The identification of untranslated regions (UTRs), introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method for mapping transcribed regions called RNA-Seq in which cDNA fragments are subjected to high throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames were also identified for many yeast genes. We also found unexpected 3′ end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.

A major challenge in genomics is to identify all genes, exons and their boundaries. This information is crucial in order to understand the functional elements of the genome, determine when they are expressed, and how they are regulated. Often genes are identified through the presence of large open reading frames (ORFs), sequence conservation or cDNA probing of DNA tiling microarrays (1–5). These methods often fail to identify short exons, do not precisely reveal boundaries of UTRs and/or have high false positive rates.

In order to better map the transcribed regions of the yeast genome, we developed an RNA-Seq approach presented in Fig. 1. Briefly, poly(A) RNA was isolated from yeast cells grown in rich media (see SOM text) and used to generate double stranded cDNA by reverse transcription with either random hexamers or oligo(dT) primers. The double stranded cDNA was fragmented and subjected to high throughput Illumina sequencing in which 35 bp of sequence was determined from one end of the fragments. Two technical and two biological replicates were performed for each hexamer and oligo(dT) primed cDNA sample for a total of 14,125,182 and 15,787,335 reads, respectively. These sequence reads were analyzed with an algorithm that maps unique (i.e. single copy) sequences to the genome and allows for up to two mismatches (see SOM). Of 29,912,517 total reads, 15,870,540 (56%) were mapped to unique genomic regions (fig. S1A and table S1).

We assessed our results by several criteria. First, none of the 29,912,517 sequence reads matched the deleted 3.5 kb regions of the genome (Fig. 1C), and very few, if any, matched the nontranscribed centromeres (fig S1B; SOM); thus our method is specific. Second, our replicates were in close agreement with one another with a 0.99 Pearson correlation coefficient for technical replicates and 0.93-0.95 for biological replicates (fig. S2). The data generated from hexamers and oligo(dT) primers were also closely correlated (0.97) and showed similar patterns of expression (fig. S2). Therefore, we merged all of these data sets, and the subsequent analyses were performed using the merged data.

RNA-Seq analyses revealed extensive expression of the whole yeast genome (Fig. 2A); 74.5% of the genome is expressed as RNA-Seq tags (Fig. 2B). We detected more reads from the 3′ ends than from the 5′ ends of annotated genes (fig. S3), presumably due to enrichment of 3′ sequences during poly(A) purification as well as enhanced priming at 3′ ends. Despite this bias, the deep sequencing allowed detection of signal across the entire gene. Overall, 85% of the bases detected as expressed by RNA-Seq overlapped with those found with DNA tiling microarrays (6).

We investigated the overall transcriptional activity and found that 4,666 of the 5,099 annotated ORFs (91.5%) in the Saccharomyces Genome Database (SGD) were expressed as tags above background (SOM). For this analysis we removed 1,178 ORFs whose 3′ ends lie within 100 bp of one another and whose transcripts might overlap. In addition, 327 ORFs that were not unique in their 3′-ends were not analyzed. We observed high expression for 20% of the genes; Gene Ontology (GO) analysis revealed that genes involved in...
biosynthetic pathways and ion transport were specifically enriched in the highly expressed category (\( P < 2.3 \times 10^{-58} \); see table S2 for complete list). Medium and low expression levels were observed for 39% and 33% of the genes, respectively. As expected we did not detect expression of genes involved in meiosis, mating, cell differentiation, sugar transport, and vitamin metabolism, whose function is not required during vegetative growth (7).

The majority of yeast genes have been annotated primarily with ORFs and, to a lesser extent, with cDNA sequencing (8); thus the 5′ and 3′ boundaries and UTRs of most yeast genes have not been precisely defined. To map the 5′ ends of genes with RNA-Seq, the 5′ ends of 1,331 genes were first determined by generating and sequencing RACE PCR products (table S3). We then used 125 RACE ends to optimize parameters for determining 5′ ends by searching RNA-Seq data for a sharp signal reduction in the transcribed region; applying these parameters revealed the 5′ boundary regions for 4,665 transcribed genes. Genes with very low levels of expression were excluded from the analysis. Comparison of these results with 1,025 boundaries mapped with 5′ RACE showed that both methods identified 5′ boundaries within 50 bp of one another for 786 genes (77.9%; Fig. 3A). Combining the 5′ RACE results with the RNA-Seq results defined the 5′ boundaries of 4,835 yeast genes (see example Fig. 3B). The median length of 5′ UTRs was 50bp with a range of 0 to 990 bp (Fig. 3A; top right panel). 241 genes contained an ATG less than 10 bp from the 5′ end; we do not know if these ATGs represent true initiation codons.

