SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1: Resolution of the classifiers and of cellular transcription. Each panel shows a heatmap illustrating the % usage of each base as a TSS/CPA site across the aligned region for each gene, sorted by average TSS/CPA site position (weighted by the percent usage). Overlaid, is the average % usage across the region for all genes (i.e. the average of the data in the heatmap). **(A, B)** The % usage of each base as a **(A)** TSS and **(B)** CPA site, aligned to the most frequently used base. **(C, D)** The % usage of each base as a **(C)** TSS and **(D)** CPA site, aligned to the centre of the initiation and termination classifier peaks, respectively. **(E, F)** The % usage of each base as a **(E)** TSS and **(F)** CPA site, aligned to initiator and cleavage site motifs found within initiation and termination classifier peaks, respectively.

Supplementary Figure 2: Precision-Recall curves for minimal and complete classifiers. Data for the initiation classifier with all features, minimal initiation classifier, termination classifier with all features, and minimal termination classifier is shown.

Supplementary Figure 3: Feature selection and feature importance for the initiation and termination classifiers. (A, B) The change in error rate as each additional feature is included in the classifier, where the dashed lines indicate the feature inclusion threshold (3%) for each of the **(A)** initiation and **(B)** termination classifiers. Retained features are shown in bold. For each, only the top 40 features are shown as no further features attained the 3% threshold. Brackets indicate motif IDs when multiple motifs for the same factor are used and refer to **Supplementary Tables 2 & 3**. *The Stb2 motif appears to be a variant of the Reb1 Motif. **For the termination classifier, the minimal error rate is achieved upon inclusion of the fourth feature, so no more features were considered. **(C, D)** The ΔAUROC for models created without each feature type, relative to the minimal (reduced feature set) **(C)** initiation and **(D)** termination classifiers. **(E)** The important *trans*-acting factors for the initiation classifier tend to regulate distinct promoter subsets. The predicted binding of the *trans*-acting factors to all promoters. Predicted binding Z-scores are iteratively sorted in decreasing order until a Z-score of 1.5 is reached, for each feature.

Supplementary Figure 4: Initiation classifier can predict the genes that will be affected by *trans*-acting factor mutants. This plot shows the ROC curve for how genes that are predicted to be controlled by a factor are stratified from those that are not by the expression changes in the corresponding *trans*-acting factor mutant. Thus, data above the line Y=X indicates that the "predicted-controlled" genes have a higher expression level in the mutant, while data below this

line indicates these genes have a generally lower expression in the mutant. AUROC and rank sum P-values are as indicated. [1] (Badis et al. 2008) [2] (Alper et al. 2006).

Supplementary Figure 5: Structure and example predictions of RNA-seq-based transcript identifiers. (A) The structure of the HMM-R HMM which uses RNA-seq to identify transcript structure. "IG" represents the intergenic state. **(B)** The predictions of HMM-R, TSS-R, and CPA-R are depicted on a part of chromosome 1. Here, TSS-R scores for the forward and reverse strands are shown in light green, with CPA-R predictions immediately below. The two HMM-R tracks depict merged transcript state maps, where, when both agree, both sets include the transcript and where they conflict, as in the case of *FUN14* and *ERP2*, the gene on the forward strand is included in “top” and the reverse-strand gene in “bottom”. In the centre, gene annotations are shown in dark blue, with thinner bars representing the UTRs.

Supplementary Figure 6: Precision-Recall curve for the UM.

Supplementary Figure 7: Novel predicted transcripts are often transcribed. (A) dUTP RNA-seq data from (Levin et al. 2010), aligned across 1,286 transcripts predicted on chromosomes 1-8 that do not correspond to known transcripts or other genomic features. **(B)** Average across the predicted transcripts for both the RNA-seq data shown in **(A)**, and NET-seq data from (Churchman and Weissman 2011), smoothed using a 10 bp sliding window. Surrounding bases (upstream and downstream of predicted transcripts) are frequently known genes and tend to be more highly expressed than the predicted transcripts, and so have a greater read count and PolII occupancy.

Supplementary Figure 8: Predicted transcript structure and measured expression of the other two randomly-generated 6kb fragments. Tracks and labelling as in **Figure 6**.

