Supplementary information for

Extensive transcriptional heterogeneity revealed by isoform profiling

Vicent Pelechano[1]†, Wu Wei[1,2]† and Lars M. Steinmetz[1,2]*.

[1]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg,Germany.

[2]Stanford Genome Technology Center, Stanford University, Palo Alto, CA 94304, USA.

*Correspondence to: larsms@embl.de

†These authors contributed equally to this work

This PDF file includes

Supplementary Figures and Legends S1-S24

Supplementary Methods

Supplementary Discussion

Supplementary Tables S1-S5

Supplementary Data S1-S10

Supplementary References (29-49)

Supplementary Figures and Legends

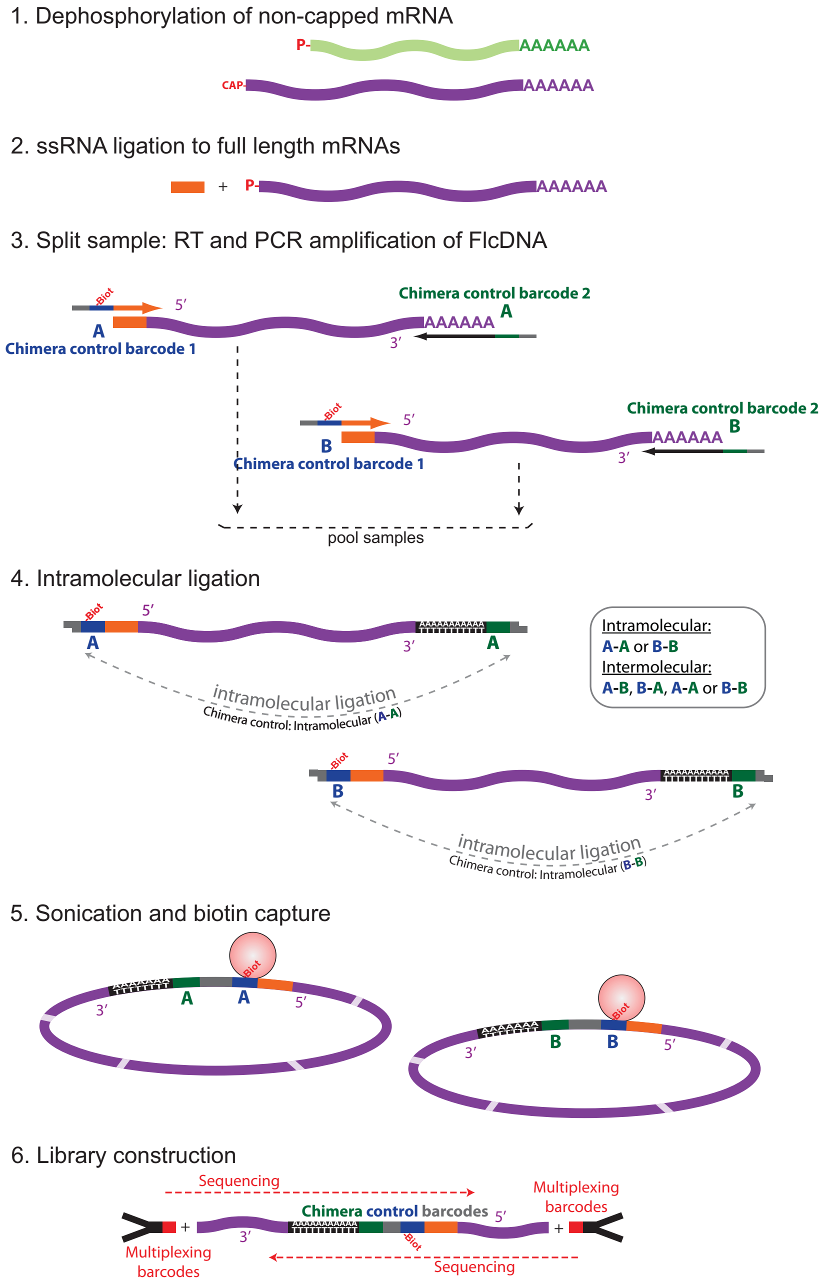


Figure S1. Detailed TIF-Seq protocol. (1) Non-capped RNA molecules are dephosphorylated. (2) Using the oligo-capping method[29] a known oligo (orange) is ligated to the capped mRNA molecules. (3) Sample is split and barcoded full-length cDNA is produced by reverse transcription and PCR amplification. (4) Sticky ends are produced by NotI digestion; samples are pooled and circularized by intermolecular ligation. (5) Circularized molecules are purified and fragmented by sonication. 5'-3' junctions are captured by biotin-streptoavidin purification. (6) Illumina libraries are produced and sequenced using 105-nucleotide paired-end reads.

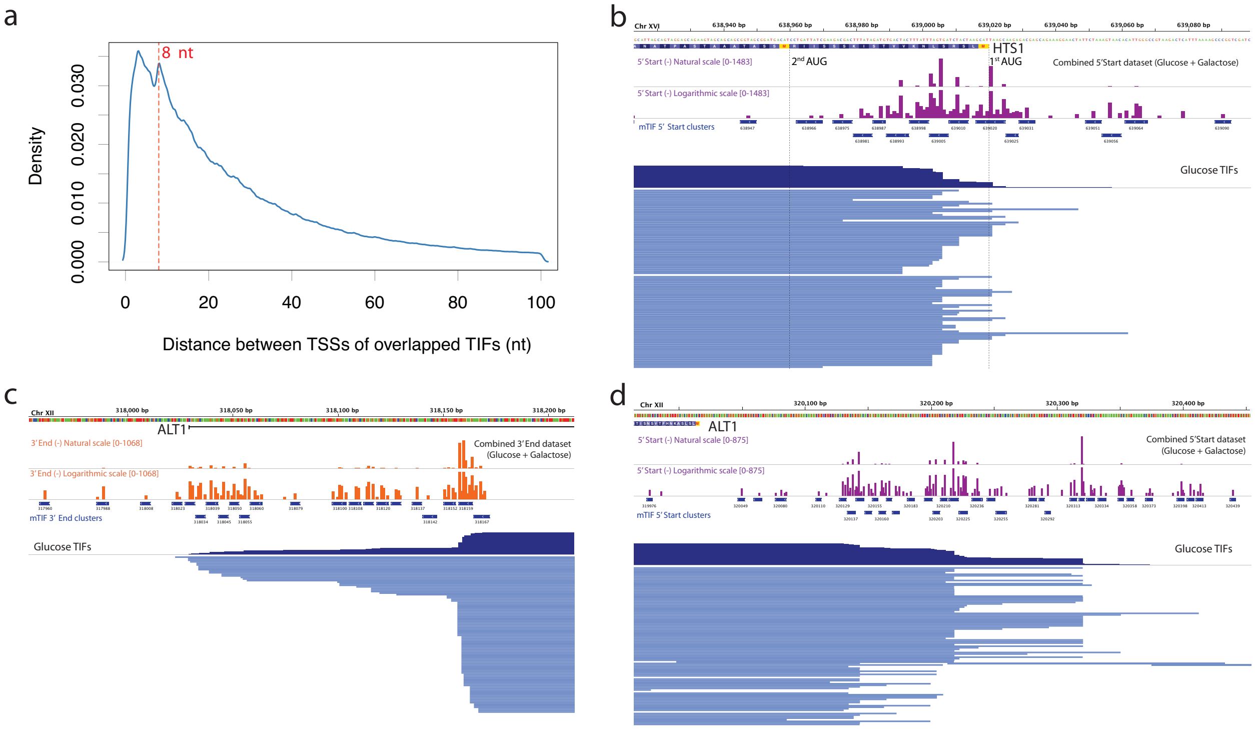


Figure S2. Clustering of transcript boundary positions for the definition of major transcript isoforms (mTIFs). a, Criteria for mTIF clustering. When comparing distances between TSSs from all overlapped TIFs, start sites separated by 8 nucleotides (red line) are overrepresented. A 5 nucleotide clustering distance for mTIFs was chosen to stay well above our technical precision while maintaining the resolution necessary to identify the 8 nucleotide-spaced start sites, which could have a biological basis. In the following panels, individual transcription start sites (purple) or polyadenylation sites (orange) are displayed in natural and logarithmic scale; sites co-occurring within 5 nucleotides were clustered into mTIFs (dark blue boxes, Supplementary Methods). Individual TIFs are displayed as blue lines, coverage as dark blue. **b**, Example of small boundary variations with large functional consequences: 5' end variations of only a few nucleotides in *HTS1* determine whether or not the 1st AUG is included and thus whether the transcript encodes the mitochondrial or cytoplasmic version of histidyl-tRNA synthetase respectively[30], as the former requires an N-terminal signal peptide (also shown in Fig. S20 with other examples of isoforms encoding truncated proteins). **c,** The same clustering was applied to 3' ends, illustrated here for *ALT1;* the logarithmic plot demonstrates the prevalence of isoforms with minor variations that would have been discarded with a larger window for clustering. **d,** In some cases transcription start sites can also span hundreds of nucleotides, shown here for *ALT1*. Protein coding sequences (in b and d) are represented as in Fig. 4a.

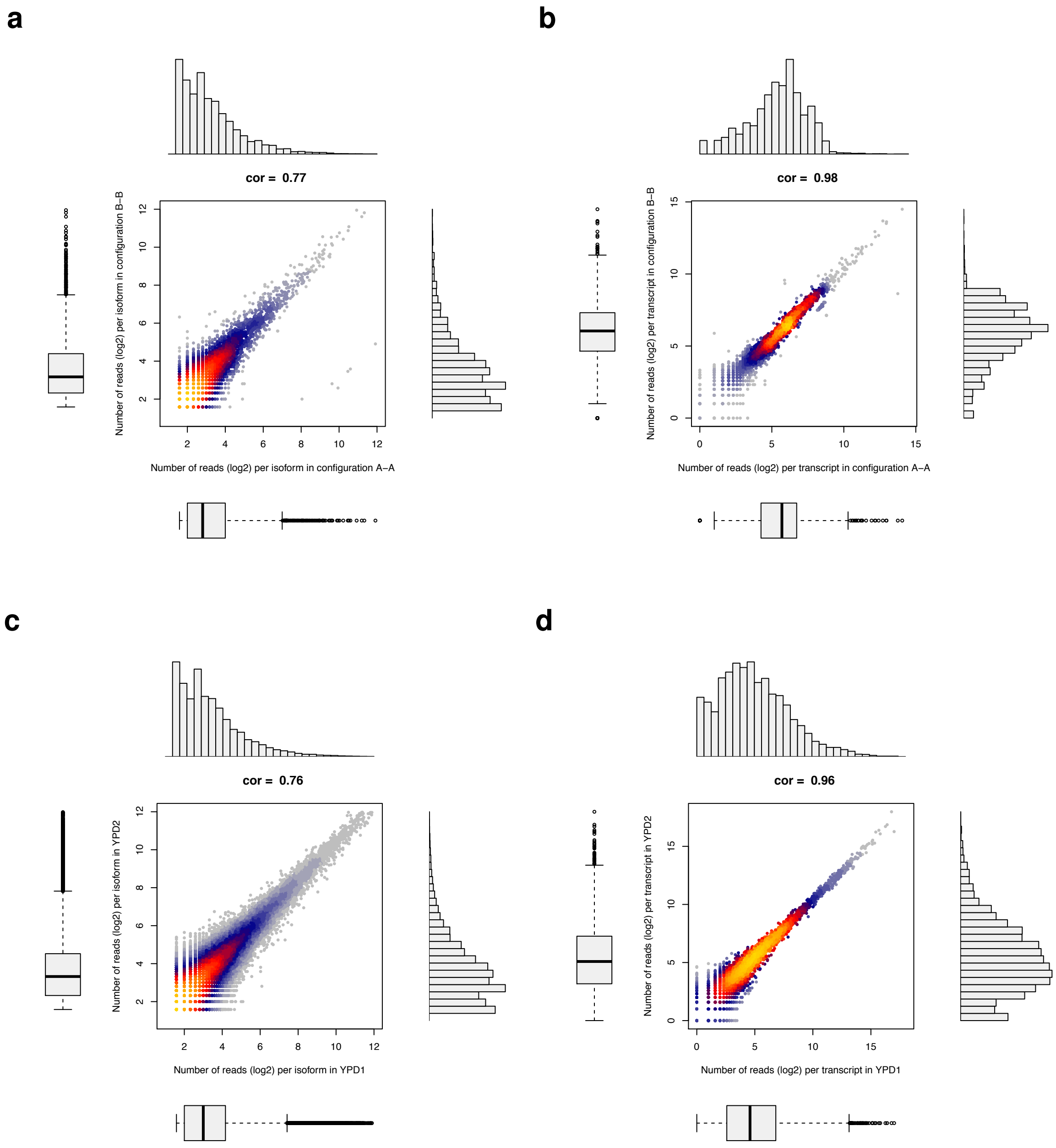


Figure S3. Reproducibility of the TIF-Seq method. Correlation between technical replicates of chimera control configurations (A-A) and (B-B) of one sample in YPD (YPD1, see Supplementary text); and biological replicates YPD1 and YPD2. Spearman correlation values (cor) are shown. **a,** Correlation between the numbers of exactly identical TIFs (identical start and end sites) in A-A and B-B; **b,** Correlation between the numbers of TIFs overlapping with annotated transcripts in A-A and B-B; **c,** Correlation between the numbers of exactly identical TIFs (identical start and end sites) in biological replicates YPD1 and YPD2; and **d,** Correlation between the numbers of TIFs overlapping with annotated transcripts[5] in YPD1 and YPD2.

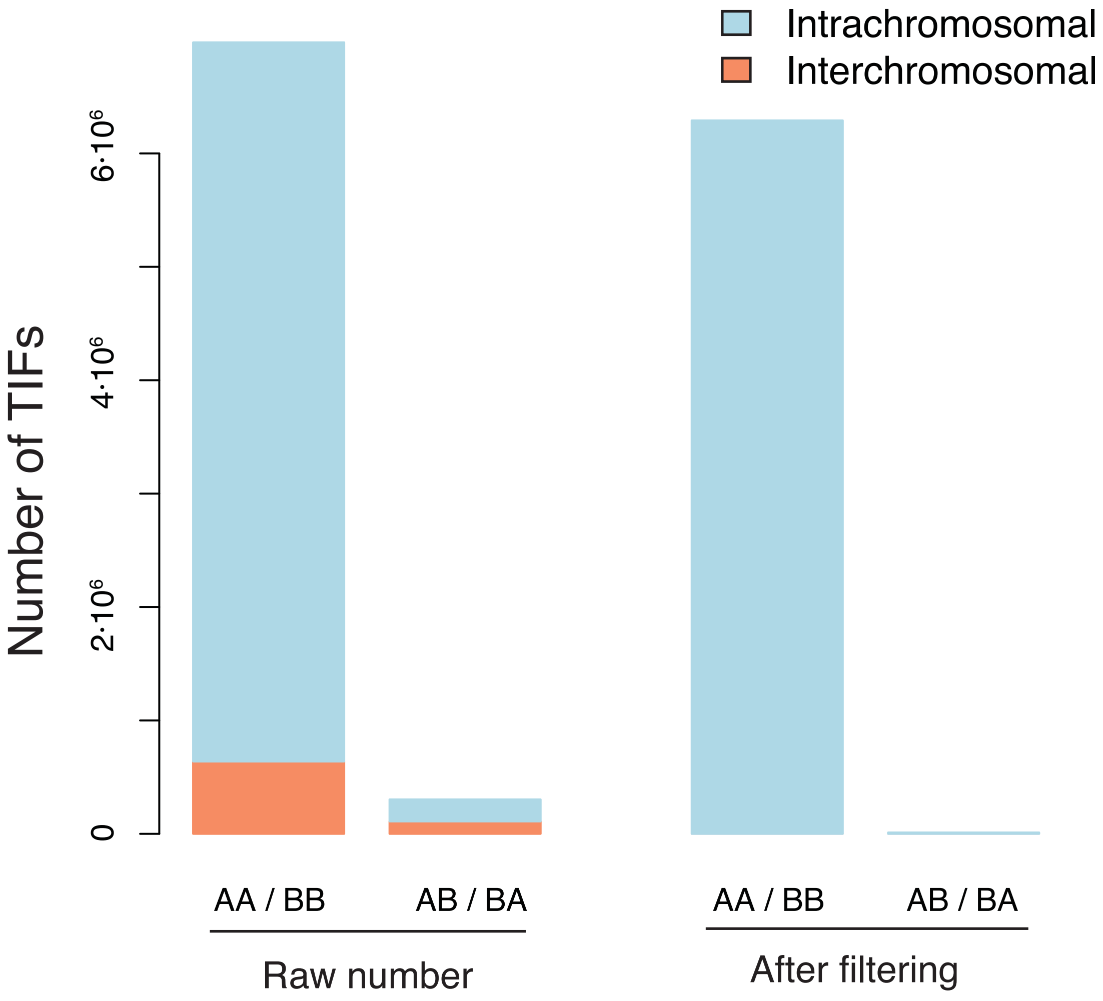


Figure S4. Control for chimeric intermolecular ligation events. Comparison of TIF sequences with putatively intramolecular configuration (A-A or B-B) to those with known intermolecular configuration (A-B and B-A) for YPD data. Number of TIFs mapping to the same (blue) or different (orange) chromosomes are shown. After bioinformatic filtering (right, see Supplementary Methods), a maximum of 0.19% intermolecular ligations are estimated to remain as false positive TIFs in our dataset.

