 8 T (DDT) | 9 - 12 C + 9 - 10 T (TempliPhi) | heterogeneity |
| 1055627-1055635 | TPESAMD_0969 | 9 C (DDT) | 9 C (DDT) | heterogeneity |

C, cytosine; G, guanine; T, thymine; Number, number of homopolymers within a tract

[a] The Samoa D genome has been finished by Marie Zobaníková. Only information on the presence of heterogeneity is shown.

[b] The number of nucleotides in a tract was determined by Illumina sequencing.

[c] Up to 8 colonies were cloned and sequenced using TempliPhi Amplification kit and DDT sequencing.

[d] The number of nucleotides in a tract was determined by DDT sequencing from XL PCR product.

[e] The locus was deeply sequenced due to Small Insert Library construction.

In the Gauthier genome, intra-strain variability was determined in TPEGAU_0012 (10-11 G), TPEGAU_0040 (10-11 G), IGR TPEGAU_0107-0108 (10-11 C), TPEGAU_0179 (16-17 G), TPEGAU_0279 (9-10 G), TPEGAU_0347 (13-18 G), IGR TPEGAU_0347-0348 (12-17 G), IGR TPEGAU_0381-0383 (13-17 C), TPEGAU_0859 (17-20 G) and TPEGAU_0865 (a combination of 9-12 C followed by 9-10 thymine homopolymeric tract).

Furthermore, the intra-strain variability was also present in the Samoa D genome in TPESAMD_0040, IGR TPESAMD_0107-0108, IGR TPESAMD_0121-0122, IGR TPESAMD_0381-0383, TPESAMD_0859, TPESAMD_0865 and TPESAMD_0969.

In summary, the observed intra-strain variability is strain specific. The most affected regions are located in paralogous sequences in the proximity of tpr genes with nucleotide substitutions and variable number of nucleotides in homolopolymeric tracts. The CDC-2 was found to be the most intra-strain variable genome. However, many intra-strain variations were found due to SMIL construction. Since this method was performed only in the CDC-2 strain, this conclusion on the most variable genome may be misinterpreted.

### 4.6. Inter-strain heterogeneity between the TPE Samoa D, CDC-2 and Gauthier strains

To further understand the relationship between TPE strains, whole genome alignments were performed using SeqMan and Cross_match software. The complete genome sequences of the TPE Samoa D, CDC-2 and Gauthier strains consist of 1,139,330 bp, 1,139,744 bp and 1,139,417 bp, respectively. All three genomes are collinear except for some paralogous regions and small insertions/deletions (indels). The core TPE genome comprises 1115 genes, genes which do not contain a frameshift mutation in all TPE orthologues, including 54 non-translated genes and 1058 ORFs.

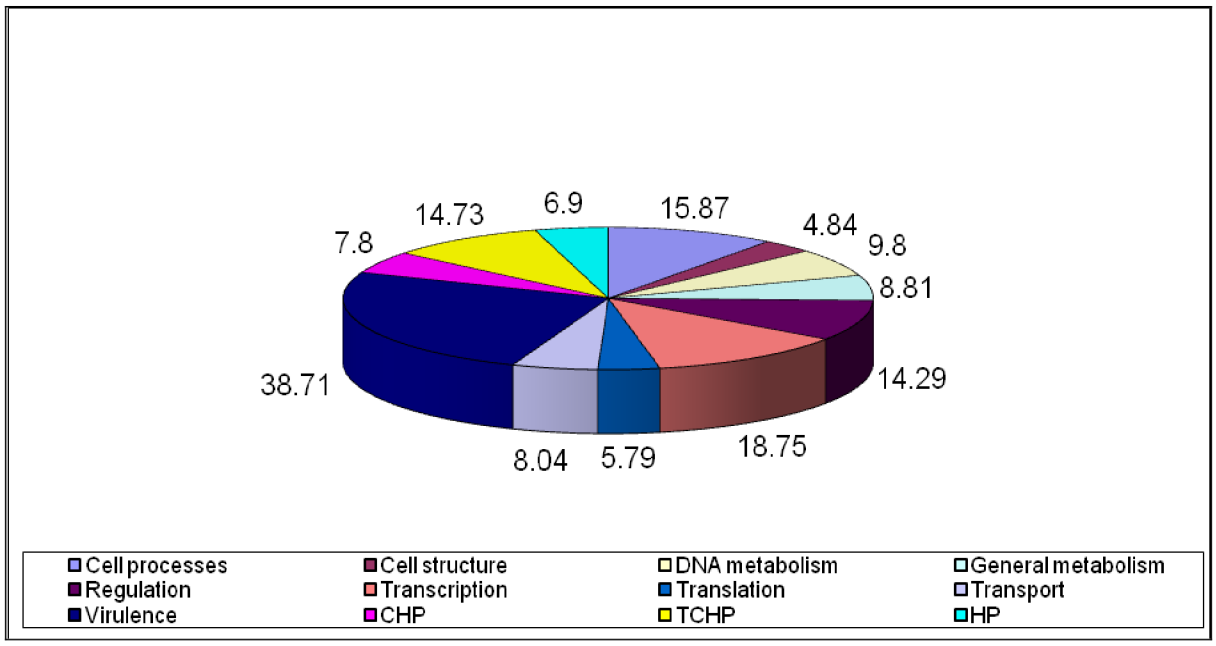
The comprehensive comparative analysis is shown in Table S5. TPE prefix is used for TPE orhologues. Because of high sequence heterogeneity, *tpr*K (TPE_0897) and *tpr*D (TPE_0131) orthologues were excluded from the analysis in this section. At the DNA level, the analysis indicates 160 nucleotide substitutions, 20 indels, a reciprocal translocation and four multiple rearrangements of SNPs and indels (within TPE_0136, TPEGAU_0858/TPE_0856a, TPE_0967, TPE_0326 genes) among individual TPE strains. There are 16 changes located in the intergenic regions inbetween two ORFs, including 8 indels, 7 substitutions, and 1 reverse translocation of *rrn* operons. Regarding the coding

sequences, 108 TPE ORFs are altered due to nucleotide differences; 14 indels and 178 single nucleotide changes were observed. Out of 14 indels located within genes, 7 resulted in a frameshift while 7 indels had no affect on the gene frame. 178 single nucleotides substitutions were divided into read-through stop codon mutations (1 gene), two-nucleotide mutations (30), synonymous mutations (43) and non-synonymous mutations (101). The latter were further subdivided into conservative (7) and non-conservative changes (94).

With respect to gene functional groups, the most affected category is the group involved in virulence processes with 38.71% of affected genes (Figure 3). All *tpr* genes were included in this category. The frequency of heterogenous alleles is also high in the groups engaged in transcription, cell processes, gene regulation and in genes encoding treponemal conserved hypothetical proteins with 18.75, 15.87, 14.29 and 14.73 percentages, respectively.

In addition to genes causing frameshift (Tables 6, 7) and heterogenous *tpr*K gene, *tpr*D genes, genes encoding proteins with more than 2AA replacements (14 genes, Table 11) are most variable genes between TPE strains examined. The genes encoding proteins with more than 2AA changes were considered genes with multiple sequence changes (MSC).

**Figure 3.** Percentual distribution of TPE inter-strain variable genes within functional groups.




Large indels were found among genes harboring tandem repeats of 60 and 24 bp, in TPE_0433 (*arp*, acidic repeat protein) and TPE_0470 (CHP), respectively. While 5' and

**Table 11.** Proteins with the highest inter-strain heterogeneity among TPE genomes.

| Affected gene (gene function; gene product) | Diverse genome(s) | Gene length in Samoa D genome (AA) | Type of change | Z-test of selection (p)[a] | Note |
|---|---|---|---|---|---|
| TPE_0067[b] (unknown; CHP) | Samoa D | 289 | 101 AA deletion | neutral (1.000) | insertion terminated by 12-bp direct repeats in CDC-2 and Gauthier; only 1 repeat present in Samoa D |
| TPE_0131 (virulence; Tpr protein D) | Gauthier | 598 | allele *tpr*D3 | recombination[c] | |
| TPE_0136 (virulence; TCHOMP) | CDC-2 | 470 | 11 AA insertion | positive (0.071) | insertion within 33-bp tandem repeats, in CDC-2 strain 2 complete and 1 partial (23-bp) repeats |
| TPE_0316 (virulence; Tpr protein F) | CDC-2, Samoa D | 609 | in total, 5 AA changed | recombination | |
| TPE_0317 (virulence; Tpr protein G) | Samoa D | 756 | 4 AA changed | recombination | |
| TPE_0322 (transport; sugar ABC superfamily ATP binding cassette transporter, membrane protein) | CDC-2 | 400 | 9 AA changed | neutral (1.000) | |
| TPE_0326 (virulence; OMP) | all | 833 | in total, 5 AA changed and indel of 1 AA | positive (0.016) | |
| TPE_0346 (unknown; CHP) | CDC-2 | 233 | 2 AA changed | positive (0.073) | |
| TPE_0433 (unknown, Arp protein) | all | 564 | indel up to 160 AA | neutral (1.000) | indels of 60 bp repeats |
| TPE_0470 (unknown; CHP) | all | 329 | indel up to 200 AA | neutral (1.000) | indels of 24 bp repeats |
| TPE_0488 (cell processes; methyl-accepting chemotaxis protein) | all | 845 | in total, 10 AA changed | positive (0.001) | |
| TPE_0548 (unknown; TCHMP) | all | 432 | in total, 14 AA changed | positive (0.000) | |
| TPE_0746 (general metabolism; phosphate dikinase) | Gauthier, Samoa D | 901 | in total, 2 AA changed | positive (0.075) | |
| TPE_0865 (unknown; TCHOMP) | all | 481 | 4 AA changed | positive (0.057) | |
| TPE_0967 (unknown; TCHP) | Samoa D | 515 | 6 AA deletion | neutral (1.000) | deletion within 9-bp tandem repeats, in CDC-2 and Gauthier 3 complete and 1 incomplete repeats |

Arp, acidic repeat protein; CHP, conserved hypothetical protein; indel, insertion or deletion; OMP, outer membrane protein; TCHMP, treponemal conserved hypothetical membrane protein; TCHP, treponemal conserved hypothetical protein; TCHOMP, treponemal conserved hypothetical outer membrane protein

[a] The selection test was calculated using the Kumar model within MEGA4 software.
[b] TPESAMD_0067, TPECDC2_0067 and TPEGAU_0067 orthologs
[c] The recombination event at this locus was predicted using RDP3 software.

3' termini of both genes are conserved, the number of tandem repeats varies. In the *arp* gene, 4 (CDC-2), 10 (Gauthier) and 12 (Samoa D) repeats were found while in the TPE_0470 gene, 12 (Samoa D), 26 (Gauthier) and 37 (CDC-2) tandem repeats were observed.

A deletion of 303 bp was detected in the TPESAMD_0067 (CHP). While two 12-bp repeats surround a unique sequence (291 bp) in CDC-2 and Gauthier, the unique sequence of 291 bp and a 12-bp repeat has been deleted in Samoa D genome. The other 12-bp repeat remained in the Samoa D genome. Recently, the gene was renamed as gene coding for cell division protein based on TPA protein-protein interactome study (Titz *et al.*, 2008).

An insertion of 33 bp was observed in TPECDC2_0136 (treponemal conserved hypothetical outer membrane protein, TCHOMP), gene encoding fibronectin-binding protein (Brinkman *et al.*, 2008). The DNA sequence of TPECDC2_0136 comprises a cluster of 89 bp with a repeating 33-bp motif (2 x 33 + 23 bp). TPE_0136 counterparts consist of 56-bp cluster with an incomplete 33-bp repeat in Samoa D and Gauthier strains (33 + 23 bp, CP002374.1: 158172-158229). While repeating motif is stable throughout the CDC-2 cluster, a single nucleotide change (SNP) occurred in both, Samoa D and Gauthier othologous sequences. When all three clusters were aligned, a SNP at position 15 in Samoa D (CP002374.1: 158187, Table S5) and a SNP at position 30 in Gauthier (CP002374.1: 158202, Table S5) were revealed. Moreover, TPE_0136 carries other single nucleotide changes and was found most variable gene among pathogenic and closely related non-pathogenic treponemes (Flasarová *et al.*, 2006; Matějková *et al.*, 2008; Šmajs *et al.*, 2012).

A deletion of 18 bp was found in TPESAMD_0967 (TCHP) when compared to other TPE_0967 genes. Within a gene, four 9-bp tandem repeats are located in CDC-2 and Gauthier. Exactly two repeats were deleted in Samoa D genome.

Additional 3 genes with MSC belong to *tpr* family of paralogous genes. The *tpr*D (TPE_0131) gene was found most variable between investigated TPE strains. Whereas Samoa D and CDC-2 strains carry allele *tpr*D2, Gauthier strain harbors an allele *tpr*D3 (Centurion-Lara *et al.*, 2000b). Indeed, all three variable *tpr* genes, *tpr*D, *tpr*F (TPE_0316) and *tpr*G (TPE_0317) are prone to recombination in TPE genomes (Gray *et al.*, 2006). Thus, it is impossible to estimate the selection type on these genes.

Altogether, three genes were found to be under positive selection ($p < 0.05$) in TPE genomes. TPE_0548 (treponemal conserved hypothetical membrane protein, TCHMP), TPE_0326 (*tp*92, outer membrane protein) and TPE_0488 (*mcp*, methyl-accepting chemotaxis protein) vary in 14, 5 and 10 AA, respectively. Then, it is of interest that another membrane protein with 9 different AA encoded by TPE_0322 (sugar ABC superfamily ATP binding

cassette transporter, membrane protein) is under neutral selection (p = 1.00). The TPECDC2_0322 gene carries an insertion followed by a downstream deletion changing in a total of 9 AA.

Remaining high variable TPE genes are as follows: TPE_0346 (CHP), TPE_0746 (phosphate dikinase) and TPE_0865 (TCHOMP). All three genes are under mild positive selection constraint ($0.05 < p < 0.10$).

In summary, many differences between TPE genomes regard some repetitive motif, either in homopolymeric tract, or in tandem as well as in dispersed repeats. Membrane proteins and proteins involved in transport, chemotaxis and cell division were found among highly heterogenous proteins in TPE genomes.

### 4.7. Structure of *rrn* operons in treponemes

Two *rrn* operons (16S-23S-5S) were described in pathogenic *Treponema* genomes with the 16S-23S intergenic spacer region (ISR) comprising genes coding for either tRNA-Ala or tRNA-Ile (Fraser *et al.*, 1998; Fukunaga *et al.*, 1992; Giacani *et al.*, 2010a; Šmajs *et al.*, 2011). Using XL-PCR, *rrn* operons were amplified from 20 treponemal strains (Table 1) including 11 TPA strains, 5 TPE strains, an unclassified simian isolate, 2 strains of *T. pallidum* ssp. *endemicum* (TEN), and a rabbit *T. paraluiscuniculi* (TPc) isolate. The XL-PCR products were obtained for all 40 investigated regions. However, the assembled sequence of *rrn*2 operon of Iraq B (TEN) was repeatedly ambiguous at several positions, probably due to low DNA quality. The amplification and sequencing was provided by Marie Zobaníková, while author of this Doctoral thesis exerted the analysis.

In the individual TPA genomes, the amplified *rrn*1 and *rrn*2 regions were identical in 5141 bp (Table 12) including the DNA regions 212 bp upstream of 16S rDNA, 16S rDNA (1537 bp), 23S rDNA (2951 bp), 5S rDNA (110 bp), 23S-5S ISR (50 bp), and a region of 54 bp downstream of 5S rDNA. Additional identical sequences were located within the 16S-23S ISR downstream of 16S (120 bp) and upstream of 23S (118 bp) rDNA genes (Figure 4). Alternate sequences within the 16S-23S ISR, encoding tRNA-Ile or tRNA-Ala, comprised an additional 64 bp or 74 bp, respectively (Figure 4). To extend comparative analysis over all

**Table 12.** DNA sequence polymorphisms found between 20 pathogenic *Treponema* strains in the *rrn* operons. Highlighted text in yellow shows single nucleotide polymorphisms, whereas text highlighted in orange shows translocation of tRNA encoding genes.

| Position downstream (D) or upstream (U) or within gene coding for rRNA or tRNA | | Treponemal homologous sequences of rDNA operons | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | IGR (212 bp) | | | 16S rDNA (1537 bp) | | | | IGR (117 or 116 bp)[a] | tRNA (74 bp)* | IGR (111 or 122 bp)[a] | 23S rDNA (2951 bp) | | | | | | | IGR (50 bp) | 5S rDNA (110 bp) |
| | | 171-167 U | 96 U | 93 U | 647 | 1134 | 1375 | 1441 | 71 D | | 21 U | 458 | 763 | 766 | 1092 | 1359 | 1546 | 2104 | 47 D | 81 |
| TPA | Bal 73-1 (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Bal 73-1 (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Chicago (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Chicago (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | DAL-1 (*rrn1*) | GGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | DAL-1 (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Grady (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Grady (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Haiti B (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Haiti B (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Madras (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Madras (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Mexico A (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Mexico A (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | MN-3 (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | MN-3 (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Nichols (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Nichols (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Philadelphia 1 (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | Philadelphia 1 (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Philadelphia 2 (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | A | C | C |
| | Philadelphia 2 (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | A | C | C |
| | SS14 (*rrn1*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ile | G | G | G | G | G | A | A | G | C | C |
| | SS14 (*rrn2*) | GGGGG | A | A | G | G | G | C | G | tRNA-Ala | G | G | G | G | G | A | A | G | C | C |
| TPE | CDC-1 (*rrn1*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| | CDC-1 (*rrn2*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| | CDC-2 (*rrn1*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| | CDC-2 (*rrn2*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| | Gauthier (*rrn1*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| | Gauthier (*rrn2*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| | Samoa D (*rrn1*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| | Samoa D (*rrn2*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| | Samoa F (*rrn1*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| | Samoa F (*rrn2*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| simian isolate | Fribourg-Blanc (*rrn1*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ala | G | A | G | A | G | A | A | A | C | C |
| | Fribourg-Blanc (*rrn2*) | GGGGG | A | G | G | G | G | C | G | tRNA-Ile | G | A | G | A | G | A | A | A | C | C |
| TEN | BosniaA (*rrn1*) | GGGGG | A | A | G | G | A | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| | BosniaA (*rrn2*) | GGGGG | A | A | G | G | A | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| | Iraq B (*rrn1*) | GGGGG | A | A | G | G | A | C | G | tRNA-Ala | G | G | G | A | G | A | A | A | C | C |
| | Iraq B (*rrn2*) | GGGGG | A | A | G | G | A | C | G | tRNA-Ile | G | G | G | A | G | A | A | A | C | C |
| TPc | Cuniculi A (*rrn1*) | GGGGG | G | A | A | A | G | T | A | tRNA-Ile | A | G | A | A | A | G | G | A | T | A |
| | Cuniculi A (*rrn2*) | GGGGG | G | A | A | A | G | T | A | tRNA-Ala | A | G | A | A | A | G | G | A | T | A |

[a] The size between 16S and 23S rDNA (both excluded) varies based on the presence of tRNA-Ile (117+74+111, in total 302 bp) or tRNA-Ala (116+74+122, in total 312 bp) gene.