We also globally mapped the 3′ boundaries of yeast genes by searching for a rapid transition in RNA-Seq signal as well as by identifying end tags with poly(A) sequences containing a novel stretch of 3 or more consecutive A’s lying next to a genomic yeast sequence (SOM). Using these methods we mapped the 3′ boundaries of 5,212 transcribed genes and deduced the transcribed strand (Fig. 3, C and D; table S4). The end tags allowed precise assignment of 3′ boundaries even when transcripts are overlapping at their 3′ ends (Fig. 4). Surprisingly, we found evidence that the transcription of a large number of yeast genes overlaps with transcription from the other strand. Of 4,646 verified expressed ORFs, 275 transcribed pairs (11.8% of expressed genes) contained overlapping 3′ ends. Pervasive overlapping transcripts may be unique to S. cerevisiae and other organisms lacking miRNA/siRNA processing components which might otherwise degrade double stranded RNAs. Moreover, overlapping transcription at 3′ ends could be a form of gene regulation by which neighboring genes can potentially influence (positively or negatively) the expression of one another.

The median length of yeast 3′ UTRs is 104 bp with a range of 0 to 1,461 bp (Fig. 3A). Interestingly although most yeast genes have a single precise 3′ end (within 1-2 bp) many genes had heterogeneous 3′ end sequences. These occurred either in localized regions (usually 2-10 bp) suggesting some variability in 3′ end processing at the polyA signal or, for the case of 540 genes, in multiple peaks greater than 10 bp apart suggesting that more than one poly(A) site is used (see an example in fig. S4).

In yeast the first ATG at the 5′ end of an ORF is usually annotated as the start codon. However, in some databases the second ATG is annotated when the predicted amino terminal protein coding sequence was not conserved (9, 10). RNA-Seq analysis revealed 35 genes with 5′ ends upstream of an ATG that is 5′ and in-frame to the annotated ATG initiation codon, suggesting that these proteins are potentially longer than predicted. We also found 29 genes whose 5′ end is located downstream of the annotated ATG suggesting that these proteins are shorter than previously predicted; the 5′ ends for 4 of the latter genes were confirmed by RACE/Sequencing.

We also examined our data for the presence of introns; in yeast introns are typically identified on the basis of sequence conservation, microarray analyses or other studies. We detected 240 of 306 known introns with an algorithm that detects introns through the presence of discontinuous sequences whose boundaries contained GT and AG/AC as well as sequence tags spanning the intron boundary demonstrating that transcripts from these genes are spliced. For the 66 introns not validated with the algorithm we examined 30 whose genes were expressed. In four cases, the intron sequences were clearly expressed at similar levels as the adjacent exons (Fig. 3E and fig. S5) and unspliced products, but not splice junction tags, were identified. Thus, transcripts from these genes are probably not spliced at appreciable levels in vegetative cells (see also 11). In two of these cases the presence or lack of an intron affects the predicted protein sequence.

Recent analysis predicts that many eukaryotic 5′ UTRs may contain upstream ORFs (uORFs) (12). uORFs have been shown to regulate protein expression (13) and mRNA degradation (14) and have been annotated for 17 yeast genes. Our data predict uORFs upstream of the start codon for 321 (6%) of yeast gene transcripts (Fig. 5 and table S6). GO analysis of these genes reveals that genes encoding DNA binding proteins (\( P < 0.0027 \), FDR adjusted) and anatomical structure and development (e.g. sporulation; \( P < 0.0045 \), FDR adjusted) were significantly enriched for uORFs (Fig. 5B). The presence of uORFs in DNA binding proteins suggests that these important genes are likely to be highly regulated.

We also searched for novel transcription in intergenic regions (Fig. 5D) by identifying stretches of 150 bp or greater with significant expression above surrounding regions (Fig. 5F; SOM). Of 487 expressed regions identified, 204 had not been observed by microarray analyses or cDNA studies (6,
For 18 regions found by RNA-Seq but not other methods we tested expression with Real-Time PCR and confirmed transcription in 16 cases (table S5).

