A high-resolution map of transcription in the yeast genome

Lior David*†, Wolfgang Huber†‡, Marina Granovskaia§, Joern Toedling‡, Curtis J. Palm*, Lee Bofkin‡, Ted Jones*, Ronald W. Davis*¶, and Lars M. Steinmetz*§¶

*Stanford Genome Technology Center and Department of Biochemistry, Stanford University, Palo Alto, CA 94304; ‡European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge CB10 1SD, United Kingdom; and §European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Contributed by Ronald W. Davis, February 10, 2006

There is abundant transcription from eukaryotic genomes unaccounted for by protein coding genes. A high-resolution genome-wide survey of transcription in a well annotated genome will help relate transcriptional complexity to function. By quantifying RNA expression on both strands of the complete genome of *Saccharomyces cerevisiae* using a high-density oligonucleotide tiling array, this study identifies the boundary, structure, and level of coding and noncoding transcripts. A total of 85% of the genome is expressed in rich media. Apart from expected transcripts, we found operon-like transcripts, transcripts from neighboring genes not separated by intergenic regions, and genes with complex transcriptional architecture where different parts of the same gene are expressed at different levels. We mapped the positions of 3′ and 5′ UTRs of coding genes and identified hundreds of RNA transcripts distinct from annotated genes. These nonannotated transcripts, on average, have lower sequence conservation and lower rates of deletion phenotype than protein coding genes. Many other transcripts overlap known genes in antisense orientation, and for these pairs global correlations were discovered: UTR lengths correlated with gene function, localization, and requirements for regulation; antisense transcripts overlapped 3′ UTRs more than 5′ UTRs; UTRs with overlapping antisense tended to be longer; and the presence of antisense associated with gene function. These findings may suggest a regulatory role of antisense transcription in *S. cerevisiae*. Moreover, the data show that even this well studied genome has transcriptional complexity far beyond current annotation.

tiling array | transcriptone survey | gene architecture | segmentation | antisense regulation

Proteins constitute most structural and functional components of cells. The assumption has been that protein-encoding genes are also the main controllers of cellular processes. Recent evidence challenges this assumption, suggesting a wide-spread involvement of noncoding RNA in regulation, including through the activity of untranslated regions of mRNAs (1), antisense transcripts (2, 3), and isolated noncoding RNAs such as microRNA that control transcript levels or their translation (4).

High-resolution transcriptome analysis in higher eukaryotes using tiling arrays has improved ORF annotations and exon-intron predictions and discovered many new transcripts of currently unknown function (5–7). However, these studies have encountered challenges, due to noise, limited resolution, lack of strand-specific signal, and drawbacks in the analysis methods (8). Sequencing of cloned cDNAs has also revealed a high level of transcriptional complexity, including the presence of many new transcripts, alternative promoter usage, splicing, and polyadenylation, as well as the presence of many sense–antisense transcript pairs (3, 9). However, because of the cost and labor of large-scale sequencing, this approach has been limited. Therefore, there is a need to develop high-throughput, precise, and high-resolution technology to map the full transcriptional activity. Yeast is a simple and relatively small eukaryotic genome that provides opportunities to rapidly characterize novel findings.

We developed an oligonucleotide array for *Saccharomyces cerevisiae* that contains 6.5 million probes and interrogates both strands of the full genomic sequence with 25-mer probes tiled at an average of eight nucleotide intervals on each strand (17 nucleotides overlap) and a four nucleotides offset of the tile between strands. This design enables a 4-nt resolution for hybridization of double stranded targets and an 8-nt resolution for strand-specific targets. We profiled transcription during exponential growth in rich media, the standard laboratory growth condition, to generate a comprehensive map of transcription.

Results and Discussion

Microarray Experiments and Analysis. We hybridized first-strand cDNA synthesized using random primers from polyadenylated [poly(A)] and total RNA. To calibrate the sequence-specific probe effect (10–12), we background-corrected and adjusted (13) the signal of each probe by sequence-specific parameters, estimated from a calibration set of genomic DNA hybridizations (Fig. 5, which is published as supporting information on the PNAS web site). This method allowed us to quantitatively compare the signal from probe to probe on the array.

The Transcriptome. To address the question of how much of the genome is transcribed, we analyzed the coding regions of 5,654 ORFs that were annotated as verified or uncharacterized genes in the *Saccharomyces* Genome Database (SGD, www.yeastgenome.org) and represented by unique probes on the array. Significant expression above background was detected for 5,104 ORFs (90%) (Binomial test, false discovery rate = 0.001; Fig. 6, which is published as supporting information on the PNAS web site). As expected, genes that were not detected have functions not required in this condition such as meiosis, sporulation, mating, sugar transport, and vitamin metabolism [hypergeometric test for gene ontology (GO) annotation enrichment, unadjusted $P \leq 3 \times 10^{-9}$]. In addition, analyzing 11,412,997 bp of unique genomic sequence, we detected expression above background on either strand for 85%. Comparing this to existing annotation, which covers ≈75% of the genome, shows that 16% of the transcribed base pairs had not been annotated before.

To obtain an unbiased map of the position, abundance, and architecture of transcripts, the hybridization signals were examined along their chromosomal position for each strand (Fig. 1). The profiles were partitioned into segments of constant hybridization intensity, separated by change points demarcating transcript bound-

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviation: GO, gene ontology.

