LETTERS

Bidirectional promoters generate pervasive transcription in yeast

Zhenyu Xu[1]*, Wu Wei[1]*, Julien Gagneur[1], Fabiana Perocchi[1], Sandra Clauder-Münster[1], Jurgi Camblong[2], Elisa Guffanti[3], Françoise Stutz[3], Wolfgang Huber[4] & Lars M. Steinmetz[1]

Genome-wide pervasive transcription has been reported in many eukaryotic organisms[1–7], revealing a highly interleaved transcriptome organization that involves hundreds of previously unknown non-coding RNAs[8]. These recently identified transcripts either exist stably in cells (stable unannotated transcripts, SUTs) or are rapidly degraded by the RNA surveillance pathway (cryptic unstable transcripts, CUTs). One characteristic of pervasive transcription is the extensive overlap of SUTs and CUTs with previously annotated features, which prompts questions regarding how these transcripts are generated, and whether they exert function[9]. Single-gene studies have shown that transcription of SUTs and CUTs can be functional, through mechanisms involving the generated RNAs[10,11] or their generation itself[12–14]. So far, a complete transcriptome architecture including SUTs and CUTs has not been described in any organism. Knowledge about the position and genome-wide arrangement of these transcripts will be instrumental in understanding their function[8,15]. Here we provide a comprehensive analysis of these transcripts in the context of multiple conditions, a mutant of the exosome machinery and different strain backgrounds of *Saccharomyces cerevisiae*. We show that both SUTs and CUTs display distinct patterns of distribution at specific locations. Most of the newly identified transcripts initiate from nucleosome-free regions (NFRs) associated with the promoters of other transcripts (mostly protein-coding genes), or from NFRs at the 3′ ends of protein-coding genes. Likewise, about half of all coding transcripts initiate from NFRs associated with promoters of other transcripts. These data change our view of how a genome is transcribed, indicating that bidirectionality is an inherent feature of promoters. Such an arrangement of divergent and overlapping transcripts may provide a mechanism for local spreading of regulatory signals—that is, coupling the transcriptional regulation of neighbouring genes by means of transcriptional interference or histone modification.

To obtain a comprehensive survey of the structure and expression level of transcripts across the yeast genome, we used tiling arrays[3] to profile wild-type transcriptomes in ethanol (YPE), glucose (YPD, SDC) and galactose (YPGal), which together encompass the main laboratory growth conditions of yeast (Supplementary Tables 1 and 2). Transcript start and end positions were mapped to the genome by a segmentation algorithm[16] and subsequent manual curation. To identify CUTs, profiles were measured for a deletion mutant of *RRP6*, coding for an important component of the nuclear exosome, which is involved in the degradation of CUTs[17,18]. Transcripts specific to the *rrp6*Δ mutant were designated as CUTs (Methods). We assigned systematic names to all SUTs and CUTs. Expression profiles are provided in a searchable web database (http://steinmetzlab.embl.de/NFRsharing).

Altogether, 7,272 transcripts were identified, comprising 5,171 verified or uncharacterized open reading frame transcripts (ORF-Ts), 847 SUTs and 925 CUTs (Fig. 1 and Supplementary Table 3). We took advantage of data from different conditions to disambiguate cases of overlapping or immediately adjacent transcripts (Methods). We only used transcripts with confidently mapped 5′ ends for analyses involving start sites (5,084 ORF-Ts, 823 SUTs and 704 CUTs; Methods and Supplementary Table 4). For validation, we compared our data to transcript start sites (TSSs) mapped by 5′ RACE (rapid amplification of complementary DNA ends)[19]. Eighty-one per cent (1,039 out of 1,281) of TSSs agreed within 50 bases with the 5′ RACE results (Supplementary Fig. 1), 3% higher than a recent Solexa sequencing approach[19]. Furthermore, a comparison of our 3′ ends with the Solexa data set showed agreement of 61% (2,774 out of 4,551) within 50 bases. In addition, we tested several CUT boundaries and they agreed well with our real-time polymerase chain reaction and 5′ RACE validations (Supplementary Fig. 2 and Supplementary Table 5). Altogether, 102 SUTs had a higher expression level in the *rrp6*Δ mutant compared to wild type (Supplementary Table 6), indicating that the distinction between CUTs and SUTs is in some cases condition-dependent, as, for example, the CUT on the antisense strand of *PHO84*, which is stabilized in old cells[11]. CUTs were, overall, shorter (median length 440 bases) than SUTs (median length 761 bases; $P < 2 \times 10^{-16}$, Wilcoxon test).

