Science
AAAS

Supplementary Materials for

Transcriptional neighborhoods regulate transcript isoform lengths and expression levels

Aaron N. Brooks[1†‡], Amanda L. Hughes[1†], Sandra Clauder-Münster[1], Leslie A. Mitchell[2§], Jef D. Boeke[2,3], Lars M. Steinmetz[1,4,5 *]

Correspondence to: lars.steinmetz@stanford.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S12
Tables S5 to S7

Other Supplementary Materials for this manuscript include the following:

Table S1 to S4

Materials and Methods

Strains

SynIXR SCRaMbLE strains were previously generated and sequenced (*15,16*). JS94 is the parental strain containing synIXR prior to SCRaMbLE (-SCRaMbLE). Haploid BY4741 and diploid SLS045-A derivatives of S288C, in which the synIXR strains were generated, were used as wildtype controls. Galactose-inducible, transcription factor overexpression strains developed by the Andrews lab were purchased from Dharmacon Horizon Discovery (YSC4515-202484524, YSC4515-202484850, YSC4515-202485149, YSC4515-202487711, YSC4515-202485944) (*21*). Overexpression of these transcription factors is sufficient to induce expression of their target genes without the need for specific growth conditions (*24*).

Construction of the tetOFF strains was performed via homologous recombination, integrating the transactivator into the TRP1 locus and then the tetO7 promoter upstream (-913 to -1121) of the *YIR018C-A* coding region. The transactivator domain was amplified from the pUG-tTA plasmid with Trp1-tailed primers (Table S5) (tTA.trp1 and tTA.trp2); the tetO7 promoter was amplified from the pCM325 plasmid with primers containing homology to the region upstream of *YIR018C-A* (YIR018C-AtetFe and YIR018C-AtetRc) (Table S5). As both tTA and tetO7 integration use KanMX as a selection marker, Cre-induced recombination at loxP sites flanking the KanMX cassette in the tTA insertion was used to recycle the marker prior to integration of the tetO7 promoter (*25*). The Cre recombinase was expressed from the GAL promoter of the pNatCre plasmid during 6 hours of growth in YP-2%Gal media, after which cells were grown on YP-2%fructose/0.025%glucose plates and replica plated to YPAD+Geneticin plates to select for colonies which had lost the KanMX cassette (*26*). Loss of the pNatCre plasmid was confirmed by lack of growth on YPAD+CloNat plates. Yeast were transformed with 0.5-2 µg PCR product or pNatCre plasmid via high-efficiency transformation with lithium acetate and polyethylene glycol.

Growth rate

Strains were grown in YPAD in 96-well deep well plates. Cultures were diluted to approximately OD_{600} of 0.05 in 100 µL YPAD and grown in a Synergy HTX multi-mode microplate reader (BioTek) at 30 °C with continuous, 3mm linear shaking. Absorbance measurements (600 nm) were taken every 15 min for up to 36 h. Growth curves from eight plates were analyzed using the growthcurver R package (*27*). The mean doubling times ± standard deviation are reported.

Strain cultivation and RNA isolation

For direct RNA sequencing, yeast cultures (50 mL) were grown at 30 °C with 180 rpm shaking in a CERTOMAT® BS-1 incubator (Sartorius Stedim Biotech Gmbh) to mid-log phase (OD_{600} 0.7-0.95). Cell pellets were collected via centrifugation, flash frozen in liquid nitrogen, and stored at -80 °C. Doxycycline induction was performed over 24 hours on wildtype (BY4741), transactivator-expressing (tTA), and tetOFF strains. YPAD with and without 10 µg/mL doxycycline was inoculated with culture to approximately 0.15 OD_{600}, and 25 mL was removed, pelleted, and flash frozen at 2, 4, and 8 hour timepoints. After 8 hours of growth, the culture was diluted 1:1000 into fresh 25 mL YPAD (+/- doxycycline) and grown for an additional 16 hours, pelleted, and flash frozen.

Transcription factor overexpression was induced by 3 h growth in SC-URA media containing 2% galactose as the sugar source after growth in SC-URA containing 2% raffinose. Saturated cultures grown overnight in raffinose media were diluted to approximately 0.2 OD_{600} in 50mL raffinose media, these cultures were diluted in 100mL raffinose media to maintain log phase growth overnight, after which cultures were split equally for growth in 50mL raffinose media (control) or 50mL galactose media (overexpression). Cells were collected after 3 h and flash frozen. Gcn4 and Msn2 overexpressing strains were also collected after 30 min and 1 h of galactose-induced overexpression.

RNA was isolated with the MasterPure™ Yeast RNA Purification Kit (Lucigen), including DNase I treatment to remove genomic contamination. mRNA was enriched using Dynabeads™ Oligo (dT)25 beads (Thermo Fisher). The quality of the total and polyA RNA was determined on an Agilent Bioanalyzer using the Agilent RNA 6000 Nano kit, and Qubit™ HS RNA assays were used to quantitate the RNA.

Oxford Nanopore direct RNA sequencing

The direct RNA sequencing protocol (SQK-RNA002) was performed on 500 ng polyA mRNA. Briefly, the RTA adapter was ligated to the 3′-end of polyA RNA by 3000 U T4 DNA Ligase (NEB) in a 15 µL reaction volume of 1x NEBNext Quick Ligation buffer, including the RCS standard. After a 10 min incubation at room temperature,

5x First Strand buffer and DTT were added to 1x and 10 mM concentrations, respectively, in a 40 µL reaction volume and the ligated RNA was reverse transcribed from the adapter with SuperScript® III reverse transcriptase (NEB) for 50 min at 50 °C. The reverse transcription was heat inactivated at 70 °C for 10 min and cooled to 4 °C before the RNA:cDNA was purified with 1.8 volumes Agencourt RNAclean XP beads (Beckman). The 20 µL eluate was then ligated to the RMX adapter in 1x NEBNext Quick Ligase buffer in a 40 µL reaction with 6000 U T4 DNA Ligase at room temperature for 10 min. The ligation reaction was purified with 1 volume Agencourt RNAclean XP beads using WSB buffer (SQK-RNA002) in the wash steps and 21 µL EB (SQK-RNA002) for the elution. Qubit HS dsDNA assay was used to quantitate 1 µL of the eluate, and the remainder was mixed with 17.5 µL water and 37.5 µL RRB (SQK-RNA002) before loading on primed (EXP-FLP001) MinION flow cell (FLO-MIN106D R9) and run on GridION with MinKNOW 3.1.8. On average, fewer than 2 million reads were sequenced per strain.

Illumina short-read RNA sequencing

Short-read, stranded mRNA sequencing was used to quantitate transcripts. Total RNA quality and concentration were assessed on a Fragment Analyzer (Agilent). RNA was diluted to 100 ng/µL, containing 5 µL of 1:500 diluted ERCC Spike-In mix (Thermo Fisher), in a 50 µL volume. RNAs from 67 samples (64 SCRaMbLE strains, the JS94 -SCRaMbLE control, and wildtype BY4741 and SLS045-A strains) were prepared in parallel with robotic handling using the stranded, NEBNext Ultra II RNA library preparation method with i7 barcoding and pooled for one lane of paired end, 75bp sequencing on the NextSeq 500. Library preparation included RNA fragmentation and random priming to generate read coverage throughout transcripts. Illumina RNA sequencing was performed in triplicate.

For RNA 3′-end sequencing in transcription factor overexpression strains, libraries were prepared from 500 ng total RNA using QuantSeq3′ REV kit (Lexogen 016.24). Following denaturation at 85°C, cDNA was synthesized at 42°C from the polyT-containing FS1. Second strand synthesis was initiated following RNA removal. Samples were purified with the QuantSeq purification module. Libraries were amplified for 12 cycles with i7 barcoding. Libraries were purified with the purification module and analyzed for fragment size distribution and concentration by Agilent HS DNA Bioanalyzer and Qubit HS dsDNA assays. Equimolar quantities of 8 samples with balanced indices were pooled for single read, 75bp sequencing on the NextSeq 500.

Gene-specific cDNA sequencing

Oxford Nanopore's PCR-cDNA Barcoding was adapted to perform gene-specific cDNA sequencing of full-length *YIR018W* and *YIR018C-A* transcripts in strains with *YIR018C-A* under the control of a tetracycline-repressible promoter. An oligo(dT) primer, containing UMI (VNP+UMI: ACTTGCCTGTCGCTCTATCTTCCGGTGTNNNNNNNNNVTTTTTTTTTTTTTTTTTTTTVN) was used for reverse transcription of 10 ng polyA RNA by SuperScript® III reverse transcriptase (NEB) in a 20µL standard reaction at 50 ℃ for 45 min. Following heat inactivation (70 ℃, 15 min), RNA was removed via treatment with 1 µL RNaseH (NEB) at 37 ℃ for 20 min. A tenth of the cDNA (2 µL) was used directly for primer extensions with a mixture of UMI-containing, gene-specific primers (Table S6) in a 20 µL LongAMP Taq (NEB) reaction with 2 µM each primer and 0.2 mM dNTPs; after 30 sec denaturation at 95 ℃, 2 cycles of 95 ℃ for 15 sec, 50 ℃ for 15 sec, and 65 ℃ for 7 min 30 sec were used for second strand synthesis. Exonuclease I (NEB) treatment at 37 ℃ for 15 min, followed by heat inactivation at 80 ℃ for 15 min removed ssDNA.

