LETTER

doi:10.1038/nature12121

Extensive transcriptional heterogeneity revealed by isoform profiling

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

Transcript function is determined by sequence elements arranged on an individual RNA molecule. Variation in transcripts can affect messenger RNA stability, localization and translation[1], or produce truncated proteins that differ in localization[2] or function[3]. Given the existence of overlapping, variable transcript isoforms, determining the functional impact of the transcriptome requires identification of full-length transcripts, rather than just the genomic regions that are transcribed[4,5]. Here, by jointly determining both transcript ends for millions of RNA molecules, we reveal an extensive layer of isoform diversity previously hidden among overlapping RNA molecules. Variation in transcript boundaries seems to be the rule rather than the exception, even within a single population of yeast cells. Over 26 major transcript isoforms per protein-coding gene were expressed in yeast. Hundreds of short coding RNAs and truncated versions of proteins are concomitantly encoded by alternative transcript isoforms, increasing protein diversity. In addition, approximately 70% of genes express alternative isoforms that vary in post-transcriptional regulatory elements, and tandem genes frequently produce overlapping or even bicistronic transcripts. This extensive transcript diversity is generated by a relatively simple eukaryotic genome with limited splicing, and within a genetically homogeneous population of cells. Our findings have implications for genome compaction, evolution and phenotypic diversity between single cells. These data also indicate that isoform diversity as well as RNA abundance should be considered when assessing the functional repertoire of genomes.

Transcript isoform variation and its functional relevance have been studied in detail for several single genes. For example, pluripotent cells express a dominant, truncated version of p53 that inhibits the function of the full protein, thereby promoting cell proliferation[3]. However, the genome-wide characterization of isoform variation has been limited. Identifying either 5′ or 3′ transcript boundaries individually[6–8] cannot determine the respective co-occurrence of start and end sites, which is essential for ascertaining the functional potential of a transcript. Thus, most studies have attributed variations at either transcript end to changes in the full-length messages. This interpretation is inaccurate in general, owing to transcripts that could arise from neighbouring genes, short abortive transcripts, bicistronic messages, and transcripts with differing lengths that overlap a gene. Thus, an important dimension of transcriptome complexity has remained largely unexplored. Here, we characterized the heterogeneity of transcript isoforms in *Saccharomyces cerevisiae* by jointly sequencing the 5′ and 3′ ends of each RNA molecule using an approach we term transcript isoform sequencing (TIF-Seq).

To capture *S. cerevisiae* transcript isoforms, capped and polyadenylated RNAs were converted into full-length complementary DNA molecules that were subjected to intramolecular ligation, fragmentation and capture of the 5′–3′ junctions through a biotin tag (Fig. 1a, Supplementary Fig. 1 and Supplementary Tables 1 and 2). The start and end sites of individual RNA molecules were then identified at single-nucleotide resolution by paired-end sequencing of the tagged fragments. We applied TIF-Seq to wild-type yeast grown in two conditions (with glucose (YPD) or galactose (YPGal) as the carbon source). We identified the exact 5′ cap and 3′ polyadenylation sites of more than 19 million individual RNA molecules (Fig. 1b, c). These transcripts are arranged in a remarkably complex, overlapping pattern across the genome (Fig. 1d). In addition to genes with variations in their untranslated regions (UTRs) (for example, *CBK1*), we discerned overlapping tandem genes (for example, *GIM3–YCK2*) and bicistronic transcripts (for example, *PGA1–IGO1*) (Fig. 1d). A comparison of our data with separate 5′ and 3′ end maps illustrates that the former cannot distinguish mono- from bicistronic transcripts, and the latter cannot distinguish 3′ UTR variation from short, overlapping 3′ end transcripts (for example, *YNL155W* and the antisense transcript of *YCK2* in Fig. 1d).

Altogether, in a genome containing only ~6,000 open reading frames (ORFs)[9], we detected over 1.88 million unique transcript isoforms (TIFs) (or 776,874 supported by at least two sequencing reads, Supplementary Data 1) that are defined by a unique combination of 5′ and 3′ end sites at single-nucleotide resolution. To enable analysis of major differences, we clustered the transcripts with each of their 5′ and 3′ end sites co-occurring within 5 nucleotides, and selected the highest expressed TIF per cluster as the representative mTIF (major transcript isoform, see Methods), yielding 371,087 mTIFs genome-wide (Supplementary Fig. 2 and Supplementary Data 2). This total corresponds to about half of all TIFs supported by two or more sequencing reads, demonstrating that there are both minor and substantial variations in transcript boundaries. Our further analysis uses TIFs and mTIFs supported by at least two sequencing reads. These numbers represent conservative estimates for transcript diversity, as our detection of isoforms is limited by sequencing depth, RNA abundance and length (Supplementary Information). We verified the accuracy of our transcript isoform mapping with extensive controls and independent confirmations (Supplementary Information, Supplementary Figs 3–8 and Supplementary Table 3).

Our data set reveals the extent to which different classes of transcripts are affected by isoform variation (Fig. 2a and Supplementary Fig. 9). We detected a median of 26 mTIFs (48 TIFs) that cover the coding region per verified or uncharacterized ORF in glucose and galactose (Fig. 2b–e). These mTIFs display a median positional variation of 75 nucleotides (26 for the 5′ start and 36 for the 3′ end, when considered independently) (Supplementary Fig. 10). Notably, this diversity is not dominated by a few highly abundant isoforms: a median of 10 mTIFs (or 29 TIFs) per gene is required to explain 80% of the mRNA population (Supplementary Fig. 11). Isoform heterogeneity is also found in non-coding genes, including an average of 7 mTIFs per stable unannotated transcript (SUT[4]) (Fig. 2b and Supplementary Fig. 12). In addition, we detected thousands of multicistronic mTIFs that cover two or more ORFs (Fig. 2a). Although the number of TIFs is probably higher than what we observed, we estimate a maximum of ~100 mTIFs (or 500 TIFs) per gene (Supplementary Fig. 13). Altogether,

[1]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Stanford Genome Technology Center, Stanford University, Palo Alto, California 94304, USA.