Nucleosome-free promoter regions (or 5′ NFRs), which facilitate transcription by allowing RNA polymerase to bind to DNA, have been reported as hallmarks of gene promoters[20–24]. To test whether unannotated transcripts have such hallmarks, we compared our transcript positions with nucleosome maps[22,25]. Consistent with promoter activity at NFRs, all classes of transcripts—ORF-Ts, CUTs and SUTs—showed depletion of nucleosomes upstream of their TSS (Fig. 2a). Furthermore, no nucleosome was detected between 422 out of the 666 (63%) non-overlapping divergent transcript pairs involving at least one unannotated transcript (Methods and Supplementary Table 7). This indicates that these pairs share a single 5′ NFR that may function as a bidirectional promoter.

To investigate further the set of potential bidirectional promoters in the yeast genome, we analysed all 1,049 non-overlapping divergent transcript pairs that shared a single 5′ NFR. The distribution of distances between their TSSs had an estimated mode (defined as the point with the highest density of the distribution) at 180 bases, whereas their shared NFR lengths had a mode at 131 bases (Fig. 2b). The size of the shared 5′ NFRs increased with the inter-transcript distances, in a relationship consistent with a model of a single NFR surrounded by two regions inside the flanking nucleosomes from which transcripts initiate[22,25].

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK. [3]Department of Cell Biology, Sciences III, University of Geneva, 30 Quai E. Ansermet, 1211 Geneva 4, Switzerland. [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK.
*These authors contributed equally to this work.