Supplementary Figure 9: Context dependence of promoter identity. Several examples of promoters occurring within gene bodies that appear to be repressed by normal transcription of the gene and de-repressed when the gene is no longer expressed in a *trans*-acting factor mutant, for **(A)** Rsc3, **(B)** Abf1, and **(C)** Rap1 mutants (Badis et al. 2008). In each case, three tracks are shown for both strands, ORFs in dark green, expression level, shown relative to wild type, in black, and the initiation score in green. The blue, red, and purple bars in the center represent the binding sites for Rsc3, Abf1, and Rap1 (respectively) in the promoter regions of each gene. De-repressed transcripts are boxed.

Supplementary Figure 10: Predictions of models across non-traditional transcripts. Each graph shows the two classifier scores, as well as the UM predictions averaged across the aligned transcripts for **(A)** CUTs and **(B)** SUTs (Xu et al. 2009), and **(C)** antisense transcripts (Yassour et al. 2010).

SUPPLEMENTARY TABLES

Supplementary Table 1: Known features of promoters and their occurrence in the initiation classifier bins.

Feature	Reference	-500 :-200	-200 :-150	-150 :-80	-80 :-50	-50 :TSS	TSS :+100
TATA-box	(Erb and van Nimwegen 2011)				X		
Most TFs (e.g. Reb1, Cbf1, Abf1)	(Lee et al. 2007; Erb and van Nimwegen 2011)			X			
Some TFs (e.g. Pbf1/2, Mcm1)	(Erb and van Nimwegen 2011)	X	X		X		
+1 Nucleosome	(Lee et al. 2007)						X
NFR (most genes)	(Lee et al. 2007)		X	X	X	X	
NFR (TATA genes)	(Erb and van Nimwegen 2011)	X	X	X	X	X	
G/C content/ DNA-structural differences	(Lee et al. 2007)		X	X	X	X	X

Supplementary Tables 2 and 3 are provided in a separate Excel file.

Supplementary Table 4: UM transition probabilities, representing the probability of transitioning from source states (first column) to destination states (other columns).

Source	IG	Gene+	TSS+	CPA+	CPA-	Gene-	TSS-	CPA+/-
IG	0.99777	0	0.001246	0	0.000986	0	0	1.19E-06
Gene+	0	0.999	0	0.000865	0	0	0	0.00013796
TSS+	0	0.047471	0.9525	0	0	0	0	2.42E-05
CPA+	0.025805	0	0.002395	0.96983	5.35E-05	0	0	0.0019219
CPA-	0	0	0	0	0.96983	0.030175	0	0
Gene-	0	0	0	0	0	0.999	0.001003	0
TSS-	0.04424	0	0	0	0.003249	0	0.9525	6.05E-06
CPA+/-	6.39E-05	0	9.13E-06	0	0.003934	0.009849	3.65E-05	0.98611

Supplementary Table 5: Final UM observation distributions. These represent means and variances of Box-Cox-transformed data (using lambdas of 0.050942 for initiation and -0.006859 for termination), which are normally distributed ($N(\mu,\sigma^2)$) for each state.

State	initiation+	termination+	initiation-	termination-
IG	N(-3.305, 2.8073)	N(-6.2889, 6.6776)	N(-3.305, 2.8073)	N(-6.2889, 6.6776)
Gene+	N(-3.655, 2.3565)	N(-7.2694, 4.4256)	N(-3.8655, 2.3784)	N(-7.1731, 5.0172)
TSS+	N(-0.4001, 2.0736)	N(-4.5387, 5.0823)	N(-4.4165, 2.5278)	N(-6.0073, 4.6229)
CPA+	N(-2.5527, 2.4549)	N(-0.63571, 7.4256)	N(-2.6374, 2.6348)	N(-1.9809, 6.6425)
CPA-	N(-2.6374, 2.6348)	N(-1.9809, 6.6425)	N(-2.5527, 2.4549)	N(-0.63571, 7.4256)
Gene-	N(-3.8655, 2.3784)	N(-7.1731, 5.0172)	N(-3.655, 2.3565)	N(-7.2694, 4.4256)
TSS-	N(-4.4165, 2.5278)	N(-6.0073, 4.6229)	N(-0.4001, 2.0736)	N(-4.5387, 5.0823)
CPA+/-	N(-3.0298, 2.1931)	N(-0.084816, 4.8119)	N(-3.0298, 2.1931)	N(-0.084816, 4.8119)