**Figure 4.** Alignment of 16S-23S intergenic spacer regions in TPA Nichols *rrn* operons. The gene coding for tRNA-Ile (TP_t12) is shown in red while the gene coding for tRNA-Ala-3 (TP_t15) is in blue.

```
ISR 16S-23S-rrn1      CCTTTCTAAGAGAAAGGGTATGGGCATGGCATGGTGCCGTGTTCGCCGTGTGGCGGAAGC 60
ISR 16S-23S-rrn2      CCTTTCTAAGAGAAAGGGTATGGGCATGGCATGGTGCCGTGTTCGCCGTGTGGCGGAAGC 60
                      ************************************************************

ISR 16S-23S-rrn1      CACACGGTAGGTTTTTCTGCTCCTGCACGGCAGTCTCTCCCCTTCCCTTTTGAAAAGGGG 120
ISR 16S-23S-rrn2      CACACGGTAGGTTTTTCTGCTCCTGCACGGCAGTCTCTCCCCTTCCCTTTTGAAAAGGGG 120
                      ************************************************************

ISR 16S-23S-rrn1      CTTGTAGCTCAGTTGGTTAGAGC-ACTTCTCTGATAAGGAAGGGGTCATTAGTTCAA--- 176
ISR 16S-23S-rrn2      GT-ATAGCTCAACTGGCTAGAGCGACGGCTTTG-CAGGTCGTAGGTCAGGGGTTCGAGTC 178
                       *  *******  *** ****** **  ** **  * *     *****   **** *

ISR 16S-23S-rrn1      CTCTAATC--------AAGCCCACTATTATTCTTTATGTCCCTTTGTTTTGTTTATGGGG 228
ISR 16S-23S-rrn2      CCCTTATCTCCAGGGAAAGCCCACTATTATTCTTTATGTCCCTTTGTTTTGTTTATGGGG 238
                      * ** ***        ******************************************

ISR 16S-23S-rrn1      TAAGGAGTAGGTGGTAGGTGATTTTTGAGAGTATTAGGGTGGGGTGTGAAGTTGAGAAGG 288
ISR 16S-23S-rrn2      TAAGGAGTAGGTGGTAGGTGATTTTTGAGAGTATTAGGGTGGGGTGTGAAGTTGAGAAGG 298
                      ************************************************************

ISR 16S-23S-rrn1      GATGGATAATATGG 302
ISR 16S-23S-rrn2      GATGGATAATATGG 312
```

available data, TPA Chicago sequences of *rrn* operons (CP001752.1) (Giacani *et al.*, 2010a) were added to the sequences of 20 strains used in this study.

When compared to TPA strains, a deletion of 33 bp was found in homologous regions of the *rrn*2 region in the TPE, TEN, and the simian strains. The TPc strain contained a 17-bp deletion at the same position. These deletions resulted in shortening (33-bp deletion) or truncation (17-bp deletion) of TP0266 orthologs. Among all investigated strains, in addition to the observed deletions, we found only 17 heterogenous sites within the entire region, excluding the 16S-23S ISR encoding tRNA-Ile or tRNA-Ala. Sixteen sites were single nucleotide changes and one was a single bp deletion (Table 12). The *rrn*1 operon of the reference TPA Nichols genome (AE000520.1) (Fraser *et al.*, 1998) showed a deletion within 16S rDNA (data not shown), whereas all other strains, including the Nichols strain examined in our study, did not. That deletion represent a sequencing error present in the reference Nichols genome (AE000520.1). On the contrary, a 1-bp deletion in the TPA DAL-1 genome, upstream of 16S rDNA gene in *rrn*1 operon, was repeatedly confirmed by DDT and Illumina sequencing. In a parallel whole genome sequencing project, 31% of Illumina reads from TPA DAL-1 showed a deletion (Zobaníková *et al.*, 2012) verifying the heterogeneity upstream of

16S rDNA genes. No such heterogeneity was found in other invastigated treponemal strains (Čejková *et al.*, 2012; Pětrošová *et al.*, 2012; Šmajs *et al.*, 2011). The identified nucleotide change in the 2104$^{th}$ position of the 23S rDNA gene (differentiating the SS14 strains from other investigated strains) corresponds to the mutation causing macrolide resistance in treponemal strains (Stamm & Bergen, 2000b).

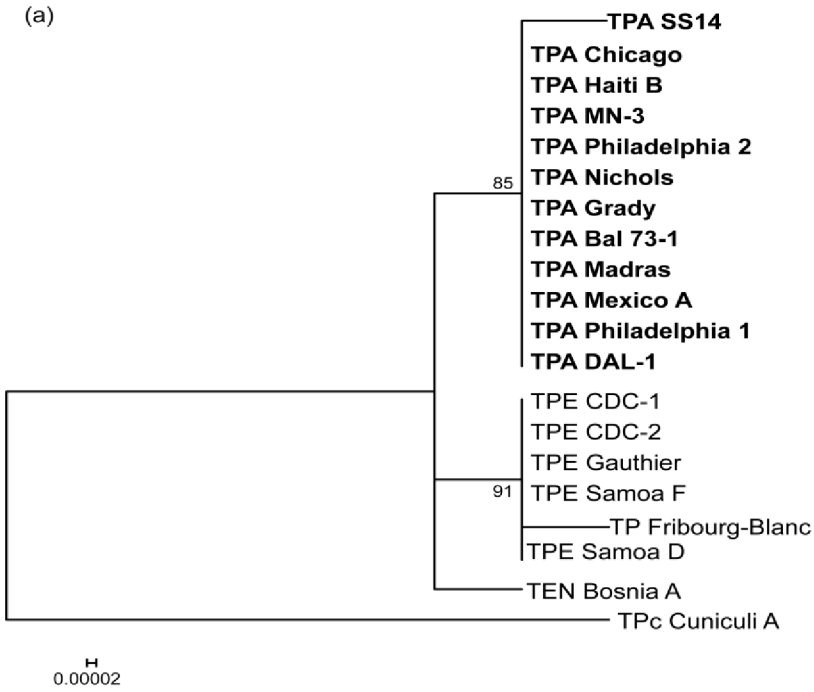
All TPA strains differ from other pathogenic treponemes by a nucleotide change at the 766$^{th}$ position of 23S rDNA genes. TPE strains and the simian isolate Fribourg-Blanc can be distinguished from other pathogenic treponemes by a SNP localized 93 bp upstream of 16S rDNA genes. TPE strains can be differentiated from the simian isolate by a nucleotide sequence change in the 23S rDNA genes (458$^{th}$ bp). TEN showed a nucleotide change in the 16S rDNA genes, and *T. paraluiscuniculi* showed 12 nucleotide changes in investigated *rrn* region sequences (Table 12).

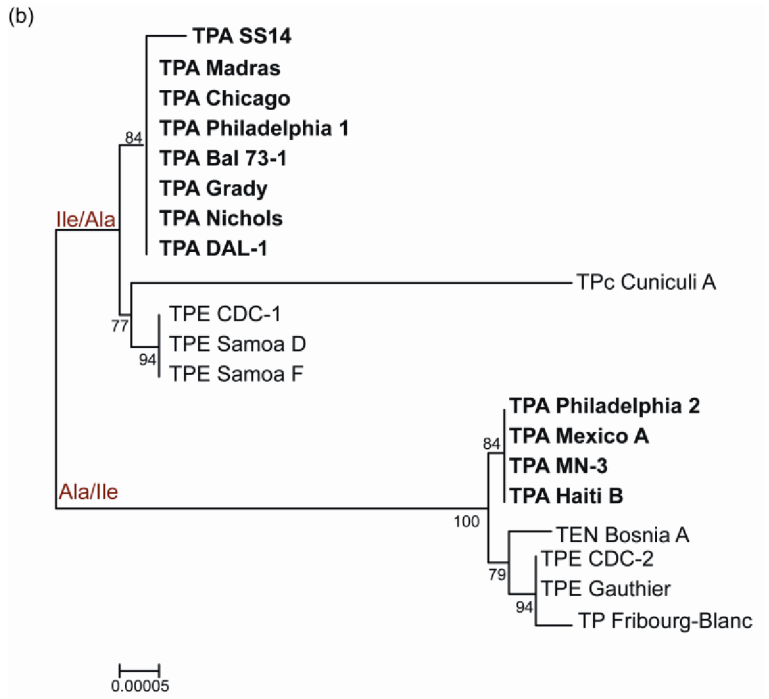
Contrary to the phylogenetically conserved SNP distribution in the repetitive sequences of *rrn* operons, the genes coding for tRNA did not show the same evolutionary pattern (Table 12, Figure 5). In this study, two 16S-23S ribosomal intergenic spacer patterns were observed. The spacer pattern Ile/Ala includes the tRNA-Ile gene within the *rrn*1 region and the tRNA-Ala gene within the *rrn*2 region. The Ile/Ala pattern was observed in the following strains: TPA Nichols, Bal 73-1, Grady, SS14, Chicago, DAL-1, Philadelphia 1, and Madras; TPE Samoa D, CDC-1, and Samoa F; and TPc Cuniculi A. The reverse intergenic spacer pattern Ala/Ile consists of the tRNA-Ala gene within the *rrn*1 region, and the tRNA-Ile gene within the *rrn*2 region. The Ala/Ile pattern was found in TPA Mexico A, MN-3, Philadelphia 2, and Haiti B strains; in TPE Gauthier and CDC-2; in an unclassified treponeme Fribourg-Blanc; and in TEN Iraq B and Bosnia A genomes.

The concatenated *rrn* operons, excluding tRNA coding genes and their vicinity, cluster according to the species/subspecies classification (Figure 5a). The TEN Iraq B strain was omitted from the analysis because we were unable to obtain unambiguous *rrn*2 operon sequence. Nevertheless, the *rrn*1 operon was identical to another TEN strain Bosnia A. On the contrary, the trees showing concatenated *rrn* operons including tRNA coding genes (Figure 5b) are branched according to the composition of tRNA coding genes in individual *rrn* operons, and then according to the species/subspecies classification. This phenomenon can be explained by recombination events that occurred between *rrn* operons.

To predict recombination hot spot sites within *rrn* operons, four methods from the Recombination Detection Program (RDP3) were applied. All four methods predicted four

**Figure 5.** Panel (a). An unrooted tree constructed from the concatenated sequences of *rrn* operons excluding heterologous tRNA coding genes. The *rrn* operons cluster according to the species/subspecies classification of treponemes. Bar scale represents 0.00002 nucleotide substitutions per site. Panel (b). An unrooted tree constructed from the sequences of *rrn* operons including heterologous tRNA coding genes. The *rrn* operons cluster according to the intergenic spacer pattern. Bar scale represents 0.00005 nucleotide substitutions per site. Bootstrap values based on 1,000 replications are shown next to branches. TPA strains causing syphilis are shown in bold.



recombination sites, two sites in each *rrn* operon (Table 13). Predicted sites correspond to the same positions within 16S (783) and 23S (324) rDNA genes in both *rrn* operons.

**Table 13.** Predicted recombination hot spot sites using RDP3 program.

| Predicted recombination hot spot site | | Prediction algorithm used in the RDP3 program (p-value) | | | |
|---|---|---|---|---|---|
| begin* | end* | RDP | GENECONV | MaxChi | Chimera |
| 231656 | 233036 | 1.33E-58 | 2.20E-55 | 8.06E-13 | 7.80E-13 |
| 280058 | 281448 | 7.22E-20 | 1.12E-20 | 5.91E-04 | 1.33E-03 |

*Whole genome TPE Samoa D coordinates are shown (GenBank acc. no. CP002374.1) (Čejková *et al.*, 2012)

## 4.8. Comparative phenomics applied on BAC library of the TPA Nichols DNA

BiOLOG Phenotype MicroArrays (PMs) were performed to compare metabolic profiles of individual BAC library clones with different TPA Nichols DNA fragments cloned (Šmajs *et al.*, 2002).

In total, 190 substrates were tested as a sole carbon sources for individual BAC clones in *E. coli*, and for *E. coli* DH10B and *E. coli* DH10B pBeloBAC11, the negative controls of the experiment. In addition to carbon sources, TPA Nichols BAC growth conditions were tested in additional 192 pH sensitivity and 240 antibiotic resistance assays. PMs were designed to test resistance to antibiotic in four two-fold dilution series, thus 10 PM plates of 96 wells were used for 240 antibiotioc assays.

Prior to compare phenotypes between BAC libraries, the phenotypes of E. *coli* DH10B and *E. coli* DH10B pBeloBAC11 were compared. Only a difference in resistance to chloramphenicol was detected. A chloramphenicol and two other chloramphenicol derivates are included in three different PM plates. In all cases, *E coli* DH10B harboring chloramphenicol resistant pBeloBAC11 plasmid was resistant, while *E. coli* DH10B was sensitive. This test also revealed reproducibility of experimental results. Throughout all 14 PM panels tested (1344 wells), only several wells showed variability in response between replicates. For example, individual replicates of the same strain showed variable resistance/sensitivity responses to chloroxylenol (PM16 panel, H05-H08 wells). Those wells were omitted from further analyses.

In general, increased resistance to antibiotics (Table 14) and differences in carbon utilization (Table 15) were found between phenotype profiles of several BAC clones and *E. coli* DH10B pBeloBAC11 (reference) strain.

An increased (at least 2-fold) resistance to lincomycin, ceftriaxone, cephalotin, cinoxacin, D,L-serine hydroxymate, guanazole and 2,4-diamino-6,7-diisopropyl-pteridine was derived from phenotype profiles of *E. coli* DSTP001, *E. coli* DSTP094 and *E. coli* DSTP334 (Table 14), consistently found in all replicates of strain analyses. *E. coli* DSTP001, which carries TPA Nichols TP0084 – TP0137 intact genes, was found more resistant to lincomycin and cephalosporins (ceftriaxone, cephalotin). The PM revealed *E. coli* DSTP0334 (TP0597 - TP0633 intact genes) to be more resistant to cinoxacin and 2,4-diamino-6,7-diisopropyl-pteridine and *E. coli* DSTP054 (TP0538 - TP0595 intact genes) more resistant to D,L-serine hydroxamate and guanazole than the reference strain and other BAC clones.

**Table 14.** Metabolic profile comparison between individual TPA Nichols BAC clones and negative control *(E. coli* pBeloBAC11) strain depicting an elevated antibiotic resistance of BAC clones.

| Elevated antibiotic resistance (at least 2-fold) | TPA Nichols BAC clone in *E. coli* | Intact genes included in BAC clone | PM position |
|---|---|---|---|
| Lincomycin | DSTP001 | TP0084-0137 | PM11-A12 |
| Ceftriaxone | DSTP001 | TP0084-0137 | PM11-G03 |
| Cephalotin | DSTP001 | TP0084-0137 | PM11-H03 |
| Cinoxacin | DSTP334 | TP0597-0633 | PM16-D10 |
| D,L-serine hydroxamate | DSTP094 | TP0538-0595 | PM12-C11 |
| Guanazole | DSTP094 | TP0538-0595 | PM18-G06 |
| 2,4-diamino-6,7-diisopropyl-pteridine | DSTP334 | TP0597-0633 | PM12-E04 |

Whereas reference strain and other BAC clones showed utilization, *E. coli* DSTP001 and *E. coli* DSTP0334 repressed utilization of several sole carbon sources (Table 15), including succinic acid, L-aspartic acid, L-proline, D-alanine, L-asparagine, α-keto-butyric acid, M-tartaric acid, fumaric acid, bromo-succinic acid and D-malic acid. It its of interest that all listed carbons are utilized via the Krebs cycle, a cycle which components are missing in treponemal strains.

Moreover, other nutrients (D,L-α-glycerol-phosphate, D-glucose-6-phosphate, D-glucose-1-phospate and D-fructose-6-phosphate) were not utilized as sole carbon sources in DSTP021 (TP0203 – TP0240 intact genes) and *E. coli* DSTP055 (TP0966 – TP1026 intact genes) while PMs revealed utilization in other investigated strains. All these nutrients are metabolized via glycolysis pathway.

On the other hand, the growth rates of *E.coli* DSTP094 (TP0538 – TP0595 intact genes) and *E. coli* DSTP198 (TP0628 – TP0686 intact genes) were very slow for direct comparison with other strains examined in PMs. Even extended cultivation to 72 hours showed no or low utilization and respiration in all examined PM array wells. Thus, the list of all suppressed metabolic profiles is not shown.