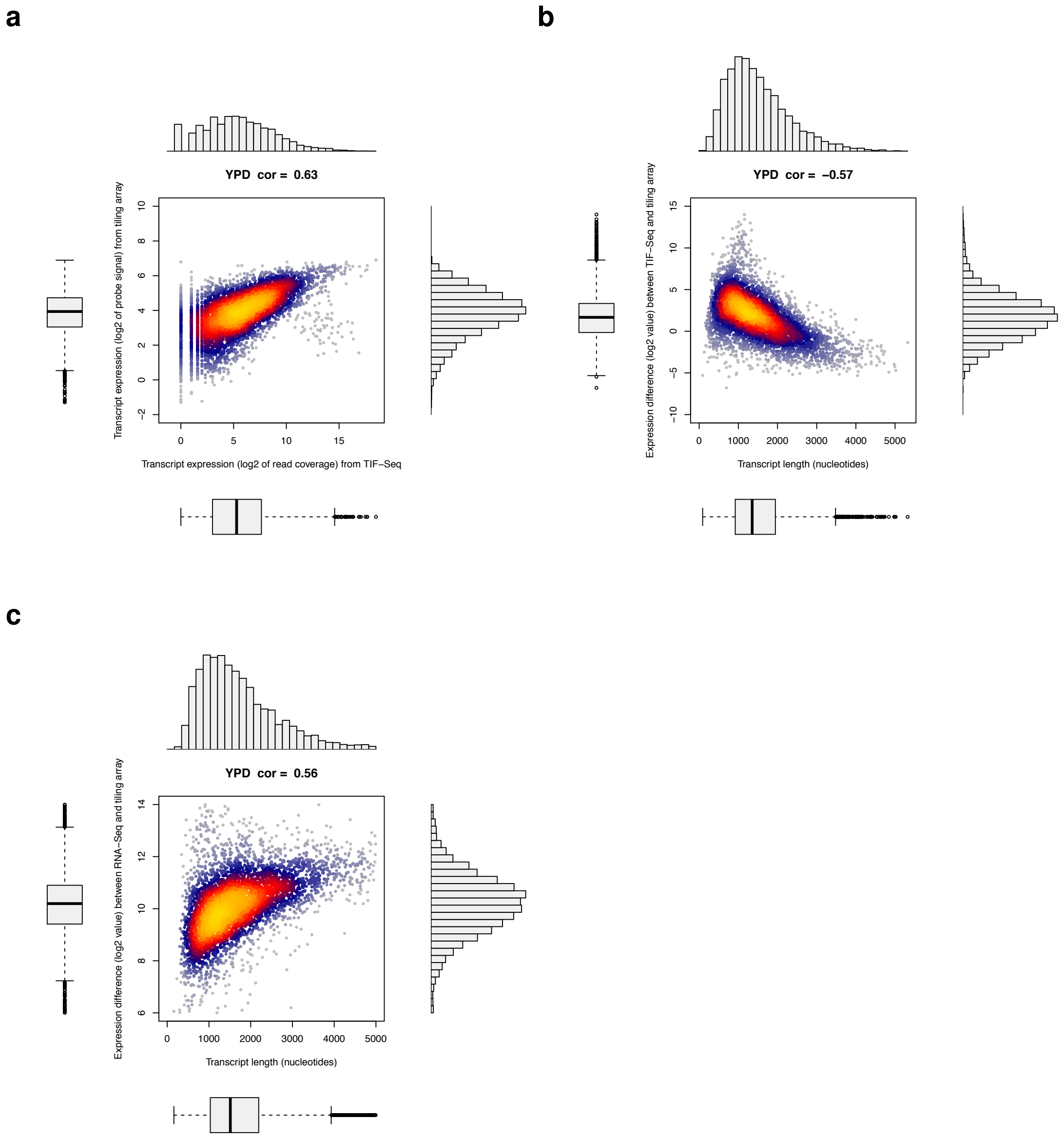


Figure S5. TIF-Seq detection limit is biased towards shorter transcripts. a, Raw TIF-Seq coverage by reads overlapping at least 80% of the ORF-T (transcripts covering annotated ORFs)[5] annotated region correlates well with mRNA abundance estimated by tiling arrays[5]. **b,** Most of the discrepancies between raw TIF-Seq and mRNA abundance from tiling arrays (shown here for YPD data) are due to the difficulty in producing full-length cDNA from long mRNA molecules. This can be overcome by enriching for longer RNA molecules (Fig. S6). **c,** Discrepancies between raw RNA-Seq[26] and tiling array measurements are also due to a length-dependent bias, and a certain transcript length must be assumed to estimate the mRNA abundance from RNA-Seq data. The number of reads in RNA-Seq is dependent on both mRNA abundance and transcript length. Contrary to TIF-Seq, however, short transcripts produce less reads and are more difficult to detect.

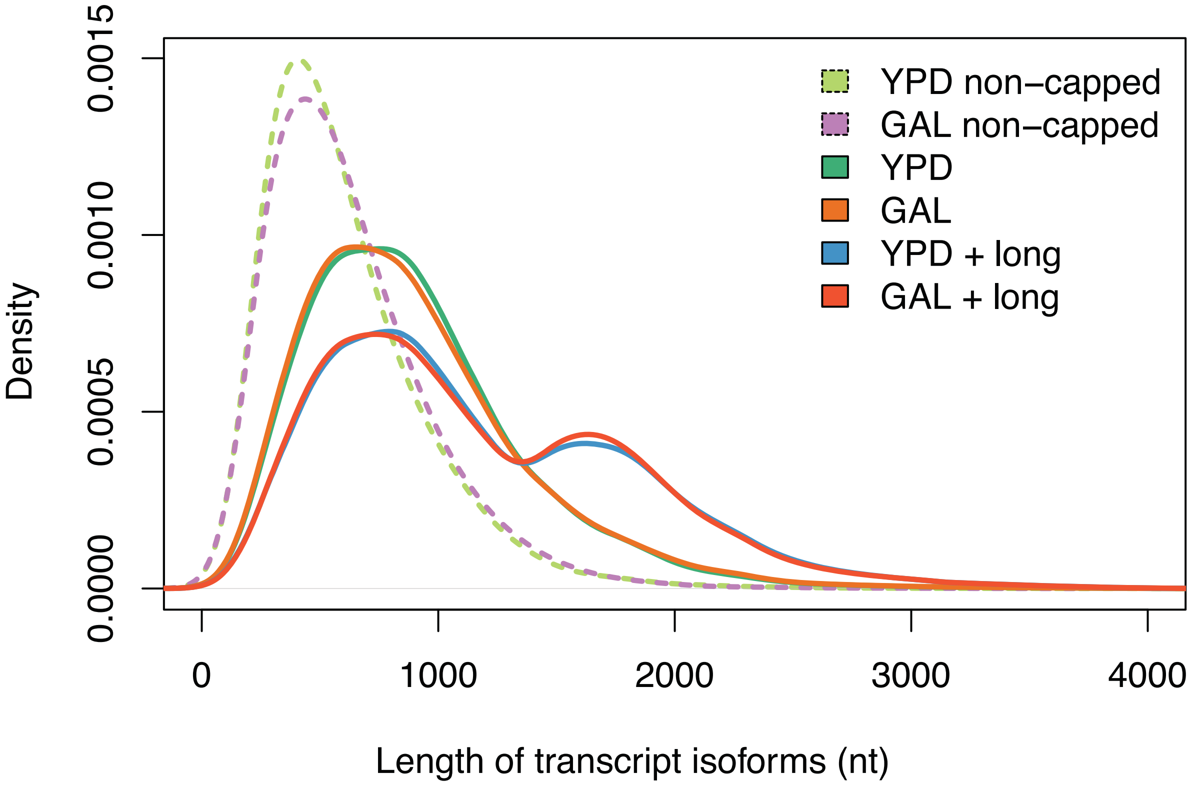


Figure S6. Size distribution of mapped TIF-Seq reads. Size distribution of RNA molecules identified in glucose (YPD) and galactose (GAL) with the standard TIF-Seq protocol. When enrichment for long molecules (>2kb) by gel size selection is performed (YPD + long and GAL + long), signals for longer RNAs are recovered. As expected since they are likely degradation products, samples of non-capped molecules (dashed lines) are enriched for short RNA molecules.

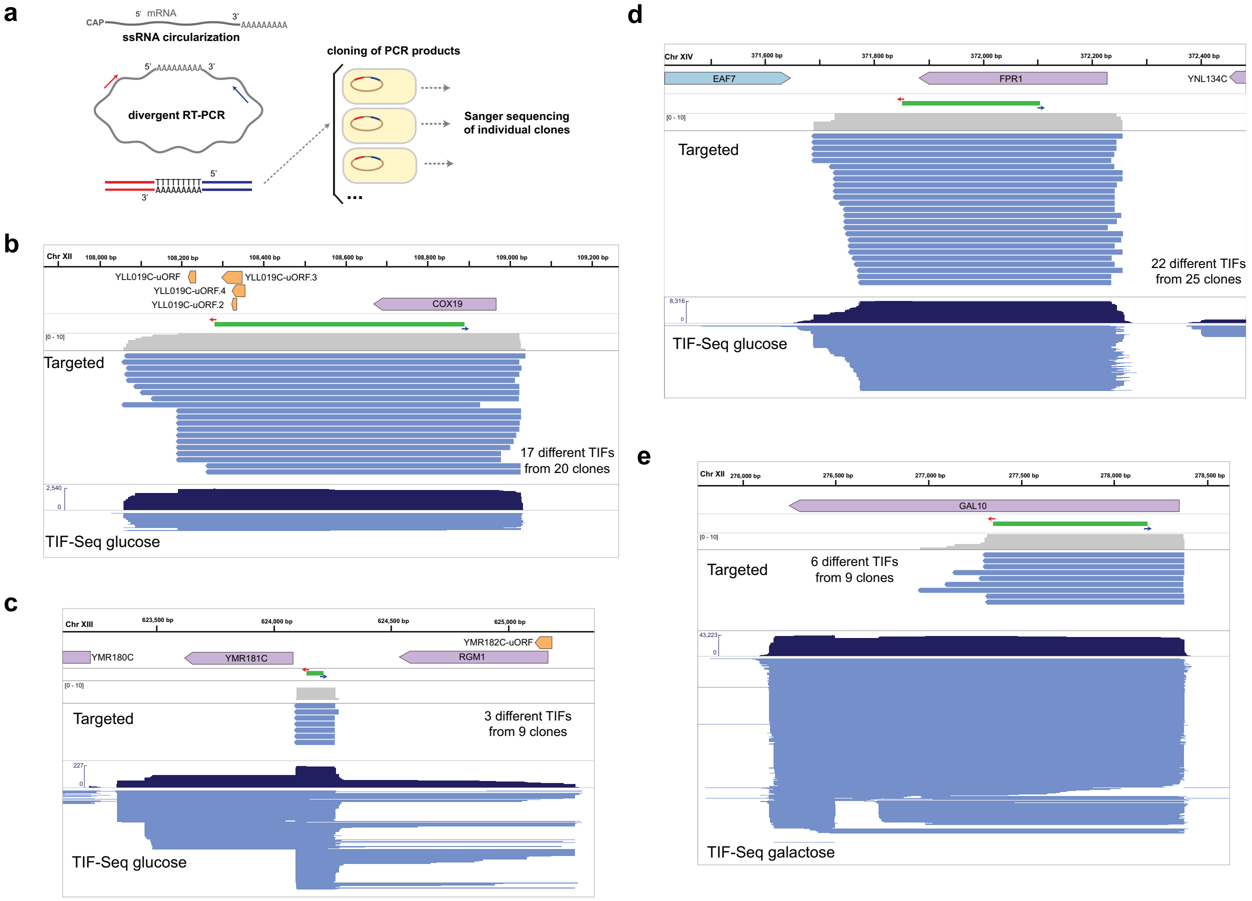


Figure S7. Confirmation by targeted sequencing. a, Experimental design for independent confirmation of TIF variability at single-nucleotide resolution. ssRNA was circularized and subjected to reverse transcription. Divergent PCR was performed for selected genes, and the products were cloned into bacteria. Individual clones were sequenced to determine the mRNA start and end sites. **b,** Confirmations of misannotated "uORFs" (in orange) located in the 3' UTR of *COX19*, which we reclassified as downstream ORF (dORFs) due to their inclusion in transcripts containing an upstream ORF. Individual TIF clones mapped by Sanger sequencing (Targeted) and collapsed reads for TIF-Seq are displayed (TIF-Seq glucose/galactose). The distance between the divergent oligos used for PCR is depicted in green. **c,** Confirmation of ncRNAs produced in the overlap region between *RGM1* and *YMR181C*. **d-e,** Confirmation of diverse isoforms expressed from *FPR1* and *GAL10*. All experiments were performed in glucose except *GAL10,* which was performed in galactose.

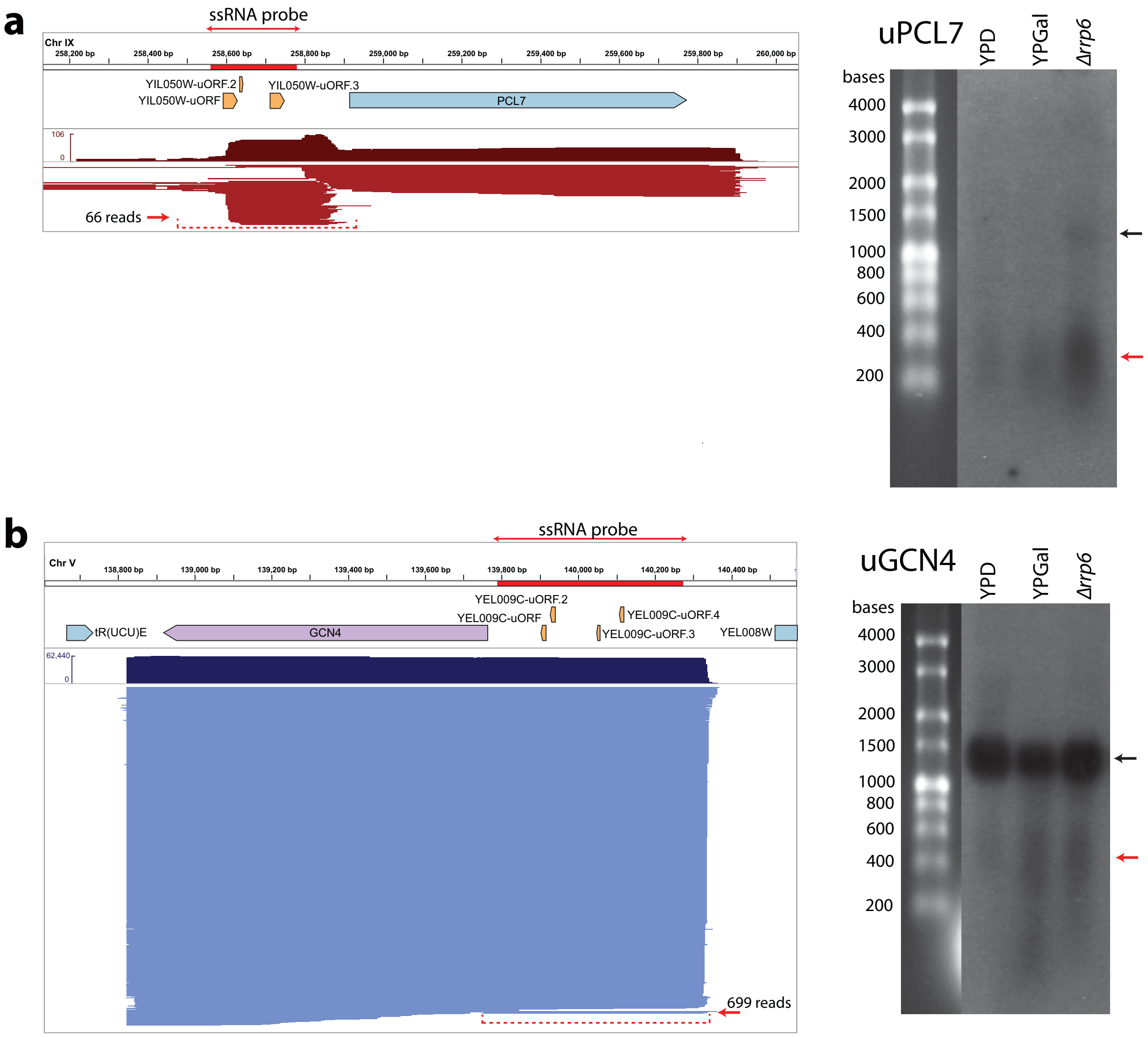


Figure S8. Validation of short coding RNAs and independent uORFs by Northern blot. a, Example of an independent scRNA upstream of *PCL7*. **b,** Alternative independently expressed uORF for *GCN4*. Annotation and TIF-Seq data in glucose are displayed. The UV image of an RNA-ladder stained with SYBR Green II and a northern blot labeled with a strand-specific DIG-RNA probe (depicted in red above the gene annotation) is shown on the right. Bands corresponding to the independent feature (red) and the canonical mRNA (black) are labeled.