SUPPLEMENTARY METHODS AND ANALYSIS

Creation and Analysis of Initiation and Termination Classifiers

Transcript Annotations

We created a new map of yeast transcripts by first manually identifying transcript boundaries based on a combination of gene annotations, RNA-seq and tiling array expression data (David et al. 2006; Nagalakshmi et al. 2008; van Bakel et al. 2013). This gave us approximate transcript boundaries which we further refined using available RNA-seq data. We automatically searched for the best representative TSS within 100 bp of the manually annotated start using RNA-seq data in which reads cluster at the TSS (Lipson et al. 2009). Using these data, we identified the site of the greatest increase in log(read-density) as the best representative TSS, where the number of $\log_{10}$(reads) had to be at least 1.5. We defined CPA sites as the position with the highest read density of poly-A reads from (Nagalakshmi et al. 2008) within 300 bp of the manually-curated transcript end, and we required that this site follow the stop codon. This procedure produced transcript TSS and CPA site annotations for 5,010/5,772 of the non-dubious ORFs.

Positive and Negative Promoter and CPA regions

We took the promoter region as the sequence from -500 to +100 relative to the TSS, as defined above (see **Supplementary Table 1** and below for justification of this definition). The "positive" promoter set consisted of the 5,010 promoters from our transcript annotations. To generate "negative" promoter examples, we first identified all TSS regions, ORF starts, rRNAs, transposons, tRNAs, ARSs, telomeres, and ncRNAs. We then identified regions which contained none of these features and, among those greater than 601 bp, we divided the region up into a maximal number of negative examples, allowing 400 bp of overlap, and spaced out these examples within the region. For instance, a 1,004 bp region would generate three negative examples (1-602, 202-803, 403-1004). In all cases, we chose negative examples in a strand-specific manner, such that a positive example on one strand would not preclude a negative example at the same position on the opposite strand. This yielded 72,276 negative promoter examples.

We defined the CPA region as -75 to +75 relative to the annotated CPA sites of transcripts, and we calculated features over three 50 bp bins as shown in **Figure 1A**. We chose these bins to capture the known positional preferences of known cleavage elements (Guo and Sherman 1996; Dichtl and Keller 2001; Ozsolak et al. 2010). The "positive" examples included the CPA sites from our annotated transcripts (as described above), and we attained "negative" examples by dividing ORFs (excluding 100 bp near the 3' end) into a maximal number of (non-CPA)

examples that could overlap by no more than 100 bp. This yielded 5,010 “positive” and 155,093 “negative” examples for the termination classifier.

Initiation Classifier Features

We calculated feature values independently over six sequence windows (bins) shown in **Figure 1A**. We determined the bin locations manually, on the basis of the characteristic architecture of yeast promoters (see **Supplementary Table 1**).

Features can generally be grouped into four broad categories: motifs for TFs, motifs for RBPs, DNA structural features and base content, and nucleosome excluding sequences. These features are fully described in **Supplementary Table 2**. For the initiation classifier, we only included RBPs and strand-specific base content for the 100 bp following the TSS, since these are the only bases that are transcribed. We included all other features in all bins. For both RBPs and TFs, we calculated features using position frequency matrices (PFMs) representing the factor’s specificity. The “score” for a sequence bin is the probability of the factor binding anywhere in the sequence given the PFM, which we calculated using the method described previously (Chen et al. 2007). For RBPs, we performed all motif scans strand-specifically, whereas we scanned both strands for TFs. Two exceptions to this are the TATA-box (including only the forward orientation) and Rap1 (including forward, reverse, and strand-unspecific orientations), since these factors had been shown to have asymmetrical binding effects (Smale and Kadonaga 2003; Beer and Tavazoie 2004). Initial analyses were complicated by the fact that there are many motifs of varying quality for each yeast TF; this motivated us to develop a database of expert curated yeast TF motifs (de Boer and Hughes 2012). We also generated new motifs for Hrp1, Npl3, and Yra1 using our RNAcompete method (Ray et al. 2013) to expand our existing catalogue of RNA features (Kessler et al. 1997; Takagaki and Manley 1997; Tacahashi et al. 2003; Kim Guisbert et al. 2005; Kim Guisbert et al. 2007; Deka et al. 2008; Millevoi and Vagner 2009; Li et al. 2010; Pancevac et al. 2010).