**Table 15.** Metabolic profile comparison between individual TPA Nichols BAC clones and negative control *(E. coli* pBeloBAC11) strain showing suppresseed metabolic activity of BAC clones.

| Suppressed metabolism | TPA Nichols BAC clone in *E. coli* | Intact genes included in BAC clone | Carbon source |
|---|---|---|---|
| Krebs cycle | DSTP001 | TP0084-0137 | D-alanine |
| Krebs cycle | DSTP001 | TP0084-0137 | L-asparagine |
| Krebs cycle | DSTP001 | TP0084-0137 | succinic acid |
| Krebs cycle | DSTP001 | TP0084-0137 | α-keto-glutaric acid |
| Krebs cycle | DSTP001 | TP0084-0137 | D-malic acid |
| Glycolysis | DSTP021 | TP0203-0240 | D-fructose-6-phosphate |
| Glycolysis | DSTP021 | TP0203-0240 | D-gluconic acid |
| Glycolysis | DSTP021 | TP0203-0240 | D-glucose-1-phosphate |
| Glycolysis | DSTP021 | TP0203-0240 | D-glucose-6-phosphate |
| Glycolysis | DSTP055 | TP0966-1026 | D,L-α-glycerol-phosphate |
| Krebs cycle | DSTP334 | TP0597-0633 | bromo succinic acid |
| Krebs cycle | DSTP334 | TP0597-0633 | Fumaric acid |
| Krebs cycle | DSTP334 | TP0597-0633 | L-aspartic acid |
| Krebs cycle | DSTP334 | TP0597-0633 | L-proline |
| Krebs cycle | DSTP334 | TP0597-0633 | M-tartaric acid |

Suppressed metabolism for DSTP094 and DSTP198 clones not shown

# 5. DISCUSSION

## 5.1. Whole genome sequencing and annotation of TPE strains

The whole genome sequencing of TPE strains was established in order to investigate differences between and within *T. pallidum* subspecies and other closely related strains. Limited samples of TPE isolates are available worldwide. The Samoa D strain was isolated in Samoa in 1953 (Turner & Hollander, 1957), the CDC-2 in Ghana in 1980 (Liska *et al.*, 1982), and the Gauthier strain in the Republic of Congo in 1960 (Gastinel *et al.*, 1963). Because the three strains were isolated from different places at different times, they might be expected to show a reasonable amount of variation for the whole genome comparison studies.

A combination of next-generation (NGS) and dideoxy-terminator (DDT) sequencing techniques has been applied to reveal the complete genomes of TPE strains. Pooled segment genome sequencing (PSGS) has been developed and applied in order to determine the Gauthier genome. PSGS enables one to sequence samples with limited quality and quantity of gDNA (Strouhal, 2010). If Multiplex Identifiers (MIDs, indexes) are added, the different libraries can be indexed separately, and thus recognized in analysis. Four pools were made from PCR products of the Gauthier DNA and multiplexed to avoid misassembly of paralogous regions (*tpr* genes and vicinity).

Because the Samoa D genome was independently sequenced by three NGS techniques (Čejková *et al.*, 2012), the final genome sequence was expected to be of the highest quality. In addition, the complete genome was experimentally confirmed by physical mapping (Čejková *et al.*, 2012; Strouhal, 2010), thus the sequencing error rate appeared to be on the order of $10^{-4}$ or better. Using the final genome sequence as a scaffold, the advantages and limitations of each NGS technique could be assessed (Zobaníková, unpublished results). All techniques have their limitations. The CGS technique needs a closely related reference sequence to design the oligonucleotide chip, while 454 and Illumina can sequence new genomes. The TPA Nichols sequence (AE000520.1) was chosen as the reference for TPE Samoa D CGS sequencing. Because CGS identifies only heterogenous loci between two genomic DNAs and because it was found later that AE00520.1 contains about 200 single nucleotide errors (Giacani *et al.*, 2012a; Šmajs *et al.*, 2012), the sequencing errors of the reference genome can be adapted to a query genome sequence as in case of TPA SS14 strain (CP000805.1) (Matějková *et al.*, 2008). Other than these types of errors, the major errors of the CGS

technique were detected in single nucleotide substitutions. In addition, the CGS method cannot detect larger insertions. The major disadvantages of the 454 pyrosequencing technique was found to be incorporated indels, especially insertions within or close to homopolymeric tracts. In total, 11.97 kb (1.1%) of the TPA Nichols genome (AE000520.1) correspond to regions containing more than 6 homopolymers in a tract. The Illumina technique showed low error rates with an even distribution between indels and single nucleotide changes. Despite the high depth of coverage of Illumina reads, the length coverage was lower than in 454 pyrosequencing: the number of Illumina gaps were tripled compared to the number of 454 gaps.

Every NGS method has pros and cons. The combination of several techniques leads to high-quality complete genome sequence. Even then, several loci in a genome typically need further sequencing by the DDT technique, in order to finish paralogous regions, loci with variable numbers of tandem repeats and variable numbers of homopolymers within a homopolymeric tract. Both NGS methods, 454 and Illumina, were also applied for other treponemal genomes, including TPA DAL-1 (Zobaníková *et al.*, 2012), re-sequenced TPA Nichols, re-sequenced SS14 (Pětrošová, unpublished results), and TPc Cuniculi A (Šmajs *et al.*, 2011) strains. Only the TPA Mexico A strain was sequenced by the Illumina technique alone (Pětrošová *et al.*, 2012).

The synteny and GC content for all genomes was identical, and genome size was very similar to other sequenced genomes of pathogenic treponemes (Šmajs *et al.*, 2012). The difference between the smallest and largest TPE genomes is 414 bp, while the maximal difference within TPA strains is 2027-bp difference between Mexico A and Nichols (AE000520.1) strains (Fraser *et al.*, 1998; Pětrošová *et al.*, 2012). However, only a 757-bp maximal difference between Chicago and Mexico A genomes has been confirmed based on high-quality NGS and/or whole genome fingerprinting (WGF) techniques (Giacani *et al.*, 2010a; Mikalová *et al.*, 2010; Strouhal *et al.*, 2007). When compared to the published Nichols genome (AE000520.1), the re-sequenced version contains an additional 420 bp in the *arp* gene (other 7 tandem repeats of 60 bp) and an insertion of 1204 bp between the TP0126 and TP0127 genes (Strouhal *et al.*, 2007). The latter insertion is similar to the *tpr*K gene and was observed only in a subpopulation of the Nichols strain (Šmajs *et al.*, 2002; Strouhal *et al.*, 2007). The simian isolate Fribourg-Blanc had a similar genome size to the TPA and TPE genomes (Mikalová *et al.*, 2010), while *T. paraluiscuniculi* (TPc) strain revealed a genome reduction, probably as a result of adaptation to a rabbit host (Šmajs *et al.*, 2011).

Due to increasing evidence of sequencing errors in the TPA Nichols genome (AE000520.1) (Matějková *et al.*, 2008; Šmajs *et al.*, 2011) and the availability of improved bioinformatics tools, the TPE genomes were annotated *de novo*. In addition, functions have been assigned for several hypothetical proteins since the TPA Nichols genome was first analyzed (Fraser *et al.*, 1998), including proteins binding to extra-cellular matrix (ECM) components (Cameron, 2003; Cameron *et al.*, 2004), transporters (Desrosiers *et al.*, 2007), regulators (Brett *et al.*, 2008; Giacani *et al.*, 2009), flagellar proteins (Titz *et al.*, 2006), and outer or periplasmic proteins (Brinkman *et al.*, 2008; Desrosiers *et al.*, 2011; Hazlett *et al.*, 2005; Cha *et al.*, 2004). All of the new functions were incorporated in the re-annotation of TPE genomes.

The genome annotations confirmed a high degree of sequence homology among pathogenic treponemes. No large rearrangements or acquired or lost gene material were identified when the new annotation was compared to the AE000520.1 annotation. However, 50 Nichols (AE000520.1) genes were fused into 24 TPE orthologues. Re-sequencing of the Nichols strain revealed that 46 Nichols genes (4.43% of all annotated genes) were improperly annotated due to sequencing errors. Thus, a relatively low error rate at the DNA level can result in a broader effect at the proteome level and have a greater impact on subsequent proteomic studies (Brinkman *et al.*, 2006; McKevitt *et al.*, 2003; McKevitt *et al.*, 2005; Titz *et al.*, 2008).

## 5.2. Intra-strain variability and comparative genomics of TPE strains

With the exception of the *tpr*K (TPE_0897) and *tpr*D (TPE_0131) genes, a total of 190 differences were found between individual TPE strains, 4 multiple rearrangements of single nucleotide exchanges and indels, 1 translocation, 20 indels and 160 single nucleotide exchanges. Whereas multiple rearrangements were observed only in the coding regions, the translocation affected genes coding for tRNA and nearby regions. About 4.38% of single nucleotide exchanges and 40% of indels were located in the intergenic regions. Since the intergenic regions comprise about 4.6% of the TPE genome size, the biased percentage of indels may indicate some effect on the expression of TPE (or *T. pallidum*) genes (Radolf & Desrosiers, 2009).

The highest intra-strain heterogeneity in the TPE strains was observed in the *tpr*K gene. The *tpr*K gene also showed high heterogeneity among laboratory TPA strains (Giacani *et al.*, 2012b; Stamm & Bergen, 2000a), and during the course of human (LaFond *et al.*, 2003) and experimental infection (LaFond *et al.*, 2006). Moreover, higher heterogeneity was observed among than within TPA isolates (LaFond *et al.*, 2003). Indels and nucleotide substitutions were detected in the TPA strains in seven discrete variable loci, although no TPA allelic variant disrupting an open reading frame has been identified (Stamm & Bergen, 2000a). The heterogeneity is likely mediated by segmental non-reciprocal homologous recombination, defined as gene conversion (Centurion-Lara *et al.*, 2004). The donor sites for gene conversion were predicted in the 5' and 3' flanking regions of the *tpr*D (TP0131) gene, including the *tpr*K-like region. The ability of the pathogens to continuously modify genes coding for outer membrane antigens allows them to circumvent or at least postpone the host immune response (Vink *et al.*, 2011). It has been shown that variable regions of the *tpr*K gene encode epitopes (Giacani *et al.*, 2010b), although outer membrane localization of TprK has not been confirmed (Cox *et al.*, 2010; Hazlett *et al.*, 2001). It is known that syphilis infection may persist within hosts, thus it is likely that antigenic variation of TprK helps to evade the humoral immune response.

Gene conversion mechanism has been described in other known pathogens (Vink *et al.*, 2011), including *Mycoplasma pneumoniae* (the agent of respiratory infections; gene conversion of the major adherence protein) (Spuesens *et al.*, 2009; Spuesens *et al.*, 2011), *Borrelia hermsii* (relapsing fever; antigenic lipoproteins) (Dai *et al.*, 2006), *Borrelia burgdorferi* (Lyme disease; outer membrane lipoprotein) (Zhang & Norris, 1998) and *Neisseria gonorrhoeae* (gonorrhea; pilin proteins) (Meyer *et al.*, 1982; Zhang *et al.*, 1992). The gene conversion of pilin genes in *Neisseria* is mediated by the *rec*F recombination pathway (Sechman *et al.*, 2005). In addition, a promoter sequence upstream of the *pil*E gene forms a guanine quadruplex which is important for antigenic variation of this pilin gene (Cahoon & Seifert, 2009). The guanine quadruplex (G4 DNA) is a G-rich DNA sequence able to form a four-stranded structure (Sen & Gilbert, 1988). G4 DNAs have been found in promoters (Duquette *et al.*, 2004; Simonsson *et al.*, 1998) and in repetitive sequences, including telomeres (Parkinson *et al.*, 2002), rDNAs (Hanakahi *et al.*, 1999), G-rich minisatelites (Weitzmann *et al.*, 1997), and immunoglobulin heavy-chain switch regions (Dunnick *et al.*, 1993). While G4 DNAs are resistant to the attack of endonucleases, both prokaryotic and eukaryotic cells can unwind G4 DNA by a class of RecQ helicases (Bachrati & Hickson, 2003), enzymes essential for genome stability. In addition to genome stability

control, G4 DNA is believed to regulate gene expression in prokaryotic (Rawal *et al.*, 2006) and eukaryotic cells (Simonsson *et al.*, 1998). Parallel to *N. gonorrhoeae*, the *tpr*K promoter region in pathogenic treponemes is also able to assemble into a guanine quadruplex (Giacani *et al.*, 2012b). Moreover, the *rec*F pathway has been recognized in TPA and TPE genomes (Čejková *et al.*, 2012; Fraser *et al.*, 1998; Pětrošová *et al.*, 2012). Of interest, the TPA_0103 genes coding for RecQ helicase are truncated at 3' end in the TPA genomes, while TPE_0103 othologous genes code for intact protein in the TPE genomes.

The highest inter-strain heterogeneity between TPE strains was observed in the *tpr*D gene. While Samoa D and CDC-2 carry allele *tpr*D2, the Gauthier strain carries allele *tpr*D3 in the TPE_0131 locus (Centurion-Lara *et al.*, 2000b). The Gauthier strain carries allele *tpr*D3 also in the TPE_0117 locus (*tpr*C) and gene conversion of the *tpr*C allele onto the *tpr*D locus has been postulated (Gray *et al.*, 2006). In addition to Gauthier, other non-TPA strains, TPE Samoa D and CDC-2, simian Fribourg-Blanc and TEN Iraq B, harbor a *tpr*D3 or *tpr*D3-like allele at the *tpr*C locus (Gray *et al.*, 2006).

Among other members of the paralogous *tpr* family, slight variability between TPE strains was also observed in the *tpr*C, *tpr*E (TPE_0313), *tpr*F (TPE_0316), *tpr*G (TPE_0317), *tpr*I (TPE_0620), *tpr*J (TPE_0621) and *tpr*L (TPE_1031) genes. Phylogenetic analysis using *tpr*C and *tpr*I loci showed clear differentiation between TPA strains and other human non-TPA strains, although recombination events were predicted (Gray *et al.*, 2006). TprC, TprD, and TprI, proteins with predicted porin activity, are supposed to be anchored in the outer membrane (Anand *et al.*, 2012; Giacani *et al.*, 2005b). Interestingly, the *tpr*F gene carries a frameshift mutation in the TPA strains (Fraser *et al.*, 1998; Giacani *et al.*, 2005a; Giacani *et al.*, 2010a; Pětrošová *et al.*, 2012; Zobaníková *et al.*, 2012) while full ORFs were annotated in the TPE strains (Čejková *et al.*, 2012). In addition to the *tpr*F gene, an authentic frameshift mutation of the *tpr*A (TP0009) gene is present in TPA but not in TPE strains (Fraser *et al.*, 1998; Giacani *et al.*, 2005a; Giacani *et al.*, 2010a; Pětrošová *et al.*, 2012; Zobaníková *et al.*, 2012). A rabbit TPc Cuniculi A genome contains only 4 intact *tpr* genes (*tpr*A, *tpr*B, *tpr*H, and t*pr*L) (Šmajs *et al.*, 2011).

Within the TPE strains, a variable number of nucleotides in homopolymeric tracts was found upstream of *tpr*E, *tpr*F, *tpr*G, *tpr*I and *tpr*J genes. Among the TPE strains, a variable number of nucleotides in homopolymeric tracts was observed upstream of *tpr*C, *tpr*D, *tpr*G, and *tpr*I. Similar findings were also reported by Giacani *et al.* (Giacani *et al.*, 2005a). In general, variability is mediated by slipped-strand mispairing during DNA replication (van der

Ende *et al.*, 1995) and bacteria carrying this feature can alter the gene expression of downstream gene(s) (Torres-Cruz & van der Woude, 2003). Whereas one bacterial subpopulation expresses the gene(s), another subpopulation within the host does not express it, resulting in phase variation (Jonsson *et al.*, 1991). The phase variation is considered to be a virulence mechanism because phase variation of genes coding for surface exposed proteins have been mainly observed in pathogenic bacteria (van der Woude & Baumler, 2004). In pathogenic treponemes, phase variations of *tpr*E, *tpr*G and *tpr*J genes were confirmed experimentally (Giacani *et al.*, 2007a).

Both intra-strain and inter-strain heterogeneity within homopolymeric tracts was also found upstream of the TPE_0126 genes resulting in the extension of the open reading frame (ORF) in Samoa D by 705 bp (TPESAMD_0126), harboring a homopolymeric tract within an ORF. The phase variation was experimentally confirmed and a shorter protein variant with a homopolymeric string located upstream (as in Gauthier and CDC-2) was predicted using a reporter study (Giacani *et al.*, 2012b). Moreover, the shorter protein is homologous to *E. coli* OmpW, an outer membrane protein (Giacani *et al.*, 2012b).

The *arp* gene (TPE_0433; coding for acidic repeat protein) and TPE_0470 (coding for conserved hypothetical protein) were found among the highly variable genes in TPE strains. Both *arp* and TPE_0470 genes harbor tandem repeats of 60 bp an 24 bp, respectively. The number of tandem repeats in the *arp* gene varies between 4 and 12, while the TPE_0470 gene has between 12 and 37 repeats. Both genes are also variable among other treponemal strains (Harper *et al.*, 2008a; Mikalová *et al.*, 2010; Strouhal *et al.*, 2007). In addition, two tandem repeats of 9 bp were deleted in TPESAMD_0967 (coding for a treponemal conserved hypothetical protein), whereas CDC-2 and Gauthier orthologues carry 3 complete and one incomplete tandem repeats. Again, the different number of tandem repeats is considered to be a virulence mechanism that helps the pathogen escape from host immune pressure (Coil *et al.*, 2008). Both proteins were found to be immunogenic in TPA strains (Brinkman *et al.*, 2006; Liu *et al.*, 2007; McKevitt *et al.*, 2005). Interestingly, the Arp protein binds to fibronectin (Liu *et al.*, 2007) and different repeat motifs in the *arp* gene were observed in venereal treponemes, while non-venereal treponemes carried only one repeat motif (Harper *et al.*, 2008a; Liu *et al.*, 2007). The TPE_0470 protein contains a tetratricopeptide repeat domain, a domain that is usually responsible for protein-protein interactions (Das *et al.*, 1998).

