Original article

The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database

Stacia R. Engel* and J. Michael Cherry

Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA

*Corresponding author: Tel: +650 725 8956; Email: stacia@stanford.edu

Submitted 30 November 2012; Revised 28 January 2013; Accepted 19 February 2013

Citation details: Engel,S.R. and Cherry,J.M. The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces* Genome Database. Database (2013) Vol. 2013: article ID bat012; doi: 10.1093/database/bat012

The first completed eukaryotic genome sequence was that of the yeast *Saccharomyces cerevisiae*, and the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) is the original model organism database. SGD remains the authoritative community resource for the *S. cerevisiae* reference genome sequence and its annotation, and continues to provide comprehensive biological information correlated with *S. cerevisiae* genes and their products. A diverse set of yeast strains have been sequenced to explore commercial and laboratory applications, and a brief history of those strains is provided. The publication of these new genomes has motivated the creation of new tools, and SGD will annotate and provide comparative analyses of these sequences, correlating changes with variations in strain phenotypes and protein function. We are entering a new era at SGD, as we incorporate these new sequences and make them accessible to the scientific community, all in an effort to continue in our mission of educating researchers and facilitating discovery.

Database URL: http://www.yeastgenome.org/

Brief history of yeast genomics

A diverse set of *Saccharomyces cerevisiae* genomes have been sequenced, encompassing a variety of commercial and laboratory strains, as well as wild isolates, many of which have been made available from the *Saccharomyces* Genome Database (SGD). Here we present a description of the isolation and uses of these budding yeast strains, their current incorporation into SGD and our plans for future developments in their annotation and analysis.

The first completed eukaryotic genome sequence was that of the yeast *S. cerevisiae* strain S288C, completed through the effort of a worldwide sequencing consortium (1). S288C has a complex genealogy, but is derived primarily (~88% of its genome) from strain EM93, which was isolated from a rotting fig in Central California in 1938 (2). The remaining 12% of the S288C genome comes from five different progenitors: two natural isolates (EM126 isolated in 1939 also from a rotting fig in Central California, and NRRL YB-210 isolated from rotting bananas from Costa Rica in 1942) and three commercial baking strains (Yeast Foam, FLD and LK). S288C is a widely used laboratory strain, designed by Mortimer for biochemical studies, and specifically selected to be non-flocculent with a minimal set...
of nutritional requirements (2). In the years since the publication of the S288C genome, dozens of yeast genome sequences have been published, laying the groundwork for giant leaps in our understanding of chromosomal evolution and the great plasticity of the eukaryotic genome.

The first new genomes arrived a decade after the consortium completed S288C. In 2005 came RM11-1a, a haploid derivative of Bb32(3), a wild isolate collected from a California vineyard. Published in 2007, YJM789 is the haploid form of an opportunistic pathogen derived from a patient isolated from the lung of an immunocompromised patient in 1989 (3, 4). YJM789 is useful for infection studies and quantitative genetics owing to its divergent phenotype, which includes flocculence, heat tolerance and deadly virulence (4). With the publication of these second and third S. cerevisiae genomes, comparative yeast genomics was born. Researchers began investigating the functional significance of genetic variation on a genomic scale.

Wei et al. (4) demonstrated almost 60,000 single nucleotide polymorphisms (SNPs) and ~6000 insertion/deletions (indels) between YJM789 and S288C, with heterogeneity in polymorphism density along chromosomes, and also within specific genes. An especially dramatic example of sequence changes contributing to an altered lifestyle, in this case, pathogenicity, is PDR5, which encodes an ABC transporter involved in the pleiotropic drug response. Wei et al. (4) also published the first chromosome-by-chromosome sequence comparison for yeast, identifying a large inversion in chromosome XIV, spanning a region just >30 kb long and flanked by transposable elements and tRNAs. Similarly, an inversion between RM11-1a and S288C exists on chromosome III, also bounded by long terminal repeats and tRNA genes. These two new genomes also allowed the first comparisons of genome-wide evolutionary rates, with Gu et al. (5) reporting increased rates of protein evolution in S288C compared with YJM789, and Ronald et al. (6) reporting even faster evolution in RM11-1a relative to the other two strains in pairwise comparisons.