The extension reaction was purified with 1.5 vol AMPure XP beads (Beckman), washed with 70% ethanol twice, and eluted in 10 µL water. The eluate was used in a final PCR with barcoded primers (SQK-PCB109): after initial denaturation at 95 ℃ for 30 sec, 18 cycles of denaturation at 95 ℃ for 15 sec, annealing at 62 ℃ for 15 sec, and extension at 65 ℃ for 2 min 30 sec were performed. The PCR reactions were exonuclease treated as above and purified with 1 vol AMPure XP beads, washed twice with 70% ethanol, and eluted in 10 µL water. Qubit HS dsDNA assays quantitated DNA concentration, and samples were combined in equimolar ratios to a total of 120 fmoles DNA (assuming ~900bp DNA length, ~585000 mol weight), which was re-purified with 1 vol AMPure XP beads and eluted in 12 µL EB (SQK-PCB109). The mixed sample was quantitated via Qubit HS dsDNA assay, and the remaining 11 µL were treated with 1 µL Rapid Adapter (RAP, SQK-PB109) at room temperature for 5 min. The RAP-treated sample was mixed with 37.5 µL sequencing buffer (SQB) and 25.5 µL loading beads (LB) (SQK-PB109), loaded on a primed (EXP-FLP001) MinION flow cell (FLO-MIN106D R9), and run on GridION.

RT-qPCR

RT-qPCR was used to assess the expression of *YIR018C-A* and *YIR018W* upon doxycycline-induced repression, using *ACT1* as an internal standard. polyA RNA (40 ng) from 24 h growth with and without doxycycline

were used for 20 µL reverse transcription reaction with SuperScript® III reverse transcriptase (NEB) using a 2 µM mixture of gene-specific oligos, harboring a 5′ tag (Table S7). The reaction was treated with RNaseH (NEB) at 37 °C for 20 min, and then diluted with 30 µL water. The diluted cDNA (2 µL) was used directly in a 20 µL qPCR reaction (Applied Biosystems, 2x SYBR), containing 250 µM primer sets, where the forward primer matched the mRNA sequence ~100-200bp upstream of the RT primer and the reverse primer matched the tag introduced during the RT (Table S7). After an initial denaturation at 95 °C for 10 min, 2-step cycling (95 °C for 15 sec and 62 °C for 1 min) was performed on a StepOne™ machine, followed by melting curve analysis. Analyses were performed in triplicate for each primer set. Quantitation relative to a control strain, expressing only the tTA transactivator, was calculated as -ΔΔCt and is reported as mean +/- standard deviation.

Feature mapping, novel junction detection and annotation

All features assigned to the synIXR chromosome in BioStudio (*28*) were transferred by alignment to each SCRaMbLE strain. SynIXR sequences in SCRaMbLE strains were inferred from putative genome order based on DNA sequencing in (*15*). For strains with more than one possible reconstruction (22 out of 64 strains), the first putative segment order was assumed. Since ambiguities that lead to there being multiple possible genome assemblies are identical at the level of local junction configurations, any errors from incorrect genome assembly would not affect our results. Features greater than 50 bp in length were aligned using pymummer 0.11.0 (*29*). All other features were aligned using GSNAP v.2019-01-24 (*30*). In cases where an alignment could not be found, the segment order was used to infer feature location. The feature mapper was implemented as a custom python script available at: git.embl.de/brooks/scramble-transcriptome/scrambleMapper.py and on Zenodo (DOI 10.5281/zenodo.5676293) (*23*).

Novel junctions were identified by non-contiguous segment order. At each novel junction, the feature types were assigned by identifying the annotations closest to the junction in a restricted annotation setting where only select annotations were retained. Genic features (5′-UTR, 3′-UTR, and gene) were prioritized such that in cases of multiple annotations occurring within 250 bp of a junction, the genic features would be selected, regardless of whether the other feature type was closer to the junction. The order of prioritization was, from high to low: gene, 3′-UTR, 5′-UTR, pseudogene, ncRNA_gene, SUT, CUT, XUT, and centromere, with all genic features having equal weight. In instances where no annotation was associated with a novel junction (e.g. small segments in between convergent genes), annotations from the adjacent segment were used up to a maximum of 10kb from the junction. Since synIXR chromosomes are circular, novel junctions were also detected between the first and final segments of the linear sequence. This 1:44 junction that is present even in the parental -SCRaMbLE strain was not considered a novel junction. The novel junction mapper was implemented as a custom python script available at: git.embl.de/brooks/scramble-transcriptome/scrambletools.py and on Zenodo (DOI 10.5281/zenodo.5676293) (*23*).

Base calling, quality-filtering and long-read alignment

Nanopore long reads were basecalled with Guppy v3.2.2. Adapter sequences were trimmed with Porechop v0.2.4. Reads were quality filtered to a minimum average Phred score >= 6 using fastp v0.20.0 (*31*). Reads were aligned to SCRaMbLE genomes with minmap2 v2.17 (r941) using the settings recommended for direct RNA sequencing (-ax splice -uf -k14) and a maximum intron size of 1500 nucleotides. (-G 1500) (*32*). Multi-mapping reads were filtered to keep only reads with the best alignment scores. Alignments with identical mapping scores were retained. Long reads were used to define transcript models used in transcript-level quantification as described below. In cases where no long read corresponding to a gene was observed in the long-read dataset, the original gene boundaries were used for the transcript model.

Gene expression quantification

NEBNext Ultra II RNA was quantified at the transcript-level with Salmon v0.12.0 (*33*) using transcript models defined from the long reads as described above. Salmon was run with both sequence and position bias modeling enabled. Copy number effects were corrected by fitting a linear model to the relationship between copy number and expression level and assigning gene expression levels to the residuals from this model.

Clustering reads into isoforms

Direct RNA reads were clustered into isoforms that represent repeated measurement of reads with co-occurring transcript start site (TSS) and transcript end site (TES) positions. Reads were ranked by their abundance and then recursively collapsed based on TSS and TES co-occurrence within a 25 nt window. Isoforms were further collapsed across strains by chromosome, TSS and TES locations in the manner described above. We retained all

isoforms supported by 2 or more full-length reads. A range of thresholds for collapsing isoforms were tested before choosing a 25 nt cutoff to ensure robust isoform calling (fig. S2B to C). The 25 nt threshold chosen is less stringent than the 5 nt threshold used previously for our transcript isoform sequencing method (TIF-seq) (*19*) due to a documented 3′ bias associated with Nanopore direct RNA sequencing (*34*).

Isoform type and mTIF assignment

Isoforms were assigned a type based on the feature the isoform overlapped and the extent to which it covered that feature as in (*19*). Isoforms that covered an entire feature, with a 25 bp buffer on either side, were classified as 'Covering', e.g. 'Covering one intact ORF' or 'Covering >= 2 ORFs'. Isoforms that covered only one side of a feature were classified as 'Overlap', e.g. 'Overlap 5′ of one ORF' or 'Overlap 3′ of one ORF'. Isoforms that did not overlap any annotated feature were classified as intergenic. All isoform type assignments were strand specific.

mTIFs from (*19*) were assigned to long reads and isoforms by matching TSS and TES coordinates. A read or isoform was considered to match an mTIF if both the TSS and TES coordinates of a read or isoform were within 25 bp of an mTIF's TSS and TES coordinates. In cases where multiple mTIFs could map to an isoform, only the mTIF with the average closest distance to both the TSS and TES was kept. In cases where a matching mTIF could not be found, the nearest mTIF was reported.

Feature and transcriptional neighborhood similarity

Feature and transcriptional neighborhood similarity were calculated using cosine similarity on full-length isoform coverage vectors. Coverage vectors consist of the read coverage (count) at each position within a defined interval. Cosine similarity was calculated as:

$$\text{similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

Where A is the feature coverage vector of interest and B is the reference coverage vector in BY4741. For quantification on individual TUs, reads covering at least 10% of the feature's length were included in the coverage vector. To compare across strains, coverage vectors were aligned with respect to the gene CDS and zero padded so that vector lengths were equal. For quantification of transcriptional neighborhoods, coverage vectors were calculated for a fixed 3 kb window upstream and downstream of a gene on each strand.

Dissimilarity was calculated as 1 - cosine similarity. Cosine similarity is defined on the interval [-1,1], however, given the properties of our data we only observe values in the range [0,1] when computing cosine similarity on stranded reads. However, computing similarities using reads on both strands resulted in values defined on the full interval, as in Figure 4C.

Isoform and neighboring transcriptional similarity for ohnologs

Ohnolog gene pairs were obtained from the Yeast Gene Order Browser (Version 7, August 2012) (*35*). Likewise, 5X the number of ohnologs or 2,735 random gene pairs were sampled. For all gene pairs, the aligned TU, upstream (3kb) and downstream (3kb) transcriptional similarity were computed as described above.

