Revisiting the *Saccharomyces cerevisiae* predicted ORFeome

Qian-Ru Li,[1,6] Anne-Ruxandra Carvunis,[1,2,6] Haiyuan Yu,[1,6] Jing-Dong J. Han,[1,6,7] Quan Zhong,[1] Nicolas Simonis,[1] Stanley Tam,[1] Tong Hao,[1] Niels J. Klitgord,[1] Denis Dupuy,[1] Danny Mou,[1] Ilan Wapinski,[3,4] Aviv Regev,[3,5] David E. Hill,[1] Michael E. Cusick,[1] and Marc Vidal[1,8]

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; [2]TIMC-IMAG, CNRS UMR5525, Faculté de Médecine, 38706 La Tronche Cedex, France; [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; [4]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA; [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Accurately defining the coding potential of an organism, i.e., all protein-encoding open reading frames (ORFs) or "ORFeome," is a prerequisite to fully understand its biology. ORFeome annotation involves iterative computational predictions from genome sequences combined with experimental verifications. Here we reexamine a set of *Saccharomyces cerevisiae* "orphan" ORFs recently removed from the original ORFeome annotation due to lack of conservation across evolutionarily related yeast species. We show that many orphan ORFs produce detectable transcripts and/or translated products in various functional genomics and proteomics experiments. By combining a naïve Bayes model that predicts the likelihood of an ORF to encode a functional product with experimental verification of strand-specific transcripts, we argue that orphan ORFs should still remain candidates for functional ORFs. In support of this model, interstrain intraspecies genome sequence variation is lower across orphan ORFs than in intergenic regions, indicating that orphan ORFs endure functional constraints and resist deleterious mutations. We conclude that ORFs should be evaluated based on multiple levels of evidence and not be removed from ORFeome annotation solely based on low sequence conservation in other species. Rather, such ORFs might be important for micro-evolutionary divergence between species.

[Supplemental material is available online at www.genome.org.]

Comparative genomics, involving homology searching of genome sequences between evolutionarily related species, is a powerful tool for predicting functional regions in a genome sequence without prior biological knowledge. To date, complete genome sequences are available for more than 500 different organisms across all three domains of life (Liolios et al. 2006). Comparative genomics of bacteria, yeast, worm, fly, and human have led to extensive revision of complete sets of predicted protein-encoding open reading frames (ORFs), or "ORFeomes" (McClelland et al. 2000; Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003; Stein et al. 2003; Clamp et al. 2007; Clark et al. 2007). Removal from earlier versions of predicted ORFeomes of ORFs that are poorly or not conserved in other species ("orphan ORFs") is a critical revision proposed by these comparative genomic studies. The principle underlying removal of orphan ORFs is that selective constraints on functional DNA sequences should prevent deleterious mutations from occurring (Hardison 2003).

However, lack of evolutionary conservation does not guarantee lack of functional significance. It may be imprudent to eliminate putative ORFs from predicted ORFeomes solely based on lack of cross-species conservation. Different species, no matter how evolutionarily close, might express distinct ORF products. In support of this possibility, the pilot Encyclopedia of DNA Elements (ENCODE) project on 1% of the human genome has revealed that experimentally identified functional elements are not necessarily evolutionary constrained (Birney et al. 2007). In addition, although evolutionary conservation implies functionality for the product of a predicted ORF, experimental validation is required to demonstrate its biological significance. Therefore, cautious experimental reinvestigation of the functionality of predicted ORFs is needed to improve the accuracy of genome annotation.

To this end we set out to examine potential functionality of orphan ORFs in *Saccharomyces cerevisiae* based on available experimental evidence. Three independent comparative genomic analyses (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003) have predicted 648 annotated ORFs as "spurious" or "false," representing 10% of originally annotated ORFs. Notably, 10 out of these 648 orphan ORFs have since been validated as functional by small-scale experiments. For example, although YDR504C lacks clear orthologs in other yeast species, its deletion causes lethality upon exposure to high temperature while in stationary phase (Martinez et al. 2004). Given the time-consuming efforts of traditional "one-gene-at-a-time" inquiries, many predicted ORFs have not been individually characterized. However, as the first sequenced eukaryotic organism, *S. cerevisiae* has been used inten-

[6]These authors contributed equally to this work.
[7]Present address: Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China.
[8]Corresponding author.
E-mail marc_vidal@dfci.harvard.edu; fax (617) 632-5739.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.076661.108.

18:1294–1303 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

sively for functional genomics and proteomics studies, providing valuable functional evidence that allow further evaluation of coding potential of the orphan ORFs.

Using currently available functional genomics and proteomics data sets, we collate functional evidence for a significant portion of *S. cerevisiae* orphan ORFs, finding that many orphan ORFs produce detectable transcripts and/or translated products. Using a naïve Bayes model, we predict the likelihood that any *S. cerevisiae* ORF encodes a functional product and show that the number of orphan ORFs with potential functional significance is higher than expected by chance. Notably, we provide experimental verification for strand-specific transcription of many orphan ORFs. Finally, we report that interstrain intraspecies genome sequence variation is lower across orphan ORFs than in intergenic regions. Altogether our results demonstrate that orphan ORFs should not be excluded from current ORFeome annotation simply because they fail to show interspecies sequence conservation. We suggest that orphan ORFs should be included in future genome-wide experimental studies to reveal their bona fide identity either as functional ORFs or as randomly occurring misannotated ORFs.