We calculated DNA structural features as the average value across the sequence bin for each mono-, di-, or trinucleotide, which we mapped to the corresponding structural value (Satchwell et al. 1986; Lu and Olson 2003), as previously described (Lee et al. 2007). The mappings are available on the accompanying website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/transcription_model/). For nucleosome excluding sequences (Lee et al. 2007), we simply counted the number of occurrences of the hexanucleotide within the sequence bin. We calculated the poly-A feature as the total length of all poly-A tracts in the bin, only considering those with a length of at least five.

We calculated these features across the six bins for every example as input to the classifier algorithms. This yielded a matrix of 77,286 examples and 1,698 features which we then used to create the initiation classifier.

Termination Classifier Features

The termination classifier features included base content and RBP motifs. We calculated the RBP features in a strand-specific manner, as described above for the RBPs included in the initiation classifier. Base content included strand-specific counts of A, T, G, and C, as well as G/C content, and the ratio of As to Ts and Gs to Cs. We calculated each feature over each of the three bins, yielding a matrix of 160,103 examples by 147 features that we used to create the termination classifier.

Random Forest Classifiers

We created both initiation and termination classifiers using the randomForests (version 4.6-6) R module (Breiman et al. 2008). We created forests as four replicates, each with 50 trees, which we averaged to make a forest of 200 trees but facilitating distributed computation. We used regression, sampling with replacement, and a minimum node size of five to make the forests. We used the resulting classifiers to classify new data and they provided a score between zero and one for how promoter/CPA site-like the sequence is. When scanning the test data (e.g. chromosomes 1-8), we averaged the predictions from all forests trained on the training data (e.g. trained on chromosomes 9-16, for each of 9-16 held out).

Model Refinement

We iteratively rebuilt the classifiers including only the top N most important features (the “mean decrease in node impurity” provided by Random Forests) for increasing N. Each time we added another feature, we calculated the change in the error rate (ER) based on the held-out chromosome. The ER is defined as the probability that a randomly selected negative example has a score greater than a randomly selected positive example, and corresponds to one minus the area under the ROC curve (1-AUROC). We only retained features whose inclusion reduced the ER by at least 3% relative to the last ER. For the termination classifier, the ER was minimal after inclusion of the fourth feature, and so no further features were retained. This procedure allowed us to gauge how beneficial the addition of each additional feature is to the classification ability of each model, and allowed us to remove redundant features, since these would not provide an additional decrease in the error rate. Because this procedure was computationally intensive, we only considered the most important features when identifying the critical features (top 150 for initiation and 100 for termination). However, no features were retained after the 23rd and 4th most important features for the initiation and termination classifiers, respectively, indicating that few, if any, additional features beyond 150 and 100 would have been included had we continued the procedure with all features. Several different feature selection criteria yielded similar results (data not shown).

Analysis of Model and Cellular TSS/CPA Site Resolution

We wanted to compare the resolution of the initiation and termination classifiers to that achieved by the cell. To do this, we analyzed RNA-seq data specific to TSSs/CPA sites (Pelechano et al. 2013), which we normalized within promoter regions (-300 to +300, relative to the annotated TSS) to get the % usage of each base as a TSS/CPA site. First, we wanted to gauge the TSS/CPA site resolution of the cell, so we aligned the % usage around the annotated TSS/CPA site (within 100 bp) to the most frequently used site (**Supplementary Figure 1A, B**). When there were ties, we went with the closest site to the annotated site. We next aligned the classifier scores (within 100 bp of the annotated TSS/CPA site) and analyzed the TSS/CPA usage around the aligned scores (**Supplementary Figure 1C, D**). Finally, to ask if we could further narrow down the exact site of initiation/cleavage, we identified motif matches for the initiator and cleavage site motifs (which have the consensuses of CA and SAA, respectively) closest to the centre of the classifier score peaks. Here, we first searched for consensus matches to the motif within the peak and, if none were found, we searched for more and more degenerate motifs until one was found. We then aligned the % usage data to these peak-specific motif matches (**Supplementary Figure 1E, F**).

