OPEN ACCESS Freely available online

PLoS GENETICS

Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of *Saccharomyces cerevisiae*

Anthony R. Borneman[1]*, Brian A. Desany[2], David Riches[2], Jason P. Affourtit[2¤a], Angus H. Forgan[1], Isak S. Pretorius[1], Michael Egholm[2¤b], Paul J. Chambers[1]

1 The Australian Wine Research Institute, Adelaide, Australia, **2** 454 Life Sciences, A Roche Company, Branford, Connecticut, United States of America

Abstract

Human intervention has subjected the yeast *Saccharomyces cerevisiae* to multiple rounds of independent domestication and thousands of generations of artificial selection. As a result, this species comprises a genetically diverse collection of natural isolates as well as domesticated strains that are used in specific industrial applications. However the scope of genetic diversity that was captured during the domesticated evolution of the industrial representatives of this important organism remains to be determined. To begin to address this, we have produced whole-genome assemblies of six commercial strains of *S. cerevisiae* (four wine and two brewing strains). These represent the first genome assemblies produced from *S. cerevisiae* strains in their industrially-used forms and the first high-quality assemblies for *S. cerevisiae* strains used in brewing. By comparing these sequences to six existing high-coverage *S. cerevisiae* genome assemblies, clear signatures were found that defined each industrial class of yeast. This genetic variation was comprised of both single nucleotide polymorphisms and large-scale insertions and deletions, with the latter often being associated with ORF heterogeneity between strains. This included the discovery of more than twenty probable genes that had not been identified previously in the *S. cerevisiae* genome. Comparison of this large number of *S. cerevisiae* strains also enabled the characterization of a cluster of five ORFs that have integrated into the genomes of the wine and bioethanol strains on multiple occasions and at diverse genomic locations via what appears to involve the resolution of a circular DNA intermediate. This work suggests that, despite the scrutiny that has been directed at the yeast genome, there remains a significant reservoir of ORFs and novel modes of genetic transmission that may have significant phenotypic impact in this important model and industrial species.

Citation: Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, et al. (2011) Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of *Saccharomyces cerevisiae*. PLoS Genet 7(2): e1001287. doi:10.1371/journal.pgen.1001287

Editor: Gavin Sherlock, Stanford University, United States of America

Received July 22, 2010; **Accepted** December 30, 2010; **Published** February 3, 2011

Funding: The AWRI, a member of the Wine Innovation Cluster in Adelaide, is supported by Australian grapegrowers and winemakers through their investment body, the Grape and Wine Research Development Corporation, with matching funds from the Australian Government. Systems Biology research at the AWRI is performed using resources provided as part of the National Collaborative Research Infrastructure Strategy, an initiative of the Australian Government, in addition to funds from the South Australian State Government. The funders of this work had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: BAD, DR, JPA, and ME were employees of 454 Life Sciences, A Roche Company, at the time this work was performed.

* E-mail: anthony.borneman@awri.com.au

¤a Current address: Ion Torrent Systems, Guilford, Connecticut, United States of America
¤b Current address: Pall Corporation, Port Washington, New York, United States of America

Introduction

During its long history of association with human activity, the genomic makeup of the yeast *S. cerevisiae* is thought to have been shaped through the action of multiple independent rounds of wild yeast domestication combined with thousands of generations of artificial selection. As the evolutionary constraints that were applied to the *S. cerevisiae* genome during these domestication events were ultimately dependent on the desired function of the yeast (e.g baking, brewing, wine or bioethanol production), these multitude of selective schemes have produced large numbers of *S. cerevisiae* strains, with highly specialized phenotypes that suit specific applications [1,2]. As a result, the study of industrial strains of *S. cerevisiae* provides an excellent model of how reproductive isolation and divergent selective pressures can shape the genomic content of a species.

Despite their diverse roles, industrial yeast strains all share the general ability to grow and function under the concerted influences of a multitude of environmental stressors, which include low pH, poor nutrient availability, high ethanol concentrations and fluctuating temperatures. In comparison, non-industrial isolates such as laboratory strains, have been selected for rapid and consistent growth in nutrient rich laboratory media, thereby producing markedly different phenotypic outcomes when compared to their industrial relatives [3]. The outcomes of these very different selection pressures are therefore most evident when comparing industrial and non-industrial yeasts. As an example, laboratory strains of *S. cerevisiae*, such as S288c, are unable to grow in the low pH and high osmolarity of most grape juices and therefore cannot be used to make wine. This is a clear difference between industrial and non-industrial strains of *S. cerevisiae*, however there are numerous subtle differences not only between industrial strains, but also between strains used within the same industry [4,5], highlighting the overall genetic diversity found in this species.

Author Summary

The yeast *S. cerevisiae* has been associated with human activity for thousands of years in industries such as baking, brewing, and winemaking. During this time, humans have effectively domesticated this microorganism, with different industries selecting for specific desirable phenotypic traits. This has resulted in the species *S. cerevisiae* comprising a genetically diverse collection of individual strains that are often suited to very specific roles (e.g. wine strains produce wine but not beer and vice versa). In order to understand the genetic differences that underpin these diverse industrial characteristics, we have sequenced the genomes of six industrial strains of *S. cerevisiae* that comprise four strains used in commercial wine production and two strains used in beer brewing. By comparing these genome sequences to existing *S. cerevisiae* genome sequences from laboratory, pathogenic, bioethanol, and "natural" isolates, we were able to identify numerous genetic differences among these strains including the presence of novel open reading frames and genomic rearrangements, which may provide the basis for the phenotypic differences observed among these strains.

There have been several attempts to characterize the genomes of industrial strains of *S. cerevisiae* which have uncovered differences that included single nucleotide polymorphisms (SNPs), strain-specific ORFs and localized variations in genomic copy number [6–14]. However, the type and scope of genomic variation documented by these studies were limited either by technology constraints (e.g arrayCGH relying on the laboratory strain as a "reference" genome), or by the resources required for the production of high-quality genomic assemblies which has limited the scope and number of whole-genome sequences available for comparison. In addition, to limit genomic complexity to a manageable level, previously published whole-genome sequencing studies on industrial strains used haploid representations of diploid, and often heterozygous, commercial and environmental strains [9–13].

We sought to address these shortcomings by sequencing the genomes of four wine and two brewing strains of *S. cerevisiae* in their industrially-used forms. The industries of winemaking and brewing were targeted for this work as they have the longest association with *S. cerevisiae* (measured in the thousands of years) and each industry has accumulated large numbers of phenotypically distinct strains for which genetic comparisons can be made. This study demonstrates that industrial yeasts display significant genotypic heterogeneity both between strains, but also between alleles present within strains (i.e. heterozygosity). This variation was manifest as SNPs, small insertions and deletions, and as novel, strain and allele-specific ORFs, many of which had not been found previously in the *S. cerevisiae* genome and may provide the basis for novel phenotypic characteristics. Interestingly, several ORFs were shown to comprise a gene cluster that was present in multiple copies and at a variety of genomic loci in a subset of the strains examined. Furthermore, this cluster appears to have integrated into genomic locations by a novel circular intermediate, but without employing classical transposition or homologous recombination, which we believe represents the first time such an element has been characterized in *S. cerevisiae*.

Overall, this work suggests that, despite the scrutiny that has been directed at the yeast genome, there remains a significant reservoir of ORFs and novel modes of genetic transmission which may have significant phenotypic impact in this important model and industrial species.

