NIH Public Access
Author Manuscript
*Yeast*. Author manuscript; available in PMC 2011 February 16.

Published in final edited form as:
*Yeast*. 2006 September ; 23(12): 857–865. doi:10.1002/yea.1400.

# *Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis

**Dianna G. Fisk**[1], **Catherine A. Ball**[2], **Kara Dolinski**[3], **Stacia R. Engel**[1], **Eurie L. Hong**[1], **Laurie Issel-Tarver**[4], **Katja Schwartz**[1], **Anand Sethuraman**[1], **David Botstein**[3], **J. Michael Cherry**[1,*], and **The *Saccharomyces* Genome Database Project**

[1]Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA

[2]Department of Biochemistry, School of Medicine, Stanford University, Stanford, CA 94305-5307, USA

[3]Lewis-Sigler Institute for Integrative Genomics, Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

[4]Ohlone College, Biology Department, Fremont, CA 94539, USA

## Abstract

The *S. cerevisiae* genome is the most well-characterized eukaryotic genome and one of the simplest in terms of identifying open reading frames (ORFs), yet its primary annotation has been updated continually in the decade since its initial release in 1996 (Goffeau *et al*., 1996). The *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) (Hirschman *et al*., 2006), the community-designated repository for this reference genome, strives to ensure that the *S. cerevisiae* annotation is as accurate and useful as possible. At SGD, the *S. cerevisiae* genome sequence and annotation are treated as a working hypothesis, which must be repeatedly tested and refined. In this paper, in celebration of the tenth anniversary of the completion of the *S. cerevisiae* genome sequence, we discuss the ways in which the *S. cerevisiae* sequence and annotation have changed, consider the multiple sources of experimental and comparative data on which these changes are based, and describe our methods for evaluating, incorporating and documenting these new data.

### Keywords

*S. cerevisiae*; genome sequence; genome annotation; comparative genomics; exon/intron boundaries

## Introduction

In the original *S. cerevisiae* genomic annotation (c. 1993–1996), protein encoding genes were simply annotated as the longest possible open reading frame of 100 or more codons.

Copyright © 2006 John Wiley & Sons, Ltd.

*Correspondence to: J. Michael Cherry, Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA, cherry@stanford.edu.

**DNA accession numbers**

The *S. cerevisiae* S288C genome sequence and annotations are maintained by SGD and archived at NCBI within the Reference Sequence (RefSeq) collection. The Accession Nos for the 16 nuclear chromosomes and the mitochondrial genome are: NC_001133, NC_001134, NC_001135, NC_001136, NC_001137, NC_001138, NC_001139, NC_001140, NC_001141, NC_001142, NC_001143, NC_001144, NC_001145, NC_001146, NC_001147, NC_001148, and NC_001224.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

These annotations have now been subjected to a decade of testing by thousands of scientists worldwide, using a large range of experimental and comparative methods. In particular, the genome-wide comparisons published by Brachat *et al*. (2003), Cliften *et al*. (2003), and Kellis *et al*. (2003) provided an excellent opportunity to review the entire *S. cerevisiae* gene model, both in sequence and interpretation. In these studies, the sequenced species were so closely related to *S. cerevisiae* as to allow the expectation of very close conservation of ORF size, location and intron/exon structure. Not surprisingly, there have been many suggested changes: new ORFs have been identified, and existing ORFs have been 'removed' and revised (Figure 1).

Most newly identified ORFs have been smaller than 100 codons. This is simply due to the fact that the *S. cerevisiae* genome sequencing project did not annotate ORFs of fewer than 100 codons that did not have significant sequence similarity to a previously identified gene. This approach was necessary because there is a high probability that ORFs of this size are just fortuitous sequences of nucleotides: only 342 (2%) of the 15 000 ORFs in the genome between 50 and 99 codons in length are currently thought to encode proteins within the yeast cell. As a consequence, any ORF under 100 codons is treated as spurious until proved otherwise through either experimental or comparative work.

However, length alone does not guarantee that an ORF is genuine, and the total number of biologically significant *S. cerevisiae* ORFs has been the subject of debate since the completion of the genomic sequence (Termier and Kalogeropoulos, 1996; Zhang and Wang, 2000; Malpertuy *et al*., 2000; Wood *et al*., 2001; Mackiewicz *et al*., 2002; Brachat *et al*., 2003; Cliften *et al*., 2003; Kellis *et al*., 2003). At the heart of this debate is the basic principle that it is virtually impossible to demonstrate experimentally that an ORF is nonfunctional; there is always a chance that a suspect ORF encodes a protein of extremely low abundance or that is produced only under some specific environmental condition. Fortunately, the availability of genomic sequences from other fungi provides a positive test for the relevance of experimentally uncharacterized ORFs: evolutionary conservation among very closely related species. This has allowed for a separation of significant ORFs from those that are likely to be spurious.