In 2008, three more genomes were published, doubling the number of sequenced yeast genomes from three to six. This is notable because the community had waited 9 years for the completion of the second yeast genome, and another 2 years for the third. Then in the span of just 12 months, the next three new genomes appeared. M22 was collected in an Italian vineyard, whereas YPS163 came from the soil beneath an oak tree in a natural woodland area in southern Pennsylvania in 1999 (7, 8). Not surprisingly, YPS163 is freeze tolerant, a phenotype associated with its increased expression of aquaporin AQY2 (9). AWRI1631 is Australian wine yeast, a robust fermenter and haploid derivative of industrial wine strain N96 (10). Researchers began comparing three different genomes at a time, with similar findings emerging. Doniger et al. (7) reported in excess of 88,000 polymorphisms between the combined genome alignments of M22, YPS163 and S288C. Of these polymorphisms, many of which are strain specific with a decidedly non-random genomic distribution, 93% are SNPs and the remainder are indels. Doniger et al. (7) also confirmed a reciprocal translocation between chromosomes VIII and XVI in vineyard isolate M22 relative to S288C. The specific reciprocal translocation identified is one that is common in wine strains, and produces increased sulfite resistance (11), an intriguing result considering that vineyards are routinely dusted with elemental sulfur as a fungicide. Borneman et al. (10) saw a mosaic pattern of differences when comparing AWRI1631 with both YJM789 and S288C, such that while substantial conservation exists throughout much of the genome, many regions exhibit high degrees of interstrain variation. Furthermore, Borneman et al. (10) also reported a reciprocal translocation, this time in YJM789 as compared with S288C and AWRI1631, between chromosomes VI and X, as well as a large inversion in chromosome XIV in YJM789.

By the end of 2009, sequences of entire yeast genomes were being published one after another. JAY291 is a non-flocculent haploid derivative of Brazilian bioethanol strain PE-2; it produces high levels of ethanol and cell mass, and is tolerant to heat and oxidative stress (12). Argueso et al. (12) determined that JAY291 is highly divergent to S288C, RM11-1a and YJM789, and contains well-characterized alleles at several genes of known relation to thermostolerance and fermentation performance. EC1118, a diploid commercial yeast, is probably the most widely used wine-making strain worldwide based on volume produced. In the Northern hemisphere, it is also known as Premier Cuvee or Prise de Mousse; it is a reliably aggressive fermenter, and makes clean but somewhat uninteresting wines. Novo et al. (13) found EC1118 more diverged from S288C and YJM789 than from RM11-1a and AWRI1631, and also reported three unique regions from 17 to 65 kb in size in the EC1118 genome on chromosomes VI, XIV and XV, encompassing 34 genes related to key fermentation characteristics, such as metabolism and transport of sugar or nitrogen. They also identified >100 genes present in S288C that are missing from EC1118. The release of the Sigma1278b genome was notable, as it was the second widely used laboratory strain to be sequenced, but also because of the concurrent production of a systematic deletion collection in this strain background. Dowell et al. (14) reported 75 genes in Sigma1278b that are absent from S288C, as well as sets of ‘conditional essentials’, which are genes required for viability in one background but not the other, demonstrating decisively that phenotypes are influenced by background-specific modifiers.

The following year, five new S. cerevisiae genomes became available (15). Foster’s O and Foster’s B are commercial ale yeasts. VIN13 is a cold-tolerant South African
wine strain, a strong fermenter that is good for making aromatic white wines. AWRI796 is another South African wine strain, but ferments more successfully at warmer temperatures and is more suited to the production of reds. CLIB215 was isolated in 1994 from a bakery in Taranaki in the North Island of New Zealand. Borneman et al. (15) identified different large chromosomal copy number variations (CNV) in the various industrial strains. Some genomes appear to have whole-chromosome amplifications: chromosome I in AWRI796, chromosome III in Foster’s O and chromosomes II, V and XV in Foster’s B. Several partial chromosomal CNV amplifications hundreds of kilobases long were also identified, as were some reductions in copy number. Borneman et al. (15) also reported dozens of novel open reading frames (ORFs) in each strain, some of which are shared between strains, for a total of 218 in this non-degenerate set of ORFs that are not present in the S288C reference genome (Table 1).

