Research

Open Access

Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*

Evan H Hurowitz[*] and Patrick O Brown[*†]

Addresses: [*]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305-5307, USA. [†]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305-5428, USA.

Correspondence: Patrick O Brown. E-mail: pbrown@cmgm.stanford.edu

Published: 22 December 2003

Genome ***Biology*** 2003, **5:**R2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/5/1/R2

Received: 1 October 2003
Revised: 18 November 2003
Accepted: 21 November 2003

Abstract

Background: Although the protein-coding sequences in the *Saccharomyces cerevisiae* genome have been studied and annotated extensively, much less is known about the extent and characteristics of the untranslated regions of yeast mRNAs.

Results: We developed a 'Virtual Northern' method, using DNA microarrays for genome-wide systematic analysis of mRNA lengths. We used this method to measure mRNAs corresponding to 84% of the annotated open reading frames (ORFs) in the *S. cerevisiae* genome, with high precision and accuracy (measurement errors ± 6-7%). We found a close linear relationship between mRNA lengths and the lengths of known or predicted translated sequences; mRNAs were typically around 300 nucleotides longer than the translated sequences. Analysis of genes deviating from that relationship identified ORFs with annotation errors, ORFs that appear not to be *bona fide* genes, and potentially novel genes. Interestingly, we found that systematic differences in the total length of the untranslated sequences in mRNAs were related to the functions of the encoded proteins.

Conclusions: The Virtual Northern method provides a practical and efficient method for genome-scale analysis of transcript lengths. Approximately 12-15% of the yeast genome is represented in untranslated sequences of mRNAs. A systematic relationship between the lengths of the untranslated regions in yeast mRNAs and the functions of the proteins they encode may point to an important regulatory role for these sequences.

Background

Messenger RNAs can carry a great deal of information in addition to the sequence of the protein they encode. The untranslated regions (UTRs) of mRNAs encode regulatory signals for translation, stability and subcellular localization [1]. Previous genomic studies of RNA transcripts have been limited to large scale cDNA sequencing [2-4] and computational modeling [5,6]. Although computational methods for predicting coding sequences from genomic sequences generally perform well, our ability to recognize and predict the boundaries of transcripts from genome sequences is still relatively unreliable. Genome-scale measurement of the lengths of transcripts corresponding to each gene could provide important constraints for modeling boundaries and splicing patterns from genome sequence, and give us insights into the characteristics of the untranslated sequences of mRNAs.

We have developed a DNA microarray-based method for measuring transcript length on a genomic scale. This method, called the Virtual Northern, is a complementary approach to cDNA sequencing. It is lower in resolution, but relatively

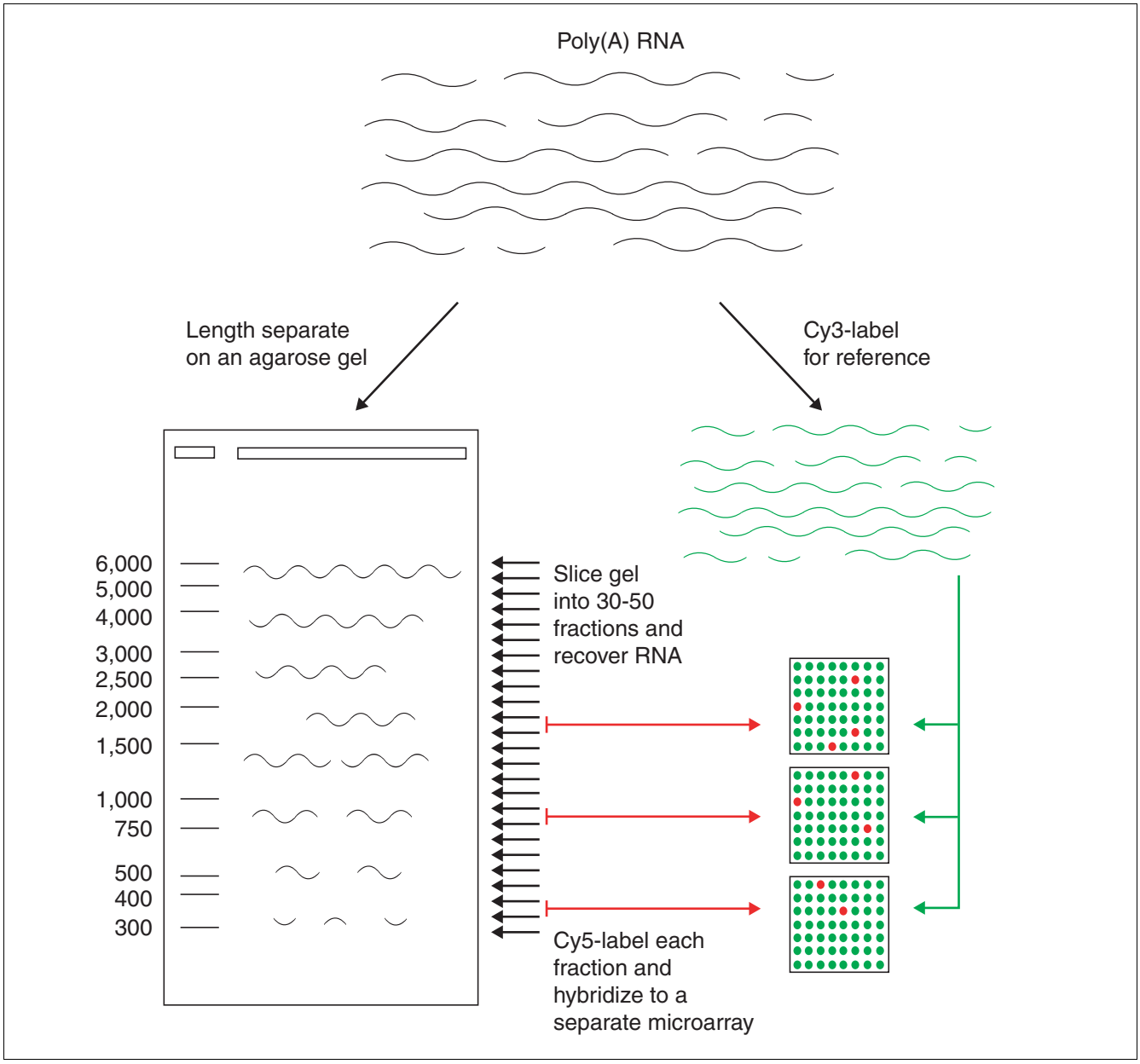

Figure 1
Virtual Northern scheme.

rapid and cheap, and less sensitive to the inabundance of rare messages or the difficulties in obtaining and cloning full length cDNAs [7].

The Virtual Northern scheme is summarized in Figure 1. Poly(A) RNA is separated by length on an agarose gel. The gel is sliced into a large number of thin slices, each of which contains RNAs from a narrow range of lengths [8]. RNA from each slice is then recovered, fluorescently labeled, and hybridized to a separate DNA microarray. Systematic analysis of the variation in the hybridization signal - across the series of gel slices for each DNA sequence represented on the microarrays - gives the length profile of all the transcripts that contain that sequence (Figure 2). Transcript lengths can then be determined from the peaks in those profiles. The method is essentially a Northern blot in reverse: instead of a gene probe labeled and hybridized to size-separated RNA immobilized on a membrane, size-separated RNA is labeled and hybridized to gene probes immobilized on a glass slide.