Even many bona fide ORFs have required updating. Revisions of ORF annotation fall into two major categories: those in which the nucleotide sequence is corrected; and those in which the nucleotide sequence remains the same but its interpretation is altered. Changes in the first category often affect the start codon, stop codon, reading frame or coding sequence for that ORF, while changes in the second category include annotation of different start codons and intron/exon structure.

Although automated data processing is an important element in the process of revising and updating genomic sequence annotation, human evaluation is also essential. In making any changes to the genome sequence, SGD curators evaluate and synthesize all available types of evidence, including that generated by individual gene-specific experiments, by large-scale analyses and by cross-species comparisons.

Because SGD strives to provide rapid access to new information, individual updates are integrated into the genome sequence and released to the community as soon as possible. As a result, genome updates have been made gradually and released continually, rather than as rare scheduled updates encompassing multiple changes. While this approach provides the fastest means of disseminating the updates, alerting the research community to the changes has proven to be a continuing challenge. Here, we describe the types of changes that have been incorporated into the *S. cerevisiae* genome annotation, how SGD handles each type of change and how the research community can access the updated information.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

## Results and discussion

### New ORFs

Over the last decade, 522 new ORFs have been added to the *S. cerevisiae* gene catalogue. Prior to the year 2001, most new small ORFs were discovered individually during the course of focused experimental research. These ORFs were annotated because they encoded proteins that were isolated from complexes (e.g. TIM9/YEL020W-A; Koehler *et al*., 2000), discovered in traditional genetic screens (e.g. SAE3/YHRO79C-A; McKee and Kleckner, 1997) or identified in focused comparative analyses (e.g. YAL044W-A; Valerie Wood, personal communication). More recently, researchers have applied large-scale approaches, both computational and experimental, to the problem of finding the biologically significant small ORFs (Basrai *et al*., 1997; Blandin *et al*., 2000; Kumar *et al*., 2002; Oshiro *et al*., 2002; Brachat *et al*., 2003; Cliften *et al*., 2003; Kessler *et al*., 2003). These large-scale studies produced 65% of the new additions to the *S. cerevisiae* ORF catalogue.

SGD curators examined each proposed new ORF to insure its validity as a potential gene. In most instances, the new ORF was accepted as proposed, but some cases required more extensive analysis. For example, several of the new ORFs proposed by Blandin *et al*. (2000), Brachat *et al*. (2003) and Cliften *et al*. (2003) contained introns; while these three groups often predicted new intron-containing ORFs in the same regions, they sometimes differed on the exact location of the exon/intron boundaries. These conflicts were resolved by examining the *sensu stricto Saccharomyces* data published by Kellis *et al*. (2003) and determining which proposed exon/intron structure was conserved in other closely related species. In a few other cases, the new ‘ORFs’ were subsequently shown to be part of previously annotated ORFs rather than independent new ORFs.

### Classification of open reading frames

The ascomycete species sequenced by Brachat *et al*. (2003), Cliften *et al*. (2003) and Kellis *et al*. (2003) largely contain the same ORFs as does *S. cerevisiae*, in the same order. Thus, lack of conservation in the closely related species constitutes evidence against the biological significance of an *S. cerevisiae* ORF. All three of these groups applied this test independently, using their own datasets, and generated three partially overlapping lists of potentially spurious ORFs. Brachat *et al*. (2003), Cliften *et al*. (2003) and Kellis *et al*. (2003) recommended that 368, 496 and 515 ORFs, respectively, be deleted.

Because even sophisticated computation is no substitute for actual laboratory experiments, SGD takes a cautious approach towards the removal of ORFs from the *S. cerevisiae* genomic catalogue. ORFs recommended for deletion are not actually eliminated from the genome annotation, but are simply labelled ‘dubious’. This approach results in an *S. cerevisiae* gene model of relatively high certainty, while still allowing further testing on the set of questionable, ‘dubious’ ORFs. The ‘dubious’ designation is prominently displayed on Locus Summary pages and is indicated by colour on graphical displays of chromosome maps. Dubious ORFs are also excluded from sets of ORFs considered biologically significant; they are not included in the comprehensive file of *S. cerevisiae* Gene Ontology annotations (**gene_association.sgd**) that SGD provides to the public, and they are not included in the *S. cerevisiae* reference sequence (RefSeq) entries that SGD maintains and provides to NCBI.