As RNA-Seq is a quantitative method it can potentially be used to quantify RNA levels. We determined the median signal in a 30 bp window located immediately upstream of the 3′ ends of the annotated stop codon (table S4) after removing genes with overlapping 3′ ends in this region. The expression levels of 34 genes predicted to be expressed at a low level and were better than those obtained with standard microarrays (R = 0.72; fig. S6C) (15) or tiling DNA microarray analysis (R = 0.48; fig. S6B) (6). Moreover, the dynamic range of RNA-Seq is at least 8000 fold; compared to ~60 fold for DNA microarrays (see fig. S6 scale of panels B and C).

RNA-Seq offers several advantages compared to existing technologies. First, it allows interrogation of all unique sequences of the genome, including those that are closely related; as long as unique bases exist that can be monitored. Second, because a large number of reads can readily be obtained, the method is very sensitive and offers a large dynamic range. Thus, RNA-Seq can detect and quantify expression levels of RNAs expressed at very low levels compared to qPCR. We found a strong correlation (R = 0.98) between the qPCR and RNA-Seq data (fig. S6A). As expected, discrepancies were the largest among the genes expressed at a low level and were better than those obtained with standard microarray analysis (R = 0.48; fig. S6B) (6). Moreover, the dynamic range of RNA-Seq is at least 8000 fold; compared to ~60 fold for DNA microarrays (see fig. S6 scale of panels B and C).

RNA-Seq allows accurate determination of exon boundaries. Although we precisely mapped 3′ ends of many yeast genes using RNA-Seq we did not attempt to determine nucleotide resolution of 5′ boundaries because yeast 5′ ends are often heterogeneous (8, 16), and we performed an amplification step. Instead, an approximate location was deduced by a sharp transition in signal over a small interval. Nonetheless, overall RNA-Seq provides a useful map of exon boundaries. In conclusion, our novel RNA-Seq method allowed us to map the transcriptional landscape of the yeast genome and define UTRs and novel transcribed regions. In the future, application of this method should aid in determining precise transcriptional landscape of other genomes.

References and Notes
17. We thank S. P. Dinesh-Kumar for comments. Funded by NIH (grant #) and the CT Stem Cell Fund. The GEO accession number for sequences is GSE11209.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1158441/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 to S6
References
31 March 2008; accepted 22 April 2008
Published online 1 May 2008; 10.1126/science.1158441
Include this information when citing this paper.

Fig. 1. (A) and (B) Flowchart of RNA-Seq method. (C) RNA-Seq signals are not evident at a deleted gene (LEU2) but are abundant at an expressed neighboring gene (YCL017C).

Fig. 2. (A) The genome distribution of transcribed regions in yeast. Colors represent different transcription levels for each base (log2 tag count). (B) A summary of the transcription level of the transcriptome.

Fig. 3. (A) Size differences between mapped 5′-UTR with RNA-Seq and RACE (top left) and 3′ UTR data with RNA-Seq and cDNA sequencing (8) (bottom left) next to the size distributions of the 5′-UTR (top right) or 3′-UTR (bottom right). (B) The 5′-UTR as determined by RNA-Seq or by 5′-RACE for gene YKL004W. In all figures a colored box represents ORF with an arrow indicating transcription direction. (C) 3′-UTR determined by RNA-Seq on the basis of end tags for genes YDR460W, YDR461W, and YDR461C-A, or YDR461W by cDNA sequencing (8). Endtag_W (blue) and Endtag_C (red) represent RNA-Seq reads containing polyA tails on either Watson or Crick strands, respectively, and the bars represent confidence scores of each 3′-end (SOM). (D) 3′-UTR determined by RNA-Seq based on sharp expression decrease, comparing to cDNA data (8). (E) An
SGD annotated intron (box in brown color) in GCR1 not supported by RNA-Seq.

**Fig. 4.** New annotations of UTRs in a relatively poorly annotated chrVI region. ORFs identified by this study are denoted by dotted lines, with arrows denoting transcription direction. UTRs- green shaded boxes. cDNA transcripts in red are of high confidence and those in blue are low (defined in 8).

**Fig. 5.** (A) RNA-Seq reveals putative upstream start codons (uATG, in red) relative to existing annotations (blue ATG). (B) Significantly more upstream ORFs (uORFs) were found relative to annotated ORFs for certain biological functions. P-values are False Discovery Rate adjusted. (C) An example of uORF (boxed and in red). (D) Size distribution of previously undescribed transcribed regions. (E) Transcribed regions previously covered by cDNA sequencing (8) in percentages. (F) An example of a previously undescribed transcribed region containing a polyA signal (shaded in red).