An integrated approach for finding overlooked genes in yeast

Anuj Kumar[1], Paul M. Harrison[2], Kei-Hoi Cheung[3], Ning Lan[2], Nathaniel Echols[2], Paul Bertone[2], Perry Miller[3], Mark B. Gerstein[2], and Michael Snyder[1,2]*

We report here the discovery of 137 previously unappreciated genes in yeast through a widely applicable and highly scalable approach integrating methods of gene-trapping, microarray-based expression analysis, and genome-wide homology searching. Our approach is a multistep process in which expressed sequences are first trapped using a modified transposon that produces protein fusions to β-galactosidase (β-gal); non-annotated open reading frames (ORFs) translated as β-gal chimeras are selected as a candidate pool of potential genes. To verify expression of these sequences, labeled RNA is hybridized against a microarray of oligonucleotides designed to detect gene transcripts in a strand-specific manner. In complement to this experimental method, novel genes are also identified *in silico* by homology to previously annotated proteins. As these methods are capable of identifying both short ORFs and antisense ORFs, our approach provides an effective supplement to current gene-finding schemes. In total, the genes discovered using this approach constitute 2% of the yeast genome and represent a wealth of overlooked biology.

At present, genome sequences (in draft or finished forms) are available for >800 different organisms. While this unprecedented volume of raw sequence data is undeniably valuable, the ultimate utility of genomic sequence as an interdisciplinary information resource depends largely on the accuracy and completeness with which it is annotated. Toward this end, a variety of computational and experimental approaches have been utilized to identify genes within genome sequence. Recent computational approaches have employed either probabilistic or pattern-based schemes to score candidate genes[1]; however, such predictive approaches have met with varying degrees of success as a means of annotating eukaryotic genomes. Within eukaryotes, evidence of sequence homology and/or expression has served as the standard criterion by which genes are identified[2]. Eukaryotic genes are typically annotated by means of sequence, motif, and structure comparison against known proteins and translated expressed sequence tags[3]. Genes may also be identified experimentally through traditional methods of gene cloning or by random analysis of complementary DNA (cDNA) clones. Typical large-scale approaches to gene identification employ screening methods or microarray-based technologies to characterize expressed sequences from representative cDNA libraries[4,5]. No single technique, however, is comprehensive. Not all genes can be identified on the basis of exhibited homology or sequence characteristics; weakly expressed genes may be underrepresented in cDNA libraries. Even applied in combination, these approaches have failed to result in the exhaustive annotation of one single eukaryotic genome.

To illustrate this point, consider the intensively studied genome of *Saccharomyces cerevisiae*. In 1996 an international consortium of 16 research teams completed a collaborative project to determine the nucleotide sequence of all 16 chromosomes constituting the nuclear genome of *S. cerevisiae* lab strain S288c (ref. 6). As released in 1997, this 13-megabase genome was predicted to encode a total of 6,274 genes[7].

Annotated sequences encompassed both previously known genes identified experimentally as well as putative genes selected using simple gene-finding algorithms[7,8]. The algorithms and criteria employed in this process (though not uniformly applied to all chromosomes) selectively identified ORFs extending at least 100 codons in length from start to stop codon. Shorter ORFs were annotated only if they corresponded to known genes or exhibited strong sequence similarity to known proteins. Overlapping ORFs satisfying these criteria were also annotated; however, ORFs nested within longer ORFs on either the same or complementary strand were excluded[8]. By these criteria, most—but not all—yeast genes were identified. Over the last four years, 65 previously non-annotated genes have been discovered in yeast, largely as by-products of data from functional or comparative genomic studies[9–11]. Systematic methods designed specifically to identify such genes will be necessary in order to more comprehensively annotate genomes from most eukaryotes, including yeast.

We present here a study integrating both experimental and computational methods as a means of discovering previously overlooked genes in yeast. All genes identified by this approach satisfy stringent criteria for gene annotation[2]; that is, all genes are identified as such upon evidence of expression or homology to a known protein. As outlined in Figure 1, candidate genes are first identified by large-scale insertional mutagenesis using a modified transposon as a simple gene trap. Expression of each candidate gene is independently verified by microarray analysis such that gene-coding sequence can again be identified in a strand-specific manner. No cDNA is required for this analysis: rather, labeled RNA is simply hybridized against a microarray of oligonucleotides representing both strands of each putative gene locus. Only sequences detected by both gene-trapping and RNA analysis are considered further—potentially as bona fide genes. As transposon-based gene-trapping is random, systematic computational approaches were