During the initial analysis, individual ORFs were designated ‘dubious’ if they met the following criteria: (a) the ORF was identified as potentially spurious by at least one of the comparative studies above; (b) there were no well-controlled, small-scale, published experiments demonstrating that detectable mRNA or protein was produced from this ORF; (c) any mutant phenotype described for the ORF could be ascribed to mutation of an

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

overlapping gene; and (d) the ORF did not contain an intron. The last condition was necessary because none of the three groups annotated introns in the related fungal species, and comparison of 'spliced' *S. cerevisiae* ORFs with exon fragments in other species could result in the artificial appearance of non-conservation. The majority of the ORFs identified as spurious by Brachat *et al*. (2003), Cliften *et al*. (2003) and Kellis *et al*. (2003) met these four criteria and were assigned a 'dubious' designation by SGD. For a small number of ORFs in this group, SGD curators found evidence suggesting that they represented functional genes. For example, all three groups recommended that *AUA1* /YFL010W-A is not a protein-encoding ORF because it is not conserved, and has substantial overlap with a characterized gene, *WWM1* /YFL010C. However, the transcription and mutant phenotype of *AUA1* have been characterized (Sophianopoulou and Diallinas, 1993) and were not easily attributed to *WWM1*.

At the same time that SGD began labelling spurious ORFs 'dubious', we also implemented a further classification of conserved ORFs, according to the certainty that they actually encode proteins. ORFs that contained an intron, or that were identified as conserved by all three of the large-scale comparative studies, were designated either 'uncharacterized' or 'verified', depending on available experimental evidence. Because the *S. cerevisiae* nomenclature system allows yeast ORFs to be assigned a genetic name only after being described in a publication, named ORFs were automatically classified as 'verified'. Unnamed ORFs were designated 'uncharacterized' unless there were published data supporting a 'verified' classification, such as mRNA or protein detection, or a mutant phenotype not ascribable to an overlapping gene.

Unfortunately, the comparative analyses done by Brachat *et al*. (2003), Cliften *et al*. (2003) and Kellis *et al*. (2003) were concurrent with many of the other large-scale analyses that identified new small ORFs. As a consequence, most of these new ORFs have not yet been assessed for conservation in closely related species. In addition, many of the new ORFs overlap with other genes, making analysis of conservation problematic. When clear evidence for conservation was not available, new ORFs that overlapped existing ORFs were assigned 'dubious' designations, while all others were classified as 'uncharacterized'.

Thus, all *S. cerevisiae* ORFs are now categorized into one of three groups: 'dubious', referring to those ORFs that are unlikely to encode a protein; 'uncharacterized', those that are likely, but not yet fully established, to encode a protein; and 'verified', those for which there is clear experimental evidence for the presence of a protein-encoding gene. It should be noted that these ORF classifications are not static properties and are expected to change as new data become available for each ORF. In the almost 3 years since the original analysis, the classifications of 299 ORFs have been updated; 90% of these changes have been from 'uncharacterized' to 'verified'. Very few 'dubious' ORFs (19 of 832 nuclear ORFs) have been reclassified as either 'uncharacterized' or 'verified'. Experimental evidence supporting the validity of these classifications is beginning to accumulate. For example, Raisner *et al*. (2006) reported that the variant histone protein H2A.Z is associated with the 5′ ends of 'verified' and 'uncharacterized' ORFs, but not with the 5′ ends of silenced genes or 'dubious' ORFs.

## Sequence changes and ORF revision

Any large-scale analysis will include some percentage of errors, and large-scale sequencing projects are no exception. During the last decade, a total of 185 ORFs have been revised due to the correction of demonstrated sequencing errors (Figure 2).

The ORF revisions and underlying sequence corrections vary widely in nature. They range from single nucleotide changes that alter the nature of a single critical amino acid (e.g.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

MCM6/YGL201C; Andrea Duina, personal communication; Gen-Bank Accession No. AY258324); to multiple changes, insertions and deletions resulting in a C-terminal extension and a new stop codon (e.g. SAL1/YNL083W; Belenkiy *et al*., 2000; Brachat *et al*., 2003); to the insertion of a 220 bp region that had not been included in the original sequence (HSP150/YJL159W; Moukadiri and Zueco, 2001; Brachat *et al*., 2003).

As with new small ORFs, the errors in the reference sequence were typically discovered during the course of focused experimental research. However, the recent large-scale genomic comparisons have allowed for much more rapid identification of a particular subset of sequencing errors. When identifying orthologues in closely-related species, Blandin *et al*. (2000), Brachat *et al*. (2003), Cliften *et al*. (2003) and Kellis *et al*. (2003) noticed many cases in which a gene was largely conserved across species in sequence and position, but the *S. cerevisiae* gene contained extensions or deletions relative to its predicted orthologues, suggesting that sequencing errors might have led to incorrect annotation of its 5′ or 3′ boundary.

