

RESEARCH ARTICLE

# Genome Sequencing and Comparative Analysis of *Saccharomyces cerevisiae* Strains of the Peterhof Genetic Collection

Polina B. Drozdova<sup>1,2\*</sup>, Oleg V. Tarasov<sup>1,3</sup>, Andrew G. Matveenko<sup>4,1,5</sup>, Elina A. Radchenko<sup>1,2<sup>‡a</sup></sup>, Julia V. Sopova<sup>4,6,1</sup>, Dmitrii E. Polev<sup>7<sup>‡b</sup></sup>, Sergey G. Inge-Vechtomov<sup>4,1</sup>, Pavel V. Dobrynin<sup>8,2</sup>

**1** Dept. of Genetics and Biotechnology, Saint Petersburg State University, St. Petersburg, Russia, **2** Bioinformatics Institute, St. Petersburg, Russia, **3** St. Petersburg Scientific Center of RAS, St. Petersburg, Russia, **4** St. Petersburg Branch, Vavilov Institute of General Genetics of the Russian Academy of Sciences, St. Petersburg, Russia, **5** Laboratory of Amyloid Biology, Saint Petersburg State University, St. Petersburg, Russia, **6** Institute of Translational Biomedicine, Saint Petersburg State University, St. Petersburg, Russia, **7** Research Resource Center for Molecular and Cell Technologies, Research Park, Saint-Petersburg State University, St. Petersburg, Russia, **8** Theodosius Dobzhansky Center for Genome Bioinformatics, Saint Petersburg State University, St. Petersburg, Russia

<sup>‡a</sup> Current address: Department of Biology, Tufts University, Medford, MA, United States of America

<sup>‡b</sup> Current address: Research Resource Center Bio-Bank, Research Park, Saint Petersburg State University, St. Petersburg, Russia

\* [p.drozdova@spbu.ru](mailto:p.drozdova@spbu.ru); [drozdovapb@gmail.com](mailto:drozdovapb@gmail.com)



OPEN ACCESS

**Citation:** Drozdova PB, Tarasov OV, Matveenko AG, Radchenko EA, Sopova JV, Polev DE, et al. (2016) Genome Sequencing and Comparative Analysis of *Saccharomyces cerevisiae* Strains of the Peterhof Genetic Collection. PLoS ONE 11(5): e0154722. doi:10.1371/journal.pone.0154722

**Editor:** Edward J Louis, University of Leicester, UNITED KINGDOM

**Received:** January 11, 2016

**Accepted:** April 18, 2016

**Published:** May 6, 2016

**Copyright:** © 2016 Drozdova et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Raw sequence data obtained in this paper are available at the NCBI SRA database (PRJNA296913, SRP064279). De novo assemblies are available at the NCBI database (PRJNA296913, LPTZ00000000-LPUD00000000). SNV data, genome assemblies and annotation are available as a custom hub at the UCSC genome browser (<http://genome.ucsc.edu/cgi-bin/hgHubConnect#publicHubs>) and at the GARfield genome browser (<http://garfield.dobzhanskycenter.org/cgi-bin/hgHubConnect>). Custom scripts used to analyze the data are available at <https://github.com/>

## Abstract

The Peterhof genetic collection of *Saccharomyces cerevisiae* strains (PGC) is a large laboratory stock that has accumulated several thousands of strains for over than half a century. It originated independently of other common laboratory stocks from a distillery lineage (race XII). Several PGC strains have been extensively used in certain fields of yeast research but their genomes have not been thoroughly explored yet. Here we employed whole genome sequencing to characterize five selected PGC strains including one of the closest to the progenitor, 15V-P4, and several strains that have been used to study translation termination and prions in yeast (25-25-2V-P3982, 1B-D1606, 74-D694, and 6P-33G-D373). The genetic distance between the PGC progenitor and S288C is comparable to that between two geographically isolated populations. The PGC seems to be closer to two bakery strains than to S288C-related laboratory stocks or European wine strains. In genomes of the PGC strains, we found several loci which are absent from the S288C genome; 15V-P4 harbors a rare combination of the gene cluster characteristic for wine strains and the *RTM1* cluster. We closely examined known and previously uncharacterized gene variants of particular strains and were able to establish the molecular basis for known phenotypes including phenylalanine auxotrophy, clumping behavior and galactose utilization. Finally, we made sequencing data and results of the analysis available for the yeast community. Our data widen the knowledge about genetic variation between *Saccharomyces cerevisiae* strains and can form the basis for planning future work in PGC-related strains and with PGC-derived alleles.

[drozdovapb/code\\_chunks/tree/master/Peterhof\\_strains\\_seq](https://github.com/drozdovapb/code_chunks/tree/master/Peterhof_strains_seq) and <https://github.com/drozdovapb/myBedGtfGffVcfTools>.

**Funding:** PBD acknowledges the Russian Foundation for Basic Research ([www.rfbr.ru](http://www.rfbr.ru)) for grant 14-04-31265. OVT and SGIV acknowledge the Russian Foundation for Basic Research for grant 15-29-02526. JVS acknowledges the Russian Science Foundation ([www.rscf.ru](http://www.rscf.ru)) for grant 14-50-00069 and the Saint-Petersburg State University for grant 1.38.426.2015. PBD, AGM, EAR, and JVS acknowledge the Saint-Petersburg State University for research grant 1.37.291.2015. PBD and OVT acknowledge the Saint-Petersburg City Committee on Science and High School ([knvsh.gov.spb.ru](http://knvsh.gov.spb.ru)) for grants 15404 and 15919, respectively. PBD, AGM, JVS, and SGIV acknowledge the Saint-Petersburg State University for research grant 15.61.2218.2013. PBD acknowledges the Saint-Petersburg State University for research grant 1.42.1394.2015. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Introduction

*Saccharomyces cerevisiae* is a widely used model organism. The S288C strain is the ancestor to many commonly used yeast laboratory strains [1, 2] and provided the first eukaryotic genome to be completely sequenced [3]. S288C and related strains originate from the Carbondale breeding stock of C. Lindgren [4], which resulted from crosses between different strains of *S. cerevisiae* as well as other *Saccharomyces* species [1, 5]. To date, genomes of more than 150 yeast strains of different origins have been sequenced [6–8]. Comparison of such a variety of genomes helps to clarify the natural history of yeast populations and allows to identify genomic elements that are selected under specific conditions. Strains distant from S288C may provide new insights in various fields of yeast genetics as it was demonstrated in studies of genetic control of metabolism and chromosome recombination [9–13]. The Peterhof genetic collection (PGC) contains several strains that became widely used in the field of translation termination ([14–16], and other works) and yeast prion research ([17–19], and other works).

The PGC originates from a Russian industrial distillery lineage (“race XII”) that is thought to be distant from the populations that gave rise to the S288C lineage. In contrast to S288C and most known laboratory strains, the progenitor strain of the PGC is presumed to derive from a single yeast population [20]. The collection is maintained with separate registration of diploids obtained from mating of strains ascending to the progenitor (designated with capital ‘P’ and a consecutive number) and those of hybrid (mosaic) origin (designated with ‘D’ or some other letter in a similar way), and thus enables tracing of ancestors of any strain [21].

A number of genetic variations between Peterhof and S288C-derived strains has been identified, but whole genome data for this big collection of strains are scarce. Thus, we aimed to characterize the genomes of several PGC strains.

## Results and Discussion

### Origin of the strains

In this work, we analyzed genomes of five PGC-related *S. cerevisiae* strains. The PGC came from the initial industrial lineage XII with low ascospore viability (*ca.* 0.7%) through 7 generations of intratetrad self-fertilization. The resulting heterothallic diploid strain with high ascospore viability (*ca.* 90%), XII<sub>7</sub>, should be considered the *bona fide* progenitor of the Peterhof lineage of strains. A haploid prototroph selected after 3 more inbred crosses, 15V-P4, gave rise to the core part of the collection; thus, strains ascending only to 15V-P4 and XII<sub>7</sub> are considered pure Peterhof strains [20, 22]. Apart from 15V-P4, we analyzed genomes of four haploid laboratory strains. 25-25-2V-P3982 (the full name of the strain is 25-25-dU8-132-L28-2V-P3982) [23] is a laboratory strain of a presumably pure Peterhof origin, while 1B-D1606 [15], 74-D694 [24], and 6P-33G-D373 Asp<sup>+</sup> [25, 26] descend from hybrids between Peterhof and S288C-derived strains (further referred to as strains of hybrid origin). The pedigree of strains selected for whole genome sequencing is shown in [S1 Fig](#). For brevity, these strains are henceforth referred to as 25-25, 1B, 74, and 6P-33G; however, as these are not legitimate strain names, we strongly discourage from using these shortened names while mentioning these strains elsewhere.

Genomes of these strains were of particular interest for a number of reasons. 25-25, 1B, and 6P-33G are derived from strains that have been widely used to study termination of translation, and these or closely related strains were the source of all sequenced PGC alleles of the translation termination factor genes as well as auxotrophy markers [15, 16, 27–29]. These sequences are instrumental in validating the quality of the next generation sequencing results. 74 and its derivatives have been exploited to study the [PSI<sup>+</sup>] prion [24], and genomic data for a [PSI<sup>+</sup>]

**Table 1. De novo assembly statistics.**

Strain	15V-P4	25-25	1B	74	6P-33G
Number of contigs*	1,165	891	480	1,514	3,039
Largest contig, bases*	86,918	122,204	252,839	71,636	22,054
Total length, bases*	11,666,974	11,614,012	11,567,449	11,330,585	10,013,534
N50*	19,288	26,385	72,884	11,948	4,341
S288C genome fraction, %*	91.87	92.78	94.17	92.53	76.01
Median S288C genome coverage	19x	22x	35x	42x	16x
Number of S288C genes found (%) **	5,115 (87%)	5,000 (85%)	4,993 (84%)	5,058 (86%)	4,004 (68%)
Complete CEGMA core genes found (%)	241 (97%)	240 (97%)	243 (98%)	243 (98%)	210 (85%)
Partial CEGMA core genes found (%)	244 (98%)	241 (97%)	243 (98%)	243 (98%)	226 (91%)

\* Calculated with Quast.

\*\* Calculated as one-to-one orthologs with ProteinOrtho clustering.

doi:10.1371/journal.pone.0154722.t001

variant of this strain was published earlier [30]. 6P-33G was particularly interesting because it has been recently shown to be disomic for chromosome VIII [26] and could therefore serve as a control in copy number variation analysis. 1B and 6P-33G are closely related (S1 Fig), which might provide material to study recombination patterns. Finally, these strains had a number of phenotypes lacking known molecular basis (see below).

### Genome assembly and gene annotation

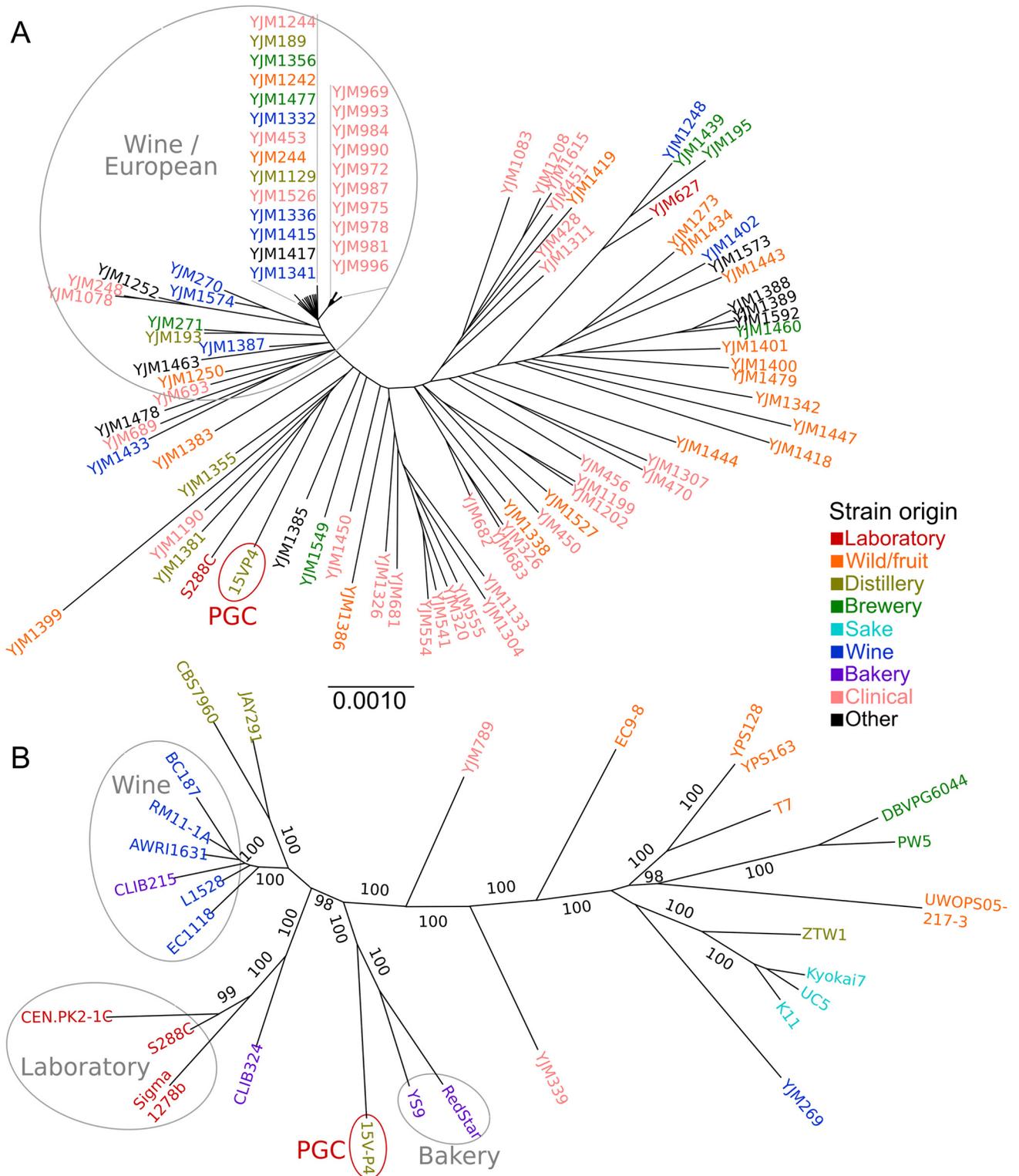
Raw reads produced with either Ion Torrent PGM (15V-P4, 25-25, 1B, and 6P-33G) or with Illumina GAII (74) were assembled *de novo* (Table 1). The resulting assemblies were characterized by varying quality as assessed by Quast [31], with the best results for the 1B genome and the worst for the 6P-33G genome. Quality of assemblies produced with Ion Torrent data increased according to the coverage, while Illumina data for the 74 genome produced a lower quality assembly despite higher coverage, which was probably due to shorter reads. Contigs were scaffolded to produce pseudochromosomes using the S288C genome as a reference.