In 2011, the most prolific year, the number of available genomes doubled from 14 to 29. CBS7960 was isolated from a cane sugar ethanol factory in Sao Paulo, Brazil. PW5 came from fermented sap of a Raphia palm tree in Nigeria in 2002. CLIB324 is a Vietnamese baker’s strain collected in 1996 from Ho Chi Minh City. CLIB382 came from beer brewed in Ireland sometime before 1952. EC9-8 is a haploid cadmium-resistant derivative of a yeast isolated from the valley bottom of Evolution Canyon at Lower Nahal Oren, Israel (18). T7 was isolated from oak tree exudate in Missouri’s Babler State Park. T73 is from a Mourvedre (aka Monastrell) red wine made in Alicante, Spain, in 1987. T73 has low nitrogen requirements, high alcohol tolerance and low volatile acidity production, making it ideal for fermenting robust structured reds grown in hot climates. UC5 came from Sene sake in Kurashi, Japan, sometime before 1974. VL3 was isolated in Bordeaux, France, and is most suited to the production of premium aromatic white wines with high thiol content (citrus and tropical fruit characters). Borneman et al. (15) reported a whole-chromosome amplification of chromosome VIII in VL3, as well as >50 ORFs in VL3 that are missing from S288C (Table 1). Kyokai No. 7 (K7) is the most extensively used sake yeast, and was first isolated from a sake brewery in Nagano Prefecture, Japan, in 1946 (17). Akao et al. (17) reported two large inversions in K7 on chromosomes V and XIV, both flanked by transposable elements and inverted repeats, two CNV reductions on chromosomes I and VII and a similar mosaic-like pattern and non-random distribution of variation compared with S288C as seen by other researchers in other strains. They also identified 48 ORFs in K7 that are absent in S288C, and 49 ORFs in S288C that are missing from K7 (Table 1). Also in 2011 came the genome of QA23, a cold-tolerant Portuguese wine strain from the Vinho Verde region. QA23 has low nutrient and oxygen requirements, and exhibits high β-glucosidase activity, a combination that makes beautiful Sauvignon blancs. Y10 was isolated from a coconut in the Philippines, sometime before 1973. YJM269 came from red Blauer Portugieser grapes in Austria in 1954. FL100 is the third laboratory strain to be sequenced, and very soon thereafter followed W303. Ralser et al. (19) reported that the W303-derivative K6001, a key model organism for research into aging, shares >85% of its genome with S288C, differing at >8000 nucleotide positions, causing changes to the sequences of 799 proteins. These differences are distributed non-randomly throughout the genome, with chromosome XVI being almost identical between the two strains, and chromosome XI the most divergent. Ralser et al. (19) also noted that some of the non-S288C regions in W303 are also present in Sigma1278b, which exhibits six times the rate of sequence divergence to S288C as seen in W303, and which is identical to S288C at less than half its genome.

In 2012, genome sequences for an additional four strains became available, such that by now, dozens of genomes have been published, from yeasts with all different kinds of jobs and lifestyles (Figure 1; Table 2). BY4741 and BY4742 are the S288C-derivative strains used for the systematic deletion collection, and variation between these strains and S288C is miniscule (T. Yamaguchi and F. Roth, personal communication). ZTW1 was isolated from corn mash used for industrial bioethanol production in China in 2007. CEN.PK113-7D is a laboratory strain derived from parental strains ENY.WA-1A and MC96A, and is popular for use in systems biology studies. Nijkamp et al. (16) found six duplicated regions in CEN.PK113-7D relative to S288C, two on chromosome II, and one each on chromosomes III, VII, VIII and XV, including an enrichment of maltose

Table 1. Various S. cerevisiae genomes contain ORFs that are not present in the S288C reference genome

| Strain         | ORFs not in S288C | Reference |
|----------------|-------------------|-----------|
| AWRI796        | 74                | 15        |
| CEN.PK113-7D   | 83                | 16        |
| EC1118         | 77                | 15        |
| FostersB       | 36                | 15        |
| FostersO       | 48                | 15        |
| JAY291         | 16                | 12        |
| Kyokai No.7    | 48                | 17        |
| QA23           | 110               | 15        |
| RM11-1a        | 38                | 15        |
| Sigma1278b     | 75                | 14        |
| VIN13          | 45                | 15        |
| VL3            | 54                | 15        |
| YJM789         | 34                | 15        |
metabolism genes. Also present in CEN.PK113-7D, which Nijkamp and coworkers found to be a biotin prototroph, are genes required for biotin biosynthesis. They also identified >20,000 SNPs between the two strains, two-thirds of which are within ORFs. Almost 5000 of these result in altered sequences of >1400 proteins. Nijkamp et al. (16) also reported >2800 small indels averaging 3 bp each, and more than 400 of these were found in coding regions. An additional 83 genes were identified that are absent from S288C, including the ENA6 sodium pump that is also found in YJM269, and others that are present in both YJM269 and PW5. Nijkamp et al. (16) also presented a phylogenetic analysis of whole-genomic distances of the strains mentioned above (Figure 1; Table 2).