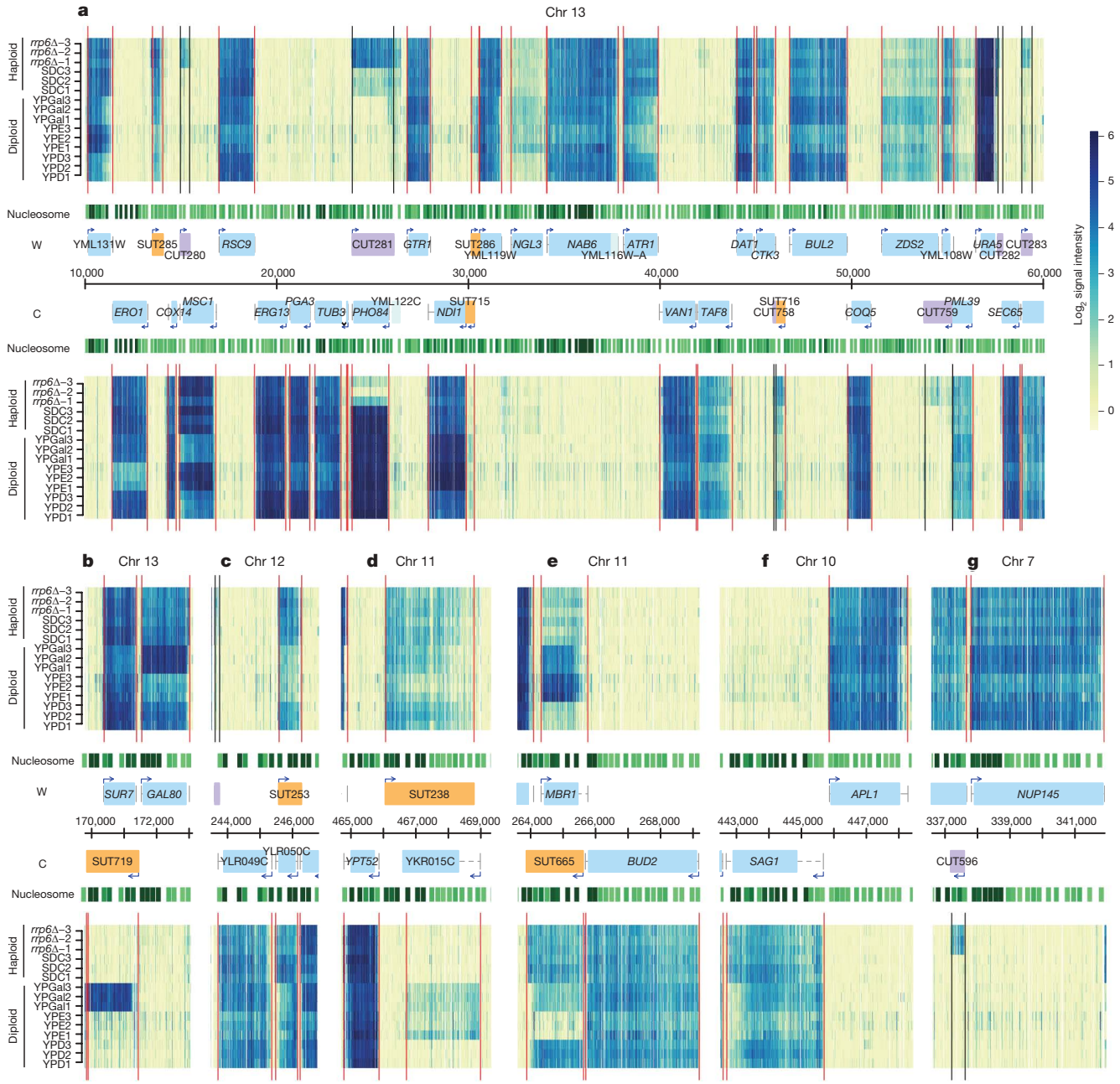

Figure 1 | Transcript maps. **a**, Expression data along 50 kb of chromosome 13 (*x* axis) for the Watson (W, top) and the Crick (C, bottom) strands. (Larger views are available for the whole genome in our searchable web database.) Normalized signal intensities are shown for the profiled samples (*y* axis): three replicates each for the *rrp6*Δ S96 haploid strain (*rrp6*Δ-1–3), the S96 haploid strain in SDC (SDC1–3), the S1003 diploid strain in YPGal (YPGal1–3) and the S1003 diploid strain in YPE (YPE1–3), and three rows (summarizing nine replicates) for the S1003 diploid strain in YPD (YPD1–3). Vertical lines represent inferred transcript boundaries. Nucleosome positions (green tracks, darker for more significant scores[22]) and genome annotations are shown in the centre: annotated ORFs (blue boxes) and their mapped UTRs (dashed grey lines), SUTs (orange boxes), CUTs (purple boxes) and transcript start sites (arrows). Coordinates are indicated in base pairs. **b**–**g**, Examples of transcriptional arrangements; layout is as in **a**. **b**, Tandem gene pair with antisense transcript: *GAL80* shares a NFR with our detected SUT719, an antisense transcript of *SUR7*; **c**, antisense transcript SUT253 originating from both a 5′ NFR (of YLR049C) and a 3′ NFR (of YLR050C); **d**, antisense transcript SUT238 originating from a 5′ NFR (of *YPT52*); **e**, SUT665 originating from a 3′ NFR (of *BUD2* and *MBR1*); **f**, divergent promoter of two ORF-Ts with putatively long UTRs; **g**, CUT596 originating from a 5′ NFR (of *NUP145*).

In our analysis, 612 out of 931 non-overlapping divergent protein-coding transcript pairs were found to share a single 5′ NFR (66%, Supplementary Table 7). This fraction is considerably higher than the 30% of divergent ORF pairs that were previously estimated to share promoters[26]. Previous studies may have underestimated the number of bidirectional promoters by considering only distances between ORF start codons. Indeed, for divergent ORF-T pairs sharing a 5′ NFR (Fig. 2c, red dots), the total untranslated region (UTR) length increased with the distance between the start codons, consistent with a typical size of the inter-transcript distance of a shared promoter being ~180 bases, as evident from Fig. 2b. This relationship holds for a wide range of inter-ORF distances, including cases greater than 1,000 bases, such as *SAG1* and *APL1* (also known as *YAP80*) (Fig. 1f). In contrast, divergent ORF-T pairs separated by multiple NFRs showed no correlation between total UTR length and distance separating start codons (Fig. 2c, black dots). Moreover, most of these pairs were separated by more than 452 bases, which is approximately the minimal size of a region spanned by two NFRs (2 × 131 bases), a nucleosome (146 bases) and two intra-nucleosome regions (2 × 22 bases; Supplementary Fig. 3). These results indicate that bidirectional promoter