Additionally, two TPE genes (TPE_0067, TPE_0136) carry an insertion terminated by direct repeats. Interestingly, both open reading frames are disrupted by frameshift mutations

in the TPA DAL-1 strain. TPE_0067 codes for a cell division protein (Titz *et al.*, 2008), and TPE_0136 encodes a treponemal conserved hypothetical outer membrane protein (Brinkman *et al.*, 2008). The TPA_0136 orthologous gene was highly expressed (Šmajs *et al.*, 2005), coding for lipidated (Setubal *et al.*, 2006) antigen (McKevitt *et al.*, 2005) with binding affinity to fibronectin and laminin (Brinkman *et al.*, 2008), which is located in the periplasmic space (Cox *et al.*, 2010). The gene is highly heterogenous between treponemal strains and is a candidate for molecular typing of pathogenic treponemes (Flasarová *et al.*, 2006).

In total, three genes were found to be under positive selection among TPE strains including *tp*92 (TPE_0326; coding for an outer membrane protein), *mcp*2 (TPE_0488; methyl-accepting chemotaxis protein) and TPE_0548 (treponemal conserved hypothetical membrane protein). The highly heterogenous TPA_0548 gene has been applied in molecular typing studies (Flasarová *et al.*, 2006; Flasarová *et al.*, 2012; Matějková *et al.*, 2009; Woznicová *et al.*, 2007). Although the protein contains a transmembrane domain, the cellular localization has not yet been determined (Cox *et al.*, 2010). Both *tp*92 and *mcp*2 genes were under positive selection within TPE and TPA strains and in the TPA-TPE comparison. However, the TPA-TPc comparison showed purifying and neutral constraints, respectively. Interestingly, both genes carry a mosaic character in the TPA Mexico A genome (Pětrošová *et al.*, 2012). The authors postulated an inter-strain homologous recombination event between the TPA and TPE strain. Tp92 is a highly immunogenic protein (Brinkman *et al.*, 2006; Van Voorhis *et al.*, 2003). The outer membrane localization predicted due to opsonic activity (Cameron *et al.*, 2000) was recently confirmed by a protease-surface accessibility assay (Desrosiers *et al.*, 2011) and a surface immunolabeling technique (Cox *et al.*, 2010). Moreover, Desrosiers *et al.* (Desrosiers *et al.*, 2011) demonstrated a Tp92 influence in outer membrane biogenesis. Mcp2 contains a protease-like domain which is homologous to the PrtB protease of *T. denticola* (Arakawa & Kuramitsu, 1994; Seshadri *et al.*, 2004). Similar chemotaxis proteins were found in other pathogenic bacteria, including *Bacillus anthracis* (Read *et al.*, 2003) and *Clostridium botulinum* (Sebaihia *et al.*, 2007). High expression (Šmajs *et al.*, 2005), positive selection and proteolytic activity might indicate a virulence function for Mcp2 in *T. pallidum*.

Fifteen additional genes (0086, 0143, 0346, 0462, 0506, 0515, 0556, 0559, 0698, 0729, 0771, 0831, 0865, 0968, 0993) were found to be under positive selection in the TPA-TPE gene comparison (Čejková *et al.*, 2012). Most of these genes code for hypothetical proteins or proteins involved in bacterial virulence. Of interest, several of these hypothetical proteins showed a high transcription rate in the TPA Nichols strain (Šmajs *et al.*, 2005)

including TP0086 (conserved hypothetical protein, transcription rate 2.497), TP0346 (CHP, 2.874), TP0462 (CHP, 4.894), TP0698 (hypothetical membrane protein, 1.813), TP0968 (treponemal conserved hypothetical protein, 3.256) and TP0993 (rare lipoprotein A, 2.487). An increased gene expression and positive selection constraint thus lead us to the suggestion of fully functional hypothetical proteins involved in pathogenesis of venereal treponemes (Čejková *et al.*, 2012; Šmajs *et al.*, 2012).

In summary, many differences between TPE genomes involve a repetitive motif, either in a homopolymeric tract, or in tandem or dispersed repeats. Thus it is likely that a polymerase slippage mechanism (Levinson & Gutman, 1987) has been a major force in the evolution of TPE genomes and phase variation could be a powerful tool in the combat with the host immune response. Membrane proteins and proteins involved in transport, chemotaxis and cell division were found among the highly heterogenous proteins in TPE genomes. Considering their importance in tissue penetration, substrate acquisition and propagation of the pathogen, these genes might represent virulence factors in pathogenic non-cultivable treponemes (Weinstock *et al.*, 1998).

The nucleotide divergence ($D_{xy}$) between TPA and TPE subspecies was 4-6 times higher, respectively, than the nucleotide diversity ($\pi$) within TPA and TPE strains. Based on the genome sequence comparison of 3 TPE (Samoa D, CDC-2, Gauthier) and 4 TPA (re-sequenced Nichols, re-sequenced SS14, DAL-1, Chicago) strains, a 99.8% identity between TPA and TPE was calculated (Čejková *et al.*, 2012). Only high-quality complete genomes can reveal differences between such genetically monomorphic bacteria (Achtman, 2008). Thus, a combination of next-generation sequencing techniques with additional DDT sequencing was a necessary methodology to obtain complete genome seguences for comparative genomics of pathogenic treponemes.

### 5.3. Structure and reciprocal translocation of *rrn* operons in pathogenic treponemes

In total, *rrn* operons were examined in 20 pathogenic treponemal strains. Additionally, 30 clinical samples were tested for the presence of treponemal DNA (Flasarová *et al.*, 2012) and *rrn* operon patterns were determined for positive samples. All investigated strains contained two copies of *rrn* operons. Two *rrn* operons with the same composition have also been described in other human and animal treponemes, except for *T. vincentii*, which contains

only 1 *rrn* operon (Fraser *et al.*, 1998; Matějková *et al.*, 2008; Seshadri *et al.*, 2004; Stamm *et al.*, 2002).

Our results confirmed little diversity within genes coding for rRNA and within intergenic spacer regions (ISR). However, our data showed that *rrn* operon structure displays blocks of conserved and polymorphic sites. The TPA DAL-1 strain showed a 1-bp deletion upstream of the 16S rDNA gene in the *rrn*1 operon. It is known that the TPA DAL-1 strain grows more rapidly in rabbits than other pathogenic strains (Wendel *et al.*, 1991) and it is possible that a different promoter DNA conformation could affect the expression of the *rrn*1 operon.

Gurtler & Stanisich (1996) used the 16S-23S ISR for classification of bacteria. The 16S-23S ISRs have been used for this purpose in several studies (de Vries *et al.*, 2006; Lan & Reeves, 1998; Lebuhn *et al.*, 2006), including treponemal (Centurion-Lara *et al.*, 1996; Stamm *et al.*, 2002) and borrelian samples (Bunikis *et al.*, 2004; Comstedt *et al.*, 2009). Centurion-Lara *et al.* (1996) examined TPA Nichols and TPE Gauthier strains, and found no difference. However, they did not examine the genomic positions of individual 16S-23S ISRs. Interestingly, the 16S-23S ISR typing of *Borrelia burgdorferi* strains agrees with the *osp*C gene typing system (Hanincová *et al.*, 2008; Wormser *et al.*, 2008). The *osp*C gene, coding for a protein involved in the initiation of infection in warm-blood animals, is located on plasmid DNA while the *rrn* operon is on chromosomal DNA. Different 16S-23S ISR genotypes have been associated with different degrees of invasivity due to their strong linkage disequilibrium with *osp*C genotypes (Wormser *et al.*, 2008).

Despite the low heterogeneity of *rrn* operons, two different intergenic spacer patterns were observed in pathogenic treponemal samples. Whereas, detection of specific nucleotide changes may be of interest in identification of treponemal diseases, the detection of the tRNA-Ile and tRNA-Ala coding gene arrangement in the 16S-23S ribosomal intergenic spacer appears to be of limited use for typing of clinical samples. All clinical samples from the Czech Republic showed the Ile/Ala spacer pattern in *rrn* operons (data not shown).

Due to the conserved machinery for protein synthesis, rDNA genes are expected to be under strong purifying selection, and are exposed to the intragenomic homogenization process via gene conversion (Liao, 2000; Nei & Rooney, 2005). Several studies (Acinas *et al.*, 2004; Pei *et al.*, 2009; Pei *et al.*, 2010) have shown that homogenization of multiple rDNA genes is very common among bacteria. In addition, Harvey & Hill (1990) successfully constructed several *E. coli* strains with recombined inverted *rrn* operons and found that the recombinants tended to recover the original configuration. The *rrn* operons of treponemal strains are direct

repeats: the tRNA-Ala gene is replaced by tRNA-Ile (and *vice versa*), and the recombination is a common event with no correlation to the otherwise-determined phylogenetic relationship among tested treponemes. It has been postulated that recombination between direct repeats leads to the duplication or deletion of a repeat (Petes & Hill, 1988; Petit, 2005). Whereas the tRNA-Ile (TP_t12) is a unique gene in sequenced treponemal genomes, there are three predicted tRNA-Ala genes (TP_t15, TP_t41, TP_t45; (Fraser *et al.*, 1998). Since both the tRNA-Ile (TP_t12, AE000520.1) and tRNA-Ala (TP_t15, AE000520.1) genes need to be maintained in the genomes of pathogenic treponemes, reciprocal translocation, rather than gene conversion, appears to be the mechanism for the observed *rrn* heterogeneity among tested strains. Such a process would require double cross-overs in both *rrn* operons, and therefore is much less common than insertion/deletion or gene conversion events (Harvey & Hill, 1990; Hashimoto *et al.*, 2003). Predicted recombination hot spot sites are located in the 16S and 23S rDNA genes, genes with two identical copies within every strain examined in our study.

During replication of direct repeat regions, DNA polymerase might lead to strand slippage, thus collapsing a replication fork formation. To continue in the DNA replication, recombination enzymes are involved in DNA repair mechanism (Darling *et al.*, 2008; Santoyo & Romero, 2005). Although only the *rec*F recombination pathway was originally predicted in the TPA Nichols genome (Fraser *et al.*, 1998), the *rec*F pathway favors the gene conversion mechanism (Kobayashi, 1992; Takahashi *et al.*, 1992). The reciprocal recombination in pathogenic treponemes may be accompanied by crossing-over, a repair mechanism implemented by the *rec*BCD pathway in *E. coli* (Kobayashi, 1992). Recently, *rec*BCD orthologs (*add*A and *add*B) were predicted for several investigated treponemal genomes (Čejková *et al.*, 2012; Giacani *et al.*, 2012a; Šmajs *et al.*, 2011), and were composed of TP0898 and fused TP0899-0900 genes. It is therefore likely that these genes are responsible for this recombination. However, it would be extremely difficult to experimentally prove the *rec*BCD-mediated crossing-over mechanism in *T. pallidum.*

### 5.4. Comparative phenomics applied to the BAC library of TPA Nichols DNA

In total, 1342 different phenotype asssays were tested on 19 clones of the BAC library of TPA Nichols DNA using the Phenotype MicroArray (PM) technology. Overall, 190 substrates were tested as sole carbon sources and 192 pH and osmotic sensitivity assays and

960 antibiotic resistance assays were performed. The data showed high reproducibility, except for a few of the examined cellular phenotypes (data not shown), which were excluded from further analyses.

Two slow growing BAC clones (DSTP094 and DSTP198) were impossible to compare with other clones even when the incubation time was extended. These clones, containing TP0538-0595 and TP0628-0686 Nichols genes, turned off almost all alternative metabolic pathways for carbon utilization. Other four BAC clones did not utilize several carbon sources which the other BAC clones and the control host *E. coli* cells did. Although it is known that TPA strains can utilize only glucose as a sole carbon source (Nichols & Baseman, 1975; Schiller & Cox, 1977), the growth rate is likely the key factor of the observed reduced phenotypes in the host *E. coli* clones

Effective treatment is the crucial step to control any disease, including syphilis. Although penicillin is the drug of choice for syphilis patients (Workowski & Berman, 2006), it cannot cross the blood-brain barrier (Marra, 2009). People allergic to penicillin need to be treated with alternative drugs, such as macrolides (erythromycin and azithromycin), cephalosporins (ceftriaxone), lincomycin, and tetracyclines (tetracycline and doxycycline) (Workowski & Berman, 2006). It is of interest that DSTP001 clone showed an increased resistance to lincomycin and cephalosporins when compared to other BAC library clones and to the host *E. coli* strain. After applying these drugs, several failures in treatment of treponemal infections have been reported (Augenbraun & Workowski, 1999; Duncan & Knox, 1971; Woznicová *et al.*, 2007).

Cephalosporins are ß-lactam antibiotics disrupting the cell wall synthesis by inhibiting the penicillin-binding proteins (Zapun *et al.*, 2008). Several strategies have been developed in bacterial cells to interrupt the effect of ß-lactam antibiotics: (i) alternation of penicillin-binding proteins to produce proteins with reduced affinity to penicillin; (ii) production of ß-lactamase; (iii) structural changes in porins proteins to prevent penicillin uptake; and (iv) increased efflux system to reduce the antibiotic concentration within the cell (Zapun *et al.*, 2008). The antibiotic resistance is widely acquired through horizontal gene transfer. Neither penicillin-binding proteins (TP0500, TP0574, TP0705, TP0760), nor proteins with ß-lactamase activity (TP0489, TP0574, TP0705) are located in the DSTP0001 clone containing the TP0084-0137 genes (Čejková *et al.*, 2012; Cha *et al.*, 2004; Weigel *et al.*, 1994). Moreover, horizontal gene transfer has not been reported in the TPA genomes (Fraser *et al.*, 1998; Giacani *et al.*, 2010a; Matějková *et al.*, 2008; Zobaníková *et al.*, 2012). The reason for increased resistance to the antibiotics thus remains unknown.

Lincomycin binds to the bacterial 50S ribosome subunit, preventing bacterial protein synthesis. The mechanism of action is thus similar to that of macrolides: however, lincomycin and macrolides are not related to each other (Stamm, 2010). The resistance to lincomycin could be acquired due to (i) incorporation of a novel mutation or methylation in the 23S rDNA gene, and/or (ii) an increased efflux system (Leclercq, 2002). The macrolide resistant treponemes have been reported worldwide (Martin *et al.*, 2009; Stamm *et al.*, 1988; Woznicová *et al.*, 2010) and two point mutations (A2058G and A2059G) in the 23S rDNA gene were identified as being associated with macrolide resistance in treponemes (Matějková *et al.*, 2009; Stamm & Bergen, 2000b). It was shown that both point mutations can cause lincomycin resistance in other spirochetes (Karlsson *et al.*, 2004; Karlsson *et al.*, 1999). Of note, the TPA SS14 strain, carrying an A2058G mutation, is more resistant to lincomycin than the TPA Nichols strain (Stamm & Bergen, 2000b; Stamm *et al.*, 1988). However, the Nichols strain also showed partial lincomycin resistance (Stamm *et al.*, 1988) and the DSTP001 clone of Nichols DNA library does not carry the gene coding for 23S rRNA.

The genetic basis accounting for the increased resistance observed in BAC clones needs to be further examined. To validate results from PM panels, the minimum inhibitory concentration for drugs of interest should be determined in the BAC clones with elevated resistance alongside other BAC clones. The minimal clone set of the BAC library, examined in the PM assay, contains 19 clones with an average of 54 intact TPA Nichols genes. However, the whole BAC library consists of 334 clones. Other clones (called “subclones”) contain smaller fragments of the Nichols chromosome than the clones in the minimal set. The minimum inhibitory concentration should be determined also in the corresponing “subclones” to find genes responsible for elevated resistance to these antibiotics.

# 6. CONCLUSIONS

- The high-quality genome sequences of the TPE CDC-2 and Gauthier strains have been compiled by a combination of 454, Illumina and DDT sequencing techniques.
- The whole genome sequence of TPE Samoa D was annotated *de novo*. Using the Samoa D gene pattern, gene annotation has been adapted for the TPE CDC-2 and Gauthier, and TPA DAL-1 genomes, as well. When necessary, manual annotation was performed for several loci.
- The intra- and inter-strain variability was estimated between and within the TPE Samoa D, CDC-2 and Gauthier strains.
- Two different *rrn* spacer patterns (Ile/Ala and Ala/Ile) were determined to be randomly distributed across the time and place of original isolation of treponemal strains. The random distribution of tRNA coding genes is likely caused by reciprocal translocation between repetitive sequences mediated by a *rec*BCD-like system.
- High-throughput metabolic profiles of individual TPA Nichols BAC clones were compared. Three BAC clones showed higher resistance to six antibiotics, which will be subject to further examination.

# 7. REFERENCES

.

**Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. (2004).** Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *Journal of bacteriology* **186**, 2629-2635.

**Achtman, M. (2008).** Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol* **62**, 53-70.

**Akrawi, F. (1949).** Is bejel syphilis? *Br J Vener Dis* **25**, 115-123.

**Albert, T. J., Dailidiene, D., Dailide, G., Norton, J. E., Kalia, A., Richmond, T. A., Molla, M., Singh, J., Green, R. D. & other authors (2005).** Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nat Methods* **2**, 951-953.

**Amin, R., Sattar, A., Basher, A. & Faiz, M. A. (2010).** Eradication of yaws. *Journal of Clinical Medicine and Research* **2**, 049-054.

**Anand, A., Luthra, A., Dunham-Ems, S., Caimano, M. J., Karanian, C., LeDoyt, M., Cruz, A. R., Salazar, J. C. & Radolf, J. D. (2012).** TprC/D (Tp0117/131), a trimeric, pore-forming rare outer membrane protein of *Treponema pallidum*, has a bipartite domain structure. *Journal of bacteriology* **194**, 2321-2333.