Analysis of Predicted *trans*-Acting Regulators and Corresponding Mutants

To categorize each gene as “controlled” or “not controlled” by a factor, we calculated the difference between the initiation score for the full model and when the corresponding feature is set to the median value for non-promoters, effectively simulating that feature’s absence. If the score decreased by at least 0.1, we considered the model to predict that the gene’s promoter is controlled by the corresponding factor (“predicted-controlled”). We then compared these to the expression changes in the corresponding *trans*-factor mutants (Alper et al. 2006; Badis et al. 2008) and gauged the significance of the association using the rank sum test (**Supplementary Figure 4**). To compare this to simply using predicted binding sites, we used the predicted binding of each TF given its motif (the same values as the inputs to the initiation classifier). For binding sites, we considered the highest ranking binding events to be targets and used the same number of targets for each TF as was used for the classifier scores.

Combinatorial Promoter Library Construction and Analysis

We synthesized the promoter library as complimentary oligonucleotides in three separate segments that contained unique complimentary ends. These segments correspond to the -150:-80, -80:TSS, and TSS:+80 regions and are designated A, B, and C, respectively. We selectively phosphorylated (to prevent unwanted ligation products), and annealed the segments by denaturing at 95˚C for 5 minutes and cooling slowly to 4˚C (1˚C/min). We then pooled the resulting double stranded promoter fragments, ligated them together at room temperature for 24 hours, and purified the full length promoters by gel extraction. After ligation and gel purification, we ligated the combinatorial promoter library into a modified GFP expression vector (Kainth et al. 2009) (pTH7638) using BamHI and NheI sites. We then transformed the resulting mixture

into *E. coli* by electroporation. We generated approximately 6 million transformants and, by pooling these, we isolated the plasmid library and transformed yeast *en masse* using the lithium acetate method. Approximately 2 million yeast transformants were generated and we pooled these to yield the final yeast promoter library.

We grew the pooled yeast library in SC-Leu overnight and isolated the cells, washed them in water, and resuspended them in 1xTE to an OD of 0.68. We then sorted the cells on a FACSAria flow cytometer into six bins of GFP fluorescence and used gates to select for single cells. After sorting, we isolated the cells, resuspended them in SC-Leu, grew them overnight, and isolated the plasmids. We then barcoded the promoters for multiplexing using PCR, and sequenced the barcoded promoter libraries in one lane on the Illumina HiSeq platform, using paired end sequencing, reading 111 and 114 bases from the ends to ensure we could uniquely identify the source promoters. We mapped reads to the promoter sequences using Bowtie (Langmead et al. 2009) and only considered promoters that had > 50 reads (summing over all bins) for further analysis. We normalized the read count per promoter by the number of reads per bin to correct for differing numbers of reads per bin and multiplied this normalized count by the proportion of cells ending up in each bin. From this, we calculated the expression level by weighting the proportion of each promoter in each bin by the average fluorescence of the bin.

Creation and Evaluation of the Unified Model

Transcript Map Used to Derive Unified Model Parameters

Overview. The HMM parameters included the mean and variance of the "observations" (the two classifier scores for both strands) in each state (the nodes shown in **Figure 4**), as well as the probabilities of transitioning between the states shown in **Figure 4**. To obtain these parameters, we needed a map of the occurrence of the HMM states to use as training data. For reasons outlined in the main text, we could not use the original transcript maps employed for the initiation and termination classifiers. We therefore used RNA-seq data (Nagalakshmi et al. 2008; Lipson et al. 2009; Levin et al. 2010; Ozsolak et al. 2010) to create a state label map. The state labels needed to have the same states as the UM. We therefore used a procedure similar to that used to derive the initiation and termination classifiers, and the UM, but instead based on RNA-seq data. We first developed classifiers that identify the range of TSS/CPA sites, which we then combined into a HMM capable of predicting transcript structure. Finally, we used this HMM to label the bases of the genome with the corresponding UM states, based entirely on RNA-seq data.