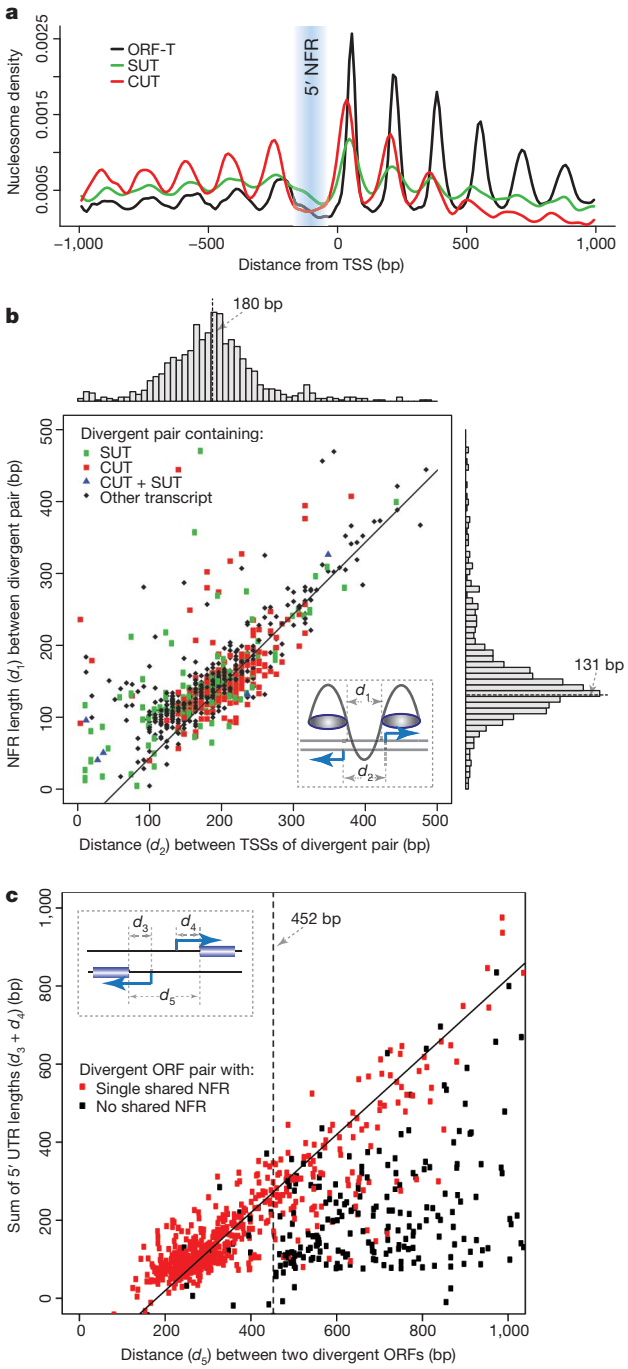

Figure 2 | Properties of divergent transcript pairs. **a**, Nucleosome density[22] relative to TSSs, averaged over ORF-Ts (black line), SUTs (green line) and CUTs (red line). **b**, Scatter plot and histograms of shared NFR length (d_1) and distance between TSSs (d_2) of divergent pairs sharing a 5′ NFR. The line corresponds to the regression $d_1 = d_2 - 2c$, where the value $c = 22$ bases was determined from the mode of the distribution of differences between d_1 and d_2, and corresponds to a typical distance between the NFR and TSS. **c**, Scatter plot of the sum of 5′ UTR lengths ($d_3 + d_4$) versus the distance (d_5) between coding sequences of divergent ORF-T pairs. The solid line corresponds to the regression $d_5 = d_3 + d_4 + b$, where the value $b = 180$ bases for the typical TSS distance between divergent pairs is taken from **b** above. The vertical dashed line at $d_5 = 452$ bases is an estimate of the minimal distance for two ORFs to have separate NFRs.

usage is frequent for divergent transcript pairs involving unannotated transcripts and protein-coding genes in any combination.

To determine how many of the 5′ NFRs initiate transcripts bidirectionally, we selected all nucleosome-depleted regions longer than 80 bases immediately upstream of TSSs, defining a set of 3,965 5′ NFRs (Methods and Supplementary Fig. 4). Of these, 1,318 (33%) were bidirectional, involving half of all transcripts with a mapped 5′ NFR (2,656 of 5,339, Supplementary Tables 8–10). The sequences of NFRs detected as bidirectional promoters did not differ significantly from the other 5′ NFRs in content of palindromic sequences or GC nucleotides. Among transcripts with mapped 5′ NFRs, 61% of unannotated transcripts and 48% of protein-coding transcripts initiated bidirectionally from shared 5′ NFRs rather than initiating from their own promoters (Fig. 3b). Of the unannotated transcripts, 90% shared the 5′ NFR with a protein-coding transcript. These results indicate that bidirectionality is an inherent property of promoters. In addition to bidirectional transcription, a small number of transcripts was found to initiate in tandem orientation from shared 5′ NFRs (Fig. 3b). This number is probably underestimated, however, because of the difficulty of distinguishing immediately adjacent tandem transcripts by microarray hybridization. Altogether, our results indicate that multiple transcripts often initiate from NFRs at promoters in yeast. Additional transcripts will probably be detected by profiling alternative conditions or mutants other than *rrp6*Δ.

