STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
ANTI-FLAG M2 affinity gel	Sigma-Aldrich	Cat# A2220; RRID: AB_10063035
Rabbit polyclonal anti-Sua7 (TFIIB)	Gift from Joseph Geisberg, Harvard Medical School	N/A
Chemicals, Peptides, and Recombinant Proteins		
3X FLAG Peptide	Sigma-Aldrich	Cat# F4799
Deposited Data		
Raw sequencing data	This paper	GEO: GSE87735
Experimental Models: Organisms/Strains		
S. cerevisiae: (Strain background: AB1380) MATa ura3-52 trp1-289 lys2-1 ade2-1 can1-100 his5 ρ+ ψ+ *RPB3-3XFLAG::NAT*	This paper	N/A
K. lactis: (Strain background: CLIB209) *KLLA0D16170 g-3XFLAG::NAT*	This paper	N/A
D. hansenii: (Strain background: NCYC2572) *DEHA2C07546 g-3XFLAG::NAT*	This paper	N/A
YAC1_1: (Strain background: AB1380) MATa ura3-52 trp1-289 lys2-1 ade2-1 can1-100 his5 ρ+ ψ+ *RPB3-3XFLAG::NAT* 128 kb YAC [*K. lactis* chromosome F 872404∼1000550]	This paper	N/A
YAC2_1: Strain background: AB1380) MATa ura3-52 trp1-289 lys2-1 ade2-1 can1-100 his5 ρ+ ψ+ *RPB3-3XFLAG::NAT* 143 kb YAC [*K. lactis* chromosome C 339713∼482935]	This paper	N/A
YAC3_1: Strain background: AB1380) MATa ura3-52 trp1-289 lys2-1 ade2-1 can1-100 his5 ρ+ ψ+ *RPB3-3XFLAG::NAT* 136 kb YAC [*K. lactis* chromosome C 443175∼578764]	This paper	N/A
YAC6_1: Strain background: AB1380) MATa ura3-52 trp1-289 lys2-1 ade2-1 can1-100 his5 ρ+ ψ+ *RPB3-3XFLAG::NAT* 115 kb YAC [*D. hansenii* chromosome C 1165392∼1280355]	This paper	N/A
YAC7_1: Strain background: AB1380) MATa ura3-52 trp1-289 lys2-1 ade2-1 can1-100 his5 ρ+ ψ+ *RPB3-3XFLAG::NAT* 216 kb YAC [*D. hansenii* chromosome D 1148162∼1364529]	This paper	N/A
Software and Algorithms		
FIDDLE	Eser and Churchman, 2016	N/A
HMMER	Wheeler and Eddy, 2013	N/A
Tophat2	Kim et al., 2013	N/A
Bowtie2	Langmead and Salzberg, 2012	N/A
Prinseq	Schmieder and Edwards, 2011	N/A
Cutadapt	Martin, 2011	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, L. Stirling Churchman (churchman@genetics.med.harvard.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strains

Yeast strains used in this study were listed in Key Resources Table. Rpb3 of strains subjected to NET-seq analysis was epitope-tagged at C terminus with 3X-Flag tag and expressed from its endogenous locus. In order to accommodate alternative codon usage

in *D. hansenii* (*i.e.*, CUG translated as serine, instead of leucine) (Moura et al., 2007), codons of the epitope tag and selection marker were modified accordingly. To promote higher efficiency of gene targeting through homologous recombination in *K. lactis* and *D. hansenii*, extra long homologous regions adjacent to the target site (up to 1000 bp) were used to flank the epitope tag and selection marker. Electroporation based transformation method were also optimized to achieve high-efficiency transformation in *K. lactis* and *D. hansenii*. Further details of strain construction are available upon request.

Growth Conditions

K. lactis and *D. hansenii* were grown in custom medium containing: SC Amino Acid mix (Sunrise Science) (0.2%), Yeast extract (1.5%), Peptone (1%), Dextrose (2%), Adenine (0.01%), Uracil (0.01%), and Tryptophan (0.01%), as previously described (Tsankov et al., 2010). All YAC containing culture was grown in slightly modified medium containing: SC –Tryptophan –Uracil mix (Sunrise Sciences) (0.2%), Yeast extract (1.5%), Peptone (1%), Dextrose (2%), and Adenine (0.01%) (Hughes et al., 2012). All yeast culture was grown at 30°C, except that *D. hansenii* was cultured at 28°C.