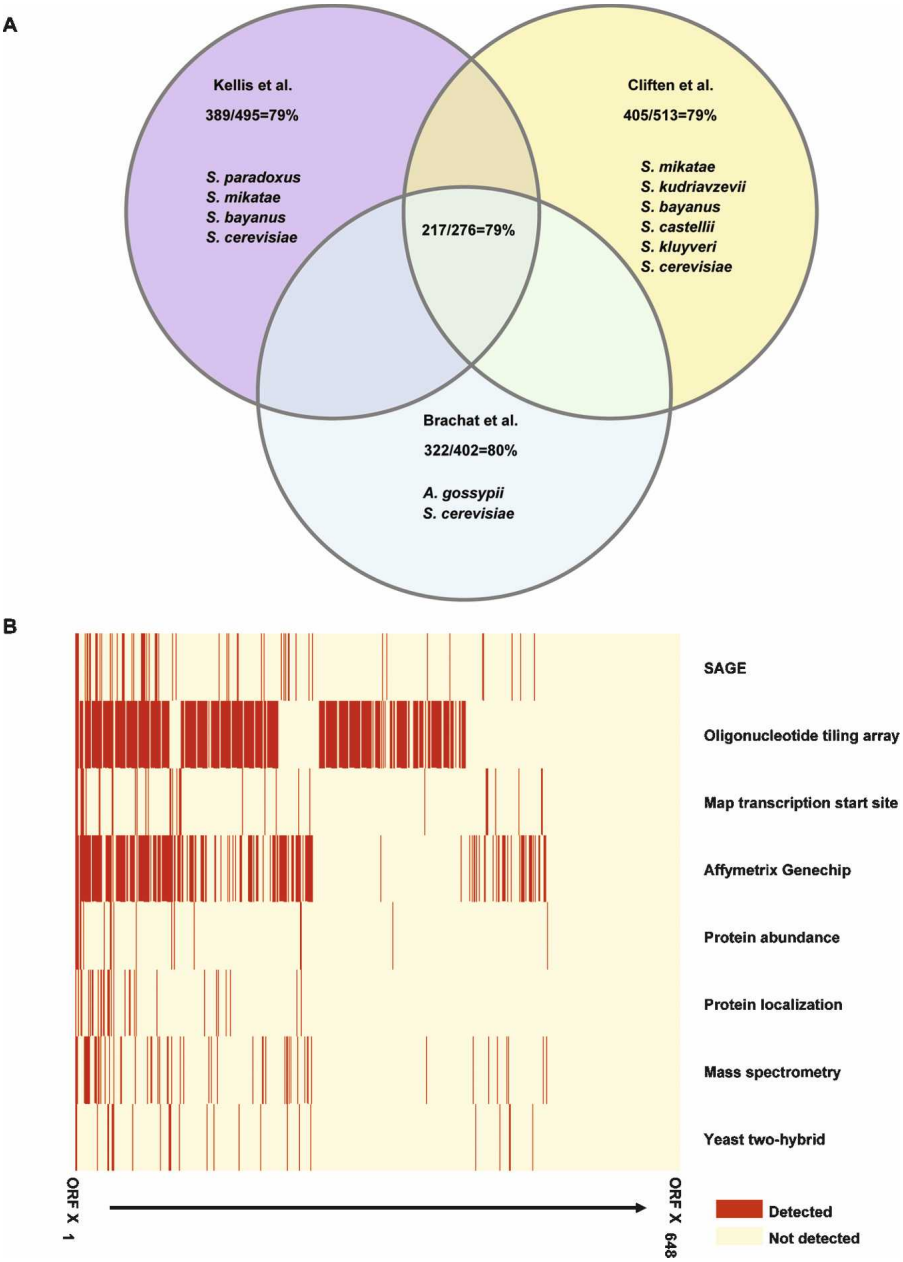

Figure 1. Experimental evidence for *S. cerevisiae* orphan ORFs. (*A*) Percentages indicate proportions of orphan ORFs detected at least in one of 13 functional genomics and proteomics data sets (Table 1). Note that ORFs rejected by all three comparative genomic studies analyzed here (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003) show similar percentages. (*B*) Supporting experimental evidence for each of 648 ORFs observed as orphan by three comparative genomic studies (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003). Complete lists of ORFs and supporting experimental evidence are in Supplemental Table 2. Columns are ordered from the ORF with most evidence (ORF X_1; *left*) to the one with the least evidence (ORF X_{648}; *right*). Data sets were grouped together by type of experimental approach, transcriptional on *top* and translational at the *bottom*. In total, there are 477 orphan ORFs with transcriptional evidence, 180 with translational evidence, and 145 with both transcriptional and translational evidence.

Results

Evidence for biological significance of *S. cerevisiae* orphan ORFs

The genome annotation of *S. cerevisiae* has undergone continuous modification through computational and experimental efforts since the original release in 1996 (Goffeau et al. 1996; Fisk et al. 2006). Three independent comparative genomic analyses compared the conservation of DNA or predicted protein sequences among several ascomycete species (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003), recommending that 402, 513, and 495 ORFs, respectively, be removed from the *S. cerevisiae* predicted ORFeome because their putative counterparts in other yeast species accumulate stop codons and frame-shift mutations (Fig. 1A). The union of these three comparative analyses is a set of 648 orphan ORFs called "spurious" or "false" in these studies (Fig. 1A).

High-throughput functional genomics and proteomics approaches have recently accelerated functional characterization of predicted ORFs. Several of these genome-wide approaches, such as gene-expression profiling or in vivo characterization of protein complexes, have detected transcripts or translated products of orphan ORFs. For example, in a proteome-wide purification of yeast protein complexes (Krogan et al. 2006), 85 proteins identified by mass spectrometry were encoded by orphan ORFs.

To provide a systematic reanalysis of *S. cerevisiae* orphan

ORFs, we collected 13 large-scale studies (Table 1) informing on either transcription or translation of orphan ORFs. The transcriptome studies included tiling arrays (David et al. 2006), high-density Affymetrix chip analysis (Holstege et al. 1998), SAGE analysis (Velculescu et al. 1997), and cDNA sequencing (Miura et al. 2006). Because many (69%) of the orphan ORFs overlap with another annotated ORF, we only included transcriptome studies able to detect strand-specific transcripts. Protein–protein interaction studies included proteome-scale yeast two-hybrid screens (Uetz et al. 2000; Ito et al. 2001) and affinity pull-downs of tagged proteins followed by mass spectrometry (Gavin et al. 2002, 2006; Ho et al. 2002; Krogan et al. 2006). For yeast two-hybrid studies, we considered an ORF being translated only if its product was involved in a protein–protein interaction as a prey. Protein expression studies included global surveys of protein abundance (Ghaemmaghami et al. 2003) and subcellular localization (Kumar et al. 2002; Huh et al. 2003).

Out of the 648 orphan ORFs, most (79%) have been detected in at least one of these data sets. The proportion of orphan ORFs detected was nearly the same for ORFs rejected by each of the three comparative genomics analyses independently (80% for Brachat, 79% for Cliften, and 79% for Kellis) and for the 276 orphan ORFs discarded by all three (79%) (Fig. 1A). Among the 648 orphan ORFs, many were detected by more than one approach. In total, 145 orphan ORFs (22%) were both detected as transcripts and translated products (Fig. 1B). A similar distribution of functional evidence was observed for the orphan ORFs rejected by all three comparative genomic analyses (Supplemental Fig. 1).

Evaluating biological significance of *S. cerevisiae* ORFs by a naïve Bayes approach

High-throughput approaches have inherently limited coverage (not all ORFs are detectable) and precision (detection of some ORFs might be artifactual). Therefore information from large-scale data sets needs to be accepted cautiously. We chose a naïve Bayes model to quantify the observations reported above, because this approach can integrate dissimilar types of data sets into a common probabilistic framework with maximal coverage and precision (Jansen et al. 2003; Yu et al. 2004). By use of such an integration scheme, evidence (i.e., features) from several data types can be accumulated to estimate with increasing confidence the likelihood that an ORF encodes a functional product.

As with any machine learning algorithm, naïve Bayes models need a training set of gold standard positives (GSPs) and negatives (GSNs). The *Saccharomyces* Genome Database (SGD), the arbiter of genome annotation for budding yeasts, has categorized all *S. cerevisiae* ORFs into three major groups based on conservation across species and on available experimental characterization: "verified" (4449 ORFs), "uncharacterized" (1333 ORFs), and "dubious" (823 ORFs) (Fisk et al. 2006). Both verified ORFs and uncharacterized ORFs are conserved across species. Verified ORFs have clear small-scale experimental evidence for the existence of functional ORF products, but uncharacterized ORFs do not. Dubious ORFs are thought not to encode a functional product due to (1) lack of conservation across species, and/or (2) absence of any small-scale experiment demonstrating detectable mRNA or protein production or phenotypic effects. We used all 4449 verified ORFs as the GSPs and all 823 dubious ORFs as the GSNs. Although ideally the GSNs should be depleted of functional ORFs, this cannot exactly be true for the dubious set. However, the dubious set is likely enriched with nonfunctional ORFs. It is common practice to use an "enriched" set of negatives in training data sets (Miller et al. 2005; Xia et al. 2006).