The assemblies were annotated with several *de novo* and alignment-based gene finders, and then united annotations were obtained with the MAKER2 pipeline [32]. Genes found with this pipeline were mapped to the known S288C genes with ProteinOrtho [33] clustering. For all the genomes excluding 6P-33G we were able to find at least about 85% of reference genome genes. We also assessed quality of assemblies with CEGMA [34] and were able to find almost 100% of common eukaryotic core genes in all assemblies except for 6P-33G and over 90% in all genomes. These results signify that the assemblies can be used in downstream analyses.

### Origin of the PGC in a phylogenetic context

Since the PGC was established independently from the Carbondale breeding stock [4, 20], we were interested in determining the phylogenetic relationships of the PGC progenitor and other *S. cerevisiae* strains. To assess this, we applied two complementary approaches.

First, we used the largest *S. cerevisiae* tree presently available, which is based on a set of highly conserved regions from all nuclear chromosomes [8]. We extracted the corresponding sequences from the 15V-P4 genome assembly and constructed an alignment with 217,304 positions from 95 genomes to re-infer the phylogenetic tree (Fig 1A). The overall tree topology is similar to that reported originally [8]. According to this tree, the closest strains to 15V-P4 are YJM1190, YJM1381, YJM1399, S288C and YJM1355. Even though these strains are of different



**Fig 1. 15V-P4 position in the phylogeny of *S. cerevisiae* strains.** (A) Neighbor joining phylogenetic tree of 95 strains including 15V-P4 inferred from alignment of conservative chromosome regions. (B) Phylogenetic tree of 29 strains including 15V-P4 inferred from sequences of 807 common genes under the GTR+G model and tested with 500 bootstrap replicates. Branch bootstrap values greater than 95 are indicated. In both trees, strain names are colored according to functional origin. Grey circles highlight either the population group (A) or common functional origin (B). Branch lengths are given in the same scale on both trees. PGC, the Peterhof genetic collection.

doi:10.1371/journal.pone.0154722.g001

geographic origins, two of them, YJM1355 and YJM1381, are of distillery origin [8], similar to 15V-P4. In addition, we merged SNV data for 15V-P4 with the available dataset [8] and assessed the population ancestry of the PGC with STRUCTURE [35, 36]. 15V-P4 appeared to possess an admixed genome with most probable ancestry to wine and human-associated populations (S2A Fig), as well as strains close to it on the tree.

As these data did not elucidate the PGC origin, we turned to an alternative approach. We sampled coding sequences of 15V-P4 and 28 strains of different origin from the *Saccharomyces* Genome Database (SGD). Only genomes with more than 3500 genes annotated were selected for the analysis, and 807 genes were found in all of them. Total alignment based on sequences of these genes (S1 Table) included 852,372 nucleotide positions. The resulting phylogenetic tree inferred from common ORFs (Fig 1B) is generally similar in topology to those inferred previously from total genomic SNVs or non-reference ORFs [7, 37]. The tree shows a major clade including three groups: the first one uniting common laboratory strains (e.g. S288C), the second comprising commercial wine and bioethanol strains, and the third consisting of 15V-P4 and two bakery strains, YS9 and RedStar (Fig 1B). We added the latter two strains to the SNV-based 95-genomes tree and confirmed this result, as 15V-P4 was closer to YS9 and RedStar than to S288C (S2B Fig). Thus, the distillery lineage ancestral to the PGC might have itself originated from a bakery strain.

### Non-reference genes in PGC genomes

Newly sequenced *S. cerevisiae* strains are frequently found to contain genes absent from the genome of the reference strain (see [7]). To determine whether the Peterhof strains possess such genes, we divided all the annotated genes found into known (i.e. those found in the reference genome) and novel (non-reference) ones. The list of novel genes was used as a BLAST query, and the BLAST output was manually curated; presence of genes from other strains or species was re-confirmed with Exonerate protein2genome search for the best BLAST hit against genome assemblies. We found a total of 11 non-reference genes in the 15V-P4 genome (Table 2); some of these genes were inherited by the other strains.

All five strains studied possess the *KHR1* gene, which encodes a killer toxin of unknown nature [38]. In 25-25 and 1B, this gene was annotated on the same contigs as known chromosome IX genes, which corroborates findings of Wei *et al.* [39], who localized this gene on chromosome IX in the YJM789 strain.

**Table 2. Genes absent from the S288C genome but found in 15V-P4.**

Gene name	Protein function	S288C	15V-P4
<i>AMI1-A</i>	Amidase	No	Yes
<i>KHR1</i>	Killer toxin	No	Yes
<i>RTM1</i>	Lipid exporter	No	Yes
<i>SCY_1426</i>	Hypothetical zinc finger transcription factor	No	Yes
<i>SUCX*</i>	Invertase	<i>SUC2</i> only	≥ 2 <i>SUCX</i> genes
Wine12	5-oxo-L-prolinase	No	Pseudogene?
Wine23	Nicotinic acid permease	No	Yes
Wine34	Flocculin	No	Yes
Wine45	Transcription factor	No	Yes
Wine56	Transcription factor	No	Yes

Yes/No denotes presence/absence of the corresponding gene.

\* Members of the *SUC* family, see S2 Table.

doi:10.1371/journal.pone.0154722.t002

All the strains analyzed except for 74 possess the *RTM1* gene annotated on its own contig. It was identified in a BLAST search as *S. bayanus*, *S. carlsbergensis*, or *S. pastorianus rtm1* and then re-confirmed with Exonerate search for the Rtm1 protein sequence from YJM789 (Genbank accession EDN59063.1). *RTM1* encodes a lipid-translocating exporter and is known to be advantageous for strains growing on molasses [40, 41]. The *RTM1* gene is a member of a subtelomeric three-gene locus found in several clinical, industrial, and environmental isolates [40]. In the strains harboring *RTM1*, the same contig contains the second member of this cluster encoding a *ca.* 750 amino acid long hypothetical zinc finger transcription factor.

The *RTM1* cluster is usually found in association with genes of the *SUC* (sucrose utilization) family [40]. The *SUC* genes of *S. cerevisiae* fall into two categories, either *SUC2* (YIL162W) or others found in subtelomeric regions [42]. The S288C strain possesses the *SUC2* gene but not the subtelomeric *SUC* genes. Analysis of the reads aligned to the region of chromosome IX corresponding to the *SUC2* (YIL162W) ORF revealed at least two different *SUC* genes (S2 Table), although we were unable to determine their exact number. This finding agrees with recently reported presence of *SUC2*, *SUC5*, and *SUC8* in the XII<sub>7</sub> strain (parental to 15V-P4) revealed with DNA-DNA hybridization of PFGE-resolved chromosomes with a *SUC2* probe [43].

In 15V-P4, but not in the rest of the strains, we also found the so-called 'wine cluster' consisting of five genes (Wine12–Wine56, see Table 2) initially identified in wine strains [44]. Sequence analysis suggests that the 5-oxo-L-prolinase gene (Wine12) is a pseudogene as it contains two frameshifts while the other four genes may be active. Interestingly, 15V-P4 appears to be the first non-wine yeast strain reported to obtain simultaneously the *RTM1* cluster and the wine-specific cluster; genomes sequenced so far rarely contain both clusters [7, 45]. Wine cluster is supposed to move within yeast genomes easily, therefore it could be lost quickly during laboratory breeding [44].

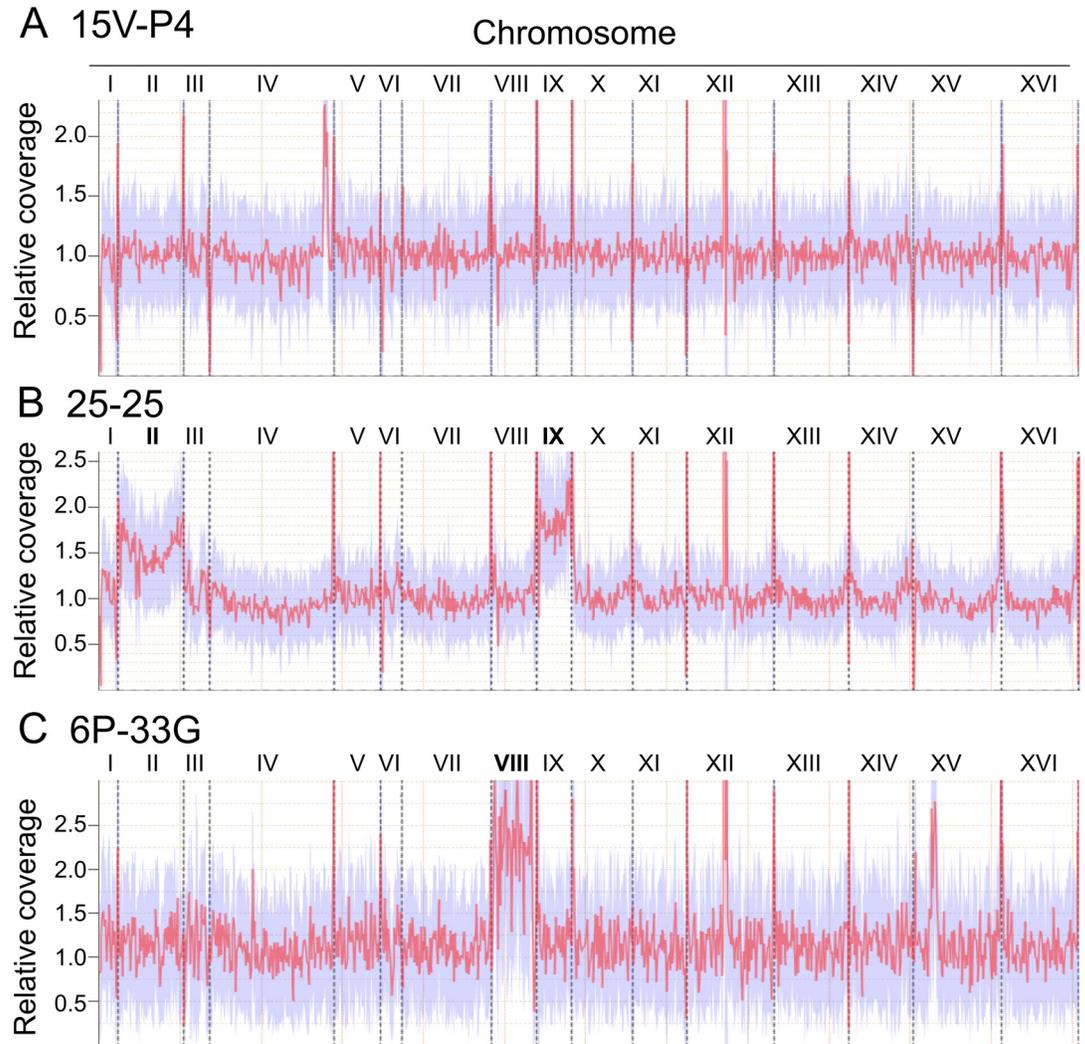
In addition, we found a *Saccharomyces pastorianus* amidase gene *AMII-A* (Uniprot A9CMR9) on its own contig in 15V-P4 but not in any other PGC strain. We also detected this gene in several other *S. cerevisiae* genomes (e.g., RedStar and Kyokai7).

Thus, we showed that the PGC progenitor possesses a unique combination of non-reference genes; however, other PGC strains lost many of them which is presumably a common effect of a laboratory breeding.

In addition to non-reference genes, we looked for regions that could have been introgressed into the 15V-P4 genomes from closely related *Saccharomyces* species. We employed two alternative methods, search for ORFs that are more similar to one of the available *Saccharomyces sensu stricto* genomes than to S288C and alignment of 15V-P4 short reads to concatenated *S. sensu stricto* genomes. In the first analysis, we did not find any regions covering the whole gene and being more similar to a non-*cerevisiae* genome (S3 Table). In the second analysis, the overall alignment of 15V-P4 was very similar to S288C and dissimilar to YJM248 [8], a positive control for introgression (S3 Fig). Thus, we could not reliably identify any introgressed regions, and this result argues against possible interspecific hybridization in the original distillery lineage.