Results

Six industrial yeasts were chosen for genomic analysis, comprising four commercial wine strains and two brewing strains used for the production of ales (ale strains are primarily *S. cerevisiae*, while lager-style brewing strains are *S. pastorianus*, a hybrid of *S. cerevisiae* and *S. bayanus* [15,16]). These six strains were sequenced to an average coverage of 20 fold with a combination of shotgun and paired-end methods using the GS FLX Titanium series chemistry [17], which resulted in six high quality genomic assemblies (Table 1).

Large chromosomal variations in industrial yeast strains

Rather than being strictly diploid, many industrial yeast strains display chromosomal copy number variation (CNV) [18]. In order to catalogue CNV in the industrial yeast genomes, the depth of sequencing coverage determined for each sequence contig were calculated such that areas of CNV could be detected as localized variations in that coverage (Figure 1). There were several large areas of increased copy number across the strains including six potential whole-chromosome amplifications (chrI of AWRI796, chrVIII of VL3, chrIII of FostersO and chrIII, V and XV of FostersB) and one potential reduction in chromosomal copy number (chrXIV of FostersO). There were also several partial chromosomal CNVs, including amplification of 200 kb of chrXIV in AWRI796, 600 kb of chrII and 200 kb of chrX in FostersO and a 400 kb reduction from chrVII of FostersO (Figure 1). However, while the ale strains had a higher number of large CNVs than wine strains, the overall fold change of these CNVs was generally reduced. This reduction can be most easily explained by the brewing strains having a polyploid genetic base while the wine strains are diploid, an observation which has been seen previously in these industrial yeasts [18].

Heterozygosity in industrial strains

As existing published industrial yeast genome sequences were either generated from haploid derivatives of industrial strains [9–12] or had heterozygous regions discarded during analysis [13], the level of genome-wide heterozygosity present in industrial strains remains largely unknown. However, as the assemblies performed in this study retained genomic heterozygosity, it was possible to determine the level of allelic differences within each of these strains (Table 2). While every industrial strain contained heterozygous single nucleotide polymorphisms (SNPs), the proportion of these varied over thirty-fold between wine strain AWRI796 (1041 total heterozygous bp) and the brewing strain FostersB (33071 bp). Heterozygous insertions and deletions (InDels) were also present and ranged from single base pair variants to large InDels of up to 35.3 kb. Strains were also shown to contain heterozygous instances of Ty element insertion, although, due to the repetitive nature of these elements, their presence in the genome could generally only be estimated through paired-end information (data not shown).

Nucleotide variation present in *S. cerevisiae*

In addition to the intra-strain variation that was present between homologous chromosomes within individual strains, there was also significant nucleotide variation between strains. As seen for the allelic variation, both SNPs and InDels were found between strains, with inter-strain InDels of up to 45 kb being observed. Many of the smaller InDels (both heterozygous and homozygous) were located in regions comprising tandem repeats (Figure 2A, Table S1) and primarily in the expansion and contraction of di- and tri-nucleotide tandem repeats (Figure 2B). Indeed, when using

Table 1. Strains sequenced in this study.

Strain	Industry	Supplier	Contigs[a]	N50[a] (kb)	Scaffolds[a]	Assembly size[a]	Genbank Accession[b]
Lalvin QA23	wine	Lallemand Inc.	96	185	39	11.6 Mb	ADVV00000000
AWRI796	wine	Maurivin	49	409	31	11.6 Mb	ADVS00000000
Vin13	wine	Anchor Bio-Technologies	80	308	29	11.5 Mb	ADXC00000000
FostersO	brewing (ale)	Fosters Group Ltd.	95	219	35	11.4 Mb	AEEZ00000000
FostersB	brewing (ale)	Fosters Group Ltd.	78	209	25	11.5 Mb	AEHH00000000
VL3	wine	Laffort	70	316	29	11.4 Mb	AEJS00000000

a Excluding repetitive sequencing contigs such as sub-telomeric regions and Ty elements.
b These Whole Genome Shotgun projects have been deposited at DDBJ/EMBL/GenBank. The versions described in this paper are the first versions; ADVV01000000, ADVS01000000, ADXC01000000, AEEZ01000000, AEHH01000000 and AEJS01000000.
doi:10.1371/journal.pgen.1001287.t001

chromosome XVI as an example, over 86% of the instances of di- and tri-nucleotide repeats displayed variable length in at least one of the strains. As the size of tandem repeats has been associated with differences in gene expression [19], this suggests that there are both strain and allele-specific differences in the expression of genes proximal to these repeat-associated InDel events.

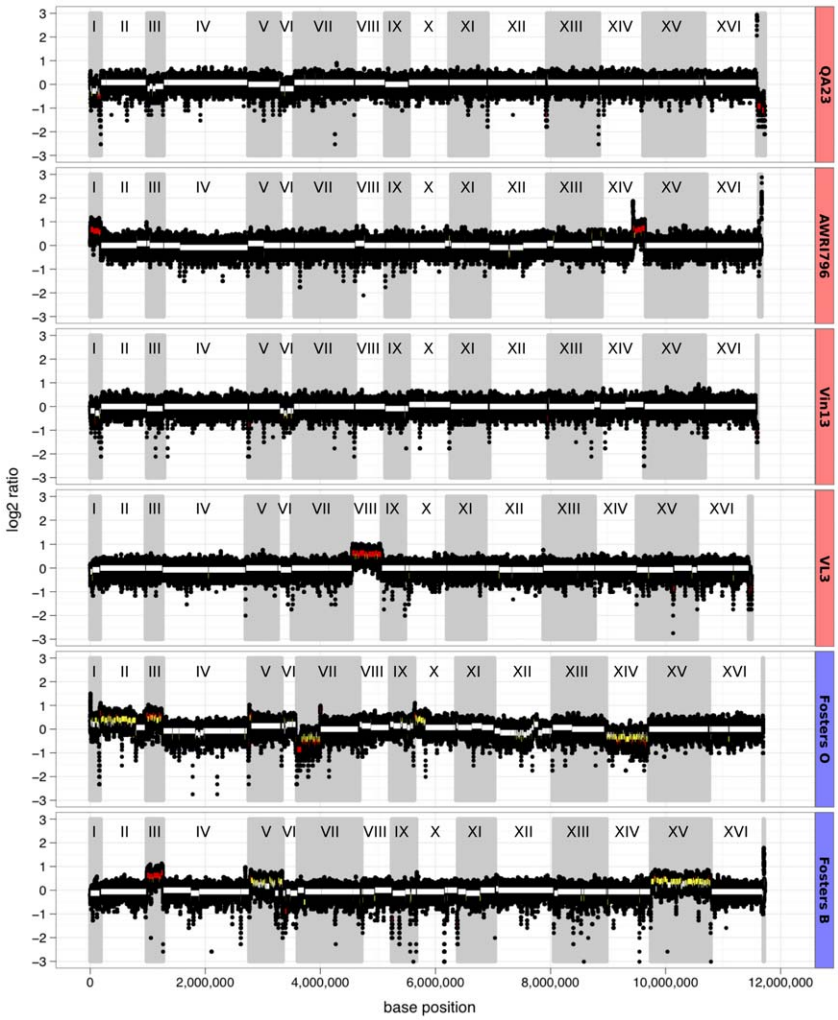

Figure 1. Chromosomal aneuploidy determined by whole-genome sequencing coverage. Sequencing coverage was determined for each contig using a sliding window of 1001 bp, with a 100 bp step frequency and plotted in chromosomal order (black circles). Regions of copy number variation were scored as either being greater than 1.25-fold (yellow lines; approximating either three or five copies in a tetraploid genome) or 1.5-fold (red lines; one or three copies in a diploid genome) different to the median coverage for that strain. Strains are shaded according to their industry (wine, red; ale, blue).
doi:10.1371/journal.pgen.1001287.g001

Table 2. Heterozygosity in industrial *S. cerevisiae* strains.