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, P.O. Box 208103, New Haven, CT 06520-8103. [2]Department of Molecular Biophysics and Biochemistry, Yale University, P.O. Box 208114, New Haven, CT 06520-8114. [3]Center for Medical Informatics, Department of Anesthesiology, Yale University School of Medicine, New Haven, CT 06510. *Corresponding author (michael.snyder@yale.edu).
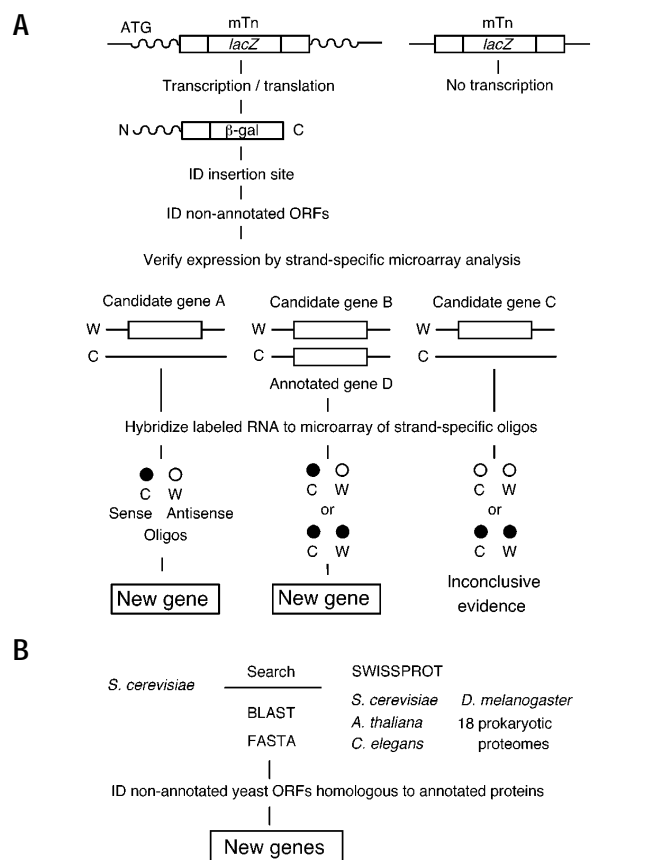
Figure 1. Overview of the integrated approach for finding overlooked yeast genes. (A) Experimental approaches. Expressed sequences were first identified by means of gene trapping using a minitransposon (mTn) bearing a *lacZ* reporter missing both its start codon and promoter. Using a high-throughput assay, over 150,000 mTn-mutagenized yeast strains were screened for β-gal activity so as to identify those strains containing an insertion within a region of the genome that is transcribed and translated. Transposon insertion sites were identified within 15,360 strains exhibiting β-gal activity, and a promising subset of non-annotated ORFs translated as β-gal chimeras were selected for further study as candidate genes. Expression of each candidate gene was verified by microarray analysis of RNA levels using strand-specific oligonucleotides. For this study, poly(A) RNA was directly labeled and hybridized to an oligonucleotide microarray. Two 60-mer oligos were used to interrogate each putative gene—a sense oligo and its complementary antisense oligo. Example results are shown for three candidate genes. Candidate gene A is classified as a gene, as mRNA is preferentially bound to its sense oligo. Candidate gene B (coding sequence on the Watson strand) is oriented opposite annotated gene D. Candidate gene B is also classified as a gene if RNA hybridization to its sense oligo is observed; as gene D may also be expressed, preferential hybridization to the sense oligo (Crick strand) relative to the antisense oligo (Watson strand) is not a meaningful prerequisite in this case. Candidate gene C is not classified as a gene: its sense and antisense oligos are bound equally, whether at low or high levels. (B) Computational approaches. The yeast genome was systematically searched in translation against the SWISS–PROT protein sequence database supplemented with sequence data from the 22 annotated proteomes listed. Non-annotated yeast ORFs homologous to annotated proteins were identified; those ORFs exhibiting strong and extended similarity to a previously annotated protein are classified here as genes.

also employed to identify novel genes more comprehensively. In total, this integrated approach revealed 137 previously overlooked genes in yeast—more than twice as many new yeast genes as had been identified over the last four years combined. The majority of these new genes are either short or overlap a previously annotated gene on the opposite strand—two gene classes present, yet severely underrepresented, among most sequenced genomes to date. As our approach is well suited to
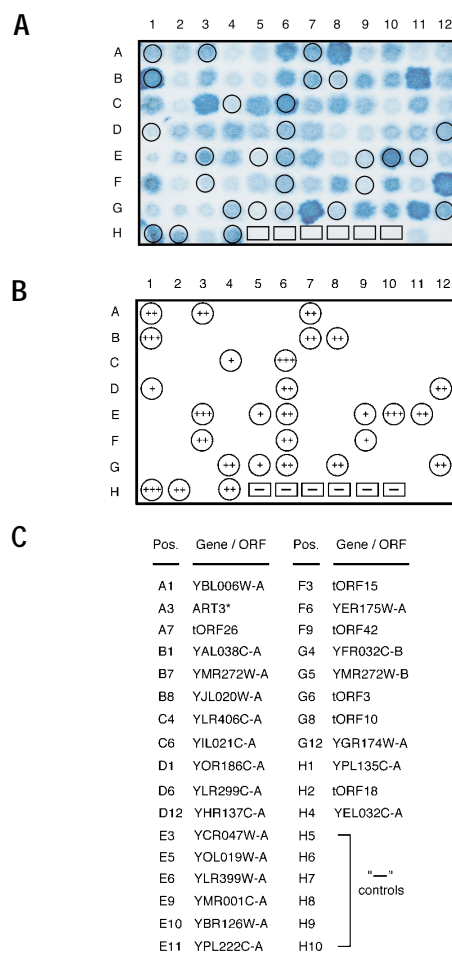

