FEMS Yeast Research, 19, 2019, foz064

https://doi.org/10.1093/femsyr/foz064
Advance Access Publication Date: 4 October 2019
Commentary

COMMENTARY

The pan-genome of *Saccharomyces cerevisiae*

Gang Li[1,†], Boyang Ji[1] and Jens Nielsen[1,2,3,*]

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden, [2]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark and [3]BioInnovation Institute, Ole Måløes Vej 3, DK-2200 Copenhagen N, Denmark

[*]**Corresponding author:** Kemivägen 10, SE-412 96, Göteborg, Sweden. Tel: +46 31 772 3804; Fax: +46 31 772 3801; E-mail: nielsenj@chalmers.se

One sentence summary: This article describes the reconstruction and application of the yeast pan-genome in resolving genotype–phenotype relationship by a machine learning-assisted approach.

Editor: Hyun Ah Kang

[†]Gang Li, http://orcid.org/0000-0001-6778-2842

ABSTRACT

Understanding genotype–phenotype relationship is fundamental in biology. With the benefit from next-generation sequencing and high-throughput phenotyping methodologies, there have been generated much genome and phenome data for *Saccharomyces cerevisiae*. This makes it an excellent model system to understand the genotype–phenotype relationship. In this paper, we presented the reconstruction and application of the yeast pan-genome in resolving genotype–phenotype relationship by a machine learning-assisted approach.

Keywords: *Saccharomyces cerevisiae*; pan-genome; genotype–phenotype relationship; machine learning

Saccharomyces cerevisiae is one of the most well-studied model organisms and is a widely used cell factory for production of fuels, chemicals and pharmaceuticals (Nielsen 2015). Since the release of the first complete genome sequence of strain S288C (Goffeau *et al*. 1996), there has recently been a rapidly increasing number of sequenced genomes for different *S. cerevisiae* strains due to the development of next-generation sequencing technology (Gallone *et al*. 2016; Peter *et al*. 2018). This large number of genomes enables us to define the pan-genome, which accounts for a set of all genes across all strains within this important species. Even though many efforts have been devoted to the construction of the pan-genome of *S. cerevisiae* (Song *et al*. 2015; Gallone *et al*. 2016; Peter *et al*. 2018; McCarthy and Fitzpatrick 2019), a holistic construction has not yet been accomplished. Inconsistent pan-genome generated is due to different dataset or different clustering methodology used. A comprehensive and high-quality pan-genome is therefore desirable.

We collected 1392 genomes of *S. cerevisiae* isolates from GenBank and published literatures (Fig. 1A; Note S1, Supporting Information), which is the most comprehensive list to date for the pan-genome construction. Genes in these genomes were annotated by a combination of a homology-based method and an *ab initio* gene prediction method (Note S2, Supporting Information). After removal of genomes that showed low completeness or that were highly fragmented (Note S3, Supporting Information), 1364 genomes with 8 947 177 predicted proteins were collected for analysis (Fig. 1A). Most of the genomes have 6000–7000 protein-encoding genes (Fig. 1B). There are 208 genomes with >7000 protein-encoding genes, which is higher than the number of 6705 protein-encoding genes in the reference strain *S. cerevisiae* S288C. There are several possible explanations linked to the high number of genes in our dataset: (i) *De novo* assembly from Illumina sequencing. The *de novo* genome assemblies may result in incomplete and fragmented contigs/scaffolds containing misassembled regions or errors, which can also explain the presence of numerous fragmented genes in the assembled genomes (supplemented dataset on Zenodo). (ii) Ploidy and aneuploidy variation in *S. cerevisiae*. Two hundred and one out of two hundred and eight genomes are with known state of ploidy and aneuploidy: they all have a ploidy number of at least 2. Seventy seven strains are aneuploid and one hundred and twenty four

