Sequencing genomes of *Saccharomyces cerevisiae* strains belonging to the Peterhof Genetic Collection helps elucidate the origin of several widely used laboratory strains

Oleg Tarasov, Polina Drozdova, E. Radchenko, D. Polev, P. Dobrynin, Sergey Inge-Vechtomov

Saint Petersburg State University, Russia, Saint-Petersburg, Universitetskaya emb. 7-9
ovtarasov@gmail.com

Introduction

Saccharomyces cerevisiae is a widely used model organism. The haploid *S. cerevisiae* strain S288c is the progenitor to many of commonly used yeast laboratory strains and gave the first sequenced eukaryotic genome. S288c and its relatives originate from Carbondale Breeding stocks of O. Winge and C. Lindegren, which resulted from crosses between not only *S. cerevisiae* itself but also other *Saccharomyces* species [1, 2]. The Peterhof genetic collection of yeast is unrelated to S288c and originates from an industrial distillery strain [3]. Some strains of this collection are widely used in the field of yeast prion research [4, 5, and other works]. A number of genetic variations between Peterhof and S288c-related strains was identified but the whole genome data for Peterhof strains are scarce.

To date, genomes of more than 150 yeast strains of different origin have been sequenced. Comparison of such a variety of genomes helps clarify the natural history of yeast populations and allow to characterize genomic elements that are selected under specific conditions. Thus, we aimed to characterize the genomes of some yeast strains from Peterhof genetic collection.

Results

In this work, we analyzed genomes of four *S. cerevisiae* strains. 15V-P4 is one of the haploid progenitors of the Peterhof genetic collection; it originates from the initial industrial strain XII through 7 generations of intertetrad self-fertilization and 3 subsequent inbred crosses. 25-25-2V-P3982 is a laboratory strain of pure Peterhof origin, and 1B-D1606 and 74-D694 are hybrid descendants of both Peterhof and S288c-derived strains. All strains sequenced are haploid.

Genomes of 15V-P4, 25-25-2V-P3982, and 1B-D1606 were sequenced with Ion Torrent PGM using unpaired reads. Raw reads for the 74-D694 genome produced with Illumina GAII were retrieved from http://bioinf.nuim.ie/ wp-content/uploads/2011/10/74D_sequence.txt.zip. Trimming of reads was performed with fastx_toolkit v0.0.13.1. Trimming length was chosen according to the basic statistics calculated with FastQC. Genome coverage was from about 12 to 40X for different strains.

We *de novo* assembled the reads with SPAdes v3.1.0 and Mira v4.0 and found that quality of the SPAdes assemblies was generally higher (Table 1 and data not shown). Thus, we used SPAdes assemblies for further analysis.

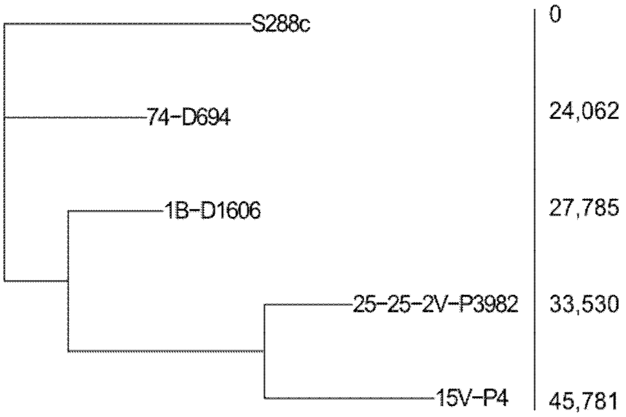
Table 1. Quast statistics of SPAdes assemblies.

	15V-P4	25-25-2V-P3982	1B-D1606	74-D694
Number of contigs	1,188	927	480	1,684
Largest contig, bases	92,741	102,076	252,839	71,636
Total length, bases	11,686,072	11,613,898	11,567,380	11,310,898
N50	18,500	25,573	73,204	11,948
Genome fraction, %	92.10	92.68	94.33	92.36
Number of S288c genes found	2,753 + 743 partial	4,832 + 863 partial	3,363 + 206 partial	2,848 + 630 partial

In order to assess the difference between our strains and the reference strain S288c, alignment of short reads to the reference genome was performed with bowtie v2.1.0. Then, SNP calling was performed with samtools v1.0 mpileup command with subsequent filtering with vcftools v1.0. All indels were filtered out with vcftools, and variations in the repeat regions (identified with RepeatMasker v4.0.2) were also filtered out.