**Andersson, B., Wentland, M. A., Ricafrente, J. Y., Liu, W. & Gibbs, R. A. (1996).** A "double adaptor" method for improved shotgun library construction. *Anal Biochem* **236**, 107-113.

**Antal, G. M. & Causse, G. (1985).** The control of endemic treponematoses. *Rev Infect Dis* **7 Suppl 2**, S220-226.

**Antal, G. M., Lukehart, S. A. & Meheus, A. Z. (2002).** The endemic treponematoses. *Microbes Infect* **4**, 83-94.

**Arakawa, S. & Kuramitsu, H. K. (1994).** Cloning and sequence analysis of a chymotrypsinlike protease from *Treponema denticola*. *Infect Immun* **62**, 3424-3433.

**Asiedu, K., Amouzou, B., Dhariwal, A., Karam, M., Lobo, D., Patnaik, S. & Meheus, A. (2008).** Yaws eradication: past efforts and future perspectives. *Bull World Health Organ* **86**, 499-499A.

**Augenbraun, M. & Workowski, K. (1999).** Ceftriaxone therapy for syphilis: report from the emerging infections network. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* **29**, 1337-1338.

**Bachrati, C. Z. & Hickson, I. D. (2003).** RecQ helicases: suppressors of tumorigenesis and premature aging. *Biochem J* **374**, 577-606.

**Baker-Zander, S. A. & Lukehart, S. A. (1983).** Molecular basis of immunological cross-reactivity between *Treponema pallidum* and *Treponema pertenue*. *Infect Immun* **42**, 634-638.

**Baker-Zander, S. A. & Lukehart, S. A. (1984).** Antigenic cross-reactivity between *Treponema pallidum* and other pathogenic members of the family *Spirochaetaceae*. *Infect Immun* **46**, 116-121.

**Bamford, C. V., Francescutti, T., Cameron, C. E., Jenkinson, H. F. & Dymock, D. (2010).** Characterization of a novel family of fibronectin-binding proteins with M23 peptidase domains from Treponema denticola. *Mol Oral Microbiol* **25**, 369-383.

**Baseman, J. B., Nichols, J. C., Rumpp, J. W. & Hayes, N. S. (1974).** Purification of *Treponema pallidum* from Infected Rabbit Tissue: Resolution into Two Treponemal Populations. *Infect Immun* **10**, 1062-1067.

**Bayon, H. (1913).** A NEW SPECIES OF *TREPONEMA* FOUND IN THE GENITAL SORES OF RABBITS. *Br Med J* **2**, 1159.

**Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004).** Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

**Bennett, S. (2004).** Solexa Ltd. *Pharmacogenomics* **5**, 433-438.

**Blanco, D. R., Walker, E. M., Haake, D. A., Champion, C. I., Miller, J. N. & Lovett, M. A. (1990).** Complement activation limits the rate of in vitro treponemicidal activity and correlates with antibody-mediated aggregation of *Treponema pallidum* rare outer membrane protein. *J Immunol* **144**, 1914-1921.

**Bochner, B. R. (2009).** Global phenotypic characterization of bacteria. *Fems Microbiology Reviews* **33**, 191-205.

**Bochner, B. R. & Savageau, M. A. (1977).** Generalized indicator plate for genetic, metabolic, and taxonomic studies with microorganisms. *Appl Environ Microbiol* **33**, 434-444.

**Bochner, B. R., Gadzinski, P. & Panomitros, E. (2001).** Phenotype MicroArrays for high-throughput phenotypic testing and assay of gene function. *Genome Research* **11**, 1246-1255.

**Bourell, K. W., Schulz, W., Norgard, M. V. & Radolf, J. D. (1994).** *Treponema pallidum* rare outer membrane proteins: analysis of mobility by freeze-fracture electron microscopy. *J Bacteriol* **176**, 1598-1608.

**Brett, P. J., Burtnick, M. N., Fenno, J. C. & Gherardini, F. C. (2008).** *Treponema denticola* TroR is a manganese- and iron-dependent transcriptional repressor. *Mol Microbiol* **70**, 396-409.

**Brinkman, M. B., McKevitt, M., McLoughlin, M., Perez, C., Howell, J., Weinstock, G. M., Norris, S. J. & Palzkill, T. (2006).** Reactivity of antibodies from syphilis patients to a protein array representing the *Treponema pallidum* proteome. *J Clin Microbiol* **44**, 888-891.

**Brinkman, M. B., McGill, M. A., Pettersson, J., Rogers, A., Matějková, P., Šmajs, D., Weinstock, G. M., Norris, S. J. & Palzkill, T. (2008).** A novel *Treponema pallidum* antigen, TP0136, is an outer membrane protein that binds human fibronectin. *Infect Immun* **76**, 1848-1857.

**Bunikis, J., Garpmo, U., Tsao, J., Berglund, J., Fish, D. & Barbour, A. G. (2004).** Sequence typing reveals extensive strain diversity of the Lyme borreliosis agents *Borrelia burgdorferi* in North America and *Borrelia afzelii* in Europe. *Microbiology* **150**, 1741-1755.

**Cahoon, L. A. & Seifert, H. S. (2009).** An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*. *Science* **325**, 764-767.

**Cameron, C. E. (2003).** Identification of a *Treponema pallidum* laminin-binding protein. *Infect Immun* **71**, 2525-2533.

**Cameron, C. E. (2006).** *T. pallidum* outer membrane and outer membrane proteins. In *Pathogenic Treponema: Molecular and Cellular Biology*, pp. 237-284. Edited by J. D. Radolf & S. Lukehart. Norfolk, England: Caister Academic Press.

**Cameron, C. E., Castro, C., Lukehart, S. A. & Van Voorhis, W. C. (1999).** Sequence conservation of glycerophosphodiester phosphodiesterase among *Treponema pallidum* strains. *Infect Immun* **67**, 3168-3170.

**Cameron, C. E., Brown, E. L., Kuroiwa, J. M., Schnapp, L. M. & Brouwer, N. L. (2004).** *Treponema pallidum* fibronectin-binding proteins. *J Bacteriol* **186**, 7019-7022.

**Cameron, C. E., Lukehart, S. A., Castro, C., Molini, B., Godornes, C. & Van Voorhis, W. C. (2000).** Opsonic potential, protective capacity, and sequence conservation of the *Treponema pallidum* subspecies *pallidum* Tp92. *J Infect Dis* **181**, 1401-1413.

**Cameron, C. E., Kuroiwa, J. M., Yamada, M., Francescutti, T., Chi, B. & Kuramitsu, H. K. (2008).** Heterologous expression of the *Treponema pallidum* laminin-binding adhesin Tp0751 in the culturable spirochete *Treponema phagedenis*. *J Bacteriol* **190**, 2565-2571.

**Capuano, C. & Ozaki, M. (2011).** Yaws in the Western pacific region: a review of the literature. *Journal of tropical medicine* **2011**, 642832.

**Castellani, A. (1905).** FURTHER OBSERVATIONS ON PARANGI (YAWS). *Brit Med Jour*, 1330-1331.

**Castellani, A. (1907).** Experimental Investigations on Framboesia Tropica (Yaws). *J Hyg (Lond)* **7**, 558-569.

**Čejková, D., Zobaníková, M., Chen, L., Pospíšilová, P., Strouhal, M., Qin, X., Mikalová, L., Norris, S. J., Muzny, D. M. & other authors (2012).** Whole Genome Sequences of Three *Treponema pallidum* ssp. *pertenue* Strains: Yaws and Syphilis Treponemes Differ in Less than 0.2% of the Genome Sequence. *PLoS neglected tropical diseases* **6**, e1471.

**Centurion-Lara, A., Castro, C., van Voorhis, W. C. & Lukehart, S. A. (1996).** Two 16S-23S ribosomal DNA intergenic regions in different *Treponema pallidum* subspecies contain tRNA genes. *FEMS Microbiol Lett* **143**, 235-240.

**Centurion-Lara, A., Godornes, C., Castro, C., Van Voorhis, W. C. & Lukehart, S. A. (2000a).** The *tpr*K gene is heterogeneous among *Treponema pallidum* strains and has multiple alleles. *Infect Immun* **68**, 824-831.

**Centurion-Lara, A., Castro, C., Castillo, R., Shaffer, J. M., Van Voorhis, W. C. & Lukehart, S. A. (1998).** The flanking region sequences of the 15-kDa lipoprotein gene differentiate pathogenic treponemes. *J Infect Dis* **177**, 1036-1040.

**Centurion-Lara, A., Sun, E. S., Barrett, L. K., Castro, C., Lukehart, S. A. & Van Voorhis, W. C. (2000b).** Multiple alleles of *Treponema pallidum* repeat gene D in *Treponema pallidum* isolates. *J Bacteriol* **182**, 2332-2335.

**Centurion-Lara, A., Arroll, T., Castillo, R., Shaffer, J. M., Castro, C., Van Voorhis, W. C. & Lukehart, S. A. (1997).** Conservation of the 15-kilodalton lipoprotein among *Treponema pallidum* subspecies and strains and other pathogenic treponemes: genetic and antigenic analyses. *Infect Immun* **65**, 1440-1444.

**Centurion-Lara, A., Castro, C., Barrett, L., Cameron, C., Mostowfi, M., Van Voorhis, W. C. & Lukehart, S. A. (1999).** *Treponema pallidum* major sheath protein homologue Tpr K is a target of opsonic antibody and the protective immune response. *J Exp Med* **189**, 647-656.

**Centurion-Lara, A., LaFond, R. E., Hevner, K., Godornes, C., Molini, B. J., Van Voorhis, W. C. & Lukehart, S. A. (2004).** Gene conversion: a mechanism for generation of heterogeneity in the *tpr*K gene of *Treponema pallidum* during infection. *Mol Microbiol* **52**, 1579-1596.

**Centurion-Lara, A., Molini, B. J., Godornes, C., Sun, E., Hevner, K., Van Voorhis, W. C. & Lukehart, S. A. (2006).** Molecular differentiation of *Treponema pallidum* subspecies. *J Clin Microbiol* **44**, 3377-3380.

**Clark, E. G. & Danbolt, N. (1955).** The Oslo study of the natural history of untreated syphilis; an epidemiologic investigation based on a restudy of the Boeck-Bruusgaard material; a review and appraisal. *J Chronic Dis* **2**, 311-344.

**Clyne, B. & Jerrard, D. A. (2000).** Syphilis testing. *J Emerg Med* **18**, 361-367.

**Coil, D. A., Vandersmissen, L., Ginevra, C., Jarraud, S., Lammertyn, E. & Anne, J. (2008).** Intragenic tandem repeat variation between *Legionella pneumophila* strains. *BMC Microbiol* **8**, 218.

**Comstedt, P., Asokliene, L., Eliasson, I., Olsen, B., Wallensten, A., Bunikis, J. & Bergstrom, S. (2009).** Complex population structure of Lyme borreliosis group spirochete *Borrelia garinii* in subarctic Eurasia. *PLoS One* **4**, e5841.

**Condon, C., Philips, J., Fu, Z. Y., Squires, C. & Squires, C. L. (1992).** Comparison of the expression of the seven ribosomal RNA operons in *Escherichia coli*. *The EMBO journal* **11**, 4175-4185.

**Cox, D. L. (1994).** Culture of *Treponema pallidum*. *Methods Enzymol* **236**, 390-405.

**Cox, D. L. & Radolf, J. D. (2006).** Metabolism of the *Treponema*. In *Pathogenic Treponema: Molecular and Cellular Biology*, pp. 61-100. Edited by J. D. Radolf & S. Lukehart. Norfolk, England: Caister Academic Press.

**Cox, D. L., Moeckli, R. A. & Fieldsteel, A. H. (1984).** Cultivation of pathogenic treponema in tissue cultures of SflEp cells. *In Vitro* **20**, 879-883.

**Cox, D. L., Chang, P., McDowall, A. W. & Radolf, J. D. (1992).** The outer membrane, not a coat of host proteins, limits antigenicity of virulent *Treponema pallidum*. *Infect Immun* **60**, 1076-1083.

**Cox, D. L., Riley, B., Chang, P., Sayahtaheri, S., Tassell, S. & Hevelone, J. (1990).** Effects of molecular oxygen, oxidation-reduction potential, and antioxidants upon in vitro replication of *Treponema pallidum* subsp. *pallidum*. *Appl Environ Microbiol* **56**, 3063-3072.

**Cox, D. L., Luthra, A., Dunham-Ems, S., Desrosiers, D. C., Salazar, J. C., Caimano, M. J. & Radolf, J. D. (2010).** Surface immunolabeling and consensus computational framework to identify candidate rare outer membrane proteins of *Treponema pallidum*. *Infect Immun* **78**, 5178-5194.

**Dai, Q. Y., Restrepo, B. I., Porcella, S. F., Raffel, S. J., Schwan, T. G. & Barbour, A. G. (2006).** Antigenic variation by *Borrelia hermsii* occurs through recombination between extragenic repetitive elements on linear plasmids. *Molecular Microbiology* **60**, 1329-1343.

**Darling, A. E., Miklos, I. & Ragan, M. A. (2008).** Dynamics of genome rearrangement in bacterial populations. *PLoS genetics* **4**, e1000128.

**Das, A. K., Cohen, P. W. & Barford, D. (1998).** The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *EMBO J* **17**, 1192-1199.

**de Vries, M. C., Siezen, R. J., Wijman, J. G., Zhao, Y., Kleerebezem, M., de Vos, W. M. & Vaughan, E. E. (2006).** Comparative and functional analysis of the rRNA-operons and their tRNA gene complement in different lactic acid bacteria. *Systematic and applied microbiology* **29**, 358-367.

**Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999).** Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636-4641.

**Desrosiers, D. C., Sun, Y. C., Zaidi, A. A., Eggers, C. H., Cox, D. L. & Radolf, J. D. (2007).** The general transition metal (Tro) and Zn2+ (Znu) transporters in *Treponema pallidum*: analysis of metal specificities and expression profiles. *Mol Microbiol* **65**, 137-152.

**Desrosiers, D. C., Anand, A., Luthra, A., Dunham-Ems, S. M., Ledoyt, M., Cummings, M. A., Eshghi, A., Cameron, C. E., Cruz, A. R. & other authors (2011).** TP0326, a *Treponema pallidum* Beta-Barrel Assembly Machinery A (BamA) Ortholog and Rare Outer Membrane Protein. *Mol Microbiol*.

**Dickerson, M. T., Abney, M. B., Cameron, C. E., Knecht, M., Bachas, L. G. & Anderson, K. W. (2012).** Fibronectin Binding to the T*reponema pallidum* Adhesin Protein Fragment rTp0483 on Functionalized Self-Assembled Monolayers. *Bioconjugate chemistry* **23**, 184-195.

**Duncan, W. C. & Knox, J. M. (1971).** Cephalosporin antibiotics in venereal disease. *Postgrad Med J* **47**, Suppl:119-122.

**Dunnick, W., Hertz, G. Z., Scappino, L. & Gritzmacher, C. (1993).** DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Res* **21**, 365-372.

**Duquette, M. L., Handa, P., Vincent, J. A., Taylor, A. F. & Maizels, N. (2004).** Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev* **18**, 1618-1629.

**Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmermann, J., Erfle, H., Caskey, C. T. & other authors (1990).** Automated DNA sequencing of the human HPRT locus. *Genomics* **6**, 593-608.

**Engelkens, H. J., Vuzevski, V. D., ten Kate, F. J., van der Heul, P., van der Sluis, J. J. & Stolz, E. (1991).** Ultrastructural aspects of infection with *Treponema pallidum* subspecies *pertenue* (Pariaman strain). *Genitourin Med* **67**, 403-407.

**Falabella, R. (1994).** Nonvenereal treponematoses: yaws, endemic syphilis, and pinta. *Journal of the American Academy of Dermatology* **31**, 1075.

**Fanella, S., Kadkhoda, K., Shuel, M. & Tsang, R. (2012).** Local transmission of imported endemic syphilis, Canada, 2011. *Emerg Infect Dis* **18**, 1002-1004.

**Farnsworth, N. & Rosen, T. (2006).** Endemic treponematosis: review and update. *Clin Dermatol* **24**, 181-190.

**Fegan, D., Glennon, M. J., Thami, Y. & Pakoa, G. (2010).** Resurgence of yaws in Tanna, Vanuatu: time for a new approach? *Trop Doct* **40**, 68-69.

**Fenno, J. C., Muller, K. H. & McBride, B. C. (1996).** Sequence analysis, expression, and binding activity of recombinant major outer sheath protein (Msp) of *Treponema denticola*. *J Bacteriol* **178**, 2489-2497.

**Fieldsteel, A. H., Cox, D. L. & Moeckli, R. A. (1981).** Cultivation of virulent *Treponema pallidum* in tissue culture. *Infect Immun* **32**, 908-915.

**Fitzgerald, J. J., Johnson, R. C. & Smith, M. (1976).** Accidental laboratory infection with *Treponema pallidum*, Nichols strain. *J Am Vener Dis Assoc* **3**, 76-78.

**Fitzgerald, T. J., Miller, J. N. & Sykes, J. A. (1975).** *Treponema pallidum* (Nichols strain) in tissue cultures: cellular attachment, entry, and survival. *Infect Immun* **11**, 1133-1140.

**Flasarová, M., Šmajs, D., Matějková, P., Woznicová, V., Heroldová-Dvořáková, M. & Votava, M. (2006).** [Molecular detection and subtyping of *Treponema pallidum* subsp. *pallidum* in clinical specimens]. *Epidemiol Mikrobiol Imunol* **55**, 105-111.

**Flasarová, M., Pospíšilová, P., Mikalová, L., Vališová, Z., Dastychová, E., Strnadel, R., Kuklová, I., Woznicová, V., Zakoucká, H. & other authors (2012).** Sequencing-based Molecular Typing of *Treponema pallidum* Strains in the Czech Republic: All Identified Genotypes are Related to the Sequence of the SS14 Strain. *Acta dermato-venereologica*.