METHOD DETAILS

NET-Seq

NET-seq libraries were constructed and sequenced as previously described with minor modifications (Churchman and Weissman, 2012). Briefly, log phase yeast culture (OD_{600} = 0.6-0.8) were harvested by filtration and flash frozen using liquid nitrogen. Frozen cells were lysed via pulverization using mixer mill. Nascent RNA was purified from immunoprecipitated RNA polymerase (precipitated using ANTI-FLAG M2 affinity gel and eluted with 3X FLAG peptide), followed by library construction. An improved version of NET-seq DNA linker featured 6 random nucleotides at 5′ terminus was used to further increase ligation efficiency, as well as minimizing ligation bias and library amplification bias (Harlen et al., 2016; Mayer and Churchman, 2016). 3′ end sequencing of NET-seq libraries was performed on Illumina sequencing platform.

RNA-Seq

Total RNA from yeast culture grown to mid-log phase was isolated using standard hot phenol-chloroform extraction protocol. mRNA was purified and fragmented, followed by cDNA synthesis. Library construction was carried out as previously described (Wong et al., 2001).

ChIP-Seq

Sheared chromatin from mid-log phase yeast culture was prepared (Fan et al., 2010). Chromatin immunoprecipitation was conducted using antibody against TFIIB. Barcoded sequencing libraries from ChIP DNA were constructed (Wong et al., 2001).

QUANTIFICATION AND STATISTICAL ANALYSIS

Processing and Alignment of Sequencing Reads

To remove adaptor sequences from NET-seq fastq files, we used cutadapt (Martin, 2011). Remaining fastq files were further cleaned by Prinseq (Schmieder and Edwards, 2011). We then aligned the remaining sequences to sacCer3 genome using Bowtie2 and Tophat2 (Kim et al., 2013; Langmead and Salzberg, 2012). Only the positions matching the 5′ end of the sequencing reads corresponding to the 3′ end of the nascent RNA fragments were recorded. Reads that align to the same genomic position and contain identical barcodes are considered PCR duplication events and are removed.

TSS detection for native and YAC species

To detect the transcription start sites of the genes in *S. cerevisiae*, we trained our deep learning model, FIDDLE (Eser and Churchman, 2016), by providing inputs from DNA sequence, NET-seq (this study), MNase-seq (Hughes et al., 2012), RNA-seq (this study) and TFIIB ChIP-seq (this study) data and the target from TSS-seq data (Malabat et al., 2015). After successfully training the model, we input the region that spans 1 kb upstream of the coding start site to predict where the TSSs are for native and other YAC species. The output of the model is a probability distribution which peaks around the TSS.

Determining nucleosome depleted regions within the coding sequence of D. hansenii YACs

After smoothing the MNase-seq data with 50bp windows, we detected the peaks that are higher than the 10% of the maximum peak value found within the coding sequence. Then, we selected the regions that are located between the detected peaks and have at least 250bp peak-to-peak distance.

Directionality score calculation

After annotating the TSS for the *S. cerevisiae* genome, we first removed the overlapping genes, then selected the promoter regions of tandemly oriented genes, where divergent transcription is non-coding. For the aggregate plot, we then calculated the transcriptional activity within ± 50bp region around each nucleotide by taking 10% trimmed mean of the NET-seq reads which contain outliers due to

the Pol II pausing. Then the profiles are calculated by aligning the tandem genes to their TSS and recording the average number of reads for all positions 1kb upstream and downstream of TSS.

To quantify the coding and divergent non-coding transcriptional activity, we took 500 bp window upstream antisense and downstream sense of the TSS and recorded the maximum window-averaged values for coding and divergent transcription, respectively. We selected the promoter regions who have a signal greater than 0.1 (at least 5 reads are expected within 50bp-averaging window) in any directions. Then the directionality scores of promoter regions are calculated by taking the $\log_{10}$ ratio of these coding and divergent transcriptional activities for those who have signal in both directions. Otherwise, they are called sense transcription and antisense transcription if they lack antisense and sense signal, respectively.

