Entering GATTACA

Yeast genomes: Analysis, insights and applications

Daran, Jean Marc G.

DOI
10.1093/femsyr/foaa064

Publication date
2020

Document Version
Accepted author manuscript

Published in
FEMS Yeast Research

Thematic issue FEMS Yeast Research –

Entering GATTACA: Yeast genomes: analysis, insights and applications

Jean-Marc G. Daran, Delft University of Technology, Department of Biotechnology, van der Maasweg 9, 2628HZ, Delft, The Netherlands.

J.g.daran@tudelft.nl

A quarter of a century ago, publication of the 12 Mb genome sequence of *Saccharomyces cerevisiae* S288C transformed yeast research forever and further cemented the strength and status of this yeast as a versatile laboratory model for eukaryotic cells. Knowledge on the gene complement and genome sequence of *S. cerevisiae* inspired yeast scientists to address the next challenge: elucidation of the biological functions of all its 6000 genes (Goffeau, et al. 1996). This was no trivial challenge, as the sequencing project had just revealed a thousand genes with uncharacterised function.

The EU-funded EUROFAN project, in which functions of *S. cerevisiae* genes were investigated via a systematic and hierarchical approach, yielded one of the best annotated genomes available at the time (Dujon 1998). In addition, the *S. cerevisiae* genome soon took centre stage in seminal genome-wide expression studies and systems-biology research, as exemplified by many early studies with Affymetrix DNA microarrays (ter Linde, et al. 1999) and the first genome-scale yeast metabolic models (Forster, et al. 2003), respectively. Availability of this resource not only contributed to acceleration of fundamental and applied research on *S. cerevisiae* itself, but also stimulated research on other, 'non-conventional' yeast species. However, for over a decade, the *S. cerevisiae* genome remained the only high-quality, accurately annotated genome sequence available to the global yeast research community.

The advent of Next Generation Sequencing (NGS) technologies set off a new revolution in yeast genomics. In contrast to the Sanger sequencing technology used for the original yeast genome sequencing project, NGS does not require laborious construction of genome libraries. Instead, it relies on generation of hundreds of thousands of random fragments of the genome of interest, which are

subsequently sequenced simultaneously. This innovation enabled an enormous increase in sequencing throughput and an equally important massive reduction of genome sequencing costs. Commercial NGS services have short turn-around cycles, with the time between DNA extraction and the delivery of the sequencing data to the customer varying from 1 to 10 weeks and at costs around 150-300 €. euro.

NGS methodologies currently available are based on different physico-chemical principles, but can be roughly divided in two categories based on the length of the sequencing reads they generate. The first category, sometimes referred to as second-generation sequencing, mainly includes methods that generates short read sequences. The best known example is the short-read Illumina technology that can produce Gigabases of data from a single yeast genome in the form of 100- to 300-nucleotide fragments. Third-generation sequencing encompasses single-molecule sequencing methods, as represented by Single Molecule Real-Time (SMRT) Sequencing from Pacific Biosystem (PacBio) and the nanopore based single-molecule sequencing platforms developed by Oxford Nanopore Technology (ONT). Although based on different sequencing principles, these methods share the capacity to generate long sequencing reads whose length can cover tens of thousands of bases.

The different genome sequencing technologies now available to yeast researchers have specific advantages and applications. Short-read technologies are excellently suited for re-sequencing of mutants of strains for which a high-quality reference genome is available and for population genomics, since their high sequencing accuracy enables reliable detection of single-nucleotide variations. In addition, the high coverage of short-read sequencing methods can be used to infer chromosomal or segmental copy number variation, reveal ploidy changes and enable resolution of heterozygosity within sequenced genomes.

Although short reads can be used for de novo genome assembly, the resulting assemblies are generally fragmented in contigs that are much shorter than the actual chromosome sizes. In particular, complex genome features such as repeated regions (repeats) or haplotypes in polyploid genomes cannot be faithfully assembled and instead are collapsed into a single sequence by genome-assembly algorithms. This limitation was eliminated when long-read sequencing technologies became available, due to the

unambiguous mappability of multi-kb fragments. This approach generates highly reliable *de novo* genome assemblies as well as identification of chromosomal rearrangements. In addition, single-molecule sequencing are very useful for long-range analyses such as nucleotide variation phasing, a step towards haplotyping in non-haploid strains. Despite continual improvement, long-read sequencing platforms still exhibit a higher rate of sequencing errors than short-read platforms. Combination of second and third generation sequencing technologies has therefore become a standard approach in studies aimed at obtaining fully assembled, highly accurate genome sequences. Today, over one third of known budding yeast species have been sequenced, allowing robust and time-scaled phylogeny of the Saccharomycotina subphylum (Shen, et al. 2018). Moreover, several thousands of full genome sequence datasets of wild, domesticated and genetically modified *Saccharomyces* strains are available and retrievable from public repositories, thereby providing unprecedented options to experimentally explore the genetic basis of their natural and man-made phenotypic diversity.