Received: 21 June 2019; **Accepted:** 1 October 2019

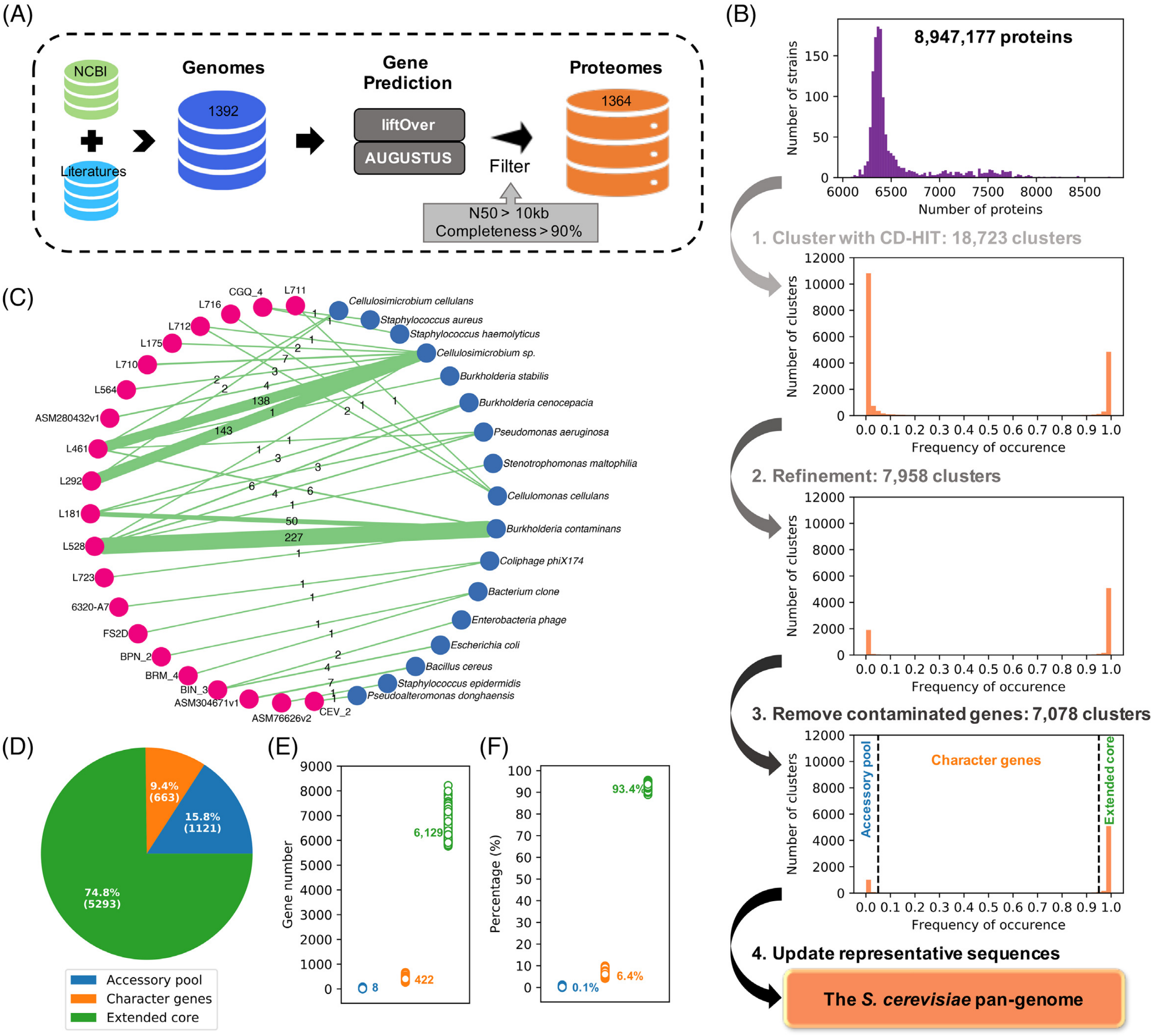

Figure 1. The construction of S. *cerevisiae* pan-genome. **(A)** Collection of genomes and prediction of proteomes. **(B)** The pipeline for protein sequence clustering. **(C)** Detection of genome contamination by prokaryotes. Pink dots denote contaminated S. *cerevisiae* genomes. Blue dots denote the suspected contamination source. The number on the linkage denotes the numbers of contigs that are suspected to be contaminated. **(D)** The fraction/number of genes in each of three categories. **(E, F)** Genome compositions of 1360 genomes. A, Accessory pool; C, Character genes; E, Extended core. The numbers labeled in (E) and percentage in (F) showed average gene numbers and average percentage of genes in a genome belonging to each category, respectively.

are euploid. They are from very diverse isolation sources (Table S1, Supporting Information).