In addition to NFRs at promoters, NFRs downstream of stop codons have been reported for most ORFs and are suspected to have a role in transcription termination as well as in the generation of transcripts antisense to the ORF[22] (Fig. 3a). To characterize better such NFRs, we selected all nucleosome-depleted regions longer than 80 bases immediately downstream of stop codons of all verified and uncharacterized ORFs that we detected expressed, defining a set of 2,616 3′ NFRs (Supplementary Table 9). Of these, 827 (32%) initiated a transcript. We observed that 27% of unannotated transcripts with a mapped 5′ NFR initiated from the 3′ NFR of an ORF (Fig. 3b). Together, 3′ and 5′ shared NFRs thus accounted for most (73%) SUT or CUT initiation, and for most (61%) ORF-T initiation (Fig. 3b, Supplementary Tables 10 and 11 for a list of all pairs). Altogether, these results show a surprisingly high level of NFR sharing, not only in bidirectional promoters but also in 3′ NFRs.

The high level of NFR sharing may explain a large extent of antisense transcription[3], that is, transcription on opposite strands. Seventy per cent of all antisense transcripts with mapped 5′ NFRs initiated from a shared NFR. For example, 269 unannotated transcripts initiating from the 3′ NFR of an ORF were transcribed antisense to the ORF (for example, YLR050C and *MBR1*, Fig. 1c, e). Another recurrent configuration is an antisense transcript starting from the 5′ NFR of a downstream tandem transcript. These configurations associate three transcripts; an example is *GAL80*, the 5′ NFR of which initiates a transcript antisense to its upstream gene *SUR7* (Fig. 1b). Notably, the level of *SUR7* was lowest in YPGal medium, in which the *SUR7* antisense transcript and *GAL80* had the highest expression (18 further examples are given in Supplementary Table 12). To generalize these observations, we analysed expression correlations across growth conditions among transcript pairs involving at least one SUT. We observed significant expression anti-correlation between sense–antisense pairs, whereas bidirectional pairs of transcripts showed a tendency for co-expression (Supplementary Fig. 5 and Supplementary Table 13; all P values $< 10^{-7}$, Pearson's product moment correlation test). The anti-correlation between sense and antisense transcripts fits the pattern displayed by individual cases of transcriptional interference or inhibitory histone modifications[10–14,27].

The extent to which the genome-wide set of unannotated transcripts have a biological role, or are merely transcriptional side products (noise) originating from nucleosome-depleted regions[9], is unknown. The action of transcription itself can be functional even if the transcription product is not. This is the case, for example, with the transcription of the non-coding RNA *SRG1* and the *IME4* antisense transcript, which mediate transcriptional silencing[13,14]. To explore the conservation of transcription initiation from 5′ and 3′ NFRs, we profiled the transcriptome of YJM789 (ref. 28), a highly diverged relative of the laboratory strain S288c. In rich media (YPD),