Gradient boosted regression trees (GBRTs) gene expression fold-change, TSS and TES predictions

10x, 5-fold cross-validated GBRTs were trained to predict gene expression fold-change and isoform TSSs and TESs. Isoform TSSs and TESs were defined as the distance from the isoform TSS/TES to the CDS start/end. Models were trained using sequence and transcriptional neighborhood features within 3 kb. Sequence features were categorical and included the identity and orientation of upstream and downstream genes. Transcriptional neighborhood features were continuous and included gene expression (TPM), gene expression fold change relative to BY4741, distance to the closest isoform, and cosine similarity calculated on either the same or opposite strand. GBRTs were trained with XGBoost (*36*) using regression and a squared error objective. For each model, 10,000 trees were trained at a learning rate of 0.01, maximum tree depth of 5, subsampling ratio equal 0.4, L1 regularization equal 0.75, L2 regularization equal 0.45, and subsampling ratio equal 0.6. Parameters were chosen by manual tuning to maximize model performance. Additional models with the same parameterizations were trained using sequence features only, transcriptional features only, and the top two transcriptional features identified from the original model. Performance metrics (MSE and ΔMSE) were computed using each of the 10X, 5-fold cross validated models. Feature importance scores were averaged across the top 10 performing models.

Predicted pairwise segment transcriptional similarities, co-clustering and comparison to observed similarities and effects

Cosine similarity between all pairs of segments along synIXR was computed from transcriptional data in the parental synthetic -SCRaMbLE strain, JS94, as described above. Transcriptional coverage was only considered within the segment itself, without extending to the TSS/TES of reads. For upstream comparisons, the segment vectors were right-aligned with a zero buffer extending to the maximal length of the longest coverage vector. For downstream comparisons, the segment vectors were left-aligned. In both cases this was done to mimic the expected adjacent transcriptional patterns, independent of segment length. Pairwise similarity comparisons were performed for all pairs of segments in both an upstream and downstream configuration as well as altered orientation where one of the two is inverted. To compute the inverted coverage vector, the native coverage vector was position and value inverted (flipped and multiplied by -1 to "switch" the strand). Spectral co-clustering with cluster number set to 2 was performed separately for upstream and downstream similarities to organize the pairwise values. Inverted pairwise comparisons were co-clustered according to the non-inverted cluster hierarchy. Transcriptional alterations occurring at rearrangements observed in SCRaMbLE strains were compared to the predicted transcriptional similarities.

3′-end mapping, quantification and detection of TES in galactose-inducible transcription factor overexpression strains

mRNA 3′ sequencing reads were trimmed and quality filtered using fastp v0.20.1 (*31*). Filtered reads were aligned to the yeast genome with bowtie2 v2.4.2. Gene expression levels were quantified using Salmon v1.4.0. (*33*). 3′-ends of each read were assigned to the closest gene using bedtools closest v2.30.0. Only reads mapping within a range -250 nt to 500 nt from the 3′-end of the CDS were retained for subsequent analysis. Overexpressed genes were identified directly from the data using a minimum read cutoff of 100 reads and a minimum 20-fold increase in expression in galactose relative to raffinose. For each overexpressed gene, we determined whether its downstream (in the direction of its transcription) neighbor was tandemly or convergently oriented. For each gene adjacent to an overexpressed gene, we tested whether TES selection was significantly altered in galactose relative to raffinose using the Kolmogorov-Smirnov test on normalized TES site counts. For each gene, TES site counts were normalized relative to the unit length of the 3′-UTRs observed across conditions to allow comparison of TES changes, despite different absolute TES to CDS distances. The difference between the area under the curve (Δ auc) of TES cumulative distributions in raffinose and galactose were used to determine whether transcripts tended to shift towards shorter or longer isoforms.

Statistical tests

The Mann-Whitney U test (or Wilcoxon Rank Sum Test) was used throughout as a nonparametric test for the equality of means in two independent samples. We used the test to identify significant differences in isoform boundaries and expression levels between rearranged contexts. Levene's test for equality of variance was used to determine significant alterations in variability of TSS/TES distributions (Fig. 1B). The Kolmogorov-Smirnov test was used to determine whether TES distributions were significantly different for a gene after galactose-induced transcription factor overexpression (Fig. 5F). The flipped appearance in some boxplots (e.g. fig. S9) is expected when the 95% confidence intervals extend beyond the first and/or third quartiles.

Supplementary Text

SCRaMbLE generates novel recombinations of genomic features

SCRaMbLE-induced deletions and duplications massively altered the size of the synthetic genome, with synIXR chromosomes ranging from half up to three times the original size in SCRaMbLE strains (*16*). Using MUMmer (*29*) and GSNAP (*30*), we were able to annotate 612 novel junctions across the 64 strains (fig. S1A). Novel junctions were present at a frequency of 1 to 48 novel junctions per strain. In total, genes were recombined with 219 novel upstream features and 526 novel downstream features across the strain collection (*16*). Specific novel junctions rarely occurred in multiple strains: 89% of specific recombination events between two loxPsym sites were represented only once or twice (fig. S1B). Taking into account the repetition of specific recombination events in the SCRaMbLE strains, there are a total of 363 distinct novel junctions. Only two novel junctions are present in more than 10 strains and both feature loss of a small intergenic region that could be the result of genome assembly errors. With respect to the feature composition at these novel junctions, we discovered, reassuringly, that nearly 7 out of 10 novel junctions represent plausible genomic rearrangements, including new convergent gene pairs (+gene:-gene, 38%), genes with alternative 3′-UTRs (+gene:+ 3′ -UTR, 14%), tandem gene pairs (+gene:+gene, 8%), and genes with alternative 5′-UTRs (+5′-UTR:+gene, 8%) (fig. S1A). The remaining junctions include unusual feature combinations (e.g. +gene:- 3′-UTR) and novel junctions that include less common genomic elements, like ncRNAs

(annotated SUTs, CUTs and XUTs) and the centromere. Given that we prioritized novel junction assignments with genic features (gene, 5′-UTR and 3′-UTR), it is unsurprising that features like ncRNAs are underrepresented.

Identifying RNA isoforms with Nanopore direct RNA sequencing

With Illumina 75bp paired end reads, only ~10% of reads mapped to the synthetic chromosome fully-spanned loxPsym junctions with an additional 30 nt of flanking region. In contrast, 88% of Nanopore reads aligning to the synthetic chromosome met these criteria and were uniquely mappable (fig. S2A). Other than increasing median 3′-UTR length by 34 nt, incorporating 34 bp loxPsym sites downstream of all non-essential genes did not affect TES distributions, indicating that the placement of loxPsym sequences at the end of CDSs did not globally disrupt termination and polyadenylation site recognition (Fig. 1B and fig. S5B). Long reads also identified single molecule TSSs and TESs, allowing us to ascertain changes in isoform usage. Given tradeoffs between depth, quantification accuracy, reproducibility and cost across these two platforms, we chose to use Illumina short-read sequencing for gene quantification and Nanopore long-read direct RNA sequencing for isoform detection.

To identify and quantify divergence between isoform usage in SCRaMbLE strains genome-wide, we first clustered single-molecule reads into isoforms, which represent repeated measurement of transcripts with co-occurring TSS and TES locations. As previously described (*19*), we clustered reads ranked by their abundance, recursively collapsing reads based on TSS and TES co-occurrence within a 25 base pair window at each end. Alternative thresholds were tested to maximize isoform detection while minimizing the window size (fig. S2B to C). All together, we identified 264,899 isoforms supported by 2 or more reads. Half of all isoforms, however, are supported by 5 or more reads (fig. S2E). The full set of isoforms are provided as Table S2. In our dataset, protein-coding genes are expressed as 6 distinct isoforms on average, which is substantially fewer than the 26 isoforms per-protein coding gene observed previously (*19*) (fig. S2D); however, this is likely due to a combination of the reduced sequencing depth in our long-read dataset and the relaxed clustering threshold. Having sequenced <2 million transcripts (i.e. long-reads) by direct RNA sequencing per strain on average, we can detect isoforms with a minimum abundance of approximately 1 transcript per million at a minimum cutoff of 2 reads per transcript (fig. S2F).

Direct RNA isoforms correspond to mTIFs identified by TIF-seq

To determine whether direct RNA sequencing isoforms accurately report on transcript TSSs and TESs, we compared both the reads and the isoforms we identified to 371,087 major transcript isoforms (mTIFs) discovered by transcript isoform sequencing (TIF-Seq) (*19*). This comparison is possible because 95% of isoforms identified are on non-synthetic chromosomes that are not rearranged by SCRaMbLE. As above, we used a 25 nt co-occurrence threshold at each end to map long reads and isoforms to mTIFs. We find that full-length direct RNA reads are consistent with those identified by TIF-Seq. 77% of mTIFs, for example, are covered by a direct RNA read (fig. S3A). Interestingly, however, only 56% of long reads correspond to an mTIF (fig. S3A). This difference can be explained by several possibilities, including: (1) that direct RNA sequencing produces spurious reads, including 5′-end degradation products, which would be discarded by our clustering approach; (2) that transcription on the non-synthetic chromosomes is indirectly affected by SCRaMbLE; or (3) that direct RNA sequencing detects transcripts that were inaccessible to TIF-Seq. Here we note that direct RNA sequencing has far fewer handling and preparation steps compared to TIF-Seq, which may reduce some types of artifacts. At the isoform level, there is also general agreement between the datasets. 69% of the 48,253 long-read isoforms covering a single, complete ORF on non-synthetic chromosomes have a corresponding mTIF (fig. S3A). The largest discrepancies between TIF-Seq and direct RNA sequencing are for intragenic species and unstable ncRNAs (2% and 20% mapping rate, respectively), as well as transcripts that do not fully cover an entire genomic feature (fig. S3B). Interestingly, we also observe a marked increase in the number of polycistronic isoforms detected with long read sequencing (an additional 4,909 isoforms, a ~60% increase; fig. S3B). Together, these results suggest that direct RNA sequencing accurately captures the complexity of the yeast transcriptome as observed previously with other technologies, especially for high-confidence isoforms that span entire genomic features. Additionally, increases in the abundance of long polycistronic transcripts in our dataset could represent discovery of transcripts that were not identifiable with TIF-Seq due to technological limitations.