## Copy number variations in PGC genomes

Genome content variations such as chromosomal rearrangements and aneuploidy were found in different *S. cerevisiae* strains [8, 46, 47]. We exploited reference genome coverage to estimate relative sequencing depth of each chromosome. It was mostly uniform for 15V-P4, 1B and 74 (Fig 2A, S4 Fig) but not for 25-25 and 6P-33G. Interestingly, a region on the right arm of chromosome IV seemed to be duplicated in 15V-P4, as well as a region of the left arm of chromosome XV in 6P-33G. In the 25-25 genome, chromosomes II and IX had higher coverage than



**Fig 2. Genome coverage across reference for representative strains.** (A) 15V-P4, (B) 25-25, (C) 6P-33G. Dashed lines signify chromosome borders.

doi:10.1371/journal.pone.0154722.g002

the others (Fig 2B), which suggests that the population of this strain includes a significant number of aneuploid cells. In case of 6P-33G, chromosome VIII coverage was about 2-fold higher compared to the other chromosomes (Fig 2C). This finding perfectly agrees with the earlier reported data on chromosome VIII disomy in this strain [26].

Next, we used the mrCaNaVaR pipeline to analyze possible segmental duplications or deletions more precisely (Table 3). For full list of the regions annotated as amplified or deleted, see

**Table 3. Lengths of regions annotated as amplified or deleted in each strain and counts of genes included into each of these regions.**

Strain	15V-P4	25-25	1B	74	6P-33G
Total length of amplified regions, bp	458,407	856,981	474,369	613,763	1,054,510
Total number of amplified genes	179	392	141	159	499
Total length of deleted regions, bp	84,475	53,117	39,722	24,334	36,913
Total number of deleted genes	20	9	5	6	24

doi:10.1371/journal.pone.0154722.t003

[S4 Table](#). In accordance to the reference genome coverage data, much more regions were annotated as amplified in 25-25 and 6P-33G genomes. This tendency becomes even clearer if the numbers of genes included into the amplified regions are compared (about 150 in euploid strains and about 400 to 500 in strains with a tendency to aneuploidy; see [S5 Table](#)). 101 amplified genes were common in the amplified regions of all five genomes but almost all of them have close paralogs and may thus represent false positive findings. We conclude that the results of analysis of amplified regions are very noisy and should be interpreted with caution. There are at least two possible reasons, the great number of recently amplified genes in yeast due to the whole genome duplication in the lineage leading to *S. cerevisiae* [48, 49] and presence of aneuploid strains in our analysis.

Analysis of deleted regions should not be prone to such noise. Importantly, we were able to confirm all known whole-ORF deletions, *i.e.* *URA3* deletion in 25-25, *HIS3* deletion in 74 and *SUP35* deletion in 6P-33G (see [S5 Table](#), and [S1 Appendix](#)). In addition, we looked for other deleted genes. Two genes, *FLO10* and *NFT1*, were presumably deleted in all the strains. These genes are adjacent on the right arm of chromosome XI, and their absence might indeed represent a common feature of PGC-related strains.

## Single nucleotide variations

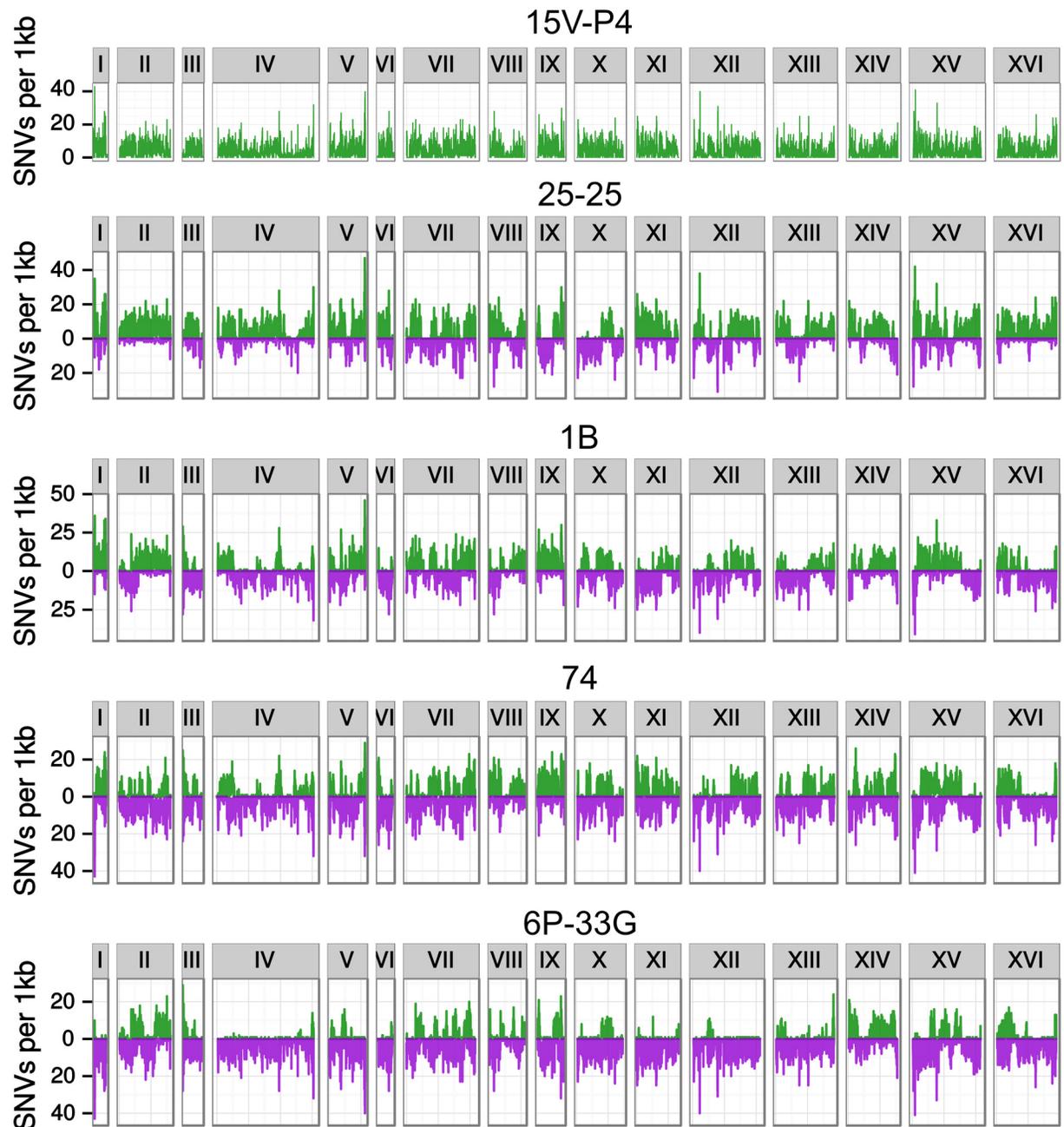
In order to assess the difference between Peterhof strains and the reference strain S288C, we aligned short reads to the S288C genome. Typically, about 95% of reads were aligned. Then, we called single nucleotide variations (SNVs), and filtered out low quality differences and differences in repeat regions.

First, we analyzed the distribution of substitutions in the ancestor strain of the PGC. The distribution of polymorphic sites in 1 kb windows across the S288C chromosomes seemed quite uniform ([Fig 3](#), upper panel). Functional classification of substitutions performed with SNPeff enabled us to find 97 nonsense, 10675 missense, and 18534 silent mutations, as well as 16020 intergenic variations. It directly translates to dN/dS = 0.58 hinting at presence of selection pressure.

Then, we estimated the number of short indels compared to reference in each of the genomes analyzed. The Ion Torrent technology is prone to errors in homopolymer regions [50, 51]. However, these errors are random and should not reproduce in all the reads aligned to a particular position. Thus, we filtered only indels supported by all reads aligning to this position (100% supported indels) as they are less likely to represent sequencing errors. Unlike the total number of indels, the number of 100% supported indels was roughly proportional to the number of SNVs ([S6 Table](#)), which consolidates our approach.

Many strains of the Peterhof genetic collection are known to be of hybrid origin, *i.e.* to originate from at least two yeast genetic lines, Peterhof and Carbondale breeding stocks. Using 15V-P4 as the reference Peterhof strain and S288C as a common reference strain, we called all variations between each strain and two reference strains. The results of this analysis are presented at [Fig 3](#). As expected, strains ascending to 'D' diploids showed long tracts of either non-Peterhof or non-S288C substitutions, as we would expect for a mosaic genome. Surprisingly, the same kind of analysis for the 25-25 strain indicates that it has hybrid origin even though it was previously described as a pure Peterhof strain [23].

We estimated genetic difference between PGC strains and S288C as the number of pairwise SNVs using the genome of S288C as a common reference and neighbor joining algorithm ([S5 Fig](#)). 15V-P4 and S288C differ by 45,842 SNVs which is comparable to the level of divergence of about 50,000 SNVs between distant *S. cerevisiae* populations reported previously [6]. As expected, the 25-25 strain is the most similar to 15V-P4. However, these two strains have much more



**Fig 3. Distribution of variable sites shown in chromosomal coordinates of S288C.** Green: SNVs compared to S288C. Purple: SNVs compared to 15V-P4. Each chromosome is framed.

doi:10.1371/journal.pone.0154722.g003

pairwise SNVs than we expected, which supports the idea that this strain should have had a non-Peterhof ancestor. 1B and 74 are roughly half as distant from S288C as from 15V-P4; this result is consistent with their known hybrid origin. 6P-33G appears to be closer to S288C than to 15V-P4.

### Selected SNVs and associated phenotypes

Since a number of genetic and phenotypic differences between particular Peterhof strains and S288C-derived strains had been identified previously ([15, 27] and other works), we employed

**Table 4. Selectable marker mutations in the PGC strains.**

Allele	Known change	Variations detected in this study*	Strains
<i>ade1-14</i>	G732A (→TGA) [52–54], G554A [54]	G732A (→TGA)	25-25, 74, 1B
<i>ade2-144,717 (ade2-144,791)**</i>	Nonsense ( <i>ade2-144</i> ) [55] + downstream non-suppressible ( <i>ade2-717 = 791</i> ) [56] mutation	T422A + G423A (→TAA), C1517T (→P506L)	6P-33G
<i>his3Δ200</i>	Deletion (-205 to +172) [57]	Deletion (-205 to +172)	74
<i>his7-1</i>	A229T (→TAA) [58, 59]	A229T (→TAA)	25-25, 6P-33G, 1B
<i>leu2-B2 (leu2-1, leu2-1A, leu2-01)**</i>	Unidentified non-suppressible mutation	G748A (→D250N)	25-25
<i>leu2-3,112</i>	249insG, 792insG [60]	249insG, 792insG	74, 1B
<i>lys2-87 (L28)**</i>	Nonsense mutation, TGA [52]	G3465A (→TGA)	25-25
<i>lys9-A21</i>	T605A (→TAA) [58, 59]	T605A (→TAA)	6P-33G, 1B
<i>pheA10 (pha2P-A10)***</i>	Nonsense mutation [22], TAA [61]	A481T (→TAA)	6P-33G
<i>thr4-B15</i>	Nonsense mutation, TGA [52]	A1180T (→TGA)	25-25
<i>trp1-289</i>	C403T (→TAG) [59, 62]	C403T (→TAG)	6P-33G, 74, 1B
<i>ura3Δ(dU8)</i>	Complete deletion of <i>URA3</i> (unpublished data)	Deletion (-188 to +76)	25-25
<i>ura3-52</i>	Ty1 insertion (transcribing left to right) at 121 [63]	Ty insertion at 121****	6P-33G, 74, 1B

Nucleotide positions in 5' UTR are preceded with the minus sign while those in 3' UTR with the plus sign; numbers indicate distance from the beginning or the end of the ORF, respectively. Stop codon type or amino acid substitution are indicated after an arrow for mutations that must lead to known auxotrophic phenotypes.

\* Only differences from the corresponding wild type alleles are listed. For complete list of substitutions, see [S1 Appendix](#).

\*\* Synonymous designations.

\*\*\* Assigned to the *PHA2* locus in this work.

\*\*\*\* Includes duplication of insertion flanking sequence (GTACC).

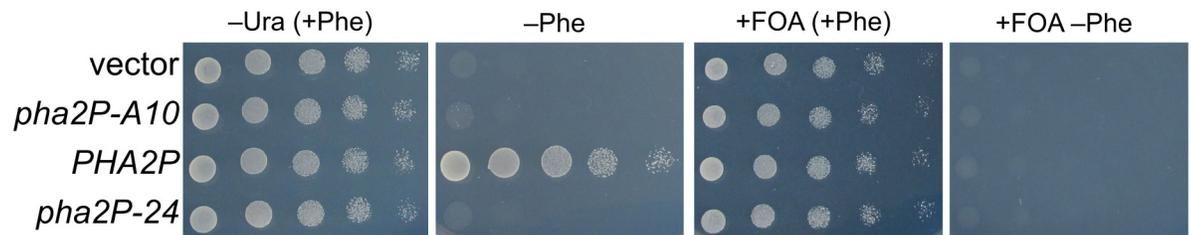
doi:10.1371/journal.pone.0154722.t004

these data in our analysis by looking for already known variations. This search served two purposes. First, we used it to validate our pipeline. Second, as variations in strains close to the PGC progenitor (e.g., 15V-P4) have never been analyzed, this approach enabled us to assess whether known differences converge to the common ancestor of the Peterhof genetic collection or were attained during the laboratory breeding of the strains.

We searched the Peterhof strains for the known genetic variations in several selectable marker genes. The whole genome sequencing results conform to the previous data and complete the missing information about precise mapping of some mutations ([Table 4](#); [S1 Appendix](#)).

Some PGC strains have been extensively used to obtain large collections of strains with suppressor mutations in release factor genes *SUP35* (*SUP2*) and *SUP45* (*SUP1*) [15, 16, 28, 64]. Their sequences were previously identified in dU8-132-L28-2V-P3982 and 1B, respectively [15, 27]. We detected all the mutations we were aware of ([S1 Appendix](#)). In 15V-P4, we found all the SNVs identified previously in wild type Peterhof *SUP35* and *SUP45* alleles. Thus, we proved that these alleles had been inherited from the common ancestor of the PGC.