*These authors contributed equally to this work.

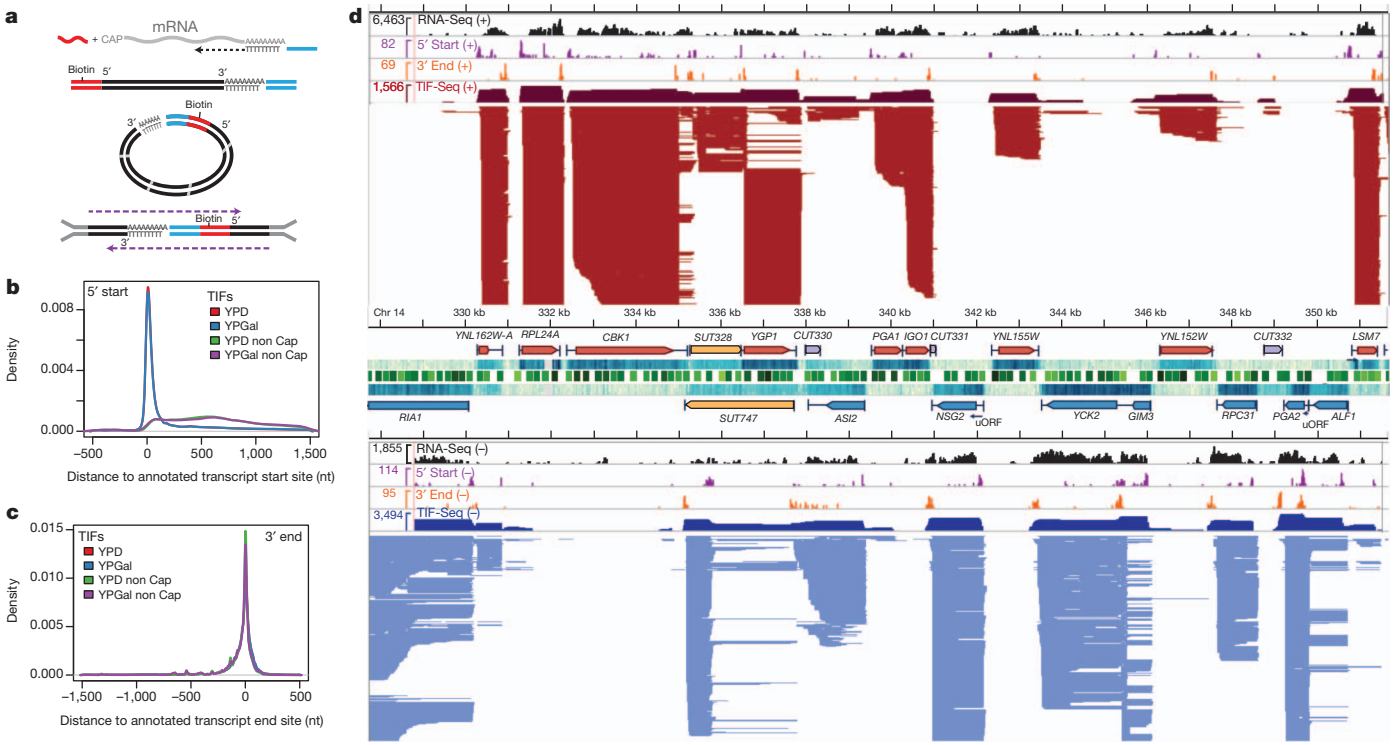

Figure 1 | Genome-wide measurement of transcript isoform diversity using TIF-Seq. **a**, The TIF-Seq method consists of RNA oligonucleotide capping, generation of full-length cDNA, circularization and paired-end sequencing. **b**, **c**, TIF boundaries agree overall with previous determinations of transcript 5′ starts (**b**) and 3′ ends (**c**) derived from tiling array annotations[4]. As expected, TIF-Seq of non-capped mRNAs does not produce many 5′ reads at the annotated transcript start sites (**b**). nt, nucleotides. **d**, Complex landscape of the yeast transcriptome in glucose, showing strand-specific RNA-Seq[25] in comparison to TIF-Seq 5′ start and 3′ end profiles, as well as TIF-Seq coverage in logarithmic scale (dark red/blue upper tracks); the maximum number of reads is indicated in each track. Individual TIFs are represented by red or blue lines (Watson (+) or Crick (−) strand, respectively), each line designating one TIF. Nucleosome positions (green track, darkness indicates significance[26]), expression measured by tiling arrays (blue heat map; darkness indicates expression level), and genome annotation[4] are shown in the centre: annotated ORFs (red and blue boxes for Watson and Crick strands, respectively), their UTRs (black lines), SUTs (yellow boxes), and CUTs (purple boxes). kb, kilobases. SUT, stable unannotated transcript; CUT, cryptic unstable transcript.

5,211 ORFs were covered by at least one mTIF, including 86% of verified or uncharacterized ORFs and 223 dubious ORFs[9] (Supplementary Data 3).

Most TIFs begin as expected: downstream of annotated transcription preinitiation complex (PIC) sites[10] and within the +1 nucleosome (Fig. 2c, Supplementary Discussion and Supplementary Figs 14–17). Notably, some interdependence between transcript start and end sites was observed in 382 protein-coding genes (false discovery rate (FDR) < 10%, Supplementary Fig. 18, Supplementary Data 4 and Supplementary Discussion), supporting the existence of interactions between promoters and terminators[11].