Next-generation sequencing methods have, by this time, become so mainstream that whole genomes are now being analyzed en masse to answer specific questions. Wenger et al. (20) used high-throughput sequencing in conjunction with bulk segregant analysis to investigate the distribution of the ability to ferment xylose among 600+ strains of S. cerevisiae, and found that this ‘xylose-positive’ phenotype, which was present in ~5% of the tested strains, clustered within wine yeasts. They further determined the presence of a novel xylitol dehydrogenase gene XDH1 in the Simi White strain, in the same sub-telomeric 65-kb insert on chromosome XV that Novo et al. (13) had previously identified in wine industry workhorse EC1118. Note that while Simi White and EC1118 share the same large insertion, the xylose utilization locus itself in EC1118 is pseudogenic (13). Libkind et al. (21) combined comparative genomics with population ecology of >2000 natural isolates to resolve questions of taxonomy and systematics, ultimately identifying S. cerevisiae and the novel cryotolerant species Saccharomyces eubayanus as progenitors of Saccharomyces pastorianus, shedding light on the evolution and domestication of lager yeasts. At the same time, Nguyen et al. (22) were also using comparative genomics to study the hybridization history of lager Saccharomyces, finding mosaic genomes and patterns of introgression between Saccharomyces bayanus, Saccharomyces uvarum and S. cerevisiae. The same novel species named S. eubayanus by Libkind and coworkers (21) was identified by Nguyen et al. (22) and called Saccharomyces lagerae. Borneman et al. (23) studied the wine yeast VIN7, which is widely used in cool-temperature fermentations to produce premium Sauvignon blancs and Semillons, and confirmed that its genome is a complex allotriploid of S. cerevisiae and cryotolerant Saccharomyces kudriavzevii. VIN7 most likely arose through a mating between a diploid S. cerevisiae and a haploid S. kudriavzevii, and exhibits evidence of translocation and recombination events occurring between alleles of both progenitors. Erny et al. (24) performed genomic analyses of Alsatan industrial wine yeast Eg8, which tolerates cool temperatures and elevated alcohol concentrations, and is ideal for fermentation of Semillons and Muscats. Erny et al. (24) found that Eg8 is also a chimeric allotriploid hybrid between a diploid S. cerevisiae and haploid S. kudriavzevii, and further identified the same translocation between chromosomes VIII and XVI that Doniger et al. (8) reported in vineyard isolate M22 that leads to increased sulfite resistance, and which is common in wine yeast strains. Peris et al. (25) investigated genomic compositions of various S. cerevisiae × S. kudriavzevii natural hybrids isolated from wine and beer fermentations, including VIN7. They found different chromosome complements and rearrangements in the different yeasts, although all shared a common set of S. kudriavzevii genes and lacked a common set of S. cerevisiae genes. The rich

---

Figure 1. Phylogram depicting relationships among S. cerevisiae strains based on Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering of whole-genomic distances as calculated by Nijkamp et al. (16). Redrawn from Nijkamp et al. (16).
content of genomic sequences available to serve as a background against which to compare has changed the way we study genomes.