Figure 2. Analysis of β-gal activity in transposon-tagged yeast. (A) Filter-based assay of 96 transposon-mutagenized yeast strains under conditions of vegetative growth. Circled strains contain insertions within previously non-annotated ORFs. Boxed strains contain insertions resulting in no detectable levels of β-gal production. (B) Intensities of β-gal staining were scored as follows: –, no detectable β-gal activity; +, light blue staining; ++, blue staining; +++, intense blue staining. (C) A sampling of novel genes detected by transposon tagging. Genes identified within each circled strain are listed alongside their corresponding position in filter 2A (rows A–H, columns 1–12). New genes have been provided with systematic names in accordance with community-accepted standards: each gene was named by adding a hyphenated letter to the name of its centromere-proximal adjacent ORF. The "tORF" nomenclature is used to designate any non-annotated ORF (detected in this study by both transposon tagging and expression analysis) that we do not classify as a gene (see text). Repeated genes are indicated with an asterisk. In this study, genes identified antisense to rDNA are named *ART*; β-gal analysis of a *lacZ* fusion to *ART3* is shown here. Complete β-gal assay results for all genes (and tORFs) identified in this study may be accessed online at bioinfo.mbb.yale.edu/genome/ yeast/orfome/new-genes.

identify these and other genes, it serves as a method by which over-looked genes may be identified within annotated genomes and as a par-adigm for large-scale gene-finding within newly sequenced genomes.

Results and discussion

Gene trapping. Expressed ORFs were identified using a Tn3-derived transposon containing a promoter-less and 5′-truncated *lacZ* reporter[12]. Transposon insertions were introduced into the yeast genome at random by means of shuttle mutagenesis[13,14]. Those insertions resulting in transcription and translation of *lacZ* fusions under conditions of vegetative growth and sporulation were identified using a high-throughput filter-based assay for β-gal activity (Fig. 2)[15]. The
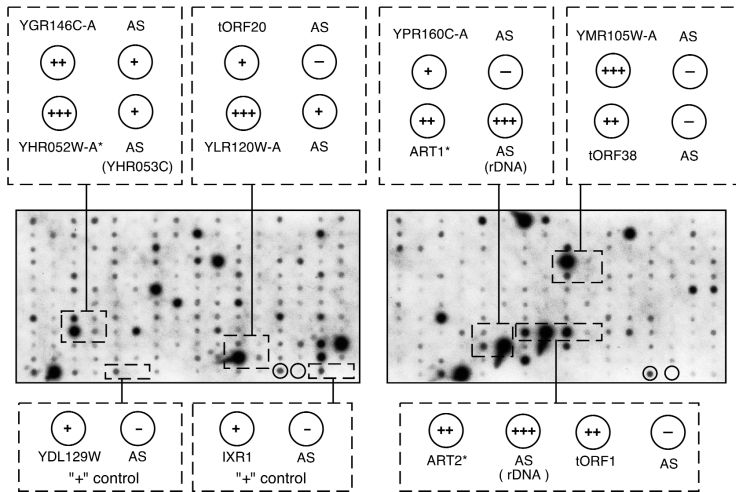
Figure 3. Microarray-based expression analysis with strand-specific oligos. Arrays were generated using 50- to 60-mer oligonucleotides ordered in a 12 × 16 grid on glass slides mounted with nylon membrane (50 mm in length and 19 mm in width). Biotinylated poly(A) RNA was hybridized to each array; bound probe was visualized by chemiluminescent detection. Arrays are ordered such that sense and antisense (AS) oligonucleotides used to interrogate a given transcript are positioned horizontally adjacent to each other. Example oligonucleotide pairs are highlighted in boxes, with accompanying diagrams indicating each assayed ORF and its hybridization intensity (– to +++; threshold detection of transcripts expressed at 0.2 copies/cell). As this hybridization assay cannot distinguish between individual members of repeated gene families, repeated sequences are marked with an asterisk. *YDL129W* and *IXR1* (bottom, left) are known genes included here as positive controls. Circled spots correspond to identical oligonucleotides included on both slides as an indication of assay reproducibility.

identity of each fusion protein was determined by sequencing its corresponding plasmid-borne insertion allele at the transposon–yeast DNA junction. Sequence data from 15,360 alleles identified insertions affecting 2,251 annotated genes and an additional set of previously non-annotated ORFs. We selected for further study all non-annotated ORFs that were (1) greater than 25 codons in length (2) located either within an intergenic region of the yeast genome or oriented opposite a previously annotated gene, and (3) identified by multiple productive transposon insertions and/or strong levels of β-gal activity. In total, 196 non-annotated ORFs satisfying all three of these criteria were chosen as candidate genes; of these sequences, 123 were identified by multiple transposon insertions, and 137 exhibited strong levels of β-gal activity as a fusion to *lacZ*.

Microarray-based expression analysis. To verify expression of these 196 sequences, we subjected each ORF to dot-blot analysis using strand-specific oligonucleotides in microarray format (Figs 1A, 3). For this analysis, poly(A) RNA was extracted from vegetatively growing diploid cells. This RNA was subsequently biotinylated and hybridized to a microarray of long oligonucleotides (50- to 60-mers) spotted onto a membrane-coated glass slide (see Experimental Protocol). Oligonucleotides were used to interrogate both strands of each putative gene so as to discriminate between genes oriented opposite each other. In addition to providing a means by which sense and antisense transcripts may be distinguished, the inclusion of oligonucleotides corresponding to each strand at a given locus serves to identify and reduce "background" hybridization. Selinger *et al.*[16] have detected a low level of transcriptional activity present throughout much of the *Escherichia coli* genome; the possibility exists that transcription may be similarly "leaky" within eukaryotes. By comparing hybridization intensities between both sense and antisense oligonucleotides, we can identify preferentially transcribed sequences, presumably indicative of biological relevance. We similarly decrease gene-finding artifacts by treating as spurious those sense and antisense sequence pairs yielding

equivalent binding intensities, assuming the oligonucleotide sequences were drawn from a predicted intergenic region of the yeast genome. Stringent hybridization conditions and careful oligonucleotide design further aid in improving the specificity of this approach. In particular, 50-mer oligonucleotides have been shown to generate little cross-hybridization to "nontarget" transcripts of <75% sequence identity, provided the nontarget sequences do not possess any contiguous stretches of complementary sequence >15 bases in length[17]. Note that all repeated sequences assayed in this study are indicated as such within Figures 2–5.