Although it is unclear how much of the isoform variation is functional, we discovered several cases where phenotypic consequences would be expected. First, we observed considerable variation in post-transcriptional regulatory elements. Most genes (66.9% in glucose, 71.9% across both conditions) with putative RNA-binding protein (RBP) sites[12] express mTIFs with different combinations of binding sites (Supplementary Fig. 19). Furthermore, RBP sites are enriched in regions that vary between isoforms ($P < 2.2 \times 10^{-16}$, Supplementary Information). Second, we observed significant variation in upstream ORFs (uORFs), short coding regions in the 5′ UTR that modulate translation efficiency of the downstream gene[13]. The standard understanding is that uORFs are transcribed along with the downstream gene, as both elements must be on the same RNA to interact. Yet over half of the genes with annotated uORFs[14] (703, 59% in Fig. 3a) expressed alternative mTIFs both with and without the uORF (for example, *ICY1* in Fig. 3a, b, Supplementary Data 5). This previously undetected occurrence, in addition to the variation in RNA binding sites, exemplifies transcriptional control of post-transcriptional regulatory potential: the precise isoform transcribed dictates the regulation that can be imposed on the gene.

Notably, our data set reveals that uORFs are not only translational regulators, but can also have an independent identity. Isoforms containing only the uORF were detected for 48% (567) of genes with known uORFs (Fig. 3c and Supplementary Fig. 20). Using ribosome profiling data[15], we found that genes containing genuine uORFs (where the mTIFs always span both the uORF and the main ORF) are significantly less translated than those where the uORFs are in fact independent, misannotated transcripts ($P < 2 \times 10^{-4}$) (Fig. 3d). This is consistent with the expected absence of translational repression by uORFs in the latter case. In addition to re-annotating uORFs, we detected the first downstream ORFs (dORFs), defined as short coding sequences within TIFs that also cover the upstream coding gene (for example, *COX19*, Supplementary Fig. 7 and Supplementary Data 6).

We confirmed the existence of several short transcripts previously misannotated as uORFs by northern blot (for example, *PCL7*, Fig. 3e and Supplementary Fig. 8). The fact that these transcripts have a canonical mRNA structure (5′ capped and polyadenylated), are bound by ribosomes[14], and are evolutionarily conserved (Supplementary Fig. 21) suggests that they are new short coding RNAs (scRNAs, Supplementary Data 5). Short peptides can perform crucial functions, as has recently been described in cellular differentiation[16]. The capacity of TIF-Seq to detect potentially peptide-encoding scRNAs opens new avenues for studying their function and regulation.

We also analysed the impact of transcript variation on protein diversity. Previous studies have identified alternative transcript isoforms that skip the first start codon, leading to loss of amino-terminal signal peptides and to alternative protein localization (for example,