**New directions**

While next-generation sequencing has been taking the yeast genetics community by storm, SGD has been preparing for this shift into the new modern era of yeast genomics. In February 2011, SGD put into place an updated reference sequence of increased quality based on these modern sequencing technologies, and henceforth, we anticipate very few sequence updates for S288C (Engel et al., in preparation). We will continue to provide the definitive reference genome sequence for *S. cerevisiae* as well as variant sequences from these other sequenced strains, and are moving increasingly toward the representation of sequence variation and allelic differences. We have already

| Strain     | Year | Provenance                                      | NCBI BioProject | Contig N50a | Scaffold N50a |
|------------|------|-------------------------------------------------|-----------------|-------------|---------------|
| S288C      | 1996 | Laboratory strain                               | PRJNA128        | N/A         | N/A           |
| RM11-1a    | 2005 | Haploid derivative of California vineyard isolate | PRJNA13674      | 263 288     | 795 018       |
| YJM789     | 2007 | Haploid derivative of opportunistic human pathogen | PRJNA13304      | 429 709     | N/A           |
| M22        | 2008 | Italian vineyard isolate                        | PRJNA28815      | 2207        | N/A           |
| YPS163     | 2008 | Pennsylvania woodland isolate                   | PRJNA28813      | 2901        | N/A           |
| AWRI1631   | 2008 | Haploid derivative of South African commercial wine strain N96 | PRJNA30553      | 7704        | N/A           |
| JAY291     | 2009 | Haploid derivative of Brazilian industrial bioethanol strain PE-2 | PRJNA32809      | 64 336      | N/A           |
| EC1118     | 2009 | Commercial wine strain                          | PRJEA37863      | 776 014     | N/A           |
| Sigma1278b | 2009 | Laboratory strain                               | PRJNA39317      | 365 700     | N/A           |
| Foster’s O | 2010 | Commercial ale strain                           | PRJNA48567      | 195 316     | N/A           |
| Foster’s B | 2010 | Commercial ale strain                           | PRJNA48569      | 204 208     | 626 897       |
| VIN13      | 2010 | South African white wine strain                 | PRJNA48563      | 308 189     | 700 638       |
| AWRI796    | 2010 | South African red wine strain                   | PRJNA48559      | 403 341     | 565 854       |
| CLIB215    | 2010 | New Zealand bakery isolate                      | PRJNA60143      | 16 813      | 47 217        |
| CBS7960    | 2011 | Brazilian bioethanol factory isolate            | PRJNA60391      | 18 761      | 65 099        |
| CLIB324    | 2011 | Vietnamese bakery isolate                       | PRJNA60415      | 4260        | 24 472        |
| CLIB382    | 2011 | Irish beer isolate                              | PRJNA60145      | 840         | 2711          |
| E9-8       | 2011 | Haploid derivative of Israeli canyon isolate    | PRJNA73985      | 15 539      | 541 605       |
| FL100      | 2011 | Laboratory strain                               | PRJNA60147      | 4244        | 26 506        |
| Kyokai No.7| 2011 | Japanese sake yeast                             | PRJNA45827      | 120 978     | 902 266       |
| QA23       | 2011 | Portuguese Vinho Verde white wine strain        | PRJNA48561      | 182 942     | 182 942       |
| PW5        | 2011 | Nigerian Raphia palm wine isolate               | PRJNA60181      | 14 234      | 393 105       |
| T7         | 2011 | Missouri oak tree exudate isolate               | PRJNA60387      | 147 205     | 476 142       |
| T73        | 2011 | Spanish red wine strain                         | PRJNA60195      | 2945        | 36 287        |
| UC5        | 2011 | Japanese sake yeast                             | PRJNA60197      | 17 142      | 356 094       |
| VL3        | 2011 | French white wine strain                        | PRJNA48565      | 293 399     | 656 188       |
| W303       | 2011 | Laboratory strain                               | PRJNA167645     | 149 943     | 367 966       |
| Y10        | 2011 | Philippine coconut isolate                      | PRJNA60201      | 2730        | 22 204        |
| YJM269     | 2011 | Austrian Blauer Portugieser wine grapes          | PRJNA60389      | 23 452      | 58 353        |
| BY4741     | 2012 | S288C-derivative laboratory strain              | N/A             | N/A         | N/A           |
| BY4742     | 2012 | S288C-derivative laboratory strain              | N/A             | N/A         | N/A           |
| CEN.PK 113-7D | 2012 | Laboratory strain                             | PRJNA52955      | 48 196      | 918 791       |
| ZTW1       | 2012 | Chinese corn mash bioethanol isolate            | PRJNA174065     | 556 921     | N/A           |

aContig and scaffold N50a lengths are common genome statistics that indicate the minimum length in the set of individual contiguous sequences (contigs or scaffolds), which contain half of all bases in the assembly.