Data deposition: The array data have been deposited in ArrayExpress database (accession no. E-TABM-14).

†L.D. and W.H. contributed equally to this work.

¶To whom correspondence may be addressed. E-mail: dbowe@stanford.edu or larsms@embl.de.

www.pnas.org/cgi/doi/10.1073/pnas.0601091103

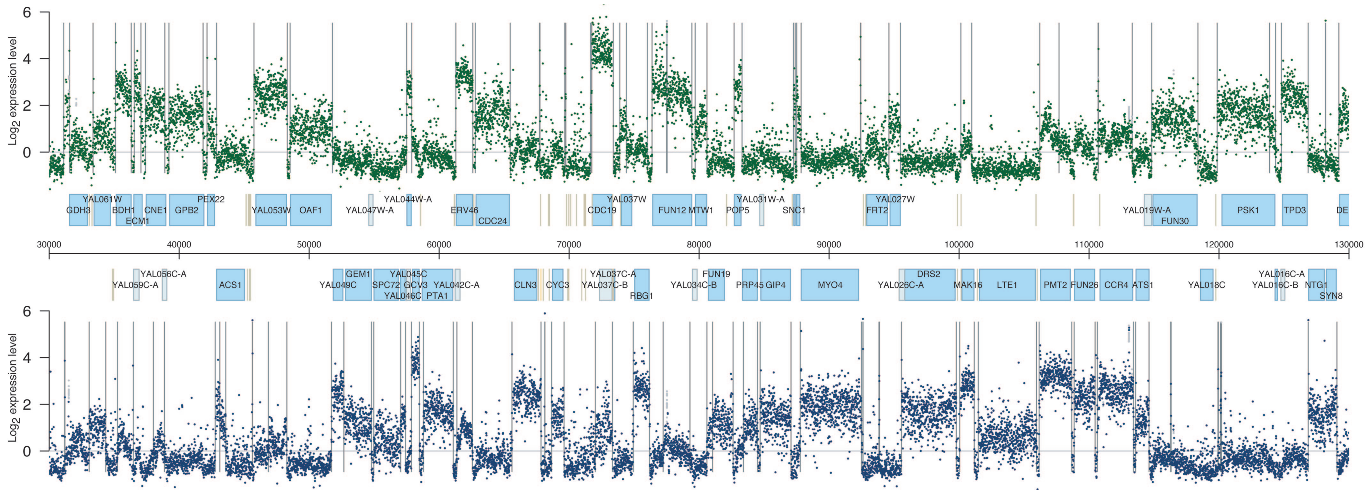

Fig. 1. Visualization of yeast tiling array intensities along 100 kb of chromosome 1, corresponding to ≈1% of the genome. The plot shows the normalized hybridization intensities (*y* axis) along genomic coordinates (*x* axis in bp). Each dot corresponds to a probe, Watson strand in green and Crick strand in blue. Probes with more than one perfect match in the genome are colored gray. Annotated ORFs are shown as blue boxes, dubious ORFs are shown as light blue boxes, and transcription factor binding sites are shown as gray bars. Vertical lines are segment boundaries. The background threshold ($y = 0$) is shown as a horizontal line.

aries. We used a change point detection algorithm that determines the global maximum of the log-likelihood of a piece-wise constant model by dynamic programming (14, 15). Compared to running-window approaches, it finds more accurate estimates of change point locations and depends on fewer user-defined parameters. Segments were determined separately for poly(A) and total RNA (Tables 3 and 4, respectively, which are published as supporting information on the PNAS web site). Segments from poly(A) and total RNA were remarkably concordant, and many noncoding RNA (ncRNA) were also found in the poly(A) data (Table 5, which is published as supporting information on the PNAS web site). Overall, the poly(A) RNA hybridization data were cleaner and therefore were the focus of our analysis.

The automated segmentation algorithm provides an unbiased global analysis, but the data complexity invites additional manual curation. Profiles for all genomic regions are provided in a database that is searchable by gene symbol or chromosomal coordinate (www.ebi.ac.uk/huber-srv/queryGene). We encourage readers to explore the database along with the examples discussed below.

Examples from this map of transcription are shown in Fig. 2. The hybridization data accurately separates exons from spliced introns,