In many instances, the authors tested their predictions by resequencing genes themselves (Brachat *et al*., 2003; Kellis *et al*., 2003). In some additional cases, other researchers independently predicted, tested and confirmed the same sequencing errors (Schmalix and Bandlow, 1994; Beh *et al*., 2001; Xiao *et al*., 1998; Treton *et al*., 2000; Angus-Hill *et al*., 2001; Moukadiri and Zueco, 2001; Kaliraman *et al*., 2001; Palmer *et al*., 2001; Robben *et al*., 2002; Jaspersen *et al*., 2002; Denis and Cyert, 2002; Muller *et al*., 2003; Charlie Boone, personal communication; Jim Brown, personal communication; Clyde Denis, personal communication; Tim Formosa, personal communication; Claude Gaillardin and Aaron P. Mitchell, personal communication; Gerard Manning, personal communication). In all remaining cases, SGD curators examined and tested the recommended sequence changes. Upon close examination of the sequence alignments and available literature for each gene, some of the proposals were rejected due to inadequate or unconvincing alignments with related fungal sequences, but in most cases, it was straightforward to predict a sequence change that would produce a highly conserved ORF.

## Annotation changes and ORF revision

In the original *S. cerevisiae* genomic annotation, each ORF was simply annotated as the longest possible reading frame. However, comparison with closely related species suggested that for some ORFs, the methionine codon that produced the longest possible reading frame might not actually represent the translational start. In these cases, the conserved start codon in the orthologues aligned with a downstream, in-frame methionine codon, rather than the start codon annotated in *S. cerevisiae*. Changing the *S. cerevisiae* annotation to use the downstream, conserved start codon effectively produces a 5′ truncation of these ORFs, relative to their previous annotation. Kellis *et al*. (2003) recommended 120 such changes. In some cases published data, such as protein size determination or N-terminal sequencing, corroborated the new predictions (Adzuma *et al*., 1984; Taylor *et al*., 1987; Dean-Johnson and Henry, 1989; Hanes *et al*., 1989; Sanni *et al*., 1991; Wang *et al*., 1992; Poon and Storms, 1994; Sanders and Herskowitz, 1996; Horazdovsky *et al*., 1997; Nothwehr and Hindes, 1997; Zheng *et al*., 1997; Mori *et al*., 1998; Davis *et al*., 2000; Kurtz *et al*., 2002; Willer *et al*., 2003; Rodney Rothstein, personal communication). In the absence of published experimental data, the new start codon was accepted only if it was the predicted start in at least three of the four available *Saccharomyces sensu stricto* species (*S. bayanus, S. paradoxus, S. mikatae* or *S. kudriavzevii* ). Of the recommended start site changes, 87 (72%) met these criteria and were incorporated into SGD. Four more were later added because they were confirmed by Zhang and Dietrich (2005), who also discovered an additional four start codon changes that Kellis *et al*. (2003) had not predicted. Although this number is small in comparison, it does illustrate the point that the work done by Kellis *et al*. (2003) was not

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

saturating, and we can expect that focused experimental work may identify even more start codon corrections.

The original annotation for the budding yeast genome contained 225 genes with introns. Introns are rare in yeast, tend to be in the extreme 5′ end of the gene, and typically include a perfect match to the branch site consensus (UACUAAC; Spingola *et al*., 1999). Since 1996, only 39 new introns and exons have been identified. The majority of these were identified by Brachat *et al*. (2003) and Cliften *et al*. (2003), who proposed that a combined total of 24 existing ORFs be updated with new introns and exons, such that the reading frame of the original ORF was preserved but the new intron and exon effectively added an extension at either the 5′ or the 3′ end. In some instances, the intron/exon predictions were directly tested (Brachat *et al*. 2003). For the remainder, SGD curators examined the *sensu stricto Saccharomyces* data published by Kellis *et al*. (2003), which was not used for the intron predictions by Brachat *et al*. (2003) and Cliften *et al*. (2003). The new intron/exon structure was annotated only if the reading frame, the start and stop codons and the branch site splicing signals were conserved in the other species.

In a few cases, examination of the evidence led to revision of the proposed change. For example, based on sequence conservation between *Ashbya gossypii* and *S. cerevisiae*, Brachat *et al*. (2003) proposed an intron and a new 3′ exon for *SEF1* /YBL066C. However, when the *SEF1* sequences from four *Saccharomyces sensu stricto* species were compared to *S. cerevisiae*, the comparative data argued against the presence of an intron. Instead it appeared that the *S. cerevisiae* sequence contained a large number of sequencing errors in this gene. SGD resequenced the 150 base pairs spanning the divergent region and found that 37 nucleotide insertions and four nucleotide substitutions were necessary to correct the reference sequence. Once these errors were corrected, the *S. cerevisiae SEF1* ORF displayed close conservation with the other *Saccharomyces sensu stricto* orthologues, none of which was predicted to contain an intron.

## Documentation

Sequence and annotation changes are announced regularly on SGD's homepage and in our quarterly newsletter. All changes are also tracked and posted in a more permanent manner, on SGD web pages and at our FTP site.