These protein sequences were first clustered with CD-HIT (Li and Godzik 2006) and then refined with a pairwise alignment to regroup those clusters that are separated due to the existence of insertions/deletions or a number of unknown amino acids in the sequence (Note S4, Supporting Information; Fig. 1B). We noticed that some genomes contain a large number of genes that are from prokaryotic organisms (Note S5, Supporting Information; Figure S1, Supporting Information). After carefully comparing those contigs that only contain prokaryotic genes with NCBI nr database (Note S5, Supporting Information), 20 genomes in our collection were found to contain at least 1 suspected contaminated contig. Particularly, 4 of them contained >50 contigs (Fig. 1C), and these 4 genomes were therefore removed from our genome collection. For other contaminated genomes, only genes from the contaminated contigs were removed. This resulted in 7078 clusters. This estimated pan-genome size is smaller than 7796 reported in Peter *et al.* (2018) and 7750 reported in McCarthy and Fitzpatrick (2019). This is partly due to the use of different clustering strategies and removal of contaminated genes. We refer these 7078 protein clusters as 7078 genes in the pan-genome hereafter.

Using the similar definition as in Lapierre and Gogarten (2009), we categorized the final gene clusters into three groups: extended core (genes that are present in at least 95% of strains), accessory pool (genes that are present in only 5% or less of strains) and character genes (genes that are present in 5–95% of strains). The reason for the use of extended core instead of core is that almost all of the genomes we are using are incomplete, which may lead to incomplete predicted proteomes. This factor has not been considered in most of the previous