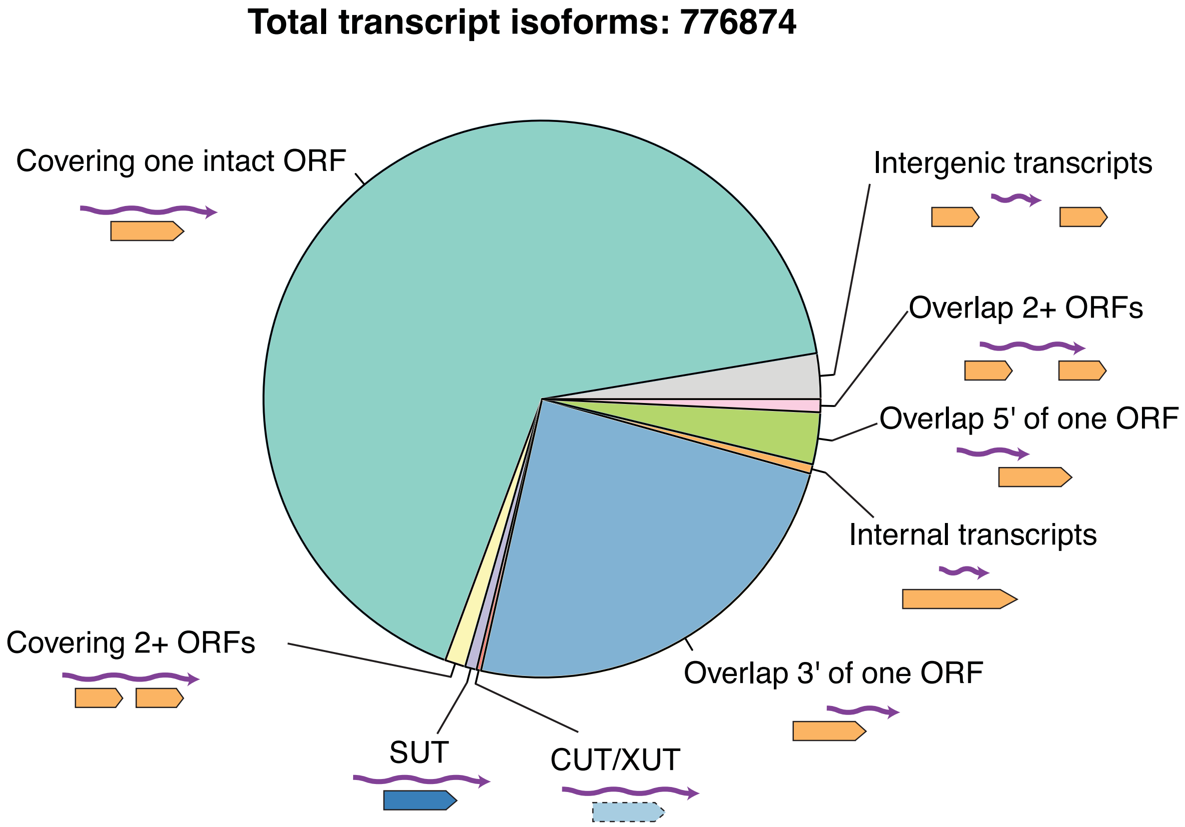


Figure S9. Number of transcript isoforms classified according to their overlap of annotated genomic features. Only TIFs supported by at least 2 sequencing reads are included. SUT, stable annotated transcript[5]; CUT, cryptic unstable transcript[5]; XUT, *xrn1Δ*-sensitive unstable transcript[31].

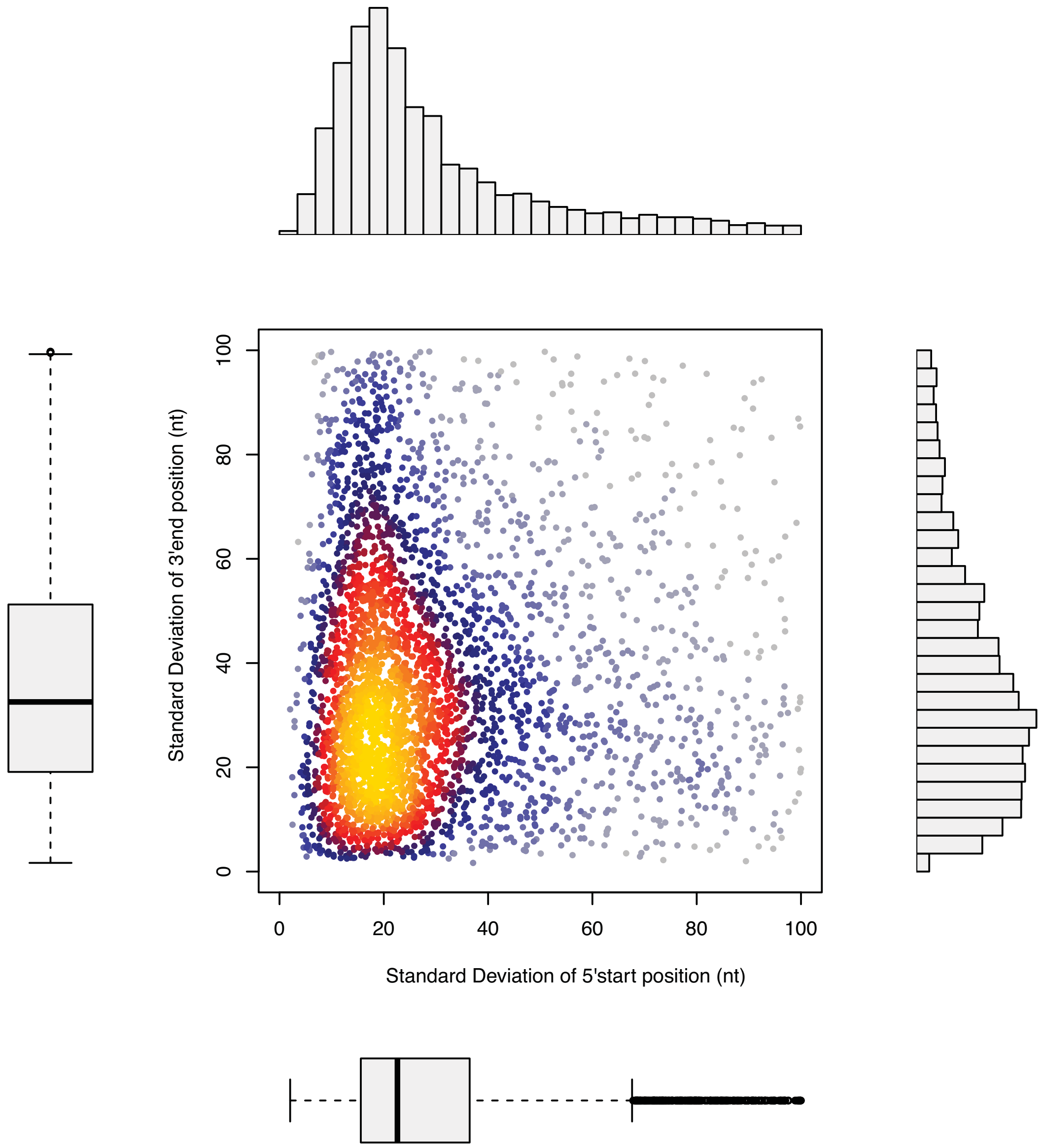


Figure S10. Variation in transcript start and end sites for all ORFs in the yeast genome. Standard deviation of the TSS and TTS positions for all TIFs per ORF. Each point represents one ORF. Barplots and boxplots reflect frequencies of given standard deviations; median values are represented by the central lines in the boxplots. Only TIFs that completely cover non-dubious ORF coding regions are included.

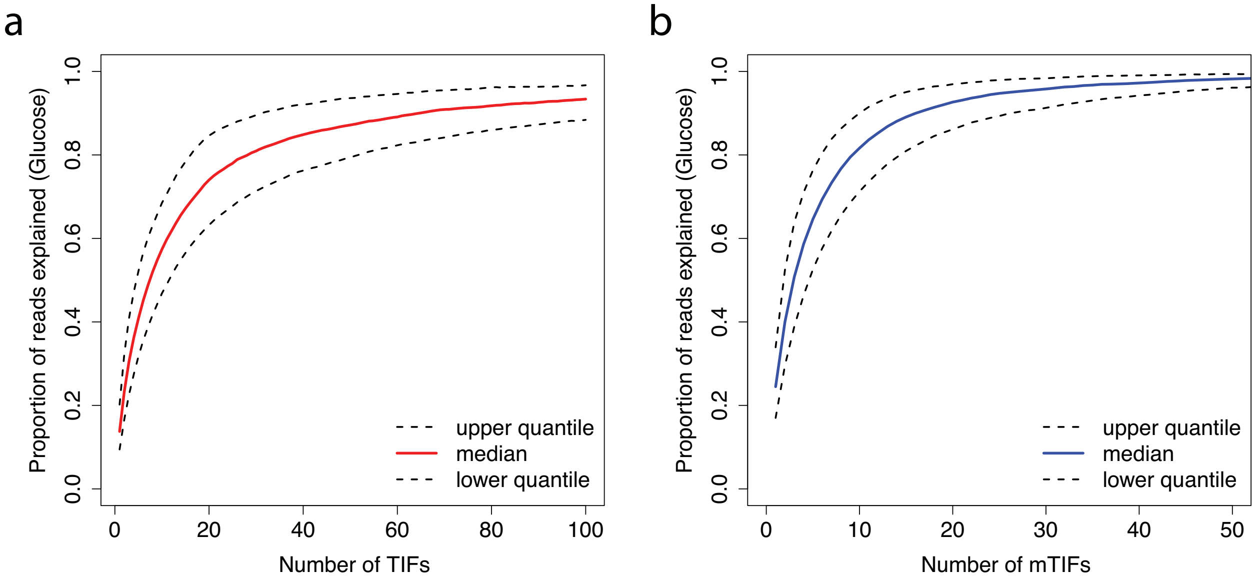


Figure S11. Estimation of the number of distinct isoforms required to explain TIF-Seq coverage for each gene. Number of distinct TIFs (**a**) and major TIFs (mTIFs, **b**) necessary to explain the population of transcripts for each gene. For each point, only genes with at least the stated number of isoforms were taken into account. To explain 80% of the coverage it is necessary to consider at least 29 TIFs or 10 mTIFs. To avoid noise due to low sequencing coverage, the analysis was restricted to the 3318 coding genes with 50 or more TIF-Seq reads.

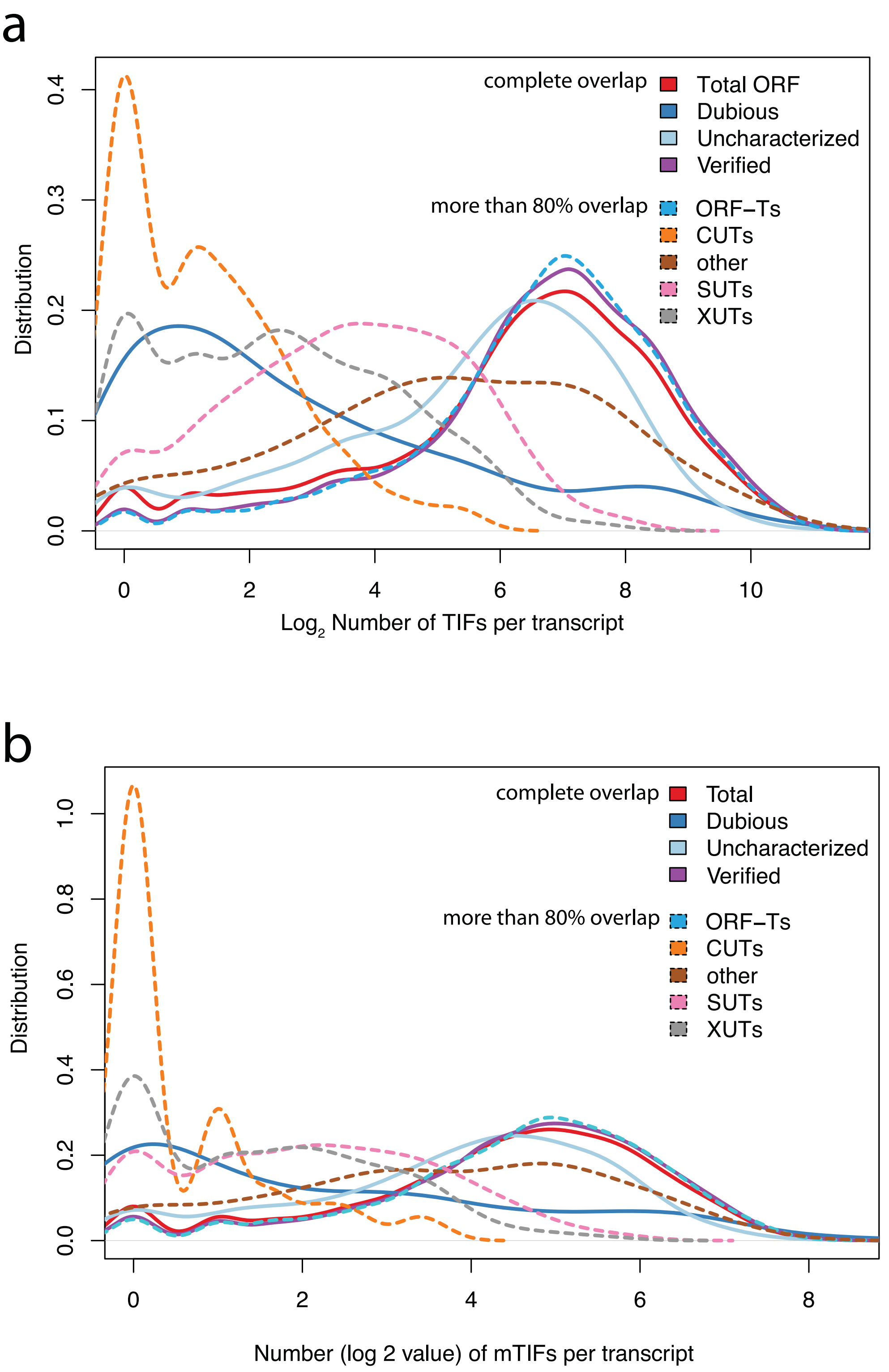


Figure S12. Number of TIFs and mTIFs covering genomic features. TIFs (**a**) or major TIFs (**b**) completely covering ORF coding regions (solid lines); TIFs overlapping more than 80% with previously reported transcript annotations for coding and non-coding RNAs (dashed lines). RNA heterogeneity exists in all types of coding and non-coding RNAs.

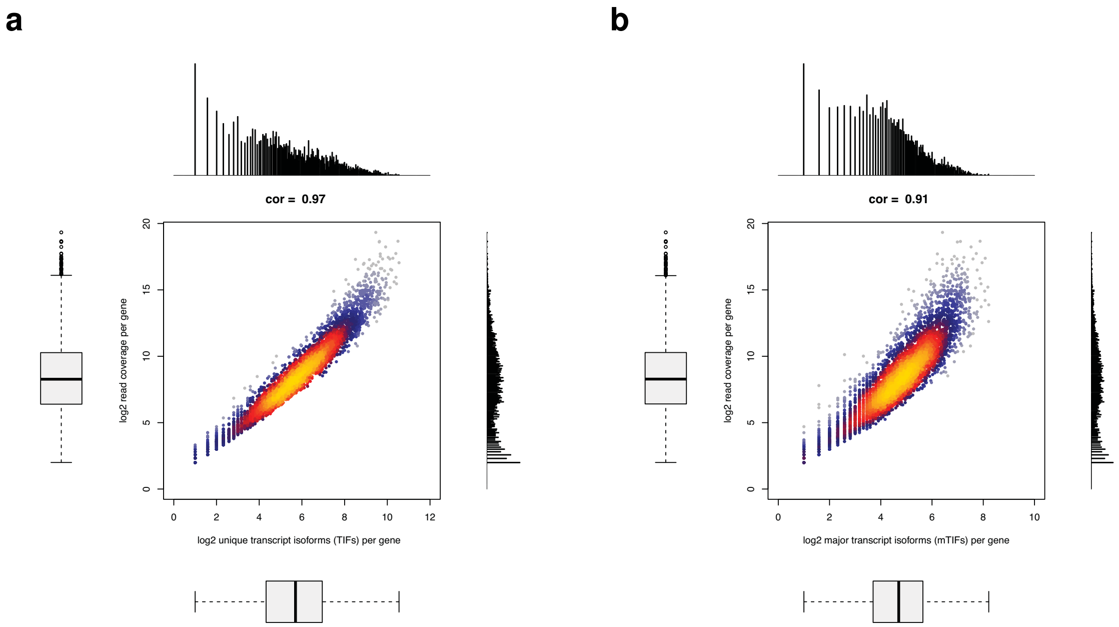


Figure S13. Estimation of the upper limit of isoforms per gene. **a,** The number of reads covering each ORF (as a measure of sequencing coverage) is represented relative to the number of unique transcript isoforms (different structure) per ORF. For genes with high sequencing coverage, a maximum of ~500 isoforms per gene ($\log_2$(TIFs) ≈ 9) is estimated. **b,** same representation as a, but using the clustered mTIFs. A maximum of ~100 major TIFs ($\log_2$(mTIFs) ≈ 7) is estimated per gene.