Composition of direct RNA isoforms

SCRaMbLE-related isoforms display different feature composition compared to those from native chromosomes or in the parental synthetic strain. On average across all strains, 48% of isoforms that map to the synthetic chromosome cover a single intact ORF, as compared to 60% on native chromosomes. We also found that

19% of isoforms overlap only the 3′-end of an ORF on the synthetic chromosome, which is likely due to the previously described Nanopore direct RNA 3′-end bias, in addition to a lack of selection against 5′-end degraded transcripts (*34*). A major change in the transcriptional landscape of rearranged chromosomes is an increase in polycistronic isoforms. Polycistronic isoforms account for more than twice the fraction of isoforms in SCRaMbLE strains compared to the parental synthetic strain and are far more abundant than on native chromosomes, suggesting that degeneration of evolved linear gene order may favor generation of polygenic species.

Expression levels of transcript isoforms

Genomic context impacts gene expression. One obvious product of SCRaMbLE is an increase in gene copy number. Most genes (53 out of 59) on the synthetic chromosome are present in 2 or more copies in at least one SCRaMbLE genome. As has been widely reported previously, copy number increases result in increased gene expression. A linear model was fit between copy number and gene expression in the SCRaMbLE strains to remove the confounding effects of copy number in duplicated genes (fig. S6A). The corrected expression data, or the residuals from that model, focus on changes due to rearrangements. We quantified gene expression using transcript-level quantification with Salmon on short-read sequencing data collected from three biological replicates in every strain. We decided to use Salmon-based quantification after benchmarking several alternative approaches, including quantification with direct RNA reads (fig. S6B).

Neighborhood affects RNA isoform expression levels and lengths

Rearrangements generally led to a higher degree of expression variability, except for cases of rearrangements at both ends (Fig. 1C). Consistent with highly variable expression, we identified 141 cases where rearrangement led to a complete loss of gene expression and 18 cases where a gene that was previously not expressed became expressed (Table S4). Importantly, the genes that lost expression had been abundantly expressed (with a median ~25 transcripts per million [TPM]), and those that gained expression achieved an average level (median ~25 TPM). Thus, these observed changes were likely not due to lowly expressed genes being variably represented in sequencing samples. Since gene promoters and 5′-UTRs are preserved by SCRaMbLE, these findings indicate that gene expression levels are not hardwired into the promoter or coding sequence itself but could be conditioned by context.

Transcript isoform alterations, such as the striking 3′-end lengthening seen for the rearranged gene *YIR018W* (Fig. 1E and fig. S8) and the altered composition of the *RPR2* polycistron (Fig. 3A), do not appear to be explained by changes in sequence alone but do appear to be associated with changes in transcriptional environment. In general, neighboring transcription appears to create environments that support and even define local gene expression.

We can observe these effects on both transcript isoform boundaries (fig. S9B) and gene expression (fig. S9A). Beyond inducing changes in novel contexts, however, these results entail that similar transcriptional environments should create analogous transcriptional patterns. As a consequence, transcription from genes in rearranged genomic regions should not only be altered when the transcriptional environment is dissimilar to its native environment, but transcription should also adopt characteristics associated with its new environment. Consistent with this, we observe isoform lengthening on both 5′- and 3′-ends associated with the proximity of neighboring transcription (fig. S9B). More specifically, 5′-ends shorten as upstream transcripts become closer, while 3′-ends lengthen as downstream transcripts are further removed. The alteration of transcripts in rearranged contexts mirrors the patterns observed in native contexts, implying that rearranged genes generate novel isoforms with characteristics inherited from their new transcriptional environments. Likewise, relocating a gene to a region where neighboring gene expression is high, generally leads to significant increases in gene expression (fig. S9A). This observation, which is consistent with clustering of co-expressed genes and formation of permissive chromatin and TADs seen in higher eukaryotes (*37-41*), occurs across a broad range of expression values and appears to function in all forms of genic rearrangements except for newly formed bidirectional promoters. Together, our results suggest that transcriptional environments influence, and may even determine, isoform choice and gene expression levels.

Gradient Boosted Regression Trees predict TSS, expression and TES

Profiling gene expression levels and isoform boundaries in more than 600 novel genomic contexts suggests that the transcriptional environment into which a gene is rearranged has a significant impact on transcript levels and lengths (Fig. 1B to C, Fig. 5B to D, and fig. S9). This hypothesis, however, is based on multiple aggregate associations, as presented previously. A complex combination of features, including neighboring sequence,

expression levels, and isoform characteristics may be required to explain transcriptional changes observed at any particular locus. To disentangle these relationships and to understand the extent to which neighboring sequences and transcriptional features explain gene expression and isoform choice, we built a predictive model based on regression trees. We trained Gradient Boosted Regression Trees (GBRTs), a non-parametric statistical learning technique consisting of an ensemble of 10,000 regression trees per model, to predict either gene expression fold changes, isoform TSSs, or isoform TESs in rearranged genomic contexts. 10 replicate GBRTs with 5-fold cross-validation each were produced for each prediction task to assess model reproducibility and overfitting. We trained the models using sequence features (upstream and downstream gene identity and orientation) and transcriptional neighborhood features (upstream and downstream gene expression, fold changes, and isoform changes, including cosine similarities on both strands and distance to the closest isoform) within 3 kb of a TU. All of the features included for training, with the exception of sequence features, are not highly correlated (fig. S12A). Each of the models was robust to the training input, indicating that models were not overfit, and able to predict fold change, isoform TSS or isoform TES sites with a reasonable degree of accuracy, particularly given the limited size of the training set and complexity of the prediction task. Fold-change was predicted with an average error of 3.9-fold, while isoform TSSs were predicted within 97 bp on average and isoform TESs within 174 bp (fig. S12B). Average errors were computed as the median of the mean error ($\sqrt{MSE}$) for each model trained. We urge caution in interpreting these errors, however, as outliers in the dataset contribute disproportionately. We likewise note that there is reasonable predictive performance across the entire dynamic range for each prediction task (fig. S12C). Given that we were not interested in the predictions themselves but rather in assessing feature importance scores, which indicate how important specific features were for making the prediction, we found this level of predictive accuracy acceptable.

Feature importance in the GBRT models provided valuable insight into factors affecting transcription in novel genomic contexts. To establish robust estimates of feature scores, we averaged feature importance scores from the top ten performing models. These scores confirm previously described hypotheses and observations, as well as suggest new associations (Fig. 4A). For example, upstream features, including the identity of the upstream gene, its fold change in expression, and the cosine similarity on the same strand are most predictive of how gene expression levels will change in a new genetic context, consistent with associations described in supplementary Figure 9A. Similarly, isoform TSSs tend to be influenced predominantly, although not exclusively, by their upstream context, while downstream context appears to have a greater influence on isoform TESs. There were also additional associations for which the mechanism is less clear, such as the effect of downstream isoform proximity on gene expression. As anticipated given the complexity of the prediction tasks, nearly every feature is important for prediction in at least one case. There are, however, clear trends that identify contextual features important for establishing transcriptional characteristics in new genomic locations.

Our previous observations about the importance of up- and downstream transcriptional environments for establishing local transcription characteristics were supported by the GBRTs. To determine to what extent neighboring sequence information relative to transcriptional environment contribute to the predictability of transcriptome alterations in SCRaMbLEd genomes, we again trained 10x, 5-fold cross-validated GBRTs, but this time with either sequence features alone or transcriptional environment features alone. While models lacking sequence features performed nearly as well as the full model (Fig. 4B), models trained with sequence features alone performed significantly worse. To ensure that this was not due to there being fewer features for training in the sequence only models, we trained additional GBRTs using only the top two environmental features for each prediction task to match the number of sequence features included in the sequence only models. These reduced models performed significantly worse and on par with the sequence only model, indicating that the reduced performance of the sequence only model could be due to the reduced number of training features (fig. S12D). All together, these results suggest that gene isoforms and gene expression are predictable in yeast, even in novel contexts, and that adjacent transcriptional information is at least as important as adjacent sequence features for making that prediction.

Predicting neighboring transcriptional changes from a reference transcriptome

The GBRT models suggested that TSS, expression and TES are predictable given measurement of transcription flanking a gene. However, the requirement for measuring adjacent transcription in a novel context to predict local transcriptional changes poses a significant and unnecessary burden, especially because local gene expression changes (length and/or level) could be observed at the same time. We therefore sought to understand whether neighboring transcriptional changes are predictable given: (1) specification of the genomic segments to be recombined, (2) their position and orientation relative to the gene of interest and (3) measurement of transcription

for each segment in a reference strain, e.g. the parental SCRaMbLE strain. We predicted the transcriptional similarity (cosine similarity) between all pairs of possible rearrangements, in both upstream and downstream contexts and when one of the segments had been inverted relative to the other. Using these predictions of altered up- and down- stream transcriptional environment, we observed general agreement with the directionality of changes in local transcriptional properties (Fig. 4C).