bN/A = not available.
incorporated the new \textit{S. cerevisiae} genomes mentioned above into SGD and will continue to expand the amount of information available to researchers for the growing number of laboratory and industrial strains and wild isolates. New tools are being developed that will provide access to this compendium of allelic and variation information and allow a newly determined sequence to be compared with the reference strain, as well as with the sequences of several widely used and commonly studied \textit{S. cerevisiae} strains. Current tools in place include precomputed protein and coding DNA alignments (ClustalW) for each ORF, as well as ORF-specific dendrograms, which depict the degree of similarity of that ORF sequence among the set of strains in which it was identified (Figure 2). From each Locus Summary page, the protein and DNA sequences are accessible via a pair of pull-down menus in the Sequence Information section, while the alignments can be reached via links in the Analyze Sequence section, or through the ‘Strains and species’ item in the Sequence menu at the top of most SGD web pages. Protein or DNA sequences can also be downloaded in batch from the alignment pages, or one-by-one from each Locus Summary. We also have the genomes of the various strains incorporated into the Basic Local Alignment Search Tool (BLAST) datasets, available for searching against genomic and coding DNA, as well as protein sequences. BLAST can be useful for finding evidence of fission/fusion events in which ORFs, such as YNR066C, are split in some strains but not others (Figure 3). The BLAST tool is accessible via the Sequence menu at the top of most SGD pages. All the strain DNA and protein sequences are available for download so that researchers can perform their own analyses (http://downloads.yeastgenome.org). Furthermore, we continue to associate information regarding sequence variation with functional effects and phenotypic variations. SGD has been curating phenotypes in the different strain backgrounds for several years, as genes shared across strains and species can produce different phenotypes, revealing genetic variation and possibly uncovering new models of disease (26).
New technologies and approaches are pushing *S. cerevisiae* annotation past the limits of a system based exclusively on a single reference sequence. Next-generation sequencing methods will determine the genomic sequences of hundreds, if not thousands, of different *S. cerevisiae* industrial strains, laboratory strains and natural isolates in the coming years. Comparative genomics can provide a clearer picture of the full constituent parts of a species’ genome and provide for the identification of sequence features such as binding motifs, regulatory regions and non-coding RNAs. As described above, these new genomes vary not only at specific nucleotides, but also in the complement of genes they carry and the architecture of their chromosomes, as genomic elements can be shuffled, amplified, lost or gained as populations adapt to different environments. To provide a more comprehensive view of the genetic repertoire of yeast, SGD is compiling the virtual *S. cerevisiae* genome, or pan-genome, that will comprise all genes found within the various sequenced *S. cerevisiae* strains.

A pan-genome more accurately describes the genetic content of a species, and can be much larger than any single constituent genome. Each gene can be binned into one of three categories. Core genes are those present in every genome, and include conserved essential genes for proteins such as actin, or polymerases, histones and ribosomal constituents required for some of the most basic cellular processes such as replication and translation. Frequent genes are those found in some genomes but not others; they are commonly involved in adaptation to specific environments or applications, such as metabolism of specific sugars or fermentation of specific carbon sources. In bacterial genomics, this intermediate class goes by various names: ‘character’, ‘dispensable’, ‘peripheral’, ‘variable’ or ‘flexible’ genes (27–30). They tend to evolve more quickly than the conserved essential genes, but more slowly than the individual genomes themselves. The *S. cerevisiae* pan-genome contains hundreds of frequent genes that are found in some strains but not others. Examples include the MAL (maltose fermentation) family of multigene loci, each of which encodes a maltose permease, a maltase and a trans-acting MAL activator (31). As mentioned earlier, Nijkamp et al. (16) found the genome of strain CEN.PK113-7D to be enriched in the MAL genes. Rare genes are those that are present in only a small number of genomes, possibly even unique to a single strain, and often are of unknown function. Rare genes tend to be rapidly evolving and especially mutable, exhibiting high rates
of gene birth and death. In bacterial genomics, these genes are sometimes called ‘accessory’ genes (27). A recently reported rare gene in S. cerevisiae is the novel XDH1 xylose utilization gene mentioned earlier (20). Other examples include PRM8 and PRM9, both of which encode non-essential pheromone-regulated transmembrane proteins of the DUP240 family (32). These three sets together—core, frequent and rare—make up the pan-genome that we want to describe, and will in the future provide a valuable resource for the annotation of newly determined budding yeast genomes and for the functional analysis and comparison of observed variation within S. cerevisiae.