The Locus Summary page, the basic unit of the SGD website, includes a 'Sequence Information' section located near the bottom of the page. This section lists sequence and coordinate details for that feature, including the dates when each was last updated. A detailed description of each update is provided on the Locus History page (accessible from a tab at the top of the Locus Summary page).

The Locus Summary provides focused update information on a gene-by-gene basis, but this information is also available via the web in more comprehensive forms. The Chromosome History pages (http://www.yeastgenome.org/chromosomes) provide a complete list of changes for each chromosome. The Advanced Search tool can be used to generate lists of all currently annotated ORFs of each classification (verified, uncharacterized, dubious) as well as lists of any other type of annotated chromosomal feature.

Comprehensive information is also available for download via the SGD site (ftp://ftp.yeastgenome.org/yeast/). Sequences for these features, as well as for entire chromosomes and intergenic regions, can be found in the 'genomic_sequence' directory.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

## Conclusion

Incorporation of sequence and annotation changes over the past decade has resulted in a significantly more accurate reference sequence for *S. cerevisiae*. However, although the recent large-scale comparative analyses (Blandin *et al*., 2000; Brachat *et al*., 2003; Cliften *et al*., 2003; Kellis *et al*., 2003) have provided a bonanza of sequence and annotation corrections, we expect that more errors will be discovered lurking within the reference sequence. The broad scope of these analyses revealed gross errors in genomic annotation, such as mistakes in intron/exon structure or ORF boundaries. A narrower focus will be required for the detection of more subtle errors that likely exist in both coding and intergenic regions, and we anticipate a continually refined reference sequence and its annotation.

## Acknowledgments

This work was made possible by the support of Gavin Sherlock in performing nucleotide sequencing within his laboratory. We appreciate the efforts of the many scientists who have corresponded with SGD about errors in the genome sequence and who have, in many cases, resequenced genomic regions in order to ensure the quality of the reference sequence. We thank Maria C. Costanzo and all the members of the SGD group for helpful discussions and critical reading of the manuscript. SGD is supported by a P41 grant, Genome Research Resource (No. HG001315), from the National Human Genome Research Institute at the US National Institutes of Health.

## References

Adzuma K, Ogawa T, Ogawa H. Primary structure of the RAD52 gene in *Saccharomyces cerevisiae*. Mol Cell Biol 1984;4:2735–2744. [PubMed: 6098821]

Angus-Hill ML, Schlichter A, Roberts D, et al. A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control. Mol Cell 2001;7:741–751. [PubMed: 11336698]

Basrai MA, Hieter P, Boeke JD. Small open reading frames: beautiful needles in the haystack. Genome Res 1997;7:768–771. [PubMed: 9267801]

Beh CT, Cool L, Phillips J, Rine J. Overlapping functions of the yeast oxysterol-binding protein homologues. Genetics 2001;157:1117–1140. [PubMed: 11238399]

Belenkiy R, Haefele A, Eisen MB, Wohlrab H. The yeast mitochondrial transport proteins: new sequences and consensus residues, lack of direct relation between consensus residues and transmembrane helices, expression patterns of the transport protein genes, and protein–protein interactions with other proteins. Biochim Biophys Acta 2000;1467:207–218. [PubMed: 10930523]

Blandin G, Durrens P, Tekaia F, et al. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. FEBS Lett 2000;487:31–36. [PubMed: 11152879]

Brachat S, Dietrich FS, Voegeli S, et al. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. Genome Biol 2003;4:R45. [PubMed: 12844361]

Cliften P, Sudarsanam P, Desikan A, et al. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 2003;301:71–76. [PubMed: 12775844]

Davis CA, Grate L, Spingola M, Ares M Jr. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast. Nucleic Acids Res 2000;28:1700–1706. [PubMed: 10734188]

Dean-Johnson M, Henry SA. Biosynthesis of inositol in yeast. Primary structure of myo-inositol-1-phosphate synthase (EC 5.5.1.4) and functional analysis of its structural gene, the INO1 locus. J Biol Chem 1989;264:1274–1283. [PubMed: 2642902]

Denis V, Cyert MS. Internal Ca(2+) release in yeast is triggered by hypertonic shock and mediated by a TRP channel homologue. J Cell Biol 2002;156:29–34. [PubMed: 11781332]

Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567. [PubMed: 8849441]

Hanes SD, Shank PR, Bostian KA. Sequence and mutational analysis of *ESS1*, a gene essential for growth in *Saccharomyces cerevisiae*. Yeast 1989;5:55–72. [PubMed: 2648698]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Hirschman JE, Balakrishnan R, Christie KR, et al. Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. Nucleic Acids Res 2006;34:D442–445. [PubMed: 16381907]