Strain	Origin	Ploidy	Homozygous SNPs[a]	Heterozygous SNPs[a]
S288C	Lab	1n	41708	0
YJM789	Human isolate	1n	40675	0
JAY291	Bioethanol	1n	25648	0
RM11-1a	Vineyard isolate	1n	10825	0
EC1118	Wine	1n[b]	13241	0
AWRI1631	Wine	1n	9935	0
QA23	Wine	2n	4913	18861
AWRI796	Wine	2n	8996	1041
Vin13	Wine	2n	3544	15216
VL3	Wine	2n	5108	9904
FostersO	Ale	>2n[c]	25802	27215
FostersB	Ale	>2n[c]	23125	33071

a SNPs were calculated relative to the most common base across all twelve strains at each position.
b EC1118 is a diploid commercial strain but the available sequence is a haploid representation of this genome.
c As estimated from overall sequencing coverage.
doi:10.1371/journal.pgen.1001287.t002

SNP variation was also common throughout the strains with a total of 165,913 non-degenerate SNPs (unique points of nucleotide variation) that were present in at least one allele of the twelve strains investigated (~1.3% of the total genome length). However, given the influence of large, strain-specific InDels (which were filtered out of the SNP analysis) the apparent SNP density is much higher than 1.3%, such that these SNPs were shown to display a median inter-SNP distance of only 37 bp.

By using the number of SNPs separating any two isolates as an estimation of their relatedness (Figure 3A), we were able to show that industrial yeasts are distinct from both the laboratory and human pathogenic strains and were also found to group by industry. This was especially true of the brewing strains which displayed a high degree of genetic distance not only from the laboratory and human isolates, but also from the wine and bioethanol strains. The only exception to this pattern of grouping by industry or environment niche was with the 'natural' isolate RM11-1a which grouped closely with wine strains. However, given that it is descended from a strain sourced from a vineyard, RM11-1a may well share genetic origins with those strains used in winemaking.

In order to put the genetic variation observed in these genomic alignments in a larger population context, twelve strains were selected to represent each of the six main *S. cerevisiae* population groups as proposed by Liti et al [12] for further SNP comparison (Figure 3B). In this broader context, wine strains sequenced in this study were shown to also group tightly with the wine/European strains DBVPG1106 and DBVPG1373, showing that the data produced across these two studies are directly comparable. However, while the ale strains were still shown to be distinct from the wine isolates they were found to be far closer to the wine strains than isolates such as those used in sake production, which display the greatest level of nucleotide diversity when compared to the wine strains. Indeed, when the SNP data from these additional strains in included in the calculations of SNP density, the total number of non-degenerate SNPs increases to 216,207 (~1.7%) with a median inter-SNP distance of only 27 bp. However, despite comparisons to eighteen other diverse strains of *S. cerevisiae* 15,576 of these SNPs were found solely in this study (2,501 in more than one strain) and with the vast majority of these SNPs being present in a heterozygous form (only 1,864 novel SNPs were homozygous in at least one strain).

ORF conservation across *S. cerevisiae*

To determine how inter-specific variation at the nucleotide level translated into protein-coding differences, the predicted coding potential of each strain was compared. ORFs were predicted from each sequence (including the pre-existing whole genome sequences) using Glimmer [20] and compared using a combination of BLAST [21] homology matches and genomic synteny to differentiate instances of orthology from gene duplication (Table S2). When using the laboratory strain S288c as a reference, there was an average of 92% ORF coverage across the strains. The majority of S288c ORFs without a match in other strains were shown to be located in repetitive regions of the *S. cerevisiae* genome such as in the sub-telomeric zones or the numerous Ty retrotransposons that are present in S288c genome relative to other strains. Due to the repetitive nature of these regions it was

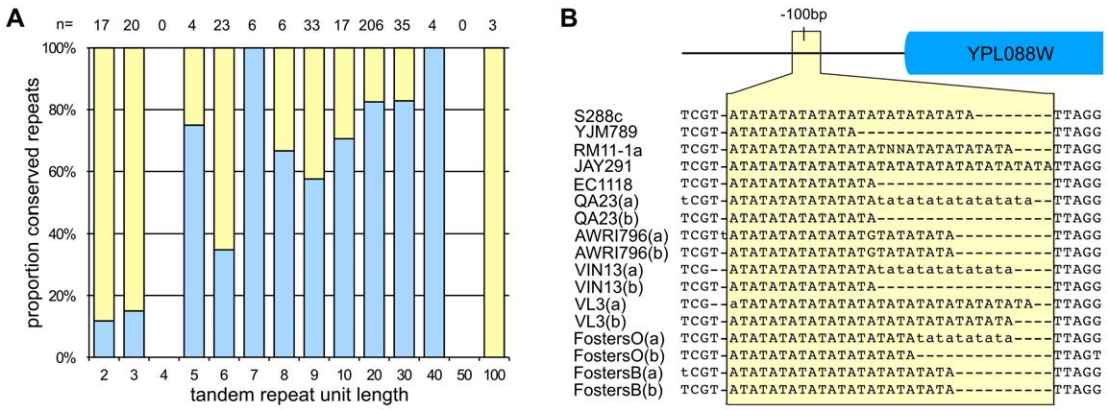

Figure 2. Nucleotide variation in *S. cerevisiae*. (A) InDels associated with tandem repeats. Histogram showing the proportion of tandem repeats of various sizes (repeated size indicated on x-axis) present on chrXVI that were either conserved in repeat length (blue) or contained strain-specific InDels (yellow). The total number of repeat loci present in each class is listed above the histogram. (B) An example of a strain- and allele-specific InDel in a tandem repeat in the promoter region of YPL088W.
doi:10.1371/journal.pgen.1001287.g002