**PHA2.** 6P-33G, as well as its direct ancestor, 33G-D373, is known to bear a phenylalanine auxotrophy mutation *pheA10* [65]. This mutation has been shown to be a TAA nonsense, as it was suppressible by ochre suppressors ([61] and unpublished data) but has never been mapped to a particular gene ([S1 Appendix](#)). So, we looked for mutations in phenylalanine biosynthesis



**Fig 4. Only *PHA2P* but not *pha2* mutant alleles compensate for the *phaA10* phenylalanine auxotrophy.** 33G-D373 was transformed with plasmids bearing indicated *PHA2* alleles. Series of 5-fold dilutions on synthetic media are shown. Vector, pRS316.

doi:10.1371/journal.pone.0154722.g004

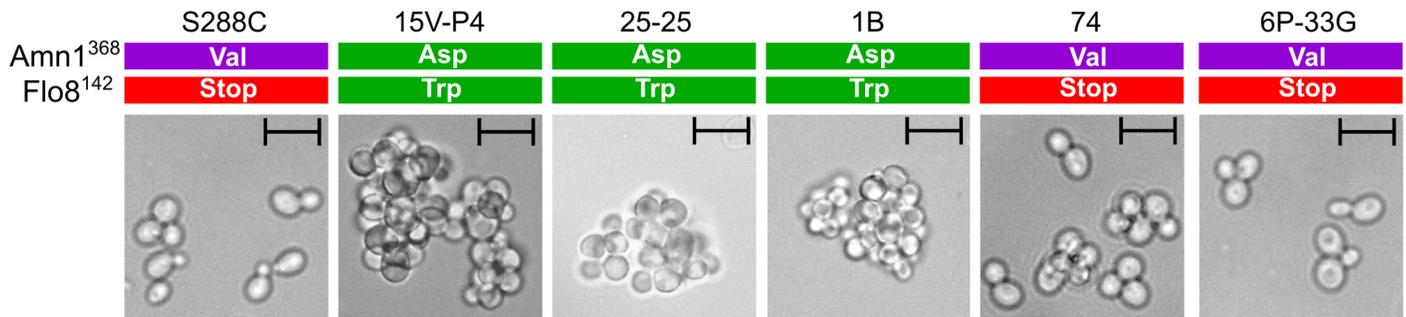
genes and found a premature termination codon (PTC) in the *PHA2* gene (see [S6 Fig](#) and [S1 Appendix](#) for details).

To find whether this nonsense mutation in *PHA2* is responsible for the phenylalanine auxotrophy we cloned either the wild type PGC allele *PHA2P* or the mutant allele (designated as *pha2P-A10*) into a centromeric *URA3* vector. Introduction of *PHA2P*, but not *pha2P-A10*-containing plasmid into 33G-D373 restored growth on media lacking phenylalanine ([S6C Fig](#)). Furthermore, loss of the plasmid-borne *PHA2P* allele on 5-FOA-containing medium led to immediate loss of phenylalanine prototrophy ([Fig 4](#)). We also obtained a *pha2* double missense mutation (*pha2P-24*) which was unable to compensate for phenylalanine auxotrophy in 33G-D373 ([S6C Fig](#), [Fig 4](#)). Thus, not only *pha2P-A10*, but other defects in *PHA2P* may lead to a phenylalanine auxotrophy, which is consistent with previous findings [66] and supports *pha2* usefulness as a selectable marker. We also showed that level of *pha2P-A10* suppression is higher in *Asp<sup>-</sup>* than in *Asp<sup>+</sup>* derivative of 6P-33G ([S6D Fig](#)), consistent with comparative levels of suppression of other nonsense mutations in the two derivatives [26]. Thus, this allele might also be employed to study nonsense suppression in yeast.

At the next step, we looked for novel nonsense mutations as their effect is the easiest to predict. We found a total of 16 to 78 genes with PTCs in the Peterhof strains (5 of these genes were common for all 5 strains) and 2 genes, *FLO8* and *CRS5*, in which stop codons present in S288C were absent from PGC strains ([S7 Table](#)). Among those, we further investigated absence of a PTC in *FLO8* and presence of a PTC in *MSN4*.

**Clumping.** Cells of Peterhof-derived strains tend to clump together in liquid medium (unpublished data). Cell aggregation is a very complex trait but some genes contributing to its control are known. Flo8 is a transcription factor contributing much in flocculation, diploid filamentous growth, and haploid invasive growth in yeast [67–70]. Several PGC-related diploid strains were shown to form pseudohyphae on solid medium and to contain the *FLO8* allele encoding the full length protein [71, 72]. Amn1 is another transcriptional regulator with a clear link between the sequence variant and the cell aggregation phenotype [70]. S288C and closely related strains with Amn1<sup>368Val</sup> and Flo8<sup>142Stop</sup> do not form clumps, while variants Amn1<sup>368Asp</sup> and Flo8<sup>142Trp</sup> contribute much to the change from non-clumping to clumping phenotype [68, 70]. We observed the same tendency in PGC strains: those with Amn1<sup>Asp368</sup> and Flo8<sup>Trp142</sup> showed clear clumping phenotype while those with known loss-of-function variants were much less prone to form cell aggregates ([Fig 5](#)). Unfortunately, we could not assess the effect of the two variable positions separately.

We also addressed the suppressibility of the *flo8* stop codon by two modifiers of translation termination, the [*PSI<sup>+</sup>*] prion and the *Asp<sup>+</sup>* determinant. We found no difference in clumping efficiency between isogenic strains with different suppressor phenotypes (data not shown). This is consistent with previous data showing almost complete absence of the *flo8* stop codon bypass [73].



**Fig 5. Cell aggregation phenotypes of strains analyzed correlate with *AMN1* and *FLO8* alleles.** The scale bar indicates 10  $\mu$ m. Amn1 and Flo8 variants are shown in color (green: associated with “clumping” phenotypes; red and purple: “non-clumping”). Representative microphotographs out of five fields of view of yeast liquid medium cultures in early stationary phase are shown.

doi:10.1371/journal.pone.0154722.g005

**MSN4.** The *MSN4* gene encodes a transcription factor with many targets including heat shock proteins. A PTC in *MSN4* in the 74 genome was first attested in Fitzpatrick *et al.* [30]. As we used the same data, our analysis produced the same result. In addition, we found the same mutation in 25-25 and 15V-P4. Thus, the other two strains probably inherited this substitution from 15V-P4.

To test whether this mutation has any associated phenotype, we cloned *MSN4* into a centromeric vector pRS316 and transformed 74 with this construct. As slight difference in thermotolerance was shown for [*PSI*<sup>+</sup>] and [*psi*<sup>-</sup>] derivatives of 74 [74], we exploited both strains. We could not see any change in thermotolerance upon plasmid addition (S7 Fig).

**Multiple substitutions in the *GAL* locus.** Several PGC strains, including 1B, are Gal<sup>-</sup>, *i.e.* they manifest no growth on media containing galactose (even with raffinose) as a sole carbon source (unpublished data). We found that in 1B and 6P-33G lengthy regions of chromosome II, which include the *GAL* locus, are enriched in different sets of SNVs which are neither 15V-P4- nor S288C-derived. We suppose that these regions may have been inherited from some ancestors other than S288C or 15V-P4; therefore, the comparison of the *GAL* locus sequences may provide additional information on genealogy of PGC strains.

The *GAL* locus encodes three enzymes of galactose metabolism (Gal7, Gal10, and Gal1). To determine possible origin of this locus in PGC strains and mutation(s) causing the galactose utilization defect in 1B, we compared the *GAL* locus sequences (chromosome II from 274,427 to 280,607) of 5 PGC genomes and 38 strains from SGD. The *GAL* locus of 6P-33G seemed to be identical to that of JK9-3d, SEY6210, and YPH499, the ancestor of the other two strains (S8A Fig) [75, 76]. YPH499 originates from a strain congenic to S288C [77] but is known to have some non-S288C SNVs [78]. In 1B, the *GAL* locus was almost identical to that of D273-10B (S8A Fig) and FL100 strains which have common origin from F. Sherman’s lab [79, 80]. Together, these data imply possible lineages of laboratory yeast strains that might have left their footprints in the history of PGC.

The only SNV unique for 1B is a missense mutation *GAL10*<sup>C287T</sup> (Gal10<sup>Ala96Val</sup>) (S8A Fig). D273-10B and FL100 are known to be Gal<sup>+</sup> [81, 82]; therefore, this substitution may be responsible for the Gal<sup>-</sup> phenotype of 1B. To test this assumption, we transformed 1B with plasmids containing the complete *GAL* locus of S288C or its fragments and found that only the plasmids containing *GAL10* reverted 1B to Gal<sup>+</sup> (S8B Fig). Gal10 has an UDP-galactose-4-epimerase (GALE) activity [83]. The residue 96<sup>Ala</sup> in *S. cerevisiae* Gal10 (93<sup>Ala</sup> in human GALE) is located in highly conservative NAD and UDP-hexose binding pocket [84, 85]. In human GALE, substitution of the adjacent 94<sup>Val</sup> with Met (S8C Fig) leads to severe

galactosemia [86, 87]. Thus, we presume that Gal10<sup>A96V</sup> is associated with inability to utilize galactose as a carbon source.

## Conclusions

The Peterhof genetic collection (PGC) of yeasts is an almost unique example of a laboratory stock developed independently of the Carbondale breeding stock (S288C-related strains) and including several thousands various strains which can be used in different types of experiments. We have characterized genomes of five PGC strains and made the data available for the yeast community. It allowed us to investigate the phylogenetic relationship of PGC strains with other *S. cerevisiae* strains. Interestingly, phylogenetic analysis places the progenitor strain, 15V-P4, together with two bakery strains even though it originates from a distillery lineage.

SNV analysis showed that the genetic difference between the progenitor strain of PGC and S288C is approximately the same as the difference between distant yeast populations reported earlier [6]. Importantly, the genetic distances between the strains generally are in good agreement with their pedigree. However, new data imply that one presumably pure Peterhof strain, 25-25, is of hybrid origin.

Strains of PGC possess several loci absent from S288C. None of these loci are unique for PGC strains but their combination such as in 15V-P4 has not been reported yet. To the extent of our knowledge, it is the first sequenced non-wine strain with *RTM1* and wine clusters at the same time.

We were able to find out the exact sequence differences corresponding to most previously known phenotypes. Particularly, we mapped the *pheA10* nonsense mutation to the *PHA2* gene and identified a missense mutation in *GAL10* as the reason behind galactose utilization defect in 1B. We also found and validated some genetic variations providing insight into physiological differences between PGC and S288C-derived strains. We saw very good agreement between allelic states of *FLO8* and *AMN1* with cell clumping pattern. Dissimilar to commonly used S288C-based laboratory strains, Peterhof strains can be used to study aggregation phenotypes and pseudohyphal growth [71, 72], and our data further support this usage.

Together, our data widen the knowledge about genetic variation between *Saccharomyces cerevisiae* strains, link some previously known phenotypes to newly identified sequence differences and form the basis for planning future work in PGC-related strains and with PGC-derived alleles.

## Materials and Methods

### Yeast strains

Yeast strains used in this work are listed in the Table 5 and are available upon request at the Department of Genetics and Biotechnology of the Saint Petersburg State University.

### DNA extraction and genome sequencing

Raw reads for the genome of 74-D694 [*PSI*<sup>+</sup>] variant originating from Yury Chernoff lab were produced with Illumina GAI [30] and downloaded from [http://bioinf.nuim.ie/wp-content/uploads/2011/10/74D\\_sequence.txt.zip](http://bioinf.nuim.ie/wp-content/uploads/2011/10/74D_sequence.txt.zip). Single end libraries for genomes of 15V-P4, 25-25-2V-P3982, 1B-D1606, and 6P-33G-D373 Asp<sup>+</sup> strains were sequenced with Ion Torrent PGM™ machine. Raw reads are available at the NCBI Sequence Read Archive [PRJNA296913, SRP064279].

DNA extraction was performed with mechanical disruption of yeast cells as described in [89]. YPD was supplemented with 100 to 250 mg/L adenine in case of *ade1* and *ade2* mutant strains.

**Table 5. Yeast strains used in this work.**

Name	Known genotype	References
15V-P4	<i>MAT a prototroph</i>	[20]
25-25-2V-P3982* (25-25)	<i>MAT a ade1-14 his7-1 lys2-87 ura3Δleu2-B2 thr4-B15 sup35-25 sup45-400</i>	[23]
6P-33G-D373 Asp <sup>+</sup> (6P-33G)	<i>MATα pha2P-A10 (pheA10) ade2-144,717 his7-1 lys9-A21 trp1-289 ura3-52 leu2-3,112sup35Δ:: PmSUP35-LEU2** CUP1//CUP1 (chromosome VIII disomy)</i>	[25, 26]
6P-33G-D373 Asp <sup>-</sup>	<i>MATα pha2P-A10 (pheA10) ade2-144,717 his7-1 lys9-A21 trp1-289 ura3-52 leu2-3,112 sup35Δ:: PmSUP35-LEU2**</i>	[25, 26]
33G-D373	<i>MATα pha2P-A10 (pheA10) ade2-144,717 his7-1 lys9-A21 trp1-289 ura3-52 leu2-3,112</i>	[65]
74-D694	<i>MAT a ade1-14 trp1-289 his3Δ200 ura3-52 leu2-3,112 [psi<sup>-</sup>] [PIN<sup>+</sup>]</i>	[24]
P-74-D694***	<i>MAT a ade1-14 trp1-289 his3Δ200 ura3-52 leu2-3,112 [PSI<sup>+</sup>] [PIN<sup>+</sup>]</i>	K. Volkov, unpublished; [61, 89]
1B-D1606	<i>MATα ade1-14 his7-1 lys9-A21 trp1-289 ura3-52 leu2-3,112 gal10-1B****</i>	[15]
S1 (isogenic to S288C)	<i>MATα gal2 flo8-1 hap1</i>	[1, 8]

\* The full name of the strain is 25-25-dU8-132-L28-2V-P3982.