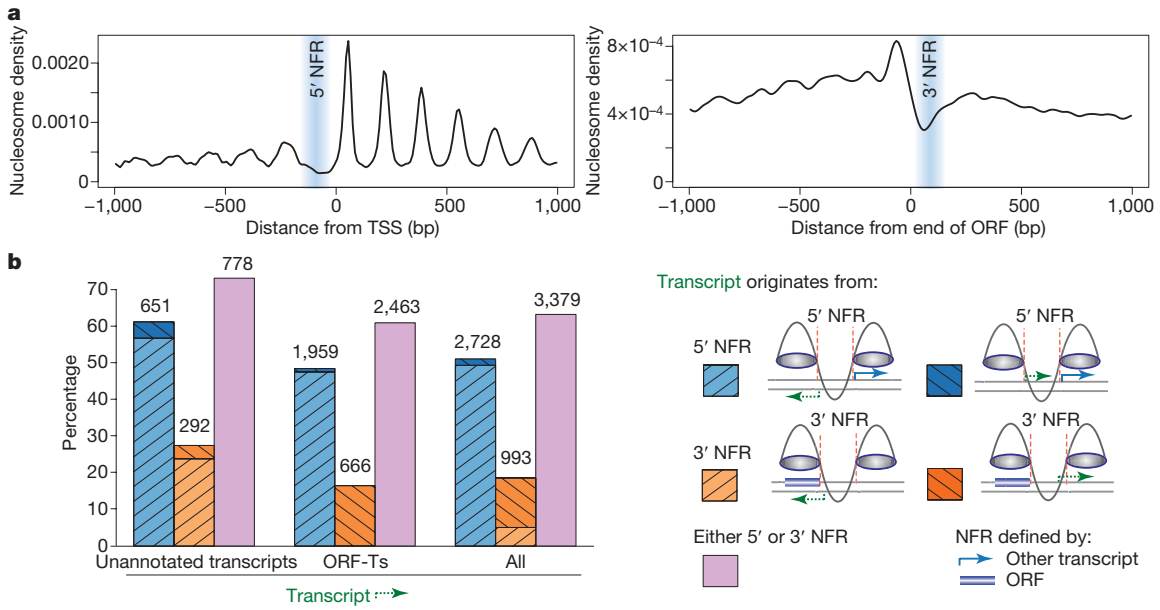

Figure 3 | 5′ and 3′ NFR sharing. a, Nucleosome density relative to TSSs, averaged over all transcripts (left panel), and relative to translation stop sites, averaged over all ORF-Ts (right panel). **b**, Transcripts initiating from 5′ or 3′ NFRs of other transcripts. The first block of bars corresponds to unannotated transcripts (1,063), the second to ORF-Ts (4,039) and the third to all transcripts (5,339) with mapped 5′ NFRs. Within each block, the bars correspond to different orientations of the transcript relative to the 5′ or 3′ NFR it originates from: divergently from a 5′ NFR (light blue), in tandem from a 5′ NFR (dark blue), in antisense to an ORF from a 3′ NFR (light orange), in tandem to an ORF from a 3′ NFR (dark orange), or in any orientation from a 5′ or 3′ NFR (pink). See Supplementary Table 11 for a list of these pairs.

about 50% (380 out of 769) of the SUTs expressed in S288c were also found expressed in YJM789 (Methods). The frequencies with which these 380 conserved SUTs were observed sharing NFRs with other transcripts were similar to those in the overall data set. These results indicate that the interlaced architecture of transcript initiation from 5′ and 3′ NFRs is conserved between these strains of *Saccharomyces cerevisiae*. Why some of the unannotated transcripts are stable and others are unstable remains to be explored. The parasite *Giardia lamblia* produces an abundance of antisense transcripts originating bidirectionally from promoters[29]; consistent with our *rrp6*Δ results, its genome lacks orthologues to several nuclear exosome components. Likewise, the function of bidirectional transcription requires further exploration. One hypothesis is that bidirectional transcription has a role in maintaining an open chromatin structure at promoters. In other instances the combined action of bidirectional promoters and transcriptional regulation by these transcripts, or their generation, may provide a mechanism to spread transcriptional regulatory signals locally in the genome.

METHODS SUMMARY

Complementary DNA for hybridization was prepared using random- or random-plus-oligo-dT-priming with the addition of actinomycin D during reverse transcription[30]. The hybridization data were normalized and segmented using the Bioconductor package 'tilingArray'[16]. Segments were then manually curated. Further details can be found in Methods and Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 September; accepted 19 December 2008.
Published online 25 January 2009.

1. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306,** 2242–2246 (2004).
2. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
3. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103,** 5320–5325 (2006).
4. Dutrow, N. *et al.* Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA–DNA hybrid mapping. *Nature Genet.* **40,** 977–986 (2008).
5. Li, L. *et al.* Genome-wide transcription analyses in rice using tiling microarrays. *Nature Genet.* **38,** 124–129 (2006).
6. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306,** 655–660 (2004).
7. Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453,** 1239–1243 (2008).
8. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8,** 413–423 (2007).
9. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14,** 103–105 (2007).
10. Berretta, J., Pinskaya, M. & Morillon, A. A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in *S. cerevisiae*. *Genes Dev.* **22,** 615–626 (2008).
11. Camblong, J. *et al.* Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in *S. cerevisiae*. *Cell* **131,** 706–717 (2007).
12. Bird, A. J., Gordon, M., Eide, D. J. & Winge, D. R. Repression of ADH1 and ADH3 during zinc deficiency by Zap1-induced intergenic RNA transcripts. *EMBO J.* **25,** 5726–5734 (2006).
13. Hongay, C. F., Grisafi, P. L., Galitski, T. & Fink, G. R. Antisense transcription controls cell fate in *Saccharomyces cerevisiae*. *Cell* **127,** 735–745 (2006).
14. Martens, J. A., Laprade, L. & Winston, F. Intergenic transcription is required to repress the *Saccharomyces cerevisiae SER3* gene. *Nature* **429,** 571–574 (2004).
15. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
16. Huber, W., Toedling, J. & Steinmetz, L. M. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22,** 1963–1970 (2006).
17. Davis, C. A. & Ares, M. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **103,** 3262–3267 (2006).
18. Wyers, F. *et al.* Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121,** 725–737 (2005).
19. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
20. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genet.* **39,** 1235–1244 (2007).
21. Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **6,** e65 (2008).
22. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18,** 1073–1083 (2008).
23. Whitehouse, I., Rando, O. J., Delrow, J. & Tsukiyama, T. Chromatin remodelling at promoters suppresses antisense transcription. *Nature* **450,** 1031–1035 (2007).
24. Yuan, G. C. *et al.* Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309,** 626–630 (2005).
25. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* **446,** 572–576 (2007).
26. Hermsen, R., ten Wolde, P. R. & Teichmann, S. Chance and necessity in chromosomal gene distributions. *Trends Genet.* **24,** 216–219 (2008).