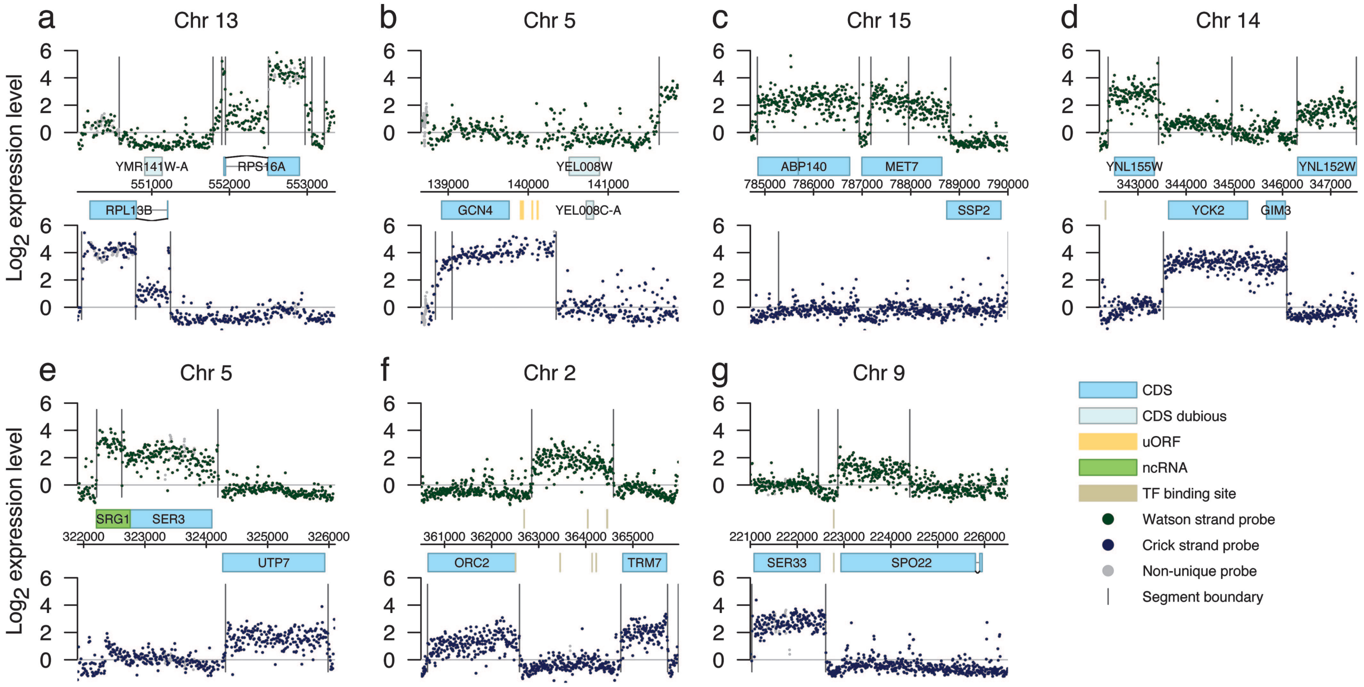

Fig. 2. Examples of transcriptional architecture. (*a*) Detection of spliced transcripts. (*b*) Long 5′ UTR of *GCN4* including its cotranscribed upstream ORFs. (*c*) Complex transcript architecture of *MET7*. (*d*) Overlapping transcripts of two ORFs. (*e*) Adjacent transcripts of *SER3* and the noncoding *SRG1*. (*f*) Nonannotated isolated transcript. (*g*) Transcript antisense to *SPO22*. CDS refers to coding sequence; uORF, upstream ORF; ncRNA, noncoding RNA; TF, transcription factor. Plot layout as in Fig. 1.

for which lower but significant levels of transcription were found, as shown for *RPS16A* (top strand) and *RPL13B* (bottom strand) (Fig. 2*a*). The segmentation-mapped UTRs of coding transcripts (e.g. *GCN4*, Fig. 2*b*) identified complex transcriptional architectures, such as uneven transcript levels for different regions of a single ORF (*MET7*, Fig. 2*c*), determined transcripts spanning multiple ORFs (*YCK2*, *GIM3*, Fig. *2d*), and identified adjacent transcripts for neighboring genes, uninterrupted by untranscribed intergenic regions (*SRG1*, *SER3*, Fig. 2*e*). Moreover, we observed transcripts in regions of the genome lacking prior annotation (Fig. 2*f*), as well as transcripts opposite annotated features (Fig. 2*g*). For each of the above, many instances were identified in the genome and we discuss them below.

UTR Boundaries. To map UTRs, we compared ORF boundaries with segment boundaries. We automatically determined UTR lengths for verified or uncharacterized nuclear-encoded genes whose annotated coding sequence was fully contained within a single segment. A total of 2,223 segments passed a confidence filter that required a sharp decrease in signal on both sides of the segment. UTR coordinates are given in Tables 3 and 4. We proceeded with analysis of the 2,044 poly(A)-determined UTRs because the poly(A) hybridization data were cleaner and yielded most of the UTR determinations (Fig. 7, which is published as supporting information on the PNAS web site). For many remaining genes that did not pass the confidence filter, the UTRs can be mapped by closer inspection.

We found that 3′ UTRs were significantly longer than the 5′ UTRs, with a median of 91 vs. 68 nt (Fig. 3*a*). Longer 3′ UTRs are consistent with them containing posttranscriptional regulatory regions that influence mRNA stability, localization, and translation (16), and with findings from other species (17). The mean sum of 3′ and 5′ UTR lengths was 211 nt and similar to a mean of 256 nt found by a gel-mobility assay (18). We computed 662 3′ UTRs from ESTs (19) and compared them to 435 ORFs that had UTRs in our data set. The Pearson correlation coefficient between the UTR length estimates was 0.63. A contribution to the differences is that in the EST data the longest transcript was chosen, whereas the array measures the average transcript abundance at each probe position.

We compared UTR lengths with transcript levels and coding sequence (ORF) lengths. Although transcript level was generally lower for genes with long coding sequences, neither transcript levels nor ORF lengths were significantly associated with UTR lengths. We also compared length distributions of UTRs for different functional and localization categories (GO annotations) and detected significant correlations (Fig. 3*b*). The longest 3′ UTRs were found for transcripts of proteins that are targeted to the mitochondrial electron transport chain, the plasma membrane, and the cell wall. These longer 3′ UTRs may contain mRNA localization signals, as has been well demonstrated for mitochondrial targeted proteins (20, 21). Genes involved in phosphorylation, transporter activity, ion transport, and specific stages of the mitotic cell cycle had both ends longer. Genes involved in RNA processing, rRNA metabolism, and ribosome biogenesis had both ends shorter. Therefore, genes with longer UTRs seem to fall into categories that require regulation, whereas genes with short UTRs seem to fall into categories with a reduced need for posttranscriptional regulation, such as housekeeping genes.