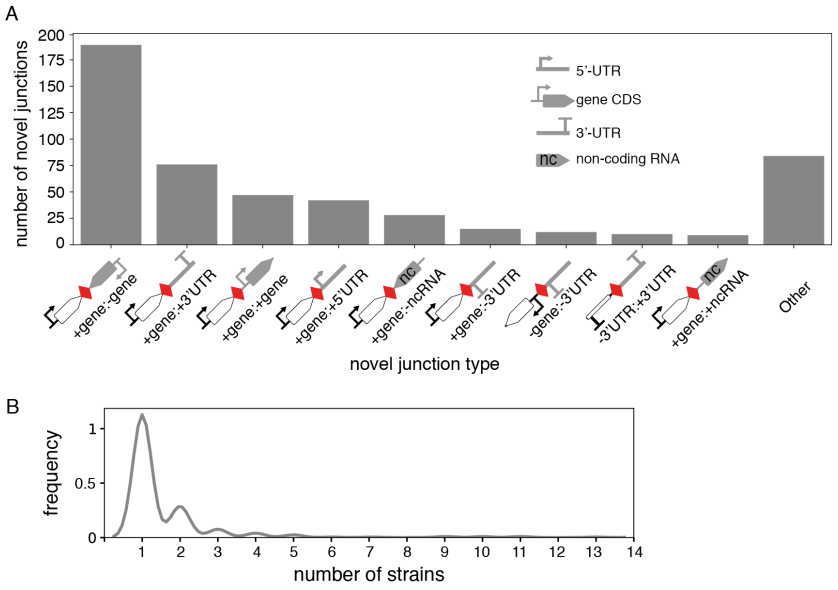

Fig. S1. SCRaMbLE induces large-scale structural variation in synthetic genomes.

(A) SCRaMbLE rearrangements at loxPsym sites 3 bp downstream of a gene CDS create combinations of diverse functional genomic elements. The number of rearrangements from the 612 novel junctions found across all 64 synIXR SCRaMbLE strains are plotted for each junction type. (B) Specific novel junctions (i.e., specific recombination events between the same two loxPsym sites) are typically represented in only one or two strains. The frequency of a novel junction's occurrence across all strains is shown.

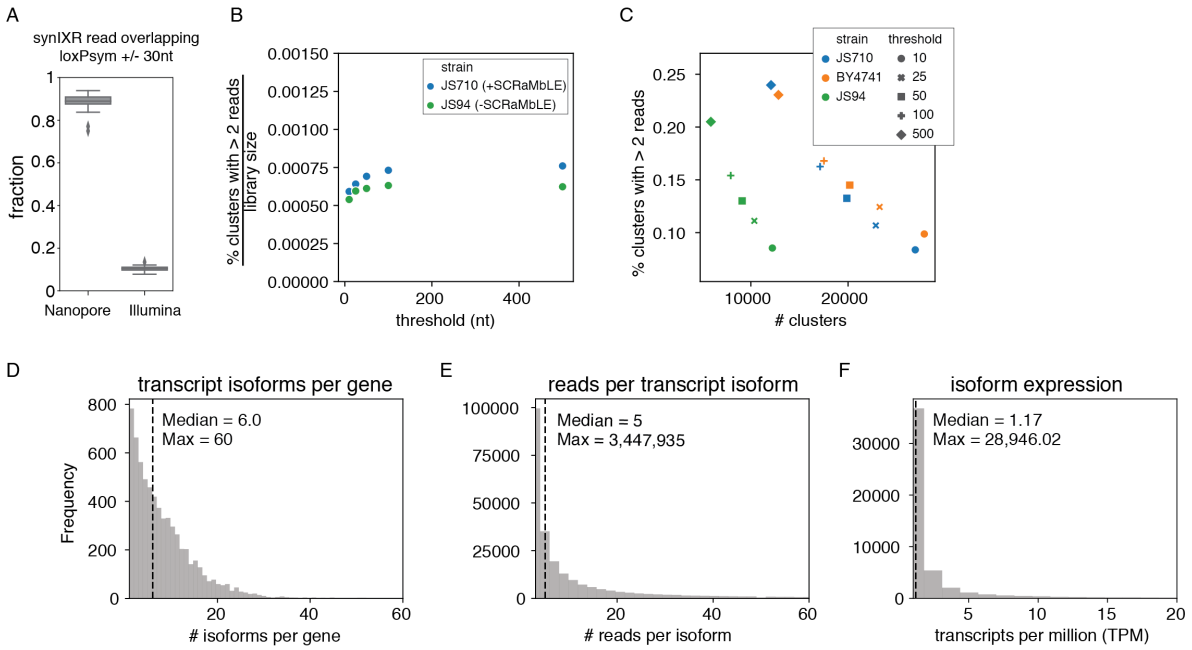

Fig. S2. Long-read sequencing allows for transcript isoform mapping.
(A) Long-reads from Nanopore direct RNA sequencing improve mapping across novel junctions. Fraction of reads aligned to the synthetic chromosome that contain a full loxPsym site plus 30 nt on either side from Nanopore and Illumina sequencing data from all strains. Data are represented as the median and interquartile range with whiskers extending ±1.5 times the interquartile range. (B) Several different thresholds (10, 25, 50, 100, and 500 nt) for window size were tested for clustering reads with co-occurring TSSs and TESs into transcript isoforms from chrI or synIXR. Fraction of clusters supported by 2 or more reads using these thresholds is shown for two strains (JS710 (+SCRaMbLE), blue; JS94 (-SCRaMbLE), green), adjusting for library size (number of reads collected for each strain). (C) Window size selection sought to maximize the number of transcript isoforms (clusters supported by more than 2 reads), while minimizing the window size. The fraction of clusters supported by more than 2 reads and number of total clusters generated with different thresholds for window size is shown. 25 nt was chosen as the window size for clustering reads into transcript isoforms. (D) Number of total transcript isoforms per gene in all strains. (E) Number of long reads supporting each isoform. (F) Expression of each isoform based on short-read data.

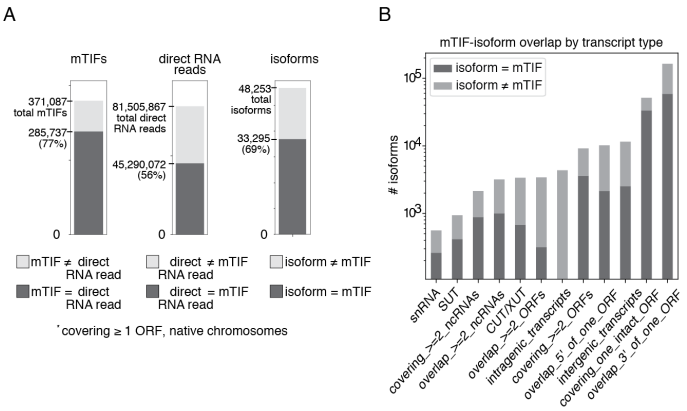

Fig. S3. Validating isoforms from direct RNA sequencing against major transcript isoforms (mTIFs). (A) Long-reads that map to the non-synthetic chromosomes and the isoforms defined from them correspond well with mTIFs. Fraction of mTIFs covered by a direct RNA read (left), fraction of direct RNA reads with a corresponding mTIF (middle), and transcript isoforms covering one ORF with a corresponding mTIF (right) are shown. (B) Transcript isoforms defined by direct RNA sequencing are separated by isoform type, defined by the genomic features to which they map. Shading indicates the fraction of isoforms that correspond to an mTIF.

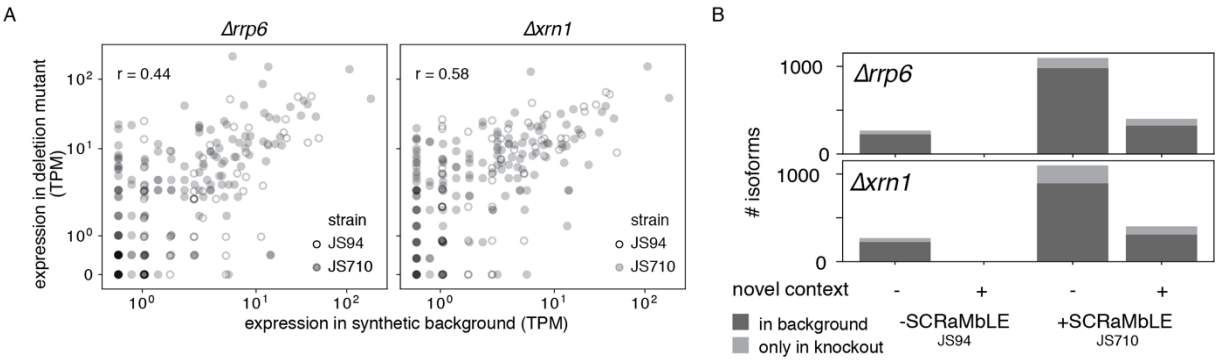

Fig. S4. Transcript isoforms detected in SCRaMbLE strains are stable.
(A) Nanopore-sequencing determined isoform expression levels in exosome component deletion mutants (Δ*rrp6* and Δ*xrn1*) compared to the synthetic strain backgrounds. JS94 (-SCRaMbLE) and JS710 (+SCRaMbLE) backgrounds were used. Pearson correlation coefficients (r) between expression levels in exosome deletion mutants and synthetic strain backgrounds are shown. (B) Exosome deletion has little effect on the number of transcript isoforms detected. Number of transcript isoforms detected in Δ*rrp6* or Δ*xrn1* knockout strains in +/-SCRaMbLE backgrounds are shown. Transcript isoforms are divided into those also observed in the synthetic background (dark gray) or unique to the knockout (light gray) in JS94 and JS710 strains and are separated as arising from native or novel contexts.