We calculated the ratio of the fraction of GSPs present in each of the 13 functional genomics and proteomics data sets divided by the fraction of GSNs present in each data set, which measures the confidence levels (Supplemental Table 1). The product of these ratios of the 13 data sets for each ORF is defined as the likelihood ratio (*LR*) of an ORF, i.e., the likelihood of each ORF to encode a functional product (see Methods). We used the base 10 logarithmic form of *LR* (*LLR*) as final prediction scores (Supplemental Table 2). Out of the large-scale studies integrated, several did measure similar biological features of ORFs and ORF products. However, we treated all 13 data sets as independent features, due to the low correlation between them (Supplemental Tables 3, 4).

To evaluate the performance of the naïve Bayes model, we used threefold cross-validation (see Methods). After randomly dividing both the GSPs and GSNs into three separate equal sets, we used two of the three sets as the training set to calculate *LLR*s and the remaining set as the test set to identify positives and negatives. The true-positive rate (TP rate: fraction of GSPs that are predicted to be functional) and the false-positive rate (FP rate: fraction of GSNs that are predicted to be functional) were calculated at different *LLR* cutoffs. The resulting couplets (TP rate–FP rate) were used to plot a receiver operating characteristic (ROC) curve. We ran this process three times so that each of the three sets was a test set and the remaining two constituted the training set. Each ROC curve looked similar (Supplemental Fig. 2), which

Table 1. Thirteen functional genomics and proteomics data sets integrated in our analysis

Functional genomics and proteomics data sets	Evidence detected	Approach category
Velculescu et al. 1997: Transcriptome characterized by SAGE	mRNA transcript	SAGE
David et al. 2006: Transcriptome characterized by oligonucleotide tiling array	mRNA transcript	Oligonucleotide tiling array
Miura et al. 2006: Full-length cDNA analysis	mRNA transcript	Map transcription start site
Holstege et al. 1998: Measurement of the transcripts abundance	mRNA transcript	Affymetrix GeneChip
Ghaemmaghami et al. 2003: Expression of TAP-tagged proteins	Protein expression	Protein abundance
Huh et al. 2003: Localization of GFP-tagged proteins	Protein localization	Protein localization
Kumar et al. 2002: Subcellular localization of transposon-tagged proteins	Protein localization	Protein localization
Gavin et al. 2002: Protein complexes characterization	Peptide sequence	Mass spectrometry
Ho et al. 2002: Protein complexes characterization	Peptide sequence	Mass spectrometry
Gavin et al. 2006: Protein complexes characterization	Peptide sequence	Mass spectrometry
Krogan et al. 2006: Protein complexes characterization	Peptide sequence	Mass spectrometry
Ito et al. 2001: Protein–protein interaction mapping by yeast two-hybrid	Protein physical interaction	Yeast two-hybrid
Uetz et al. 2000: Protein–protein interaction mapping by yeast two-hybrid	Protein physical interaction	Yeast two-hybrid

validated the overall quality of our training set. A final ROC curve was plotted by using potential *LLR* cutoffs from all three training subsets and their associated TP rate and FP rate based on the predictions from the complete training set (Fig. 2A). The significant deviation of the final ROC curve from the 45° random ROC line indicates that our model has substantial predictive value (area under ROC curve = 0.982). To assess the contribution of each data set to the final prediction scores, we successively omitted one data set and repeated the training and cross-validation procedures. We plotted ROC curves for all procedures (Supple-