We chose to test the Virtual Northern method on the genome of the budding yeast *Saccharomyces cerevisiae* for several reasons. As the first fully-sequenced eukaryotic genome, the *S. cerevisiae* genome is complete, well-studied and well-

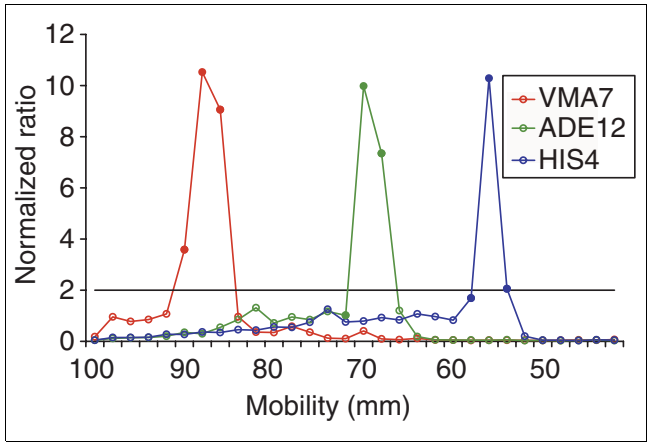

Figure 2
Examples of length profiles. Each length profile is a plot of the normalized ratio from all 30 microarrays. The x-axis is the distance of the midpoint of each gel slice from the origin. The black line indicates the threshold fluorescence ratio for peak recognition of 2.0, and the closed circles represent the three points used to calculate the midpoint of each peak.

annotated [9]. Furthermore, the genome is small and compact enough to allow the whole genome to be printed on a single microarray [10]. Finally, very little was known about the lengths or diversity of lengths of *S. cerevisiae* transcripts.

Results and discussion

Evaluation of the Virtual Northern method

The Virtual Northern method was applied to the genome of *S. cerevisiae*. After data quality filtering and array normalization, each length profile was searched for peaks at a ratio threshold of 2.0, resulting in 9,867 profiles with at least one and as many as six peaks (Tables 1 and 2). The threshold of 2.0 was deliberately low to provide a measurement for as many genes as possible. An ordered map of all of the known transcribed features of the *S. cerevisiae* genome was constructed based on a download of all annotated features from the *Saccharomyces* Genome Database (SGD) [11] as of 25 January 2003. The 9,867 profiles were matched against the ordered map, and as many peaks as possible were matched to the appropriate gene based solely on three criteria: features overlapped by the spot, genes directly adjacent to the spot on the ordered map, and the potential of a spot to crosshybridize to other genes with which it has high sequence similarity. As a result, at least one peak from 8,928 (90%) profiles could be matched to an annotated gene (Table 1). Of the 6,214 annotated non-Ty/non-pseudogene open reading frames (ORFs), 5,189 (84%) were matched to at least one and as many as six peaks (Table 3). Multiple measurements for the gene were typically provided by duplicated spots, overlapping genes and intergenic regions.

The precision of the Virtual Northern method was estimated based on the deviations between transcript lengths inferred from the Virtual Northern and their independently measured lengths using rapid amplification of cDNA ends (RACE) and on the variation among multiple measurements for the same gene. A major limitation to the precision with which the method can measure length is the means with which the length separation is achieved. In this case, 30 equal-width slices, representing RNAs of length 300-4,400 nucleotides (nts), were excised from an agarose gel. Due to the approximately exponential relationship between gel mobility and RNA length, the range of lengths represented by each equal-width slice increased from 35 to 326 nucleotides with decreasing distance from the origin over the range analyzed. However, as a fraction of RNA length each slice remained a constant 9%. Nevertheless, the average absolute deviation between the 94 RACE-analyzed transcripts and their Virtual Northern measurements was only 5.2%. Additionally, the

Table 1

Results of data filtering and profile inspection

Spot type	Number of spots on each array	Number of profiles after filtering (threshold 2.0)	Number of profiles matched to a gene
ORF	7,492 (6,398)	6,031 (5,498)	5,846 (5,304)
Intergenic	6,389	3,689	2,998
Intron	157 (133)	32	28
LTR	141	37	6
Mitochondrial	138 (71)	40 (28)	18 (12)
snRNA/snoRNA	44	3	2
tRNA	42	3	0
Ty	31 (28)	12	12
rRNA	23 (19)	17 (14)	17 (14)
Centromeric	16	1	0
Telomeric	10	2	1
Total	14,483 (13,291)	9,867 (9,311)	8,928 (8,377)

As these microarrays contained duplicated spots, the parentheses represent the number of unique spots or profiles in the dataset.

Table 2

Number of peaks per profile

Number of peaks	Number of profiles (threshold 2.0)
1	7,253
2	1,963
3	505
4	112
5	29
6	5
Total	9,867

Table 3

Number of times each gene is represented by a separate peak

Number of times represented	Number of genes
1	2,337
2	2,069
3	675
4	88
5	17
6	3
Total	5,189

deviations between multiple measurements for the same gene were calculated to estimate the error introduced by comparing data from different spots; this average absolute deviation was calculated to be only 0.93%. Therefore, we estimate the average error in length measurement to be approximately 6-7% (5.2% + 0.93%), the majority of which is due to the level of precision with which the gel was sliced.

We estimated the accuracy of the Virtual Northern method from the fraction of cases where a transcript's length measurement was completely spurious and unrelated to its true length. Several alternate sources of transcript length information were compared to the Virtual Northern dataset, and the fraction of discordant measurements was calculated for each source. Measurements were considered discordant when they differed from their alternate measurement by more than the average error of 7%. Firstly, six RACE-measured transcripts were excluded from the conversion of gel mobility to transcript length because they diverged too greatly from the best fit line. Six discrepancies out of 100 gives an error rate of 6.0%. Data from three articles [12-14] that reported traditional Northern blot analyses of large numbers of yeast genes were compared to our dataset. The published data of Richard *et al.*, Naitou *et al.* and Shiratori *et al.* showed five discrepancies out of 90 genes (5.6%), six out of 89 (6.7%) and three out of 38 (7.9%), respectively, for a total error rate of 6.5% (14 out of 217). Our dataset was also compared to data from a test run of the Virtual Northern method which covered a much smaller overall length range. The comparison showed 133 discrepancies out of 3,516 genes for an error rate of 3.8%. Finally, 939 out of the 9,867 profiles in the dataset (9.5%) were not matched to any known gene (Table 1). Comparison to the test run data probably provides an underestimate of the true error rate since the two agarose gels were run under identical conditions and any artifacts consistently associated with gel electrophoresis would be shared. On the other hand, the fraction of unexplained profiles is likely to be an overestimate of the true error rate since unexplained peaks could be due to novel or anomalous transcripts. Therefore we estimate the error rate to be approximately 6-7%, as indicated by the RACE and traditional Northern comparisons, which should be the most unbiased alternate sources of length information.

Relationship of transcript length to genome sequence

The transcript lengths of the 5,189 non-Ty/non-pseudogene ORFs for which we obtained a measurement were compared to their ORF length and to a theoretical maximum length. ORF lengths were taken as annotated in the SGD and corrected by subtracting the lengths of any annotated introns. The theoretical maximum length for each ORF was calculated to be the intron-corrected ORF length plus the lengths of its flanking intergenic regions. This parameter is only a simplistic theoretical maximum since it assumes that the transcript from one ORF never overlaps adjacent ORFs, which is clearly untrue; for example, there are cases where the ORFs themselves overlap [15].

The relationship between ORF length and transcript length is shown in Figure 3. Transcript length shows an extremely good fit to a linear relationship with ORF length. The parameters of the best fit line also show that, even over the entire range from 300 to 4,000 nucleotides, transcript length closely approximates the ORF length plus a fixed length of approximately 300 nucleotides. With an average ORF length for the entire yeast genome of 1,385 base pairs (bps), the average combined 5'- and 3'-UTR length can be calculated to be 256 nucleotides (1,385 times 0.0031, plus 312 minus 60 nucleotides for a poly(A) tail). A yeast mRNA can be well modeled as the ORF plus an average UTR length of 256 nucleotides and a 60 nucleotide poly(A) tail. This agrees well with the RACE data which shows an average combined UTR length of 276 nucleotides. Furthermore, since the mean intergenic region in the yeast is 536 bps, this simple mRNA model predicts that, on average, the transcribed portions of adjacent genes are nonoverlapping and separated by 280 bps (536 bps minus 256 bps).