27. Uhler, J. P., Hertel, C. & Svejstrup, J. Q. A role for noncoding transcription in activation of the yeast *PHO5* gene. *Proc. Natl Acad. Sci. USA* **104,** 8011–8016 (2007).
28. Wei, W. *et al.* Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc. Natl Acad. Sci. USA* **104,** 12825–12830 (2007).
29. Teodorovic, S., Walls, C. D. & Elmendorf, H. G. Bidirectional transcription is an inherent feature of *Giardia lamblia* promoters and contributes to an abundance of sterile antisense transcripts throughout the genome. *Nucleic Acids Res.* **35,** 2544–2553 (2007).
30. Perocchi, F., Xu, Z., Clauder-Münster, S. & Steinmetz, L. M. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.* **35,** e128 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Akhtar, A. Ladurner, S. Blandin, R. Aiyar, E. Mancera and E. Fritsch for comments on the manuscript, J. Toedling for discussion and for the template of the website, C. Girardot for data submission to ArrayExpress, N. Proudfoot for access to experimental equipment, and the contributors to the Bioconductor (www.bioconductor.org) and R (http://www.r-project.org) projects for their software. This work was supported by grants to L.M.S. from the National Institutes of Health and Deutsche Forschungsgemeinschaft, by a SystemsX fellowship to E.G., by a Roche fellowship to J.C. and by grants to F.S. from SNF and NCCR Frontiers in Genetics.

Author Contributions L.M.S., Z.X. and W.W. designed the research; Z.X. and W.W. annotated the transcripts with the help of J.G. and F.P.; W.W. and Z.X. performed analysis of the transcripts with the help of J.G.; F.P. and S.C.-M. performed the array hybridizations; J.C. E.G. and F.S. provided samples for the *rrp6Δ* mutant, and designed and performed validation polymerase chain reaction with reverse transcription and 5′ RACE experiments; L.M.S., J.G., F.S. and W.H. supervised the research; and L.M.S., Z.X., W.W., J.G. and W.H. wrote the manuscript.

Author Information Raw data are available from ArrayExpress (http://www.ebi.ac.uk/arrayexpress) under accession number E-TABM-590. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.M.S. (larsms@embl.de).

METHODS

Strains and media. *S. cerevisiae* strains used in this work were isogenic to either S288c or YJM789 (Supplementary Table 1). Strains were grown to mid-exponential phase (D_{600} ~1.0) in YPD (2% peptone, 1% yeast extract, 2% dextrose), YPGal (2% peptone, 1% yeast extract, 2% galactose), YPE (2% peptone, 1% yeast extract, 2% ethanol) or synthetic complete (SDC) medium (0.67% yeast nitrogen base without amino acids, with ammonium sulphate, 2% dextrose and amino acid supplements) (Supplementary Table 2).

RNA extraction and hybridization to arrays. Total RNA was extracted from yeast cultures with standard hot phenol protocol and processed for array hybridizations as described previously[30] (Supplementary Methods). Importantly, to remove reverse transcription artefacts, first-strand cDNA was synthesized in the presence of 6.25 μg ml^{-1} actinomycin D. The labelled cDNA samples were denatured and processed for hybridizations[30]. Our analysis is based on replicate hybridizations (Supplementary Table 2).

Genome sequence and annotation. Sequence and feature files (.gff files) for S288c were obtained from the *Saccharomyces* Genome Database on 4th September 2007.