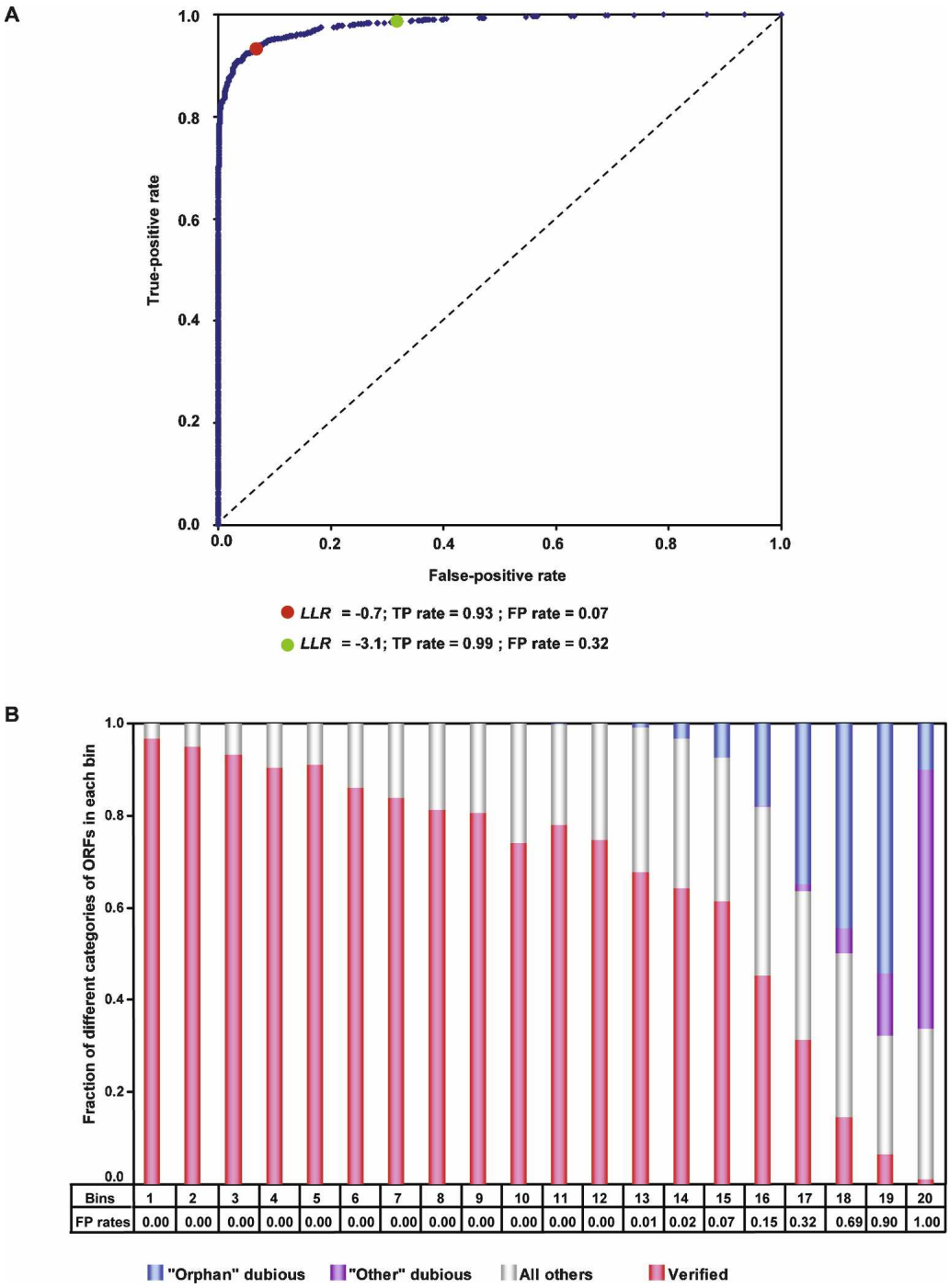

Figure 2. Evaluating functionality of *S. cerevisiae* ORFs. (*A*) ROC curve (blue) for naïve Bayes predictions based on 13 functional genomics and proteomics data sets. The diagonal (black dotted line) is the expected ROC curve for random, where the TP rate equals the FP rate. The two *LLR* cutoffs highlighted on the curve were used later as thresholds for categorizing orphan ORFs. (*B*) All 6718 *S. cerevisiae* ORFs were divided into 20 bins by decreasing *LLR*. Each bin has similar numbers of ORFs. The false-positive rates associated with the minimum *LLR* in each bin are listed. Distributions of verified ORFs, orphan dubious ORFs, "other" dubious ORFs, and all other ORFs in each bin are shown. Orphan dubious ORFs tend to have a higher *LLR* than ORFs classified as dubious for other reasons.

mental Fig. 3) and observed little difference when excluding any single data set. Thus it seems that no single data set dominates the prediction.

We divided all 6718 *S. cerevisiae* ORFs into 20 bins ranked by decreasing *LLR*, with each bin containing similar numbers of ORFs. Verified ORFs localized mostly in the higher *LLR* bins (92.5% of all verified ORFs distributed between bin 1 and bin 15), while dubious ORFs localized in lower *LLR* bins (only 4.98% of dubious ORFs distributed between bin 1 and bin 15) (Fig. 2B). Such segregation between verified ORFs and dubious ORFs was expected, given that the ORFs used in the training as GSPs (verified ORFs) are bound to have a higher *LLR* than the ones used in the training as GSNs (dubious ORFs). An unanticipated result of the naïve Bayes predictions is that orphan dubious ORFs have overall higher *LLR* ($P < 10^{-15}$ by Mann-Whitney *U* test) (Fig. 2B) than ORFs classified as dubious for reasons other than strict lack of interspecies sequence conservation (e.g., a mutant phenotype described for the ORF could be ascribed to mutation of an overlapping well-characterized ORF) (Fisk et al. 2006). This suggests that orphan dubious ORFs might be more likely to encode functional products than "other" dubious ORFs.

For an ORF to be considered "most-likely" functional in our naïve Bayes predictions, its posterior odds (the product of the prior odds and the likelihood ratio) has to be larger than 1 (see Methods). We can estimate that the prior odds for any given ORF to be most-likely functional is ~5.4 (4449 GSPs divided by 823 GSNs). Hence, we used $LLR = \log_{10}(1/5.4) = -0.7$ (FP rate = 0.07) as the cutoff for an ORF to be most-likely functional (bins 1–15). Among the 648 orphan ORFs, 54 ORFs with $LLR \geq -0.7$ were thus assigned to a set of most-likely functional orphan ORFs. Although the percentage of verified ORFs decreased significantly from bin 16 to bin 20 compared with the first 15 bins (Fig. 2B), there were still 3.4% and 2.5% of verified ORFs (152 and 111 ORFs) in bins 16 and 17, respectively. We classified the 199 orphan ORFs in bins 16 and 17, with an *LLR* between -0.7 (FP rate = 0.07) and -3.1 (FP rate = 0.32), as "moderately-likely" to encode a functional product. The remaining 395 orphan ORFs distributed between bins 18 and 20 were called "least-likely" functional ORFs. Detectability limitations in the large-scale data sets integrated in our predictions may have biased against these least-likely ORFs. Integration of new lines of experimental evidence in the future could still potentially identify promising functional ORF candidates among the least-likely ORFs.

Experimental evidence for expression of *S. cerevisiae* orphan ORFs

We next experimentally measured mRNA expression for orphan ORFs using reverse transcription–polymerase chain reaction (RT-PCR) (Fig. 3A). Strand specificity was needed to ensure that the transcripts detected were transcribed from the predicted DNA strand and to exclude artifacts caused from read-through transcription on the opposite strand (Craggs et al. 2001).

We tested strand specificity on two verified *S. cerevisiae* ORFs that both contain introns: YER133W (*GLC7*) and YBR078W (*ECM33*) (see Methods). Given the presence of introns in these ORFs, the sense-strand transcripts should be appreciably shorter in length than the antisense-strand transcripts. Spliced transcripts of the expected sizes were obtained in reactions where strand-specific primer was added for cDNA synthesis (Fig. 3B). No RT-PCR products were obtained in reactions without RT, demonstrating absence of contaminating genomic DNA in the poly(A) mRNA template preparation. No RT-PCR products were observed in the absence of cDNA primer for first-strand cDNA synthesis, demonstrating that the second step of standard PCR amplification contained no active reverse transcriptase for the synthesis of incorrect strand cDNA from antisense strand–specific primer. The identities of RT-PCR products were confirmed by sequencing.

Thereafter we applied our strand-specific RT-PCR to 201 orphan ORFs that do not overlap any other annotated ORF. The requirement for nonoverlap further reduces the false-positive rate, because it is less likely that there would be any transcription from the incorrect strand. Among 201 nonoverlapping orphan ORFs tested under conditions of growth on rich media, RT-PCR products of expected size were obtained for 105 ORFs (Supplemental Table 2). Although the available supporting experimental evidence for these 105 ORFs is not strikingly different from the ORFs whose transcripts were not detected by strand-specific RT-PCR (Supplemental Fig. 4), the detected ORFs have a significantly higher average *LLR* (-3.4 ± 0.2) than the ones undetected by RT-PCR (-3.8 ± 0.2, $P = 0.03$ by Mann-Whitney *U* test) (Fig. 3C), demonstrating the validity and robustness of our predictions for positives. In particular, YJL199C, a dubious ORF, has the highest *LLR* among 201 tested ORFs and was detected by RT-PCR. YJL199C was recently predicted to encode a metabolic protein based on large-scale protein–protein interaction studies (Samanta and Liang 2003).

Notably, out of 49 orphan ORFs tested that had not been detected by any of the 13 data sets (Table 1), 29 were expressed (Supplemental Table 2), among which YPR096C was recently found to encode a ribosome-interacting protein (Fleischer et al. 2006) and YOR235W was shown through a genome-wide phenotypic analysis to be involved in DNA recombination events (Alvaro et al. 2007). Therefore, we suggest that more experimentation is needed before rejecting ORFs from the *S. cerevisiae* ORFeome annotation.