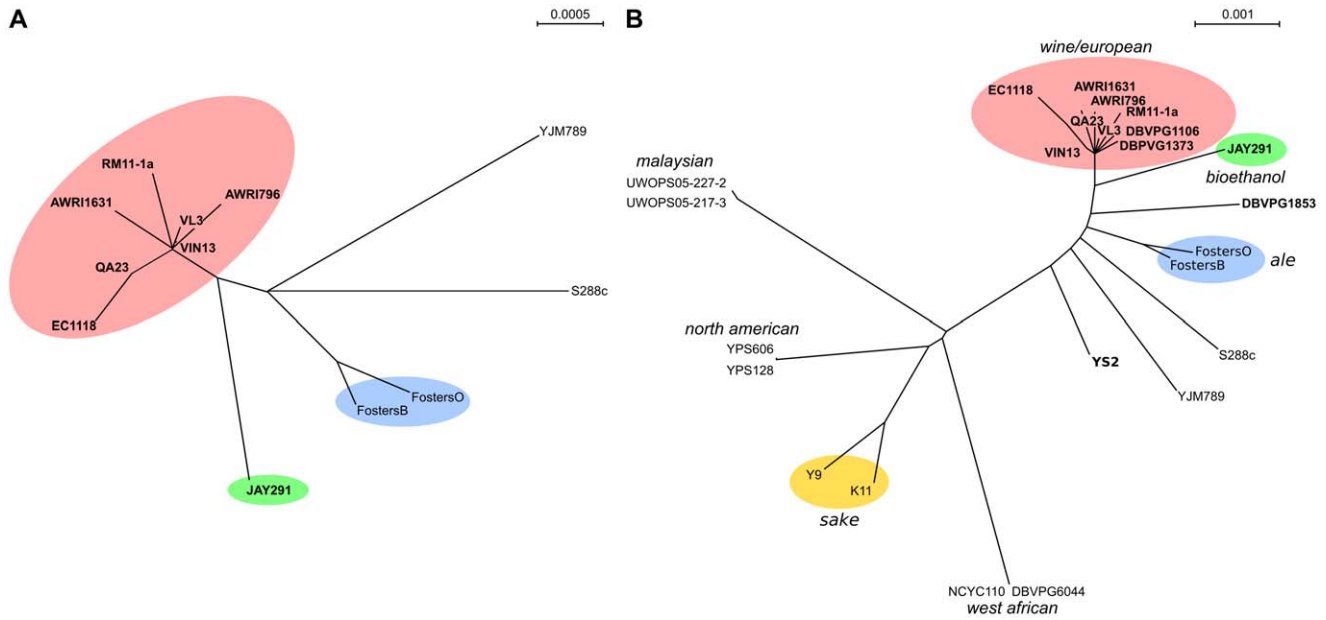

Figure 3. Nucleotide relationships between *S. cerevisiae* strains. (A) A neighbor joining tree representing the genetic distance between strains as calculated from the total SNP diversity present in whole genome alignments. (B) A neighbor joining tree representing the genetic distance between strains presented in part (A) and representative strains from several *S. cerevisiae* geographical populations [12]. Industrial strains are color-coded based upon their primary industry (wine/European, including RM11-1a, pink; ale, blue; bioethanol, green; sake, yellow). Strains that are predicted to contain the heterogeneous five-gene cluster are labeled in bold.
doi:10.1371/journal.pgen.1001287.g003

often impossible to unambiguously position these sequences in the industrial yeast genome assemblies and they remain within repetitive, unmappable contigs in the various genome assemblies. It therefore appears that, due to its persistent propagation in the laboratory, the genome of S288c may represent a reduced genomic state as it does not appear to contain additional genes that provide unique metabolic or cellular potential outside of those present in other strains. It does however contain a far greater number of Ty transposons relative to all of the other strains suggesting that transposon proliferation occurred on at least one occasion during the development of this laboratory strain.

Novel ORFs

While the laboratory strain S288c is considered the reference for the genomic complement of *S. cerevisiae*, it is becoming apparent that it lacks a multitude of ORFs which exist in other strains of *S. cerevisiae* [9–13,22,23]. This is confirmed n the present study with between 36 (FostersB) and 110 (Lalvin QA23) ORFs lacking significant homology to the S288c genome but for which there were clear matches to sequences in other *S. cerevisiae* strains or microbial species (Table S2). Orthologs of 102 out of 218 of the non-degenerate set of these 'non-S288c' ORFs have been identified previously in *S. cerevisiae* strains, mainly through whole-genome sequencing of AWRI1631, EC1118 and RM11-1a and YJM789 [8,9,13] (Table S2). These include genes encoding proteins such as the Khr1 killer toxin [24] which is found in YJM789, EC1118, Vin13, VL3, FostersB and FostersO and orthologs of the *MPR1* stress-resistance gene (which was originally identified in the Sigma 1278b strain[23]) in RM11-1a, EC1118, AWRI1631, JAY291, QA23 and VL3.

Interestingly, in addition to these ORFs there were at least three proteins present in the human pathogen YJM789 and the FostersB and FostersO ale strains but which were lacking from the wine, biofuel and laboratory strains (Figure 4C). These included the YJM-GNAT GCN5-related N-acetyltransferase [8] and a separate gene cluster which is predicted to contain both *RTM1*, which was identified previously as a distillery-strain specific gene that provides resistance to an inhibitory substance found in molasses [22], and a large ORF of around 2.3 kb which, despite its large size and high-degree of conservation across the brewing and human pathogenic strains, lacks significant homology to any other protein sequences except for six isolates from the large *S. cerevisiae* population genomic screen which also appear to encode this protein [12] (Figure S1). In addition to these two conserved ORFs, in the ale strains this cluster also appears to encode an invertase that would be expected convert sucrose into the sugars glucose and fructose.

Despite the presence of at least two existing high-coverage wine strain sequences and at least an additional six low coverage genomes, the entire repertoire of ORFs present in wine strains of *S. cerevisiae*, let alone the species as a whole, is far from complete. In addition to expanding the strain range of previously identified non-S228c proteins, it was possible to identify at least eleven ORFs that lacked homology to existing proteins from *S. cerevisiae*, in addition to many new paralogs of existing *S. cerevisiae* genes. These novel ORFs often clustered in large InDels, the largest of which was a 45 kb fragment in the wine strain AWRI796. This novel genomic region is located adjacent to a large repetitive element present on chromosomes XIII, XV and XVI, which hampered initial efforts to assign this region to a specific chromosome. However, through the application of a 20 kb paired-end library, it was possible to bridge the repetitive region and position this novel region at the end of the right arm of chromosome XV. This fragment is predicted to encode nineteen ORFs (Figure 4A), three of which are predicted to encode aryl-alcohol dehydrogenases (AADs). AADs have been extensively characterized in filamentous fungi where they catalyze the reversible reduction of aldehydes and ketones to aromatic alcohols during lignin-degradation