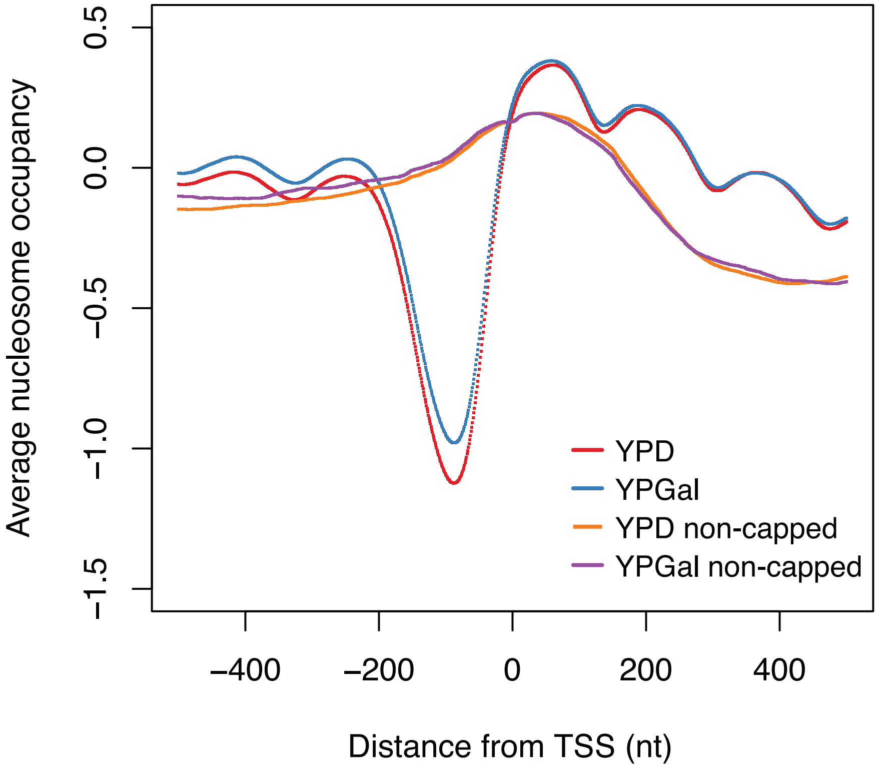


Figure S14. Nucleosome occupancy relative to the transcript start (TSS) and end (TTS) sites. Nucleosome-depleted regions near the TSS were observed for capped molecules but not for non-capped. This suggests non-capped molecules arise from posttranscriptional events. Nucleosome data are from Kaplan et al.[28].

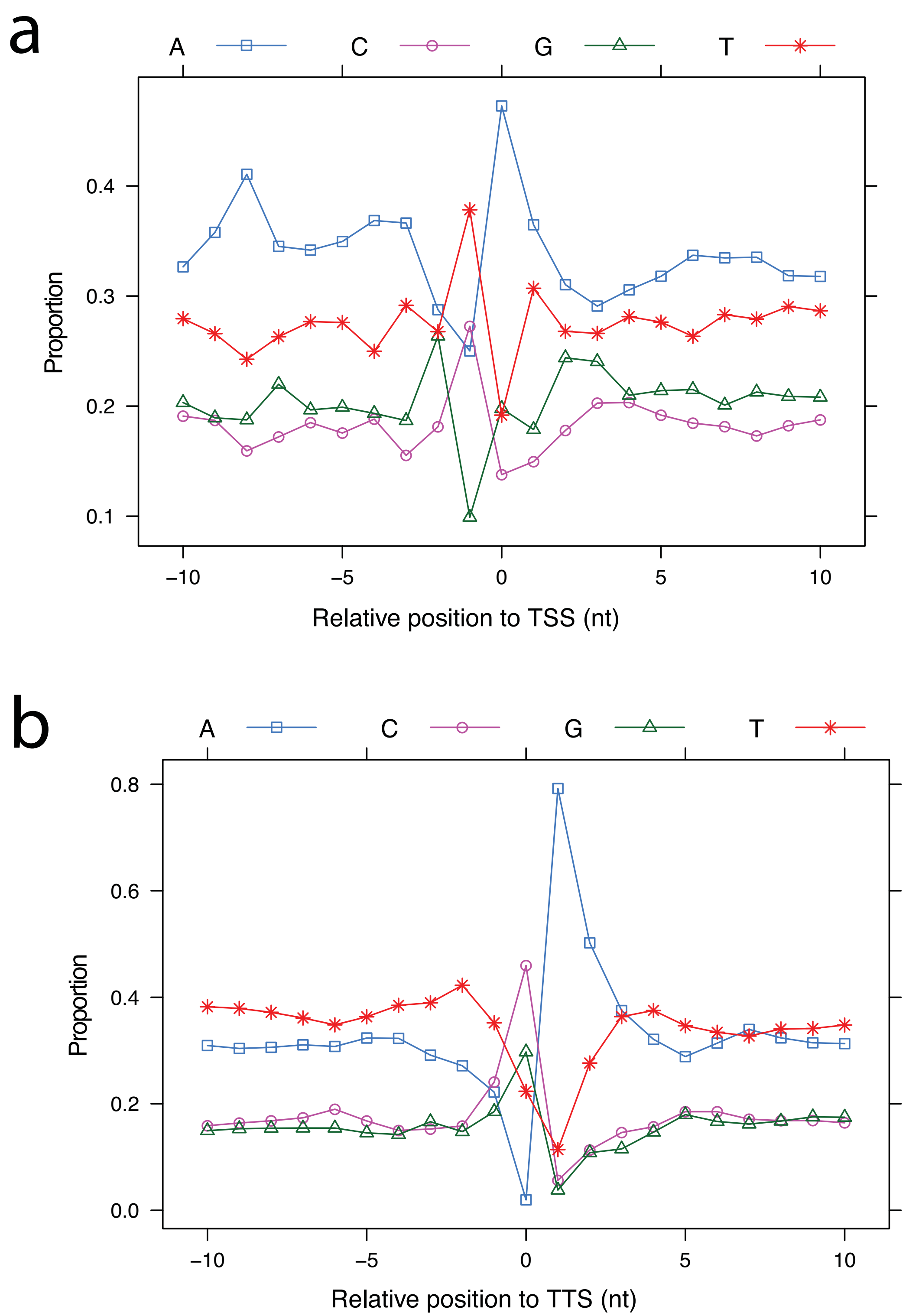
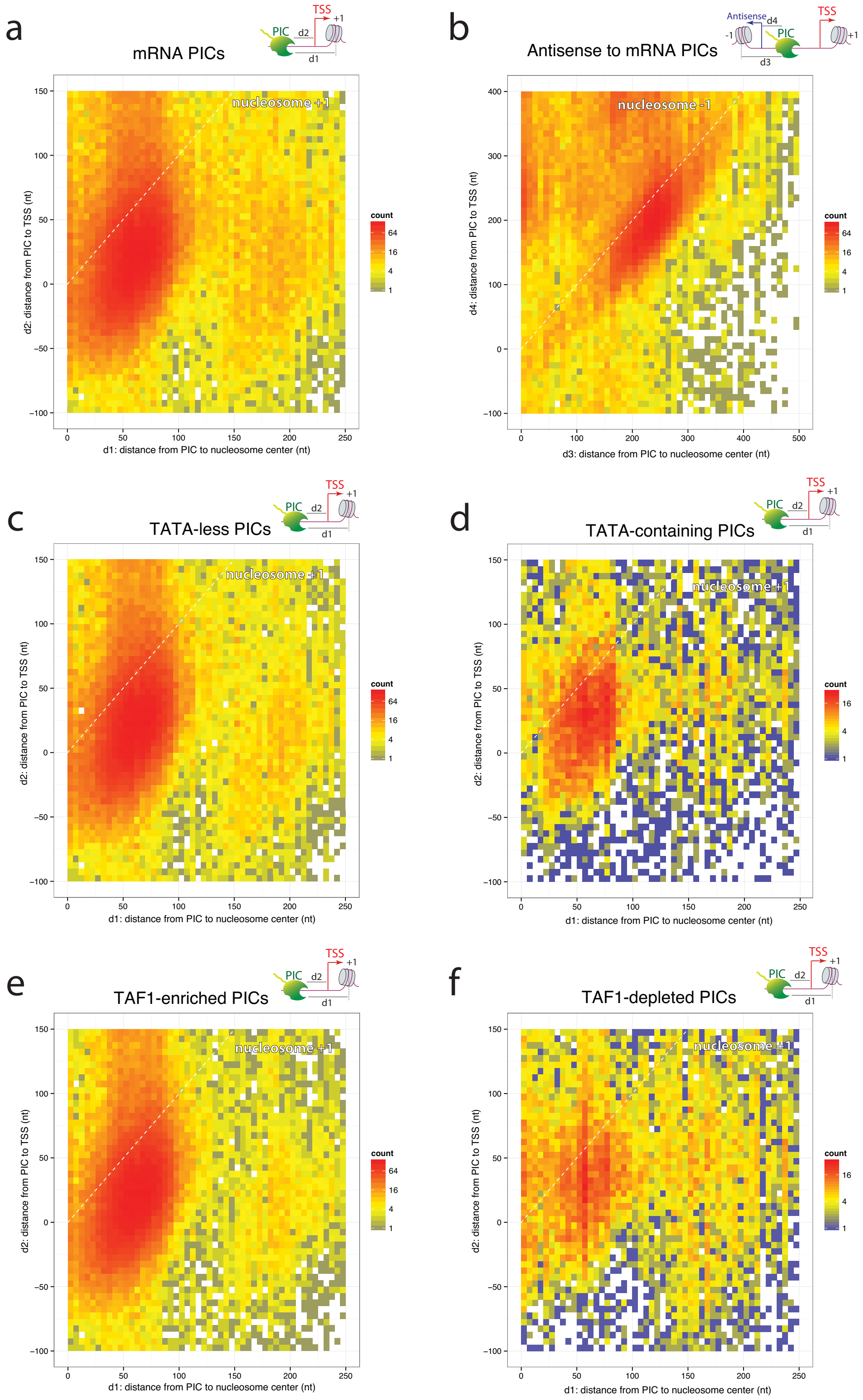


Figure S15. Sequence motifs of transcript start and termination sites. **a,** Sequence motif of TSSs. **b,** Sequence motif of TTSs. Position 0 corresponds to either the first nucleotide of the 5' end of the transcript (a) or the last nucleotide before the poly(A) site (b). Negative and positive values refer respectively to upstream and downstream orientation.



a
mRNA PICs
PIC
d2
TSS
+1
d1
nucleosome +1
d2: distance from PIC to TSS (nt)
d1: distance from PIC to nucleosome center (nt)
150
100
50
0
−50
−100
0
50
100
150
200
250
count
64
16
4
1
b
Antisense to mRNA PICs
Antisense
d4
-1
PIC
TSS
+1
d3
nucleosome -1
d4: distance from PIC to TSS (nt)
d3: distance from PIC to nucleosome center (nt)
400
300
200
100
0
−100
0
100
200
300
400
500
count
64
16
4
1
c
TATA-less PICs
PIC
d2
TSS
+1
d1
nucleosome +1
d2: distance from PIC to TSS (nt)
d1: distance from PIC to nucleosome center (nt)
count
64
16
4
1
d
TATA-containing PICs
PIC
d2
TSS
+1
d1
nucleosome +1
d2: distance from PIC to TSS (nt)
d1: distance from PIC to nucleosome center (nt)
count
16
4
1
e
TAF1-enriched PICs
PIC
d2
TSS
+1
d1
nucleosome +1
d2: distance from PIC to TSS (nt)
d1: distance from PIC to nucleosome center (nt)
count
64
16
4
1
f
TAF1-depleted PICs
PIC
d2
TSS
+1
d1
nucleosome +1
d2: distance from PIC to TSS (nt)
d1: distance from PIC to nucleosome center (nt)
count
16
4
1

Figure S16 .Transcript start sites are mainly defined by nucleosome positioning. Heatmap mTIF TSS positions, where the distance from annotated PICs[11] to the first downstream nucleosome[27] is represented on the x-axis (d1), and the distance from the PICs to the mTIF TSS measured in this study is represented on the y-axis (d2). The TSS tends to appear downstream of the PIC, just before the first downstream nucleosome (nucleosome +1 or -1 represented as a dashed white line). However, a significant number of TSSs also appear before the annotated PIC, suggesting the existence of non-annotated upstream secondary PICs. **a,** Representation of all the PICs associated with mRNA transcripts. A secondary accumulation of TSSs can be observed to the right, where the +1 nucleosome was likely unstable and not properly detected. **b,** Representation of antisense transcripts produced from the same nucleosome-depleted region. In this case, a dashed line represents the -1 nucleosome and distances are measured relative to the -1 nucleosome (d3) and the antisense TSS (d4). **c-f,** same representation used in (a**)**, but for genes annotated as TATA-less, TATA-containing, TAF1-enriched and TAF1-depleted[11]. TATA-containing genes and TAF1-depleted genes form a second population of TSSs downstream of the PIC that does not appear to depend as strongly on the presence of the +1 nucleosome (can be seen on the right side and on Supplementary Fig. S17a-b). An alternative explanation is that the +1 nucleosome, which is known to be unstable in these cases, has not been properly localized. However, this nucleosome-independent (or dependent on unstable nucleosomes) TSS selection described for TATA-containing genes[11] represents only a minor population of all the transcript start events. Only TSSs supported by major transcript isoforms (mTIFs) are included.

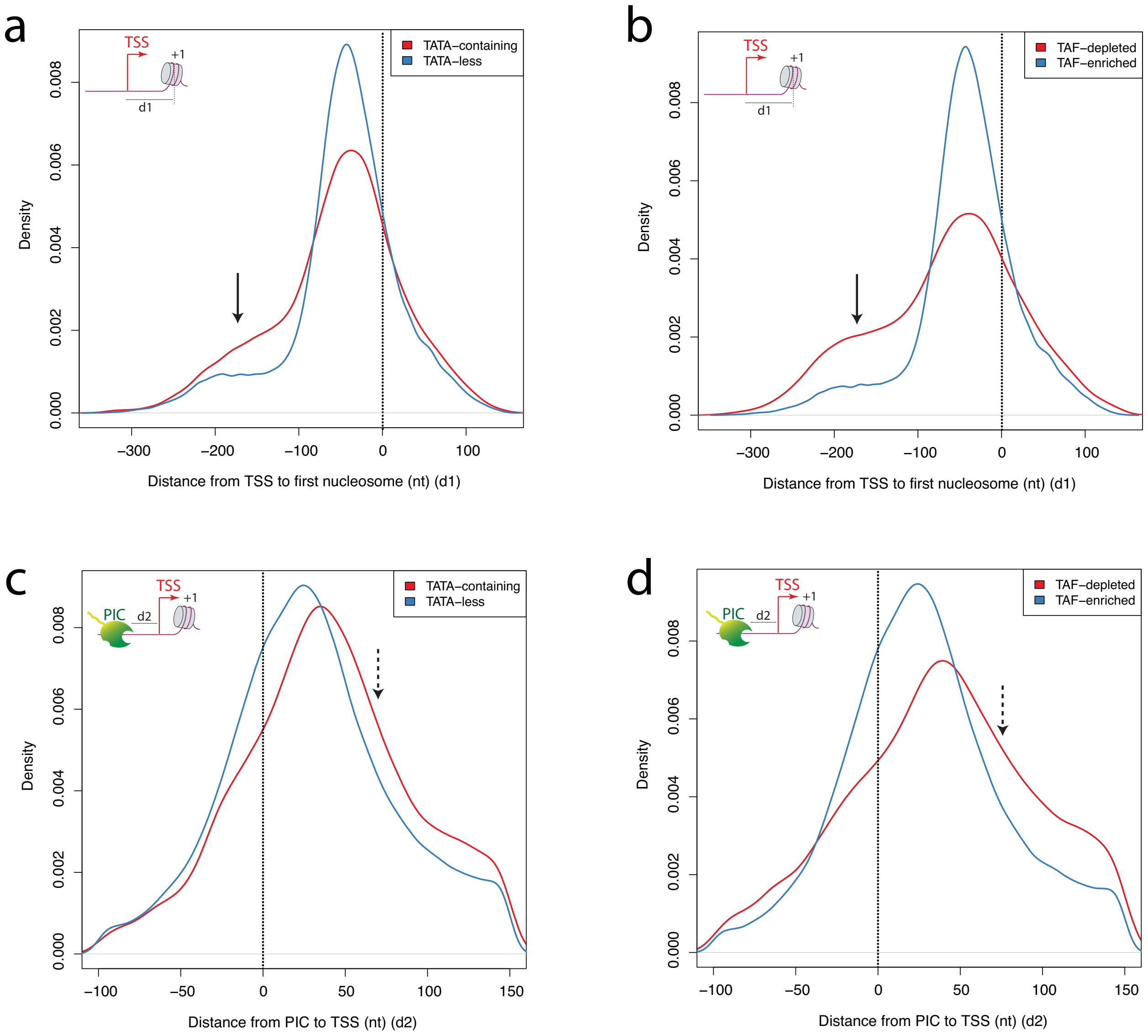


Figure S17. Distribution of transcript start site with respect to the +1 nucleosome and preinitiation complex. Frequencies of start sites associated with TATA-less, TATA-containing, TAF1-enriched and TAF1-depleted genes[11] are displayed with respect to the +1 nucleosome (**a, b**) or the preinitiation complex (PIC)[11] (**c, d**). TATA-containing and TAF1-depleted genes present a second population of TSSs (solid arrow) further upstream of the nucleosome +1 (d1) and longer polymerase scanning distance (d2, dashed arrow).