Interstrain intraspecies sequence conservation for *S. cerevisiae* orphan ORFs

The available experimental evidence from large-scale data sets, combined with our experimental support for many orphan ORFs, implies that lack of interspecies conservation does not necessarily dispel the bona fide functionality of an ORF. Functional orphan ORFs may have a relaxed selective constraint due to their dispensable roles in other species and may therefore rapidly lose sequence similarity even in closely related species (Schmid and Aquadro 2001). However, select species–specific functions may stringently constrain sequence divergence of functional orphan ORFs within species (Domazet-Loso and Tautz 2003). Therefore, we examined the intraspecies conservation of orphan ORFs in *S. cerevisiae,* using single nucleotide polymorphism (SNP) information from genome resequencing of multiple strains of *S. cerevisiae* by the *Saccharomyces* Genome Resequencing Project (SGRP) (http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml). Among the 37 currently available strain sequences, four (SK1, W303, Y55, and DBVPG6765) have been sequenced at twofold coverage or higher. We used the SNP data from these four genomes to assess nucleotide variation in different genomic regions across *S. cerevisiae* strains. We compared nucleotide divergence among three genomic features: orphan ORFs, nonorphan ORFs, and intergenic regions, considering only the regions that do not overlap with any other annotated ORF (see Methods).

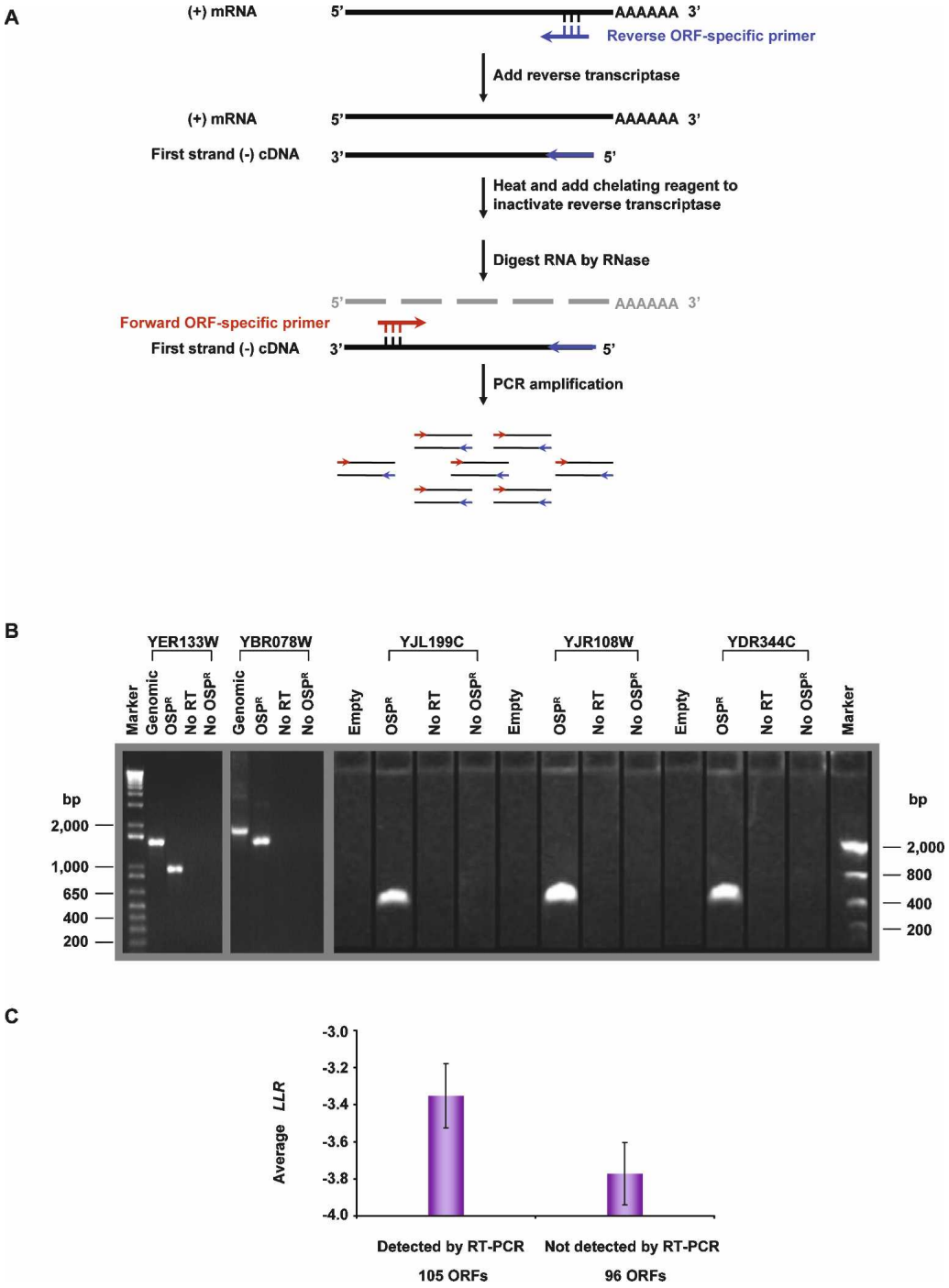

Figure 3. Two-step strand-specific RT-PCR. (*A*) Schematic diagram of the strand-specific RT-PCR procedure. (*B*) Electrophoretic analysis of strand-specific RT-PCR products. Reverse ORF-specific primers (OSP^R), with sequences complementary to the ORF-coding strand, were used for first-strand cDNA synthesis. Second-step PCR amplifications used a pair of forward (OSP^F) and reverse ORF-specific primers (OSP^R). As controls, the first step of RT-PCR was performed without reverse transcriptase for detecting contamination by genomic DNA, or without the OSP^R primer for detecting residual reverse transcriptase activity in second-step PCR reactions. Two intron-containing verified ORFs, YER133W (genomic DNA length: 1464 bp; coding sequence length: 939 bp) and YBR078W (genomic DNA length: 1737 bp; coding sequence length: 1407 bp), were used to test the strand specificity. An extra control for these two verified ORFs was a standard PCR action using yeast genomic DNA as template and the same pair of ORF-specific primers. The observed difference in the length of PCR products amplified from genomic DNA versus poly(*A*) mRNA manifested the strand specificity. Strand-specific RT-PCR results of 201 nonoverlapping orphan ORFs were analyzed on 1% agarose E-gel (Invitrogen). Of the reactions 53% (105 ORFs) gave rise to visible RT-PCR products of the expected sizes. Three orphan ORFs, YJL199C (327 bp), YJR108W (372 bp), and YDR344C (444 bp), are shown as examples of successful RT-PCR reactions. (*C*) Comparison of the average *LLR* between nonoverlapping ORFs detected and undetected by strand-specific RT-PCR. Error bars, SEM.