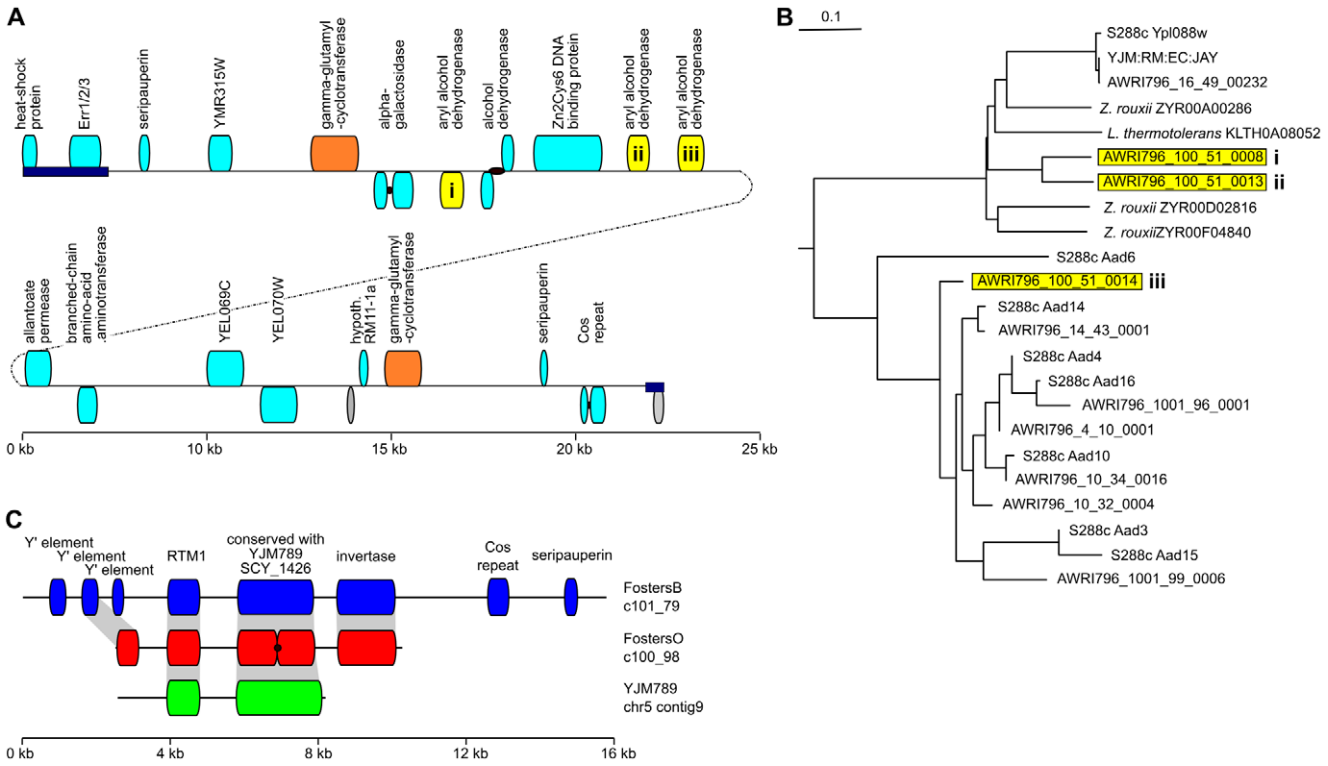

Figure 4. Novel genes found in industrial strains. (A) A 45 kb strain-specific region in AWRI796 which is predicted to encode at least 21 ORFs (full ORF sequences are listed in Dataset S12). ORFs with homology to AADs are highlighted in yellow. The extreme 5′ and 3′ ends of this cluster are homologous to a repetitive region present in the sub telomeric regions of chrXIII, XV and XVI (dark blue boxes). Black dots within ORFs represent potential frameshifts in the sequence of these regions. (B) Clustalw dendrogram produced by aligning AAD proteins from S288c, AWRI796 and the top five matches to the highly divergent AWRI796 proteins AAD(i) and AAD(ii). (C) The region in the brewing strains FostersO and FostersB containing *RTM1* [22] and the conserved hypothetical ORFs are also found in the human pathogen YJM789 [8].
doi:10.1371/journal.pgen.1001287.g004

[25,26]. These new AAD homologs are phylogenetically distinct from other AAD enzymes that have been identified, including the seven predicted AADs that are present in the S288c genome [27,28] (Figure 4B).

Characterization of a novel, and potentially transmissible, gene cluster

One particularly curious feature of many of the industrial yeast strains analyzed in this study, was a cluster of five conserved ORFs that was present in all of the wine strains, RM11-1a and the bioethanol strain JAY291, and potentially in at least four of the strains present in the Liti et al [12] study (Figure 3). This cluster is predicted to encode two potential transcription factors (one zinc-cluster, one C_6 type), a cell surface flocullin, a nicotinic acid permease and a 5-oxo-L-prolinase, and has been suggested to be horizontally acquired by *S. cerevisiae* from *Zygosacharomyces spp* [13]. In this study we have been able to show that while the sequences of the individual genes within this cluster are highly conserved between strains, the cluster itself is actually highly diverse with respect to copy number, genomic location and overall gene order (Figure 5, Table S3). The cluster was present in one to at least three copies across strains, with individual clusters being located in at least seven different genomic loci (Figure 5A). For example, wine strain Lalvin QA23 was shown to contain at least three copies of the cluster, found in three different genomic loci and with at least two copies being heterozygous. However, despite this diversity, the sequence of the ORFs and intergenic regions of the cluster were highly conserved, with only fifteen nucleotide substitutions (0.01%) recorded across the eleven known copies of the cluster (Figure 5B, Figure S2).

In addition to the differences in copy number and location, the exact order of the ORFs within the cluster differed in a location dependent manner (Figure 5B, 5C). However, all of these different ORF arrangements could be resolved into a syntenically-conserved order if the linear genomic copy of each cluster resulted from the differential resolution of a common circular intermediate, with a unique breakpoint in this circular arrangement being observed for each genomic location (Figure 5B–5D). However, despite the differential location of these clusters these integration events appear to select for functional conservation of the genes with the majority of the breakpoints being located within intergenic regions (Figure 5B). Of the two exceptions to this, one of these events occurs at the extreme 3′ end (~100 bp from the predicted stop codon) of one ORF such that a functional protein is likely to still be produced from this gene.

Adding further interest to the mode of transfer of this cluster, its integration into the genome appears to occur without the production of the terminal repeated sequences that would be expected if integration of this element occurred by either homologous recombination or classical mobilization via a transposon-like mechanism. In fact, for at least three of the seven different integration events characterized in this study, integration of the cluster has occurred between two directly adjacent, conserved nucleotides, with a further two events showing only single nucleotide indels at the junction between the cluster and the flanking genomic sequences (Figure 5E).