\*\* *PmSUP35*, *Pichia methanolica SUP35*.

\*\*\* A [PSI<sup>+</sup>] derivative of 74-D694. In addition, other [PSI<sup>+</sup>] derivatives of 74-D694 (OT55 and OT56) [61, 88] and their [psi<sup>-</sup>] isolates were used for the clumping experiments.

\*\*\*\* *gal10-1B* designates the *gal10<sup>C287T</sup>* allele from 1B-D1606 discovered in this work.

doi:10.1371/journal.pone.0154722.t005

Genomic DNA library was prepared using Ion Plus Fragment Library Kit, according to the manufacturers recommendations (User Guide Publication Number 4471989, Revision N). Template-positive particles for genomic DNA sequencing were prepared using Ion PGM™ Template OT2 400 Kit according to the user guide (Publication number MAN0007218, revision 3.0). Sequencing was conducted using Ion PGM™ Sequencing 400 Kit and Ion 318™ Chip v2, following the manufacturer’s user guide (Publication Number MAN0007242, Revision 2.0). Sanger sequencing was performed with ABI Prism 3500xl.

All sequencing reactions were performed at the Research Resource Center for Molecular and Cell Technologies of the Saint Petersburg State University.

## Data analysis

Quality control of reads was performed with FastQC [90]. Trimming of reads was performed with fastx\_toolkit v0.0.13.1 [91] and cutadapt [92]. Trimming length was chosen according to the basic statistics calculated with FastQC [90].

*De novo* genome assembly was performed with SPAdes [93] v3.1.0 with IonHammer (option—iontorrent for homopolymer correction) for Ion Torrent data and with SPAdes v3.6.0 for Illumina data. Reference-assisted assembly of scaffolds was performed with chromosomer [94]. The S288C genome (Release R64-1-1, downloaded from the *Saccharomyces* Genome Database [95]) was used as a reference throughout this work. Quality of assemblies was estimated with Quast [31] and CEGMA [34].

Genome assemblies were annotated with Exonerate 2.2.0 with protein2genome and est2genome models using sets of S288C proteins and mRNAs, and the following *de novo* gene finders: Augustus 3.0.3 [96], GeneMark-ES v4.21 [97], and SNAP (version 2013-11-29) [98].

All sets of annotations were united using the MAKER2 pipeline v2.31.8 [32]. Genes found with this pipeline were divided into novel and known ones with an in-house Python3 script. This script employs a renaming table produced with proteinortho [33] by clustering the whole sets of

ORFs from the reference strain (S288C) and the strain analyzed. RepeatMasker v4.0.5 [99] and Tandem Repeat Finder v4.07 [100] were used to identify and mask repeated sequences.

Introgression analysis was carried out in two ways. First, we looked for S288C ORFs for which the best BLAST hit in the 15V-P4 genome had < 96% identity, extracted the corresponding regions from the 15V-P4 genome and looked for better (> 98% identity) BLAST hits in *S. sensu stricto* genomes, as described by Strobe *et al.* [8]. Second, short reads for the 15V-P4, YJM248 and S288C genomes were aligned to a reference consisting of concatenated *S. kudriavzevii* ZP 591, *S. bayanus* var. *uvarum* CBS 7001, *S. cerevisiae* S288C, *S. kudriavzevii* IFO1802<sup>T</sup>, *S. mikatae* IFO1815<sup>T</sup>, *S. paradoxus* CBS432, *S. eubayanus* FM1318, and *S. arboricolus* H-6 genomes, similarly to the analysis reported in [7]. Depth of coverage was reported with qualimap v2.1 [101]. YJM248 ([8]; GenBank accessions CP004414, CP004618, CP006335, CP004664, CP004758, CP004894, CP005198, CP006123, CP004986, CP005118, CP005295, CP006439, CP005396, CP005498, CP005592, CP006215, CP006478, CP004505, and SRA accession SRR800768) was used as a positive control for introgression in both cases. *S. sensu stricto* assemblies, reported in [102], were retrieved from [saccharomycesgenomes.org](http://saccharomycesgenomes.org). Short reads for S288C were reported in [103] and retrieved from SRA (SRR027936). *S. arboricolus* and *S. eubayanus* genome sequences, reported in [104, 105], respectively, were retrieved from the NCBI Genome portal (40941, 577061).

Mapping of short reads to the reference genome was performed with Bowtie v2.1.0 [106] for analysis of single nucleotide variation and with mrFast [107] for analysis of copy number variation. Quality control of bam files was performed with qualimap v2.2 [101]. Alignments were visualized with UGENE [108, 109] for manual check.

SNV calling on alignments was performed with samtools [110] v1.0 mpileup command with subsequent filtering of low quality ( $q < 30$ ) and low coverage ( $DP < 3$ ) positions with vcfutils [111] v1.0. Heterozygous indels and variations in the repeat regions were also filtered out.

SNVs were annotated with snpEff 4.1 [112]. snpEff output was used to infer the effect of mutations and dN/dS number. The NJ tree was built with hierarchical clustering in R [113]. To address the difference between individual Peterhof strains by SNVs according to the S288C genome bedtools-intersect [114] with the -v option was used. SNV distribution in the genome was visualized with the ggplot2 package for R [115].

Copy number variation was estimated with the mrCaNaVaR pipeline v0.51 [107]. Subsequent analysis was performed with R v3.2 [113]. 1kb windows with normalized copy number above 1.8 were considered as amplified while those with copy number below 0.3 were considered as deleted. These windows were merged to calculate length of amplified or deleted regions and intersected with reference genome annotation to produce lists of presumably amplified or missing genes. The resulting gene lists were analyzed with YeastMine [116].

Genome tracks for nucleotide variation and genome assemblies visualized with UCSC Genome Browser [117] and GARfield are available at <http://genome.ucsc.edu/cgi-bin/hgHubConnect#publicHubs> and at <http://garfield.dobzhanskycenter.org/cgi-bin/hgHubConnect>, respectively.

Conservative chromosome regions were extracted from the 15V-P4, YS9 and RedStar assemblies with lastz [118] with default settings and manually curated. The YS9 and RedStar assemblies were downloaded from the *Saccharomyces* genome database. The corresponding sequences from the other strains were reported in [8] and downloaded at <https://github.com/daskelly/yeast100genomes/>. Multiple alignment of these regions from 95 or 97 strains was performed with MAFFT v7.182 [119, 120] in fftnsi mode. Neighbor-joining tree was also constructed with MAFFT.

For the ORF-based tree, ORF sets for different strains were downloaded from the *Saccharomyces* Genome Database. MAKER2 [32] was used to collect 15V-P4 ORFs, and in-house

scripts were used to match them to the known reference genes, to intersect ORF sets and to distribute them into separate files, one for the each gene. Multiple alignment of common 807 ORFs was performed with MAFFT v7.182 [119, 120] in E-INS-i mode. Poorly aligned segments were filtered out with Gblocks v0.91b [121, 122] with a minimum block length equaling 6 bases and only positions where 50% or more of the sequences had a gap treated as a gap position. Maximum likelihood tree was constructed with RAxML v7.2.8 using rapid bootstrap analysis (-f a option) [123, 124].

Phylogenetic trees were visualized with the ape package for R [113] and with Figtree v1.4.0 [125].

Population structure analysis was performed with STRUCTURE v2.3.4 [35, 36]. For this, raw reads for the 15V-P4 genome were aligned to the reference S288C genome with BWA v0.6.1-r104 [126] in samse mode. SNV calling was performed with freebayes v1.0.2-6-g3ce827d [127] in haploid mode; SNVs with quality below 10 were filtered out. After that, these data were added to the data for 24,360 positions from the 100-genomes project [8] downloaded from <https://github.com/daskelly/yeast100genomes/> with vcftools [111] and distributed into 10 equally spaced minor allele frequency bins with plink v1.9 [128, 129]. Then, 121 positions were randomly extracted from each of the bins 3 times and merged to create 3 datasets with equal representation of each frequency bin. These datasets were recoded with PGDSpider v2.1.0.0 [130] and used as input to STRUCTURE. STRUCTURE was run for 1,000,000 iterations with 200,000 burnin period for 6 clusters. The results were united with CLUMPP v1.1.2 [131] and visualized with R [113].

Custom scripts used for data analysis are available at [https://github.com/drozdovapb/code\\_chunks/tree/master/Peterhof\\_strains\\_seq](https://github.com/drozdovapb/code_chunks/tree/master/Peterhof_strains_seq) and <https://github.com/drozdovapb/myBedGtfGffVcfTools>.

## Plasmids

Plasmids YGPM27n09, YGPM11114 and YGPM11e21 from the The Yeast Genomic Tiling Collection [132] were used to test complementation of the Gal<sup>-</sup> phenotype. Multicopy *LEU2* vector YEp351 [133] was used as a control.

pRS316-MSN4 was constructed by subcloning the 3.2 kb BamHI-EcoRI fragment from YGPM14i02 (The Yeast Genomic Tiling Collection; [132]) into the same sites of pRS316 [77]. pRS316-PHA2P, pRS316-pha2P-A10 and pRS316-pha2P-24 were constructed by cloning PCR products amplified with PHA2-F-BamHI and PHA2-R-BamHI (S8 Table) into the BamHI site of pRS316 [77]. *PHA2P* and *pha2P-24* were amplified using genomic DNA of 25-25-2V-P3982 as template (*pha2P-24* contains two PCR-induced missense mutations, N300S and F325S); *pha2P-A10* was amplified from 33G-D373 genomic DNA. The inserts were verified by sequencing (see S8 Table for information on the primers used); the same primers were used for sequencing of *PHA2* genomic allele. Sanger sequencing data were analyzed with UGENE [108].

## Phenotypic approaches

Standard yeast media [134] with minor modifications were used.

Yeast transformation was carried out according to the standard protocol [135] with modifications.

To test yeast abilities to grow in selective conditions, cells were suspended in water to equal OD<sub>595</sub> and spotted on solid media in 5- or 10-fold serial dilutions.

To test cell aggregation, strains were inoculated in liquid YEPD medium and grown overnight at 26°C until reaching the stationary phase. Then the cultures were diluted tenfold with fresh media and grown for additional 4 hours. Aliquots were placed on microscopic slides and photographed (5 fields of view, Zeiss Primostar microscope, 400x magnification).

## Supporting Information

**S1 Fig. Pedigree of the strains.** Names of strains with sequenced genomes are shown on yellow background. The number of generations is counted as the number of meiotic events between two strains. *MAT a* strains are depicted left-budded and *MAT* are right-budded. Diploids are unbudded. Curved arrow indicates self-fertilization. Dashed arrows indicate genetic manipulations without crossing. 25-25-2V-P3982 (the full name of the strain is 25-25-dU8-132-L28-2V-P3982) is an auxotrophic and suppressor mutant derivative of 2V-P3982 [28, 29]. 6P-33G-D373 is a 33G-D373 derivative in which *SUP35* is replaced with its homolog from *Pichia methanolica* [25].

(PDF)

**S2 Fig. Phylogenetic relation of 15V-P4 to other *S. cerevisiae* strains.** (A) Population structure of 101 strains including 15V-P4 assessed with three sets of 1210 variable positions with roughly uniform minor allele frequency distribution. Populations or groups of similar populations are framed. 15V-P4 and S288C are highlighted in red. (B) Neighbor joining phylogenetic tree of 97 strains including 15V-P4, RedStar and YS9 inferred from alignment of conservative chromosome regions. Bakery strains (RedStar and YS9) are highlighted in violet, 15V-P4 and S288C are highlighted in red.

(PDF)

**S3 Fig. Coverage of *Saccharomyces sensu stricto* genomes with short reads for 15V-P4 does not reveal introgression from any of the closely related species.** Short reads for the 15V-P4 genome were aligned to concatenated genomes of *S. sensu stricto* species with Bowtie2. S288C and YJM248 were used as a negative and positive controls for introgression, respectively. *Port*, *S. kudriavzevii* ZP 591. *Sbay*, *S. bayanus* var. *uvarum* CBS 7001. *Scer*, *S. cerevisiae* S288C. *Skud*, *S. kudriavzevii* IFO1802<sup>T</sup>. *Smik*, *S. mikatae* IFO1815<sup>T</sup>. *Spar*, *S. paradoxus* CBS432. *Seub*, *S. eubayanus* FM1318. *Sarb*, *S. arboricolus* H-6.

(PDF)

**S4 Fig. Genome coverage across reference for euploid strains.** (A) 1B, (B) 74. Dashed lines signify chromosome borders.

(PDF)

**S5 Fig. Neighbour joining (NJ) clustering of the PGC strains and S288C based on number of pairwise SNVs.** Shown in right are numbers of SNVs in comparison to S288C (highlighted in different shades of green with color intensity proportional to the number of SNVs) or to 15V-P4 (similarly highlighted in shades of purple).

(PDF)

**S6 Fig. Phenylalanine auxotrophy mutation *pheA10* is allelic to *PHA2*.** (A) Short read alignment. (B) Sanger resequencing. Red frame, TAA nonsense mutation appearing at codon 161. (C) 33G-D373 plated on selective media immediately after transformation with low copy number plasmids bearing indicated *PHA2* alleles. Vector, pRS316. (D) Asp<sup>+</sup> and Asp<sup>-</sup> variants of 6P-33G-D373 spotted for nonsense suppression and copper resistance. Series of 5-fold dilutions are shown.

(PDF)

**S7 Fig. Nonsense mutation in *MSN4* does not contribute to thermosensitivity.** Introduction of a centromeric plasmid with the wild type *MSN4* allele does not influence thermotolerance in 74-D694 (*[psi<sup>-</sup>]*) and P-74-D694 (*[PSI<sup>+</sup>]*). Series of 5-fold dilutions are shown. Vector, pRS316.