This thematic issue of FEMS Yeast Research comprises a collection of nine mini-reviews that highlight important technological advances and applications in the field of yeast genome sequencing and genome analysis.

The review "**Into the wild: new yeast genomes from natural environments and new tools for their analysis**" by Libkind and co-authors (Libkind, et al. 2020b) demonstrate how high-throughput sequencing of a wide range of strains from different ecological or geographical origins can detect genomic signatures of pathogenicity and domestication and thereby shed light on the impact of the long interaction of *Saccharomyces* with mankind. They also demonstrate that these analyses can be extended to less intensively studied non-conventional yeasts. In "**Towards yeast taxogenomics: lessons from novel species descriptions based on complete genome sequences**", Libkind and co-authors (Libkind, et al. 2020a) discuss the challenges involved in the use of data-rich genome sequencing information for description of new taxa. In addition they outline how to use such information to for more reliable assessment of genetic distances and for constructing more robust

phylogenies. "**An update on the diversity, ecology and biogeography of the *Saccharomyces* genus**" by Alsammar and Delneri (Alsammar and Delneri 2020) provides an update on the geographical and ecological distribution of the *Saccharomyces* genus. This review also includes information on inter-species *Saccharomyces* hybrids, whose complex genomes can now be ever more accurately resolved by NGS. The majority of these hybrids appear to be related to human activities and, in particular, to industrial processes such as brewing and wine making.

Interestingly, hybridization seems to be a more common phenomenon in fungi. In "**Hybridization and the origin of new yeast lineages**", Gabaldon (Gabaldon 2020) demonstrates that fungal hybridization can occur across large phylogenetic distances and that yeasts belonging to the Saccharomycotina clade are particularly prone to hybridization. As observed in the genus *Saccharomyces*, hybrid formation can be promoted by interaction with humans, not only in yeast biotechnology but also as human opportunistic pathogens. In their review "**Lager-brewing yeasts in the era of modern genetics**", Gorter de Vries and co-authors (Gorter de Vries, et al. 2019) focus on the hybrid *Saccharomyces pastorianus* lager yeasts that annually produce almost 200 billion liters of lager-type beer. Evolutionary origin and genome structure of these yeasts are discussed from a historical, technical and socio-economical perspective and the authors discuss how recent developments in genome sequencing, genome editing and interspecies hybridization methods provide a new impulse to industrial strain improvement.

While an organism's genome sequence provides a blueprint of its capabilities, a real understanding of this blueprint requires analysis and integration of different types of data (e.g. transcript, protein, gene-regulation and protein–protein interaction data). Accurate, objective annotation of genomes, a key prerequisite for this field of functional yeast genomics, is discussed by Douglass et al. in "**The Methylotroph Gene Order Browser (MGOB) reveals conserved synteny and ancestral centromere locations in the yeast family Pichiaceae**" (Douglass, et al. 2019). Their study presents a comparative genome browser to perform gene orthology and synteny analysis in genomes of Pichiaceae, a family of budding yeasts that, despite its biotechnological relevance, remains understudied. MGOB complements a constellation of bioinformatics tools dedicated to specific yeast families to assist

genome annotation. In addition to their role in genome sequencing, advances in sequencing technologies have also enabled generation of high quality of RNA sequencing (RNA-seq) datasets for transcriptome analyses. Doughty and Kerkhoven, in "**Extracting novel hypotheses and findings from RNA-seq data**" (Douglass, et al. 2019), discuss a range of methodologies to extract meaningful information from multifactorial RNA-seq data. They review how, alongside the functional annotation classically employed in transcriptome analysis, application of quantifiable gene metrics are valuable to extract new information that complement traditional differential gene expression read-out in generating new testable hypothesis. They also draw attention to new opportunities offered by RNA seq to explore the role of long non-coding RNA molecules.

Proteome analyse provides an invaluable additional level of information in yeast functional genomics. In their review "**Shot-gun proteomics: why thousands of unidentified signals matter**", den Ridder and co-authors (den Ridder, et al. 2020) highlight the recent advances in microbial proteomics for unrestricted protein modification discovery, and progress in integration the resulting information to elucidate protein interaction and regulation in *S. cerevisiae*.