**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. & other authors (1995).** Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

**Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R. & other authors (1998).** Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375-388.

**Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R. & other authors (1997).** Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580-586.

**Fribourg-Blanc, A. & Mollaret, H. H. (1969).** Natural treponematosis of the African primate. *Primates Med* **3**, 113-121.

**Fribourg-Blanc, A., Niel, G. & Mollaret, H. H. (1963).** [Note on Some Immunological Aspects of the African Cynocephalus. 1. Antigenic Relationship of Its Gamma Globulin with Human Gamma Globulin. 2. Guinean Endemic Focus of Treponematosis.]. *Bull Soc Pathol Exot Filiales* **56**, 474-485.

**Fribourg-Blanc, A., Mollaret, H. H. & Niel, G. (1966).** [Serologic and microscopic confirmation of treponemosis in Guinea baboons]. *Bull Soc Pathol Exot Filiales* **59**, 54-59.

**Fukunaga, M., Okuzako, N., Mifuchi, I., Arimitsu, Y. & Seki, M. (1992).** Organization of the ribosomal RNA genes in *Treponema phagedenis* and *Treponema pallidum*. *Microbiol Immunol* **36**, 161-167.

**Gaibani, P., Pellegrino, M. T., Rossini, G., Alvisi, G., Miragliotta, L., Prati, C. & Sambri, V. (2010).** The central region of the msp gene of Treponema denticola has sequence heterogeneity among clinical samples, obtained from patients with periodontitis. *BMC Infect Dis* **10**, 345.

**Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S. & other authors (2009).** Rfam: updates to the RNA families database. *Nucleic Acids Research* **37**, D136-D140.

**Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M. & Brinkman, F. S. (2005).** PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**, 617-623.

**Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003).** ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784-3788.

**Gastinel, P., Vaisman, A., Hamelin, A. & Dunoyer, F. (1963).** [Study of a recently isolated strain of *Treponema pertenue*]. *Prophyl Sanit Morale* **35**, 182-188.

**Gerstl, S., Kiwila, G., Dhorda, M., Lonlas, S., Myatt, M., Ilunga, B. K., Lemasson, D., Szumilin, E., Guerin, P. J. & other authors (2009).** Prevalence study of yaws in the Democratic Republic of Congo using the lot quality assurance sampling method. *PLoS One* **4**, e6338.

**Giacani, L., Hevner, K. & Centurion-Lara, A. (2005a).** Gene organization and transcriptional analysis of the *tpr*J, *tpr*I, *tpr*G, and *tpr*F loci in *Treponema pallidum* strains Nichols and Sea 81-4. *J Bacteriol* **187**, 6084-6093.

**Giacani, L., Lukehart, S. & Centurion-Lara, A. (2007a).** Length of guanosine homopolymeric repeats modulates promoter activity of subfamily II *tpr* genes of *Treponema pallidum* ssp. *pallidum*. *FEMS Immunol Med Microbiol* **51**, 289-301.

**Giacani, L., Molini, B., Godornes, C., Barrett, L., Van Voorhis, W., Centurion-Lara, A. & Lukehart, S. A. (2007b).** Quantitative analysis of *tpr* gene expression in *Treponema pallidum* isolates: Differences among isolates and correlation with T-cell responsiveness in experimental syphilis. *Infect Immun* **75**, 104-112.

**Giacani, L., Godornes, C., Puray-Chavez, M., Guerra-Giraldez, C., Tompa, M., Lukehart, S. A. & Centurion-Lara, A. (2009).** TP0262 is a modulator of promoter activity of *tpr* Subfamily II genes of *Treponema pallidum* ssp. *pallidum*. *Mol Microbiol* **72**, 1087-1099.

**Giacani, L., Jeffrey, B. M., Molini, B. J., Le, H. T., Lukehart, S. A., Centurion-Lara, A. & Rockey, D. D. (2010a).** Complete genome sequence and annotation of the *Treponema pallidum* subsp. *pallidum* Chicago strain. *J Bacteriol* **192**, 2645-2646.

**Giacani, L., Molini, B. J., Kim, E. Y., Godornes, B. C., Leader, B. T., Tantalo, L. C., Centurion-Lara, A. & Lukehart, S. A. (2010b).** Antigenic variation in *Treponema*

*pallidum*: TprK sequence diversity accumulates in response to immune pressure during experimental syphilis. *J Immunol* **184**, 3822-3829.

**Giacani, L., Sambri, V., Marangoni, A., Cavrini, F., Storni, E., Donati, M., Corona, S., Lanzarini, P. & Cevenini, R. (2005b).** Immunological evaluation and cellular location analysis of the TprI antigen of *Treponema pallidum* subsp. *pallidum*. *Infect Immun* **73**, 3817-3822.

**Giacani, L., Chattopadhyay, S., Centurion-Lara, A., Jeffrey, B. M., Le, H. T., Molini, B. J., Lukehart, S. A., Sokurenko, E. V. & Rockey, D. D. (2012a).** Footprint of Positive Selection in *Treponema pallidum* subsp. *pallidum* Genome Sequences Suggests Adaptive Microevolution of the Syphilis Pathogen. *PLoS Negl Trop Dis* **6**, e1698.

**Giacani, L., Brandt, S. L., Puray-Chavez, M., Brinck Reid, T., Godornes, C., Molini, B. J., Benzler, M., Hartig, J. S., Lukehart, S. A. & other authors (2012b).** Comparative Investigation of the Genomic Regions Involved in Antigenic Variation of the TprK Antigen among Treponemal Species, Subspecies, and Strains. *J Bacteriol*.

**Gioia, J., Yerrapragada, S., Qin, X., Jiang, H., Igboeli, O. C., Muzny, D., Dugan-Rocha, S., Ding, Y., Hawes, A. & other authors (2007).** Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. *PLoS One* **2**, e928.

**Glenn, T. C. (2011).** Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**, 759-769.

**Gray, R. R., Mulligan, C. J., Molini, B. J., Sun, E. S., Giacani, L., Godornes, C., Kitchen, A., Lukehart, S. A. & Centurion-Lara, A. (2006).** Molecular evolution of the *tpr*C, D, I, K, G, and J genes in the pathogenic genus *Treponema*. *Mol Biol Evol* **23**, 2220-2233.

**Green, P. (1993).**

**Gupta, P. K. (2008).** Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26**, 602-611.

**Gurtler, V. (1999).** The role of recombination and mutation in 16S-23S rDNA spacer rearrangements. *Gene* **238**, 241-252.

**Gurtler, V. & Stanisich, V. A. (1996).** New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. *Microbiology-Uk* **142**, 3-16.

**Guthe, T. (1960).** The treponematoses as a world problem. *Br J Vener Dis* **36**, 67-77.

**Hanakahi, L. A., Sun, H. & Maizels, N. (1999).** High affinity interactions of nucleolin with G-G-paired rDNA. *J Biol Chem* **274**, 15908-15912.

**Hanincová, K., Liveris, D., Sandigursky, S., Wormser, G. P. & Schwartz, I. (2008).** *Borrelia burgdorferi* sensu stricto is clonal in patients with early Lyme borreliosis. *Appl Environ Microbiol* **74**, 5008-5014.

**Hardy, J. B., Hardy, P. H., Oppenheimer, E. H., Ryan, S. J., Jr. & Sheff, R. N. (1970).** Failure of penicillin in a newborn with congenital syphilis. *JAMA* **212**, 1345-1349.

**Harper, K. N., Liu, H., Ocampo, P. S., Steiner, B. M., Martin, A., Levert, K., Wang, D., Sutton, M. & Armelagos, G. J. (2008a).** The sequence of the acidic repeat protein (*arp*) gene differentiates venereal from nonvenereal *Treponema pallidum* subspecies, and the gene has evolved under strong positive selection in the subspecies that causes syphilis. *FEMS Immunol Med Microbiol* **53**, 322-332.

**Harper, K. N., Ocampo, P. S., Steiner, B. M., George, R. W., Silverman, M. S., Bolotin, S., Pillay, A., Saunders, N. J. & Armelagos, G. J. (2008b).** On the origin of the treponematoses: a phylogenetic approach. *PLoS Negl Trop Dis* **2**, e148.

**Harrison, L. W. (1956).** The Oslo study of untreated syphilis, review and commentary. *Br J Vener Dis* **32**, 70-78.

**Harvey, S. & Hill, C. W. (1990).** Exchange of spacer regions between rRNA operons in *Escherichia coli*. *Genetics* **125**, 683-690.

**Hashimoto, J. G., Stevenson, B. S. & Schmidt, T. M. (2003).** Rates and consequences of recombination between rRNA operons. *Journal of bacteriology* **185**, 966-972.

**Hazlett, K. R., Sellati, T. J., Nguyen, T. T., Cox, D. L., Clawson, M. L., Caimano, M. J. & Radolf, J. D. (2001).** The TprK protein of *Treponema pallidum* is periplasmic and is not a target of opsonic antibody or protective immunity. *J Exp Med* **193**, 1015-1026.

**Hazlett, K. R., Cox, D. L., Decaffmeyer, M., Bennett, M. P., Desrosiers, D. C., La Vake, C. J., La Vake, M. E., Bourell, K. W., Robinson, E. J. & other authors (2005).** TP0453, a concealed outer membrane protein of *Treponema pallidum*, enhances membrane permeability. *J Bacteriol* **187**, 6499-6508.

**Highlander, S. K., Hulten, K. G., Qin, X., Jiang, H., Yerrapragada, S., Mason, E. O., Jr., Shang, Y., Williams, T. M., Fortunov, R. M. & other authors (2007).** Subtle genetic changes enhance virulence of methicillin resistant and sensitive *Staphylococcus aureus*. *BMC Microbiol* **7**, 99.

**Ho, E. L. & Lukehart, S. A. (2011).** Syphilis: using modern approaches to understand an old disease. *The Journal of clinical investigation* **121**, 4584-4592.

**Houston, S., Hof, R., Francescutti, T., Hawkes, A., Boulanger, M. J. & Cameron, C. E. (2011).** Bifunctional role of the *Treponema pallidum* extracellular matrix binding adhesin Tp0751. *Infect Immun* **79**, 1386-1398.

**Hovind-Hougen, K., Birch-Andersen, A. & Jensen, H. J. (1976).** Ultrastructure of cells of *Treponema pertenue* obtained from experimentally infected hamsters. *Acta Pathol Microbiol Scand B* **84**, 101-108.

**Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. (2007).** Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**, R143.

**Cha, J. Y., Ishiwata, A. & Mobashery, S. (2004).** A novel beta-lactamase activity from a penicillin-binding protein of *Treponema pallidum* and why syphilis is still treatable with penicillin. *J Biol Chem* **279**, 14917-14921.

**Chamberlain, N. R., Brandt, M. E., Erwin, A. L., Radolf, J. D. & Norgard, M. V. (1989).** Major integral membrane protein immunogens of *Treponema pallidum* are proteolipids. *Infect Immun* **57**, 2872-2877.

**Chaudhuri, R. R., Sebaihia, M., Hobman, J. L., Webber, M. A., Leyton, D. L., Goldberg, M. D., Cunningham, A. F., Scott-Tucker, A., Ferguson, P. R. & other authors (2010).** Complete Genome Sequence and Comparative Metabolic Profiling of the Prototypical Enteroaggregative *Escherichia coli* Strain 042. *PloS one* **5**.

**Ihssen, J. & Egli, T. (2005).** Global physiological analysis of carbon- and energy-limited growing Escherichia coli confirms a high degree of catabolic flexibility and preparedness for mixed substrate utilization. *Environmental Microbiology* **7**, 1568-1581.

**Indra, A., Blaschitz, M., Kernbichler, S., Reischl, U., Wewalka, G. & Allerberger, F. (2010).** Mechanisms behind variation in the *Clostridium difficile* 16S-23S rRNA intergenic spacer region. *Journal of medical microbiology* **59**, 1317-1323.

**Jonsson, A. B., Nyberg, G. & Normark, S. (1991).** Phase variation of gonococcal pili by frameshift mutation in *pil*C, a novel gene for pilus assembly. *EMBO J* **10**, 477-488.

**Jun, H. K., Kang, Y. M., Lee, H. R., Lee, S. H. & Choi, B. K. (2008).** Highly conserved surface proteins of oral spirochetes as adhesins and potent inducers of proinflammatory and osteoclastogenic factors. *Infect Immun* **76**, 2428-2438.

**Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. & Krogh, A. (2003).** Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**, 1652-1662.

**Karlsson, M., Fellstrom, C., Johansson, K. E. & Franklin, A. (2004).** Antimicrobial resistance in *Brachyspira pilosicoli* with special reference to point mutations in the 23S rRNA gene associated with macrolide and lincosamide resistance. *Microb Drug Resist* **10**, 204-208.

**Karlsson, M., Fellstrom, C., Heldtander, M. U., Johansson, K. E. & Franklin, A. (1999).** Genetic basis of macrolide and lincosamide resistance in *Brachyspira* (*Serpulina*) *hyodysenteriae. FEMS Microbiol Lett* **172**, 255-260.

**Katz, K. A. & Klausner, J. D. (2008).** Azithromycin resistance in *Treponema pallidum*. *Curr Opin Infect Dis* **21**, 83-91.

**Kestelyn, P. (2010).** Venereal and endemic treponematoses in the developing world. *International ophthalmology clinics* **50**, 41-55.

**Knauf, S., Batamuzi, E. K., Mlengeya, T., Kilewo, M., Lejora, I. A., Nordhoff, M., Ehlers, B., Harper, K. N., Fyumagwa, R. & other authors (2011).** *Treponema* Infection Associated With Genital Ulceration in Wild Baboons. *Veterinary pathology*.

**Kobayashi, I. (1992).** Mechanisms for gene conversion and homologous recombination: the double-strand break repair model and the successive half crossing-over model. *Advances in biophysics* **28**, 81-133.

**LaFond, R. E., Centurion-Lara, A., Godornes, C., Van Voorhis, W. C. & Lukehart, S. A. (2006).** TprK sequence diversity accumulates during infection of rabbits with *Treponema pallidum* subsp. *pallidum* Nichols strain. *Infect Immun* **74**, 1896-1906.

**LaFond, R. E., Centurion-Lara, A., Godornes, C., Rompalo, A. M., Van Voorhis, W. C. & Lukehart, S. A. (2003).** Sequence diversity of *Treponema pallidum* subsp. *pallidum tpr*K in human syphilis lesions and rabbit-propagated isolates. *J Bacteriol* **185**, 6262-6268.

**Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T. & Ussery, D. W. (2007).** RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**, 3100-3108.

**Lan, R. T. & Reeves, P. R. (1998).** Recombination between rRNA operons created most of the ribotype variation observed in the seventh pandemic clone of *Vibrio cholerae*. *Microbiol-Uk* **144**, 1213-1221.

**Leader, B. T., Hevner, K., Molini, B. J., Barrett, L. K., Van Voorhis, W. C. & Lukehart, S. A. (2003).** Antibody responses elicited against the *Treponema pallidum* repeat proteins differ during infection with different isolates of *Treponema pallidum* subsp. *pallidum*. *Infect Immun* **71**, 6054-6057.

**Lebuhn, M., Bathe, S., Achouak, W., Hartmann, A., Heulin, T. & Schloter, M. (2006).** Comparative sequence analysis of the internal transcribed spacer 1 of *Ochrobactrum species*. *Systematic and applied microbiology* **29**, 265-275.

**Leclercq, R. (2002).** Mechanisms of resistance to macrolides and lincosamides: nature of the resistance elements and their clinical implications. *Clin Infect Dis* **34**, 482-492.

**Levinson, G. & Gutman, G. A. (1987).** Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**, 203-221.

**Levrero, F., Gatti, S., Gautier-Hion, A. & Menard, N. (2007).** Yaws disease in a wild gorilla population and its impact on the reproductive status of males. *Am J Phys Anthropol* **132**, 568-575.

**Li, H. & Durbin, R. (2009).** Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009).** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

**Liao, D. (2000).** Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol* **51**, 305-317.

**Lim, J. Y., Hong, J. B., Sheng, H., Shringi, S., Kaul, R., Besser, T. E. & Hovde, C. J. (2010).** Phenotypic diversity of *Escherichia coli* O157:H7 strains associated with the plasmid O157. *Journal of microbiology* **48**, 347-357.

**Liska, S. L., Perine, P. L., Hunter, E. F., Crawford, J. A. & Feeley, J. C. (1982).** Isolation and Transportation of *Treponema Pertenue* in Golden Hamsters. *Current Microbiology* **7**, 41-43.

**Liu, H., Rodes, B., George, R. & Steiner, B. (2007).** Molecular characterization and analysis of a gene encoding the acidic repeat protein (Arp) of *Treponema pallidum*. *J Med Microbiol* **56**, 715-721.

**Liu, J., Howell, J. K., Bradley, S. D., Zheng, Y., Zhou, Z. H. & Norris, S. J. (2010).** Cellular architecture of *Treponema pallidum*: novel flagellum, periplasmic cone, and cell envelope as revealed by cryo electron tomography. *J Mol Biol* **403**, 546-561.

**Lovell, N. C., Jurmain, R. & Kilgore, L. (2000).** Skeletal evidence of probable treponemal infection in free-ranging African apes. *Primates* **41**, 275-290.

**Lowe, T. M. & Eddy, S. R. (1997).** tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955-964.

**Luciani, F., Bull, R. A. & Lloyd, A. R. (2012).** Next generation deep sequencing and vaccine design: today and tomorrow. *Trends Biotechnol* **30**, 443-452.

**Lukashin, A. V. & Borodovsky, M. (1998).** GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-1115.

**Lukehart, S. A., Godornes, C., Molini, B. J., Sonnett, P., Hopkins, S., Mulcahy, F., Engelman, J., Mitchell, S. J., Rompalo, A. M. & other authors (2004).** Macrolide resistance in *Treponema pallidum* in the United States and Ireland. *The New England journal of medicine* **351**, 154-158.