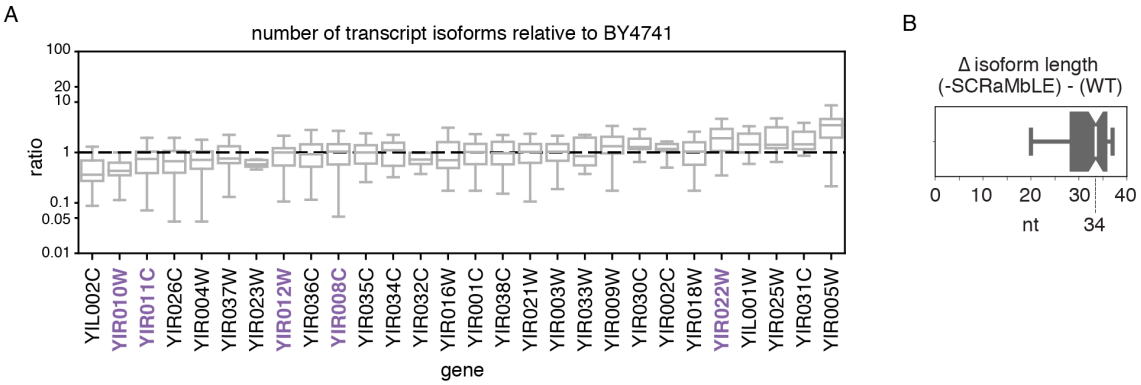

Fig. S5. SCRaMbLE alters TSSs and TESs.
(A) Number of isoforms for each gene across SCRaMbLE strains expressed as a ratio relative to WT BY4741. Dashed line indicates the same number of isoforms as found in WT for each gene. Essential genes are shown in purple. (B) Distribution of differences in isoform lengths between the WT and loxPsym-containing -SCRaMbLE strains. Median difference is 34 nt, which is the size of the loxPsym site.

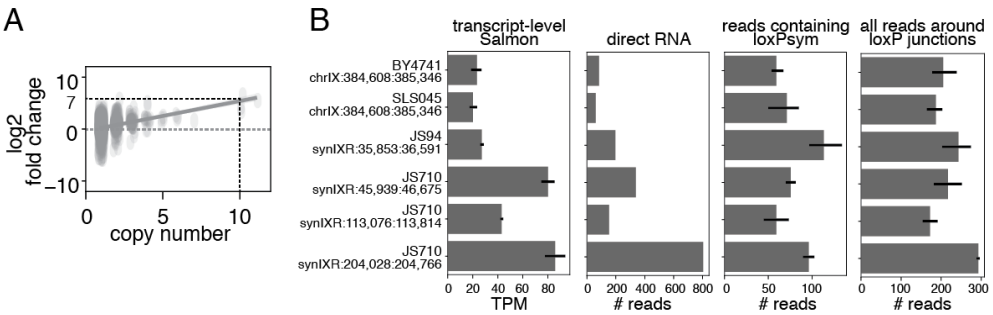

Fig. S6. SCRaMbLE leads to expression changes.
(A) SCRaMbLE-induced duplications increase gene expression sub-linearly with copy number. Standard linear regression was used to remove these confounding effects for subsequent analyses of expression levels. (B) Transcript-level quantification of Illumina short-read data by Salmon was benchmarked against other methods. Salmon quantification was compared to estimates based on direct RNA reads, and direct counting of Illumina reads using either reads that contain a loxPsym site or all reads flanking a loxPsym junction. Comparison of these quantification methods was performed for the gene *YIR108W* in WT strains, BY4741 and SLS045, JS94 (-SCRaMbLE) and three rearranged contexts in JS710 (+SCRaMbLE). In strains without loxPsym sites (BY4741, SLS045) an equivalent window corresponding to where the loxPsym site would have been was used. Bars indicate 95% confidence interval based on 3 biological replicates.

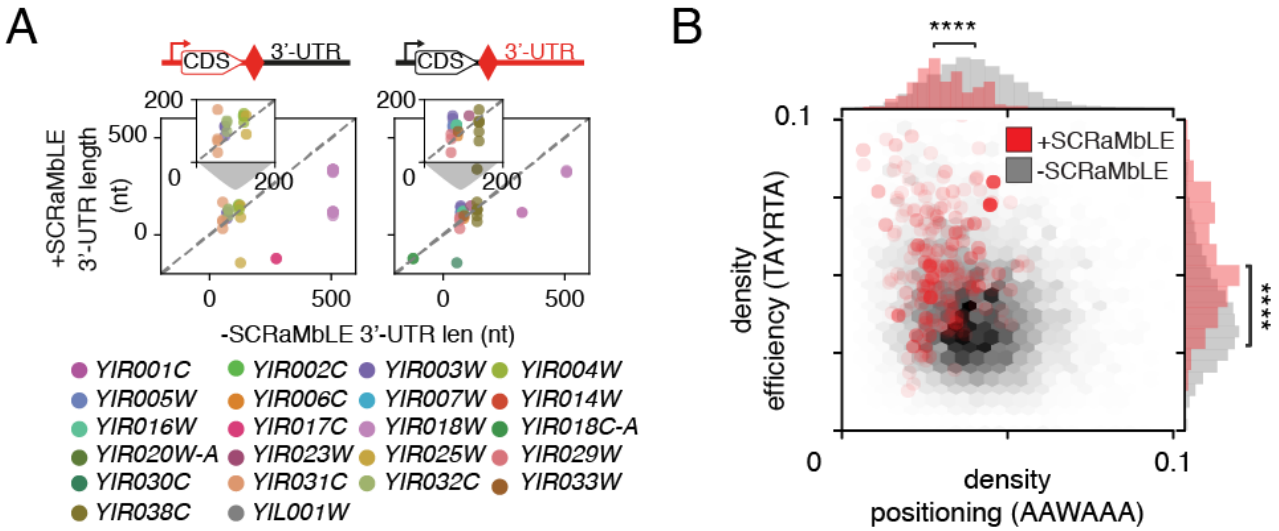

Fig. S7. 3′-UTR sequence is not sufficient to explain length alterations.
(A) Neither 3′-UTR (left) nor CDS (right) sequences maintain the 3′-UTR length of their native context when coupled to different CDSs or 3′-UTRs, respectively. TES distance from coding regions for all 3′-UTR-associated rearrangements are compared to -SCRaMbLE. These are aggregated by 3′-UTRs coupled to alternative coding regions (left) or CDSs coupled to alternative 3′-UTRs (right). Points above the diagonal line indicate increased 3′-UTR length compared to the native context, and points below the line show reduced 3′-UTR length. Colored points designate the identity of the recipient 3′-UTR (left) or gene (right). (B) Presence of polyadenylation signal (PAS) sequence motifs is altered by SCRaMbLE-induced replacement of native 3′-UTR sequences. Density of PAS efficiency (y-axis) and positioning motifs (x-axis) in a 50bp window surrounding the observed TESs of each isoform in +/-SCRaMbLE contexts. The density of positioning motifs decreases in +SCRaMbLE (left shift), while efficiency motifs increase in +SCRaMbLE (upward shift). Asterisks denote significance in Mann-Whitney U test, **** $p \leq$1e-4.