Transcript length is plotted against the theoretical 'maximum' length in Figure 4. Excluded from this plot are ORFs

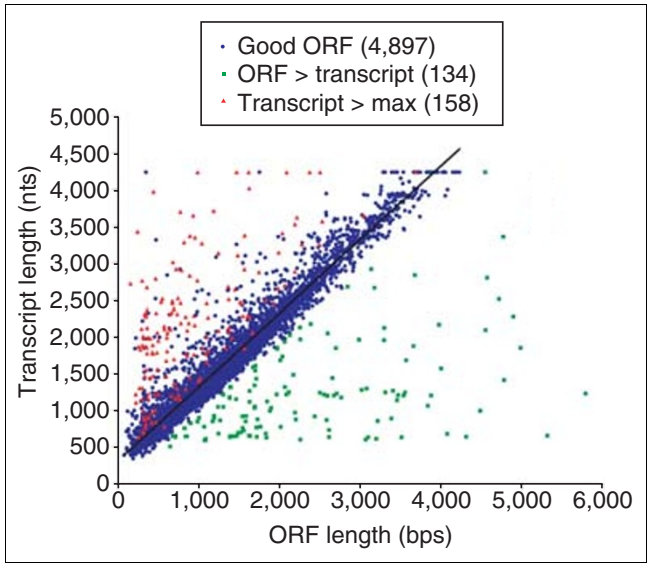

Figure 3
Relationship between ORF length and transcript length. Measured transcript length in nucleotides (nts) is plotted against ORF length in base pairs (bps). Green squares and red triangles indicate ORFs whose length is greater than their transcript length and transcripts greater than their theoretical maximum length, respectively. Parentheses indicate how many spots of each type are plotted. The black line is the linear least-squares fit to the blue 'Good' ORF circles. It has the parameters $y = 1.0031x + 311.88$ ($R^2 = 0.93$).

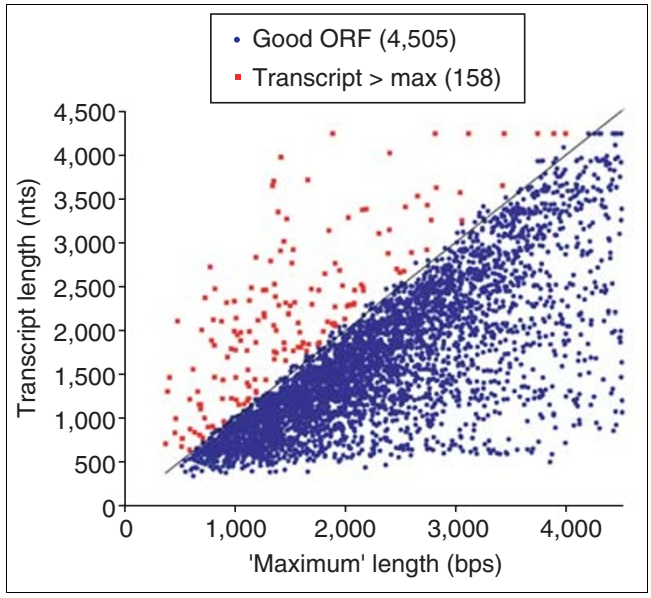

Figure 4
Relationship between the interval between flanking ORFs ('maximum' length) and transcript length. Measured transcript length in nucleotides (nts) for all nonoverlapping genes is plotted against the inter-ORF interval for each gene. Red squares indicate those transcripts whose length exceeds their theoretical maximum within the precision of the measurement. The identity line is shown in black for clarity. Parentheses indicate how many spots of each type are plotted.

which are annotated as overlapping another ORF since the transcript lengths of virtually all such ORFs exceed their theoretical maximum length. Although there is clearly no simple function which relates transcript length to maximum length, there is clearly a pronounced tendency for transcript length to be less than the theoretical 'maximum'. Accounting for the imprecision of the measurement, only 3% of the total points lie above the equality line in Figure 4. Thus, it appears that the transcribed portions of yeast genes rarely overlap adjacent ORFs.

Identification of biologically relevant groups of genes whose transcripts deviate significantly from the linear relationship between ORF and transcript length

Since transcript length showed such a good linear relationship with ORF length, we investigated the possibility that deviations from this relationship might be functionally significant. We calculated the 'expected' transcript length for each protein-coding gene from its ORF length, based on the linear equation described in the previous section, and determined the difference between the observed transcript length and the expected transcript length for every gene with a Virtual Northern measurement. We found no consistent correlation between the deviation of a transcript's length from the 'expected' length and its abundance [16], half-life [16] or translation rate [17]. We then downloaded the Gene Ontology (GO) annotation (biological process, molecular function and cellular component) for every yeast gene from the SGD [18]. We compared the distribution of the residuals, between observed and expected lengths, for every distinct GO classification containing more than five genes in our dataset to the distribution of all residuals in the dataset, using a Student's t-test (two-tailed, unequal variance). This analysis identified groups of genes with the same GO annotation whose transcripts were on average significantly longer or shorter than expected. Table 4 shows the ten most significant GO annotations along with their associated ontology and t-statistic.

As shown in Table 4, there were a number of functional classes of genes with shorter-than-average UTRs. The most significant class was the class of transcripts encoding ribosomal proteins. Those transcripts were the most significantly deviant in UTR length with an average length 222 nucleotides, approximately 63% of the genome-wide average. Strangely, there was an apparent lack of any functional classes of genes with transcripts larger than average; only the class of genes whose function was unknown were reasonably significant. Inspection of these transcripts revealed that the distribution was largely skewed by a handful of small, dubious ORFs whose signal probably derives from longer, adjacent or overlapping *bona fide* genes. When unnamed ORFs of less than 400 bps in length were filtered out, the 'unknown' class lost statistical significance.

Table 4

GO annotations whose transcripts had an average (observed minus expected) length significantly different than the average for all transcripts

GO annotation	t-statistic	Ontology	Comparison of distribution to average	Number of genes
Structural constituent of ribosome	4.39E-55	MF	Shorter	229
Chaperone activity	1.27E-10	MF	Shorter	82
Cytoplasm	4.49E-07	CC	Shorter	2,391
Unknown	5.79E-07	All	Longer	2,313
Proteasome core complex	6.53E-07	CC	Shorter	15
Response to oxidative stress	7.64E-07	BP	Shorter	35
Nucleolus	2.70E-06	CC	Shorter	220
Gluconeogenesis	3.24E-06	BP	Shorter	30
Peroxisome organization and biogenesis	8.91E-06	BP	Shorter	39
Chromatin assembly/ disassembly	1.43E-05	BP	Shorter	24

BP, biological process; MF, molecular function; CC, cellular component.

To determine if the unannotated genes were masking any functional classes with larger than average transcripts, we repeated the t-test analysis excluding the 2,313 genes lacking either a biological process or molecular function annotation. That analysis had no substantial effect on the functional classes whose transcripts were shorter than average, but it identified several functional classes with statistically significant, larger than average transcripts. The top five are listed in Table 5. These functional classes were particularly interesting because they are enriched for proteins involved in the regulation of dynamic cellular processes such as transcription, signal transduction, cell cycle control, and metabolism. Messenger RNA transcripts may be longer than expected due to the presence of additional sequences in their UTRs - important for the regulation of their translation, cellular localization, or decay.