**Lumeij, J. T. (2010).** Widespread treponemal infections of hare populations (Lepus europaeus) in the Netherlands. *Eur J Wildl Res* **56**, DOI 10.1007/s10344-10010-10428-10343.

**Lumeij, J. T., de Koning, J., Bosma, R. B., van der Sluis, J. J. & Schellekens, J. F. (1994).** Treponemal infections in hares in The Netherlands. *J Clin Microbiol* **32**, 543-546.

**Luthra, A., Zhu, G., Desrosiers, D. C., Eggers, C. H., Mulay, V., Anand, A., McArthur, F. A., Romano, F. B., Caimano, M. J. & other authors (2011).** The Transition from Closed to Open Conformation of *Treponema pallidum* Outer Membrane-associated Lipoprotein TP0453 Involves Membrane Sensing and Integration by Two Amphipathic Helices. *The Journal of biological chemistry* **286**, 41656-41668.

**Magnuson, H. J., Thomas, E. W., Olansky, S., Kaplan, B. I., De Mello, L. & Cutler, J. C. (1956).** Inoculation syphilis in human volunteers. *Medicine (Baltimore)* **35**, 33-82.

**Mardis, E. R. (2008a).** Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402.

**Mardis, E. R. (2008b).** The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-141.

**Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J. & other authors (2005).** Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

**Marra, C. M. (2009).** Update on neurosyphilis. *Curr Infect Dis Rep* **11**, 127-134.

**Marra, C. M., Sahi, S. K., Tantalo, L. C., Godornes, C., Reid, T., Behets, F., Rompalo, A., Klausner, J. D., Yin, Y. P. & other authors (2010).** Enhanced molecular typing of treponema pallidum: geographical distribution of strain types and association with neurosyphilis. *J Infect Dis* **202**, 1380-1388.

**Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. & Lefeuvre, P. (2010).** RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* **26**, 2462-2463.

**Martin, I. E., Gu, W., Yang, Y. & Tsang, R. S. (2009).** Macrolide resistance and molecular types of *Treponema pallidum* causing primary syphilis in Shanghai, China. *Clin Infect Dis* **49**, 515-521.

**Matějková, P., Flasarová, M., Zakoucká, H., Bořek, M., Křemenová, S., Arenberger, P., Woznicová, V., Weinstock, G. M. & Šmajs, D. (2009).** Macrolide treatment failure in a case of secondary syphilis: a novel A2059G mutation in the 23S rRNA gene of *Treponema pallidum* subsp. *pallidum*. *J Med Microbiol* **58**, 832-836.

**Matějková, P., Strouhal, M., Šmajs, D., Norris, S. J., Palzkill, T., Petrosino, J. F., Sodergren, E., Norton, J. E., Singh, J. & other authors (2008).** Complete genome sequence of *Treponema pallidum* ssp. *pallidum* strain SS14 determined with oligonucleotide arrays. *BMC Microbiol* **8**, 76.

**Maurice, J. (2012).** WHO plans new yaws eradication campaign. *Lancet* **379**, 1377-1378.

**McCutcheon, J. P. & Moran, N. A. (2012).** Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**, 13-26.

**McGill, M. A., Edmondson, D. G., Carroll, J. A., Cook, R. G., Orkiszewski, R. S. & Norris, S. J. (2010).** Characterization and serologic analysis of the *Treponema pallidum* proteome. *Infect Immun* **78**, 2631-2643.

**McKevitt, M., Patel, K., Šmajs, D., Marsh, M., McLoughlin, M., Norris, S. J., Weinstock, G. M. & Palzkill, T. (2003).** Systematic cloning of *Treponema pallidum* open reading frames for protein expression and antigen discovery. *Genome Res* **13**, 1665-1674.

**McKevitt, M., Brinkman, M. B., McLoughlin, M., Perez, C., Howell, J. K., Weinstock, G. M., Norris, S. J. & Palzkill, T. (2005).** Genome scale identification of *Treponema pallidum* antigens. *Infect Immun* **73**, 4445-4450.

**McLeod, M. P., Qin, X., Karpathy, S. E., Gioia, J., Highlander, S. K., Fox, G. E., McNeill, T. Z., Jiang, H., Muzny, D. & other authors (2004).** Complete genome sequence of *Rickettsia typhi* and comparison with sequences of other rickettsiae. *J Bacteriol* **186**, 5842-5855.

**Meheus, A. (1985).** Integration of yaws control and primary health care. *Rev Infect Dis* **7 Suppl 2**, S284-288.

**Meyer, T. F., Mlawer, N. & So, M. (1982).** Pilus expression in *Neisseria gonorrhoeae* involves chromosomal rearrangement. *Cell* **30**, 45-52.

**Miao, R. & Fieldsteel, A. H. (1978).** Genetics of *Treponema* - Relationship between *Treponema Pallidum* and 5 Cultivable Treponemes. *Journal of Bacteriology* **133**, 101-107.

**Miao, R. M. & Fieldsteel, A. H. (1980).** Genetic relationship between *Treponema pallidum* and *Treponema pertenue*, two noncultivable human pathogens. *J Bacteriol* **141**, 427-429.

**Mikalová, L., Strouhal, M., Čejková, D., Zobaníková, M., Pospíšilová, P., Norris, S. J., Sodergren, E., Weinstock, G. M. & Šmajs, D. (2010).** Genome analysis of *Treponema pallidum* subsp. *pallidum* and subsp. *pertenue* strains: most of the genetic differences are localized in six regions. *PLoS One* **5**, e15713.

**Mitchell, S. J., Engelman, J., Kent, C. K., Lukehart, S. A., Godornes, C. & Klausner, J. D. (2006).** Azithromycin-resistant syphilis infection: San Francisco, California, 2000-2004. *Clin Infect Dis* **42**, 337-345.

**Mitja, O., Hays, R., Lelngei, F., Laban, N., Ipai, A., Pakarui, S. & Bassat, Q. (2011).** Challenges in recognition and diagnosis of yaws in children in papua new Guinea. *The American journal of tropical medicine and hygiene* **85**, 113-116.

**Molepo, J., Pillay, A., Weber, B., Morse, S. A. & Hoosen, A. A. (2007).** Molecular typing of *Treponema pallidum* strains from patients with neurosyphilis in Pretoria, South Africa. *Sex Transm Infect* **83**, 189-192.

**Nei, M. & Kumar, S. (2000).** *Molecular Evolution and Phylogenetics*. Oxford: Oxford University Press.

**Nei, M. & Rooney, A. P. (2005).** Concerted and birth-and-death evolution of multigene families. *Annual review of genetics* **39**, 121-152.

**Nichols, H. J. & Hough, W. H. (1913).** DEMONSTRATION OF *SPIROCHAETA PALLIDA* IN THE CEREBROSPINAL FLUID. *JAMA* **60**, 108-110.

**Nichols, J. C. & Baseman, J. B. (1975).** Carbon sources utilized by virulent *Treponema pallidum*. *Infect Immun* **12**, 1044-1050.

**Noordhoek, G. T., Hermans, P. W. M., Paul, A. N., Schouls, L. M., Vandersluis, J. J. & Vanembden, J. D. A. (1989).** *Treponema Pallidum* Subspecies *Pallidum* (Nichols) and *Treponema Pallidum* Subspecies *Pertenue* (Cdc-2575) Differ in at Least One Nucleotide - Comparison of 2 Homologous Antigens. *Microbial Pathogenesis* **6**, 29-42.

**Noordhoek, G. T., Cockayne, A., Schouls, L. M., Meloen, R. H., Stolz, E. & Vanembden, J. D. A. (1990).** A New Attempt to Distinguish Serologically the Subspecies of *Treponema Pallidum* Causing Syphilis and Yaws. *Journal of Clinical Microbiology* **28**, 1600-1607.

**Norgard, M. V. & Miller, J. N. (1981).** Plasmid DNA in *Treponema pallidum* (Nichols): potential for antibiotic resistance by syphilis bacteria. *Science* **213**, 553-555.

**Norris, S. J. (1982).** In vitro cultivation of *Treponema pallidum*: independent confirmation. *Infect Immun* **36**, 437-439.

**Norris, S. J., Cox, D. L. & Weinstock, G. M. (2001).** Biology of *Treponema pallidum*: correlation of functional activities with genome sequence data. *J Mol Microbiol Biotechnol* **3**, 37-62.

**Norris, S. J., Paster, B. J., Moter, A. & Gobel, B. (2006).** The Genus *Treponema*. In *The Prokaryotes: A Handbook on the Biology of Bacteria, 3rd edition*, vol. 7, pp. 211-234. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer & E. Stackerbrandt. New York, NY, USA: Springer Science-Business media, LLC.

**Ovcinnikov, N. M. & Delektorskij, V. V. (1970).** *Treponema pertenue* under the electron microscope. *Br J Vener Dis* **46**, 349-379.

**Ovcinnikov, N. M. & Delektorskij, V. V. (1971).** Current concepts of the morphology and biology of *Treponema pallidum* based on electron microscopy. *The British journal of venereal diseases* **47**, 315-328.

**Pandori, M. W., Gordones, C., Castro, L., Engelman, J., Siedner, M., Lukehart, S. & Klausner, J. (2007).** Detection of azithromycin resistance in *Treponema pallidum* by real-time PCR. *Antimicrob Agents Chemother* **51**, 3425-3430.

**Parkinson, G. N., Lee, M. P. & Neidle, S. (2002).** Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **417**, 876-880.

**Paster, B. J. & Dewhirst, F. E. (2000).** Phylogenetic foundation of spirochetes. *J Mol Microbiol Biotechnol* **2**, 341-344.

**Pei, A., Nossa, C. W., Chokshi, P., Blaser, M. J., Yang, L., Rosmarin, D. M. & Pei, Z. (2009).** Diversity of 23S rRNA genes within individual prokaryotic genomes. *PLoS One* **4**, e5437.

**Pei, A. Y., Oberdorf, W. E., Nossa, C. W., Agarwal, A., Chokshi, P., Gerz, E. A., Jin, Z., Lee, P., Yang, L. & other authors (2010).** Diversity of 16S rRNA genes within individual prokaryotic genomes. *Applied and environmental microbiology* **76**, 3886-3897.

**Peterson, K. M., Baseman, J. B. & Alderete, J. F. (1983).** *Treponema pallidum* receptor binding proteins interact with fibronectin. *The Journal of experimental medicine* **157**, 1958-1970.

**Petes, T. D. & Hill, C. W. (1988).** Recombination between repeated genes in microorganisms. *Annu Rev Genet* **22**, 147-168.

**Petit, M.-A. (2005).** Mechanisms of homologous recombination in bacteria. In *The Dynamic Bacterial Genome*, pp. 3-32. Edited by P. Mullany. New York: Cambridge University Press.

**Pětrošová, H., Zobaníková, M., Čejková, D., Mikalová, L., Pospíšilová, P., Strouhal, M., Chen, L., Qin, X., Muzny, D. M. & other authors (2012).** Whole Genome Sequence of *Treponema pallidum* ssp. *pallidum*, Strain Mexico A, Suggests Recombination between Yaws and Syphilis Strains. *PLoS Negl Trop Dis* **6**, e1832.

**Pillay, A., Liu, H., Chen, C. Y., Holloway, B., Sturm, A. W., Steiner, B. & Morse, S. A. (1998).** Molecular subtyping of *Treponema pallidum* subspecies *pallidum*. *Sex Transm Dis* **25**, 408-414.

**Pillay, A., Chen, C. Y., Reynolds, M. G., Mombouli, J. V., Castro, A. C., Louvouezo, D., Steiner, B. & Ballard, R. C. (2011).** Laboratory confirmed case of yaws in a 10 year-old boy from the Republic of the Congo. *Journal of clinical microbiology*.

**Pillay, A., Liu, H., Ebrahim, S., Chen, C. Y., Lai, W., Fehler, G., Ballard, R. C., Steiner, B., Sturm, A. W. & other authors (2002).** Molecular typing of *Treponema pallidum* in South Africa: cross-sectional studies. *J Clin Microbiol* **40**, 256-258.

**Piruzian, A. L. (1989).** [Study of the plasmid composition of various strains of *Treponema pallidum*]. *Mol Gen Mikrobiol Virusol*, 37-42.

**Posada, D. & Crandall, K. A. (2001).** Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13757-13762.

**Pruss, B. M., Verma, K., Samanta, P., Sule, P., Kumar, S., Wu, J., Christianson, D., Horne, S. M., Stafslien, S. J. & other authors (2010).** Environmental and genetic factors that contribute to *Escherichia coli* K-12 biofilm formation. *Archives of microbiology* **192**, 715-728.

**Radolf, J. D. & Desrosiers, D. C. (2009).** *Treponema pallidum*, the stealth pathogen, changes, but how? *Mol Microbiol* **72**, 1081-1086.

**Radolf, J. D., Norgard, M. V. & Schulz, W. W. (1989).** Outer-Membrane Ultrastructure Explains the Limited Antigenicity of Virulent *Treponema Pallidum*. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 2051-2055.

**Rawal, P., Kummarasetti, V. B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S. K. & Chowdhury, S. (2006).** Genome-wide prediction of G4 DNA as regulatory motifs: role in *Escherichia coli* global regulation. *Genome Res* **16**, 644-655.

**Rawlings, N. D., Morton, F. R., Kok, C. Y., Kong, J. & Barrett, A. J. (2008).** MEROPS: the peptidase database. *Nucleic Acids Res* **36**, D320-325.

**Read, T. D., Peterson, S. N., Tourasse, N., Baillie, L. W., Paulsen, I. T., Nelson, K. E., Tettelin, H., Fouts, D. E., Eisen, J. A. & other authors (2003).** The genome

sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**, 81-86.

**Rinaldi, A. (2008).** Yaws: a second (and maybe last?) chance for eradication. *PLoS Negl Trop Dis* **2**, e275.

**Roman, G. C. & Roman, L. N. (1986).** Occurrence of congenital, cardiovascular, visceral, neurologic, and neuro-ophthalmologic complications in late yaws: a theme for future research. *Rev Infect Dis* **8**, 760-770.

**Rozen, S. & Skaletsky, H. (2000).** Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386.

**Sadeghifard, N., Gurtler, V., Beer, M. & Seviour, R. J. (2006).** The mosaic nature of intergenic 16S-23S rRNA spacer regions suggests rRNA operon copy number variation in *Clostridium difficile* strains. *Applied and environmental microbiology* **72**, 7311-7323.

**Saier, M. H., Jr., Tran, C. V. & Barabote, R. D. (2006).** TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**, D181-186.

**Saitou, N. & Nei, M. (1987).** The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**, 406-425.

**Santoyo, G. & Romero, D. (2005).** Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol Rev* **29**, 169-183.

**Sawyer, S. (1989).** Statistical tests for detecting gene conversion. *Molecular biology and evolution* **6**, 526-538.

**Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R. & other authors (2009).** Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**, D5-15.

**Sebaihia, M., Peck, M. W., Minton, N. P., Thomson, N. R., Holden, M. T., Mitchell, W. J., Carter, A. T., Bentley, S. D., Mason, D. R. & other authors (2007).** Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes. *Genome Res* **17**, 1082-1092.

**Sechman, E. V., Rohrer, M. S. & Seifert, H. S. (2005).** A genetic screen identifies genes and sites involved in pilin antigenic variation in *Neisseria gonorrhoeae*. *Molecular Microbiology* **57**, 468-483.

**Sen, D. & Gilbert, W. (1988).** Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334**, 364-366.

**Sepetjian, M., Guerraz, F. T., Salussola, D., Thivolet, J. & Monier, J. C. (1969).** [Contribution to the study of the treponeme isolated from monkeys by A. Fribourg-Blanc]. *Bull World Health Organ* **40**, 141-151.

**Seshadri, R., Myers, G. S., Tettelin, H., Eisen, J. A., Heidelberg, J. F., Dodson, R. J., Davidsen, T. M., DeBoy, R. T., Fouts, D. E. & other authors (2004).** Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc Natl Acad Sci U S A* **101**, 5646-5651.

**Setubal, J. C., Reis, M., Matsunaga, J. & Haake, D. A. (2006).** Lipoprotein computational prediction in spirochaetal genomes. *Microbiology* **152**, 113-121.

**Schaudinn, F. & Hoffman, E. (1905a).** Über *Spirochaeta pallida* bei Syphilis und die Unterschiede dieser Form gegenuber anderen Arten dieser Gattung. *Berlin Klin Wochschr* **42**, 673-675.

**Schaudinn, F. R. & Hoffman, E. (1905b).** Vorlaufiger Bericht uber das Vorkommen von Spirochaeten in syphilitischen Krankenheitsprodukten und bei Papilomen. *Arb K Gesund* **22**, 527-534.

**Schell, R. F. (1983).** Rabbit and hamster model of treponemal infection. In *Pathogenesis and Immunology of Treponemal Infection*, pp. 121-135. Edited by R. F. Schell & D. M. Musher. New York, NY, USA: Marcel Dekkker, INC.

**Schiller, N. L. & Cox, C. D. (1977).** Catabolism of glucose and fatty acids by virulent *Treponema pallidum*. *Infect Immun* **16**, 60-68.

**Simonsson, T., Pečinka, P. & Kubista, M. (1998).** DNA tetraplex formation in the control region of *c-myc*. *Nucleic Acids Res* **26**, 1167-1172.

**Šmajs, D., Norris, S. J. & Weinstock, G. M. (2012).** Genetic diversity in *Treponema pallidum*: implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases* **12**, 191-202.

**Šmajs, D., McKevitt, M., Wang, L., Howell, J. K., Norris, S. J., Palzkill, T. & Weinstock, G. M. (2002).** BAC library of *T. pallidum* DNA in *E. coli*. *Genome Res* **12**, 515-522.

**Šmajs, D., McKevitt, M., Howell, J. K., Norris, S. J., Cai, W. W., Palzkill, T. & Weinstock, G. M. (2005).** Transcriptome of *Treponema pallidum*: gene expression profile during experimental rabbit infection. *J Bacteriol* **187**, 1866-1874.