(PDF)

**S8 Fig. *GAL10*<sup>C287T</sup> mutation in the 1B strain may be responsible for the Gal- phenotype.** (A) SNVs in the *GAL* locus compared to S288C. Upper character, reference nucleotide; lower character, variant nucleotide. Nucleotides of the Watson strand are indicated. C287T substitution in *GAL10* of 1B is highlighted in blue circle. (B) The complete *GAL* locus or its *GAL10*-containing fragment but not *GAL1* alone compensates for 1B inability to grow on galactose-containing medium. 1B was transformed with multicopy plasmids containing the complete *GAL* locus (*GAL7*+*GAL10*+*GAL1*) or its fragments containing either only *GAL1* or *GAL7*+*GAL10*. Shown are series of 10-fold dilutions spotted on synthetic media lacking leucine with glucose or galactose/raffinose as a carbon source. Vector, YEp351. (C) Alignment of conservative part of UDP-galactose-4-epimerase homologs (Gal10 from *S. cerevisiae* S288C and 1B strains and GalE proteins from other species). In blue frame, Ala96Val substitution in 1B. In red frame, 94Val in human GALE.  
(PDF)

**S1 Table. Systematic names of genes used to infer the ORF-based phylogenetic tree.**  
(XLS)

**S2 Table. Summary of variable positions in the *SUCX* genes.** Positions are indicated according to S288C *SUC2* sequence. Variants are called according to short read alignment for sequenced PGC strains and to ungapped multiple alignment for known *SUC* genes (NCBI accession numbers are given in parentheses).  
(XLS)

**S3 Table. Summary of BLAST analysis for introgressed regions.** Shown are results of BLAST search (output format 6) in the 15V-P4 genome and in the YJM248 genome.  
(XLS)

**S4 Table. Genomic regions annotated as amplified or deleted in each of the genomes.**  
(XLS)

**S5 Table. Lists of genes annotated as amplified or deleted in each of the genomes and their functional characteristics.** Genes that are classified as amplified because they have close paralogs or similar sequences somewhere else in the genome are highlighted in beige, those residing in amplified chromosomes in gray, common deleted genes in orange, and known genotypic changes in green.  
(XLS)

**S6 Table. Number of SNVs and short indels in each of the genomes analyzed.**  
(PDF)

**S7 Table. List of genes with stop codons gained or lost in the strains analyzed.** Light green, known genotype.  
(XLS)

**S8 Table. Primers used in this work.**  
(XLS)

**S1 Appendix. Genetic variations previously described in 25-25, 1B, 74, and 6P-33G compared to S288C and 15V-P4.**  
(PDF)

## Acknowledgments

The authors would like to thank Aleksey Komissarov, Konstantin Okonechnikov, and Gaik Tamazian for help with data analysis, and Daniel Skelly for providing data and help with data analysis. We are grateful to Alexey Masharsky for Sanger sequencing and to John McCusker for providing the S1 strain. We would also like to thank Mikhail Belousov and Anton Gurkov for critical reading of the manuscript and the anonymous peer reviewers for their useful suggestions.

## Author Contributions

Conceived and designed the experiments: PBD OVT EAR SGIV PVD. Performed the experiments: PBD OVT AGM JVS DEP. Analyzed the data: PBD OVT AGM EAR PVD. Contributed reagents/materials/analysis tools: DEP JVS SGIV PVD. Wrote the paper: PBD OVT AGM SGIV. Established the Peterhof Genetic Collection: SGIV.

## References

1. Mortimer RK, Johnston JR. Genealogy of principal strains of the yeast genetic stock center. *Genetics*. 1986; 113(1): 35–43. PMID: [3519363](#)
2. Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, et al. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)*. 2014; 4: 389–98. doi: [10.1534/g3.113.008995](#)
3. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. *Science*. 1996; 274(5287): 546, 63–7. doi: [10.1126/science.274.5287.546](#) PMID: [8849441](#)
4. Lindegren CC. The yeast cell, its genetics and cytology. Educational Publishers, Inc, St. Louis; 1949.
5. Sherman F. Getting started with yeast. *Methods Enzymol*. 2002; 350: 3–41. doi: [10.1016/S0076-6879\(02\)50954-X](#) PMID: [12073320](#)
6. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. *Nature*. 2009; 458(7236): 337–41. doi: [10.1038/nature07743](#) PMID: [19212322](#)
7. Borneman AR, Pretorius IS. Genomic insights into the *Saccharomyces sensu stricto* complex. *Genetics*. 2015; 199(2): 281–91. doi: [10.1534/genetics.114.173633](#) PMID: [25657346](#)
8. Strobe PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, et al. The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome research*. 2015; 25(5): 762–74. doi: [10.1101/gr.185538.114](#) PMID: [25840857](#)
9. Toh-e A, Ueda Y, Kakimoto SI, Oshima Y. Isolation and characterization of acid phosphatase mutants in *Saccharomyces cerevisiae*. *J Bacteriol*. 1973; 113(2): 727–38.
10. Kane SM, Roth R. Carbohydrate metabolism during ascospore development in yeast. *J Bacteriol*. 1974; 118(1): 8–14. PMID: [4595206](#)
11. Borts RH, Lichten M, Hearn M, Davidow LS, Haber JE. Physical monitoring of meiotic recombination in *Saccharomyces cerevisiae*. *Cold Spring Harb Symp Quant Biol*. 1984; 49: 67–76. doi: [10.1101/SQB.1984.049.01.010](#) PMID: [6397320](#)
12. Malkova A, Swanson J, German M, McCusker JH, Housworth EA, Stahl FW, et al. Gene conversion and crossing over along the 405-kb left arm of *Saccharomyces cerevisiae* chromosome VII. *Genetics*. 2004; 168(1): 49–63. doi: [10.1534/genetics.104.027961](#) PMID: [15454526](#)
13. Yin Y, Petes TD. Genome-wide high-resolution mapping of UV-induced mitotic recombination events in *Saccharomyces cerevisiae*. *PLoS Genet*. 2013; 9(10): e1003894. doi: [10.1371/journal.pgen.1003894](#) PMID: [24204306](#)
14. Zhouravleva G, Frolova L, Le Goff X, Le Guellec R, Inge-Vechtomov S, Kisselev L, et al. Termination of translation in eukaryotes is governed by two interacting polypeptide chain release factors, eRF1 and eRF3. *The EMBO journal*. 1995; 14(16): 4065. PMID: [7664746](#)
15. Moskalenko S, Chabelskaya S, Inge-Vechtomov S, Philippe M, Zhouravleva G. Viable nonsense mutants for the essential gene *SUP45* of *Saccharomyces cerevisiae*. *BMC Mol Biol*. 2003; 4: 1–15.
16. Chabelskaya S, Kiktev D, Inge-Vechtomov S, Philippe M, Zhouravleva G. Nonsense mutations in the essential gene *SUP35* of *Saccharomyces cerevisiae* are non-lethal. *Mol Genet Genomics*. 2004; 272: 297–307. doi: [10.1007/s00438-004-1053-1](#) PMID: [15349771](#)

17. Chernoff YO, Derkach IL, Inge-Vechtormov SG. Multicopy *SUP35* gene induces *de-novo* appearance of *psi*-like factors in the yeast *Saccharomyces cerevisiae*. *Curr Genet*. 1993; 24: 268–70. doi: [10.1007/BF00351802](https://doi.org/10.1007/BF00351802) PMID: [8221937](https://pubmed.ncbi.nlm.nih.gov/8221937/)
18. Derkach IL, Bradley ME, Hong JY, Liebman SW. Prions affect the appearance of other prions: the story of [*PIN*<sup>+</sup>]. *Cell*. 2001; 106: 171–82. doi: [10.1016/S0092-8674\(01\)00427-5](https://doi.org/10.1016/S0092-8674(01)00427-5) PMID: [11511345](https://pubmed.ncbi.nlm.nih.gov/11511345/)
19. Du Z, Li L. Investigating the interactions of yeast prions: [*SWI*<sup>+</sup>], [*PSI*<sup>+</sup>], and [*PIN*<sup>+</sup>]. *Genetics*. 2014; 197: 685–700. doi: [10.1534/genetics.114.163402](https://doi.org/10.1534/genetics.114.163402) PMID: [24727082](https://pubmed.ncbi.nlm.nih.gov/24727082/)
20. Inge-Vechtormov SG. [New genetic lines of yeast *Saccharomyces cerevisiae*.] *Vestn Leningr Univ Ser Biol*. 1963; 21: 117–25.
21. Andrianova, VM, Samsonova MG, Sopova YV, Inge-Vechtormov SG. Catalogue of the Peterhof Genetic Collection of yeast *Saccharomyces cerevisiae*. St. Petersburg. 2003.
22. Inge-Vechtormov SG. [Identification of some linkage groups of Peterhof breeding stocks of yeast]. *Genetika*. 1971; 7(9): 113–24.
23. Aksenova AY, Volkov KV, Rovinsky NS, Svitin AV, Mironova LN. Phenotypic expression of epigenetic determinant [*ISP*<sup>+</sup>] in *Saccharomyces cerevisiae* depends on the combination of *sup35* and *sup45* mutations. *Mol Biol*. 2006; 40: 758–763. doi: [10.1134/S0026893306050104](https://doi.org/10.1134/S0026893306050104)
24. Chernoff YO, Lindquist SL, Ono B, Inge-Vechtormov SG, Liebman SW. Role of the chaperone protein Hsp104 in propagation of the yeast prion-like factor [*psi*<sup>+</sup>]. *Science*. 1995; 268: 880–4. doi: [10.1126/science.7754373](https://doi.org/10.1126/science.7754373) PMID: [7754373](https://pubmed.ncbi.nlm.nih.gov/7754373/)
25. Zadorsky SP, Inge-Vechtormov SG. [*SUP35* gene in *Pichia methanolica* is a recessive suppressor in *Saccharomyces cerevisiae*]. *Dokl Akad nauk*. 1998; 361(6): 825–9.
26. Zadorsky SP, Sopova YV, Andreichuk DY, Startsev VA, Medvedeva VP, Inge-Vechtormov SG. Chromosome VIII disomy influences the nonsense suppression efficiency and transition metal tolerance of the yeast *Saccharomyces cerevisiae*. *Yeast*. 2015; 32: 479–97. doi: [10.1002/yea.3074](https://doi.org/10.1002/yea.3074) PMID: [25874850](https://pubmed.ncbi.nlm.nih.gov/25874850/)
27. Volkov KV, Kurishko K, Inge-Vechtormov SG, Mironova LN. [Polymorphism of the *SUP35* gene and its product in the *Saccharomyces cerevisiae* yeasts]. *Genetika*. 2000; 36(2): 155–8. PMID: [10752025](https://pubmed.ncbi.nlm.nih.gov/10752025/)
28. Volkov KV, Aksenova AY, Soom MJ, Osipov KV, Svitin AV, Kurishko C, et al. Novel non-Mendelian determinant involved in the control of translation accuracy in *Saccharomyces cerevisiae*. *Genetics*. 2002; 160(1): 25–36. PMID: [11805042](https://pubmed.ncbi.nlm.nih.gov/11805042/)
29. Rogoza T, Goginashvili A, Rodionova S, Ivanov M, Viktorovskaya O, Rubel A, et al. Non-Mendelian determinant [*ISP*<sup>+</sup>] in yeast is a nuclear-residing prion form of the global transcriptional regulator Sfp1. *Proceedings of the National Academy of Sciences*. 2010; 107(23): 10573–7. doi: [10.1073/pnas.1005949107](https://doi.org/10.1073/pnas.1005949107)
30. Fitzpatrick DA, O'Brien J, Moran C, Hasin N, Kenny E, Cormican P, et al. Assessment of inactivating stop codon mutations in forty *Saccharomyces cerevisiae* strains: implications for [*PSI*<sup>+</sup>] prion-mediated phenotypes. *PLoS One*. 2011; 6: e28684. doi: [10.1371/journal.pone.0028684](https://doi.org/10.1371/journal.pone.0028684) PMID: [22194885](https://pubmed.ncbi.nlm.nih.gov/22194885/)
31. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013; 29(8): 1072–5. doi: [10.1093/bioinformatics/btt086](https://doi.org/10.1093/bioinformatics/btt086) PMID: [23422339](https://pubmed.ncbi.nlm.nih.gov/23422339/)
32. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 2011; 12: 491. doi: [10.1186/1471-2105-12-491](https://doi.org/10.1186/1471-2105-12-491) PMID: [22192575](https://pubmed.ncbi.nlm.nih.gov/22192575/)
33. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics*. BioMed Central Ltd; 2011; 12: 124. doi: [10.1186/1471-2105-12-124](https://doi.org/10.1186/1471-2105-12-124)
34. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. *Nucleic Acids Res*. 2009; 37: 289–97. doi: [10.1093/nar/gkn916](https://doi.org/10.1093/nar/gkn916) PMID: [19042974](https://pubmed.ncbi.nlm.nih.gov/19042974/)
35. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000; 155: 945–59. PMID: [10835412](https://pubmed.ncbi.nlm.nih.gov/10835412/)
36. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour*. 2009; 9: 1322–32. doi: [10.1111/j.1755-0998.2009.02591.x](https://doi.org/10.1111/j.1755-0998.2009.02591.x) PMID: [21564903](https://pubmed.ncbi.nlm.nih.gov/21564903/)
37. Song G, Dickins BJ, Demeter J, Engel S, Gallagher J, Choe K, et al. AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of *Saccharomyces cerevisiae*. *PLoS One*. 2015; 10(3): e0120671. doi: [10.1371/journal.pone.0120671](https://doi.org/10.1371/journal.pone.0120671) PMID: [25781462](https://pubmed.ncbi.nlm.nih.gov/25781462/)
38. Goto K, Iwatuki Y, Kitano K, Obata T, Kara S. Cloning and nucleotide sequence of the *KHR* killer gene of *Saccharomyces cerevisiae*. *Agric Biol Chem*. 1990; 54: 979–984. doi: [10.1271/bbb1961.54.979](https://doi.org/10.1271/bbb1961.54.979) PMID: [1368554](https://pubmed.ncbi.nlm.nih.gov/1368554/)

39. Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, et al. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci U S A*. 2007; 104(31): 12825–30. doi: [10.1073/pnas.0701291104](https://doi.org/10.1073/pnas.0701291104) PMID: [17652520](https://pubmed.ncbi.nlm.nih.gov/17652520/)
40. Ness F, Aigle M. *RTM1*: a member of a new family of telomeric repeated genes in yeast. *Genetics*. 1995; 140(3): 945–56. PMID: [7672593](https://pubmed.ncbi.nlm.nih.gov/7672593/)
41. Manente M, Ghislain M. The lipid-translocating exporter family and membrane phospholipid homeostasis in yeast. *FEMS Yeast Res*. 2009; 9(5): 673–87. doi: [10.1111/j.1567-1364.2009.00513.x](https://doi.org/10.1111/j.1567-1364.2009.00513.x) PMID: [19416366](https://pubmed.ncbi.nlm.nih.gov/19416366/)
42. Carlson M, Botstein D. Organization of the *SUC* gene family in *Saccharomyces*. *Mol Cell Biol*. 1983; 3(3): 351–9. doi: [10.1128/MCB.3.3.351](https://doi.org/10.1128/MCB.3.3.351) PMID: [6843548](https://pubmed.ncbi.nlm.nih.gov/6843548/)
43. Naumova ES, Sadykova AZ, Martynenko NN, Naumov GI. Molecular genetic characteristics of *Saccharomyces cerevisiae* distillers' yeasts. *Microbiology*. 2013; 82(2): 175–85. doi: [10.1134/S0026261713020112](https://doi.org/10.1134/S0026261713020112)
44. Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, et al. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. *PLoS Genet*. 2011; 7(2): e1001287. doi: [10.1371/journal.pgen.1001287](https://doi.org/10.1371/journal.pgen.1001287) PMID: [21304888](https://pubmed.ncbi.nlm.nih.gov/21304888/)
45. Borneman AR, Forgan AH, Kolouchova R, Fraser JA, Schmidt SA. Whole genome comparison reveals high levels of inbreeding and strain redundancy across the spectrum of commercial wine strains of *Saccharomyces cerevisiae*. *G3 (Bethesda)*. 2016; doi: [10.1534/g3.115.025692](https://doi.org/10.1534/g3.115.025692)
46. Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, et al. A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol*. 2014; 31: 872–88. doi: [10.1093/molbev/msu037](https://doi.org/10.1093/molbev/msu037) PMID: [24425782](https://pubmed.ncbi.nlm.nih.gov/24425782/)
47. Deregowska A, Skoneczny M, Adamczyk J, Kwiatkowska A, Rawska E, Skoneczna A, et al. Genome-wide array-CGH analysis reveals *YRF1* gene copy number variation that modulates genetic stability in distillery yeasts. *Oncotarget*. 2015; 6: 30650–63. doi: [10.18632/oncotarget.5594](https://doi.org/10.18632/oncotarget.5594) PMID: [26384347](https://pubmed.ncbi.nlm.nih.gov/26384347/)
48. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. 1997; 387: 708–3. PMID: [9192896](https://pubmed.ncbi.nlm.nih.gov/9192896/)
49. Wolfe KH. Origin of the yeast whole-genome duplication. *PLoS biology*. 2015; 13(8): e1002221. doi: [10.1371/journal.pbio.1002221](https://doi.org/10.1371/journal.pbio.1002221) PMID: [26252643](https://pubmed.ncbi.nlm.nih.gov/26252643/)
50. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol*. 2013; 9: e1003031. doi: [10.1371/journal.pcbi.1003031](https://doi.org/10.1371/journal.pcbi.1003031) PMID: [23592973](https://pubmed.ncbi.nlm.nih.gov/23592973/)
51. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012; 30: 434–9. doi: [10.1038/nbt.2198](https://doi.org/10.1038/nbt.2198) PMID: [22522955](https://pubmed.ncbi.nlm.nih.gov/22522955/)
52. Inge-Vechtomov SG, Tikhodeev ON, Karpova TS. [Selective systems for recessive ribosomal suppressors in yeast *Saccharomyces*]. *Genetika*. 1988; 24(7): 1159–65. PMID: [3053330](https://pubmed.ncbi.nlm.nih.gov/3053330/)
53. Bertram G, Bell HA, Ritchie DW, Fullerton G, Stansfield I. Terminating eukaryote translation: domain 1 of release factor eRF1 functions in stop codon recognition. *RNA*. 2000; 6: 1236–47. doi: [10.1017/S1355838200000777](https://doi.org/10.1017/S1355838200000777) PMID: [10999601](https://pubmed.ncbi.nlm.nih.gov/10999601/)
54. Nakayashiki T, Ebihara K, Bannai H, Nakamura Y. Yeast [*PSI<sup>+</sup>*] prions that are cross-transmissible and susceptible beyond a species barrier through a quasi-prion state. *Molecular cell*. 2001; 7(6): 1121–30. doi: [10.1016/S1097-2765\(01\)00259-3](https://doi.org/10.1016/S1097-2765(01)00259-3) PMID: [11430816](https://pubmed.ncbi.nlm.nih.gov/11430816/)
55. Inge-Vechtomov SG, Simarov BV. [Relation of supersuppression and interallelic complementation at *ad2* locus of *Saccharomyces cerevisiae*.] *Issledovaniya po genetike*. 1967; 3: 127–48.
56. Inge-Vechtomov SG, Gordenin DA, Kvasha VV. [Double mutants for *ade2* locus in yeast *Saccharomyces cerevisiae* and their utilization in intragenic mapping.] *Genetika*. 1975; 11(4): 121–133.
57. Struhl K. Genetic properties and chromatin structure of the yeast *gal* regulatory element: an enhancer-like sequence. *Proc Natl Acad Sci U S A*. 1984; 81: 7865–9. doi: [10.1073/pnas.81.24.7865](https://doi.org/10.1073/pnas.81.24.7865) PMID: [6096864](https://pubmed.ncbi.nlm.nih.gov/6096864/)
58. Sambuk EV, Ter Avanesyan MD. [Restoration of activities of acid phosphatase I and phosphoribosylaminoimidazol carboxylase as result of ochre-suppression of *pho1* and *ade2* mutations in yeast *Saccharomyces cerevisiae*.] *Genetika*. 1980; 16: 833–839.
59. Chabelskaya S, Gryzina V, Moskalenko S, Le Goff C, Zhouravleva G. Inactivation of NMD increases viability of *sup45* nonsense mutants in *Saccharomyces cerevisiae*. *BMC Mol Biol*. 2007; 8: 71. doi: [10.1186/1471-2199-8-71](https://doi.org/10.1186/1471-2199-8-71) PMID: [17705828](https://pubmed.ncbi.nlm.nih.gov/17705828/)

60. Meira LB, Henriques JA, Magaña-Schwencke N. 8-Methoxypsoralen photoinduced plasmid-chromosome recombination in *Saccharomyces cerevisiae* using a centromeric vector. *Nucleic Acids Res.* 1995; 23: 1614–20. doi: [10.1093/nar/23.9.1614](https://doi.org/10.1093/nar/23.9.1614) PMID: [7784218](https://pubmed.ncbi.nlm.nih.gov/7784218/)
61. Derkatch IL, Chernoff YO, Kushnirov VV, Inge-Vechtomov SG, Liebman SW. Genesis and variability of [PSI] Prion factors in *Saccharomyces cerevisiae*. *Genetics.* 1996; 144: 1375–1386. PMID: [8978027](https://pubmed.ncbi.nlm.nih.gov/8978027/)
62. Calderon IL, Contopoulou CR, Mortimer RK. Isolation of a DNA fragment that is expressed as an amber suppressor when present in high copy number in yeast. *Gene.* 1984; 29: 69–76. doi: [10.1016/0378-1119\(84\)90167-7](https://doi.org/10.1016/0378-1119(84)90167-7) PMID: [6092233](https://pubmed.ncbi.nlm.nih.gov/6092233/)
63. Rose M, Winston F. Identification of a Ty insertion within the coding sequence of the *S. cerevisiae* *URA3* gene. *Molecular and General Genetics MGG.* 1984; 193(3): 557–60. doi: [10.1007/BF00382100](https://doi.org/10.1007/BF00382100) PMID: [6323928](https://pubmed.ncbi.nlm.nih.gov/6323928/)
64. Inge-Vechtomov S, Andrianova V. [Recessive supersuppressors in yeast.] *Genetika.* 1970; 6: 103.
65. Chernoff IO, Derkach IL, Dagkesmanskaya AR, Tikhomirova VL, Ter-Avanesyan MD. [Nonsense-suppression during amplification of the gene coding for protein translation factor]. *Dokl Akad Nauk SSSR.* 1988; 301: 1227–9.
66. Mortimer RK, Hawthorne DC. Genetic mapping in *Saccharomyces*. *Genetics.* 1966; 53: 165–73. PMID: [5900603](https://pubmed.ncbi.nlm.nih.gov/5900603/)
67. Kobayashi O, Suda H, Ohtani T, Sone H. Molecular cloning and analysis of the dominant flocculation gene *FLO8* from *Saccharomyces cerevisiae*. *Mol Gen Genet.* 1996; 251(6): 707–15. doi: [10.1007/BF02174120](https://doi.org/10.1007/BF02174120) PMID: [8757402](https://pubmed.ncbi.nlm.nih.gov/8757402/)
68. Liu H, Styles CA, Fink GR. *Saccharomyces cerevisiae* S288C has a mutation in *FLO8*, a gene required for filamentous growth. *Genetics.* 1996; 144(3): 967–78. PMID: [8913742](https://pubmed.ncbi.nlm.nih.gov/8913742/)
69. Rupp S, Summers E, Lo HJ, Madhani H, Fink G. MAP kinase and cAMP filamentation signaling pathways converge on the unusually large promoter of the yeast *FLO11* gene. *EMBO J.* 1999; 18(5): 1257–69. doi: [10.1093/emboj/18.5.1257](https://doi.org/10.1093/emboj/18.5.1257) PMID: [10064592](https://pubmed.ncbi.nlm.nih.gov/10064592/)
70. Li J, Wang L, Wu X, Fang O, Lu C, Yang S, et al. Polygenic molecular architecture underlying non-sexual cell aggregation in budding yeast. *DNA Res.* 2013; 20(1): 55–66. doi: [10.1093/dnares/dss033](https://doi.org/10.1093/dnares/dss033) PMID: [23284084](https://pubmed.ncbi.nlm.nih.gov/23284084/)
71. Zhouravleva GA, Petrova AV. The role of translation termination factor eRF1 in the regulation of pseudohyphal growth in *Saccharomyces cerevisiae* cells. *Dokl Biochem Biophys.* 2010; 433(1): 209–211. doi: [10.1134/S1607672910040162](https://doi.org/10.1134/S1607672910040162) PMID: [20714858](https://pubmed.ncbi.nlm.nih.gov/20714858/)
72. Petrova A, Kiktev D, Askinazi O, Chabelskaya S, Moskalenko S, Zemlyanko O, et al. The translation termination factor eRF1 (Sup45p) of *Saccharomyces cerevisiae* is required for pseudohyphal growth and invasion. *FEMS yeast research.* 2015; 15(4): fov033. doi: [10.1093/femsyr/fov033](https://doi.org/10.1093/femsyr/fov033) PMID: [26054854](https://pubmed.ncbi.nlm.nih.gov/26054854/)
73. Namy O, Duchateau-Nguyen G, Hatin I, Hermann-Le Denmat S, Termier M, Rousset JP. Identification of stop codon readthrough genes in *Saccharomyces cerevisiae*. *Nucleic acids research.* 2003; 31(9): 2289–96. doi: [10.1093/nar/gkg330](https://doi.org/10.1093/nar/gkg330) PMID: [12711673](https://pubmed.ncbi.nlm.nih.gov/12711673/)
74. True HL, Lindquist SL. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature.* 2000; 407: 477–83. doi: [10.1038/35035005](https://doi.org/10.1038/35035005) PMID: [11028992](https://pubmed.ncbi.nlm.nih.gov/11028992/)
75. Robinson JS, Klionsky DJ, Banta LM, Emr SD. Protein sorting in *Saccharomyces cerevisiae*: isolation of mutants defective in the delivery and processing of multiple vacuolar hydrolases. *Mol Cell Biol.* 1988; 8(11): 4936–48. doi: [10.1128/MCB.8.11.4936](https://doi.org/10.1128/MCB.8.11.4936) PMID: [3062374](https://pubmed.ncbi.nlm.nih.gov/3062374/)
76. Heitman J, Movva NR, Hiestand PC, Hall MN. FK 506-binding protein proline rotamase is a target for the immunosuppressive agent FK 506 in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A.* 1991; 88(5): 1948–52. doi: [10.1073/pnas.88.5.1948](https://doi.org/10.1073/pnas.88.5.1948) PMID: [1705713](https://pubmed.ncbi.nlm.nih.gov/1705713/)
77. Sikorski RS, Hieter P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics.* 1989; 122(1): 19–27. PMID: [2659436](https://pubmed.ncbi.nlm.nih.gov/2659436/)
78. Kumar C, Sharma R, Bachhawat AK. Investigations into the polymorphisms at the ECM38 locus of two widely used *Saccharomyces cerevisiae* S288C strains, YPH499 and BY4742. *Yeast.* 2003; 20(10): 857–63. doi: [10.1002/yea.1012](https://doi.org/10.1002/yea.1012) PMID: [12868055](https://pubmed.ncbi.nlm.nih.gov/12868055/)
79. Lacroute F. Regulation of pyrimidine biosynthesis in *Saccharomyces cerevisiae*. *J Bacteriol.* 1968; 95(3): 824–32. PMID: [5651325](https://pubmed.ncbi.nlm.nih.gov/5651325/)
80. Dunlop PC, Roon RJ. L-asparaginase of *Saccharomyces cerevisiae*: an extracellular enzyme. *J Bacteriol.* 1975; 122(3): 1017–24. PMID: [238936](https://pubmed.ncbi.nlm.nih.gov/238936/)
81. Baldacci G, Zennaro E. Mitochondrial transcripts in glucose-repressed cells of *Saccharomyces cerevisiae*. *Eur J Biochem.* 1982; 127(2): 411–6. doi: [10.1111/j.1432-1033.1982.tb06887.x](https://doi.org/10.1111/j.1432-1033.1982.tb06887.x) PMID: [6754381](https://pubmed.ncbi.nlm.nih.gov/6754381/)