Complex Transcriptional Architectures. Many expressed segments flanked other expressed segments with different signal levels, thus making up complex transcriptional architectures. In many cases, different parts of the same gene are expressed at different levels: 921 ORFs from the poly(A) RNA sample were divided into at least two expressed segments, one covering >50% of the feature and others <50%. Such complex architectures could be due to alternative transcription initiation, termination, or alternative splicing, as has been described in mouse (9) and for several human genes (22). In yeast, it has been suggested that up to 20% of mRNAs have alternative 3′ ends (23). Complex hybridization patterns on the array could also be caused by RNA decay or variation introduced by reverse transcription, because the array captures the sum of cDNA molecules present at the time of hybridization. The explanation of our observations by such mechanisms will require a case-by-case analysis.

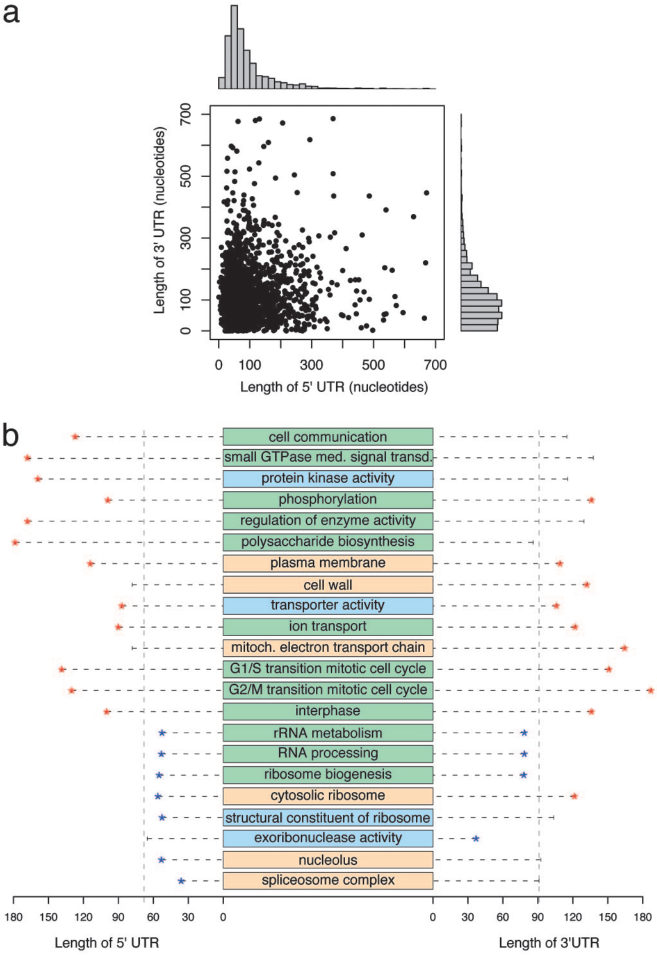

Fig. 3. Length of UTRs and functional categories with exceptional UTR length. Analyses were based on 2,044 genes from poly(A) samples. (*a*) Scatterplot and histogram of 3′ vs. 5′ UTR lengths. (*b*) Association between UTR length, cellular localization, and biological process. Length distributions between genes inside and outside of GO categories were compared, and selected significant categories are shown (orange, cellular component; green, biological process; blue, molecular function). For each category, a horizontal line shows the 5′ and 3′ median UTR lengths measured in nucleotides (*x* axis). The median over all genes is shown by a vertical dashed line. Significant medians are indicated by asterisks, red longer, blue shorter (two-sided Wilcoxon test, $P \leq 0.002$).

Here, we discuss a few cases. For *CPB1* and *RNA14*, our observed architecture matches previous results describing alternative 3′ ends in response to carbon source regulation (24). For *GCN4*, lower hybridization signal was observed at the 3′ end (Fig. 2*b*). *GCN4* is not translated during nutrient-rich growth because of the translation of the upstream ORFs encoded in the same transcript (25). 3′ end degradation due to a lack of translation could explain the lower 3′ end signal. In support, this decrease was not seen in an oligo(dT) reverse-transcribed sample, where no priming would occur on degraded poly(A) transcripts. At the 5′ end, the segment boundary matches the previously determined position to within nine bases.

For *MET7*, the annotated gene was segmented into three regions (Fig. 2*c*), suggesting a misannotation of the translation start site. A later transcription start site is supported by the multiple sequence alignment of yeast species in SGD, which shows that conservation of *MET7* starts at a later methionine (M55), whose position agrees with the transcription start site detected by the array. Also, the level difference between the central and the 3′ segment was not seen in a poly(A) sample that was reverse transcribed by using oligo(dT) primers (Fig. 8, which is published as supporting information on the PNAS web site), consistent with early transcript termination or RNA decay. Altogether, we tested 27 regions from 10 genes by quantitative real-time PCR (including *MET7*). For seven genes, the PCR results matched the architectures in the array data (Fig. 8).