Across the four strains analyzed, orphan ORFs showed higher nucleotide divergence (7.0 ± 0.4 SNPs per kb) than did nonorphan ORFs (3.7 ± 0.1 SNPs per kb, $P < 10^{-5}$ by Mann-Whitney *U* test), but less than intergenic regions (15.5 ± 0.2 SNPs per kb, $P < 10^{-15}$ by Mann-Whitney *U* test) (Fig. 4A). Such intermediate nucleotide divergence for orphan ORFs suggests that at least a portion of them are subject to significant intraspecies evolutionary constraints. Such "interstrain intraspecies" conservation of orphan ORFs indicates potential functionality of an ORF in addition to experimental evidence.

Among the 648 orphan ORFs, the most-likely functional ones displayed a significantly lower nucleotide divergence (3.8 ± 0.7 SNPs per kb) than both moderately-likely (6.4 ± 0.8 SNPs per kb, $P = 0.016$ by Mann-Whitney *U* test) and least-likely ORFs (7.7 ± 0.6 SNPs per kb, $P = 0.005$ by Mann-Whitney *U* test) (Fig. 4B). Although the moderately-likely category does have a lower nucleotide divergence than least-likely category, the difference is not significant ($P > 0.05$ by Mann-Whitney *U* test). Because different types of SNPs, such as synonymous or nonsynonymous substitutions, could have distinct effects on an ORF product, we applied another test to compare sequence conservation among the three groups, measuring the percentage of ORFs with preserved reading frames (absence of stop codons or frameshift mutations) across all four *S. cerevisiae* strains. A decreasing trend was observed from most-likely to least-likely ORFs (Fig. 4C), with significant differences among the three categories

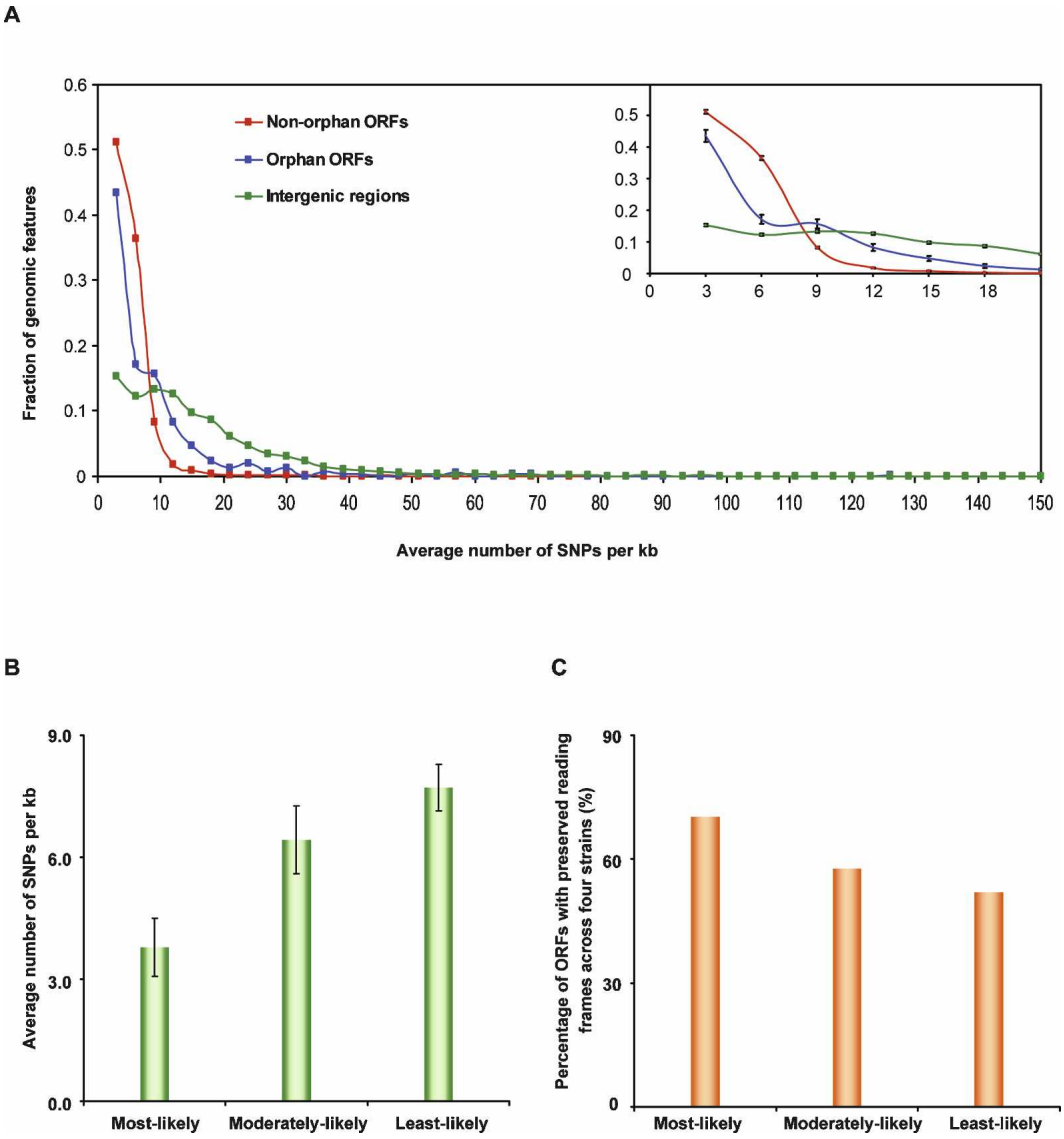

Figure 4. Interstrain intraspecies sequence conservation for orphan ORFs. (*A*) Distribution of nucleotide divergence in different genomic features. We binned three types of genomic features, (1) non-orphan ORFs (red curve), (2) orphan ORFs predicted by three comparative genomic analyses (blue curve) (Brachat et al. 2003; Cliften et al. 2003; Kellis et al. 2003), and (3) intergenic regions (green curve), using a window of an average three SNPs per kb across four *S. cerevisiae* strains. Each dot represents the fraction of genomic features in each bin. Numbers on the *X*-axis represent the maximum number of SNPs per kb in each bin. For instance the first bin collects the genomic regions that have between zero and three SNPs per kb in four strains. The inset zooms in on the 0–21 SNPs per kb range with SEM displayed. (*B*) Comparison of nucleotide divergence among three predicted categories of orphan ORFs based on their *LLR*s. Error bars, SEM in each category. (*C*) Comparison of the percentage of ORFs among the three predicted categories of orphan ORFs that have reading frames preserved across four *S. cerevisiae* strains.

($P = 0.03$ by χ^2 test). The coexistence of high interstrain intraspecies conservation with high likelihood of functionality demonstrates that some orphan ORFs face functional constraints that protect them from deleterious intraspecies mutations.

In summary, analysis of nucleotide variation in multiple *S. cerevisiae* strains, combined with multiple lines of experimental evidence, suggest that reevaluation of the functionality of all ORFs, especially orphan ORFs, is warranted.