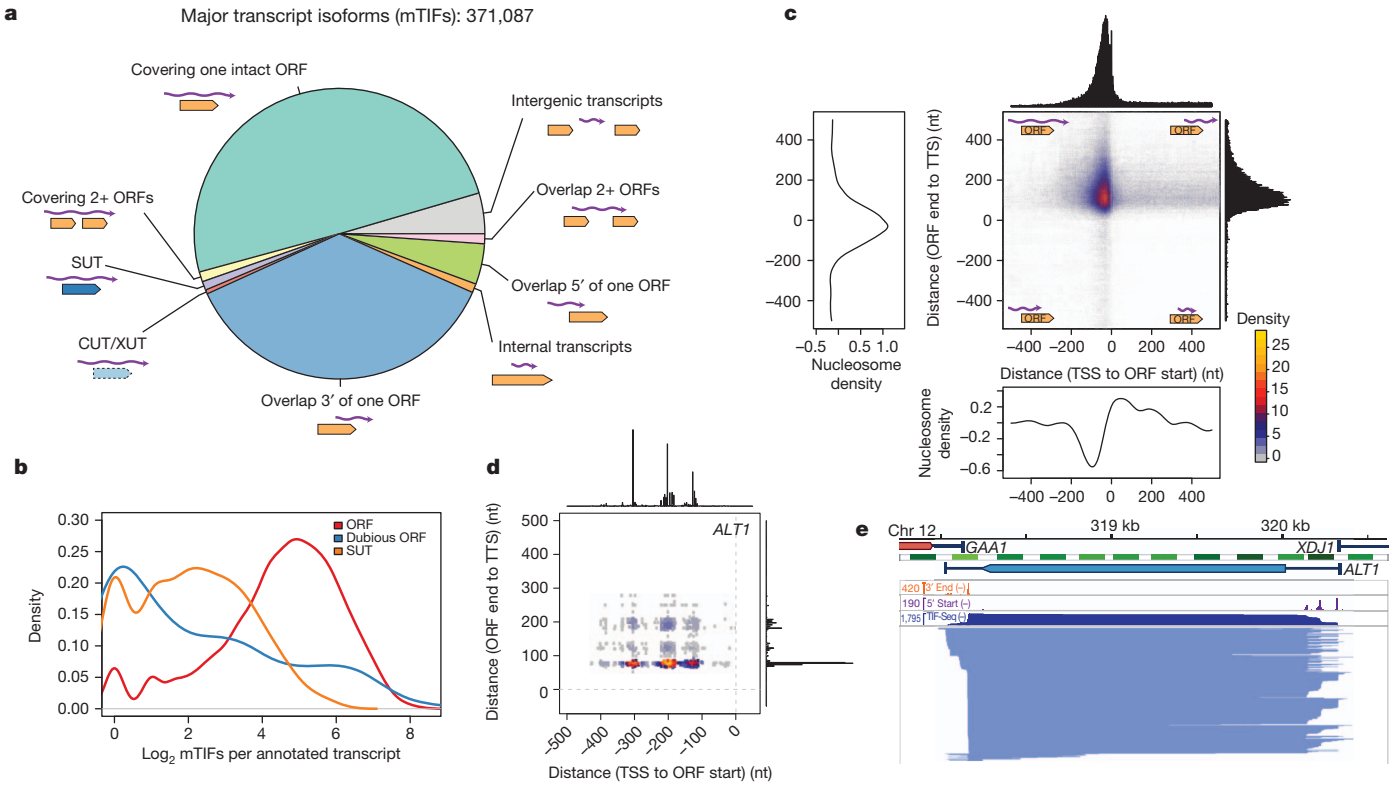

Figure 2 | Extensive isoform diversity revealed among overlapping RNA populations, both at the genomic and single-gene level. **a**, Categories of mTIFs identified in glucose and galactose. XUT, *XRN1*-sensitive unstable transcript. **b**, Log_2-scale distribution of clustered mTIFs per annotated transcript that cover characterized or uncharacterized ORFs (ORFs), dubious ORFs, or overlap more than 80% of stable unannotated transcripts (SUTs)[4]. **c**, Transcript end distance to ORF stop codon (*y* axis) versus transcript start distance to ORF start codon (*x* axis) genome-wide, revealing that most mTIFs cover the entire ORF. Decreased nucleosome density[27] coincides with peaks in transcript start and end site distributions. TTS, transcript termination site; TSS, transcript start site. **d**, Boundaries of TIFs covering *ALT1* relative to ORF boundaries (as in **c**). **e**, Structure of TIFs overlapping *ALT1* in glucose. 5′ start, 3′ end and TIF-Seq coverage in natural scale. Nucleosome and genome annotations as in Fig. 1d.

SUC2 (ref. 2), whose protein product can be either cytosolic or secreted, and *VAS1* (ref. 17)). We identified 153 additional genes in which at least half of the TIFs with coding potential skipped the first start codon (Fig. 4a and Supplementary Data 7). The translation of these truncated isoforms is supported by recent ribosome profiling data[15] (Fig. 4b and Supplementary Fig. 22), which along with recent proteomics data[18] indicate that N-terminal truncation via alternative start codon usage is a common phenomenon. This phenomenon can be transcriptionally regulated: we detected 9 genes with significantly differential truncation between glucose and galactose ($P < 10^{-3}$, FDR < 0.1, Fig. 4c and Supplementary Table 4). These findings indicate a more common production of truncated 5′ transcripts than previously appreciated, which can lead directly to increased protein diversity even without post-transcriptional regulation.

Our data set also provides evidence for the production of carboxyl-terminal protein truncation via alternative polyadenylation sites. We identified 33 genes enriched for internal polyadenylation that introduces early stop codons into the RNA that are not encoded by the DNA[19] (FDR < 10%, Supplementary Discussion, Supplementary Fig. 23, Supplementary Table 5 and Supplementary Data 8). Among them is *GAL10* ($P < 1.2 \times 10^{-9}$, Fig. 4d), which encodes a bifunctional enzyme in *S. cerevisiae* with two enzymatic domains that are encoded by two separate genes in other organisms[20]. In galactose media, such early stop codons result in additional transcripts encoding proteins with only one of these domains (Fig. 4d). Our evidence of protein truncation via alternative isoform usage represents a plausible means for organisms such as yeast, in which alternative splicing is uncommon, to increase protein diversity by selective domain truncation.

Our genome-wide map of transcript boundaries enabled us to measure the extent of transcriptional compaction on each strand of the genome. Most tandem TIFs are separated by approximately 150 base pairs (bp; Supplementary Fig. 24). However, chained arrangements between adjacent TIFs are common, where the end of one TIF coincides with the start of the TIF for the downstream gene. In fact, of 2,747 tandem ORF pairs in the genome, 27% (743) express overlapping mTIFs (for example, *GIM3–YCK2*, Fig. 1d and Supplementary Data 9) and 6.7% (185) produce bicistronic transcripts (Supplementary Data 10). Most overlapping transcripts stop within the first 100–200 bp of the downstream gene (Supplementary Discussion and Supplementary Fig. 24), indicating that upstream elongating and downstream initiating RNA polymerases are distinguished within this window[21]. This common overlap facilitates crosstalk between transcriptional units, wherein the expression of a gene can depend not only on its own promoter, but also on the expression of its neighbours[22].

Our data set reveals the extent of transcript isoform diversity in the yeast genome at unprecedented resolution. Since most yeast genes have fewer than one mRNA molecule per cell[23], the sheer number of isoforms detected here, even within a single environmental condition, indicates that every cell in a clonal population has a unique transcriptome in terms of RNA abundance, sequence and thus regulatory potential. Such cell-to-cell heterogeneity may confer evolutionary advantages, enabling more rapid adaptation of the species to unforeseen environmental challenges. The variation in transcript isoforms has functional consequences through its impact on post-transcriptional regulatory potential, as well as on protein length and localization. In addition, we discovered hundreds of short coding RNAs whose function can now be investigated. Further applications of the TIF-Seq method, or of alternative paired-end strategies such as RNA-PET[24], to additional environmental conditions, genetic backgrounds and organisms will deepen our understanding of transcriptional complexity.