Classifiers. We created classifiers that would identify TSSs and CPA sites given RNA-seq data (Nagalakshmi et al. 2008; Lipson et al. 2009; Levin et al. 2010; Ozsolak et al. 2010) as features (designated TSS-R/CPA-R, where "R" stands for "RNA-seq"). We made these classifiers with random forests and designed them to distinguish TSSs and CPA sites from non-TSS and non-CPA sites, respectively, using raw read counts and the change in counts over different windows (since read counts go up at the TSS and down at the CPA site), as well as CPA-specific reads

(i.e. poly-A reads). These classifiers provided a score reflecting how much each base in the genome resembles a TSS/CPA site based on RNA-seq data surrounding that base (see **Supplementary Figure 5B** for an example of the predictions).

Hidden Markov Model. We then used the TSS-R and CPA-R scores, as well as the raw RNA-seq data, to train a HMM (designated HMM-R) using the Hidden Markov Model Toolbox for Matlab (http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html). The HMM outputs states corresponding to those in **Supplementary Figure 5A**. In order to train the HMM-R, we required a set of state labels for every base in the genome representing the TSS, transcript, CPA site, and intergenic states (treating strands independently, see **Supplementary Figure 5A**). We derived these labels by identifying peaks in the TSS-R score upstream of ORFs, and peaks in the CPA-R score downstream of ORFs (both within 1kb of the ORF start/end) and labelling these as the TSS and CPA states, respectively. We labelled everything between these peaks as the transcript state and we labelled everything else intergenic. We then used these state labels to train the RNA-seq based HMM-R.

The observations of the HMM-R included the TSS-R and CPA-R classifier scores, as well as log read counts for RNA-seq data (Lipson et al. 2009; Levin et al. 2010) for every base. We assumed the observation distributions are Gaussian within each state, with each state having a mean and variance for each observation. We calculated the maximum likelihood parameters given the aforementioned state labels based on CPA-R/TSS-R peaks surrounding ORFs for half the chromosomes (1-8, or 9-16) and then used the resulting HMM-R to generate a transcript state map for the other half, swapping the two chromosome sets to get maps for both halves. We used the Viterbi path, corresponding to the *single most likely path* through all the HMM states given the observations, to generate two state label maps (one per strand) that correspond to our UM states. Since these are derived from RNA-seq data, they include both ORFs and non-ORF stable transcripts. We removed state labels corresponding to very small transcripts (<150 bp) since these may represent artefacts in the RNA-seq data. Next, we combined the single stranded state labels such that CPA sites were allowed to overlap (in the bidirectional terminator case), but transcripts and TSSs could not. In the case of two convergent transcripts where the ends of transcripts overlap, we treated the overlapping bases as bidirectional terminators up to an overlap of 130 bp, above which the transcripts were considered to conflict. When the state labels conflicted, we kept the two strands separate, resulting in two final state label maps designated “top” and “bottom”, which are identical except where transcripts conflict, in which case they contain the forward and reverse strand predictions, respectively (**Supplementary Figure 5B**). This yielded two state label maps for the entire genome which could then be used for training the UM. When calculating the maximum likelihood (ML) estimates for the UM parameters, we used both “top” and “bottom” state labels, which amounts to weighting the conflicting data half as much in the ML estimate.

Creating the Unified Model

We built the unified model (UM) using the Hidden Markov Model Toolbox for Matlab (http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html). For our initial parameters, we used

the ML estimates given the RNA-seq-based transcript state label map (derived as described above) and the corresponding classifier scores. This yielded a matrix of transition probabilities representing the probability of moving from every state to every other state, as well as a matrix of means and a matrix of variances representing the expected observation distributions for each classifier score in each state (see **Supplementary Tables 4 and 5**). We assumed that the observations were approximately normally distributed within each state and calculated the observation probabilities (probability of observing a given classifier scores in a given state) using a Gaussian probability density function with diagonal covariance (one mean and one variance for each classifier score for each strand). Since the DNA strand label is arbitrary, we calculated the parameters symmetrically between the strands. The classifier scores were not initially normally distributed, so we first transformed them to be approximately normal by Box-Cox power transformation.