Utilizing this expression array approach, we detected preferential hybridization to sense oligonucleotides in sequences representing 70 novel ORFs located within regions of the yeast genome previously considered to be intergenic. We further detected significant hybridization to sense oligonucleotides representing 79 non-annotated ORFs predicted to lie partially or completely opposite previously annotated genes. In 21 of these 79 cases, messenger RNA (mRNA) transcripts were bound to antisense oligonucleotides at levels above background, indicating that both overlapping genes are transcribed under the assayed growth conditions. It should be noted that, even in these 21 cases, sense and antisense transcripts from a given locus may be temporally separated, as RNA used in this study was extracted from an asynchronous population. In total, this analysis verifies expression of 149 previously non-annotated ORFs; because these ORFs were initially identified as chimeras to transposon-encoded β-gal, we refer to them here as transposon-tagged ORFs (tORFs).

New genes and tORFs. The majority of these tORFs can appropriately be classified as genes; specifically, we identify as genes those tORFs that satisfy the following criteria. All tORFs located within intergenic regions of the yeast genome have been designated for annotation, provided they are >150 base pairs from the coding sequence of any previously annotated ORF. This qualification is necessary in order to address the possibility that a given tORF resides in the 5′- or 3′-untranslated region of another ORF. Most yeast genes possess a 5′-untranslated region of <100 base pairs[18] and a 3′-untranslated region of similar length. All intergenic tORFs were also searched for consensus splice site donor, acceptor, and lariat sequences to ensure that these tORFs do not represent overlooked exons of previously annotated genes. Applying these criteria, 32 new genes were identified, ranging in length from 27 to 99 codons (including start and stop codons). Consistent with initial criteria for gene identification in *S. cerevisiae*, we have also classified as genes 15 tORFs that partially overlap an annotated ORF on the opposite strand. Possessing a mean length of 91 codons, 12 of these 15 genes are also under 100 codons in length. The novel gene *YAL038C-A*, however, is 325 codons long and has been detected as an expressed sequence by serial analysis of gene expression (SAGE analysis)[9]. Finally, on the basis of expression analysis and transposon tagging, we have identified 54 new genes wholly overlapping an annotated ORF on the opposite strand. These genes range in length from 27 to 217 codons, with moderate-to-strong expression levels as detected by RNA arrays. This complete set of 101 new genes is illustrated in Figures 4 and 5; data regarding these genes and tORFs can be found as Supplementary Tables 1 and 2 in the Web Extras page of *Nature Biotechnology* Online.

Additional lines of experimental and computational evidence suggest that these sequences are indeed transcribed and translated. Eighteen of these ORFs have been independently detected by SAGE analysis in yeast[9]. Also, 78 of 101 genes possess a codon adaptation index (CAI) of ≥0.1, indicative of sequences likely to be transcribed at moderate-to-high levels[19]. Furthermore, six of these genes encode proteins that can be localized to sites within the cell by means of epitope-tagging and immunofluorescence analysis (as shown in Fig. 6).
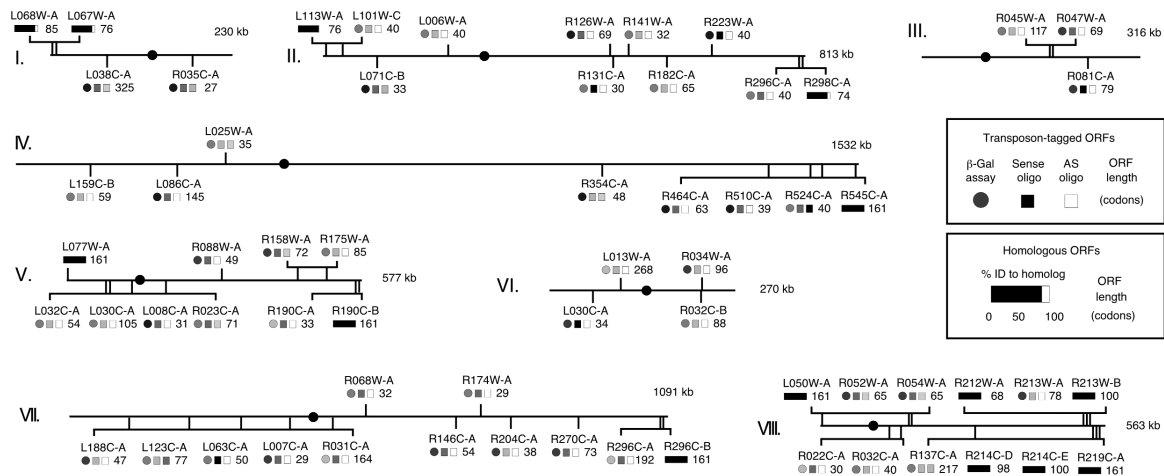
Figure 4. Distribution and analysis of previously unappreciated genes in yeast chromosomes I through VIII. Not drawn to scale. Each chromosome is represented as a linear map with its centromere (black dot) and end coordinate shown. Novel genes are indicated within either the Watson (upward) or Crick (downward) strand of each chromosome. In yeast, gene names begin with the letter "Y" followed by a letter (A–P) corresponding to chromosomes I–XVI. All genes shown here are identified by the third through ninth characters of their systematic names. For each gene identified by transposon tagging, β-gal assay results are indicated in the gray circle; levels of β-gal activity are represented as three shades of gray (from faint, ±, to dark gray, ±±±). RNA binding to sense and antisense oligos are depicted in the left-most and right-most shaded squares, respectively (see Key), with each hybridization signal represented by a shade of gray (from light gray, +, to black, +++). A white square indicates that no binding was detected. For each gene identified by homology searching, the percentage identity between it and its closest homolog (over the region of observed homology) is represented as the shaded fraction of the rectangle below its name (see Key). The length (in codons) of each gene reported in this study is indicated below its name (to the right).