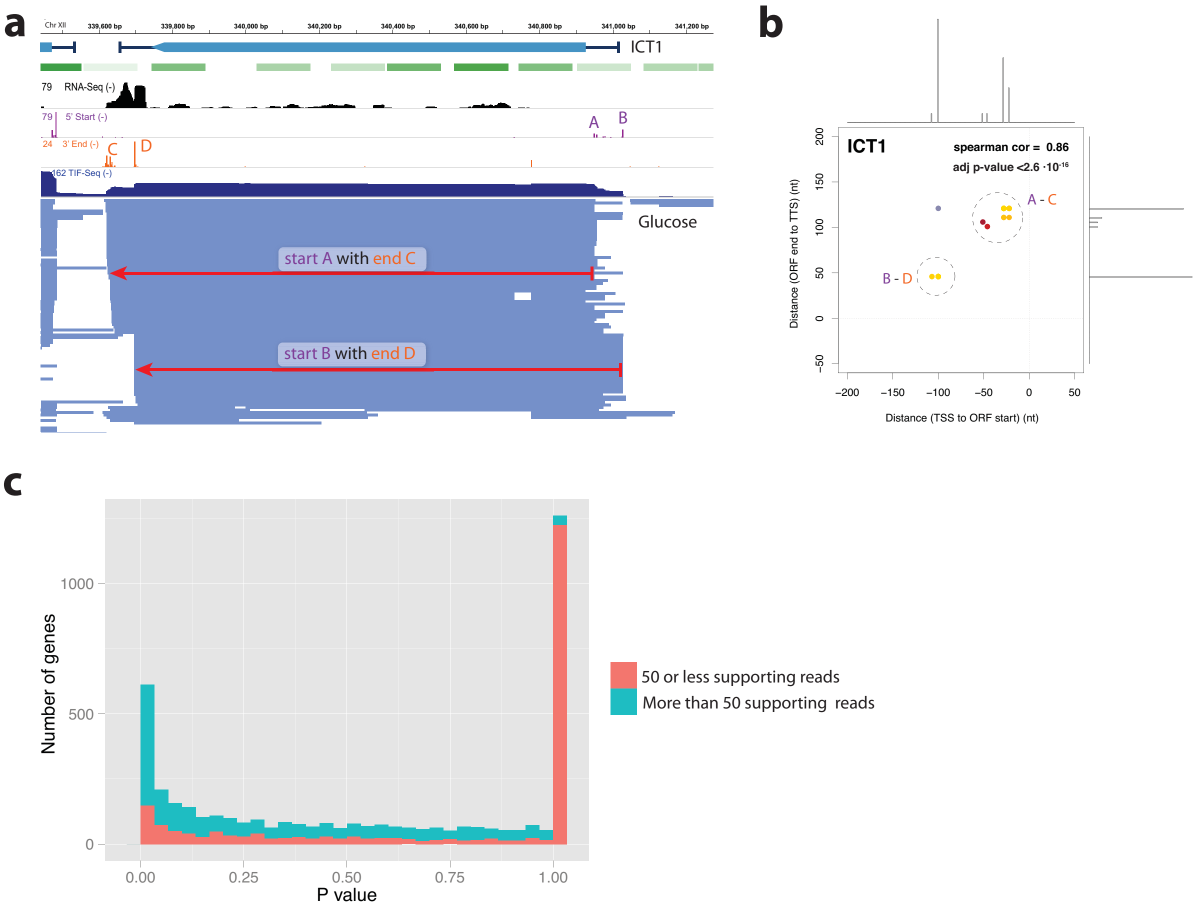


Figure S18. Example of interdependence between transcription start and end sites. a-b, The transcription start and polyadenylation sites of *ICT1* are correlated. The main TIF start-end combinations are depicted as A-C and B-D. Additional data and annotations as in Fig. 2d-e. **c,** The distribution of *P* values from correlations between start and end sites for genes with respect to their sequencing coverage. Only genes with more than 50 supporting reads (blue) were considered for statistical analysis of start-end correlations.

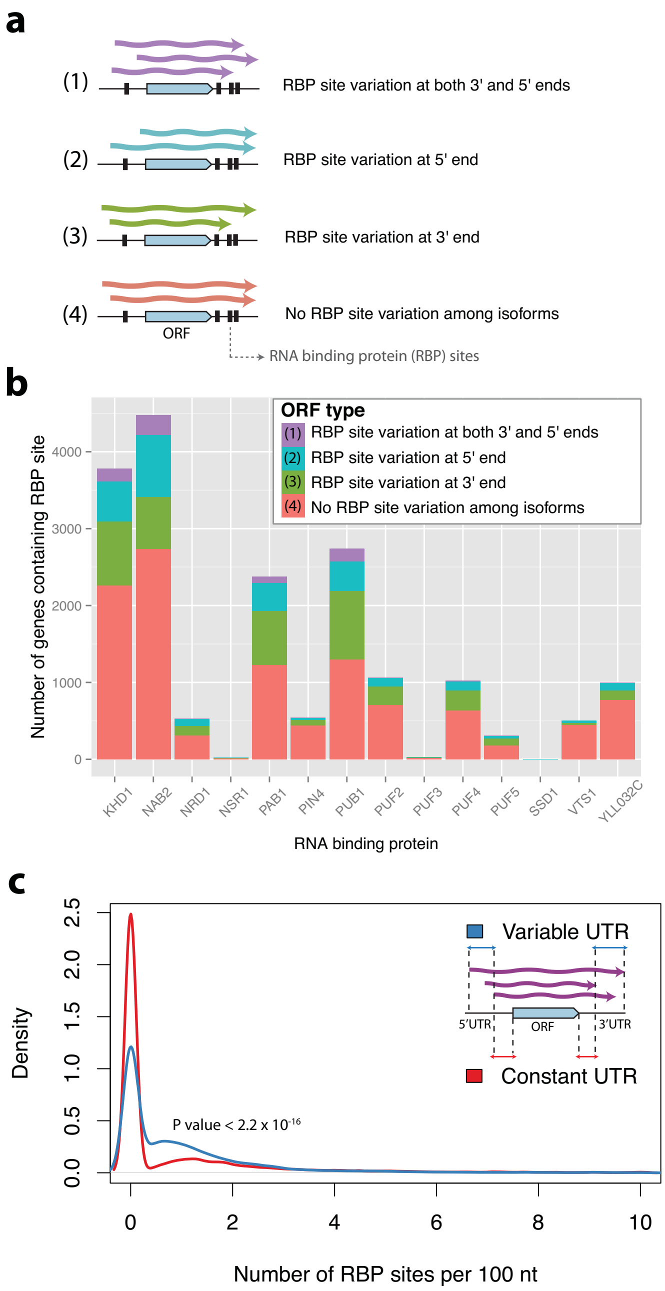


Figure S19. Number of genes with RNA binding protein (RBP) site variations due to mTIF heterogeneity. a, Illustration of the possible effects of mTIF heterogeneity on the alternative inclusion of RBP sites. Coding genes can express populations of mTIFs that vary in their inclusion of RBP sites at both 5’ and 3’ UTRs (1), only at 5’UTR (2) or 3’ UTR (3), or without alternative presence of RBP sites due to the isoform heterogeneity (4). **b,** Influence of alternative TIFs on the inclusion or exclusion of RBP sites determined by Riordan *et al.*[13], categorized as in a), for all isoforms detected in wildtype cells grown in glucose and galactose. **c**, Distributions of number of RBP binding sites per 100 nt in constant (regions that are present in every TIF) *vs.* variable UTR regions; the statistically significant difference between these distributions demonstrates that variable UTR regions are enriched for RBP sites.

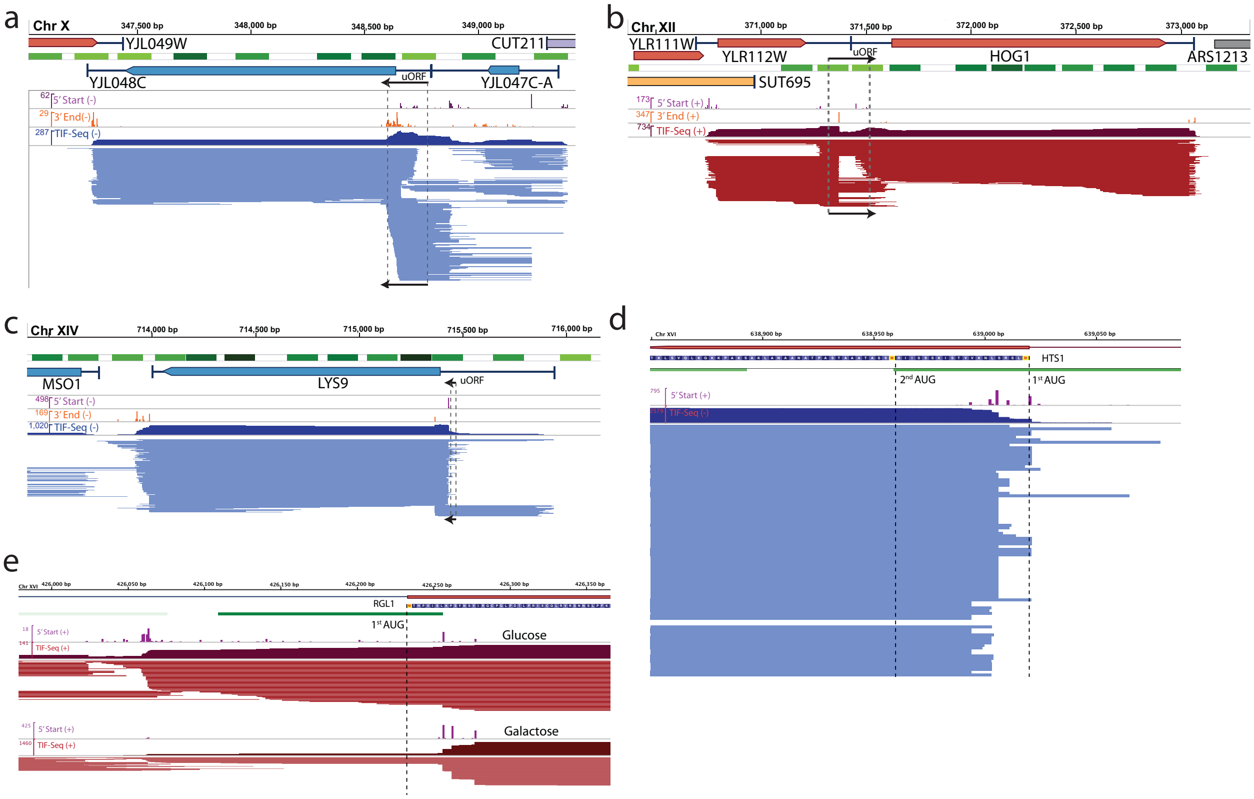


Figure S20. Examples of short coding RNAs and alternative isoforms that encode truncated proteins. a-c, Examples of scRNAs previously annotated as "uORF"s covered by independent TIFs. **d,** Small 5' end variations in *HTS1* determine whether or not the 1st AUG is included and thus whether the transcript encodes the mitochondrial or cytoplasmic version of histidyl-tRNA synthetase respectively[30], as the former requires an N-terminal signal peptide. **e,** The gene *RGL1* produces an N-terminal truncated protein isoform preferentially in galactose but not in glucose. All data are from glucose profiles, except (e) for which both glucose and galactose profiles are shown.

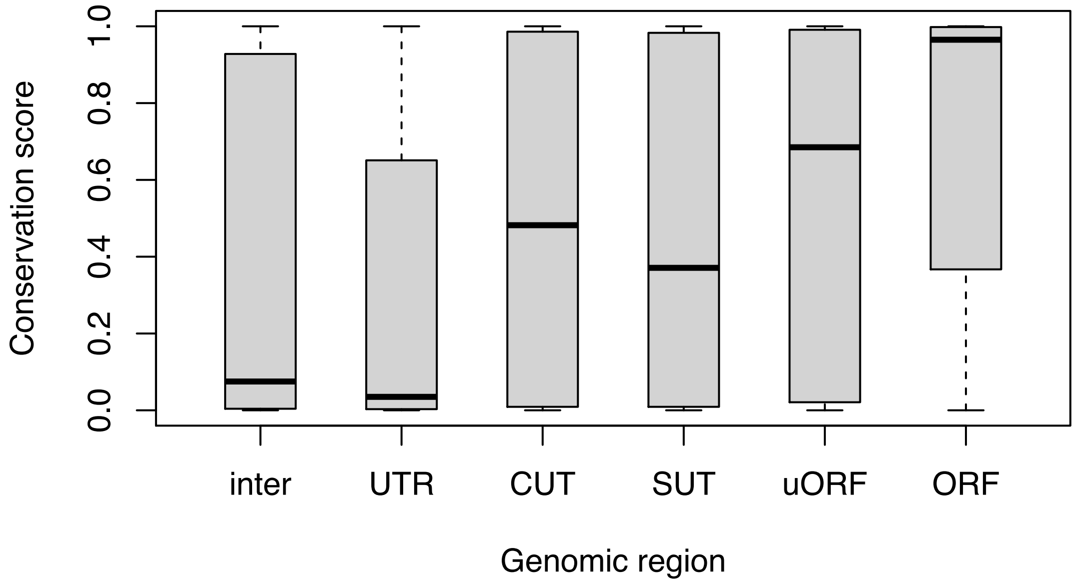


Figure S21. Conservation scores for annotated genomic regions. Distribution of conservation scores across annotated genomic features (‘inter’ refers to intergenic/unannotated regions). Conservation phastCons scores for multiple alignments of 6 yeast genomes to the *S. cerevisiae* genome were obtained from UCSC[32].