Figure 5 illustrates the features of both the functional classes whose transcripts are, on average, shorter than expected and those that are, on average, longer. The distributions of residuals between observed and expected transcript lengths for functional classes whose transcripts are shorter than average, are exemplified by the distribution for transcripts encoding ribosomal proteins. Such distributions appear to be approximately normally distributed, similar to the distribution of residuals for all transcripts, but with a lower mean. On the other hand, the distributions of residuals for functional classes whose transcripts are longer than average, exemplified by the distribution for transcripts encoding transcription factors, are generally multimodal. Although the majority of transcripts in those functional classes have UTRs comparable in length to the genome-wide average, a subset of approximately 20% of those transcripts are from 180 to over 1,000

Table 5

Top five statistically significant GO annotations associated with transcripts that are longer than expected

GO annotation	t-statistic	Ontology	Number of genes
Nucleus	8.16E-06	CC	1,485
Transcription factor activity	9.43E-05	MF	311
Protein serine/threonine kinase	1.96E-03	MF	95
DNA binding activity	3.51E-03	MF	253
G1/S transition of mitotic cell cycle	8.12E-03	BP	53

BP, biological process; MF, molecular function; CC, cellular component.

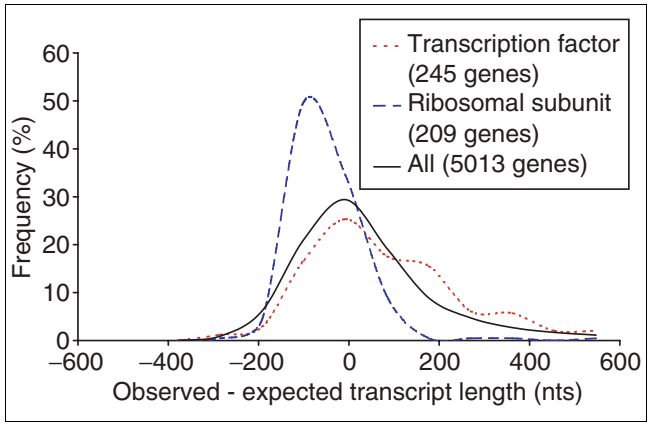

Figure 5
Distribution of residuals between observed and expected transcript lengths. The distributions of the residuals are plotted for all genes, for 209 genes annotated as ribosomal subunits and for 245 genes annotated as having transcription factor activity. The number of genes in each length bin is plotted as a percent of the total number of genes in that distribution.

nucleotides longer than expected. Whereas shorter-than-average UTRs appear to be a consistent, general feature of ribosomal subunit genes, there is nothing inherent about genes encoding transcription factors that predispose their UTRs to be longer than average; the group of genes is merely enriched for genes with long UTRs. The disproportionate representation of genes with regulatory functions among the genes with long UTRs suggests that those UTRs may contain sequences important for precise regulation. The significance of functional variation in UTR length merits further investigation.

Overlapping and questionable ORFs

The number of *bona fide* protein-coding genes in the *S. cerevisiae* genome is still unclear. Since the completion of the *S. cerevisiae* sequence in 1996, several computational and experimental approaches have been taken to identify previously undiscovered ORFs and to classify certain ORFs as questionable, dubious, spurious, bogus or hypothetical ([15,19-23] and references therein). The Virtual Northern method provides another piece of information to help distinguish true genes from ORFs erroneously annotated as hypothetical genes. Since the method can only measure the lengths of transcripts present in the experimental sample, this dataset only represents those genes expressed under the two conditions used in this experiment. Therefore, like any experiment that depends on gene expression for signal, we cannot confidently conclude that an unobserved gene does not exist. However, the presence of a transcript corresponding in length to, and hybridizing to, an ORF provides strong presumptive evidence that it represents a *bona fide* gene. As a demonstration, a list of 820 ORFs classified as dubious was downloaded from the SGD. This list represents a curation of experimental and computational data from the articles listed above [15,19-23], among others. Transcript measurements were available for 243 of those ORFs. Of those, 192 showed good agreement between ORF length and transcript length as discussed previously, indicating that these 192 of the 820 ORFs on the questionable list are likely to represent *bona fide* genes.

Overlapping ORFs encompass the majority of the questionable ORFs. Although examples of two real genes overlapping each other exist, it is likely that most overlapping ORF pairs contain only a single real gene [15]. The yeast genome has 873 non-Ty/non-pseudogene ORFs which overlap another gene. Virtual Northern profiles passed our filtering criteria for 695 of those. For virtually all available overlapping pairs, both members hybridized to a transcript of the same length, typically one that matched the predicted transcript length for one of the two ORFs. In those cases, we assigned the measured transcript to the ORF to which it more closely correlated, unless the assignment was contradicted by hybridization to an adjacent intergenic region. Overlapping ORFs with approximately the same ORF length or with distinct transcript lengths were each assigned the transcript length measurement from their respective hybridization result, and presumed to each encode a separate transcript. As a result, transcripts hybridizing to 428 ORFs were assigned to that ORF, and 213 were assigned to the overlapping ORF instead. The remaining 54 profiles had multiple peaks which matched both ORFs. Although our data suggest that there are at least 54 examples of pairs of *bona fide* genes with overlapping ORFs in yeast, most overlapping ORF pairs are represented by only a single transcript.

Gene families with highly homologous members

The effect on transcript length measurement of cross-hybridization between ORFs with high sequence similarity was examined. Sequence similarity data from an all-versus-all BLAST comparison of every annotated feature and intergenic region were available from the SGD. The profiles of all ORFs with more than 70% sequence identity to another locus in the genome were examined for the possibility of cross-hybridization. As a result, 14 separate gene families with numerous highly homologous members were identified, most of which consisted of subtelomeric ORFs (complete gene lists are available online [24]). The cross-hybridization potential of the COS, HXT, IMD, MAL, PRM, SSA and THI gene families could not be determined. Although the different members of each family all had similar measured transcript lengths, the ORFs comprising each family were all of nearly identical length, making it impossible to determine which family members are truly expressed and which cross-hybridize. On the other hand, despite an average nucleotide sequence identity of 75.6% between members, the RPP family showed no evidence of cross-hybridization: ORF lengths varied from 321 to 939 bps, and the measured transcript lengths followed suit, increasing from 499 to 1,090 nucleotides. The remaining six gene families, namely the AAD, PAU, YRF, YAR066W-like, YBL108W-like and YDR543C-like genes, are all subtelomeric. Their Virtual Northern profiles were characterized by a large number of

peaks, subsets of which are shared by their various family members, consistent with extensive cross-hybridization.

ORF lengths greater than the measured length of the cognate transcripts

As shown in Figure 3, there are 134 genes in the dataset whose ORF length is greater than their measured transcript length. In order to determine whether these measurements are spurious or biologically relevant anomalies, traditional Northern blots were performed on six of the 134 genes - YML104C, YML116W, YML123C, YMR024W, YMR054W and YLR160C. In five out of six cases, the transcript length, as measured by the traditional analysis, was greater than the ORF length and did not match the Virtual Northern measurement. In the case of ASP3 (YLR160C), however, traditional Northern analysis confirmed a transcript length of 600 nucleotides for a 1,089 bp ORF. A search of the literature showed that the transcriptional initiation site of the ASP3 gene is regulated by nitrogen catabolite levels. In conditions where the cell is not starved for nitrogen, a transcript of 600 nucleotides is produced by initiation at a downstream promoter, encoding an amino-terminally truncated protein. Only during nitrogen starvation is the full-length 1,200 nucleotide transcript produced [25]. A search of the other genes among the 134 on the list identified two genes, CBP1 (YJL209W) [26] and SIR1 (YKR101W) [27], that are also known from published studies to show transcript length regulation consistent with the Virtual Northern results. Although it is likely, based on the traditional Northern analysis, that the majority of the 134 anomalously short transcript lengths represent inaccurate measurements, many of those genes are potential examples of regulated transcript length.