Genome-wide homology searching. To complement these experimental approaches, we have also applied homology searching as a means of identifying previously unappreciated genes in yeast (Fig. 1B). The complete genome sequence of yeast was searched against the SWISS–PROT database plus the combined annotated proteomes of *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *S. cerevisiae* itself, and 18 prokaryotes (see Experimental Protocol) for any non-annotated ORFs displaying significant sequence similarity to a known protein. Excluding those ORFs exhibiting strong similarity to a "questionable" yeast gene (as classified in the Munich Information Center for Protein Sequences (MIPS)[20]), we identified 44 yeast ORFs homologous to annotated genes; for brevity, we refer to these homologous ORFs as hORFs. From these searches, we also verified 10 sequences recently proposed to be genes in a study by Blandin *et al.*[10] of conserved sequences from 13 hemiascomycetous yeast species (data not shown). Wood *et al.*[21]

recently identified three additional novel genes by homology to known *Schizosaccharomyces pombe* and human sequences; however, we did not detect these three genes in our analysis, as corresponding homologous sequences had likely not been deposited in SWISS–PROT at the time of this analysis.

Of the 44 hORFs reported here, eight align to a coding segment representing <30% of that protein's full length. While this confined region of similarity may still indicate homology, we consider these hORFs to be questionable. They may represent pseudogenes, which are thought to be rare in yeast[8,22]. The remaining 36 hORFs, however, exhibit strong and extended sequence similarity to annotated proteins. Of these 36 hORFs, 32 exhibit >80% identity to an annotated protein over the observed region of homology; this region extends over 90% of the homolog's coding sequence in 25 cases. On the basis of this homology, we classify these sequences as genes (Figs 4, 5).

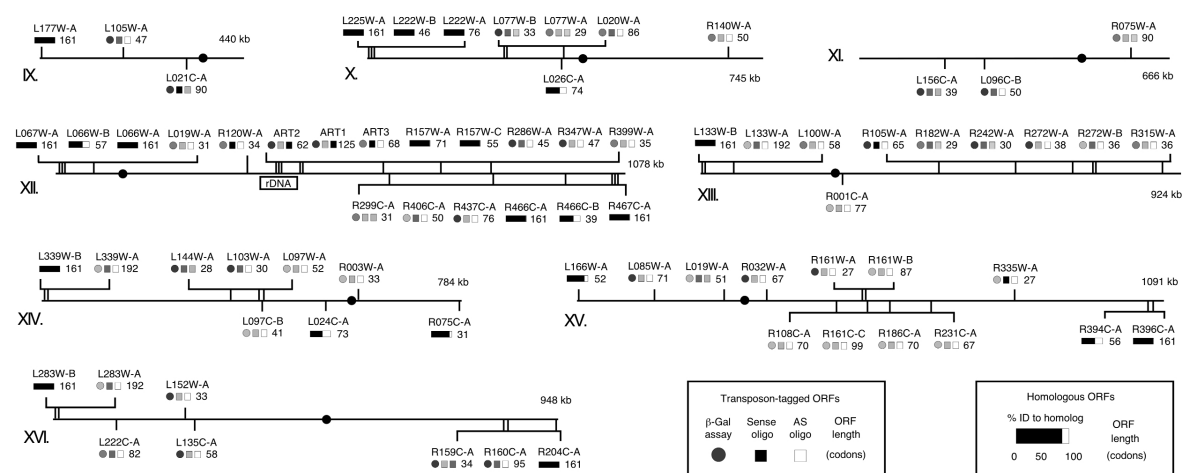The majority of these genes are paralogs of previously annotated

Figure 5. Distribution and analysis of new genes in yeast chromosomes IX through XVI. All chromosomes and new genes are represented as described in Figure 4. Three genes found opposite rDNA loci on chromosome XII are designated *ART1*, *ART2*, and *ART3*. In this diagram, the 1,260-kb rDNA repeat region of chromosome XII is represented by a single 9-kb rDNA repeat unit; the listed size of chromosome XII excludes this region of repetitive DNA.
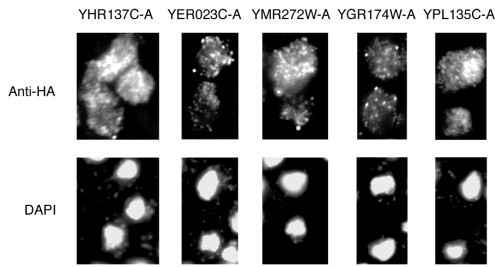
YHR137C-A YER023C-A YMR272W-A YGR174W-A YPL135C-A