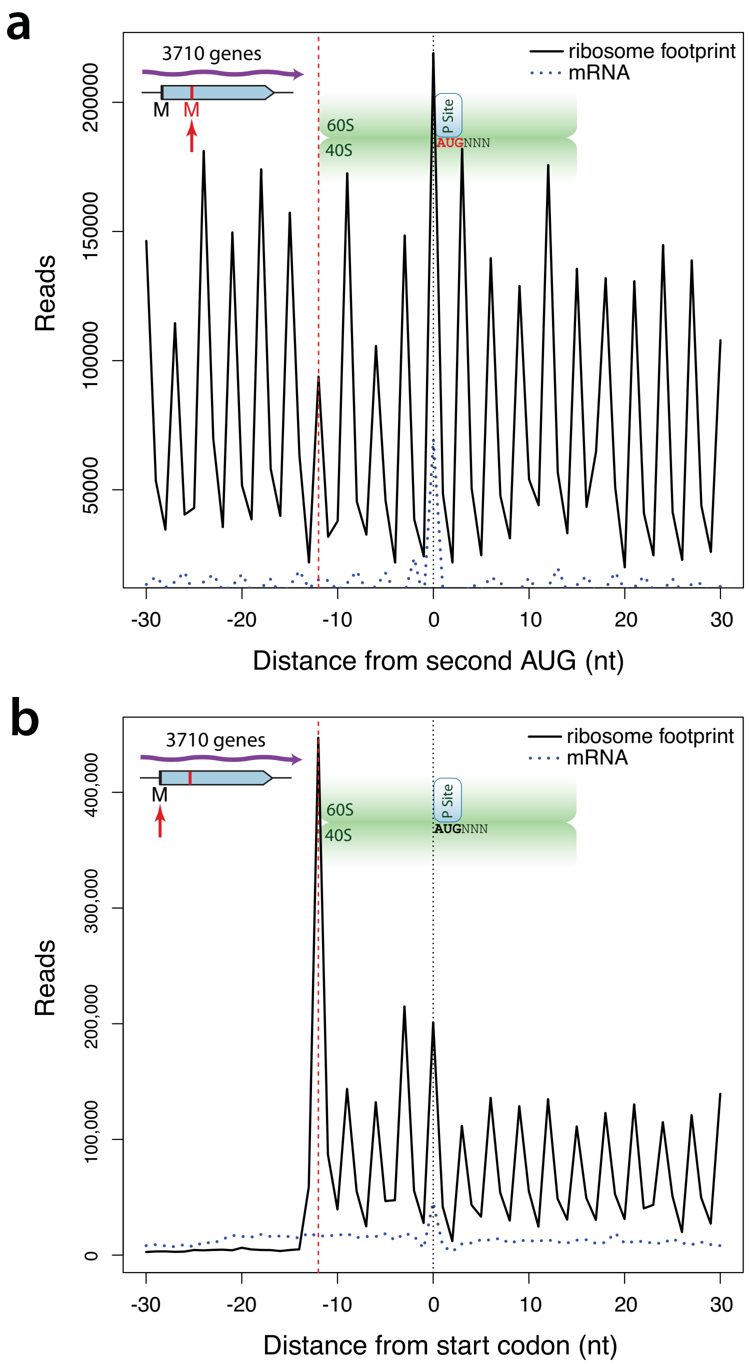


Figure S22. Ribosome profiles in cases where isoform variation could lead to N-terminal truncation of proteins. Control genes with transcripts generally arising before the annotated start codon (*i.e.*, less than 5% of their TIFs starting between the first and second Met) were aligned to the second (**a**) or the first AUG (**b**). The expected ribosome protection (60/40S green box) can only be observed when the genes are aligned using the first AUG, supporting translation initiation at the first AUG and not the second. Reads for both ribosome footprint (black lines) and control mRNA reads (dashed blue lines) are depicted. Data from Gerashchenko *et al.*[16].

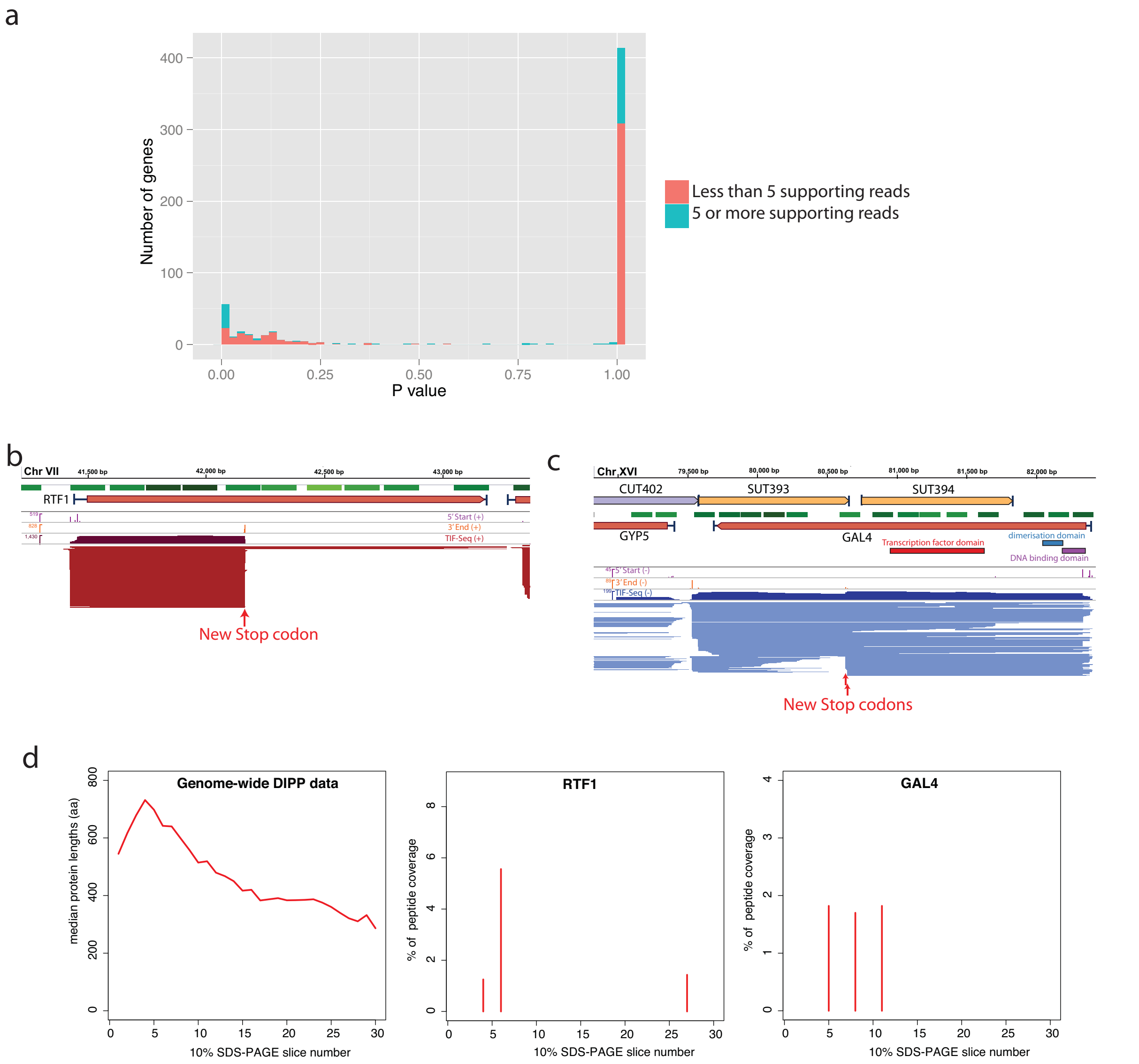


Figure S23. Examples of internal polyadenylation events that introduce early stop codons and likely lead to C-terminal truncation. a, Distribution of P values from Fisher's exact test for genes depending on the number of reads supporting C-terminal truncation. Based on this distribution, only genes with 5 or more supporting reads (blue) were used to determine whether internal polyadenylation tends to introduce stop codons. 33 genes were significantly enriched for polyadenylation events that introduced early stop codons (FDR<10%, Supplementary Table S5). **b-c,** Examples of internal polyadenylation events introducing new stop codons. *RTF1* was profiled in glucose and *GAL4* in galactose. **d,** Position of peptides for indicated proteins in a 10% SDS-PAGE fractionation followed by mass spectrometry detection (DIPP)[33]. SDS-PAGE gene slices are sorted from high to low molecular weight. The detection of peptides in different SDS-PAGE slices corresponding to different molecular weights (red bars) is in agreement with the production of truncated proteins. Aditionally, truncated GAL4 proteins have previously been detected by immunoblotting, but were assumed to be proteolytic fragments [34].

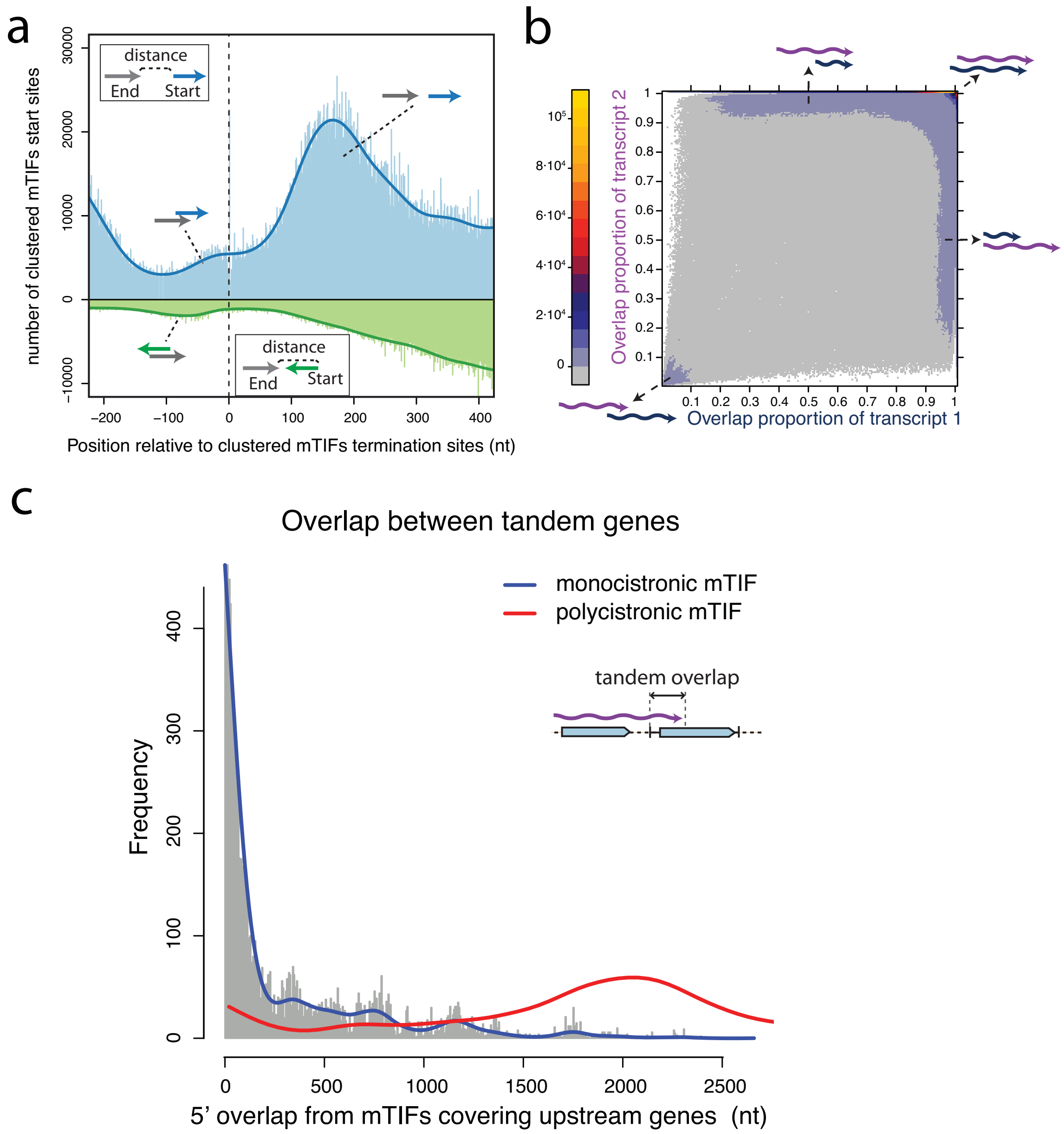


Figure S24. Same-strand overlap of transcripts from tandem genes is common. a, Number of identified mTIF start sites on sense (blue) and antisense strands (green). Both overlapping tandem (upper left) and internal antisense transcripts (lower left) were observed. **b,** Pairwise analysis of the genomic overlap of mTIFs genome-wide. mTIFs corresponding to the same gene present a high degree of overlap (top-right corner), while neighboring tandem genes express mTIFs that overlap by less than 10% of their sequence (bottom-left corner). **c,** TIFs covering distinct tandem genes typically overlap by less than 200 bp (histogram and blue line). If the overlap is larger, they tend to produce fully bicistronic RNAs (red line).

Supplementary Methods

Method overview

The TIF-Seq method developed in this study allows unambiguous identification of transcript isoforms genome-wide by the concurrent sequencing of both 5’ and 3’ ends of each mRNA molecule. A conceptual overview of the process is in Supplementary Fig. S1 and further technical details are below. TIF-Seq is conceptually similar to the independently developed RNA-PET technique[25] but, in contrast, includes critical controls to assess both its accuracy and library complexity. This has enabled us to comprehensively profile and discover new features in the budding yeast transcriptome with general implications for gene expression and protein diversity.

The first step of the protocol consists of selectively amplifying RNA molecules with both a 5’ cap structure and a poly(A) tail. To select for capped molecules, we used the popular and established oligo-capping method[29], which selectively attaches an oligo of known sequence to capped RNA molecules. First, we desphosphorylated the 5’ end of the non-capped molecules (which include *e.g.*, mRNA degradation intermediates). We then used TAP (Tobacco Acid Pyrophosphatase) to remove the 5’ cap structure of the “*bona-fide*” mRNAs and expose the 5’ phosphate group which could consequently be used for strand-specific RNA ligation. Next, the oligo with known sequence was ligated to the 5’ end of the formerly capped RNA molecules. Although alternative methods are available for 5’ end mapping[25,35], we have chosen this one because it allows us to assess the RNA quality (*e.g.*, by Agilent RNA Bioanalyzer) at this intermediate step. Additionally, it allowed us to modify the approach to map different species of RNAs (*e.g.*, control experiments for only non-capped mRNA molecules) that cannot be easily studied with a cap-trapper approach[25].

Following selection of capped molecules, the RNA was subjected to retrotranscription (optimized to produce long cDNA molecules) followed by a brief PCR amplification. This initial amplification step is essential to maintain the complexity of the sample. In fact, we produced also libraries without this initial PCR amplification (*e.g.*, as performed in RNA-PET[25]); however the complexity of the sample was dramatically decreased and an artifactual isoform homogeneity produced. One advantage of TIF-Seq over RNA-PET[25] is that we can better control for the presence of PCR duplicates and the complexity of the sequencing library. In the case of TIF-Seq we can identify, for each mapped transcript, not only the TSS and TTS but also the fragmentation points produced during the sonication and Illumina library construction (points 5-6 in Supplementary Fig. S1). By using these fragmentation positions as molecular barcodes, we were able to determine that (if the initial PCR amplification is skipped) most of the sequenced reads originated from a small fraction of the input mRNA molecules, producing an apparently homogeneous population of isoforms that in reality arose from PCR duplicates. This loss in complexity is likely due to a bottleneck in the number of mRNA molecules that reach the final library stage, and can be overcome by intermediate sample amplifications (preliminary experiments, data not shown). We intentionally avoided initial size selection of the FlcDNA (as performed in the GIS-PET approach[36] and RNA-PET[25]), as we were also interested in detecting small RNAs (*e.g.*, CUTs or the short coding RNAs described in this study). Our only size selection was an enrichment for longer (2 kb) FlcDNAs to compensate for the protocol's bias towards shorter molecules (Supplementary Fig. S5), in order to expand the range of molecule sizes in the study (Supplementary Fig. S6). To control for the production of chimeric molecules during library construction, we used a double barcoding scheme (similar to Fullwood *et al.*[37]). Briefly, the oligo-capped RNAs sample was split into two independent tubes to produce full-length cDNA (FlcDNA) using primers containing tube-specific barcodes. Next, the FlcDNA was PCR-amplified with specific primers producing molecules with particular biotinylated chimera control barcodes at the 5' and 3' ends. As those barcodes were added in two independent reactions, the produced FLcDNA molecules contain either chimera control barcodes A-A or B-B (at the 5'-3' ends respectively). After this step, to improve intermolecular ligation efficiency, sticky ends were generated by digestion with the NotI enzyme. NotI recognizes the sequence GCCGGCCGC, which is extremely rare in the AT-rich yeast genome. In fact, there are only 4 genes (*GRX3*, *UTR4*, *DOT6* and *PUF3*) out of 5886 annotated yeast ORFs that contain NotI recognition sequences in their coding region. It is possible that the number of isoforms for these genes was underestimated; however, this is not expected to have a significant impact on our description of yeast transcriptome architecture.

The FLcDNA with sticky ends was then pooled and subjected to intramolecular ligation with extensive dilution. In this step, the previously introduced chimera control barcodes allow an accurate estimation (below 0.19%, Supplementary Fig. S4) of the intermolecular (chimeric) ligations, as the presence of the barcodes A-B or B-A in the final products can only arise from molecular chimeras produced during this step. After this circularization step, the remaining non-circularized molecules were degraded by incubation with exonucleases I and III, and the sample was further purified.