Adjacent ORF anomalies

The transcript length information was matched to an ordered map of the yeast genome, and adjacent ORF pairs were identified that showed not only the same transcript length but also a transcript length anomalously long enough to contain both ORFs. Forty-five loci matching these criteria were identified, including five loci for which three adjacent ORFs are involved. Nine of these loci had been identified previously [15,21]. All 45 loci were analyzed by reverse transcriptase PCR (RT-PCR) analysis to verify that adjacent ORF pairs are in fact present on the same transcript, and 50 of the 95 ORFs were subjected to RACE analysis. All 95 of these ORFs were crosschecked against their orthologs in three other newly sequenced budding yeast species [21] using the Synteny Viewer at the SGD [28]. The results are summarized in Figure 6. In 32 loci the ORFs are arranged in parallel, that is, head to tail. In nine of those cases, the ORFs share the same transcript and their annotation as separate ORFs appears to be due to a sequence error. In five cases, the ORFs are apparently distinct genes but the 3'-UTR of the upstream ORF contains all or part of the downstream ORF. In the remaining 18 cases, one of the two ORFs is present entirely or in part in the UTR of the other gene but is not conserved in other yeast species and therefore is not likely to be a *bona fide* gene. In three loci the ORFs are in a convergent, that is, tail to tail, arrangement. As above, in two of those cases a pair of real ORFs have overlapping 3'-UTRs, and in the remaining case the 3'-UTR overlaps an ORF which is not conserved in other species. In eight loci the ORFs are in a divergent, that is, head to head, orientation. Also as above, in two cases a pair of real ORFs have overlapping 5'-UTRs, and in four cases the 5'-UTR overlaps an ORF which is not conserved in other species. In the remaining two loci, the transcripts do not overlap but are coincidentally the same (unusually long) length. Not shown in Figure 6 are two cases of transcripts apparently involving three ORFs each. The evidence suggests that all six of those ORFs may be real genes, with overlapping UTRs in one case and overlapping ORFs in the other. More detailed descriptions of each of these 95 ORFs are available online [24]. Although this class of anomaly was examined specifically for the purpose of identifying ORFs which should be fused or extended due to introns or sequencing errors, no new introns and few sequence errors were identified. The most common reason for two adjacent ORFs to share the same anomalously long transcript was that the 3' or 5' UTR of one of the two ORFs was anomalously long and significantly overlapped the other ORF (76% of the loci). In most of those cases, the ORF that lacked a distinct transcript measurement of its own was not conserved in other species and therefore may not be a *bona fide* gene (68%). In some cases, the lack of a detectable transcript may be attributable to a lack of expression under the limited range of conditions represented in our analysis.

Transcripts longer than the interval between flanking ORFs

As shown in Figure 4, there are only 158 ORFs whose transcript length exceeds the length of the interval between the flanking ORFs. Of those, 37 are among the ORFs described in the previous section on adjacent ORF anomalies. The remaining 121 cases presumably include both erroneous measurements and additional cases of unusually long transcripts that overlap adjacent ORFs. Although this set of genes was not experimentally characterized further, 40 of those transcript measurements were corroborated by hybridization results from adjacent independent (generally intergenic) spots, and are therefore likely to be accurate measurements of abnormally long transcripts. The other 81, many of which would correspond to extremely large differences between transcript and ORF length, are not corroborated by the expected hybridization pattern to adjacent sequences, and thus are more likely to be erroneous measurements.

Introns and intron-containing ORFs

As shown in Table 1, of the 133 unique introns represented on the microarrays, 32 hybridization profiles passed data filtering. Exactly 30 of those 32 intron profiles had a corresponding ORF profile in the dataset. The length measured by each of those intron profiles was compared to the transcript length measured by the corresponding ORF. The length

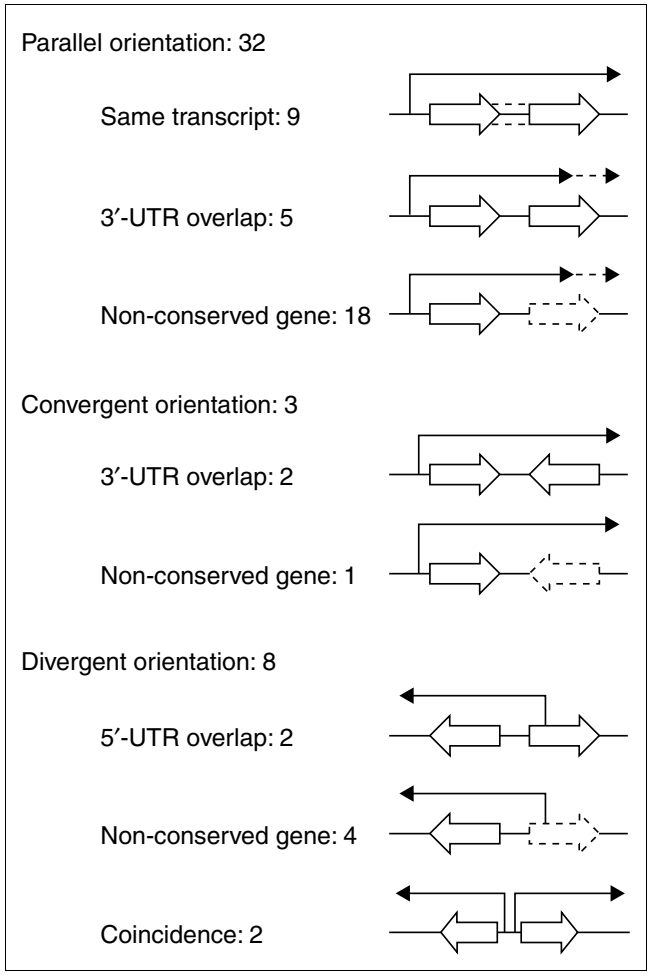

Figure 6
Adjacent ORF anomalies. Eight different classes of adjacent ORF anomalies are pictured schematically. Open boxes represent ORFs, and their arrows represent their orientation from translational start to stop. Dashed open boxes represent non-conserved ORFs. The solid lines indicate the transcripts detected by Virtual Northern and RACE analysis. The dashed portions of those lines represent variation in the extent of 3'-UTR overlap between different cases. The number following each title lists the number of loci in that class.

measurements were identical in three cases. In 25 cases (83%), the transcript length inferred from the intron's hybridization profile was consistent with the measured length of the mRNA, inferred from the ORF hybridization profile, plus the length of the intron. Thus it appeared to represent an unspliced precursor. Finally, two gave unexplained, probably erroneous lengths.

Although splicing and intron degradation are not instantaneous [29], it was uncertain, *a priori*, what signal the intron spots would detect. It was comforting to observe that the majority of those spots detected the small amount of unspliced message in the cell. More confusing were the three profiles with a peak corresponding to the length of the spliced transcript. The observation for one of those genes, HAC1, was explained by its biology. The translation of HAC1 protein is regulated by the unconventional splicing of HAC1 mRNA in response to the accumulation of unfolded proteins in the endoplasmic reticulum (ER) [30]. Thus, in unstressed cells, both ORF and intron sequence would hybridize to the same unspliced mRNA. Furthermore, the annotation of the other two genes, YHL050C and YJR079W, has recently been brought into question [21]. Comparison of those two genes to their orthologs in other yeast species has suggested that the boundaries of the intron of YHL050C are incorrectly annotated, and that YJR079W is not actually spliced. Thus, the 'intron' sequence represented by those two profiles probably includes or consists entirely of exon sequence.

Since a handful of intron spots were able to detect unspliced mRNA, the profiles of all intron-containing ORFs were examined for secondary peaks consistent with an unspliced precursor. Four ORF profiles (YDR064W, YGL103W, YKL002W and YKL156W) were found to have a secondary peak consistent with an unspliced precursor.

Transposons

The yeast genome contains five different classes of retrotransposable elements denoted as Ty1 to Ty5 [31]. Elements from the first four classes exceed 5 kb, which is above the measured length range for this dataset. Therefore, with the exception of Ty5, the Ty spots should not be able to detect full-length transposon RNA. However, examination of Ty transcripts typically identifies partial length transcripts around 3 kb in addition to full length [32]. The 12 Ty profiles along with seven Ty ORF profiles identified transcripts of 0.7 and 3.5 kb for Ty1, 3.5 kb for Ty2, and 1.9 and 3.3 kb for Ty3. No reliable transcripts were identified for Ty4 or Ty5.