Evolutionary rate profiling

GERP score quantifies the evolutionary rate of a specified position in the genome (Cooper et al., 2005a). We calculated the average GERP score for 500 bp upstream and downstream of the TSS for both directional and bidirectional tandem promoter regions using sequence alignment from seven *Saccharomyces* species (Siepel et al., 2005). We performed Kolmogorov-Smirnov test to compare the distributions of average GERP scores over directional and bidirectional promoters, *i.e.*, 500 bp upstream of TSS.

Transcription factor enrichment

We used FIMO scanning to determine the transcription factor binding sites (Grant et al., 2011). The PWM of the binding motifs are obtained from YEASTRACT (Teixeira et al., 2014). To calculate the statistics of differential enrichment, we assumed that for a given nucleotide in the genome, the probability of finding the mid-point of the specific transcription factor binding site is a Bernoulli process with a very low probability, $p << 1$. Then the probability of finding k TF binding site within large regions (tens of kb) can be approximated by Poisson process with a point mass function:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ is the average number of TF binding sites expected under the null hypothesis and k is the number of binding sites that are observed. Therefore, the probability of observing at least k_0 number of TF binding sites approximates the p value which is given by

$$p(k > k_0) = 1 - CDF$$

$$p(k > k_0) = 1 - \frac{\Gamma(k_0 + 1, \lambda)}{k_0!}$$

where Γ is the upper incomplete gamma-function. We used Scipy stats module in Python to calculate empirical cumulative density function. The chance of having a Type-I error for testing a family-wise hypothesis increases by the number of transcription factors. On the other hand, adjusting p-values for multiple hypotheses increases the change of having Type-II error. Therefore, we report both unadjusted and Bonferroni adjusted p-values (Table S3). We used statsmodels package to correct for multiple hypotheses. Note that Reb1 and Abf1 show significant enrichment in both cases.

Evolutionary retention of fortuitous promoter regions

We take the 200 bp upstream and downstream regions of all *S. cerevisiae* genes that are not overlapping with each other. Then we used the command line tool, Hmmer (Wheeler and Eddy, 2013), with the default options to search for the matches that both upstream and downstream of *S. cerevisiae* TSS within the same coding sequences of other 22 yeast species, obtained from Broad orthogroup repository (Wapinski et al., 2007). We then selected those that satisfies the correct order, *i.e.*, upstream sequence has to match upstream of the position where downstream sequence matched, minimum 100 bp distance between these matches and the maximum E-value of 0.1. Our analysis only reveals the lower limit of matches as we do not consider the 5′UTR and have margins around 200 bp from coding sequence start and end sites for the 22 target species.

Evolutionary purging of fortuitous promoter regions

We used FIMO (Grant et al., 2011) to scan the coding sequences of all other yeast species for TF binding sites whose motifs belong to *S. cerevisiae* and are obtained from YEASTRACT (Teixeira et al., 2014). Then we calculated the number of hits divided by the CDS length for each gene and averaged across the genome for each species. We then aggregated the average TF binding site density at the CDS of yeast species that diverges from the same first order branching point relative to *S. cerevisiae.* We performed two-sample Poisson intensity test (Gu et al., 2008) by comparing TF binding densities on coding sequences found in species belonging to the specified branching point to *S. cerevisiae*.

Directionality change boxplots for individual transcription factors

Directionality change for YAC promoters is calculated by subtracting the directionality score in their native environment from the one in *S. cerevisiae*. Next, the promoters are assigned to transcription factors if FIMO scanning (Grant et al., 2011) results in at least one hit

for the transcription factor motifs. Then we plotted the boxplot of directionality changes for each transcription factor, for both *K. lactis* and *D. hansenii* YACs.

Motif match score

We take consensus TATA box motifs from (Basehoar et al., 2004) and formed position weight matrix (PWM) and convolved the motif along the promoters.