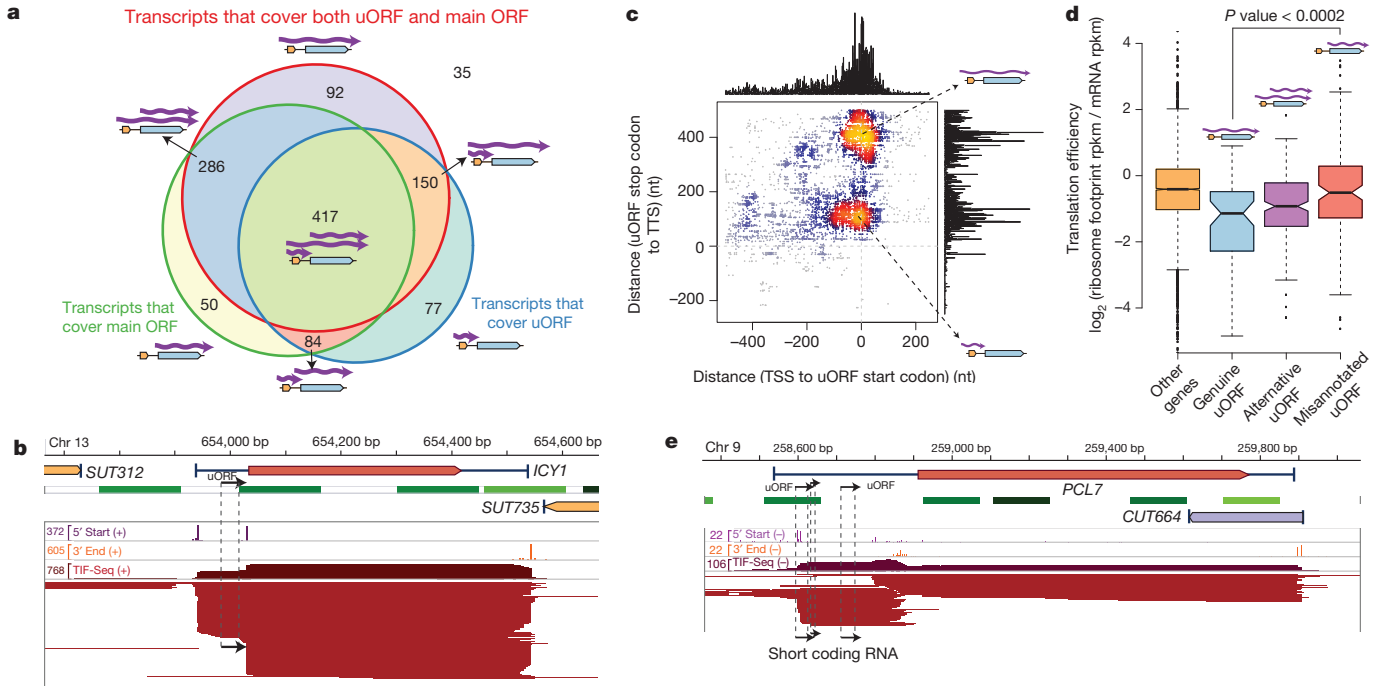

Figure 3 | Transcript isoforms with varying regulatory elements and independent short coding RNAs. **a**, Number of genes whose mTIFs overlap with previously annotated upstream ORFs (uORFs) and their associated (main) ORFs[14]. **b**, *ICY1* transcripts in glucose show alternative presence of uORFs (marked with arrows). **c**, Genome-wide plot of uORF-containing mTIFs: transcript end distance to uORF stop codon (*y* axis) versus transcript start distance to uORF start codon (*x* axis). Small coding RNAs previously misannotated as uORFs represent a separate population of short overlapping RNAs. **d**, Genes with mTIFs that always contain uORFs have lower translation efficiency[15] than those for which the uORF is independently transcribed. Genes with alternative presence of uORFs (for example, *ICY1*) have intermediate translation efficiency. Significance was computed using the Wilcoxon rank-sum test with continuity correction. **e**, Example of an scRNA that was previously misannotated as a uORF in the *PCL7* locus (glucose data shown).