Anti-HA

DAPI

Figure 6. Immunolocalization of novel gene products. A subset of gene products identified in this study were tagged with three copies of the HA epitope using a transposon-based approach described elsewhere[14]. Top panel: examples of immunofluorescence patterns in vegetative yeast cells stained with monoclonal antibodies directed against HA. Bottom panel: the same cells stained with the DNA-binding dye 4',6-diamidino-2-phenylindole (DAPI). Epitope-tagged *YHR137C-A* and *YMR272W-A* gene products localize to the cytoplasm, with slightly increased concentration around the nuclear rim and endoplasmic reticulum. Tagged alleles of *YER023C-A* and *YGR174W-A* yield punctate patterns of cytoplasmic staining. Granular staining of the cytoplasm and nucleus is evident upon immunolocalization of HA-tagged YPL135C-A protein.

yeast ORFs; however, two identified genes are most closely related to nonyeast proteins. *YNL024C-A* encodes a protein similar to the *CG14199* gene product in *D. melanogaster*, while *YJL026C-A* is an ortholog of *F ORF B*, a gene encoding a hypothetical 8.3 kDa protein in vaccinia virus. Of the remaining 34 paralogs identified here, 18 are homologous to YFL068W, an annotated, subtelomerically-encoded protein similar to a hypothetical protein in *E. coli*[22]. In total, these 19 ORFs constitute a previously overlooked family of genes encoded within subtelomeric regions of 12 different yeast chromosomes. Interestingly, a large majority of previously non-annotated genes identified in our homology searches are located toward the ends of chromosomes (Figs 4, 5). Despite their proximity to regions of telomeric repeats, the 36 genes reported here exhibit normal sequence complexity. Moreover, all but four of these genes possess a CAI >0.1. Also of note, four previously non-annotated genes (*YHR212W-A*, *YHR213W-B*, *YHR214C-D*, *YHR214C-E*) within a 12.7-kb region of chromosome VIII exhibit sequence similarity and synteny with four annotated genes (*YAR061W*, *YAR064W*, *YAR069C*, *YAR070C*) from chromosome I. These genes are likely part of a duplicated region shared between chromosomes I and VIII, corresponding to one of at least 53 segmental duplications present within the yeast genome[7].

A collection of new genes in yeast. In total, the 137 genes reported in this study represent an important supplement to the yeast genome as it is presently annotated. This gene set constitutes a rich source of short ORFs (ref. 23): 104 genes identified here are <100 codons in length. In addition to the fundamental biology associated with these genes[24], this collection expands available databases of short ORF sequences—a necessary step in order to train *ab initio* gene-finding software more effectively. The genes in this study also highlight an almost entirely overlooked class of sequences: genes nested antisense to other genes. Widely identified in prokaryotes[16,25], antisense genes have been detected in eukaryotes as well, both as protein-coding sequences[26,27] and as naturally occurring RNA molecules thought to regulate processes of transcription and translation[28]. Because the genes detected here are capable of being transcribed and translated as β-gal chimeras, this particular subset of antisense ORFs more likely encode proteins, potentially with novel cellular functions. One gene to be described in more detail elsewhere, *ART1*, is located antisense to the 25S ribosomal DNA (rDNA) locus and has recently been shown to encode a protein that localizes to mitochondria and complements point mutations of mitochondrial RNA polymerase I (P.C. Coelho *et al.*, pers. commun.).

The identification of this previously overlooked gene set offers insight relevant to the current debate regarding total gene number in

S. cerevisiae. As the genes identified here (by transposon-tagging) were drawn from data representing insertions within regions of DNA corresponding to ~40% of the yeast genome, at least an additional 150 undetected genes are probably present within the genome as a whole. Furthermore, we expect this estimate to be overly conservative, as methods of random transposon mutagenesis are less likely to identify short ORFs than longer ORFs. Interestingly, several studies have suggested that a significant number of presently annotated yeast genes are, in fact, spurious. Building upon the assumption that all genes possess similar sequence properties, Mackiewicz *et al.*[29] and Zhang *et al.*[30] have employed separate computational approaches to estimate the total population of yeast genes at 4,800 and 5,600 genes, respectively. Wood *et al.*[21] and Malpertuy *et al.*[31] have principally used comparative genomics to estimate a total gene count of ~5,600. We predict the presence of ~400 spurious yeast genes to be offset by the presence of an equal number of previously unappreciated genes, yielding a stable total population of ~6,000 genes in yeast. Considering that a similar percentage of genes have likely been overlooked and mis-annotated within other organisms, most sequenced eukaryotic genomes will require similar re-annotation within the immediate future.

Experimental protocol

Transposon tagging and β-gal assays. Shuttle mutagenesis was performed as described[15]. The Tn3-derived transposon, mTn-3xHA/*lacZ* (ref. 12), was used to mutagenize a plasmid-based library containing 50 genome equivalents of yeast DNA. Transposon-mutagenized DNA fragments were subsequently introduced into diploid strain Y800 (ref. 15) by lithium acetate–mediated transformation[32]. Yeast transformants carrying integrated transposon insertions were screened for β-gal activity using a filter-based chloroform lysis procedure described elsewhere[15]. β-gal assays were repeated for those strains predicted to contain a *lacZ*-fusion to a non-annotated ORF. Insertion alleles corresponding to each of these non-annotated ORFs were re-introduced into yeast, and three independent transformants were assayed for β-gal activity. Insertions resulting in β-gal production in at least two of three transformants were scored as "positive".