Horazdovsky BF, Davies BA, Seaman MN, et al. A sorting nexin-1 homologue, Vps5p, forms a complex with Vps17p and is required for recycling the vacuolar protein-sorting receptor. Mol Biol Cell 1997;8:1529–1541. [PubMed: 9285823]

Jaspersen SL, Giddings TH Jr, Winey M. Mps3p is a novel component of the yeast spindle pole body that interacts with the yeast centrin homologue Cdc31p. J Cell Biol 2002;159:945–956. [PubMed: 12486115]

Kaliraman V, Mullen JR, Fricke WM, Bastin-Shanower SA, Brill SJ. Functional overlap between Sgs1-Top3 and the Mms4-Mus81 endonuclease. Genes Dev 2001;15:2730–2740. [PubMed: 11641278]

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 2003;423:241–254. [PubMed: 12748633]

Kessler MM, Zeng Q, Hogan S, et al. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. Genome Res 2003;13:264–271. [PubMed: 12566404]

Koehler CM, Murphy MP, Bally NA, et al. Tim18p, a new subunit of the TIM22 complex that mediates insertion of imported proteins into the yeast mitochondrial inner membrane. Mol Cell Biol 2000;20:1187–1193. [PubMed: 10648604]

Kumar A, Harrison PM, Cheung KH, et al. An integrated approach for finding overlooked genes in yeast. Nat Biotechnol 2002;20:58–63. [PubMed: 11753363]

Kurtz JE, Exinger F, Erbs P, Jund R. The URH1 uridine ribohydrolase of *Saccharomyces cerevisiae*. Curr Genet 2002;41:132–141. [PubMed: 12111094]

Mackiewicz P, Kowalczuk M, Mackiewicz D, et al. How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome? Yeast 2002;19:619–629. [PubMed: 11967832]

Malpertuy A, Tekaia F, Casaregola S, et al. Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes-specific genes. FEBS Lett 2000;487:113–121. [PubMed: 11152894]

McKee AH, Kleckner N. A general method for identifying recessive diploid-specific mutations in *Saccharomyces cerevisiae*, its application to the isolation of mutants blocked at intermediate stages of meiotic prophase and characterization of a new gene, *SAE2*. Genetics 1997;146:797–816. [PubMed: 9215888]

Mori K, Ogawa N, Kawahara T, Yanagi H, Yura T. Palindrome with spacer of one nucleotide is characteristic of the *cis*-acting unfolded protein response element in *Saccharomyces cerevisiae*. J Biol Chem 1998;273:9912–9920. [PubMed: 9545334]

Moukadiri I, Zueco J. Evidence for the attachment of Hsp150/Pir2 to the cell wall of *Saccharomyces cerevisiae* through disulfide bridges. FEMS Yeast Res 2001;1:241–245. [PubMed: 12702350]

Muller O, Neumann H, Bayer MJ, Mayer A. Role of the Vtc proteins in V-ATPase stability and membrane trafficking. J Cell Sci 2003;116:1107–1115. [PubMed: 12584253]

Nothwehr SF, Hindes AE. The yeast VPS5/GRD2 gene encodes a sorting nexin-1-like protein required for localizing membrane proteins to the late Golgi. J Cell Sci 1997;110:1063–1072. [PubMed: 9175702]

Oshiro G, Wodicka LM, Washburn MP, et al. Parallel identification of new genes in *Saccharomyces cerevisiae*. Genome Res 2002;12:1210–1220. [PubMed: 12176929]

Palmer CP, Zhou XL, Lin J, et al. A TRP homolog in *Saccharomyces cerevisiae* forms an intracellular Ca(2+)-permeable channel in the yeast vacuolar membrane. Proc Natl Acad Sci USA 2001;98:7801–7805. [PubMed: 11427713]

Poon PP, Storms RK. Thymidylate synthase is localized to the nuclear periphery in the yeast *Saccharomyces cerevisiae*. J Biol Chem 1994;269:8341–8347. [PubMed: 8132557]

Raisner RM, Madhani HD. Patterning chromatin: form and function for H2A.Z variant nucleosomes. Curr Opin Genet Dev 2006;16:119–124. [PubMed: 16503125]

Robben J, Hertveldt K, Volckaert G. Revisiting the yeast chromosome VI DNA sequence reveals a correction merging YFL007w and YFL006w to a single ORF. Yeast 2002;19:699–702. [PubMed: 12185839]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Sanders SL, Herskowitz I. The BUD4 protein of yeast, required for axial budding, is localized to the mother/BUD neck in a cell cycle-dependent manner. J Cell Biol 1996;134:413–427. [PubMed: 8707826]

Sanni A, Walter P, Boulanger Y, Ebel JP, Fasiolo F. Evolution of aminoacyl-tRNA synthetase quaternary structure and activity: *Saccharomyces cerevisiae* mitochondrial phenylalanyl-tRNA synthetase. Proc Natl Acad Sci USA 1991;88:8387–8391. [PubMed: 1924298]