Neighboring Transcription. Additional unusual architecture was found for adjacent ORFs not separated by an unexpressed region. Such architectures can result from more than one ORF being encoded from a single transcript, like the upstream ORFs in *GCN4*, or from distinct transcripts not separated by untranscribed intergenic regions. We found the ORFs of *GIM3* and *YCK2* within one segment resembling a bicistronic transcript (Fig. 2*d*). The PhastCons multiple alignment (26) of the intergenic region with other yeast species shows high sequence conservation, but includes frame-shifting gaps, which suggests that the two ORFs are not translated as one. By reverse-transcription PCR across the gap between the ORFs, a product was obtained supporting either bicistronic or overlapping transcripts. Operon-like transcription was reported for few eukaryotic species and mostly for *Caenorhabditis elegans* (27). A bicistronic transcript had been reported previously in yeast for *YMR181C* and *RGM1* (28), and we observed different transcript levels for the ORFs, but no separation by an untranscribed region.

Fig. 2*e* shows two other adjacent transcripts, *SRG1* and *SER3*, expressed at different levels and not separated by an intergenic region. It had been proposed that *SRG1*, an upstream noncoding RNA, represses the expression of *SER3* in rich media, by reducing the binding of *SER3* transcription factors (29). In contrast, we find that *SER3* is expressed significantly above background, suggesting that even though *SRG1* is expressed at a higher level, its transcription does not prevent *SER3* from being transcribed. There are many cases of adjacent genes not separated by unexpressed, intergenic regions in our data set, and this suggests that transcription over active promoters of adjacent genes is common in yeast. Some further examples are *QCR6*, *PHO8*, *RIB3*, *HCH1*, *UBI4*, *SEC53*, *RPS26A*, and *ADE13*.

Unannotated Transcripts. Many segments with signal above background did not overlap existing annotation. They fall into two classes: nonannotated isolated segments if there was no prior annotation on either strand (Fig. 2*f*), and nonannotated antisense segments if there was an annotation on the opposite strand (Fig. 2*g*). Many are not independent transcription units: some represent UTRs of genes with complex transcriptional architecture; others are part of unannotated transcripts that are divided into multiple segments. The identification of antisense transcripts requires caution because reverse transcription can generate double stranded cDNA from secondary mispriming (30, 31). Considering these concerns, we applied a filter requiring segments to be at least 48 bp long, be flanked by segments with reduced hybridization signal on both sides, and have higher expression signal than seen on the opposite strand for at least part of their length. In these filtered categories, we obtained 427 nonannotated segments from poly(A) and 357 from total RNA hybridizations. These segments divide approximately equally into the isolated and antisense categories (Fig. 4*a*). Antisense segments and segments overlapping annotation ($\geq$50%) had similar length distributions and tended to be longer than those of the isolated categories (Fig. 4*b*). Isolated segments showed similar levels of expression as annotated segments, whereas antisense segments had lower expression levels (Fig. 4*c*).

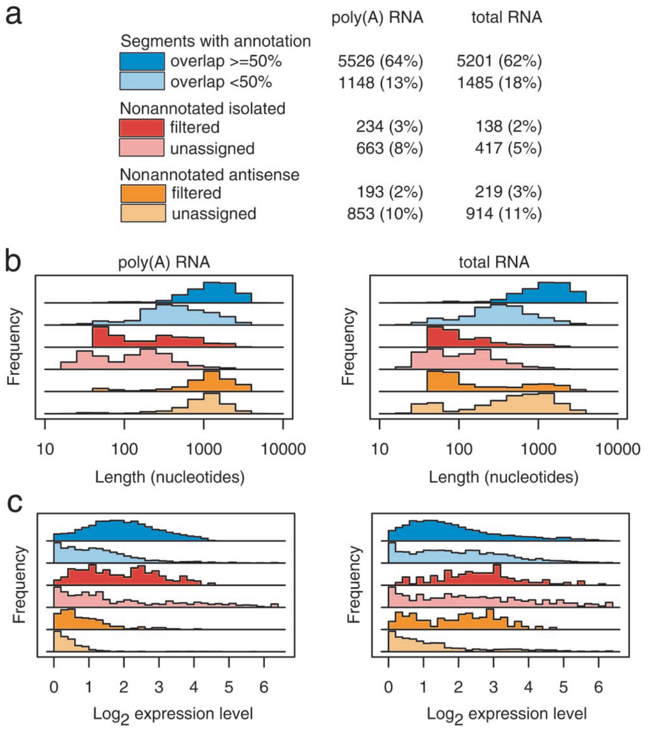

Fig. 4. Categories of expressed segments, their length, and their expression levels. (*a*) Number and percentage of the expressed segments detected from the poly(A) RNA and total RNA hybridizations. Categories ''>= 50%'' and ''<50%'' consist of segments that overlap more, or less, than half of an annotated feature, respectively. The ''nonannotated isolated'' category consists of segments that have no overlap with annotated features on either strand, whereas the ''nonannotated antisense'' category consists of those that overlap with features on the opposite strand. The ''filtered'' categories consist of the high confidence segments that passed our filter, and the ''unassigned'' categories consist of the remaining segments. Length (*b*) and transcript level (*c*) distributions for segments from the above categories are given.