Discussion

We report here that many interspecies nonconserved ORFs or orphan ORFs predicted by comparative genomic analyses in *S. cerevisiae* show evidence of transcription or translation, as reported in various functional genomics or proteomics data sets. We used a naïve Bayes probabilistic integration of a heterogeneous set of large-scale data sets to predict the likelihood that a predicted ORF encodes a functional product. Threefold cross-validation demonstrated high performance for this approach, which revealed that orphan ORFs are more likely functional than are ORFs classified as dubious for reasons other than strict lack of sequence conservation across species. Independent strand-specific RT-PCR confirmed that many orphan ORFs are indeed expressed. Although presence of transcripts is not sufficient by itself to conclude that an ORF encodes a functional product, the correspondence between our RT-PCR results and naïve Bayes prediction scores demonstrated both the potential functionality of orphan ORFs and the robustness of our prediction method. Confirming that orphan ORFs could be functional, many show signs of interstrain intraspecies negative selection, such as lower nucleotide divergence than intergenic regions and retaining an intact reading frame in multiple *S. cerevisiae* strains.

Collectively our findings argue that the likelihood that an ORF encodes a functional product is best evaluated by combining multiple lines of experimental and evolutionary evidence (Snyder and Gerstein 2003). The potential functionality of orphan ORFs in *S. cerevisiae* suggests that experimentally verified functional sequences are not always conserved across species. Such nonconserved functional sequences might be responsible for species-specific phenotypic differences, making *S. cerevisiae* "cerevisiae" and not some other species in the *Saccharomyces* genus. An alternative explanation is that there are some functional elements evolving neutrally and conferring no specific benefit to the organism (Birney et al. 2007). Either way, experimental investigation has an irreplaceable role in determining biologically relevant DNA sequences. Comparative genomics has demonstrated analytic power in predicting functional regions before availability of any experimental information (Hardison 2003). When experimental information does become available (mainly from high-throughput functional genomics and proteomics analyses), then its integration should revise the genome annotation accordingly. The naïve Bayes model implemented here can be readily applied to all organisms.

Although we provide confidence scores about the likelihood of a predicted ORF to encode a functional product, comprehensive functional characterization of an ORF needs more concrete evidence from genetics, cell biology, and biochemistry than simple evidence of transcription or translation. The functional genomics or proteomics data sets used in our naïve Bayes predictions only investigated a few growth conditions, generally growth on rich media, limiting investigation of functions unique to the development and physiology of *S. cerevisiae*. Given the limited functional information obtained so far under laboratory conditions about uncharacterized ORFs (Pena-Castillo and Hughes 2007), perhaps what is needed are studies of yeast cells outside the laboratory. Upon such a shift, data sets generated under diverse conditions will become available, and our approach will then be available to aid precise and powerful annotation of genomes.

Methods

Large-scale data sets analysis

We collected 13 published functional genomics and proteomics data sets of *S. cerevisiae*, summarized in Table 1 with references to the data sources. Only ORFs identified by the same primary SGDID in the publication and in the January 2007 version of SGD annotation were included. We assigned "presence" or "absence" of transcript or translated product of every ORF in each data set. For protein complexes characterization data sets (Gavin et al. 2002, 2006; Ho et al. 2002; Krogan et al. 2006) all proteins that were identified as peptides were considered "present," independent of further filtration by the investigators. For high-throughput yeast two-hybrid (Uetz et al. 2000; Ito et al. 2001), only proteins identified as preys were considered present. Only protein–protein interactions classified as "core" by Ito et al. (2001) were included. Transcripts identified by SAGE (Velculescu et al. 1997) and assigned to "class 1" by the investigators were considered present; all others, absent. We divided the Affymetrix Genechip data (Holstege et al. 1998) into two groups: intensity of expression strictly positive but less than or equal to 1, and intensity strictly more than 1. These two groups were treated separately in the naïve Bayes model. The normalized intensity of expression per probe (David et al. 2006) was averaged, and the percentage of probes whose intensity was higher than this average was considered as the intensity of expression of each ORF. We then extracted four groups (undetected, intensity strictly positive but less than 0.4, intensity strictly more than or equal to 0.4 but less than 0.8, and intensity strictly more than or equal to 0.8) that were treated separately in the naïve Bayes model. The remaining data sets were not reprocessed.

The naïve Bayes model

If the numbers of positives are known among the total number of ORFs, the "prior" odds of finding a positive are

$$O_{prior} = \frac{P(pos)}{P(neg)} = \frac{P(pos)}{1 - P(pos)}.$$

The "posterior" odds are the odds of finding a positive after considering N different feature data sets with values $f_1 \ldots f_N$:

$$O_{post} = \frac{P(pos|f_1 \ldots f_N)}{P(neg|f_1 \ldots f_N)}.$$

The likelihood ratio LR is defined as

$$LR(f_1 \ldots f_N) = \frac{P(f_1 \ldots f_N|pos)}{P(f_1 \ldots f_N|neg)}.$$

According to Bayes rule, the posterior odds can be expressed as

$$O_{post} = LR(f_1 \ldots f_N)O_{prior}.$$

If the N features are conditionally independent, LR can be simplified to

$$LR(f_1 \ldots f_N) = \prod_{i=1}^{N} L(f_i) = \prod_{i=1}^{N} \frac{P(f_i|pos)}{P(f_i|neg)}.$$

LR can be computed from contingency tables relating positive and negative examples with the *N* features (we binned the feature values $f_1 \ldots f_N$ into discrete intervals). Since O_{prior} is a fixed value, O_{post} is determined by *LR*. We used log-likelihood ratio ($\log_{10} LR$ or *LLR*) as the final prediction score. The higher the *LLR* of a certain ORF, the more likely it is a positive, i.e., a functional ORF.

Threefold cross-validation

We divided the whole training set into three subsets randomly. We then trained the model with two subsets and tested its performance on the third subset. We repeated this step three times so that each subset was used once to test the performance. We calculated the ROC curve with the predictions for the whole training set by combining the results from the three repeated tests.

Strand-specific RT-PCR

S. cerevisiae strain S288C was grown in yeast extract-peptone-dextrose (YPD) medium at 30°C to mid-exponential phase. Yeast cells were then harvested and used for total RNA isolation with an RNeasy kit (Qiagen). Poly(A) RNA was subsequently enriched by Oligotex mRNA kit (Qiagen). Before RT-PCR experiments, Poly(A) RNA was subjected to DNA-free DNase treatment (Ambion) to eliminate genomic DNA contamination. Genomic DNA was extracted from yeast culture by the DNeasy blood and tissue kit (Qiagen). We modified a strand-specific RT-PCR method previously described (Craggs et al. 2001), using the GeneAmp thermostable rTth reverse transcriptase RNA PCR kit (Applied Biosystems). DNase-treated poly(A) RNA sample (25 ng) was denatured for 5 min at 70°C with 2 µL of 10× rTth reverse transcriptase buffer and 1 µL of 10 µM reverse ORF-specific primer complementary to the ORF-coding strand (OSP^R). While the template and the primer were still incubating at 70°C, a preheated reaction mixture was added consisting of 2 µL of 10 mM $MnCl_2$ solution, 1.6 µL of 10 mM dNTP mix, and 2.5U of rTth polymerase. The temperature was lowered for 2 min to 55°C for annealing and then raised for 30 min to 70°C for the first-strand cDNA synthesis. After the cDNA synthesis, 20 µL of prewarmed 1× chelating buffer was added to chelate Mn^{2+} followed by heating the mixture for 30 min at 98°C to inactivate the reverse transcriptase activity of rTth. Second-step PCR reactions were performed in a 50-µL reaction volume using one-tenth of the synthesized first-strand cDNA as template, forward ORF-specific primer (OSP^F) and OSP^R as primers, and one unit of High Fidelity Platinum Taq polymerase (Invitrogen). The OSP^R complementary to the ORF-coding strand was used in both first-strand cDNA synthesis and second-step PCR amplification. The OSP^F complementary to the opposite strand was used only in the second-step PCR amplification. Both OSP^R and OSP^F were designed using the OSP Program (Hillier and Green 1991). The OSP^R starts from the last nucleotide of the termination codon, while the OSP^F starts from A of the ATG initiation codon. Primers used for RT-PCR of 201 nonoverlapping orphan ORFs are listed in Supplemental Table 5.