Genome-scale metabolic models (GEMs) have become essential platforms to organize information and to analyse, predict and redesign metabolic capabilities of yeast strains. In the final paper of this thematic issue, "**Evaluating accessibility, usability and interoperability of genome-scale metabolic models for diverse yeasts species**", Domenzain and co-authors (Domenzain, et al. 2020) address applicability of GEMs for different yeast species. Based on a comprehensive description of models available for Saccharomycotina yeasts, the authors advocate the adoption of new, standardized and ready-to-use formats to facilitate utilization of these models by non-experts.

In a period of 25 years, yeast genomics has transformed our daily work and has moved far beyond the iconic S288C sequence. Yeast genome sequencing is no longer performed by a handful of highly specialised research centres, but is a mainstay technique in many laboratories. In the design of guide RNAs for CRISPR-based genome editing, we take it for granted that accurate genome sequence information for the targeted strain is available from a database and, if not, can be quickly generated.

And, the number and complexity of genome modifications that can be achieved by new genome editing tools continues to increase, whole-genome sequencing, once an effort that required a world-wide effort, has become a routine quality control technique. Pocket-size MinION devices and equivalent equipment from other companies will drive this democratization of yeast genomics even further. As in “personalized medicine”, in which an individual’s genome sequence and derived –data are used to diagnose diseases and health risks, we are entering an era in which genomes of all yeast strains isolated, used and constructed in our labs will be rapidly and systematically sequenced and analysed to predict the strain performance.

I thank all the authors for contributing their perspective on this fast-moving field of yeast research and hope that this special issue will inspire readers to apply and extend the amazing possibilities of yeast genomics in their research.

References

Alsammar H, Delneri D. An update on the diversity, ecology and biogeography of the Saccharomyces genus. *FEMS Yeast Res* 2020;**20**;10.1093/femsyr/foaa013.

den Ridder M, Daran-Lapujade P, Pabst M. Shot-gun proteomics: why thousands of unidentified signals matter. *FEMS Yeast Res* 2020;**20**;10.1093/femsyr/foz088.

Domenzain I, F. L, J. KE, V. S. Evaluating accessibility, usability and interoperability of genome-scale 2 metabolic models for diverse yeasts species. *FEMS Yeast Res* 2020

Gabaldon T. Hybridization and the origin of new yeast lineages. *FEMS Yeast Res* 2020;**20**;10.1093/femsyr/foaa040.

Libkind D, Cadez N, Opulente DA, Langdon QK, Rosa CA, Sampaio JP, Goncalves P, Hittinger CT, Lachance MA. Towards yeast taxogenomics: lessons from novel species descriptions based on complete genome sequences. *FEMS Yeast Res* 2020a;**20**;10.1093/femsyr/foaa042.

Libkind D, Peris D, Cubillos FA, Steenwyk JL, Opulente DA, Langdon QK, Rokas A, Hittinger CT. Into the wild: new yeast genomes from natural environments and new tools for their analysis. *FEMS Yeast Res* 2020b;**20**;10.1093/femsyr/foaa008.

Douglass AP, Byrne KP, Wolfe KH. The Methylotroph Gene Order Browser (MGOB) reveals conserved synteny and ancestral centromere locations in the yeast family Pichiaceae. *FEMS Yeast Res* 2019;**19**;10.1093/femsyr/foz058.

Gorter de Vries AR, Pronk JT, Daran JG. Lager-brewing yeasts in the era of modern genetics. *FEMS Yeast Res* 2019;**19**;10.1093/femsyr/foz063.

Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, Boudouris JT, Schneider RM, Langdon QK, Ohkuma M, Endoh R, Takashima M, Manabe RI, Cadez N, Libkind D, Rosa CA, DeVirgilio J, Hulfachor AB, Groenewald M, Kurtzman CP, Hittinger CT, Rokas A. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 2018;**175**: 1533-45 e20;10.1016/j.cell.2018.10.023.

Forster J, Famili I, Fu P, Palsson BO, Nielsen J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* 2003;**13**: 244-53;10.1101/gr.234503.

ter Linde JJ, Liang H, Davis RW, Steensma HY, van Dijken JP, Pronk JT. Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of Saccharomyces cerevisiae. *J Bacteriol* 1999;**181**: 7409-13;10.1128/JB.181.24.7409-7413.1999.

Dujon B. European Functional Analysis Network (EUROFAN) and the functional analysis of the Saccharomyces cerevisiae genome. *Electrophoresis* 1998;**19**: 617-24;10.1002/elps.1150190427.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. *Science* 1996;**274**: 546, 63-7;10.1126/science.274.5287.546.