Nonannotated Isolated Transcripts. We verified the array identification of 126 nonannotated isolated transcripts by RT-PCR. All were expressed in both total and poly(A) RNA reverse transcribed by using random and oligo(dT) primers, respectively. For 10 of them, a quantitative real-time PCR analysis showed their levels to be similar to expressed ORFs in both sample types.

The 1.7-kb transcript between *ORC2* and *TRM7* (Fig. 2*f*) is an example of a nonannotated isolated segment that is highly conserved across other yeast species in the PhastCons multiple alignment (26). Nevertheless, assessment of 125 nonannotated isolated segments showed that only 48 had a multiple alignment >50 nt across four yeast species (32). In addition, median total tree lengths of the phylogenies were 0.59 for verified genes, 1.08 for undetected unannotated segments, and 1.50 for nonannotated isolated segments (Table 6, which is published as supporting information on the PNAS web site). To assess protein-coding potential, we tested for dissimilarity in evolutionary rates among first, second, and third codon positions in all reading frames. There was no protein coding signature for the 48 nonannotated segments (median likelihood-ratio statistic of 1.06, compared to 1.14 for undetected unannotated segments and 162.0 for verified genes). Conservation of DNA sequence or protein coding ability is nevertheless neither a necessary nor a sufficient attribute of transcript function.

We generated knockouts for 47 nonannotated isolated segments

Table 1. Selected GO categories found overrepresented among the 355 genes opposite filtered nonannotated antisense segments

GO term	N_g	N_{obs}	N_{exp}	Odds ratio	P
Cell wall	95	17	5.1	3.3	2×10^{-6}
M phase of meiotic cell cycle	127	21	6.8	3.1	9×10^{-7}
Transcriptional activator activity	33	9	1.8	5.1	5×10^{-6}
Transcriptional repressor activity	23	6	1.2	4.8	1×10^{-4}
Monosaccharide transporter activity	19	5	1	4.9	3×10^{-4}

N_g, number of genes in the genome annotated to this category; N_{obs}, number of genes observed in this category that were opposite an antisense segment; N_{exp}, number of genes expected if genes opposite antisense segments are randomly distributed over GO categories; *P,* hypergeometric test *P* value.

and tested for growth defects in rich media conditions (Table 7, which is published as supporting information on the PNAS web site). A growth defect was identified for two knockouts: one on chromosome 6, positions 54813–55221, the other on chromosome 7, positions 622039–622295. On chromosome 6, the deleted segment contained annotated transcription factor binding sites upstream to *ACT1*, an essential gene, which likely accounts for the observed inviability. On chromosome 7, the deletion does not overlap any annotation, and strains with deletions of the neighboring ORFs (*YGR066C*, *YGR067C*) did not have a growth defect. This segment does not appear to be evolutionarily conserved or to contain a long ORF. The proportion of growth defects found within the 47 knockouts is much lower than the ≈40% found for knockouts of protein-coding genes (33).

Nonannotated Antisense Transcription. We identified antisense transcripts opposite to 1,555 genes, of which 402 were in the filtered set from both poly(A) and total RNA samples (Tables 3 and 4). The antisense transcripts are not caused by read-through from ORFs on the opposite strand, but appear as independent transcription units. For example, antisense transcription was found opposite *SPO22*, a meiosis-specific protein induced early in meiosis (Fig. 2*g*). Upstream of this antisense transcript, there is a binding site for *CBF1*. *CBF1* is involved in regulation of DNA replication and chromosome cycle and is important for growth in rich media, suggesting that the antisense expression may be negatively correlated with the expression of *SPO22*.

Many genes with antisense transcripts had products that localize to the cell cortex and cell wall, and that function in the meiotic cell cycle and in transcriptional regulation (Table 1). Some of these categories included genes not active during growth in rich media, like meiosis. Others included genes that are active during growth in rich media, but which may need posttranscriptional regulation. Further correlations were found between UTRs and their opposite antisense segments: More antisense transcripts overlapped the 3′ UTRs than the 5′ UTRs; also, UTRs that had overlapping antisense transcripts were longer than UTRs that did not (Table 2).

The generation and significance of the many nonannotated transcripts is unclear. Regulation of gene expression by antisense transcripts was reported in prokaryotes (34) and higher eukaryotes (35). Sense/antisense transcript pairs were suggested to be frequent in mammalian genomes and to provide regulatory function (3). In *S. cerevisiae*, major components of the RNA interference machinery have not been identified (36); however, in other species, alternative mechanisms for regulation by noncoding RNAs exist (2, 37). In *Drosophila*, it has been shown that microRNA predominantly target 3′ UTRs (38) and these UTRs also tend to be longer than UTRs of genes not targeted (39). We observed similar correlations for antisense transcripts in *S. cerevisiae*, which together with their association to particular functional categories may suggest a possible regulatory role. There are experiments supporting this hypothesis: artificial antisense transcripts in *S. cerevisiae* had effects on expression of several genes (40–42), and overexpression of random genomic fragments antisense to ORFs has led in several cases to growth inhibition (43). In our data set, naturally occurring antisense transcripts were found for ≈20 of these cases.