Mitochondrial transcripts

The yeast mitochondrial genome encodes 24 tRNAs, two rRNAs, 9S RNA and 19 ORFs [33]. These genes are transcribed into approximately 14 multigene primary transcripts which are then extensively spliced and nucleolytically processed [34]. The mitochondrial sequences represented on the microarrays were therefore often theoretically capable of hybridizing to a number of different processing intermediates and alternately spliced transcripts. Interestingly, no mitochondrial sequence had a Virtual Northern hybridization profile with more than two detectable peaks. Nevertheless, only 12 unique sequences could be unambiguously assigned to a specific gene. Most of the sequences could possibly represent as many as four different genes.

rRNA, snRNA, snoRNA and tRNA

As annotated in the SGD, the nuclear yeast genome is currently known to encode 25S, 18S, 5.8S and 5S rRNAs, six snRNAs, 67 snoRNAs, five stable RNAs and 299 tRNAs. Since most stable non-protein coding RNAs are shorter than 300 nucleotides, only the 25S and 18S rRNAs, two snRNAs, eight snoRNAs, and four stable RNAs are within the measured

length range for this dataset. Of those, profiles exist for the two rRNAs, the snRNA LSR1, the three snoRNAs (SNR17A, SNR42 and SNR37), and the RNase P RNA component, RPR1. Also, profiles were obtained for the 5S rRNA, SNR44 and three tRNAs, all of which are too small to have been detected unless their electrophoretic mobilities were aberrant. The measured stable non-protein coding RNAs provide the most aberrant measurements in the dataset. The Virtual Northern profiles for the three snoRNAs matched their annotated lengths almost perfectly, with deviations of, at most, 11 nucleotides. The lengths inferred from the data for LSR1, and the 25S and 18S rRNAs, however, were much shorter than their annotated lengths. In contrast, the measured lengths of the large and small mitochondrial rRNAs exceeded their annotated length by as much as 11%. Although their aberrant migration is likely to be related to their stable folded structure, it is unclear why rRNAs deviate from their annotated length in such a nonsystematic way.

Intergenic regions

The yeast genome is currently annotated to contain 6,158 intergenic regions. After data filtering, 3,701 unique profiles remained, representing 3,547 (58%) of the total intergenic regions. The purpose of measuring the lengths of transcripts that hybridize to intergenic sequences was two-fold. Intergenic sequences can provide an independent measurement for the length of transcripts corresponding to adjacent ORFs, by virtue of hybridization of the corresponding UTRs to the 'intergenic' sequences, and they can potentially identify novel transcripts. Analysis of the entire dataset resulted in the match of 2,877 intergenic hybridization profiles to one adjacent gene, and 112 to both adjacent genes, leaving 712 unexplained. An intergenic hybridization profile was considered a match to an adjacent gene when its inferred transcript length deviated from the transcript length inferred from the adjacent ORF's profile by less than 7%.

The results for most (81%) of the intergenic sequences could therefore be explained by hybridization to the UTRs of adjacent genes. The average fluorescent hybridization signal measured for the intergenic spots was much lower than that of their ORF counterparts. Thus, a significantly smaller percentage of intergenic spots than ORF spots passed the data filtering criteria (Table 1). This was probably due to short yeast UTRs providing relatively little sequence for hybridization. Furthermore, the matches between the hybridization profiles of intergenic sequences and those of the adjacent ORFs show a strong 3'-end bias in which 78% of the matching intergenic sequences are downstream of the ORF whose transcript measurement they match. It was expected that 3'-UTRs should give a greater hybridization signal than 5'-UTRs. This is because they are on average longer [1], and because the use of dT primer, in addition to random primer, in the synthesis of the cDNA hybridized to the microarrays, may have disproportionately represented the 3'-UTRs in the fluorescently-labeled probe.

The existence of 712 intergenic profiles that do not match either adjacent gene provides the possibility of identifying novel transcripts. Since it is unlikely that 712 novel genes remain to be discovered in the *S. cerevisiae* genome, it is important to separate the most promising targets from the large number of profiles that presumably represent spurious or inaccurate measurements. First, the length measurement from each profile was compared to the length of the intergenic interval itself. Although it is possible for a novel transcript to overlap other genes or for an intergenic interval's true length to be masked by a spurious ORF, transcript length measurements that vastly exceed the length of the intergenic interval from which they derive are more likely to be erroneous. On that basis, 450 intergenic regions were excluded from further consideration. An additional 48 intergenic intervals had high sequence similarity to other loci, and hybridization profiles consistent with a cross-hybridization artifact. Finally, the remaining 214 intergenic intervals were compared to 121 unique novel genes proposed by Oshiro *et al.* [19], Kessler *et al.* [20], Brachat *et al.* [22] and Cliften *et al.* [23]. Out of those 214 intergenic regions, 48 were found to contain at least one of the proposed novel genes. Thus, our list appears to be strongly enriched for novel genes.

Conclusions

We developed a 'Virtual Northern' method for measuring RNA transcript lengths on a genomic scale, and applied it to the genome of the yeast *S. cerevisiae*. Overall, the technique worked well. Despite limiting ourselves to examining RNA representing only two growth conditions, and a length range of 300-4,400 nucleotides, we obtained 9,867 electrophoretic migration profiles, which provided a length measurement for 84% of the annotated ORFs in the yeast genome. Evaluation of the precision of the length measurements estimated measurement error at about 6-7% of the transcript's length. It is likely that further narrowing the length range of each fraction, for example by running a gel longer and cutting more slices, would not only increase precision but would also further sample each peak, providing an additional parameter which could be used to more accurately discriminate true peaks from spurious measurements. Evaluation of the accuracy of the method suggested that only about 6-7% of the measurements were spurious. A higher level of accuracy can generally be obtained by raising the threshold for peak assignment.

Virtual Northern analysis of the yeast genome has provided a number of insights into its organization. A measurement of the length of nearly every mRNA in the yeast genome has uncovered a strongly linear relationship between the length of an ORF and the length of its RNA transcript. This relationship indicates that the average mRNA consists of the protein coding sequence and 260 nucleotides of untranslated sequence, assuming a 60-nucleotide poly(A) tail [35]. Further comparisons of the measured transcript lengths to the genome

sequence have shown that, despite the compactness of the yeast genome, there are few overlaps between transcripts of adjacent genes.

Our results show that the untranslated regions of mRNAs comprise 12-15% of the yeast genome sequence, highlighting the potential importance of these sequences in the post-transcriptional regulation of mRNAs. A striking relationship between gene function and the length of the UTRs provides further evidence for such a regulatory role. Since the functional classes of genes with longer than average UTRs were predominantly involved in the regulation of cellular processes, especially transcription, we hypothesize that those transcripts have UTRs longer than average because they contain additional sequences that play a role in regulating their fate. There are now numerous examples of UTR sequences that participate in the regulation of mRNA export from the nucleus, localization in the cell, translation, or decay [1]. Thus, those transcripts whose UTRs are longer than average are good targets for the investigation of the post-transcriptional regulation of mRNA. The possibility that certain functional classes, such as the ribosomal subunit genes, are under a selective pressure for short UTR length also merits further investigation.

Our global analysis of transcript lengths has provided new evidence for the existence of some genes and against the existence of others, identified possible genome annotation errors, and revealed a novel link between UTRs and gene function. Examination of the length data from ORFs whose status as *bona fide*, transcribed genes is in question (for example, ORFs that overlap each other) has provided evidence for the existence of some of these genes and additional evidence against the existence of others. Analysis of genes whose measured transcript length was less than that of the corresponding ORF identified possible examples of regulated transcript length. Analysis of pairs of adjacent ORFs with unusually long transcript measurements has identified loci that may be misannotated. Finally, the length measurements obtained from intergenic spots identify potential novel genes.