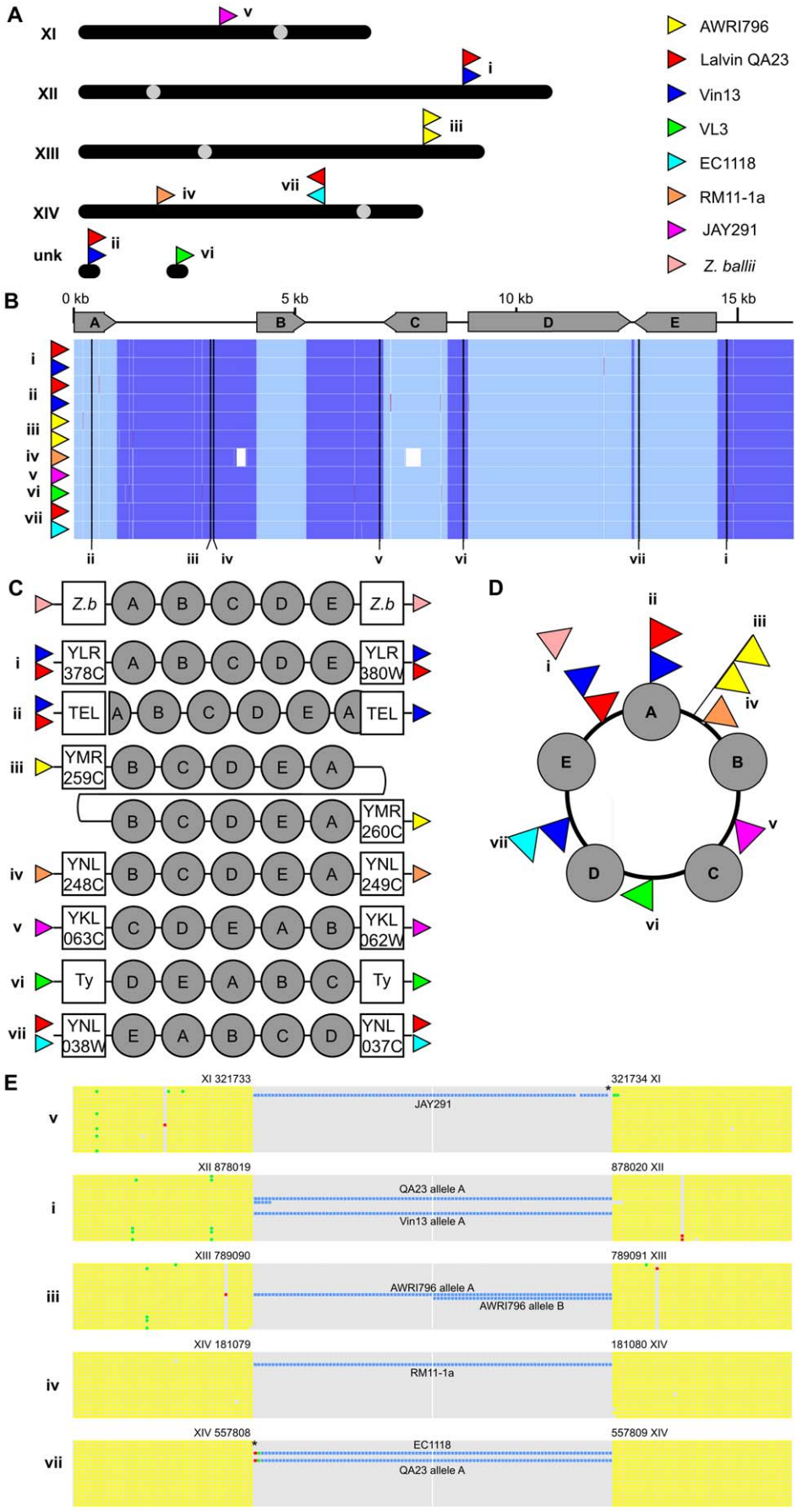
A
XI
XII
XIII
XIV
unk
v
i
iii
iv
vii
ii
vi
AWRI796
Lalvin QA23
Vin13
VL3
EC1118
RM11-1a
JAY291
Z. ballii
B
0 kb
5 kb
10 kb
15 kb
A
B
C
D
E
C
Z.b
YLR 378C
YLR 380W
TEL
YMR 259C
YMR 260C
YNL 248C
YNL 249C
YKL 063C
YKL 062W
Ty
YNL 038W
YNL 037C
D
E
XI 321733
321734 XI
JAY291
XII 878019
878020 XII
QA23 allele A
Vin13 allele A
XIII 789090
789091 XIII
AWRI796 allele A
AWRI796 allele B
XIV 181079
181080 XIV
RM11-1a
XIV 557808
557809 XIV
EC1118
QA23 allele A

Figure 5. A divergent cluster of genes with a possible circular intermediate. (A) The location and orientation of the gene cluster throughout the genomes of the industrial yeasts. Upper case roman numerals refer to standard *S. cerevisiae* chromosomes (unk – location unknown) with individual loci labeled with lower case roman numerals. (B) Nucleotide conservation of the five-gene clusters. An alignment of the nucleotide sequence of all eleven clusters is shown below a schematic depiction of the five predicted ORFs present in this nucleotide sequence (A, zinc-cluster transcription factor; B, cell-surface flocculin; C, nicotinic acid permease; D, 5-oxo-L-prolinase; E, C_6 transcription factor). In order to produce contiguous alignments, the sequence of each cluster was manually split to begin with the start codon of ORF A, with the position of each break indicated. Conserved bases are shaded blue (light blue for ORFs sequences). Insertions are highlighted in red and substitutions in green. (C) Differences in gene order within individual clusters. Each of the five genes are represented by filled circles (labeled as in partB), with the systematic name of the ORFs that border each insertion listed in open squares (*Z.b*, this cluster is present in *Z. bailii* (Accession number FN295481.1); Ty, transposon sequence; TEL, sub-telomeric repeat (COS) sequence). Colored arrows bordering each cluster indicate the strain(s) in which this insertion is present. (D) Each of the nine cluster locations and orders can be resolved through the use of a circular intermediate that integrates into the genome via breakage at locations indicated by each colored triangle. (E) Conservation of genomic sequences flanking individual cluster insertion events. Nucleotide alignments are shown for the 50 bp directly adjacent to either side of the five chromosomally-mapped insertion events (shaded yellow when conserved) in addition to the first and last 50 bp of the each cluster (shaded according to partB). Insertions are shaded in red, substitutions in green with both additionally highlighted by asterisks. Sequences used for the alignment are (from top to bottom) S228c, JAY291, RM11-1a, EC1118, AWRI1631, QA23 allele A, QA23 allele B, AWRI796 allele A, AWRI796 allele B, Vin13 allele A, Vin13 allele B, VL3 allele A, VL3 allele B, Fosters B allele A, Fosters B allele B, Fosters O allele A, Fosters O allele B. Nucleotide coordinates for the bases directly flanking the insertion are relative to the S288c genome.
doi:10.1371/journal.pgen.1001287.g005

Discussion

While *S. cerevisiae* is one of the most intensively studied biological model organisms and economically-important industrial microorganisms, many characteristics of its genome remain unknown, especially in strains other than the laboratory reference S288c. Through the analysis of six industrial strains, it was possible to show that the industrial members of this species are distinct, with wine and brewing strains being almost as distantly related at the DNA level as they are to either the laboratory or human pathogenic strains. This suggests that despite their roles in performing industrial fermentations, the two groups comprise genetically separate *S. cerevisiae* lineages. While this is a situation similar to that proposed previously for wine and sake strains of *S. cerevisiae* [2], the wine and ale strains were much more closely related to each other than to strains with origins outside of Europe [12], and this may reflect a distant common European-type ancestor. The bioethanol strain JAY291 displays an intermediate level of sequence relatedness to the wine strains (compared to ale strains) and also contains the five-gene cluster, suggesting that this strain shares at least some of its genomic origins with the wine isolates. With the relatively recent development of the bioethanol industry, it is not entirely unexpected that yeasts used in this process may well have their origins in commercial strains used in established ethanologenic industries. Wine strains would therefore make a logical choice for this starting point given their highly efficient production of ethanol and relatively high tolerance to a variety inhibitory substances, such as ethanol or polyphenols, that also exist in bioethanol fermentations [29].

In addition to mapping the relationships between these strains, this study uncovered a number of genetic elements not previously identified in the *S. cerevisiae* genome, as well as expanding the range of several strain specific elements that had been identified previously. This highlights the fact that the genetic variation that underlies the phenotypic diversity of *S. cerevisiae* goes well beyond that of SNPs or small InDels and is similar to the situation observed with many bacterial species where the pan (species-wide) genome is larger than that observed in any single strain [30]. As for the situation observed with single nucleotide variation, several of these genetic elements link strains to specific industries (e.g. the *RTM1* cluster in the ale strains and the five-gene cluster in the wine strains). It would therefore be expected that these ORFs provide selective advantage within specific industries that have favored their retention. For some of these ORFs, such as the *RTM1* cluster, the phenotypic benefits that they have historically provided in one industry may be advantageous in modern incarnations of others. For example, modern wine production generally makes use of inoculated commercial strains (rather than the historical use of wild yeast), which are produced on a large scale using molasses as a feedstock. Genes such as the *RTM1* cluster may therefore provide advantages in the production of modern commercial wine yeast, but which are lacking from the genomic complement of this group of strains due to the historical practices of winemaking.