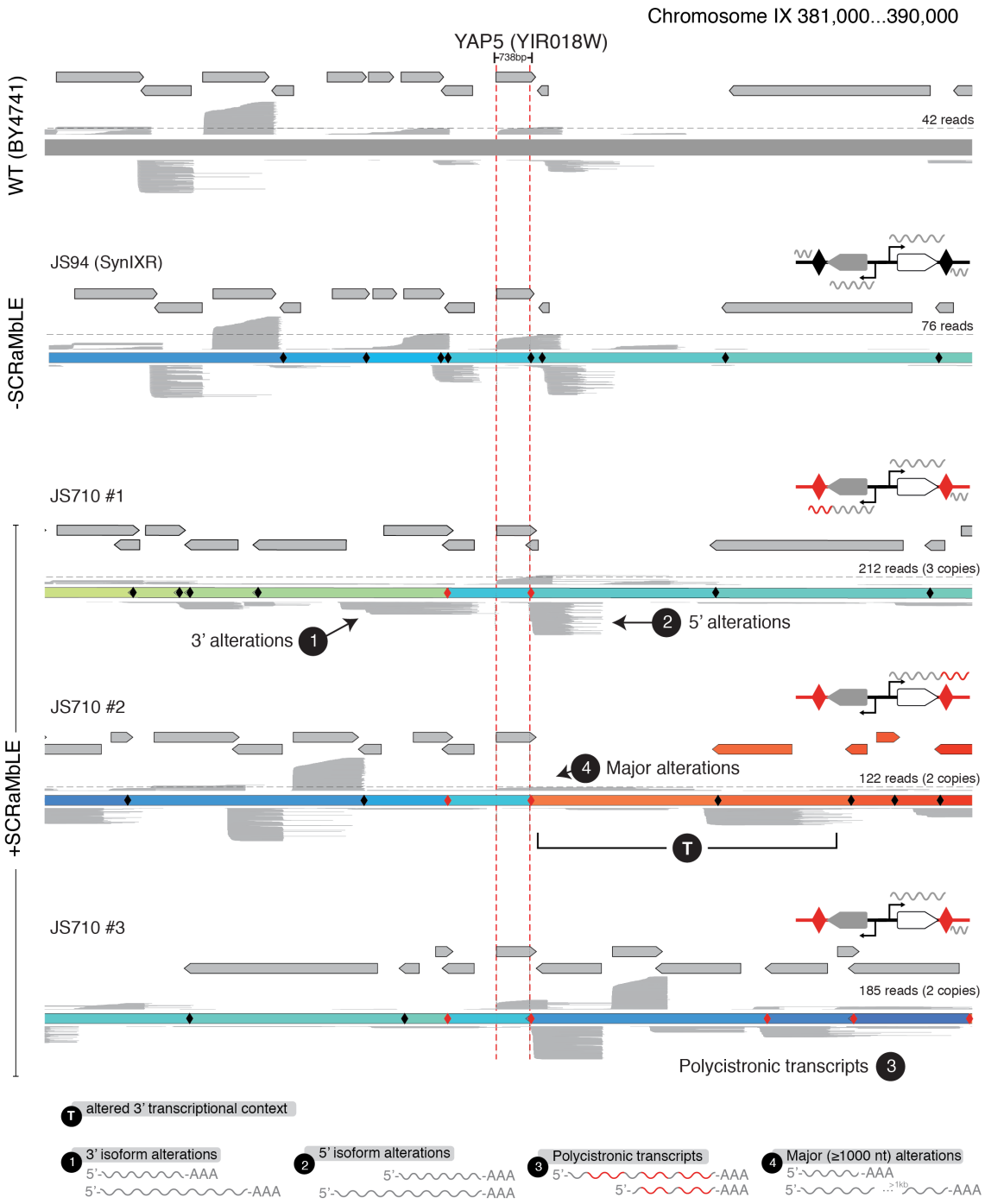

Fig. S8. Rearrangements lead to alterations of transcript isoforms.
Single-molecule direct RNA reads aligned to the genomic region around *YAP5* (*YIR018W*) in two reference strains (BY4741 and parental -SCRaMbLE strain, JS94) and three contexts in a highly-rearranged +SCRaMbLE strain (JS710). Each genomic region is aligned to the *YIR018W* coding sequence start/stop, denoted by red dashed lines. Rearrangements in SCRaMbLE strains are illustrated by color, which refers to original genomic location in the -SCRaMbLE strain as in (*16*). Transcript long-reads (thin gray lines) from the + and - strands are displayed above and below the segment track, respectively. Gene models are shown above. Cartoons on the right indicate the novel junctions and changes in proximal transcriptional environments (up- and downstream on either strand) with red diamonds and wavy lines, respectively. For the +SCRaMbLE strain, JS710, there are three unique genetic configurations of *YIR018W*, each existing in multiple copies in this genome (total copy number 7). The number of

copies and number of reads mapping to each unique rearrangement are indicated above the segment track. One rearrangement (orange segment) substantially changes the downstream transcriptional environment (marked as 'T') and leads to a major alteration (3′-UTR extension; marked as '4') in the *YIR018W* transcript isoform. In the remaining two contexts (first and third +SCRaMbLE rows), the downstream transcriptional environment is maintained, as is the *YIR018W* transcript isoform profile. Representations of the types of transcriptional alterations observed in this region are depicted below.

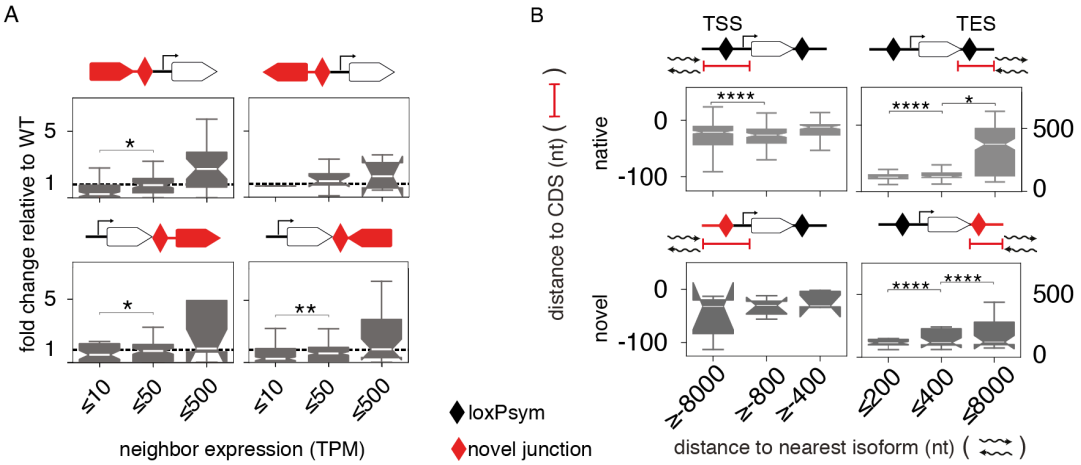

Fig. S9. Observations support individual associations from machine learning.
(A) Expression of neighboring genes affects the fold change in gene expression in rearranged contexts. Expression fold change relative to wildtype is shown for different levels of neighboring expression. Each panel shows the fold change in TU expression separately for rearrangements affecting expression of neighboring transcripts in either position and orientation (red arrows). (B) TSS and TES distances from the CDS tend to increase as neighboring transcription becomes more distant. Variation in TSS and TES distance to the CDS as a function of distance to the nearest transcript isoform in rearranged (lower) and native (upper) contexts. Adjacent black and red diamonds indicate native junctions and rearrangements, respectively, and black wavy lines indicate adjacent transcripts on either strand. Data are represented as the median and interquartile range with whiskers extending ±1.5 times the interquartile range. Notches indicate 95% confidence intervals. For all statistical tests, only adjacent bins were tested. Asterisks denote significance level in Mann-Whitney U test, * $p \leq 0.05$, ** $p \leq 0.01$, **** $p \leq$ 1e-4.

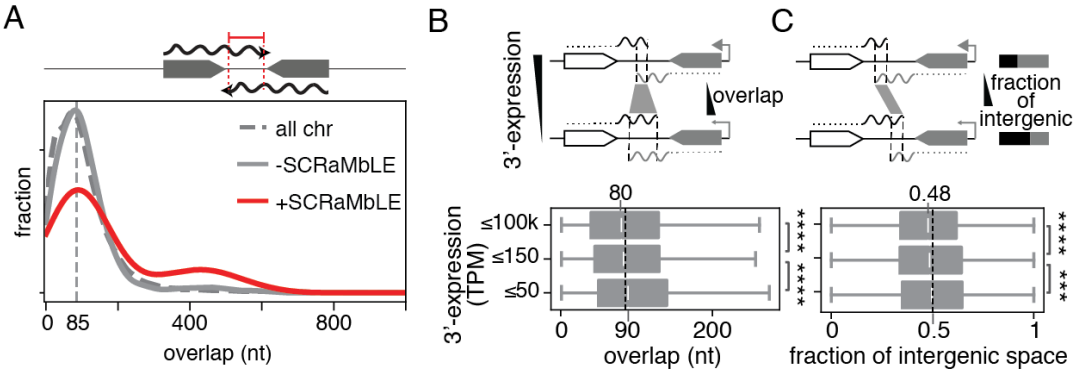

Fig. S10. 3′-UTR length of convergent transcripts is affected by intergenic distance and expression levels. (A) Transcripts arising from convergent gene pairs frequently overlap by 85 nt in the native and synthetic genomes. Distributions of the length of overlap between reads from convergent gene pairs on native chromosomes (dashed gray line), -SCRaMbLE synIXR (solid gray line), and +SCRaMbLE synIXR (red line). (B) Convergent transcripts tend to overlap to a greater extent as the downstream transcript decreases in expression. Read overlap lengths are shown as a function of downstream gene expression level. (C) Fraction of intergenic space occupied by the upstream transcript compared to the downstream transcript increases as the downstream convergent transcript decreases in expression. The fraction of the intergenic distance occupied by the upstream transcript is compared for different downstream gene expression levels. Data in (B) and (C) are represented as the median and interquartile range with whiskers extending ±1.5 times the interquartile range. Asterisks denote significance level in Mann-Whitney U test, *** $p \leq 0.001$, **** $p \leq$ 1e-4. For all statistical tests, only adjacent bins were tested.