Discriminative motif match analysis

We calculated motif match score for the transcription factor motifs from the YEASTRACT database (Teixeira et al., 2014). Then we recorded the maximum motif match score for each promoter region. We selected the transcription factors whose maximum motif match score distributions for directional and bidirectional promoter regions are significantly different (KS-test $p < 0.05$).

DATA AND SOFTWARE AVAILABILITY

All data are deposited in Gene Expression Omnibus under accession number GSE87735.

Cell

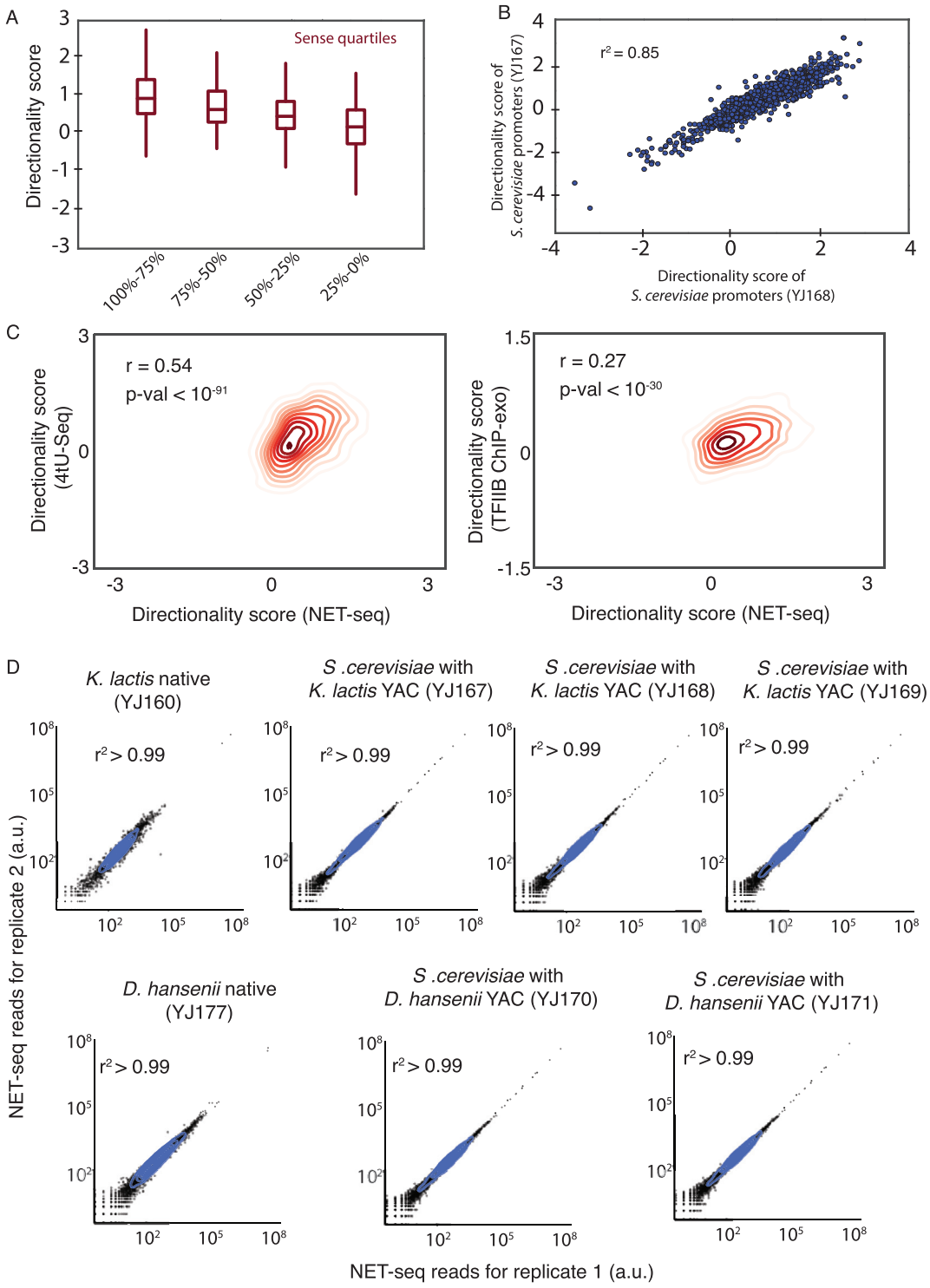

Figure S1. Directionality analysis of promoter regions in *S. cerevisiae*, Related to Figures 1 and 2

(A) Boxplots show the distribution of directionality scores for different quartiles of sense transcription.

