b, Scatterplot 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 NFR and TSS.

c, Scatterplot of the sum of 5' UTR lengths ($d_3 + d_4$) vs. 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 panel b above. The vertical dotted line at d_5 = 452 bases is an estimate of the minimal distance for two ORFs to have separate NFRs.

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), in any orientation from a 5' or 3' NFR (pink). See Supplementary Table 11 for a list of these pairs.

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 (OD_{600} ~ 1.0) in either 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 w/o amino acids with ammonium sulfate, 2% dextrose, 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 previously described[30] (Supplementary Methods). Importantly, to remove reverse transcription artifacts, first-strand cDNA was synthesized in the presence of 6.25μg/ml 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 reference[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[28].

Transcript categorization. The manually curated transcripts were overlapped with the genome annotation features and classified as (1) *SUT*, if they did not overlap with existing annotation, (2) *ORF-T*, if they overlapped with a verified or uncharacterized ORF; (3) *other*, otherwise. Transcripts detected solely in *rrp6*Δ were defined as (4) *CUT* (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: a) to not overlap any annotated feature; b) to show higher than 2 fold expression in *rrp6*Δ compared to wild type; c) 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 criteria a-c. 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 relation. The *rrp6*Δ-specific (non-overlapping) parts of these segments were classified as CUT if they fulfilled criteria a-c.

Classification of transcript ends. Ends of transcripts can be unambiguously detected from microarray signal when they are not adjacent to another transcript with higher signal. We classified all transcript ends as being mapped or unmapped. Only 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: 1. the signal in the intergenic region between the two adjacent transcripts was above background in all conditions; 2. the expression difference between the intergenic region and the downstream transcript was less than two-fold in all conditions; 3. the expression of the downstream transcript was lower than the expression of the upstream transcript signal by two-fold in all conditions. Indeed, if any of these 3 criteria was violated, we considered this as an 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[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 cut-off value of 80 bases was chosen based on 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[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.