The Virtual Northern technique should be readily applicable to the study of any nucleic acid sample from any organism. Our analyses focused on genome organization and annotation. Such analyses would be equally useful in other organisms with compact genomes. For example, the measurement of transcript length in a bacterial genome would provide a quick identification of the operon structure of that species. In species with a transcriptome more diverse and complicated than *S. cerevisiae*, genome-wide transcript length measurement could also be used to explore transcript length changes during RNA splicing and nucleolytic processing, or to assay the diversity of transcripts produced from a single gene under separate conditions or from separate tissues.

The method may also prove useful for a range of technical purposes. For example, since Virtual Northern measurements are less susceptible than cDNA libraries to underestimating transcript length, they could be used as a verification that nominally full-length cDNAs are truly full length. An analogous technique could be used for genome-wide measurement of electrophoretically (or otherwise) separated DNA fragments, such as restriction fragments. Since the length profiles of thousands of arrayed DNA elements such as cDNAs, sequence-tagged sites (STSs) or genomic DNA clones can be determined in a single experiment, the technique could be used for genome sequence assembly, high-throughput clone fingerprinting, mapping cDNAs, STSs, or clones to chromosomes, or restriction mapping whole genomes.

Materials and methods

RNA preparation

S. cerevisiae strain S288C was grown at 30°C to OD_{600} = 1.0 in yeast extract/peptone/dextrose (YPD) medium and to OD_{600} = 1.7 in yeast extract/peptone/galactose (YPGal) medium. Total RNA was prepared from each culture by hot phenol extraction [36], poly(A) purified using the Oligotex™ mRNA kit from Qiagen, and pooled.

Length fractionation

A 50 μg aliquot of the pooled poly(A) RNA and a RNA ladder (a mix of the Millennium™ and Century™ markers from Ambion) were heat denatured for 10 min at 65°C in formamide, placed on ice and then immediately loaded on a 1.2% low melting point agarose gel in 1X TAE running buffer. The poly(A) RNA was loaded in a wide (5 cm) well to prevent overloading the gel. The RNA was then separated by electrophoresis at high voltage (approximately 9 V/cm) for 2 hours. During electrophoresis, the running buffer was passively recirculated, and the apparatus was cooled to prevent the gel and running buffer from overheating. The ladder lane was excised from the gel and visualized by ethidium bromide staining. Based on the mobility of the ladder bands, a portion of the poly(A) RNA lane, corresponding approximately to RNAs of length 300-4,400 nucleotides was cut into 30 slices of 2 mm each. RNA was then recovered from each length fraction by β-agarase digestion. Each agarose slice was melted at 70°C for 10 min, and then digested with 2 Units of AgarACE enzyme (Promega) at 42°C for 2 hours. The products from each agarase reaction were recovered using a Microcon-30 column (Amicon) and washed twice with 10 mM Tris pH 7.0 buffer.

Microarray analysis

One third of the recovered material from each length fraction was reverse transcribed in the presence of amino allyl-dUTP. A mix of dT_{20} and random nonamer was used to prime the reaction in order to maximize the reverse transcription of both polyadenylated and non-polyadenylated RNAs. The resulting cDNA was fluorescently labeled by coupling reactive

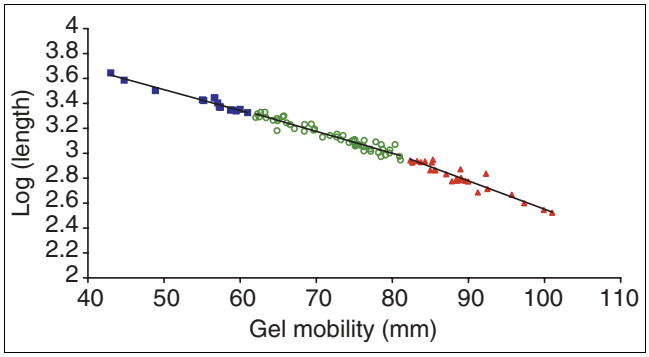

Figure 7
Calibrating the relationship between gel mobility and transcript length. The precise gel mobilities of 94 transcripts are plotted against the base 10 logs of their exact lengths based on their 5'- and 3'-ends as determined by RACE. The exponential fits to the top, middle and bottom 2 cm of gel are shown by three black lines with the best fit parameters $y = -0.0167x + 4.3456$ ($R^2 = 0.94$), $y = -0.0176x + 4.4088$ ($R^2 = 0.92$) and $y = -0.0228x + 4.8286$ ($R^2 = 0.91$) respectively.

Cy5 to the amino allyl groups on the incorporated dUTP. Whole-genome yeast microarrays containing spotted DNA segments, representing every ORF and intergenic region, were produced [10], and each labeled length fraction was hybridized to a separate microarray [37]. To provide an internal hybridization reference, a 1.5 μg aliquot of the pooled poly(A) RNA was reverse transcribed using the same protocol as the length fractions, fluorescently labeled by coupling the cDNA to Cy3, and included in each hybridization. Microarrays were scanned with an Axon Instruments (Foster City, CA) scanner, and the data collected with GENEPIX PRO 3.0 software (Axon Instruments). The Cy5/Cy3 fluorescence ratio data was then filtered for quality. Spots with defects apparent from visual inspection, or with a background-corrected intensity in the reference channel less than 225 were excluded from further analysis. For each electrophoretic profile based on the 30 measurements for each analyzed DNA sequence, the median background-corrected reference channel intensity for all spots in the profile was calculated. Profiles with more than nine missing data points, or whose median background-corrected reference channel intensity was less than 250, were excluded from further analysis.

For normalization between the length fractions, an internal standard was prepared with a pool of *in vitro*-transcribed *Bacillus subtilis* RNAs [16]. PCR products representing five different *B. subtilis* DNAs were printed onto the microarrays and *in vitro* transcribed into RNA. A mix of 400 pg of each *B. subtilis* RNA was doped into each gel slice during the agarase digestion, and into each 1.5 μg reference aliquot before labeling. The internal standard gives us a way to account for differences in the efficiency of RNA recovery between the length fractions.

Peak finding

For each normalized length profile, a particular length fraction was considered a peak if its normalized fluorescence ratio was greater than a specified threshold (Figure 2) and greater than the normalized fluorescence ratios of the two flanking length fractions. In order to estimate the gel mobility for each transcript more precisely, each peak was fit to a normal distribution. For each peak, the Gaussian equation of form $y = A*\exp[-(x-M)^2/V]$ that uniquely fits the peak fraction and its two adjacent length fractions (Figure 2) was calculated where x is the midpoint of each length fraction's position in the gel in millimeters and y is the normalized ratio. The mean of the distribution, M, for each peak was taken as the estimated mobility of the transcript represented by that peak.

RACE analysis

The precise 5'- and 3'-ends of selected transcripts were determined by the rapid amplification of cDNA ends (RACE) technique using the FirstChoice™ RLM-RACE kit from Ambion. To determine the 3'-end, the pooled poly(A) RNA was reverse transcribed using a dT primer with a 5' linker, and the resulting cDNAs are amplified by PCR using a gene-specific primer in conjunction with a primer based on the 5' linker. To determine the 5'-end, the pooled poly(A) RNA was subjected to a series of enzymatic reactions that replaced the 5'-cap with an oligoribonucleotide [38]. The resulting RNA was then reverse transcribed, and PCR amplified using a gene-specific primer in conjunction with a primer specific to the capping oligo. For additional specificity, the product of the first amplification was subjected to second round of PCR using another gene-specific primer upstream of the first. PCR amplification products from both 3'- and second round 5'-RACE were cloned using the TOPO TA cloning® kit from Invitrogen. Plasmid DNA was purified from bacterial culture with QIAprep® Spin and Turbo kits from Qiagen, cycle sequenced using ABI BigDye sequencing chemistry, and analyzed on an ABI Prism 3100 Genetic Analyzer (Applied Biosystems). Plasmid inserts containing the proper primer sequence were matched to the *S. cerevisiae* genome sequence using WU-BLAST [39] at SGD to find the precise 5'- or 3'-base pair of each transcript.