The genomic site of transposon insertion within each strain exhibiting β-gal activity was determined by directly sequencing its corresponding plasmid-borne insertion allele using a primer complementary to sequence from the 5′-end of mTn-3xHA/*lacZ*. Full-length ORF sequences were identified by extending each reading frame from the transposon insertion site downstream to the first observed stop codon and upstream to the farthest start codon. This process was automated in the custom program ORFSEEK (K. Cheung, unpublished data).

Immunolocalization of epitope-tagged proteins. Hemagglutinin (HA)-tagged proteins were analyzed by indirect immunofluorescence using mouse monoclonal anti-HA 16B12 (MMS101R, BAbCO; Eurogentec, Seraing, Belgium) and Cy3-conjugated goat anti-mouse IgG (Jackson Laboratories, Bar Harbor, ME).

RNA microarray analysis. RNA was extracted from strain Y800 grown to late-log phase in yeast extract–peptone–dextrose medium plus adenine (YPAD) using standard SDS/phenol-based protocols[33]. RNA preparations were treated with DNA-*free* (Ambion, Austin, TX) before isolation of poly(A) RNA by oligo(dT) cellulose column chromatography[33]. Poly(A) RNA was subsequently labeled using the BrightStar Psoralen-Biotin kit (Ambion). Biotinylated RNA was used to probe a microarray of 50- to 60-base oligonucleotides. For purposes of this expression analysis, oligonucleotide sequences were selected from the predicted coding sequence (both strands) of each putative gene, taking care to avoid sequence strongly similar to any nontarget regions of the yeast genome. All probe sequences were searched against yeast genomic DNA using BLAST[34]; alignments were generated using the PAM40 scoring matrix. Candidate oligonucleotide sequences exhibiting <75% identity (and no more than 15 contiguous identical bases) with any other segment of the yeast genome[17] were subsequently screened for nucleotide composition and secondary structure according to standard methods. Many gene-coding sequences identified in this study are <150 bp in length, offering limited flexibility in 50- to 60-mer oligonucleotide design. All oligonucleotides (both sense and antisense sequences) used here to identify new genes may be viewed online at bioinfo.mbb.yale.edu/genome/yeast/orfome/new-genes.

Oligonucleotides were arrayed onto nylon membrane CAST slides (Schleicher and Schuell, Keene, NH) using a flat-pin glass slide microarrayer (V&P Scientific, San Diego, CA). Oligonucleotides were arrayed at a spot-to-spot distance of 1 mm; double pin strikes were used to deposit 2 pmol oligonucleotides in a 40 nl volume. Between each spotting, arrayer pins were cleaned sequentially in 5% bleach, sterile water, and 100% ethanol after an initial 30 s sonication in sterile water. Spots exhibited a mean diameter of ~0.5 mm with a variability of <0.05 mm.

Arrayed oligonucleotides were immobilized by UV crosslinking at 120 mJ/cm². All hybridizations were carried out in buffer containing formamide at 45°C according to standard protocols. Arrays were hybridized with 200 ng biotinylated poly(A) RNA supplemented with denatured salmon sperm DNA at a final concentration of 100 μg/ml. Bound RNA was detected using the BrightStar BioDetect kit (Ambion).

Spot size and intensity were quantified with software distributed in the NIH Image package, version 1.62 (http://rsb.info.nih.gov/nih-image). All samples were arrayed and hybridized against biotinylated RNA a minimum of two times. Multiple arrays were normalized relative to each other using a set of control oligonucleotides included on each array; any local variations in background intensity were addressed manually. Oligonucleotide replicates yielded similar hybridization spot sizes and intensities (~5–10% variation in signal). The 2 pmol of oligonucleotides deposited per spot represent a molar excess as compared to predicted RNA transcript levels per gene; therefore, arrayer precision is not likely a critical factor affecting signal reproducibility, as a molar excess of oligonucleotide is typically present even if slightly reduced quantities are deposited per spot. After background subtraction and normalization, hybridization signals ranged from 24 to 3,986 (arbitrary units). Hybridization signals were scored as follows (Fig. 3): <500, no binding; 500–999, +; 1,000–1,499, ++; ≥1,500, +++. Note that these broad categories far exceed the observed variation in our samples. By this scoring strategy, we consider 20-fold enrichment over minimum signal to constitute "detectable" binding. From comparison with known transcript levels of genes included as controls in this study[35], we estimate that this level of detectable binding corresponds to ~0.2 RNA copies/cell (assuming that binding proceeds to completion). Central to this assumption, oligonucleotides must be in molar excess of target transcripts (as discussed); also, RNA binding must not be hindered by oligonucleotide secondary structure—a reasonable assumption considering the care taken in designing each oligonucleotide and the independent observation that 50-mer oligonucleotide probes yield results comparable to those obtained using 400-bp PCR probes[17].

Homology searching. Initially, we searched the SWISS–PROT protein sequence database[20] against the complete genomic sequence of *S. cerevisiae* strain S288c using the alignment program TFASTX (ref. 36). Low complexity was masked using the SEG algorithm[37]. All protein matches that overlapped annotated transposable elements were deleted. Significant protein matches (e-value <0.01 for a FASTA alignment) were reduced for mutual overlap by selecting homology segments in decreasing order of significance and flagging any others that overlap them for deletion. Matched stretches of genomic DNA were further examined by comparing to the matching protein a larger segment of the genomic DNA that had been extended at either end by the size of the matching protein sequence (in the equivalent number of nucleotides). These enlarged homology fragments were then extended into the most appropriate hORFs by searching for the nearest downstream stop codon and the farthest upstream start codon, while maintaining the correct reading frame.