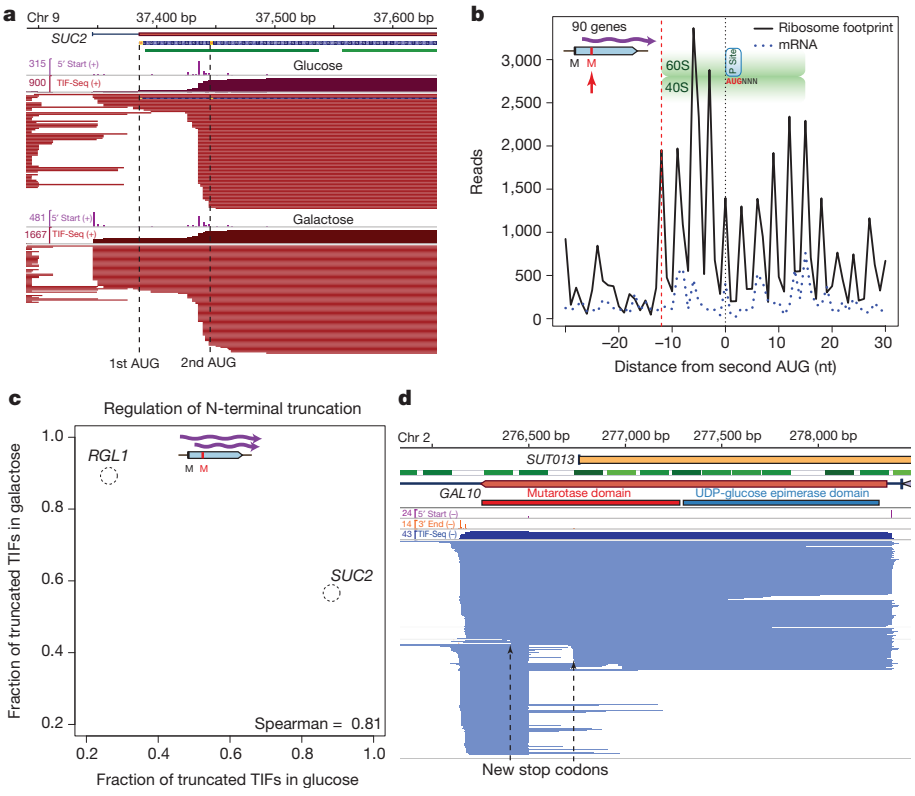

Figure 4 | Alternative transcript isoforms increase coding diversity. **a**, Differential isoform regulation between glucose and galactose produces alternative truncated proteins with differential cellular localization, as shown here for *SUC2* (ref. 2). This regulation is due to subtle variations in TSS selection (5′ start track in purple) that result in alternative inclusion of the first AUG. **b**, Genes producing truncated transcripts that skip the first AUG (80% of these TIFs start between the first and second AUG) are effectively translated and show the expected codon usage pattern and ribosomal protection (green) in ribosome profiling data[15], starting at but not before the second in-frame methionine codon. **c**, Proportion of N-terminal truncated TIFs, (that is, using the second methionine as start codon) in glucose and galactose. **d**, TIFs with internal polyadenylation events that introduce novel stop codons into the RNA encode truncated ORFs and potentially alternative protein isoforms, as shown here for *GAL10*.

In multicellular organisms, the combination of transcript boundary variation and alternative splicing is expected to amplify the diversity of transcript isoforms generated from a single genomic sequence, thereby expanding the functional repertoire of the genome.

METHODS SUMMARY

TIF-Seq library construction was performed as described in Methods and Supplementary Methods. Libraries (Supplementary Table 1) were sequenced using an Illumina HiSeq 2000. Sequencing read analysis was performed using R and Bioconductor (http://www.bioconductor.org/). Only TIFs and mTIFs supported by at least two sequencing reads were used for statistical analysis. Genome sequences (SGD R64) and annotation feature files for S288c were obtained from SGD on 26 March 2011 (http://www.yeastgenome.org/).

Full Methods and any associated references are available in the online version of the paper.

Received 18 December 2012; accepted 26 March 2013.
Published online 24 April 2013.

1. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43,** 853–866 (2011).
2. Carlson, M. & Botstein, D. Two differentially regulated mRNAs with different 5′ ends encode secreted with intracellular forms of yeast invertase. *Cell* **28,** 145–154 (1982).
3. Ungewitter, E. & Scrable, H. Δ40p53 controls the switch from pluripotency to differentiation by regulating IGF signaling in ESCs. *Genes Dev.* **24,** 2408–2419 (2010).
4. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457,** 1033–1037 (2009).
5. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
6. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genet.* **38,** 626–635 (2006).
7. Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143,** 1018–1029 (2010).
8. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5′ SAGE. *Nucleic Acids Res.* **33,** 2838–2851 (2005).
9. Cherry, J. M. *et al. Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700–D705 (2012).
10. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483,** 295–301 (2012).
11. Tan-Wong, S. M. *et al.* Gene loops enhance transcriptional directionality. *Science* **338,** 671–675 (2012).
12. Riordan, D. P., Herschlag, D. & Brown, P. O. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucleic Acids Res.* **39,** 1501–1509 (2011).
13. Hood, H. M., Neafsey, D. E., Galagan, J. & Sachs, M. S. Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi. *Annu. Rev. Microbiol.* **63,** 385–409 (2009).
14. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
15. Gerashchenko, M. V., Lobanov, A. V. & Gladyshev, V. N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl Acad. Sci. USA* **109,** 17394–17399 (2012).
16. Kondo, T. *et al.* Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* **329,** 336–339 (2010).
17. Chatton, B., Walter, P., Ebel, J. P., Lacroute, F. & Fasiolo, F. The yeast VAS1 gene encodes both mitochondrial and cytoplasmic valyl-tRNA synthetases. *J. Biol. Chem.* **263,** 52–57 (1988).
18. Fournier, C. T. *et al.* Amino termini of many yeast proteins map to downstream start codons. *J. Proteome Res.* **11,** 5712–5719 (2012).
19. Yao, P. *et al.* Coding region polyadenylation generates a truncated tRNA synthetase that counters translation repression. *Cell* **149,** 88–100 (2012).
20. Majumdar, S., Ghatak, J., Mukherji, S., Bhattacharjee, H. & Bhaduri, A. UDPgalactose 4-epimerase from *Saccharomyces cerevisiae*. A bifunctional enzyme with aldose 1-epimerase activity. *Eur. J. Biochem.* **271,** 753–759 (2004).
21. Mayer, A. *et al.* CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* **336,** 1723–1725 (2012).
22. Xu, Z. *et al.* Antisense expression increases gene expression variability and locus interdependency. *Mol. Syst. Biol.* **7,** 468 (2011).
23. Miura, F. *et al.* Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics* **9,** 574 (2008).
24. Ruan, X. & Ruan, Y. Genome wide full-length transcript analysis using 5′ and 3′ paired-end-tag next generation sequencing (RNA-PET). *Methods Mol. Biol.* **809,** 535–562 (2012).
25. Wilkening, S. *et al.* An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* **41,** e65 (2013).
26. Venters, B. J. & Pugh, B. F. A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res.* **19,** 360–371 (2009).
27. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. Aiyar for help in editing and refining the manuscript. We thank W. Huber, C. Zhu, A. I. Järvelin, S. Clauder-Münster, J. Zaugg, S. Adjalley, G. Lin and the members of the Steinmetz laboratory for helpful discussions and critical comments on the manuscript. We thank V. N. Gladyshev and C. Pineau for sharing published data. This study was technically supported by the EMBL Genomics Core Facility. This study was financially supported by the National Institutes of Health (to L.M.S.). V.P. was supported by an EMBO fellowship.

Author Contributions W.W., V.P. and L.M.S. conceived the project. V.P. developed the TIF-Seq method and performed experiments. W.W. and V.P. performed the analysis. V.P., W.W. and L.M.S. wrote the manuscript.