Array data analysis. Arrays profiled in conditions YPD, YPE and YPGal were normalized with genomic DNA as in ref. 16. Only the probes matching exactly and uniquely to the S288c genome were considered further. The normalized data were jointly segmented using a segmentation algorithm[16] and the automatically identified segments were curated using a custom web-interface (Supplementary Information). This defined the set of manually curated transcripts.

To identify CUTs, arrays for the *rrp6*Δ strain were segmented jointly with the arrays of the wild-type strain in the same condition (SDC). YJM789 arrays were normalized with YJM789 genomic DNA as a reference. Only the probes matching exactly and uniquely to the S288c-aligned part of the YJM789 sequence were considered further. The normalized data were segmented based on the alignment between S288c and YJM789 (ref. 28).

Transcript categorization. The manually curated transcripts were overlapped with the genome annotation features and classified as: (1) SUTs, if they did not overlap with existing annotation; (2) ORF-Ts, if they overlapped with a verified or uncharacterized ORF; or (3) other. Transcripts detected solely in *rrp6*Δ were defined as (4) CUTs (see next section). We refer to the union of SUTs and CUTs also as unannotated transcripts. (5) Antisense transcripts were defined as unannotated transcripts that overlapped with other transcripts on the opposite strand.

Definition of CUTs. The automatically detected segments for the *rrp6*Δ strain were overlapped with the manually curated transcripts. We defined three criteria: to not overlap any annotated feature; to show higher than twofold expression in *rrp6*Δ compared to wild type; and to be at least 100 bases long. Two types of CUTs were defined. CUTs of the first type were *rrp6*Δ segments that did not overlap any manually curated segments and fulfilled all three criteria. CUTs of the second type were derived from the *rrp6*Δ segments overlapping manually curated transcripts in either a one-to-one or a many-to-one relationship. The *rrp6*Δ-specific (non-overlapping) parts of these segments were classified as CUTs if they fulfilled all criteria.

Classification of transcript ends. Ends of transcripts can be unambiguously detected from the microarray signal when they are not adjacent to another transcript with a higher signal. We classified all transcript ends as being mapped or unmapped. Adjacent transcript ends on a same strand and separated by a distance shorter than 100 bases were investigated as potentially unmapped ends. In such configurations, the 5′ end of the downstream transcript was classified as unmapped if all the following criteria were fulfilled: the signal in the intergenic region between the two adjacent transcripts was above background in all conditions; the expression difference between the intergenic region and the downstream transcript was less than twofold in all conditions; and the expression of the downstream transcript was lower than the expression of the upstream transcript signal by twofold in all conditions. Indeed, if any of these three criteria was violated, we considered this as evidence for a transcript starting from this boundary, and considered the 5′ end mapped. An analogous definition was applied for the 3′ end of the upstream transcript.

Categorization of adjacent transcript pairs. To detect adjacent transcript pairs, transcripts were sorted according to the minimum of their start and end positions. Two consecutive transcripts were considered as adjacent pairs. The adjacent pairs were further classified as divergent if the first transcript was on the Crick strand and the second on the Watson strand, as convergent if the reverse was true, and as tandem if both transcripts were on the same strand. To estimate the mode of a (distance) distribution, we used the midpoint of the shorth (the shortest interval that covers half the values).

Nucleosome data analysis. The transcripts were compared to the nucleosome map combining the H2A.Z and H3/H4 data from http://atlas.bx.psu.edu (refs 22, 25). Two transcripts were considered as sharing a 5′ NFR if there was no nucleosome peak between their TSSs. The 5′ NFR was defined as the nucleosome-depleted region (at least 80 bases long, see below) immediately upstream of the TSS, and the 3′ NFR as the nucleosome-depleted region (at least 80 bases long) downstream of the stop codon of all verified or uncharacterized ORFs. The cutoff value of 80 bases was chosen on the basis of the nucleosome distance distribution. The nucleosome distance distribution showed two modes: one presumably corresponding to the normal nucleosome linker region (18 bases) and a second mode at around 130 bases corresponding to the NFRs (Supplementary Fig. 4).

YJM789 comparison. The SGD annotation was first converted into an alignment coordinate system between S288c and YJM789 (ref. 28). The YJM789 transcripts were categorized in the same manner as the manually verified transcripts from S288c-derived strains. S288c SUTs were also mapped into alignment coordinates and overlapped with the unannotated transcripts from YJM789. A transcript was considered expressed in both S288c-derived and YJM789 genomes if the overlap was at least 50% of the transcript lengths measured in the S288c genome.