**Šmajs, D., Zobaníková, M., Strouhal, M., Čejková, D., Dugan-Rocha, S., Pospíšilová, P., Norris, S. J., Albert, T., Qin, X. & other authors (2011).** Complete Genome Sequence of *Treponema paraluiscuniculi*, Strain Cuniculi A: The Loss of Infectivity to Humans Is Associated with Genome Decay. *PLoS One* **6**, e20415.

**Smibert, R. M. (1984).** Genus III: *Treponema* Schaudinn 1905, 1728AL. In *Bergey's Manual of Systematic Bacteriology*, pp. 49-57. Edited by N. R. Krieg & J. G. Holt. Baltimore, MD, USA: Williams&Wilkins.

**Smith, J. L., David, N. J., Indgin, S., Israel, C. W., Levine, B. M., Justice, J., Jr., McCrary, J. A., 3rd, Medina, R., Paez, P. & other authors (1971).** Neuro-ophthalmological study of late yaws and pinta. II. The Caracas project. *Br J Vener Dis* **47**, 226-251.

**Smith, J. M. (1992).** Analyzing the mosaic structure of genes. *Journal of molecular evolution* **34**, 126-129.

**Soo, V. W., Hanson-Manful, P. & Patrick, W. M. (2011).** Artificial gene amplification reveals an abundance of promiscuous resistance determinants in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 1484-1489.

**Spuesens, E. B., Oduber, M., Hoogenboezem, T., Sluijter, M., Hartwig, N. G., van Rossum, A. M. & Vink, C. (2009).** Sequence variations in RepMP2/3 and RepMP4 elements reveal intragenomic homologous DNA recombination events in *Mycoplasma pneumoniae*. *Microbiology* **155**, 2182-2196.

**Spuesens, E. B., van de Kreeke, N., Estevao, S., Hoogenboezem, T., Sluijter, M., Hartwig, N. G., van Rossum, A. M. & Vink, C. (2011).** Variation in a surface-exposed region of the *Mycoplasma pneumoniae* P40 protein as a consequence of homologous DNA recombination between RepMP5 elements. *Microbiology* **157**, 473-483.

**Stamm, L. V. (2010).** Global challenge of antibiotic-resistant *Treponema pallidum*. *Antimicrob Agents Chemother* **54**, 583-589.

**Stamm, L. V. & Bergen, H. L. (2000a).** The sequence-variable, single-copy *tpr*K gene of *Treponema pallidum* Nichols strain UNC and Street strain 14 encodes heterogeneous TprK proteins. *Infect Immun* **68**, 6482-6486.

**Stamm, L. V. & Bergen, H. L. (2000b).** A point mutation associated with bacterial macrolide resistance is present in both 23S rRNA genes of an erythromycin-resistant *Treponema pallidum* clinical isolate. *Antimicrob Agents Chemother* **44**, 806-807.

**Stamm, L. V., Stapleton, J. T. & Bassford, P. J., Jr. (1988).** *In vitro* assay to demonstrate high-level erythromycin resistance of a clinical isolate of *Treponema pallidum*. *Antimicrob Agents Chemother* **32**, 164-169.

**Stamm, L. V., Bergen, H. L. & Walker, R. L. (2002).** Molecular typing of papillomatous digital dermatitis-associated *Treponema* isolates based on analysis of 16S-23S ribosomal DNA intergenic spacer regions. *Journal of clinical microbiology* **40**, 3463-3469.

**Stamm, L. V., Kerner, T. C., Jr., Bankaitis, V. A. & Bassford, P. J., Jr. (1983).** Identification and preliminary characterization of *Treponema pallidum* protein antigens expressed in *Escherichia coli*. *Infect Immun* **41**, 709-721.

**Stewart, F. J. & Cavanaugh, C. M. (2007).** Intragenomic variation and evolution of the internal transcribed spacer of the rRNA operon in bacteria. *J Mol Evol* **65**, 44-67.

**Strouhal, M. (2010).** Restrikční mapování genomů a genomová amplifikace patogenních kmenů rodu *Treponema*. *Ph.D. thesis,* Masaryk University, Brno.

**Strouhal, M., Šmajs, D., Matějková, P., Sodergren, E., Amin, A. G., Howell, J. K., Norris, S. J. & Weinstock, G. M. (2007).** Genome differences between *Treponema pallidum* subsp. *pallidum* strain Nichols and *T. paraluiscuniculi* strain Cuniculi A. *Infect Immun* **75**, 5859-5866.

**Sturino, J., Zorych, I., Mallick, B., Pokusaeva, K., Chang, Y. Y., Carroll, R. J. & Bliznuyk, N. (2010).** Statistical Methods for Comparative Phenomics Using High-Throughput Phenotype Microarrays. *International Journal of Biostatistics* **6**.

**Sun, E. S., Molini, B. J., Barrett, L. K., Centurion-Lara, A., Lukehart, S. A. & Van Voorhis, W. C. (2004).** Subfamily I *Treponema pallidum* repeat protein family: sequence variation and immunity. *Microbes Infect* **6**, 725-737.

**Sutton, M. Y., Liu, H., Steiner, B., Pillay, A., Mickey, T., Finelli, L., Morse, S., Markowitz, L. E. & St Louis, M. E. (2001).** Molecular subtyping of *Treponema pallidum* in an Arizona County with increasing syphilis morbidity: use of specimens from ulcers and blood. *J Infect Dis* **183**, 1601-1606.

**Tachon, S., Michelon, D., Chambellon, E., Cantonnet, M., Mezange, C., Henno, L., Cachon, R. & Yvon, M. (2009).** Experimental conditions affect the site of tetrazolium violet reduction in the electron transport chain of *Lactococcus lactis*. *Microbiology* **155**, 2941-2948.

**Takahashi, N. K., Yamamoto, K., Kitamura, Y., Luo, S. Q., Yoshikura, H. & Kobayashi, I. (1992).** Nonconservative recombination in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 5912-5916.

**Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007).** MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.

**Thornburg, R. W. & Baseman, J. B. (1983).** Comparison of major protein antigens and protein profiles of *Treponema pallidum* and *Treponema pertenue*. *Infect Immun* **42**, 623-627.

**Titz, B., Rajagopala, S. V., Ester, C., Hauser, R. & Uetz, P. (2006).** Novel conserved assembly factor of the bacterial flagellum. *J Bacteriol* **188**, 7700-7706.

**Titz, B., Rajagopala, S. V., Goll, J., Hauser, R., McKevitt, M. T., Palzkill, T. & Uetz, P. (2008).** The binary protein interactome of *Treponema pallidum*--the syphilis spirochete. *PLoS One* **3**, e2292.

**Torres-Cruz, J. & van der Woude, M. W. (2003).** Slipped-strand mispairing can function as a phase variation mechanism in *Escherichia coli*. *J Bacteriol* **185**, 6990-6994.

**Turner, T. B. & Hollander, D. H. (1957).** Biology of the treponematoses based on studies carried out at the International Treponematosis Laboratory Center of the Johns

Hopkins University under the auspices of the World Health Organization. *Monogr Ser World Health Organ*, 3-266.

**Ussery, D. W. & Hallin, P. F. (2004).** Genome Update: annotation quality in sequenced microbial genomes. *Microbiology* **150**, 2015-2017.

**Vabres, P. (2011).** Endemic treponemal infections in international adoptees and immigrant children: how common are they? *Pediatric dermatology* **28**, 214-215.

**van der Ende, A., Hopman, C. T., Zaat, S., Essink, B. B., Berkhout, B. & Dankert, J. (1995).** Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the spacing between the -10 and -35 regions of the promoter. *J Bacteriol* **177**, 2475-2480.

**van der Woude, M. W. & Baumler, A. J. (2004).** Phase and antigenic variation in bacteria. *Clin Microbiol Rev* **17**, 581-611, table of contents.

**van Heerden, J., Korf, C., Ehlers, M. M. & Cloete, T. E. (2002).** Biolog for the determination of diversity in microbial communities. *Water Sa* **28**, 29-35.

**Van Voorhis, W. C., Barrett, L. K., Lukehart, S. A., Schmidt, B., Schriefer, M. & Cameron, C. E. (2003).** Serodiagnosis of syphilis: antibodies to recombinant Tp0453, Tp92, and Gpd proteins are sensitive and specific indicators of infection by *Treponema pallidum*. *J Clin Microbiol* **41**, 3668-3674.

**Vink, C., Rudenko, G. & Seifert, H. S. (2011).** Microbial antigenic variation mediated by homologous DNA recombination. *FEMS microbiology reviews*.

**Walker, E. M., Arnett, J. K., Heath, J. D. & Norris, S. J. (1991).** *Treponema pallidum* subsp. pallidum has a single, circular chromosome with a size of approximately 900 kilobase pairs. *Infect Immun* **59**, 2476-2479.

**Walker, E. M., Zampighi, G. A., Blanco, D. R., Miller, J. N. & Lovett, M. A. (1989).** Demonstration of Rare Protein in the Outer Membrane of *Treponema Pallidum* Subsp *Pallidum* by Freeze-Fracture Analysis. *Journal of Bacteriology* **171**, 5005-5011.

**Walker, E. M., Howell, J. K., You, Y., Hoffmaster, A. R., Heath, J. D., Weinstock, G. M. & Norris, S. J. (1995).** Physical map of the genome of *Treponema pallidum* subsp. *pallidum* (Nichols). *J Bacteriol* **177**, 1797-1804.

**Walker, S. L. & Hay, R. J. (2000).** Yaws-a review of the last 50 years. *Int J Dermatol* **39**, 258-260.

**Wang, X. X., Kim, Y., Ma, Q., Hong, S. H., Pokusaeva, K., Sturino, J. M. & Wood, T. K. (2010).** Cryptic prophages help bacteria cope with adverse environments. *Nature Communications* **1**.

**Weigel, L. M., Radolf, J. D. & Norgard, M. V. (1994).** The 47-kDa major lipoprotein immunogen of *Treponema pallidum* is a penicillin-binding protein with carboxypeptidase activity. *Proc Natl Acad Sci U S A* **91**, 11611-11615.

**Weinstock, G. M. (2012).** Genomic approaches to studying the human microbiota. *Nature* **489**, 250-256.

**Weinstock, G. M., Hardham, J. M., McLeod, M. P., Sodergren, E. J. & Norris, S. J. (1998).** The genome of *Treponema pallidum*: new light on the agent of syphilis. *FEMS Microbiol Rev* **22**, 323-332.

**Weitzmann, M. N., Woodford, K. J. & Usdin, K. (1997).** DNA secondary structures and the evolution of hypervariable tandem arrays. *J Biol Chem* **272**, 9517-9523.

**Wendel, G. D., Jr., Sanchez, P. J., Peters, M. T., Harstad, T. W., Potter, L. L. & Norgard, M. V. (1991).** Identification of *Treponema pallidum* in amniotic fluid and fetal blood from pregnancies complicated by congenital syphilis. *Obstet Gynecol* **78**, 890-895.

**White, R. M. (2000).** Unraveling the Tuskegee Study of Untreated Syphilis. *Arch Intern Med* **160**, 585-598.

**WHO (1998).** The World Health Report 1998 - -life in the 21century: a visiion for all. *World Health Organization*, 132.

**WHO (2001).** Global prevalence and incidence of selected curable sexually transmitted infections: overview and estimates. *Tech Report No WHO/HIV_AIDS/200102, WHO/CDS/CSR/EDC/200110*.

**Wilson, J. (1973).** Syphilis and yaws: diagnostic difficulties and case report. *N Z Med J* **78**, 18-21.

**Workowski, K. A. & Berman, S. M. (2006).** Sexually transmitted diseases treatment guidelines, 2006. *MMWR Recomm Rep* **55**, 1-94.

**Wormser, G. P., Brisson, D., Liveris, D., Hanincová, K., Sandigursky, S., Nowakowski, J., Nadelman, R. B., Ludin, S. & Schwartz, I. (2008).** *Borrelia burgdorferi* genotype predicts the capacity for hematogenous dissemination during early Lyme disease. *J Infect Dis* **198**, 1358-1364.

**Woznicová, V., Šmajs, D., Wechsler, D., Matějková, P. & Flasarová, M. (2007).** Detection of *Treponema pallidum* subsp. *pallidum* from skin lesions, serum, and cerebrospinal fluid in an infant with congenital syphilis after clindamycin treatment of the mother during pregnancy. *J Clin Microbiol* **45**, 659-661.

**Woznicová, V., Matějková, P., Flasarová, M., Zakoucká, H., Vališová, Z., Šmajs, D. & Dastychová, E. (2010).** Clarithromycin treatment failure due to macrolide resistance in *Treponema pallidum* in a patient with primary syphilis. *Acta Derm Venereol* **90**, 206-207.

**Xue, X., Sztajer, H., Buddruhs, N., Petersen, J., Rohde, M., Talay, S. R. & Wagner-Dobler, I. (2011).** Lack of the delta subunit of RNA polymerase increases virulence related traits of *Streptococcus mutans*. *PloS one* **6**, e20075.

**Zapun, A., Contreras-Martel, C. & Vernet, T. (2008).** Penicillin-binding proteins and beta-lactam resistance. *FEMS Microbiol Rev* **32**, 361-385.

**Zdobnov, E. M. & Apweiler, R. (2001).** InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848.

**Zerbino, D. R. & Birney, E. (2008).** Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829.

**Zhang, J. R. & Norris, S. J. (1998).** Genetic variation of the *Borrelia burgdorferi* gene *vls*E involves cassette-specific, segmental gene conversion. *Infect Immun* **66**, 3698-3704.

**Zhang, Q. Y., DeRyckere, D., Lauer, P. & Koomey, M. (1992).** Gene conversion in *Neisseria gonorrhoeae*: evidence for its role in pilus antigenic variation. *Proc Natl Acad Sci U S A* **89**, 5366-5370.

**Zobaníková, M., Mikolka, P., Čejková, D., Pospíšilová, P., Chen, L., Strouhal, M., Qin, X., Weinstock, G. M. & Šmajs, D. (2012).** Complete genome sequence of *Treponema pallidum* strain DAL-1. *Standards in Genomics Sciences* **7**.

# 8. ATTACHMENTS

**Attachment No. 1** – Table S1:
List of sequence coordinates used in the study of *rrn* operons

**Attachment No. 2** – Supplemental results:
Detailed information about whole genome sequencing of *Treponema pallidum* ssp. *pertenue* CDC-2 and Gauthier strains

**Attachment No. 3 -** TPESAMD.tbl file:
Submitted TPE Samoa D gene annotation depicted in .tbl file

**Attachment No. 4** – Table S2:
A set of 95 genes newly predicted in the TPE Samoa D genome when compared to TPA Nichols genome annotation (AE000520.1)

**Attachment No. 5** – Table S3:
A set of 40 genes deleted in the TPE Samoa D genome when compared to TPA Nichols genome annotation (AE000520.1)

**Attachment No. 6 –** Table S4
A set of 107 genes renamed in the TPE Samoa D genome when compared to TPA Nichols genome annotation (AE000520.1)

**Attachment No. 7** – Table S5
Genetic variability within TPE (CDC-2, Samoa D and Gauthier) strains

# 9. ACRONYMS AND ABBREVIATIONS.

A – adenosine deoxyribonucleotide
AA – amino acid
BAC – bacterial artificial chromosome
bp – base pair
C – cytosine deoxyribonucleotide
CGS - comparative genome sequencing
DDT - dideoxy-terminator
DNA – deoxyribonucleic acid
dNTP – deoxyribonucleotide
ECM - extra-cellular matrix
G – guanosine deoxyribonucleotide
G4 - guanine quadruplex
GA - Genome Analyzer
GC - guanosine and cytosine deoxyribonucleotides
gDNA – genomic deoxyribonucleic acid
HGSC - Human Genome Sequencing Center (at Baylor College of Medicine, Houston, USA)
HMP - hypothetical membrane protein
HP – hypothetical protein
CHP – conserved hypothetical protein
IGR – intergenic region
indels – insertion or deletions
ISR - intergenic spacer regions
Ka - number of nonsynonymous substitutions per a nonsynonymous site
kb – kilo base pair
kDa – kilo Dalton
Ks - number of synonymous substitutions per a synonymous site
LB – Luria-Bertani
MID - multiplex identifiers
MSP – major sheath protein
NADH – β-Nicotinamide adenine dinucleotide, reduced dipotassium salt hydrate
NCBI - National Center for Biotechnology Information
NGS - next-generation sequencing
nr - non-redundant
OM – outer membrane
OMP - outer membrane protein
ORF – open reading frame
PCR – polymerase chain reaction
PM - Phenotype MicroArray
PSGS - pooled segment genome sequencing
rDNA – gene coding for rRNA
RDP - Recombination Detection Program
RFLP - restriction fragment length polymorphism
RNA – ribonucleic acid
*rrn* - ribosomal RNA (regarding operons)
rRNA – ribosomal ribonucleic acid
S – Svedberg unit
SMIL - small insert library
SNP - single nucleotide polymorphism
T – thymine deoxyribonucleotide
TCHMP - treponemal conserved hypothetical membrane protein
TCHOMP - treponemal conserved hypothetical outer membrane protein
TCHP - treponemal conserved hypothetical protein
TEN - *Treponema pallidum* ssp. *endemicum*
TGI - The Genome Institute (at Washington University in St. Louis, Saint Louis, USA)
TPI - *Treponema pallidum* intervals
TPA - *Treponema pallidum* ssp. *pallidum*
TPc - *Treponema paraluiscuniculi*
TPE - *Treponema pallidum* ssp. *pertenue*
*tpr* / Tpr - *Treponema pallidum* repeat
tRNA – transfer ribonucleic acid
WGA - whole genome amplification
WGF - whole genome fingerprinting
WHO - World Health Organization
XL PCR – extra-large polymerase chain reaction