82. Jung PP, Fritsch ES, Blugeon C, Souciet JL, Potier S, Lemoine S, et al. Ploidy influences cellular responses to gross chromosomal rearrangements in *Saccharomyces cerevisiae*. *BMC Genomics*. 2011; 12: 331. doi: [10.1186/1471-2164-12-331](https://doi.org/10.1186/1471-2164-12-331) PMID: [21711526](https://pubmed.ncbi.nlm.nih.gov/21711526/)
83. Majumdar S, Ghatak J, Mukherji S, Bhattacharjee H, Bhaduri A. UDPgalactose 4-epimerase from *Saccharomyces cerevisiae*. A bifunctional enzyme with aldose 1-epimerase activity. *Eur J Biochem*. 2004; 271(4): 753–9. doi: [10.1111/j.1432-1033.2003.03974.x](https://doi.org/10.1111/j.1432-1033.2003.03974.x) PMID: [14764091](https://pubmed.ncbi.nlm.nih.gov/14764091/)
84. Thoden JB, Wohlers TM, Fridovich-Keil JL, Holden HM. Human UDP-galactose 4-epimerase. Accommodation of UDP-N-acetylglucosamine within the active site. *J Biol Chem*. 2001; 276(18): 15131–6. doi: [10.1074/jbc.M100220200](https://doi.org/10.1074/jbc.M100220200) PMID: [11279032](https://pubmed.ncbi.nlm.nih.gov/11279032/)
85. Thoden JB, Wohlers TM, Fridovich-Keil JL, Holden HM. Molecular basis for severe epimerase deficiency galactosemia. X-ray structure of the human V94M-substituted UDP-galactose 4-epimerase. *J Biol Chem*. 2001; 276(23): 20617–23. doi: [10.1074/jbc.M101304200](https://doi.org/10.1074/jbc.M101304200) PMID: [11279193](https://pubmed.ncbi.nlm.nih.gov/11279193/)
86. Wohlers TM, Christacos NC, Harreman MT, Fridovich-Keil JL. Identification and characterization of a mutation, in the human UDP-galactose-4-epimerase gene, associated with generalized epimerase-deficiency galactosemia. *Am J Hum Genet*. 1999; 64(2): 462–70. doi: [10.1086/302263](https://doi.org/10.1086/302263) PMID: [9973283](https://pubmed.ncbi.nlm.nih.gov/9973283/)
87. Timson DJ. Functional analysis of disease-causing mutations in human UDP-galactose 4-epimerase. *FEBS J*. 2005; 272(23): 6170–7. doi: [10.1111/j.1742-4658.2005.05017.x](https://doi.org/10.1111/j.1742-4658.2005.05017.x) PMID: [16302980](https://pubmed.ncbi.nlm.nih.gov/16302980/)
88. Lada AG, Stepchenkova EI, Waisertreiger IS, Noskov VN, Dhar A, Eudy JD, et al. Genome-wide mutation avalanches induced in diploid yeast cells by a base analog or an APOBEC deaminase. *PLoS Genet*. 2013; 9(9): e1003736. doi: [10.1371/journal.pgen.1003736](https://doi.org/10.1371/journal.pgen.1003736) PMID: [24039593](https://pubmed.ncbi.nlm.nih.gov/24039593/)
89. Newnam GP, Wegrzyn RD, Lindquist SL, Chernoff YO. Antagonistic interactions between yeast chaperones Hsp104 and Hsp70 in prion curing. *Molecular and cellular biology*. 1999; 19(2): 1325–33. doi: [10.1128/MCB.19.2.1325](https://doi.org/10.1128/MCB.19.2.1325) PMID: [9891066](https://pubmed.ncbi.nlm.nih.gov/9891066/)
90. Andrews S. FastQC. A quality control tool for high throughput sequence data. 2010. Available: <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>
91. Fastx toolkit. Available: [http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)
92. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*. 2011; 17: 10. doi: [10.14806/ej.17.1.200](https://doi.org/10.14806/ej.17.1.200)
93. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, Pyshkin AV. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*. 2012; 19(5): 455–77. doi: [10.1089/cmb.2012.0021](https://doi.org/10.1089/cmb.2012.0021) PMID: [22506599](https://pubmed.ncbi.nlm.nih.gov/22506599/)
94. Tamazian G, et al. Chromosomer: a reference-assisted genome assembly tool for producing draft chromosome sequences. 2015. Available: <https://github.com/gtamazian/chromosomer>
95. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic acids research*. 2011: gkr1029.
96. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003; 19: ii215–ii225. doi: [10.1093/bioinformatics/btg1080](https://doi.org/10.1093/bioinformatics/btg1080) PMID: [14534192](https://pubmed.ncbi.nlm.nih.gov/14534192/)
97. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*. 2005; 33(20): 6494–506. doi: [10.1093/nar/gki937](https://doi.org/10.1093/nar/gki937) PMID: [16314312](https://pubmed.ncbi.nlm.nih.gov/16314312/)
98. Korf I. Gene finding in novel genomes. *BMC bioinformatics*. 2004 14; 5(1): 59. doi: [10.1186/1471-2105-5-59](https://doi.org/10.1186/1471-2105-5-59) PMID: [15144565](https://pubmed.ncbi.nlm.nih.gov/15144565/)
99. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2015. Available: <http://www.repeatmasker.org>.
100. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999; 27: 573–580. doi: [10.1093/nar/27.2.573](https://doi.org/10.1093/nar/27.2.573) PMID: [9862982](https://pubmed.ncbi.nlm.nih.gov/9862982/)
101. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. 2015: btv566. doi: [10.1093/bioinformatics/btv566](https://doi.org/10.1093/bioinformatics/btv566)
102. Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, et al. The awesome power of yeast evolutionary genetics: New genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. *G3 (Bethesda)*. 2011. doi: [10.1534/g3.111.000273](https://doi.org/10.1534/g3.111.000273)
103. Qi J, Wijeratne AJ, Tomsho LP, Hu Y, Schuster SC, Ma H. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in *Saccharomyces cerevisiae*. *BMC Genomics*. *BioMed Central*; 2009; 10: 475. doi: [10.1186/1471-2164-10-475](https://doi.org/10.1186/1471-2164-10-475)

104. Liti G, Nguyen Ba AN, Blythe M, Müller CA, Bergström A, Cubillos FA, et al. High quality *de novo* sequencing and assembly of the *Saccharomyces arboricolus* genome. BMC Genomics. 2013; 14: 69. doi: [10.1186/1471-2164-14-69](https://doi.org/10.1186/1471-2164-14-69) PMID: [23368932](https://pubmed.ncbi.nlm.nih.gov/23368932/)
105. Baker E, Wang B, Bellora N, Peris D, Hulfachor AB, Koshalek JA, et al. The genome sequence of *Saccharomyces eubayanus* and the domestication of lager-brewing yeasts. Mol Biol Evol. 2015; 32: 2818–31. doi: [10.1093/molbev/msv168](https://doi.org/10.1093/molbev/msv168) PMID: [26269586](https://pubmed.ncbi.nlm.nih.gov/26269586/)
106. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–9. doi: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) PMID: [22388286](https://pubmed.ncbi.nlm.nih.gov/22388286/)
107. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009; 41 (10): 1061–7. doi: [10.1038/ng.437](https://doi.org/10.1038/ng.437) PMID: [19718026](https://pubmed.ncbi.nlm.nih.gov/19718026/)
108. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012; 28(8): 1166–7. doi: [10.1093/bioinformatics/bts091](https://doi.org/10.1093/bioinformatics/bts091) PMID: [22368248](https://pubmed.ncbi.nlm.nih.gov/22368248/)
109. Golosova O, Henderson R, Vaskin Y, Gabrielian A, Grekhov G, Nagarajan V, et al. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. PeerJ. 2014; 2: e644. doi: [10.7717/peerj.644](https://doi.org/10.7717/peerj.644) PMID: [25392756](https://pubmed.ncbi.nlm.nih.gov/25392756/)
110. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16): 2078–9. doi: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) PMID: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/)
111. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15): 2156–8. doi: [10.1093/bioinformatics/btr330](https://doi.org/10.1093/bioinformatics/btr330) PMID: [21653522](https://pubmed.ncbi.nlm.nih.gov/21653522/)
112. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly. 2012; 6(2): 80–92.
113. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.
114. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–2. doi: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033) PMID: [20110278](https://pubmed.ncbi.nlm.nih.gov/20110278/)
115. Wickham H. ggplot2: elegant graphics for data analysis. Springer Science & Business Media; 2009.
116. Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine—an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit. Database (Oxford). 2012; 2012: bar062. doi: [10.1093/database/bar062](https://doi.org/10.1093/database/bar062)
117. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002 1; 12(6): 996–1006. doi: [10.1101/gr.229102](https://doi.org/10.1101/gr.229102) PMID: [12045153](https://pubmed.ncbi.nlm.nih.gov/12045153/)
118. Harris RS. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University. 2007. Available: [http://www.bx.psu.edu/~rsharris/rsharris\\_phd\\_thesis\\_2007.pdf](http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf)
119. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005; 33(2): 511–8. doi: [10.1093/nar/gki198](https://doi.org/10.1093/nar/gki198) PMID: [15661851](https://pubmed.ncbi.nlm.nih.gov/15661851/)
120. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4): 772–80. doi: [10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010) PMID: [23329690](https://pubmed.ncbi.nlm.nih.gov/23329690/)
121. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17(4): 540–52. doi: [10.1093/oxfordjournals.molbev.a026334](https://doi.org/10.1093/oxfordjournals.molbev.a026334) PMID: [10742046](https://pubmed.ncbi.nlm.nih.gov/10742046/)
122. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007; 56(4): 564–77. doi: [10.1080/10635150701472164](https://doi.org/10.1080/10635150701472164) PMID: [17654362](https://pubmed.ncbi.nlm.nih.gov/17654362/)
123. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006; 22(21): 2688–90. doi: [10.1093/bioinformatics/btl446](https://doi.org/10.1093/bioinformatics/btl446) PMID: [16928733](https://pubmed.ncbi.nlm.nih.gov/16928733/)
124. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol. 2008; 57(5): 758–71. doi: [10.1080/10635150802429642](https://doi.org/10.1080/10635150802429642) PMID: [18853362](https://pubmed.ncbi.nlm.nih.gov/18853362/)
125. Rambaut A. Figtree. 2012. Available: <http://tree.bio.ed.ac.uk/software/figtree>.
126. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754–60. doi: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) PMID: [19451168](https://pubmed.ncbi.nlm.nih.gov/19451168/)
127. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. preprint arXiv. 2012;arXiv: 1207.3907 [q-bio.GN]

128. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81: 559–75. doi: [10.1086/519795](https://doi.org/10.1086/519795) PMID: [17701901](https://pubmed.ncbi.nlm.nih.gov/17701901/)
129. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience. BioMed Central;* 2015; 4: 7. doi: [10.1186/s13742-015-0047-8](https://doi.org/10.1186/s13742-015-0047-8)
130. Lischer HEL, Excoffier L. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics.* 2012; 28: 298–9. doi: [10.1093/bioinformatics/btr642](https://doi.org/10.1093/bioinformatics/btr642) PMID: [22110245](https://pubmed.ncbi.nlm.nih.gov/22110245/)
131. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* 2007; 23: 1801–6. doi: [10.1093/bioinformatics/btm233](https://doi.org/10.1093/bioinformatics/btm233) PMID: [17485429](https://pubmed.ncbi.nlm.nih.gov/17485429/)
132. Jones G, Stalker J, Humphray S, West A. A systematic library for comprehensive overexpression screens in *Saccharomyces cerevisiae*. *Nat Methods.* 2008; 5: 239–241. doi: [10.1038/nmeth.1181](https://doi.org/10.1038/nmeth.1181) PMID: [18246075](https://pubmed.ncbi.nlm.nih.gov/18246075/)
133. Hill JE, Myers AM, Koerner TJ, Tzagoloff A. Yeast/*E. coli* shuttle vectors with multiple unique restriction sites. *Yeast.* 1986; 2: 163–7. doi: [10.1002/yea.320020304](https://doi.org/10.1002/yea.320020304) PMID: [3333305](https://pubmed.ncbi.nlm.nih.gov/3333305/)
134. Kaiser C, Michaelis S, Mitchell A. *Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1994.
135. Gietz RD, Woods RA. Yeast transformation by the LiAc/SS Carrier DNA/PEG method. *Methods Mol Biol.* 2006; 313: 107–20. PMID: [16118429](https://pubmed.ncbi.nlm.nih.gov/16118429/)