(B) Scatterplot shows the directionality scores of *S. cerevisiae* endogenous promoter regions calculated in two different yeast strains, YJ167 and YJ168, containing two different YACs from *K. lactis*. Pearson correlation scores are displayed.

(C) Contour plots show the density of directionality scores calculated from 4tU-seq (Schulz et al., 2013) and TFIIB ChIP-exo (Rhee and Pugh, 2011) compared to the directionality score from NET-seq. Pearson correlation scores and the corresponding p values are shown inset.

(D) Scatterplot of total NET-seq reads of every gene for two replicate experiments. Pearson correlation scores are displayed.

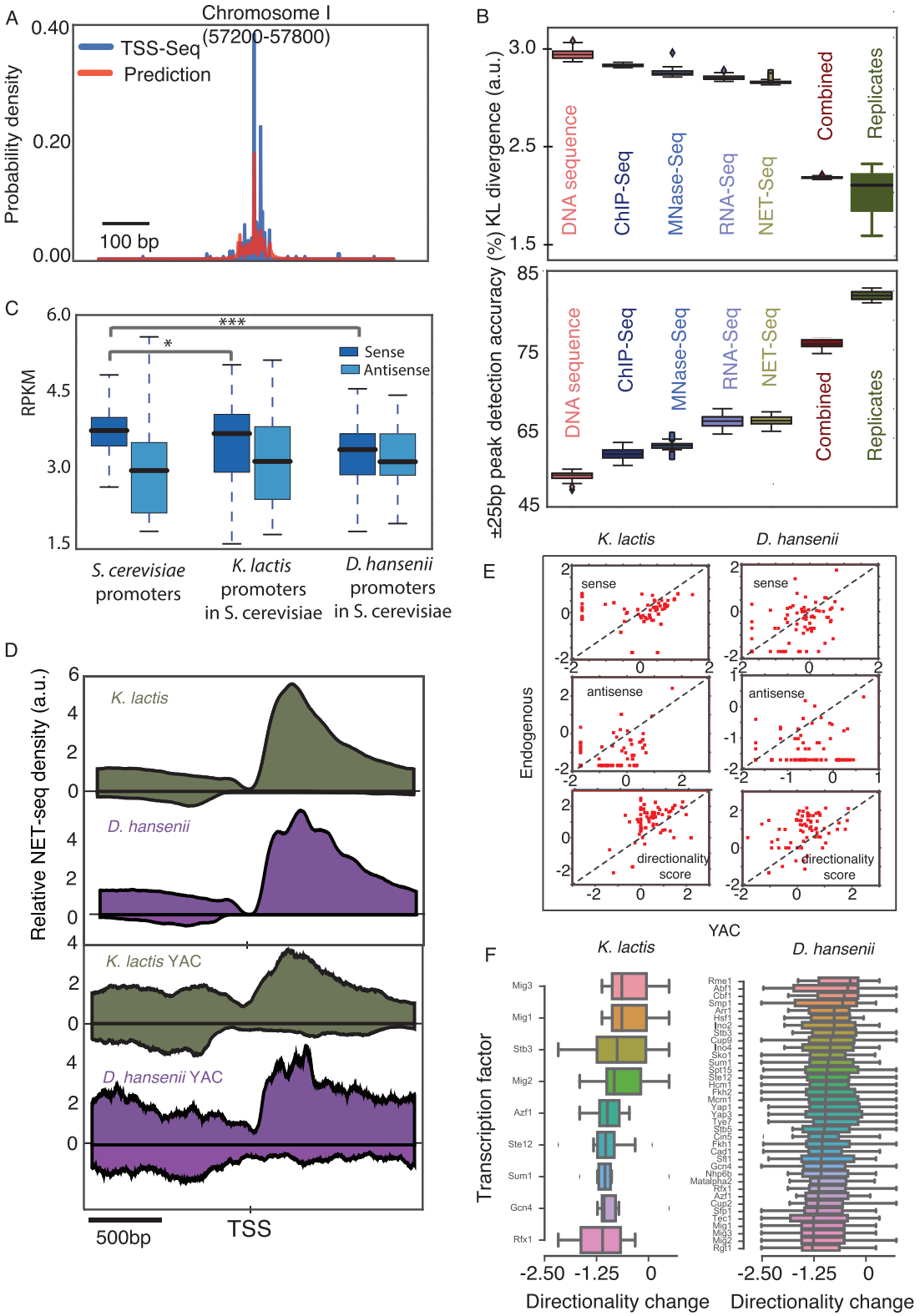

Figure S2. Directionality of Promoter Regions in a Foreign Environment, Related to Figure 2

(A) Examples of a TSS prediction using FIDDLE showing the model prediction (red) and the experimental data (gray) for TSS location (Eser and Churchman, 2016). The model is trained by using TSS-seq data from Malabat et al. as supervised data (Malabat et al., 2015). FIDDLE does not require data pre-processing, which is

(legend continued on next page)

highly common in standard methods such as peak detection, feature selection, dimensionality reduction etc. After training the model with *S. cerevisiae* TSS-seq data, we then transfer the model to *D. hansenii* and *K. lactis* as well as *S. cerevisiae* YAC containing strains.