While other strain-specific ORFs were shown to have much narrower strain ranges (often single strains), it was possible to predict industrially-relevant roles for some of these genes. For example, the novel AAD proteins that were identified in the wine strain AWRI796 may have a direct impact on the range of volatile aromas produced during fermentation, as the aromatic alcohols produced through the action of the AAD enzymes can present very different aromas profiles to their corresponding aldehydes and ketones [31]. The presence of these AADs in specific industrial yeasts may therefore alter the profile of volatile aromas produced during winemaking or brewing, contributing to strain-specific aroma characteristics that are vitally important to many flavor and aroma-based industrial applications.

The role of ORFs such as those present in the wine yeast five-gene cluster are less clear but, given the potential regulatory role for at least two of these proteins, they could produce significant phenotypic effects. The generally similar characteristics of high sugar and ethanol tolerance of *Zygosacharomyces spp* and the wine and bioethanol strains of *S. cerevisiae* [29,32], may provide a selective advantage for growth under these conditions. However, understanding the function of individual ORFs is overshadowed by questions regarding the origins of this novel cluster in addition to its effect on genome structure and dynamics. It was recently proposed that this cluster entered the *S. cerevisiae* genome from *Zygosacharomyces spp* [13]. Our data suggests that if this is the case, the transfer has either occurred on multiple occasions via a conserved circular intermediate that has integrated randomly into different genomic loci, or the fragment has entered the *S. cerevisiae* genome on a single occasion but has subsequently mobilized to new genomic locations via a circular intermediate (Figure S3). Alternatively, this cluster is a mobile feature of the *S. cerevisiae* genome that has been lost from many strains and was transferred to *Zygosacharomyces spp*. Regardless of the direction or precise mode of transfer it appears that this genetic cluster may mobilize throughout the genome via a method which has yet to be characterized in yeast and therefore provides an entirely new mechanism for the generation of variation in the *S. cerevisiae* genome.

A thorough understanding of the scope of plasticity of the yeast genome is a vital prerequisite for the systematic understanding of yeast biology or for the development of the next generation of yeasts for industrial applications. As more *S. cerevisiae* strains are sequenced, the suitability of S288c as a "reference" strain for this species is becoming less clear, especially as it appears to lack a large numbers of ORFs found in many other *S. cerevisiae* strains while containing an abnormally high number of Ty transposable elements [8,9]. Given the ubiquitous nature of the S288c genome for the design of 'omics experiments, these novel elements have generally not been considered when studying strains other than S288c. Thus, little data exists regarding the functional contributions of these proteins. As such, they represent a significant knowledge gap with respect to cellular and metabolic modeling strategies. This is especially true for proteins such as the ORF located next to *RTM1* which is large (~800 amino acids) and highly conserved but has no significant homologs outside of a small subset of *S. cerevisiae* strains on which a function can be based. Fortunately, the continued development of next generation sequencing, such as that applied in this work, have provided the means to now characterize large numbers of yeast strains to provide this information and outline the true scope and variability of this species.

Materials and Methods

Yeast strains

Each commercial strain was obtained from the original mother cultures from the supplier. Genomic DNA was prepared by zymolase digestion and standard phenol-chloroform extraction.

Sequencing and assembly

Library construction and sequencing was performed at 454 Life Sciences, A Roche Company (Branford, CT) using a pre-release development version of the GS FLX Titanium series shotgun and 3 kb paired-end protocols. Sequences were assembled using MIRA (http://sourceforge.net/apps/mediawiki/mira-assembler/index.php?title = Main_Page) and manually-edited using Seqman Pro (DNAstar).

Regions of chromosomal CNV were determined by calculating the per-base sequencing coverage across each sequencing contig with median smoothing (1001 bp window, 100 bp step size). The ratio between the coverage at each genomic location and the overall median genomic coverage was the calculated to determine the level of over-representation for each location. Large-scale chromosomal aneuploidies were detected by screening for regions in which median ratio for a contiguous stretch of at least 101 individual segments differed from the overall genomic median by either 1.25 (5:4 ratio representing at least 1 extra genomic copy in a tetraploid) or 1.4 fold (3:2 ratio representing at least 1 extra genomic copy in a diploid).

SNP prediction

Chromosomal scaffolds from each yeast strain were aligned using FSA [33]. Diploid sequences were assigned into two haploid alleles by converting any degenerate bases into their non-degenerate pairs. Heterozygous regions were divided into both an insertion and deletion allele. A chromosomal consensus was computed for the alignment based upon the most frequent allele at each position in the alignment. Nucleotides that varied from the consensus in each strain were scored as sequence variants and were subsequently divided into SNPs (nucleotide substitution) or InDels (nucleotide insertion or deletion). To enable the comparison to strains with low coverage sequences [12], SNPs that were calculated for each strain relative to S288c (imputed SNPs) were used to create synthetic S288c-based genome sequences that contain the SNPs present in these strains. The genetic relationship between the strains was calculated by editing and concatenating the nucleotide alignments of all sixteen chromosomes using Seaview [34] followed by calculating the distance tree using the NJ algorithm of Clustalw (ignoring gapped regions in the alignment). Tandem repeats were predicted from the chromosomal alignment of all twelve yeast strains using Tandem Repeats Finder [35] using default parameters (match weight, 2; mismatch, 7; indel, 7; *pM*, 0.80; *pI*, 0.10; minimum alignment score, 50; maximum period size, 500). Individual repeats were then scored as either being variable if the specific tandem repeat region contained strain- or allele- specific InDels.

ORF prediction and comparison

ORFs were predicted using Glimmer [20] with the predicted ORFs of S288c being used to build the prediction model (See Datasets S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11 for actual CDS sequences for each strain). Initial ORF designations were made by identifying the best sequence match for each ORF when compared to S288c using BLASTn [21]. Glimmer was also used to predict ORFs from the sequence of S288c (Accession numbers NC001133-NC001148) to correct for false-negatives in the predictions when compared to existing ORF designations in S288c. ORFs with no match to S288c were searched against the full list of non-redundant Genbank proteins to identify a closest existing homology match. ORFs from each strain were then arranged in syntenic order (Table S2 for a full list of ordered ORFs). For protein sequence comparisons, predicted protein sequences were aligned using Clustalw [36] (http://align.genome.jp).

Supporting Information

Dataset S1 Gllimmer-predicted ORFs from S288c.
Found at: doi:10.1371/journal.pgen.1001287.s001 (8.86 MB TXT)

Dataset S2 Gllimmer-predicted ORFs from YJM789.
Found at: doi:10.1371/journal.pgen.1001287.s002 (8.84 MB TXT)

Dataset S3 Gllimmer-predicted ORFs from JAY291.
Found at: doi:10.1371/journal.pgen.1001287.s003 (8.49 MB TXT)

Dataset S4 Gllimmer-predicted ORFs from RM11-1a.
Found at: doi:10.1371/journal.pgen.1001287.s004 (8.57 MB TXT)

Dataset S5 Gllimmer-predicted ORFs from EC1118.
Found at: doi:10.1371/journal.pgen.1001287.s005 (8.54 MB TXT)

Dataset S6 Gllimmer-predicted ORFs from QA23.
Found at: doi:10.1371/journal.pgen.1001287.s006 (8.15 MB TXT)

Dataset S7 Gllimmer-predicted ORFs from AWRI796.
Found at: doi:10.1371/journal.pgen.1001287.s007 (7.94 MB TXT)

Dataset S8 Gllimmer-predicted ORFs from Vin13.
Found at: doi:10.1371/journal.pgen.1001287.s008 (8.02 MB TXT)

Dataset S9 Gllimmer-predicted ORFs from VL3.