The circularized molecules were then subjected to sonication to produce linear fragments. The fragments spanning the connection between the 5' and 3' ends were captured by streptavidin-biotin pulldown in a way similar to the standard Illumina mate-pair approach. Next, a standard Illumina library preparation was performed using barcoded forked adaptors to allow sample multiplexing[38]. The library was subjected to a stringent size selection (~300 bp) to maximize the molecules that would produce reads long enough to map both 5' and 3' ends, while permitting the simultaneous identification of the poly(A) site and the chimera barcode combination. Most reads mapped either to the 5' or 3' end, however only reads where both ends and chimera control barcode were identified simultaneously were considered for our analysis. These correspond to ~5-10% of the reads (depending on the sequencing quality), however if less stringent criteria for chimera barcode identification or longer sequencing reads were used, this percentage would increase.

Finally, the fragments spanning the connection between 5' and 3' ends were sequenced with an Illumina HiSeq 2000 with paired-end 105 bp reads. Transcript 5' and 3' end sequences were extracted from the paired-end reads and mapped to the reference genome. In the end, transcript isoforms were reconstructed using single-nucleotide resolution maps of starts and ends sites along the genome.

Method validation

To establish TIF-Seq as a robust approach to identify transcript isoforms, we performed extensive controls throughout the protocol as well as independent validations of our findings, at both genome-wide and single-gene levels.

As a first assessment of the quality of the method, we performed both technical and biological replicates. TIF-Seq allows the reproducible quantification of transcript isoforms (Spearman correlation (ρ) > 0.77) even when considering reads per individual TIFs (isoforms mapping to the same start and end bases, Supplementary Fig. S3). As expected, this correlation improves when combining all the isoforms overlapping a certain annotated transcript region (ρ = 0.98).

While TIF-Seq read coverage depends on transcript length like RNA-Seq, it is biased towards short molecules instead of long ones (Supplementary Fig. S5). Despite the well-known tendency of longer mRNA molecules to be less efficiently reverse transcribed into full-length cDNA[39], raw TIF-Seq read coverage correlates well (ρ = 0.63) with previous absolute RNA amount estimations by tiling arrays[5]. To improve the detection of long molecules, we increased the recovery of long transcripts with an additional enrichment for long RNA molecules (>2kb) via gel size-selection after the production of full-length cDNA (Supplementary Fig. S6). However, to maintain complexity among the short transcripts, we only performed this enrichment in addition to the size-selection free approach (without replacing it).

Internal controls

To estimate the number of TIFs that could arise from intermolecular ligations during the circularization step, we used the double barcoding scheme for chimera control described above. As the FlcDNA samples contain either A-A or B-B combinations, the appearance of A-B or B-A combinations in the final library indicates the frequency of these intermolecular chimeras, which we found to be reasonably low (*e.g.*, less than 4.17% in YPD, Supplementary Fig. S4). However, chimeras could also be present in samples with A-A or B-B combinations (*e.g.*, resulting from the ligation of two molecules with A barcodes). To decrease the number of false positives in our final dataset, we performed additional bioinformatic filtering. We benchmarked the filters against the reads containing A-B or B-A configurations, as we know that all of them resulted from chimeric intermolecular ligations. After extensive optimization, we used conservative filters selecting only those TIFs mapping to the same chromosome and spanning a distance between 40 and 5000 bp. These filters specifically decreased the configurations A-B and B-A (Supplementary Fig. S4), allowing us to estimate a false positive rate in the A-A/B-B configurations of below 0.19%.

Although previous studies have used similar techniques to identify interchromosomal transcripts[25,36], we found that more instances of interchromosomal transcripts arose from the apparent fusion of highly expressed families of transcripts with high sequence homology (*e.g.*, ribosomal Protein isoforms, TDH family). This, along with the propensity of reverse transcriptase to perform template switching between homologous templates[40], suggests that the most common origin of these apparently fused transcripts is artefactual. In this study, although we cannot exclude the existence of some truly fused transcripts, we discarded all interchromosomal transcripts using the conservative approach described above.

To estimate the accuracy of our identification of the 5’ and 3’ ends, we combined *in vitro* synthesized RNA molecules from *B. subtilis* with the starting RNA sample. These molecules, which contain an encoded poly(A) tail, were purified and capped *in vitro*. Afterwards, they were ‘spiked in’ to the initial sample and underwent the entire library construction process. Assuming that most of these molecules should be identical (with a well-defined 5’ and 3’ end), we were able to estimate the accuracy of our approach. We focused our analysis on the shorter transcript (pGIBS-LYS), which is the more abundant with a size of 1kb (a size more comparable to our TIF population transcripts, Supplementary Fig. S6). 93.7% of molecules mapped exactly to the expected 3’ end, and 95.4% to the exact 5’ site. Considering both ends simultaneously, 93.5% of the TIFs mapped within a 5bp window of the expected site. It should be noted that although a production of a homogeneous population of synthetic RNAs with precise 5’ and 3’ ends was intended, during classical *in vitro* transcription, some non-templated nucleotides can be added at both ends, producing micro-heterogeneous full-length RNAs[41]. The presence of such RNAs during the *in vitro* capping reaction would produce a small proportion of capped RNAs with transcript start sites that differ from the expected ones, which we would detect due to the high sensitivity of TIF-Seq. Therefore, the technical accuracy of TIF-Seq boundaries should be even higher than what we observed in the spike-in controls.

Despite these validations of TIF-Seq accuracy, we took the conservative approach of defining the major transcript isoforms by clustering the transcripts with each of their 5' and 3' end sites co-occurring within 5 bp. This window size was chosen because when comparing the typical distance between neigbouring TSSs, start sites separated by 8nt are overrepresented (Supplementary Fig. S2a). This common occurrence can also be observed in the sequence motif associated with the start sites (secondary A peak situated at -8nt in Supplementary Fig. S15a). Thus, our 5 bp clustering window is large enough to be above our technical precision and small enough to provide information about these common transcription isoform variations (Supplementary Fig. S2).

Assessment of transcript boundaries

To independently assess the accuracy of our transcript boundary mapping, we compared our results with independent datasets available in yeast. We found that our results are in agreement with the measurement of population-level boundaries obtained previously by tiling microarrays[5] (Fig. 1b-c).

To further validate our approach and confirm that the majority of our TIFs come from *bona-fide* capped molecules, we performed two additional experiments where we modified our protocol to capture polyadenylated but non-capped molecules from cells grown in glucose and galactose media (Fig. 1b-c). The resulting TIFs have the expected features of mRNA degradation intermediates: they are shorter (Supplementary Fig. S6) and their 5' ends (not capped in this case) do not align to the canonical TSSs (Fig. 1c).

Although it is technically possible that our final dataset contains a minor contamination of fragmented molecules that escape our extensive dephosphorylation, this is not expected to noticeably affect our dataset for the following reasons: 1) In our *in vitro* transcript controls, the identified 5' ends of most (95.4%) molecules correspond to the expected position; 2) Assuming this undesired fragmentation occurs randomly, it is unlikely that such molecules are identified more than twice (*i.e.,* supported by more than two sequencing reads) with identical 5' and 3' sites, which is one of our filtering criteria; 3) We performed control experiments on non-capped molecules (Fig. 1b) and found that the resulting dataset is fundamentally different; 4) To prevent such possible contaminations from affecting our assessment of the number of mTIFs per gene, we focused only on those that completely cover the coding region. Overall, we cannot completely exclude the possibility that our raw dataset contains minor amounts of these molecules, but for the reasons above, we believe they should not affect our conclusions.

Validation of new features

Our genome-wide map of transcription with concurrently identified 5' and 3' ends corroborates the few previously identified bicistronic transcripts[42,43]. Our dataset also explains why some transcripts that appear to be bicistronic (*e.g.*, *YMR181C-RGM1*, Supplementary Fig. S7c)) according to hybridization-based techniques[44,45] could not be confirmed by full-length cDNA cloning strategies[42]: these transcripts are expressed in a complex pattern of overlapping TIFs on the same strand, but are not (or very rarely are) true bicistronic mRNAs.

Target sequencing validation through ssRNA circularization

In addition to the independent genome-wide validation of the 5' and 3' ends shown above, we validated the simultaneous 5'-3' variability for selected loci of interest (Supplementary Fig. S8). For this, we used an independent method based on ssRNA circularization, divergent RT-PCR, and cloning (Supplementary Fig. S7a). Instead of constructing full-length double-stranded cDNAs and ligating them as in our main approach, we directly circularized the capped mRNA by single-stranded RNA ligation. To do so, we first dephosphorylated all the non-capped mRNA molecules and then removed the cap structure with TAP. The mRNAs were then subjected to single-stranded RNA circularization by T4 RNA ligase. The circularized RNA was used as a template for retrotranscription and divergent PCR of the loci of interest. This produces a diversity of molecules spanning the 5'-3' connection of the TIFs of interest. The PCR products were cloned into *E. coli* and multiple colonies from each transformation (each harboring an insert coming from a different mRNA molecule) were Sanger sequenced. The sequenced regions (spanning the 5' start, poly(A) tail, and 3' end) were mapped to the genome and compared to the results obtained from the main TIF-Seq protocol. This validation confirmed the presence of 48 new isoforms detected by TIF-Seq and the existence of TIF variability at nucleotide-resolution level. It failed to detect only some long molecules that were not efficiently amplified and cloned and would likely require hundreds of clones to be sequenced. Detailed results are in Supplementary Table S3.

Finally, to confirm the existence of some of the new transcripts without using enzymatic treatment, we performed strand-specific northern blots (Supplementary Fig. S8). These blots confirmed the existence of several new short coding RNAs (scRNA) and independent transcripts for uORFs.

Summary of the samples and strains

We applied TIF-Seq to profile the transcriptome of wild-type yeast in the two main laboratory growth conditions (YPD and YPGal). Multiple biological and technical replicates were performed, as shown in Supplementary Table S1.

Assessment of TIF start and end dependence

We explored the dependence between start and end sites of 4259 genes that were fully covered in the YPD non-size selected library. Due to the possible effects of differential transcript lengths on this analysis, we restricted it to TIFs that completely cover coding regions. We computed Spearman correlations between 5' and 3' end positions. *P* values were further adjusted with FDR test by filtering out the genes with less than 50 reads covering the coding region[46].

Distribution of TIF variation and RNA binding proteins

We explored whether RNA binding protein sites are preferentially found in variable regions of UTRs (those included only in a subset of mTIFs) rather than constant ones (present in all mTIFs). For this, we defined the variable UTR regions for each gene as the maximal 5' and 3' UTR regions covered by at least one but not all isoforms, and the constant UTR regions as those covered by all isoforms. We then tested if the variable UTR regions are enriched for RNA binding protein sites compared to the constant UTR regions. Only mTIFs completely covering the coding region of the 4806 verified or uncharacterized ORFs region were considered. Binding sites from 14 RNA binding proteins were mapped according to their binding site motif[13]. The numbers of RNA binding protein sites per 100 nt present in the variable and constant UTR regions in each genes were calculated and significance was assessed using the Wilcoxon rank-sum test.

Statistical estimation of protein truncation by internal polyadenylation

To determine whether internal polyadenylation events significantly tend to introduce non-templated, early stop codons, we considered both the presence and the sequence bias of the polyadenylation events (Supplementary Table S5). As single base pair resolution was needed to resolve these events, and to avoid any possible distortion originating from the mTIF clustering, we restricted our analysis to the TIF dataset. We calculated the number of events introducing new stop codons (observed in-frame) and the events affecting the same sequence but that could not introduce stop codons because they were out of frame (observed out-frame). We performed this analysis independently for TAC/TAT codons (where a polyadenylation site located after the first T in frame would convert them to the stop codon TAA) and TGT/TGC/TGG codons (where a polyadenylation site located after the second G in frame would convert them to the stop codon TGA). We also calculated the random expectation number of in- and out-frame occurrences of the codons TAC/TAT or TGT/TGC/TGG in each gene (expected in- and out-frame). In total, 539 genes that have at least polyadenylation event inside the tested codons (in- or out-frame) were tested with Fisher's exact test for the ratio of observed and expected number of in- and out-frame codons. The distribution of *P* values suggests the existence of evolutionary conserved events

(Supplementary Fig. S23a). To increase our statistical power[46], we restricted our analysis only to those genes with 5 or more reads supporting the introduction of non-templated stop codons.

Supplementary Discussion

Origins of transcript diversity. To understand how these alternative isoforms arise, we investigated their chromatin context and promoter sequences. Most mTIFs originate at the edge of nucleosome-depleted regions[5] that contain sequence motifs consistent with previous reports[9,47] (Supplementary Fig. S15). In agreement with the density of the corresponding nucleosome-depleted regions, we observed more variability among poly(A) sites than transcript initiation sites (Fig. 2c and Supplementary Fig. S14). In addition, we found more transcript start site (TSS) variability associated with TATA-containing promoters than with TATA-less (34 vs. 24 nucleotides, Supplementary Fig. S16-17). Most mTIFs arise downstream of the annotated preinitiation complex (PIC)[11] and their TSS positioning is strongly delimited by the +1 nucleosome position (Supplementary Fig. S16-17). This supports the theory that TSS selection during RNAPol II scanning[11,48] is fine-tuned by nucleosome positioning. However, the interplay between PIC assembly and the +1 nucleosome position can only explain small TSS variations during RNAPol II scanning. Larger differences (*e.g.*, TIFs originating from different nucleosome-depleted regions) and the common existence of TSSs upstream of annotated PICs likely require the existence of alternative PICs (*e.g., ALT1,* Fig. 2d, 2e and Supplementary Fig. S16-17).

Notably, during the transcription process the use of specific TSSs and polyadenylation sites are not always independent events. We identified 382 genes with significant associations between transcript start and end positions (FDR<10%, Supplementary Fig. S18 and Data S4), where the use of specific start sites can influence the polyadenylation site that will be used. The distribution of *P* values also indicates that this might be a conserved phenomenon (Supplementary Fig. S18c). This suggest the existence of crosstalk between promoter and termination regions, in agreement with the established 3D organization of genes forming gene loops[12].

Implications of overlapping RNA heterogeneity on transcriptome architecture. The common overlap of tandem genes and the existence of bicistronic transcripts raise the question of how the distinction is made between the RNA polymerases that transcribe the upstream gene that should be terminated and the RNA polymerases transcribing the downstream gene that, if terminated, would produce a short cryptic transcript. The fact that most upstream transcripts stop within the first 100-200 bp of the downstream transcript (Fig.24c) supports a hypothesis where the RNA polymerase transcribing the downstream gene may not be able to recruit termination factors[22] and thus would be likely insensitive to any polyadenylation and termination signals. In support of this hypothesis, tandem overlapping transcripts that do not terminate within this region tend to produce bicistronic transcripts. These observations indicate that the bicistronic transcripts we detected here may result from inefficient polyadenylation and termination of the upstream transcript due to extreme gene compaction of the yeast genome. However, this mechanism can only explain extreme differences between overlapping transcription units, as is the case for tandem overlapping transcripts. Explaining more subtle differences, such as the interdependence between specific TSS and polyadenylation sites (Supplementary Fig. S18), would require other mechanims where the RNA polymerase fate could be determined during transcription initiation.

Supplementary Tables

Supplementary Table S1. Summary of the samples sequenced. A total of 7 HiSeq 2000 lanes were used. Raw data can be obtained via GEO accession number GSE39128.

Samples	Strain	Media	Biological replicates	Technical replicates	Libraries sequenced	Notes and samples
YPD	SLS045	YPD	2	3	6	ypd_bio1_lib1 ypd_bio2_lib1 ypd_bio1_lib2 ypd_bio2_lib2 ypd_bio1_lib3 ypd_bio2_lib3
GAL	SLS045	YPGal	4	2	6	ypgal_bio1_lib1 ypgal_bio2_lib1 ypgal_bio3_lib2 ypgal_bio4_lib2 ypgal_bio3_lib3 ypgal_bio4_lib3
long_YPD	SLS045	YPD	2	1	2	Enriched for long molecules: ypd_bio1_lib4 ypd_bio2_lib4
long_GAL	SLS045	YPGal	2	1	2	Enriched for long molecules: ypgal_bio1_lib4 ypgal_bio4_lib4
nc_YPD	SLS045	YPD	1	1	1	Library for non-capped RNAs: nypd_bio2_lib1
nc_GAL	SLS045	YPGal	1	1	1	Library for non-capped RNAs: nypgal_bio3_lib1

Supplementary Table S2. Oligonucleotide sequences used in this study.