In a second search for hORFs, additional matches were found using BLAST (ref. 34). We extracted all possible ORFs of size >29 codons from the yeast genome and searched them (in translation) against SWISS–PROT plus the combined annotated proteomes of *C. elegans, A. thaliana, D. melanogaster, S. cerevisiae* itself, and 18 prokaryotes. All significant protein matches (e-value <1×10⁻⁴ for a BLASTP alignment) were again selected and processed as above for the original searches to define additional hORFs, using TFASTX in the re-alignment stage.

Note: Supplementary information can be found on the Nature Biotechnology *website in Web Extras (http://biotech.nature.com/web_extras).*

1. Stein, L. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2**, 493–503 (2001).
2. Gopal, S. *et al.* Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. *Nat. Genet.* **27**, 337–340 (2001).
3. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster. Science* **287**, 2185–2195 (2000).
4. Reboul, J. *et al.* Open-reading frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans. Nat. Genet.* **27**, 332–336 (2001).
5. Shoemaker, D.D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
6. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–567 (1996).
7. Mewes, H.W. *et al.* Overview of the yeast genome. *Nature* **387 (Suppl.)**, 7–8 (1997).
8. Philippsen, P. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. *Nature* **387 (Suppl.)**, 93–98 (1997).
9. Velculescu, V.E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
10. Blandin, G. *et al.* Genomic exploration of the hemiascomycetous yeasts: 4. the genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett.* **487**, 31–36 (2000).
11. Cliften, P.F. *et al.* Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11**, 1175–1186 (2001).
12. Ross-Macdonald, P., Sheehan, A., Roeder, G.S. & Snyder, M. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA* **94**, 190–195 (1997).
13. Seifert, H.S., Chen, E.Y., So, M. & Heffron, F. Shuttle mutagenesis: a method of transposon mutagenesis for *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA* **83**, 735–739 (1986).
14. Ross-MacDonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).
15. Kumar, A., des Etages, S.A., Coelho, P.S.R., Roeder, G.S. & Snyder, M. High-throughput methods for the large-scale analysis of gene function by transposon tagging. *Methods Enzymol.* **328**, 550–574 (2000).
16. Selinger, D.W. *et al.* RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.* **18**, 1262–1268 (2000).
17. Kane, M.D. *et al.* Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **28**, 4552–4557 (2000).
18. Zhu, J. & Zhang, M.Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae. Bioinformatics* **15**, 607–611 (1999).
19. Sharp, P.M. & Li, W.H. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
20. Bairoch, A. & Apweiler, R. The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
21. Wood, V., Rutherford, K.M., Ivens, A., Rajandream, M.-A. & Barrell, B. A re-annotation of the *Saccharomyces cerevisiae* genome. *Comp. Funct. Genom.* **2**, 143–154 (2001).
22. Ball, C. *et al. Saccharomyces* Genome Database provides tools to survey gene expression and functional analysis data. *Nucleic Acids Res.* **29**, 80–81 (2001).
23. Basrai, M.A., Hieter, P. & Boeke, J.D. Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7**, 768–771 (1997).
24. Basrai, M.A., Velculescu, V.E., Kinzler, K.W. & Hieter, P. NORF5/HUG1 is a component of the MEC1-mediated checkpoint response to DNA damage and replication arrest in *Saccharomyces cerevisiae. Mol. Cell. Biol.* **19**, 7041–7049 (1999).
25. Wagner, E.G. & Simons, R.W. Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* **48**, 713–742 (1994).
26. Henikoff, S., Keene, M.A., Fechtel, K. & Fristrom, J.W. Gene within a gene: nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. *Cell* **44**, 33–42 (1986).
27. Spencer, C.A., Gietz, R.D. & Hodgetts, R.B. Overlapping transcription units in the dopa decarboxylase region of *Drosophila. Nature* **322**, 279–281 (1986).
28. Vanhee-Brossollet, C. & Vaquero, C. Do natural antisense transcripts make sense in eukaryotes? *Gene* **211**, 1–9 (1998).
29. Mackiewicz, P., Kowalczuk, M., Gierlik, A., Dudek, M.R. & Cebrat, S. Origin and properties of non-coding ORFs in the yeast genome. *Nucleic Acids Res.* **27**, 3503–3509 (1999).
30. Zhang, C.-T. & Wang, J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. *Nucleic Acids Res.* **28**, 2804–2814 (2000).
31. Malpertuy, A. *et al.* Genomic exploration of the hemiascomycetous yeasts: 19. *Ascomycetes*-specific genes. *FEBS Lett.* **487**, 113–121 (2000).
32. Ito, H., Fukuda, Y., Murata, K. & Kimura, A. Transformation of intact yeast cells treated with alkali cations. *J. Bacteriol.* **153**, 163–168 (1983).
33. Adams, A., Gottschling, D.E., Kaiser, C.A. & Stearns, T. *Methods in yeast genetics*, 1997 Edn. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; 1998).
34. Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
35. Holstege, F.C.P. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
36. Pearson, W.R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24–36 (1997).
37. Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285 (1994).