Interstrain intraspecies conservation analysis

SNP information from the four strains SK1, Y55, DBVPG6765, and W303 were extracted from the website of the Sanger Institute *Saccharomyces* Genome Resequencing Project (http://www.sanger.ac.uk/Teams/Team71/durbin/) on September 18, 2007 (R. Durbin and E. Louis, pers. comm.). The preassembly SNPs were taken into account only when their quality was "confirmed." They were mapped to the ORFeome of the reference strain S288C as annotated by SGD on January 2007, as well as to intergenic regions that are annotated as "not feature" (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/intergenic/NotFeature.fasta.gz). The nucleotide divergence of each ORF was then computed by averaging the number of SNPs per kb found in each of the four strains, counting insertions and deletions as one event independently of their length. For overlapping ORFs, only the regions unique to the ORFs themselves were considered for counting SNPs. To be considered as a preserved reading frame in our analysis, the ORF had to show neither stop codons nor frame-shift mutations in any of the four strains. The reading frame of an ORF was not considered preserved if the ORF had an insertion or deletion (indel) longer or equal to 20 bp, no matter whether the indel caused a frame-shift or not.

Acknowledgments

We thank R. Durbin and E. Louis for providing SNP information and F. Roth (HMS) for helpful discussions. We thank the members of the Vidal Lab and the Center for Cancer Systems Biology (CCSB) for their scientific and technical support, especially M. Boxem, K. Venkatesan, M. Yildirim, K. Salehi-Ashtiani, M. Dreze, S. Milstein, and C. Fraughton. This work was supported by an Ellison Foundation grant awarded to M.V. and by Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative awarded to M.V. and CCSB.

References

Alvaro, D., Lisby, M., and Rothstein, R. 2007. Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. *PLoS Genet.* **3:** e228. doi: 10.1371/journal.pgen.0030228.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Brachat, S., Dietrich, F., Voegeli, S., Zhang, Z., Stuart, L., Lerch, A., Gates, K., Gaffney, T., and Philippsen, P. 2003. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol.* **4:** R45. doi: 10.1186/gb-2003-4-7-r45.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., and Lander, E.S. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci.* **104:** 19428–19433.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Craggs, J.K., Ball, J.K., Thomson, B.J., Irving, W.L., and Grabowska, A.M. 2001. Development of a strand-specific RT-PCR based assay to detect the replicative form of hepatitis C virus RNA. *J. Virol. Methods* **94:** 111–120.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. 2006. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci.* **103:** 5320–5325.

Domazet-Loso, T. and Tautz, D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res.* **13:** 2213–2219.

Fisk, D.G., Ball, C.A., Dolinski, K., Engel, S.R., Hong, E.L., Issel-Tarver, L., Schwartz, K., Sethuraman, A., Botstein, D., Cherry, J.M., et al. 2006. *Saccharomyces cerevisiae* S288C genome annotation: A working hypothesis. *Yeast* **23:** 857–865.

Fleischer, T.C., Weaver, C.M., McAfee, K.J., Jennings, J.L., and Link, A.J. 2006. Systematic identification and functional screens of

uncharacterized proteins associated with eukaryotic ribosomal complexes. *Genes & Dev.* **20:** 1294–1307.
Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141–147.
Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636.
Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., and Weissman, J.S. 2003. Global analysis of protein expression in yeast. *Nature* **425:** 737–741.
Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546–567.
Hardison, R.C. 2003. Comparative genomics. *PLoS Biol.* **1:** e58. doi: 10.1371/journal.pbio.0000058.
Hillier, L. and Green, P. 1991. OSP: A computer program for choosing PCR and DNA sequencing primers. *PCR Methods Appl.* **1:** 124–128.
Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415:** 180–183.
Holstege, F.C.P., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717–728.
Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. 2003. Global analysis of protein localization in budding yeast. *Nature* **425:** 686–691.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.* **98:** 4569–4574.
Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302:** 449–453.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440:** 637–643.
Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., et al. 2002. Subcellular localization of the yeast proteome. *Genes & Dev.* **16:** 707–719.
Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N.C. 2006. The Genomes On Line Database (GOLD) v.2: A monitor of genome projects worldwide. *Nucleic Acids Res.* **34:** D332–D334.
Martinez, M.J., Roy, S., Archuletta, A.B., Wentzell, P.D., Anna-Arriola, S.S., Rodriguez, A.L., Aragon, A.D., Quinones, G.A., Allen, C., and Werner-Washburne, M. 2004. Genomic analysis of stationary-phase and exit in *Saccharomyces cerevisiae*: Gene expression and identification of novel essential genes. *Mol. Biol. Cell* **15:** 5295–5305.
McClelland, M., Florea, L., Sanderson, K., Clifton, S.W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R.K., and Miller, W. 2000. Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res.* **28:** 4974–4986.
Miller, J.P., Lo, R.S., Ben-Hur, A., Desmarais, C., Stagljar, I., Noble, W.S., and Fields, S. 2005. Large-scale identification of yeast integral membrane protein interactions. *Proc. Natl. Acad. Sci.* **102:** 12123–12128.
Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., and Ito, T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci.* **103:** 17846–17851.
Pena-Castillo, L. and Hughes, T.R. 2007. Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176:** 7–14.
Samanta, M.P. and Liang, S. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci.* **100:** 12579–12583.
Schmid, K.J. and Aquadro, C.F. 2001. The evolutionary analysis of "orphans" from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics* **159:** 589–598.
Snyder, M. and Gerstein, M. 2003. Defining genes in the genomics era. *Science* **300:** 258–260.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1:** e45. doi: 10.1371/journal.pbio.0000045.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.
Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243–251.
Xia, Y., Lu, L.J., and Gerstein, M. 2006. Integrated prediction of the helical membrane protein interactome in yeast. *J. Mol. Biol.* **357:** 339–349.
Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. 2004. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14:** 1107–1118.

Received January 29, 2008; accepted in revised form May 5, 2008.