TIF-Seq method

Oligo	Sequence
5oligocap	dCdAdCdTdCdTrGrArGrCrArArUrArCrC
3cDNANotI_A	TATAGCGGCCGCCTGGAGTTTTTTTTTTTTTTTTVN
3cDNANotI_B	TATAGCGGCCGCCTCCACTTTTTTTTTTTTTTTTVN
5BiotNotI_A	TATAGCGGCCGCTA[BtndT]CACTCTGAGCAATACC
5BiotNotI_B	TATAGCGGCCGCTA[BtndT]ATCACTCTGAGCAATACC
3AmpNotI_A	CATGTATAGCGGCCGCCTGGAG
3AmpNotI_B	ACATGTATAGCGGCCGCCTCCAC

Barcoded forked adaptors#

Oligo	Sequence
mp1PE1	[Phos]AGCGCTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

mp1PE2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCGCT*T
mp5PE1	[Phos]ACAGTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
mp5PE2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCACTGT*T
mp19PE1	[Phos]CGGAATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
mp19PE2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTCCG*T
mp22PE1	[Phos]CTATACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
mp22PE2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTATAG*T
mp34PE1	[Phos]GGTAGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
mp34PE2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTACC*T
mp37PE1	[Phos]GTTTCGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
mp37PE2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGAAAC*T

Divergent oligos for targeted sequencing

Oligo	Sequence
dCOX19-L	GCATGTACTTGGTGCATTCG
dCOX19-R	GAATCCTGGCTCCACATCTC
dFPR1-L	TCAACGGAGGAATCGAATTT
dFPR1-R	TATCAATTGAGCCGCCTTTC
dGAL10-L	AGGGAATGTGATGCTTGGTC
dGAL10-R	CTGAAATGGCAGACCGAGTT
dRGM1-L	GCTGCTTGTTATCTTACAGCAATG
dRGM1-R	TTTTGTTCCTCGACCCCTTA

Northern blot probe generation

Oligo	Sequence
T3- uRGM1-L	AATTAACCCTCACTAAAGGGGAGACAACCTTAATTAATTCGCACGTC
T7-uRGM1-R	TAATACGACTCACTATAGGGAGATTGTTCTGAAATCAGGAGAATTAGAG
T3- uGCN4-L	AATTAACCCTCACTAAAGGGGAGATGCCCATCAGTTTCACTAGC
T7-uGCN4-R	TAATACGACTCACTATAGGGAGATTGAGCAGACAAATTGGTAAACA
T3- uPCL7-L	AATTAACCCTCACTAAAGGGGAGATGCACTTTTCTACGGGCTCT
T7-uPCL7-R	TAATACGACTCACTATAGGGAGAATATAAAGGTTTGAGATGTTGAAGC

Notes:

dNTP - desoxyribonucleotide
rNTP - ribonucleotide
[Phos] - Phosphorylation
[BtndT] - Biotin dT
* S-linkage between the two bases
# Oligonucleotide sequences © 2006-2012 Illumina, Inc. All rights reserved.

Supplementary Table S3. Transcript isoforms identified by ssRNA circularization and targeted sequencing.

Gene of interest	Clone	Chromosome	5'start	3'end
COX19	COX19_03	chr12	109038	108065
COX19	COX19_07	chr12	109016	108192

COX19	COX19_09	chr12	109023	108065
COX19	COX19_10	chr12	108928	108059
COX19	COX19_12	chr12	109023	108059
COX19	COX19_14	chr12	109028	108192
COX19	COX19_15	chr12	109023	108088
COX19	COX19_16	chr12	109028	108192
COX19	COX19_17	chr12	109022	108130
COX19	COX19_19	chr12	108979	108192
COX19	COX19_20	chr12	108979	108192
COX19	COX19_22	chr12	109026	108264
COX19	COX19_23	chr12	109028	108069
COX19	COX19_24	chr12	109010	108192
COX19	COX19_25	chr12	109012	108069
COX19	COX19_26	chr12	109001	108192
COX19	COX19_29	chr12	109023	108193
COX19	COX19_31	chr12	109023	108191
COX19	COX19_32	chr12	109026	108264
COX19	COX19_33	chr12	109023	108103
FPR1	FRP1_01	chr14	372253	371755
FPR1	FRP1_02	chr14	372245	371728
FPR1	FRP1_03	chr14	372228	371747
FPR1	FRP1_04	chr14	372256	371772
FPR1	FRP1_06	chr14	372245	371689
FPR1	FRP1_07	chr14	372256	371689
FPR1	FRP1_08	chr14	372235	371772
FPR1	FRP1_09	chr14	372256	371756
FPR1	FRP1_10	chr14	372235	371774
FPR1	FRP1_11	chr14	372241	371738
FPR1	FRP1_12	chr14	372241	371755
FPR1	FRP1_13	chr14	372253	371747
FPR1	FRP1_14	chr14	372256	371750
FPR1	FRP1_17	chr14	372256	371728
FPR1	FRP1_18	chr14	372241	371746
FPR1	FRP1_19	chr14	372245	371747
FPR1	FRP1_21	chr14	372241	371728
FPR1	FRP1_22	chr14	372245	371689
FPR1	FRP1_23	chr14	372256	371728
FPR1	FRP1_24	chr14	372241	371721
FPR1	FRP1_25	chr14	372241	371728
FPR1	FRP1_26	chr14	372241	371766
FPR1	FRP1_27	chr14	372235	371762
FPR1	FRP1_28	chr14	372235	371689
FPR1	FRP1_31	chr14	372241	371689
RGM1	RGM1_02	chr13	624258	624093
RGM1	RGM1_04	chr13	624258	624092
RGM1	RGM1_09	chr13	624258	624093
RGM1	RGM1_11	chr13	624275	624093

RGM1	RGM1_12	chr13	624258	624093
RGM1	RGM1_6	chr13	624258	624093
RGM1	RGM1_7	chr13	624258	624093
GAL10	GAL10_10	chr2	278379	277318
GAL10	GAL10_13	chr2	278375	277098
GAL10	GAL10_17	chr2	278379	277318
GAL10	GAL10_18	chr2	278375	276956
GAL10	GAL10_3	chr2	278379	277305
GAL10	GAL10_30	chr2	278379	277305
GAL10	GAL10_6	chr2	278375	277283
GAL10	GAL10_7	chr2	278379	277142
GAL10	GAL10_9	chr2	278379	277305

Supplementary Table S4. Genes with significantly differential use of the first and second AUG between YPD and YPGal. Statistical analysis performed using DEXSeq[49] on non-size-selected TIF-Seq experiments.

Gene	ORF	mRNAs truncated in YPD	mRNAs truncated in YPGal	p- value	Adjusted p-value
RPL23A	YBL087C	1.50%	0.34%	1.05E-06	0.0012709
RIB5	YBR256C	8.72%	3.39%	1.47E-05	0.008918186
RPL23B	YER117W	0.35%	0.13%	4.68E-05	0.014161126
FUM1	YPL262W	64.91%	23.55%	3.64E-05	0.014161126
FIT3	YOR383C	12.82%	0.88%	0.000158381	0.035634893
RGL1	YPL066W	42.11%	93.25%	0.000176556	0.035634893
SUC2	YIL162W	92.52%	66.79%	0.000249776	0.043211306
CPA1	YOR303W	11.30%	0.50%	0.000359922	0.054483192
BUD20	YLR074C	2.09%	0.00%	0.000579445	0.077967491

Supplementary Table S5. Genes with significant in-frame C-terminal truncation by internal polyadenylation. Systematic ORF name, observed and expected polyadenylation events introducing non-templated stop codons (in-frame) or negative controls (out-frame), affected codon, p-value and adjusted p-value are displayed. Only genes with 5 or more events supporting C-terminal truncation by internal polyadenylation were considered.

Gene	in-frame observed	out-frame observed	in-frame expectation	out-frame expectation	Codon	p-value	Adjusted p-value
YBL042C	5	0	36	55	TAC/TAT	1.23E-02	5.98E-02
YBR019C	200	39	39	49	TAC/TAT	7.34E-12	1.18E-09
YCR023C	13	2	24	58	TAC/TAT	4.32E-05	8.69E-04
YDR119W	12	9	22	90	TAC/TAT	7.67E-04	7.18E-03
YDR213W	8	0	18	68	TAC/TAT	1.40E-05	3.23E-04
YER178W	7	0	21	23	TAC/TAT	1.02E-02	5.31E-02
YFL026W	13	0	19	37	TAC/TAT	8.99E-06	2.89E-04
YGL140C	5	0	45	91	TAC/TAT	4.90E-03	3.02E-02

YGL244W	28	0	25	27	TAC/TAT	3.11E-07	1.94E-05
YGR125W	7	2	42	96	TAC/TAT	6.61E-03	3.80E-02
YIL083C	8	0	14	34	TAC/TAT	2.25E-04	3.30E-03
YIL107C	6	0	29	73	TAC/TAT	8.48E-04	7.19E-03
YIL109C	5	0	34	60	TAC/TAT	8.05E-03	4.47E-02
YJL059W	6	0	22	37	TAC/TAT	4.56E-03	2.94E-02
YJL191W	17	6	2	14	TAC/TAT	1.84E-04	2.96E-03
YJR072C	22	0	15	17	TAC/TAT	1.20E-05	3.22E-04
YKL205W	8	2	27	100	TAC/TAT	2.81E-04	3.42E-03
YKL205W	6	0	30	96	TGT/TGC/TGG	2.98E-04	3.42E-03
YLR110C	6	0	2	11	TAC/TAT	1.03E-03	8.31E-03
YLR228C	5	0	13	65	TAC/TAT	2.95E-04	3.42E-03
YLR249W	12	0	20	64	TAC/TAT	3.62E-07	1.94E-05
YLR286C	5	0	20	59	TAC/TAT	1.72E-03	1.19E-02
YLR342W	24	13	99	162	TAC/TAT	1.78E-03	1.19E-02
YLR443W	7	0	23	44	TAC/TAT	1.13E-03	8.67E-03
YMR185W	5	0	34	104	TAC/TAT	1.24E-03	9.08E-03
YNL178W	6	0	7	12	TAC/TAT	9.69E-03	5.20E-02
YPL054W	6	0	14	19	TAC/TAT	1.19E-02	5.98E-02
YPL086C	10	0	34	41	TAC/TAT	7.93E-04	7.18E-03
YPL248C	6	2	30	95	TGT/TGC/TGG	5.06E-03	3.02E-02
YPR072W	8	0	20	40	TAC/TAT	4.21E-04	4.51E-03
YHL020C	7	0	6	42	TGT/TGC/TGG	8.46E-06	2.89E-04
YIL050W	5	0	6	28	TGT/TGC/TGG	8.02E-04	7.18E-03
YKR086W	6	0	21	99	TGT/TGC/TGG	6.01E-05	1.08E-03

Supplementary Data

Supplementary Data S1. TIFs identified in the wild-type strain in YPD and YPGal. 5' end positions (t5), poly(A) sites (t3), number of sequencing reads supporting the TIF in glucose (ypd) and galactose (gal), annotation type, and the systematic gene/transcript name for all TIFs identified in wild-type yeast.

Supplementary Data S2. Major transcript isoforms (mTIFs) identified in the wild-type strain in YPD and YPGal. 5' end positions (t5), poly(A) sites (t3), number of sequencing reads supporting the mTIF in glucose (ypd) and galactose (gal), annotation type, and the systematic gene/transcript name for all mTIFs identified in wild-type yeast. mTIFs were defined by clustering TIFs as described in Methods.

Supplementary Data S3. Number of mTIFs per ORF. For each protein-coding gene, the number of unique mTIFs identified (mTIF_number) as well as median UTR lengths (median5, median3) and standard deviation of end positions (sd5, sd3).

Supplementary Data S4. Genes with TIF start and end interdependence. For each significant protein-coding gene, the systematic name, p-value, number of reads covering the coding region, and the adjusted p-value (FDR<10%) are displayed. Only reads for the YPD non-size enriched samples and genes with more than 50 reads covering the coding region were considered for the analysis.

Supplementary Data S5. Previously annotated upstream ORF(uORF)/ORF pairs and their coverage by mTIFs in this study. For all previously annotated uORFs[15], annotated genomic positions of uORFs (uORF_first, uORF_last) and main ORFs (ORF_first, ORF_last), plus whether either ('cover uORF', 'cover ORF') or both ('cover both') features are covered by mTIFs detected in this study. uORFs present in mTIFs that never include the ORF ('cover uORF' TRUE, 'cover both' FALSE) in these conditions were considered novel short coding RNAs. Data are visually represented in Fig. 3a.

Supplementary Data S6. **Previously annotated uORFs that can also be classified as dORFs (downstream ORFs) according to TIF-Seq.** Systematic names of annotated uORFs (relabelled here as 'dORF') and their downstream genes (gene2); these features can be reannotated as downstream ORFs due to their inclusion in mTIFs containing their upstream gene (gene1).

Supplementary Data S7. **Genes with TIFs that skip the first AUG codon, thereby encoding N-terminally truncated proteins.** Systematic name of protein-coding genes, number of reads supporting full (full_reads_<condition>) and truncated (truncated_reads_<condition>) versions, and fraction of TIFs supporting N-terminal truncation (percentage_truncated_<condition>). 'Percentage of truncated TIFs' is the fraction of truncated TIFs in all possible translated TIFs (including truncated and full ones). TIFs supporting N-terminal truncation were defined by their coverage of the second AUG but not the first; only genes with >=50% reads supporting truncation in a given condition are reported.

Supplementary Data S8. TIFs with new stop codons produced by internal polyadenylation. Systematic gene name (ORF), strand, TIF coordinates (t5, t3), number of reads supporting polyadenylation events at the specified sequence (codon) and sequencing reads supporting it in wild-type conditions (glucose, ypd_reads; galactose, ypgal_reads; or combined sm) are shown. TIFs are classified according to their ability to introduce new stop codons by internal polyadenylation (frame = 0, observed in-frame) or not (frame = 1 or 2, observed out-frame).

Supplementary Data S9. Tandem gene pairs that express overlapping mTIFs. Following assignment of mTIFs to ORFs, these gene pairs expressed at least one pair of mTIFs from the same strand that overlap.

Supplementary Data S10. Tandem gene pairs covered by bicistronic mTIFs. Gene pairs that were both contained in at least one mTIF in wild-type yeast.

Supplementary References

29 Scotto-Lavino, E., Du, G. & Frohman, M. A. Amplification of 5' end cDNA with 'new RACE'. Nat Protoc 1, 3056-3061, doi:10.1038/nprot.2006.479 (2006).

30 Chiu, M. I., Mason, T. L. & Fink, G. R. HTS1 encodes both the cytoplasmic and mitochondrial histidyl-tRNA synthetase of Saccharomyces cerevisiae: mutations alter the specificity of compartmentation. Genetics 132, 987-1001 (1992).

31 van Dijk, E. L. et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. Nature 475, 114-117, doi:10.1038/nature10118 (2011).

32 Fujita, P. A. et al. The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39, D876-882, doi:10.1093/nar/gkq963 (2011).

33 Lavigne, R. et al. Direct iterative protein profiling (DIPP) - an innovative method for large-scale protein detection applied to budding yeast mitosis. Mol Cell Proteomics 11, M111 012682, doi:10.1074/mcp.M111.012682 (2012).

34 Chasman, D. I. & Kornberg, R. D. GAL4 protein: purification, association with GAL80 protein, and conserved domain structure. Mol Cell Biol 10, 2916-2923 (1990).

35 Carninci, P. Constructing the landscape of the mammalian transcriptome. J Exp Biol 210, 1497-1506, doi:10.1242/jeb.000406 (2007).

36 Ng, P. et al. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2, 105-111, doi:10.1038/nmeth733 (2005).

37 Fullwood, M. J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58-64, doi:10.1038/nature08497 (2009).

38 Lefrancois, P. et al. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. BMC Genomics 10, 37, doi:10.1186/1471-2164-10-37 (2009).

39 Carninci, P. et al. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13, 1273-1289, doi:10.1101/gr.1119703 (2003).

40 McManus, C. J., Duff, M. O., Eipper-Mains, J. & Graveley, B. R. Global analysis of trans-splicing in Drosophila. Proc Natl Acad Sci U S A 107, 12975-12979, doi:10.1073/pnas.1007586107 (2010).

41 Mörl, M., Lizano, E., Willkomm, D. K. & Hartmann, R. K. in Handbook of RNA Biochemistry 22-35 (Wiley-VCH Verlag GmbH, 2008).

42 Miura, F. et al. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proc Natl Acad Sci U S A 103, 17846-17851, doi:10.1073/pnas.0605645103 (2006).

43 Pelechano, V., Garcia-Martinez, J. & Perez-Ortin, J. E. A genomic study of the inter-ORF distances in Saccharomyces cerevisiae. Yeast 23, 689-699, doi:10.1002/yea.1390 (2006).

44 He, F. et al. Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast. Mol Cell 12, 1439-1452 (2003).

45 David, L. et al. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A 103, 5320-5325, doi:10.1073/pnas.0601091103 (2006).

46 Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A 107, 9546-9551, doi:10.1073/pnas.0914005107 (2010).

47 Zhao, J., Hyman, L. & Moore, C. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev 63, 405-445 (1999).

48 Kaplan, C. D., Jin, H., Zhang, I. L. & Belyanin, A. Dissection of Pol II Trigger Loop Function and Pol II Activity-Dependent Control of Start Site Selection In Vivo. PLoS Genet 8, e1002627, doi:10.1371/journal.pgen.1002627 (2012).

49 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. Genome Res 22, 2008-2017, doi:10.1101/gr.133744.111 (2012).