(B) FIDDLE performance: Summary statistics of the KL-divergence and the TSS prediction accuracy of models trained using individual datasets (DNA sequence, ChIP-seq etc) or all datasets together (Combined). The predictive value of a biological replicate dataset is shown for comparison as it represents the intrinsic variability of the method (Eser and Churchman, 2016). Error bars show the standard error of the mean for the convergence of 6 independent models.

(C) Boxplots show the distribution of sense and antisense transcription which is normalized by the library size and multiplied by one million. Asterisks denote the statistical significance level: *p value < 0.05, ***p value < 0.0005 by KS test. Error bars show the standard error of the mean for all tandemly oriented genes.

(D) Metagene view of aggregated NET-seq reads by aligning genes to their transcription start sites (TSS) for native *K. lactis* and *D. hansenii* species (left) and the corresponding YAC containing *S. cerevisiae* strains (right).

(E) Sense, antisense and directionality scores of YAC promoters are plotted for endogenous (y-axes) and YAC containing *S. cerevisiae* (x-axes) environments. 10-based logarithmic values are shown on the axes.

(F) Boxplots show the distribution of changes in directionality score for the promoter regions that are enriched for motifs of certain transcription factors for *K. lactis* (left) and *D. hansenii* (right) YACs. None of them are significantly different than the overall changes in directionality score. Error bars show the standard error of the mean for the set of promoter regions with a particular transcription factor binding motif.