The availability of an ever-increasing number of sequenced genomes presents a growing list of clear and present challenges that all genome databases will have to address: How will any particular approach scale up to handling hundreds of genomes? What is the best way to organize and display SNPs, larger polymorphisms and genome rearrangements? How should chromosomal coordinates and mapping information be dealt with in the context of a pan-genome? At SGD, we are expanding our scope to provide annotation and comparative analyses of all major budding yeast strains, and are moving toward providing multiple reference genomes. We are not abandoning a standard sequence, but instead determining how far one can get from a reference while still maintaining utility. It is helpful to be able to ‘shift the reference’, selecting the genome that is most appropriate and informative for a specific area of study. SGD has actively sought and obtained genome sequences for a set of strains with a substantial history of use and experimental results that will serve as reference genomes. These strains include W303, Sigma1278b, SK1, SEY6210, CEN.PK, JK9-3d and FL100, and are the genomes for which we have the most curated phenotype data, and for which we aim to curate specific functional information. High-quality genome sequences combined with detailed phenotypes and functional annotations will allow dissection of the genomic bases of phenotypic variation.

The meticulous investigation of the complexes and interactions of individual gene products in the yeast cell allows great things to be done in yeast genomics. There is no other model organism that provides such a fertile environment for understanding the basic mechanisms of biology. There is a continued need for this ‘small science’ investigating the biochemistry and cell biology of eukaryotic cells (33). SGD bridges the experimentally defined knowledge provided from investigations on the small scale over to its application during the annotation of genomic results. The power of yeast genomics is resting squarely on the shoulders of yeast genetics and biochemistry. SGD has a long history of service to yeast researchers and to the broader genetics community as a whole. As we all move through this new era of comparative yeast genomics, SGD maintains its high level of dedication to quality and remains the primary annotation resource for new strains of S. cerevisiae. We continue in our mission of educating students, enabling bench researchers and facilitating scientific discovery.

Acknowledgements

The authors thank Craig Amundsen, Shuai Wang and Benjamin C. Hitz for incorporation and analyses of strain genomes; Markus Ralser for early access to the W303 genome sequence; Takafumi Yamaguchi and Frederick P. (Fritz) Roth for early access to the BY4741 and BY4742 genome sequences; Gail Binkley and Edith D. Wong for reading the manuscript; and the SGD staff, without whose effort and dedication the SGD Project would not be possible.

Funding

SGD is funded by the National Human Genome Research Institute (NHGRI), US National Institutes of Health [5U41HG001315-18]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health.

Conflict of interest. None declared.