Conversion of gel mobility to transcript length

The estimated mobility of a set of 94 transcripts was compared to their exact length as deduced from the sequence of the 5' and 3' ends, determined by RACE and the published genome sequence [8]. The exact length of RACE-measured transcripts was taken to be the distance in nucleotides between the first and last nucleotide of each transcript relative to the genomic sequence, minus any introns, plus 60 nucleotides to represent a poly(A) tail [35]. As expected, the data showed an excellent fit to an exponential relationship between gel mobility and transcript length. The data were split into three equal mobility ranges (the top 2 cm of gel, the middle 2 cm and the bottom 2 cm), and the best fit to an exponential equation was calculated for each mobility range (Figure 7). The precise mobility of every peak within each range

was then converted to an inferred transcript length using the corresponding exponential equation.

Northern blot analysis

Traditional Northern blot analysis was performed on selected genes using the NorthernMax™ kit from Ambion. Pooled poly(A) RNA (0.5 μg) was separated in a 1%TAE agarose gel, blotted to BrightStar-Plus™ membrane (Ambion), and hybridized to labeled PCR products. The same PCR amplicons used to print the microarrays were labeled with biotinylated dATP using the reverse PCR primer. Hybridizations were performed overnight at 45°C. The resulting blots were washed and imaged using the BrightStar™ BioDetect™ kit from Ambion and autoradiographic film.

Acknowledgements

We thank Yulei Wang for providing the *B. subtilis* doping control RNAs. This work was supported by a grant from the NHGRI and by the Howard Hughes Medical Institute. P.O.B. is an associate investigator of and E.H.H. is a predoctoral fellow of the Howard Hughes Medical Institute.

References

1. Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome Biol* 2002, **3:**reviews0004.1-0004.10.
2. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H *et al.*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563-573.
3. Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y *et al.*: **Functional annotation of a full-length *Arabidopsis* cDNA collection.** *Science* 2002, **296:**141-145.
4. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S *et al.*: **A *Drosophila* full-length cDNA resource.** *Genome Biol* 2002, **3:**research0080.1-0080.8.
5. Graber JH, McAllister GD, Smith TF: **Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites.** *Nucleic Acids Res* 2002, **30:**1851-1858.
6. Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res* 2002, **12:**458-461.
7. Sugahara Y, Carninci P, Itoh M, Shibata K, Konno H, Endo T, Muramatsu M, Hayashizaki Y: **Comparative evaluation of 5'-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries.** *Gene* 2001, **263:**93-102.
8. Woolford JL Jr, Hereford LM, Rosbash M: **Isolation of cloned DNA sequences containing ribosomal protein genes from *Saccharomyces cerevisiae*.** *Cell* 1979, **18:**1247-1259.
9. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al.*: **Life with 6000 genes.** *Science* 1996, **274:**546, 563-567.
10. Lieb JD, Liu X, Botstein D, Brown PO: **Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association.** *Nat Genet* 2001, **28:**327-334.
11. ***Saccharomyces* Genome Database** [http://www.yeastgenome.org/]
12. Richard GF, Fairhead C, Dujon B: **Complete transcriptional map of yeast chromosome XI in different life conditions.** *J Mol Biol* 1997, **268:**303-321.
13. Naitou M, Hagiwara H, Hanaoka F, Eki T, Murakami Y: **Expression profiles of transcripts from 126 open reading frames in the entire chromosome VI of *Saccharomyces cerevisiae* by systematic northern analyses.** *Yeast* 1997, **13:**1275-1290.
14. Shiratori A, Shibata T, Arisawa M, Hanaoka F, Murakami Y, Eki T: **Systematic identification, classification, and characterization of the open reading frames which encode novel helicase-related proteins in *Saccharomyces cerevisiae* by gene disruption and Northern analysis.** *Yeast* 1999, **15:**219-253.
15. Wood V, Rutherford KM, Ivens A, Rajandream MA, Barrell B: **A re-annotation of the *Saccharomyces cerevisiae* genome.** *Comp Funct Genom* 2001, **2:**143-154.
16. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO: **Precision and functional specificity in mRNA decay.** *Proc Natl Acad Sci USA* 2002, **99:**5860-5865.
17. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D: **Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2003, **100:**3889-3894.
18. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G *et al.*: **Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO).** *Nucleic Acids Res* 2002, **30:**69-72.
19. Oshiro G, Wodicka LM, Washburn MP, Yates JR 3rd, Lockhart DJ, Winzeler EA: **Parallel identification of new genes in *Saccharomyces cerevisiae*.** *Genome Res* 2002, **12:**1210-1220.
20. Kessler MM, Zeng Q, Hogan S, Cook R, Morales AJ, Cottarel G: **Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome.** *Genome Res* 2003, **13:**264-271.
21. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423:**241-254.
22. Brachat S, Dietrich FS, Voegeli S, Zhang Z, Stuart L, Lerch A, Gates K, Gaffney T, Philippsen P: **Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*.** *Genome Biol* 2003, **4:**R45.
23. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301:**71-76.
24. **Virtual Northerns - web supplement** [http://microarray-pubs.stanford.edu/vnorth/]
25. Kim KW, Kamerud JQ, Livingston DM, Roon RJ: **Asparaginase II of *Saccharomyces cerevisiae*. Characterization of the ASP3 gene.** *J Biol Chem* 1988, **263:**11948-11953.
26. Mayer SA, Dieckmann CL: **The yeast CBP1 gene produces two differentially regulated transcripts by alternative 3'-end formation.** *Mol Cell Biol* 1989, **9:**4161-4169.
27. Stone EM, Swanson MJ, Romeo AM, Hicks JB, Sternglanz R: **The SIR1 gene of *Saccharomyces cerevisiae* and its role as an extragenic suppressor of several mating-defective mutants.** *Mol Cell Biol* 1991, **11:**2253-2262.
28. **Budding Yeasts Genome Comparison** [http://db.yeastgenome.org/cgi-bin/FUNGI/FungiMap]
29. Elliott DJ, Rosbash M: **Yeast pre-mRNA is composed of two populations with distinct kinetic properties.** *Exp Cell Res* 1996, **229:**181-188.
30. Kawahara T, Yanagi H, Yura T, Mori K: **Endoplasmic reticulum stress-induced mRNA splicing permits synthesis of transcription factor Hac1p/Ern4p that activates the unfolded protein response.** *Mol Biol Cell* 1997, **8:**1845-1862.
31. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF: **Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence.** *Genome Res* 1998, **8:**464-478.
32. Boeke JD, Sandmeyer SB: **Yeast transposable elements.** In *The molecular and cellular biology of the yeast Saccharomyces Volume 1*. Edited by: Broach J, Jones E, Pringle J. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 1991:193-261.
33. Foury F, Roganti T, Lecrenier N, Purnelle B: **The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*.** *FEBS Lett* 1998, **440:**325-331.
34. Pon L, Schatz G: **Biogenesis of yeast mitochondria.** In *The molecular and cellular biology of the yeast Saccharomyces Volume 1*. Edited by: Broach J, Jones E, Pringle J. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 1991:333-406.
35. Brown CE, Sachs AB: **Poly(A) tail length control in *Saccharomyces cerevisiae* occurs by message-specific deadenylation.** *Mol Cell Biol* 1998, **18:**6548-6559.
36. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J Jr, Boguski MS *et al.*: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283:**83-87.
37. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.

38. Maruyama K, Sugano S: **Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides.** *Gene* 1994, **138:**171-174.
39. Altschul SF, Gish W: **Local alignment statistics.** *Methods Enzymol* 1996, **266:**460-480.