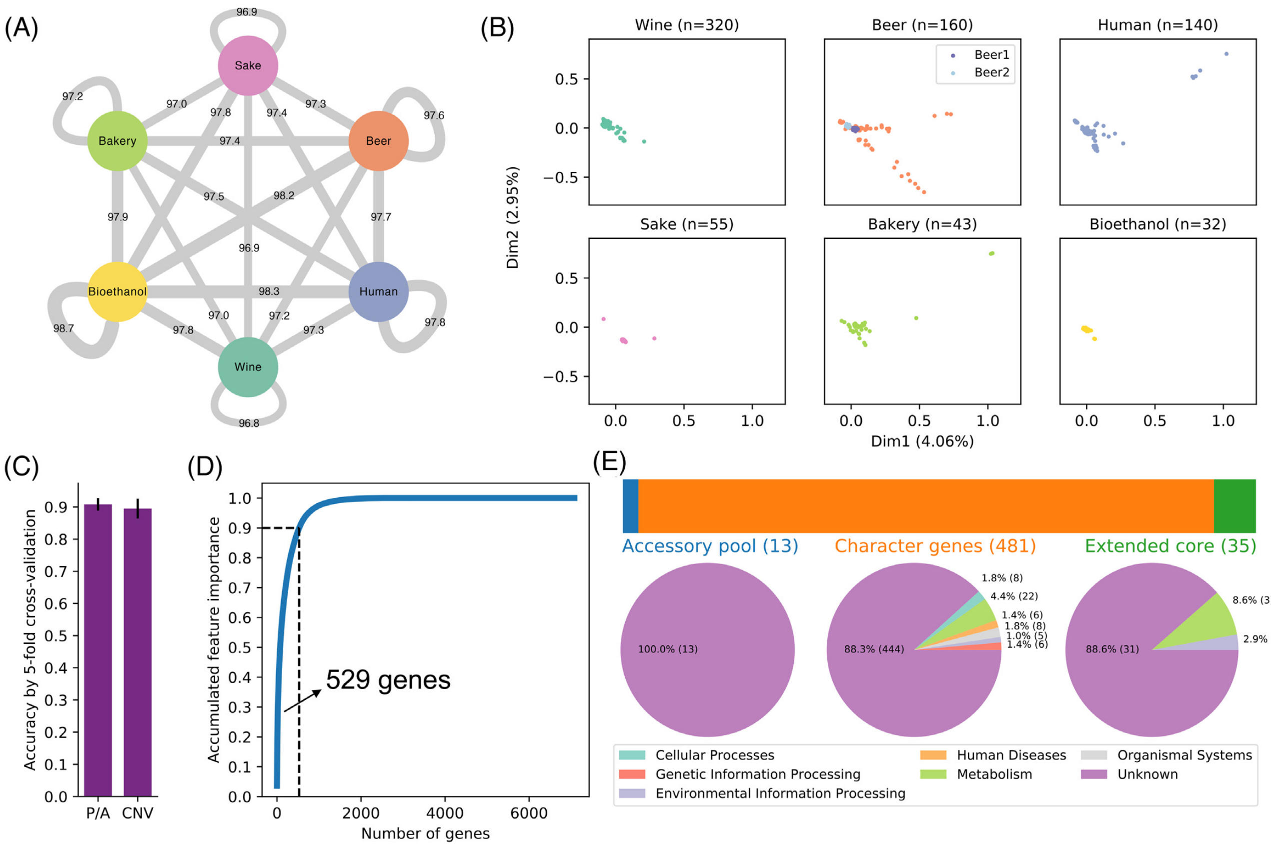

Figure 2. Analysis of genotype–phenotype relationship with machine learning in *S. cerevisiae*. **(A)** Average Jaccard similarity score between genome content of different *S. cerevisiae* types. Numbers on edges indicate the Jaccard similarity score (%). **(B)** MCA on the presence/absence of genes in pan-genome. **(C)** Accuracy score obtained by 5-fold cross-validation with a random forest classifier. P/A, gene presence/absence table. CNV, gene copy number variation. **(D)** Accumulated feature importance curve from (C) on P/A dataset. The top 527 genes contribute to 90% of the prediction power. **(E)** The pan-genome categories and KEGG function categories of the top 527 genes. See more details about methods in Note S6 (Supporting Information).

S. cerevisiae pan-genome reconstructions. This led to identification of 5293 out of 7078 gene clusters belonging to the extended core (Fig. 1D). On average, a typical *S. cerevisiae* genome is composed of 6129 protein-coding genes belonging to 5279 extended core gene groups (93.4% of all protein-coding genes in a genome), 422 belonging to 363 character gene groups (6.4%) and 8 belonging to 7 accessory gene groups (0.1%) (Fig. 1E and F).

There are several factors that affect the size of the final pan-genome from our pipeline: (i) the existence of insertions/deletions and/or unknown amino acids in the sequence; (ii) genome contamination; (iii) the identity threshold used for protein sequence clustering; and (iv) the number of genomes used. All these factors have only a slight effect on the size of the accessory pool but not on the extended core (Fig. 1B; Figures S2 and S3, Supporting Information).

Next, we tested how the genome content contributes to the differentiation of strain types. For this, 767 industrial and clinical strains belonging to 6 different types with >30 strains in each class and clear class labels were selected. To get quantitative estimation of genome content differences between strain types, average Jaccard similarities (Jaccard 1901) between strain types were calculated (Note S6, Supporting Information; Fig. 2A). It showed that these strain types showed a very similar genome content. This is due to the large fraction of shared genes in the extended core genome (Fig. 1D–F, Fig. 2A). Then, multiple correspondence analysis (MCA) (Abdi and Valentin 2007) was applied to test whether those genomes of different types could be clustered based on gene content (gene presence/absence, P/A). The results showed that most of the *S. cerevisiae* strains clustered together, while a few outlier strains showed distinct gene presence patterns (Table S2, Supporting Information), like some beer yeast (Fig. 2B; Figure S4 (Supporting Information) for all 1360 genomes). The MCA was also able to distinguish strains from sub-classes 'Beer1' and 'Beer2' as previously described (Gallone *et al.* 2016) (Fig. 2B; a zoomed in version in Figure S5 (Supporting Information)). In the next step, a random forest classifier (Ho 1998) was applied to test to what extent genome content determines the strain types. The input features were either gene content (P/A) or copy number variation (CNV) of genomes. Model accuracies were calculated via a 5-fold cross-validation approach (Note S6, Supporting Information). Both models on P/A or CNV achieved an accuracy up to 90% (Fig. 2C). This indicates that in addition to the sequence variations in the core genome (Gallone *et al.* 2016), genome content also contributes to the differentiation of different strain types. The model doesn't perform better when considering gene copy number variation than only presence/absence. This indicates that gain/loss of some genes is more likely to determine the strain types rather than increase/decrease of the gene copy number. To identify the gain/loss of what genes make differences among different strain types, all genes in the pan-genome were scored with the feature importance value from the random forest classifier trained on P/A table. Five hundred twenty seven genes were found to contribute 90% of the predictive power (Fig. 2D). The genes are mainly from the group of character genes in the pan-genome (Fig. 2E), even though only 9.4% of genes in the pan-genome are character genes (Fig. 1D). Most of the genes are poorly annotated so far. The second largest groups of the genes are metabolic genes (Fig. 2E). This makes sense as most of those selected strains were used for the production of different products in industry and hence have been selected for having distinct metabolism, e.g. for utilization of certain metabolites in the medium or producing specific flavors.