Author Information The data reported in this paper have been deposited in GEO under accession number GSE39128 and are also accessible at http://steinmetzlab.embl.de/TIFSeq. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.M.S. (larsms@embl.de).

METHODS

Biological samples. *S. cerevisiae* strain SLS045 (*MATa/α GAL2/GAL2*, S288c background) was grown to mid-log phase ($D_{600\,nm} \approx 1$) using either YPD (1% yeast extract, 2% peptone, 2% glucose) or YPGal (1% yeast extract, 2% peptone, 2% galactose). Total RNA was isolated by the standard hot phenol method and contaminant DNA was removed by DNase I treatment. Sequenced samples are shown in Supplementary Table 1.

TIF-Seq method. For the construction of the TIF-Seq libraries, we used 60 μg of DNA-free total RNA as input. As an internal control, we added capped and polyadenylated *in vitro* transcripts (ATCC 87482, 87483 and 87484). The 5′ end of the non-capped RNA molecules was dephosphorylated by treatment with 6 units of shrimp alkaline phosphatase (SAP, Fermentas) for 30 min at 37 °C in the presence of RNase inhibitor (RNasin +, Promega). The RNA was then purified by a double phenol extraction and ethanol precipitation. CAP was removed by a 1-h incubation at 37 °C with 5 units of tobacco acid pyrophosphatase (TAP, Epicentre) in the presence of RNase inhibitor. The sample was phenol:chloroform-purified and ethanol-precipitated. Finally, for oligonucleotide ligation to the 5′ end of the formerly capped molecules, the treated RNA sample was incubated overnight at 16 °C with 20 units of T4 RNA ligase 1 (NEB) in the presence of 10 mM DNA/RNA '5oligocap' oligonucleotide (Supplementary Table 2), 10% dimethylsulphoxide (DMSO) and RNase inhibitor. The RNA was column-purified (RNeasy, Qiagen) and its integrity was checked (Bioanalyzer, Agilent).

To control for the presence of chimaeras, each ligated RNA sample was divided into two aliquots and independently processed to generate two fractions of full-length cDNA (FlcDNA) with different terminal barcoding (see Supplementary information for details). Specifically, each fraction (11.2 μl) was mixed with 1 μl RT priming oligonucleotide, 1 μM either 3cDNANotI_A or 3cDNANotI_B (Supplementary Table 2), and 1 μl 10mM dNTPs. The sample was incubated at 65 °C for 5 min and transferred to ice. 4 μl of 5× First-Strand buffer (Invitrogen), 2 μl DTT 0.1 M and 0.5 μl RNase inhibitor were added to each sample, which was incubated at 42 °C for 2 min to minimize possible mispriming. 2 μl of Superscript III reverse transcriptase (200 U μl^{-1}, Invitrogen) and high temperature (55 °C) incubation was then used for the retrotranscription to minimize the effects of RNA secondary structure. The reaction was incubated for 20 min at 42 °C, 40 min at 50 °C, and 20 min at 55 °C with a final 15 min enzyme inactivation at 70 °C. RNA removal was performed by adding 0.5 μl RNase cocktail (Ambion) and 2.5 units of RNase H (NEB) to each sample and incubating at 37 °C for 30 min. Samples were purified using Agencourt AMPure XP beads (Beckman Coulter Genomics) according to the manufacturer's instructions and eluted in 19 μl EB (10 mM TrisHCl pH 8.0). 19 μl of resulting FlcDNA samples were PCR-amplified using 20 μl 2× HF Phusion MasterMix (Finnzymes), 0.5 μl biotinylated oligonucleotide 5 μM (either 5BioNotI_A or 5BioNotI_B) and 0.5 μl oligonucleotide 5 μM (either 3AmpNotI_A or 3AmpNotI_B) (Supplementary Table 2). The following thermocycler program was used: 30 s for initial denaturation at 98 °C, 10 cycles (20 s of denaturation at 98 °C, 30 s of annealing at 50 °C (+1 °C per cycle) and 5 min elongation at 72 °C (+10 s per cycle)) and a final elongation of 5 min at 72 °C. Samples were purified using AMPure XP beads. The two independently barcoded aliquots were ultimately pooled together.

To generate cohesive ends, FlcDNA samples were digested for 1 h at 37 °C with 100 units of NotI (NEB) and heat-inactivated for 20 min at 65 °C. The samples were AMPure XP-purified and DNA yield was quantified. Between 300 and 600 ng of digested FlcDNA was circularized for 16 h at 16 °C by intramolecular ligation with 20 μl of T4 DNA ligase (2,000 units μl^{-1}, NEB) in 600 μl final volume. Non-circularized molecules were degraded by incubating the samples for 20 min at 37 °C with 20 units each of exonuclease III (NEB) and exonuclease I (NEB). Enzymes were inactivated by adding 12 μl of 0.5 M EDTA and incubating the samples at 70 °C for 30 min. Circularized FlcDNAs were then phenol:chloroform-purified and ethanol-precipitated.

Purified, circularized FlcDNAs were resuspended in 130 μl EB and sonicated with a Covaris S220 (4 min, 20% duty cycle, intensity 5, 200 cycles per burst). The fragmented DNA was purified with AMPure XP beads and eluted with 20 μl EB. Biotin-containing fragments were captured by incubating the samples for 30 min at room temperature with 20 μl of Streptavidin-conjugated Dynabeads M-280 (Invitrogen) and washed according to the manufacturer's instructions.

Addition of forked barcoded adapters to the captured molecules was performed using the standard Illumina DNA-Seq library generation protocol with some minor modifications. Specifically, purifications using AMPure beads were replaced with separation on magnetic Dynabeads and NEBNext Master Mixes (NEB) were used. A 20-cycle PCR enrichment was performed using Phusion polymerase (Finnzymes). 300-bp libraries were isolated using e-Gel 2% SizeSelect (Invitrogen) and sequenced with a HiSeq 2000 (Illumina) using paired-end sequencing of 105 bp reads.