Found at: doi:10.1371/journal.pgen.1001287.s009 (7.96 MB TXT)

Dataset S10 Gllimmer-predicted ORFs from FostersO.
Found at: doi:10.1371/journal.pgen.1001287.s010 (7.71 MB TXT)

Dataset S11 Gllimmer-predicted ORFs from FostersB.
Found at: doi:10.1371/journal.pgen.1001287.s011 (7.70 MB TXT)

Dataset S12 Novel-predicted ORFs in AWRI796 contig c100.
Found at: doi:10.1371/journal.pgen.1001287.s012 (0.02 MB TXT)

Figure S1 Clustal alignment of the hypothetical, conserved gene adjacent to RTM1 in the ale yeasts AWRI1684 and AWRI1685 and the human pathogen YJM789.
Found at: doi:10.1371/journal.pgen.1001287.s013 (0.03 MB DOC)

Figure S2 Clustal alignment of the five-gene cluster present in wine yeasts.
Found at: doi:10.1371/journal.pgen.1001287.s014 (3.51 MB PDF)

Figure S3 A model for the horizontally-acquired five-gene cluster.
Found at: doi:10.1371/journal.pgen.1001287.s015 (0.04 MB PDF)

Table S1 Tandem repeat variability.
Found at: doi:10.1371/journal.pgen.1001287.s016 (0.10 MB XLS)

Table S2 Multi-strain ORF comparisons.
Found at: doi:10.1371/journal.pgen.1001287.s017 (5.23 MB XLS)

Table S3 Instances of the five-gene cluster.
Found at: doi:10.1371/journal.pgen.1001287.s018 (0.02 MB XLS)

Acknowledgments

The authors would like to thank the commercial yeast suppliers for access to original mother cultures of all strains used in this study and Paul Henschke (AWRI) for critical comments on the manuscript.

Author Contributions

Conceived and designed the experiments: ARB JPA ISP ME PJC. Performed the experiments: BAD DR AHF. Analyzed the data: ARB. Wrote the paper: ARB PJC.

References

1. Querol A, Belloch C, Fernandez-Espinar MT, Barrio E (2003) Molecular evolution in yeast of biotechnological interest. Int Microbiol 6: 201–205.
2. Fay JC, Benavides JA (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet 1: e5. doi:10.1371/journal.pgen.0010005.
3. Mortimer RK, Johnston JR (1986) Genealogy of principal strains of the yeast genetic stock center. Genetics 113: 35–43.
4. Lambrechts MG, Pretorius IS (2000) Yeast and its importance to wine aroma - a review. Sth Afr J Enol Vitic 21: 97–129.
5. Swiegers JH, Pretorius IS (2005) Yeast modulation of wine flavor. Adv Appl Microbiol 57: 131–175.
6. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. Science. New York, NY 274: 546, 563–547.
7. Dunn B, Levine RP, Sherlock G (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. BMC Genomics 6: 53.
8. Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, et al. (2007) Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proc Natl Acad Sci USA 104: 12825–12830.
9. Borneman AR, Forgan AH, Pretorius IS, Chambers PJ (2008) Comparative genome analysis of a *Saccharomyces cerevisiae* wine strain. FEMS Yeast Res 8: 1185–1195.
10. Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, et al. (2008) A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet 4: e1000183. doi:10.1371/journal.pgen.1000183.
11. Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OV, et al. (2009) Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. Genome Res 19: 2258–2270.
12. Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. (2009) Population genomics of domestic and wild yeasts. Nature 458: 337–341.
13. Novo M, Bigey F, Beyne E, Galeote V, Gavory F, et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. Proc Natl Acad Sci USA 106: 16333–16338.
14. Stambuk BU, Dunn B, Alves SL, Jr., Duval EH, Sherlock G (2009) Industrial fuel ethanol yeasts contain adaptive copy number changes in genes involved in vitamin B1 and B6 biosynthesis. Genome Res 19: 2271–2278.
15. Tamai Y, Momma T, Yoshimoto H, Kaneko Y (1998) Co-existence of two types of chromosome in the bottom fermenting yeast, *Saccharomyces pastorianus*. Yeast 14: 923–933.
16. Dunn B, Sherlock G (2008) Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. Genome Res 18: 1610–1623.
17. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
18. Mortimer RK (2000) Evolution and variation of the yeast (*Saccharomyces*) genome. Genome Res 10: 403–409.
19. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. Science 324: 1213–1216.
20. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673–679.
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. NAR 25: 3389–3402.
22. Ness F, Aigle M (1995) *RTM1*: a member of a new family of telomeric repeated genes in yeast. Genetics 140: 945–956.
23. Takagi H, Shichiri M, Takemura M, Mohri M, Nakamori S (2000) *Saccharomyces cerevisiae* sigma 1278b has novel genes of the N-acetyltransferase gene superfamily required for L-proline analogue resistance. J Bacteriol 182: 4249–4256.
24. Goto K, Iwase T, Kichise K, Kitano K, Totuka A, et al. (1990) Isolation and properties of a chromosome-dependent KHR killer toxin in *Saccharomyces cerevisiae*. Agric Biol Chem 54: 505–509.
25. Constam D, Muheim A, Zimmermann W, Fiechter A (1991) Purification and Partial Characterization of an Intracellular NADH - Quinone Oxidoreductase from *Phanerochaete chrysosporium*. J Gen Microbiol 137: 2209–2214.
26. Reiser J, Muheim A, Hardegger M, Frank G, Fiechter A (1994) Aryl-alcohol dehydrogenase from the white-rot fungus *Phanerochaete chrysosporium*. Gene cloning, sequence analysis, expression, and purification of the recombinant enzyme. J Biol Chem 269: 28152–28159.
27. Delneri D, Gardner DC, Bruschi CV, Oliver SG (1999) Disruption of seven hypothetical aryl alcohol dehydrogenase genes from *Saccharomyces cerevisiae* and construction of a multiple knock-out strain. Yeast 15: 1681–1689.
28. Delneri D, Gardner DC, Oliver SG (1999) Analysis of the seven-member *AAD* gene set demonstrates that genetic redundancy in yeast may be more apparent than real. Genetics 153: 1591–1600.
29. Pretorius IS (2000) Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. Yeast 16: 675–729.
30. Lefebure T, Stanhope MJ (2007) Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol 8: R71.
31. Ugliano M, Henschke PA (2009) Yeasts and Wine Flavour. In: Moreno-Arribas MV, Polo MC, eds. Wine Chemistry and Biochemistry. New York: Springer. pp 313–392.
32. Sponholz W (1993) Wine spoilage by microorganisms. In: Fleet G, ed. Wine Microbiology and Biotechnology. London: Taylor and Francis. pp 395–420.
33. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. PLoS Comput Biol 5: e1000392. doi:10.1371/journal.pcbi.1000392.
34. Gouy M, Guindon S, Gascuel O SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27: 221–224.
35. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. NAR 27: 573–580.
36. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. NAR 22: 4673–4680.