We estimated genetic difference between our strains and S288c based on number of pairwise SNPs (fig. 1). 15V-P4 and S288c differs by 45,781 SNPs which is comparable to the level of divergence between distant *S. cerevisiae* populations reported previously [6]. As we could predict, “pure Peterhof” 25-25-2V-P3982 strain is the most similar to 15V-P4. However, these two strains have much more pairwise SNPs than we expected. We suppose that this difference may reflect laboratory evolution of the strain. 1B-D1606 and 74-D694 are roughly half as distant from S288c as 15V-P4 which is consistent with their hybrid origin.

Figure 1. NJ clustering of our strains and S288c based on number of pairwise SNPs. Shown in right are numbers of SNPs in comparison to S288c.



We tested whether any SNPs can be attributed to known phenotypic differences between Peterhof and S288c-derived strains. Amn1 and Flo8 are transcriptional regulators of cell aggregation in yeast. Amn1^{Asp368Val} and Flo8^{Trp142Stop} alleles are known to contribute much into change from clumping to non-clumping phenotype [7]. We observe the same tendency in our strains (data not shown).

De novo genome assemblies annotated with exonerate v2.2.0 were used to search for specific genes known to be lacking in S288c but present in other strains primarily of industrial or environmental origin [8].

All four strains studied possess the *KHR* and *RTM1* genes. The *KHR* gene encodes killer toxin of unknown nature. In 25-25-2V-P3982 and 1B-D1606, this gene is annotated on the same contigs as known genes of chromosome IX. In 15V-P4 and 74-D694, it is annotated on its own contig without any neighbouring ORFs. The *RTM1* gene is a member of a three-gene cluster associated with the subtelomeric sucrose utilization (SUC) locus that is present in several clinical, industrial, and environmental isolates. It encodes a lipid-translocating exporter and is known to be advantageous for strains growing on molasses.

In 15V-P4, we also found the cluster of five genes initially identified in wine strains [9]. Based on sequence, we suppose that 5-oxo-L-prolinase gene is a pseudogene while other four genes may be active. Interestingly, 15V-P4 appear to be the first yeast strain reported to obtain simultaneously the *RTM1* gene and the wine-specific cluster. It can be associated with

distillery origin of Peterhof genetic collection. Wine cluster is supposed to move in yeast genomes easily, therefore it could be quickly lost during laboratory evolution.
Other widely distributed genes we looked for (biotin biosynthesis genes *BIO1* and *BIO6*, fructose transporter *FSY1*, cell wall component *AWA1*, epoxide hydrolase *EHL*, and N-acetyltransferase *MPR1*) were found in none of four strains analyzed.

Conclusions

The results presented above show that genomes of the strains of the Peterhof genetic collection and of the S288c-based laboratory strains differ significantly and provide insight into some physiological differences of these strains such as clumping. The genomic data obtained are consistent with known origin of Peterhof strains. These data could form the basis for planning future work in these strains.

The work was supported by RFBR grants 14-04-31265, 14-04-32213, and 15-29-02526. Genome sequencing was performed in the Research Resource Center for molecular and cell technologies, Research park of Saint Petersburg State University.

References

1. R.K.Mortimer, J.R.Johnston (1986), *Genetics*, 113(1):35–43.
2. F. Sherman (2002), *Methods Enzymol*, 350:3–41.
3. S.G.Inge-Vechtomov (1963), *Vestn. LGU*, 21:117–125 (in Russian).
4. Y.O.Chernoff *et al.* (1993), *Curr Genet*, 24:268–270.
5. L.Westergard, H.L.True (2014), *Mol Microbiol*, 92:183–193.
6. G.Liti *et al.* (2009), *Nature*, 458:337–341.
7. J.Li *et al.* (2013), *DNA Res*, 20(1):55–66.
8. A.R.Borneman, I.S.Pretorius (2015), *Genetics*, 199(2):281–291.
9. A.R.Borneman *et al.* (2011), *PLoS Genet,* 7(2):e1001287.