Most ncRNAs previously reported as novel have since been annotated in SGD, and hence do not overlap with our expressed, nonannotated segments (44, 45). We compared our data to transcriptome surveys, carried out by using serial analysis of gene expression (SAGE) (46) and ESTs (19). Thirteen percent of the nonannotated isolated and 42% of the nonannotated antisense transcripts were represented by SAGE tags. For the EST data, these numbers were 1% and 6%, respectively. Analysis of SAGE tags on microarrays described a number of novel transcripts in a mutant strain defective in the RNA degradation pathway (47); however, the eight primary examples were not found expressed in our study of wild-type yeast.

This study reveals considerable transcriptional activity in yeast that is currently not systematically annotated. Our transcription map will be useful for annotating the genome. Furthermore, the position of transcription initiation and termination sites will help in defining the promoters and transcriptional regulators of genes. Although our results suggest that not many new, long protein-coding regions will be discovered in yeast, the extensive noncoding transcription detected in regions with no prior anotation and antisense to annotated transcripts invites further investigation. Therefore, even for a genome that has been studied intensively since it was sequenced 10 years ago (48), a glimpse into the complexity of its transcriptional architecture makes this genome appear like novel territory.

Materials and Methods

Array Design and Sample Hybridization. The array was designed in collaboration with Affymetrix (Santa Clara, CA) (PN 520055). An S288c background strain S96 (MATa *gal2 lys5*) was grown in rich yeast-extract/peptone/dextrose media to mid-exponential phase. Total RNA was isolated by hot phenol extraction. Poly(A)

Table 2. Association of UTR lengths with presence of antisense transcript, and the 3′/5′ bias in position of antisense transcripts.

	Antisense		Control, no antisense
	Filtered	All	
Number of 3′ overlaps	145	783	NA
Number of 5′ overlaps	94	355	NA
Median length of 3′ UTR (no. of genes)	111 (49) $P = 0.05$	104 (408) $P = 0.00001$	87.5 (1,588)
Median length of 5′ UTR (no. of genes)	89 (28) $P = 0.08$	82 (142) $P = 0.003$	67 (1,588)

Overlap was measured with respect to the start and stop codons. Significance was calculated by comparing the length distributions of UTRs with antisense to controls where UTRs had no antisense partner by using the two-sided Wilcoxon test. NA, not applicable.

RNA was enriched by two rounds of the Oligotex mRNA kit (Qiagen). First-strand cDNA was synthesized by using random primers. Three replicate hybridizations (biological) of poly(A), two of total RNA, and three of genomic DNA were performed.

Probe Annotation. Probe sequences were aligned to the genome sequence of *S. cerevisiae* strain S288c (SGD of August 7, 2005). Perfect match probes were further analyzed.

Normalization. RNA hybridization intensities were adjusted by

$$N_{ij} = \frac{X_{ij} - B_j(A_i)}{A_i}$$

where X_{ij} is the RNA intensity of the ith probe on the jth array, A_i is the geometric mean of the intensities from the DNA hybridizations, $B_j(A)$ is a continuous function that parameterizes the estimated background of probes with gain A, and N_{ij} is the adjusted intensity. Probes were grouped into 20 strata defined by the 5%, 10%, 15%, . . . , 100% quantiles of A_i. Within each stratum, and for each array j, the midpoint of the shorth of the intensities of the probes for which no genomic feature was annotated on either strand was calculated. Linear interpolation yielded the function B_j. Dead probes (the 5% of probes with lowest signal in the DNA hybridization) were discarded. The values N_{ij} were background-adjusted and transformed to $\log_2$ scale by using VSN (13). Fig. 5 demonstrates the successive improvements in the signal-to-noise ratio during normalization.

Segmentation. Segments of approximately constant hybridization signal were defined by using a dynamic programming algorithm that, for each chromosome strand separately, minimizes the cost function

$$G(t_l, \ldots, t_s) = \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{i>=t_s}^{i<t_{s+1}} (y_{ij} - \bar{y}_{sj})^2,$$

where y_{ij} is the VSN-normalized signal of the ith probe on the jth replicate array, $\bar{y}_{sj}$ is the arithmetic mean of the signal values of array j in segment s, S is the number of segments, and $t_1, \ldots, t_S$ are the segment boundaries (15). For each chromosome, S was chosen such that the average segment length was 1,500 nt. S, the only parameter of the segmentation algorithm, controls the sensitivity–specificity tradeoff and was chosen to yield high sensitivity.

All analyses were performed with custom-written software in the language and statistics environment R (49) and BIOCONDUCTOR (14). For additional details on analyses and experimental procedure, see *Supporting Text*, which is published as supporting information on the PNAS web site.

We thank Mike Mittmann for the array design, Iain Russell and Victor Sementchenko for experimental advice, Raquel Kuehn and Michelle Nugyen for technical assistance, Roy Parker, Rafael Irizarry, Elisa Izaurralde, Steve Cohen, and Nick Goldman for helpful comments on the manuscript, and contributors to the BIOCONDUCTOR (www.bioconductor.org) and R (www.R-project.org) projects for their software. This work was funded by the National Institutes of Health (R.W.D and L.M.S.) and the Deutsche Forschungsgemeinschaft (L.M.S.).