References

1. Goffeau,A., Barrell,B.G., Bussey,H. et al. (1996) Life with 6000 genes. Science, 274, 546–567.
2. Mortimer,R.K. and Johnston,J.R. (1986) Genealogy of principal strains of the yeast genetic stock center. Genetics, 113, 35–43.
3. Tawfik,O.W., Papasian,C.J., Dixon,A.Y. et al. (1989) Saccharomyces cerevisiae pneumonia in a patient with Acquired Immune Deficiency Syndrome. J. Clin. Microbiol., 27, 1689–1691.
4. Wei,W., McCusker,J.H., Hyman,R.W. et al. (2007) Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proc. Natl Acad. Sci. U S A, 104, 12825–12830.
5. Gu,Z., David,L., Petrov,D. et al. (2005) Elevated evolutionary rates in the laboratory strain of Saccharomyces cerevisiae. Proc. Natl Acad. Sci. U S A, 102, 1092–1097.
6. Ronald,J., Tang,H. and Brem,R. (2006) Genomewide evolutionary rates in laboratory and wild yeast. Genetics, 174, 541–544.
7. Doniger,S.W., Kim,H.S., Swain,D. et al. (2008) A catalog of neutral and deleterious polymorphism in yeast. PLoS Genet., 4, e1000183.
8. Sniegowski,P.D., Dombrowski,P.G. and Fingerman,E. (2002) Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. FEMS Yeast Res., 1, 299–306.
9. Fay,J.C., McCullough,H.L., Sniegowski,P.D. et al. (2004) Population genetic variation in gene expression associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol., 5, R26.
10. Borneman, A.R., Forgan, A.H., Pretorius, I.S. et al. (2008) Comparative genome analysis of a Saccharomyces wine strain. FEMS Yeast Res., 8, 1185–1195.
11. Perez-Ortin, J.E., Quero, A., Puig, S. et al. (2002) Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Res., 12, 1533–1539.
12. Argueso, J.L., Carrozziolle, M.F., Mieczkowski, P.A. et al. (2009) Genome structure of a Saccharomyces cerevisiae strain widely used in bioethanol production. Genome Res., 19, 2258–2270.
13. Novo, M., Bigey, F., Beyne, E. et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. Proc. Natl Acad. Sci. U S A, 106, 16333–16338.
14. Dowell, R.D., Ryan, O., Jansen, A. et al. (2010) Genotype to phenotype: a complex problem. Science, 328, 469.
15. Borneman, A.R., Desany, B.A., Riches, D. et al. (2011) Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae. PLoS Genet., 7, e1001287.
16. Nijkamp, J.F., van den Broek, M., Datema, E. et al. (2012) De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology. Microb. Cell Fact., 11, 36.
17. Akao, T., Yashiro, I., Hosoyama, A. et al. (2011) Whole-genome sequencing of sake yeast Saccharomyces cerevisiae Kyokai no 7. DNA Res., 18, 423–434.
18. Chang, S.L. and Leu, J.Y. (2011) A tradeoff drives the evolution of reduced metal resistance in natural populations of yeast. PLoS Genet., 7, e1002034.
19. Ralser, M., Kuhl, H., Ralser, M. et al. (2012) The Saccharomyces cerevisiae W303-K6001 cross-platform genome sequence: insights into ancestry and physiology of a laboratory mutt. Open Biol., 2, 120093.
20. Wenger, J.W., Schwartz, K. and Sherlock, G. (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. PLoS Genet., 6, e1000942.
21. Libkind, D., Hittinger, C.T., Valero, E. et al. (2011) Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. Proc. Natl Acad. Sci. U S A, 108, 14539–14544.
22. Nguyen, H.V., Legras, J.L., Neuvoglise, C. et al. (2011) Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of Saccharomyces bayanus strains NBRC1948 and CBS380T. PLoS One, 6, e25821.
23. Borneman, A.R., Desany, B.A., Riches, D. et al. (2012) The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with Saccharomyces cerevisiae and Saccharomyces kudriavzevii origins. FEMS Yeast Res., 12, 88–96.
24. Ems, C., Raoul, P., Alais, A. et al. (2012) Ecological success of a group of Saccharomyces cerevisiae x Saccharomyces kudriavzevii hybrids in the Northern European wine-making environment. Appl. Environ. Microbiol., 78, 3256–3265.
25. Peris, D., Lopes, C.A., Belloch, C. et al. (2012) Comparative genomics among Saccharomyces cerevisiae x Saccharomyces kudriavzevii natural hybrid strains isolated from wine and beer reveals different origins. BMC Genomics, 13, 407.
26. Engel, S.R., Balakrishnan, R., Binkley, G. et al. (2010) Saccharomyces Genome Database provides mutant phenotype data. Nucleic Acids Res., 38, D433–D436.
27. Bentley, S. (2009) Sequencing the species pan-genome. Nat. Rev., 7, 258–259.
28. Conlan, S., Mijares, L.A., Becker, J. et al. (2012) Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. Genome Biol., 13, R64.
29. Liang, W., Zhao, Y., Chen, C. et al. (2012) Pan-genomic analysis provides insights into the genomic variation and evolution of Salmonella Paratyphi A. PLoS One, 9, e45346.
30. Psomopoulos, F.E., Siarkou, V.I., Papanikolaou, N. et al. (2012) The Chlamydiales pan-genome revisited: structural stability and functional coherence. Genes, 3, 291–319.
31. Chow, T.H.C., Sollitti, P. and Marmur, J. (1989) Structure of the multigene family of MAL loci in Saccharomyces. Mol. Gen. Genet., 217, 60–69.
32. Heiman, M.G. and Walter, P. (2000) Prm1p, a pheromone-regulated multispansing membrane protein, facilitates plasma membrane fusion during yeast mating. J. Cell Biol., 151, 719–30.
33. Alberts, B. (2012) The end of “small science”? Science, 337, 1583.