In this study, it was shown that strain types could be easily classified by a random forest model trained only on gene presence/absence information with a very high accuracy. However, in the case of more complicated quantitative phenotypes (like the ones described in Gallone *et al.* (2016) and Peter *et al.* (2018)), only genome content information may be insufficient to explain such phenotypic variation. More factors have to be taken into consideration, including chromosomal rearrangements (Hou *et al.* 2014), SNPs and other factors (Märtens *et al.* 2016).

In conclusion, we presented a holistic reconstruction of the *S. cerevisiae* pan-genome. We provided an application of the pan-genome by applying machine learning to resolve the genotype–phenotype relationship. The pan-genome will prove valuable for many fundamental applications, such as reconstruction of genome-scale metabolic models for this species and understanding the role of polymorphisms on strain phenotypes.

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie program, project PAcMEN (grant agreement no. 722287). We also acknowledge funding from the Novo Nordisk Foundation (grant no. NNF10CC1016517), the Knut and Alice Wallenberg Foundation and the Infect-ERA program FunComPath.

CODE AND DATA AVAILABILITY

The pan-genome datasets were available in Zenodo (http://doi.org/10.5281/zenodo.3407352). All codes are available from the authors upon request.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSYR online.

ACKNOWLEDGMENTS

The computations were performed on resources at Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC).

Author contributions: GL, BJ and JN designed research; GL performed research; GL, BJ and JN analyzed the data and GL, BJ and JN wrote the paper.

Conflicts of interest. None declared.

REFERENCES

Abdi H, Valentin D. Multiple correspondence analysis. In: Salkind N (ed). *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications, 2007, 651–7.

Gallone B, Steensels J, Prahl T *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* 2016;**166**:1397–410.e16.

Goffeau A, Barrell BG, Bussey H *et al.* Life with 6000 genes. *Science* 1996;**274**:546–67.

Ho TK. The random subspace method for constructing decision forests. *IEEE T Pattern Anal* 1998;**20**:832–44.

Hou J, Friedrich A, de Montigny J *et al.* Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in *Saccharomyces cerevisiae*. *Curr Biol* 2014;**24**:1153–9.

Jaccard P. *Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines*. Zürich, Schweiz: Bulletin de la Société Vaudoise des Sciences Naturelles, 1901.

Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. *Trends Genet* 2009;**25**:107–10.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.

Märtens K, Hallin J, Warringer J *et al.* Predicting quantitative traits from genome and phenome with near perfect accuracy. *Nat Commun* 2016;**7**:11512.

McCarthy CGP, Fitzpatrick DA. Pan-genome analyses of model fungal species. *Microb Genom* 2019;**5**:e000243.

Nielsen J. Yeast cell factories on the horizon. *Science* 2015;**349**:1050–1.

Peter J, De Chiara M, Friedrich A *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 2018;**556**:339–44.

Song G, Dickins BJA, Demeter J *et al.* AGAPE (Automated Genome Analysis PipelinE) for pan-genome analysis of *Saccharomyces cerevisiae*. *PLoS One* 2015;**10**:e0120671.