1. Wilkie, G. S., Dickson, K. S. & Gray, N. K. (2003) *Trends Biochem. Sci.* **28,** 182–188.
2. Storz, G., Altuvia, S. & Wassarman, K. M. (2005) *Annu. Rev. Biochem.* **74,** 199–217.
3. Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C. C., Suzuki, M., Kawai, J., *et al.* (2005) *Science* **309,** 1564–1566.
4. Mattick, J. S. (2004) *Nat. Rev. Genet.* **5,** 316–323.
5. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., *et al.* (2005) *Science* **308,** 1149–1154.
6. Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., *et al.* (2004) *Science* **306,** 2242–2246.
7. Yamada, K., Lim, J., Dale, J. M., Chen, H., Shinn, P., Palm, C. J., Southwick, A. M., Wu, H. C., Kim, C., Nguyen, M., *et al.* (2003) *Science* **302,** 842–846.
8. Royce, T. E., Rozowsky, J. S., Bertone, P., Samanta, M., Stolc, V., Weissman, S., Snyder, M. & Gerstein, M. (2005) *Trends Genet.* **21,** 466–475.
9. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., *et al.* (2005) *Science* **309,** 1559–1563.
10. Hekstra, D., Taussig, A. R., Magnasco, M. & Naef, F. (2003) *Nucleic Acids Res.* **31,** 1962–1968.
11. Naef, F. & Magnasco, M. O. (2003) *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **68,** 011906.
12. Wu, Z. & Irizarry, R. A. (2005) *J. Comput. Biol.* **12,** 882–893.
13. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. (2002) *Bioinformatics* **18,** Suppl. 1, S96–S104.
14. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (Springer, Heidelberg).
15. Picard, F., Robin, S., Lavielle, M., Vaisse, C. & Daudin, J. J. (2005) *BMC Bioinformatics* **6,** 27.
16. Kuersten, S. & Goodwin, E. B. (2003) *Nat. Rev. Genet.* **4,** 626–637.
17. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. (2002) *Genome Biol.* **3,** REVIEWS0004.
18. Hurowitz, E. H. & Brown, P. O. (2003) *Genome Biol.* **5,** R2.
19. Graber, J. H., Cantor, C. R., Mohr, S. C. & Smith, T. F. (1999) *Nucleic Acids Res.* **27,** 888–894.
20. Marc, P., Margeot, A., Devaux, F., Blugeon, C., Corral-Debrinski, M. & Jacq, C. (2002) *EMBO Rep.* **3,** 159–164.
21. Gerber, A. P., Herschlag, D. & Brown, P. O. (2004) *PLoS Biol.* **2,** E79.
22. Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S. & Gingeras, T. R. (2005) *Genome Res.* **15,** 987–997.
23. Graber, J. H., McAllister, G. D. & Smith, T. F. (2002) *Nucleic Acids Res.* **30,** 1851–1858.
24. Sparks, K. A. & Dieckmann, C. L. (1998) *Nucleic Acids Res.* **26,** 4676–4687.
25. Hinnebusch, A. G. (2005) *Annu. Rev. Microbiol.* **59,** 407–450.
26. Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.* (2005) *Genome Res.* **15,** 1034–1050.
27. Blumenthal, T. & Gleason, K. S. (2003) *Nat. Rev. Genet.* **4,** 112–120.
28. He, F., Li, X., Spatrick, P., Casillo, R., Dong, S. & Jacobson, A. (2003) *Mol. Cell* **12,** 1439–1452.
29. Martens, J. A., Laprade, L. & Winston, F. (2004) *Nature* **429,** 571–574.
30. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004) *Genome Res.* **14,** 331–342.
31. Johnson, J. M., Edwards, S., Shoemaker, D. & Schadt, E. E. (2005) *Trends Genet.* **21,** 93–102.
32. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003) *Nature* **423,** 241–254.
33. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., *et al.* (1999) *Science* **285,** 901–906.
34. Wagner, E. G. & Simons, R. W. (1994) *Annu. Rev. Microbiol.* **48,** 713–742.
35. Kumar, M. & Carmichael, G. G. (1998) *Microbiol. Mol. Biol. Rev.* **62,** 1415–1434.
36. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 11319–11324.
37. Vanhee-Brossollet, C. & Vaquero, C. (1998) *Gene* **211,** 1–9.
38. Lai, E. C. (2002) *Nat. Genet.* **30,** 363–364.
39. Stark, A., Brennecke, J., Bushati, N., Russell, R. B. & Cohen, S. M. (2005) *Cell* **123,** 1133–1146.
40. Xiao, W. & Rank, G. H. (1988) *Curr. Genet.* **13,** 283–289.
41. Peterson, J. A. & Myers, A. M. (1993) *Nucleic Acids Res.* **21,** 5500–5508.
42. Park, H., Shin, M. & Woo, I. (2001) *J. Biosci. Bioeng.* **92,** 481–484.
43. Boyer, J., Badis, G., Fairhead, C., Talla, E., Hantraye, F., Fabre, E., Fischer, G., Hennequin, C., Koszul, R., Lafontaine, I., *et al.* (2004) *Genome Biol.* **5,** R72.
44. Olivas, W. M., Muhlrad, D. & Parker, R. (1997) *Nucleic Acids Res.* **25,** 4619–4625.
45. McCutcheon, J. P. & Eddy, S. R. (2003) *Nucleic Acids Res.* **31,** 4119–4128.
46. Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B. & Kinzler, K. W. (1997) *Cell* **88,** 243–251.
47. Wyers, F., Rougemaille, M., Badis, G., Rousselle, J. C., Dufour, M. E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., *et al.* (2005) *Cell* **121,** 725–737.
48. 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) *Science* **274,** 563–567.
49. R Development Core Team (2005) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).