Schmalix WA, Bandlow W. *SWH1* from yeast encodes a candidate nuclear factor containing ankyrin repeats and showing homology to mammalian oxysterol-binding protein. Biochim Biophys Acta 1994;1219:205–210. [PubMed: 8086466]

Sophianopoulou V, Diallinas G. *AUA1*, a gene involved in ammonia regulation of amino acid transport in *Saccharomyces cerevisiae*. Mol Microbiol 1993;8:167–178. [PubMed: 8497191]

Spingola M, Grate L, Haussler D, Ares M Jr. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. RNA 1999;5:221–234. [PubMed: 10024174]

Taylor GR, Lagosky PA, Storms RK, Haynes RH. Molecular characterization of the cell cycle-regulated thymidylate synthase gene of *Saccharomyces cerevisiae*. J Biol Chem 1987;262:5298–5307. [PubMed: 3031048]

Termier M, Kalogeropoulos A. Discrimination between fortuitous and biologically constrained open reading frames in DNA sequences of *Saccharomyces cerevisiae*. Yeast 1996;12:369–384. [PubMed: 8701609]

Treton B, Blanchin-Roland S, Lambert M, Lepingle A, Gaillardin C. Ambient pH signalling in ascomycetous yeasts involves homologues of the *Aspergillus nidulans* genes *palF* and *paIH*. Mol Gen Genet 2000;263:505–513. [PubMed: 10821185]

Wang SS, Stanford DR, Silvers CD, Hopper AK. *STP1*, a gene involved in pre-tRNA processing, encodes a nuclear protein containing zinc finger motifs. Mol Cell Biol 1992;12:2633–2643. [PubMed: 1588961]

Willer M, Jermy AJ, Young BP, Stirling CJ. Identification of novel protein–protein interactions at the cytosolic surface of the Sec63 complex in the yeast ER membrane. Yeast 2003;20:133–148. [PubMed: 12518317]

Wood V, Rutherford KM, Ivens A, Rajandream M-A, Barrell B. A re-annotation of the *Saccharomyces cerevisiae* genome. Comp Funct Genom 2001;2:143–154.

Xiao W, Chow BL, Milo CN. Mms4, a putative transcriptional (co)activator, protects *Saccharomyces cerevisiae* cells from endogenous and environmental DNA damage. Mol Gen Genet 1998;257:614–623. [PubMed: 9604884]

Zhang CT, Wang J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. Nucleic Acids Res 2000;28:2804–2814. [PubMed: 10908339]

Zhang Z, Dietrich FS. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. Nucleic Acids Res 2005;33:2838–2851. [PubMed: 15905473]

Zheng W, Xu HE, Johnston SA. The cysteine-peptidase bleomycin hydrolase is a member of the galactose regulon in yeast. J Biol Chem 1997;272:30350–30355. [PubMed: 9374524]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript



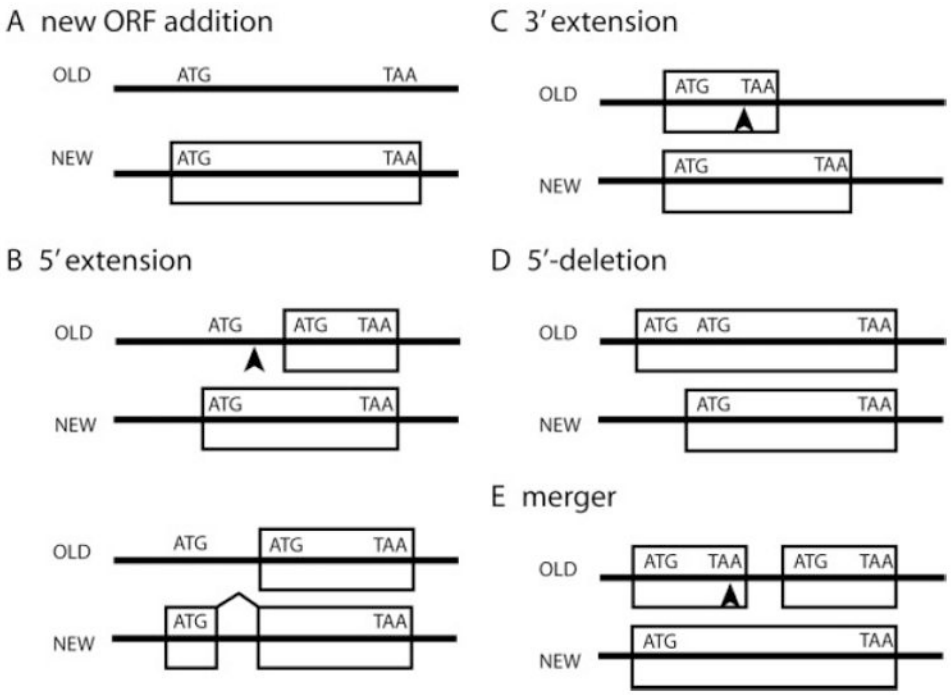


**Figure 1.**

Sequence annotation changes since 1996. Arrow symbol represents the location of an indel [inserted or deleted nucleotide(s)]. Only the most common types of change have been diagrammed. (A) New ORF addition: 523 small ORFs have been added as a result of new experimental reports, comparative genomic analysis and sequence changes. (B) 5′ Extension: 64 ORFs were extended when an indel caused an upstream ATG to be brought in frame with the existing coding region; 21 extensions resulted from the discovery of an upstream intron. (C) 3′ Extension: 51 ORFs were extended when the sequence of the ORF changed to alter the reading frame and/or STOP codon position. Note that sequence changes also produced 20 ORFs with 3′ deletions (not diagrammed). An additional four 3′ extensions resulted from the discovery of a downstream intron (not diagrammed). (D) 5′ Deletion: 113 ORFs have decreased in length as a result of comparative analysis or because experimental results showed that the first ATG was not the start of translation. (E) Merger: 20 sequence changes have merged two adjacent ORFs

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

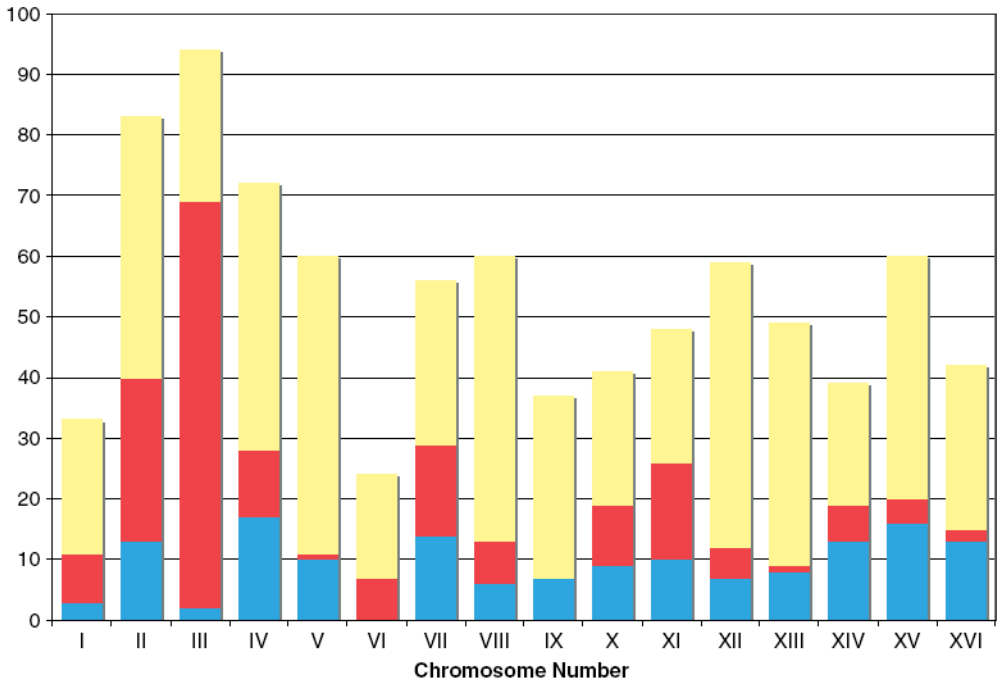


**Figure 2.**

Changes per chromosome since 1996. For each chromosome, yellow indicates the number of new ORFs; red indicates the number of ORFs revised due to changes in the chromosomal sequence; and blue indicates the number of ORFs that were revised without changes to the chromosomal sequence. Sequence changes include confirmed changes reported by the research community (Schmalix and Bandlow, 1994; Xiao *et al*., 1998; Treton *et al*., 2000; Angus-Hill *et al*., 2001; Beh *et al*., 2001; Kaliraman *et al*., 2001; Moukadiri and Zueco, 2001; Palmer *et al*., 2001; Robben *et al*., 2002; Jaspersen *et al*., 2002; Denis and Cyert, 2002; Muller *et al*., 2003; Charlie Boone, personal communication; Jim Brown, personal communication; Clyde Denis, personal communication; Tim Formosa, personal communication; Claude Gaillardin and Aaron P. Mitchell, personal communication; Gerard Manning, personal communication). Note that the graph includes only changes that were incorporated into the reference genome. A total of 43 proposed sequence changes have been rejected because re-sequencing verified the original sequence. An additional 54 proposed annotation-based revisions have also been rejected due to inadequate data