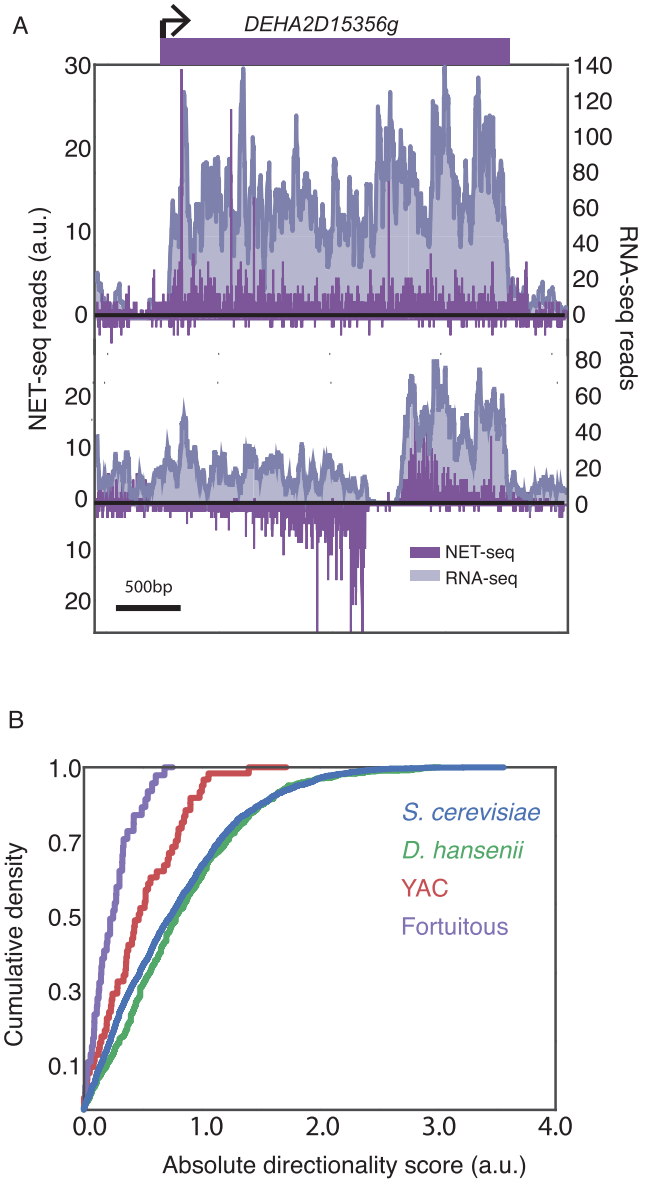

Figure S3. Bidirectional Transcription Occurs from Fortuitous Promoter Regions, Related to Figure 3

(A) Example of a fortuitous promoter region emerging within the coding sequence of a *D. hansenii* gene, DEHA2D15365 g. Gray shows the RNA-seq (unstranded) data.

(B) Cumulative density plots show the absolute value of directionality score distributions for *S. cerevisiae, D. hansenii, D. hansenii* YAC and fortuitous promoters. Distributions of the fortuitous promoters are significantly different than that of YACs (p value $< 10^{-5}$ by KS test) and *D. hansenii* (p value $< 10^{-12}$ by KS test).