TIF-Seq method for long mRNA molecules. We used the same method as described above, but introduced an additional size selection step. Specifically, after the initial PCR amplification, the FlcDNA samples were size-selected on a 1.5% agarose gel, and fragments over 2 kb were purified using QIAquick Gel Extraction Kit (Qiagen). The recovered FlcDNA samples were reamplified using 10 cycles of PCR before the NotI digestion.

TIF-Seq method for non-capped mRNA molecules (mono- and triphosphorylated mRNAs). We used the same method as described above, but modified steps before the ssRNA ligation. RNA was dephosphorylated using shrimp alkaline phosphatase as described above, but instead of proceeding to treatment with tobacco acid pyrophosphatase, RNA was rephosphorylated for 1 h at 37 °C using T4 polynucleotide kinase (NEB).

RNA circularization and targeted sequencing. Sixty micrograms of DNA-free RNA samples were SAP-and TAP-treated as described above to obtain full-length RNA molecules with 5′ phosphate ends. RNA circularization was performed in the presence of RNase inhibitor, 10% DMSO, T4 RNA ligase buffer, and 50 units of T4 RNA ligase 1 (ssRNA ligase, NEB) for 16 h at 16 °C. The circularized RNA was purified with RNeasy columns (Qiagen) and subjected to random hexamer retrotranscription. The resulting cDNAs were used as a template for standard PCR amplification with divergent oligonucleotides (Supplementary Table 2). The PCR products were cloned using the TOPO TA cloning system (Invitrogen). Individual clones were bidirectionally sequenced with Sanger sequencing to determine both 5′ and 3′ ends. Only clone sequences spanning a poly(A)-tail were taken into consideration.

Northern blot. The DIG Starter Northern kit (Roche) was used according to the manufacturer's instructions. Strand-specific RNA-DIG probes were generated by *in vitro* transcription.

Sequencing read processing and alignment. Sequencing reads were de-multiplexed and barcode sequences were removed. The presence of internal chimaera control barcodes was assessed by the Needleman–Wunsch global alignment method provided by the R Biostrings package from Bioconductor (http://www.bioconductor.org/). Samples were classified into 4 groups (putative intermolecular (A-A and B-B) or intramolecular events (A-B and B-A)). Only high-confidence reads with both chimaera control barcodes and a poly(A)-tail were considered for further analysis.

Pairs of 5′ sequences and 3′ sequences were trimmed and then separately aligned to the reference genome using Novoalign V2.07.10 (http://www.novocraft.com) using default parameters. The S288c *S. cerevisiae* genome (SGD R64, http://www.yeastgenome.org), along with the sequences of the *in vitro* transcripts that were included as spike-in controls, were used as reference sequences. Only sequences where both ends mapped to the reference were further analysed. Intermolecular pairs (A-B and B-A) were discarded. To exclude any other possible intermolecular cDNA species, only TIFs with both ends mapping to the same chromosome with a length ranging from 40 to 5,000 bp were considered for further analysis.

TIF clustering and mTIF definition. We clustered the transcripts with 5′ and 3′ end sites co-occurring within 5 bp (Supplementary Fig. 2a). Specifically, we defined TSS and TTS clusters separately. Each cluster was defined by both a window and a mTSS/TTS (the most abundant within that window). Clusters of TSS/TTSs were assigned iteratively in decreasing order of expression. In this process, each TSS/TTS site was compared to previously defined clusters, and the site was: (1) defined as a new mTSS/TTS with a 5-bp window (±2 bp up/downstream) if the window did not overlap with previous clusters; (2) defined as a new mTSS/TTS with a smaller window (<5 bp) to avoid overlap with previously defined clusters, if the TSS/TTS was ≥ 5 bp away from the closest previously defined mTSS/TTS but ≤ 2 bp from the closest cluster; and (3) merged with a previously defined cluster if the TSS/TTS was < 5 bp away from the closest mTSS/TTS, in which case the original cluster window was extended to include the newly assigned TSS/TTS (thus the maximum window size of one cluster would be 9 bp); if the TSS/TTS overlapped with 2 previously defined mTSS/TTS in < 5 bp, it was merged with the cluster with the closer (or higher expressed) mTSS/TTS. After this assignment process, only clusters defined by mTSS/TTSs with at least 3 supporting reads were considered. mTIFs were defined as connections between mTSS and mTTS supported by at least 2 reads connecting the associated clusters. All TIFs that shared a given TSS/TTS cluster were assigned to the corresponding mTIF cluster.

TIF annotation. The TIFs were aligned to genome annotation features and classified as: (1) ORFs, if they covered an intact ORF coding region; (2) bicistronic TIFs, if they covered 2 or more ORFs; (3) SUTs, if the common region between the TIF and SUT included more than 80% of the length of both the TIF and the SUT; (4) CUTs/XUTs, same as SUTs; (5) overlapping 2 ORFs, if they overlapped two ORFs but did not entirely cover either; (6) overlapping 5′ of one ORF, if they overlapped only the 5′ of one ORF; (7) overlapping 3′ of one ORF, if they overlapped only the 3′ of one ORF; and (8) intergenic TIF, if

they did not overlap with any annotated ORFs or overlapped with less than 80% of annotated SUTs, CUTs or XUTs. Annotated transcripts (ORF-Ts, SUTs and CUTs) with TSSs and TTSs are from our previous study that used tiling arrays[4].

Comparing TSSs and TTSs to nucleosome data. Nucleosome raw data are derived from ref. 27. Normalized nucleosome occupancy values for the regions flanking the TSSs or TTSs (±500 bp) were extracted and the median values in each position were calculated and plotted. TSSs or TTSs from mTIFs were used.