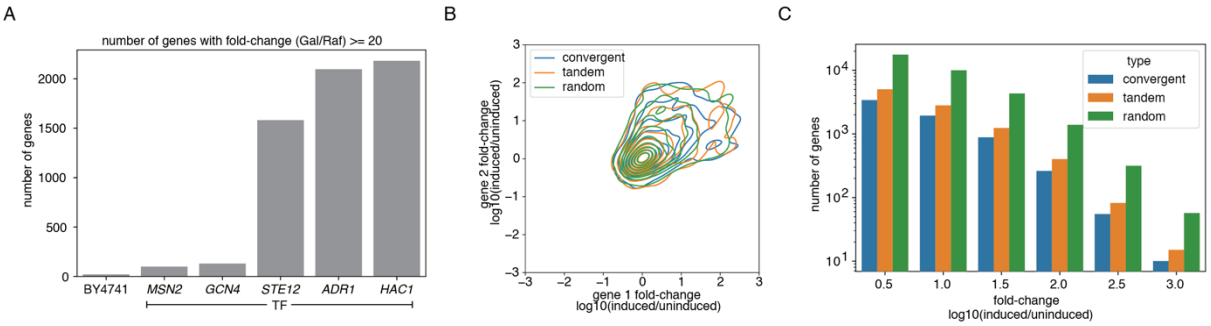

Fig. S11. Variable number of genes are overexpressed in each galactose TF overexpression (OE) strain. (A) Galactose-driven transcription factor (TF) overexpression (OE) in five TF OE strains increases expression of many target genes throughout the genome. Number of genes with a greater than 20-fold galactose-induced increase in expression in each TF OE strain and the wildtype control, BY4741. (B) Gene expression fold change correlations of gene pairs are insensitive to gene orientation. Fold change in expression between galactose (induced) and raffinose (uninduced) conditions are plotted for convergent, tandem, and random gene pairs. (C) Across all strains, thousands of genes in convergent and tandem orientations are represented among overexpressed genes. The number of genes that are overexpressed are shown for different levels of overexpression. Orientation of gene pairs are distinguished by color. Convergent and tandem gene pairs are defined within 1000 bp.

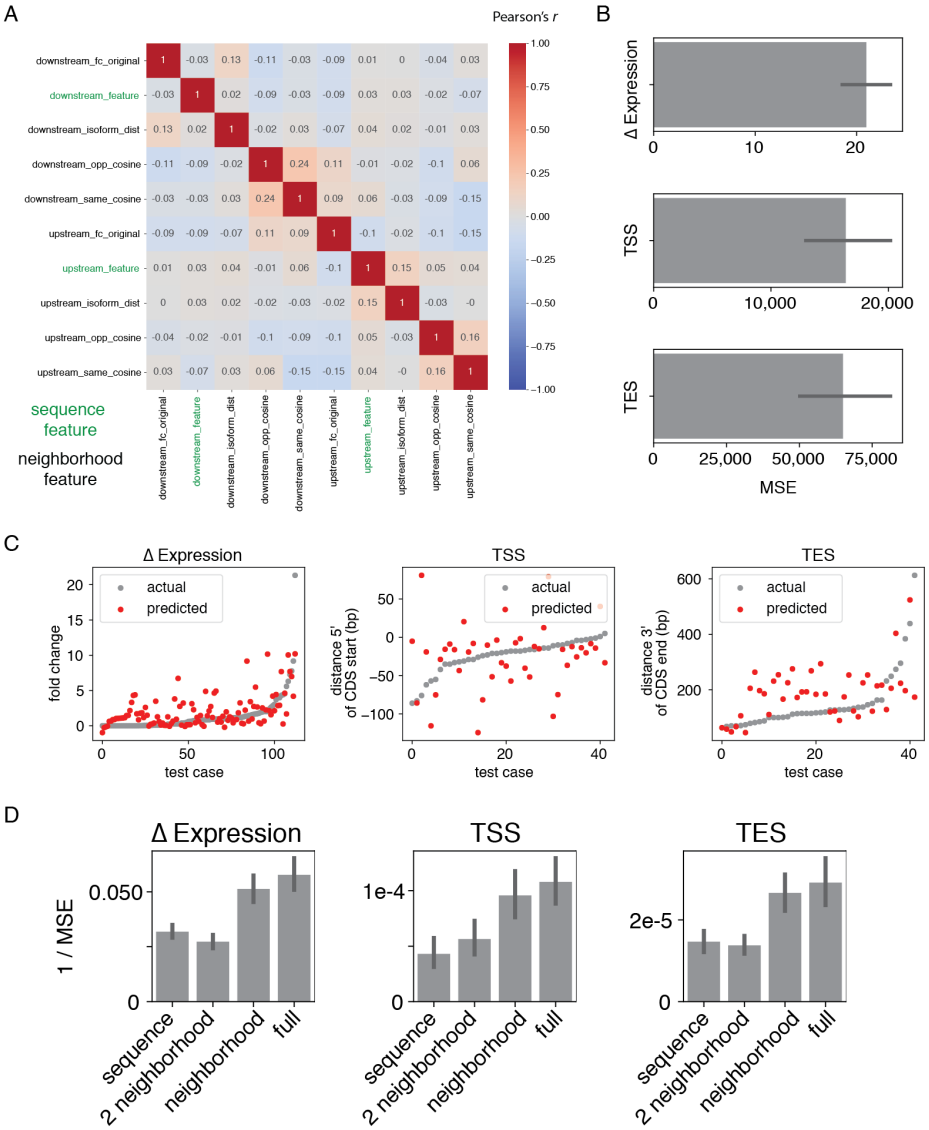

Fig. S12. Gradient Boosted Regression Tree (GBRT) models including sequence and transcriptional neighborhood information predict TSS, TES, and expression level.
(A) Features used to train GBRT models are distinct. Correlations between all features used to train the GBRTs. (B) Mean squared error (MSE) of GBRT models predicting expression change, TSS, or TES. (C) Predictions follow the same general trend as observed data for TSS (middle) and TES (right) distances from coding regions and expression fold change (left). Actual data compared to model predictions from the best model for each prediction task. (D) Models trained with sequence and transcriptional neighborhood features ('Full') and all transcriptional neighborhood features ('Neighborhood') have greater predictive performance (1/MSE, MSE: mean squared error) than those trained with sequence features only ('Sequence') or only two transcriptional neighborhood features ('2 Neighborhood'). Error bars in (B) and (D) indicate 95% confidence interval estimated from 10X, 5-fold cross-validation.

Table S1. (separate file)

Growth of synIXR SCRaMBLE strains reported as mean doubling time ± standard deviation alongside the total number of novel junctions in the strain

Table S2. (separate file)

Transcript isoforms supported by 2 or more reads mapping within 25 nt at both ends, native chromosomes

Table S3. (separate file)

Transcript isoforms supported by 2 or more reads mapping within 25 nt at both ends, synthetic chromosomes

Table S4. (separate file)

Genes that lose or gain detectable expression in SCRaMbLE strains, reporting copy number, expression level, and rearrangement status.

tTA.trp1	*TATTGAGCACGTGAGTATACGTGATTAAGCACACAAAGGCAGCTTGGAGT*C CACTAGTGGATCTGATATC
tTA.trp2	*GCAAGTGCACAAACAATACTTAAATAAATACTACTCAGTAATAACCTATTCC GCCAGCT*GAAGCTTATTA
YIR018C-AtetFe	*TACGCGCGCAGAGAGTCTGGATCGAACCTGACGGAGATATATGTCAGATC*C GTACGCTGCAGGTCGACGG
YIR018C-AtetRc	*GCGTCAGAGATCTCGAATGACAGATCGGTGGCTACCACGAAATCTTATAG*C ATAGGCCACTAGTGGATCTG

Table S5.
Primers for constructing tetO7 promoter strains.

2SP+UMI_YIR017C	TTTCTGTTGGTGCTGATATTGCTAGTTGTNNNNNNNNNNATGAGTGCGAAACAAGGG
2SP+UMI_YIR016W	TTTCTGTTGGTGCTGATATTGCTAGTTGTNNNNNNNNNNGTAATGAGTGGCACGAGG
2SP+UMI_YIR018W	TTTCTGTTGGTGCTGATATTGCTAGTTGTNNNNNNNNNNGTCATGGCTCTACCTCTG
2SP+UMI_YIR018C-A	TTTCTGTTGGTGCTGATATTGCTAGTTGTNNNNNNNNNNTGATTATACTTCCCATTACCCA
2SP+UMI_YIR018C-Aup	TTTCTGTTGGTGCTGATATTGCTAGTTGTNNNNNNNNNNTGCGTAACGTCTAACTACAG
2SP+UMI_YIR019C	TTTCTGTTGGTGCTGATATTGCTAGTTGTNNNNNNNNNNCACACTATGCAAAGACCA

Table S6.
Primers for gene-specific cDNA sequencing.

YIR018Wq3R_BC02	TCGATTCCGTTTGTAGTCGTCTGTGTGGATGATGGACCGGATGT
YIR018C-Aq2R_BC03	GAGTCTTGTGTCCCAGTTACCAGGGTTCATGGACATTGGCCGC
qPCR_ACT1R_BC04	TTCGGATTCTATCGTGTTTCCCTACCAAGGCGACGTAACATAGTTTT
YIR018Wq3F	CGACAACACTACTTGCGTTTGT
qPCR_BC02	TCGATTCCGTTTGTAGTCGTCTGT
YIR018C-Aq2F	TTGCAGGAATGTATAGGCATAGT
qPCR_BC03	GAGTCTTGTGTCCCAGTTACCAGG
qPCR_ACT1_fwd	CTCCACCACTGCTGAAAGAGAA
qPCR_BC04	TTCGGATTCTATCGTGTTTCCCTA

Table S7.
Primers for RT-qPCR.