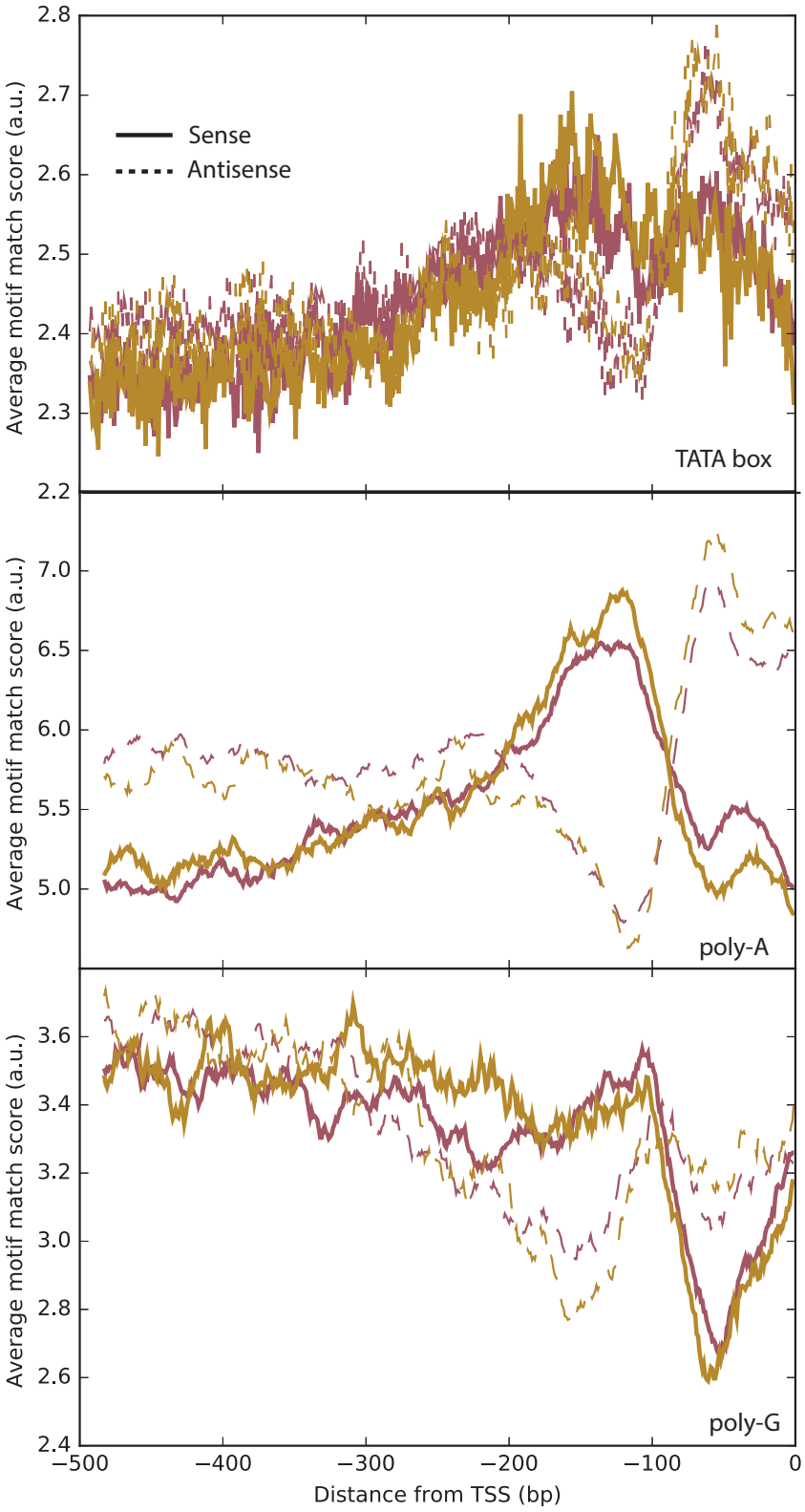

Figure S4. Motif Analysis of Directional and Bidirectional Promoter Regions in *S. cerevisiae*, Related to Figure 4

Average motif match scores for directional (red) and bidirectional (yellow) promoters in *S. cerevisiae* are shown for conservative TATA-box (top), poly-A (middle) and poly-G (bottom) motifs.

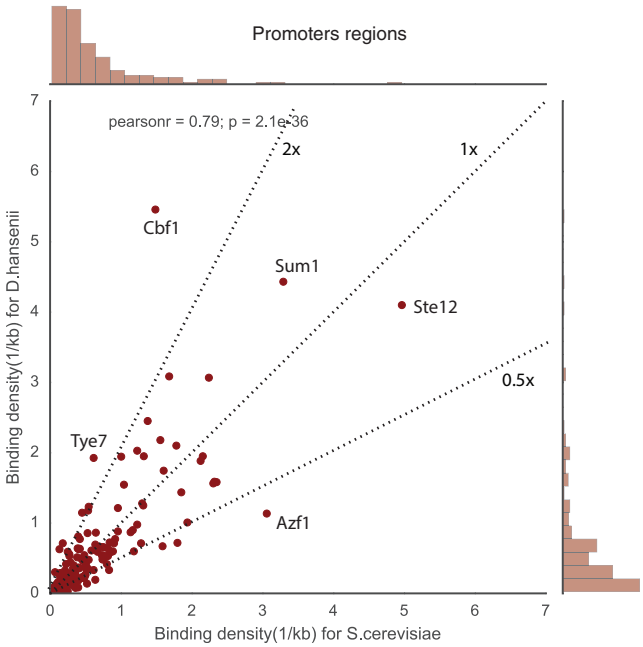

Figure S5. Density Comparison of Transcription Factor Binding Sites between *D. hansenii* and *S. cerevisiae*, Related to Figure 3

Scatterplot shows the binding site density for the transcription factors at promoter regions of *D. hansenii* and *S. cerevisiae,* calculated by FIMO scanning using YEASTRACT motifs. Dotted lines denote the 2-fold, equal and one half binding site ratio of *D. hansenii* over *S. cerevisiae*. Histograms at the top and the right are of the TF binding site densities for *S. cerevisiae* and *D. hansenii*, respectively. Scatterplot data displayed in Table S3.