We noted that the performance of the UM was initially worse than expected and hypothesized that this was because the ML parameter estimates did not accurately represent the parameters used by the cell. To address this, we optimized the observation means to maximize the Pearson correlation of the transcript predictions to observed RNA-seq read counts, on a per-base level (Levin et al. 2010), for the training chromosomes, using a hill-climbing algorithm. To reduce overfitting of the parameters, we first scaled all means together, then we scaled the initiation and termination means together, then we scaled each mean individually, each time optimizing until convergence. We tuned the means iteratively such that each was changed by at most 50% per round. Each time no additional changes to the parameters improve the correlation, we halved the amount by which the means are changed (50%, 25%, 12.5%, etc.) and repeated the process until the parameters converge. The algorithm stops when the amount by which the parameters are changed is 3.125%. This procedure did not add any additional parameters to the model, since we altered the values of existing parameters, and we found it selected a stable set of parameters and improved the quality of the model's predictions (as measured by the AUROC predicting TSSs, CPA sites, and transcribed bases). The final observation distribution parameters are shown in **Supplementary Table 5**.

Observations that are extremely unlikely in any state can tend to dominate the predictions of a HMM. For example, if an observation probability is nearly zero in all states but is many fold higher in one state compared with the others, the most probable state will be strongly favoured. For this reason, we also added a constant (0.001) to the observation probabilities of every state, which we then re-normalized to sum to one over all states, preventing these extremely improbable observations from dominating the prediction. We used the forward-backward algorithm to generate the predictions of the model, which yields the probability of being in each state at every base, given the classifier scores for the entire chromosome.

Evaluation of the UM: Comparisons to Transcripts and RNA-seq Data

We calculated the overlap of the model's predictions with RNA-seq data (Levin et al. 2010) and transcript predictions on a base-by-base basis. The positive set of transcripts included non-dubious ORF-containing genes and, when available, the boundaries represent the TSSs and

CPA sites from the transcript annotations used to train the classifiers, but when unavailable, the boundaries represent the ORF start and end. We considered a base a "predicted transcript" if the UM probability of being part of a transcript was over 0.5 as this value precludes there also being a transcript predicted on the opposite strand. For calculating percent overlaps, for instance, with antisense transcripts, we found the number of bases that were both transcribed (>10 reads) and predicted to be transcribed, and compared this to the overlap expected by chance (%transcribed x %predicted).

We calculated P-values associated with overlaps, precisions, and recalls by simulation. We first calculated the actual number of true positives (TPs) for each set. We then used simulation (1000 replicates each) to estimate the TP distribution expected by chance, which we used to convert the actual number of true positives into a Z-score. We then converted the Z-scores into P-values using the Gaussian cumulative distribution function. For simulating the number of TPs expected by chance for RNA-seq data in intergenic regions, we repeatedly shuffled the order of intergenic regions for the RNA-seq data and calculated the number of TPs when compared to the un-shuffled prediction data. The same procedure was used for RNA-seq/prediction overlap in antisense transcripts. We calculated the number of TPs expected by chance for the precision and recall of transcript species (e.g. ORF-containing, ORF-containing within 100 bp, Ty elements/sn/snoRNAs) similarly, but, here, we randomized the prediction data by shuffling the order of genes and intergenic regions. We then calculated precision and recall values as if the data had not been shuffled to estimate the background TP distribution. Each of these procedures yields a Z-score for how the actual TP values compare with the randomized TP-distribution, and from these we calculated the P-values described in the paper.

For transcript predictions used in **Supplementary Figure 7**, we first defined transcriptional units predicted by the model. We treated each strand independently and, for the forward strand, we first identified potential TSSs by, starting from the first base of the chromosome, identifying regions with an increase in the probability of being transcribed of at least 0.08 over 120 bp. We wanted to avoid excluding potential transcripts, and so intentionally set this threshold relatively low. We then identified CPA sites that followed potential TSSs using the same criteria, only instead using the decrease in the probability of being a transcript. Where suitable TSS and CPA predictions were both found, we considered the position of the maximum increase and maximum decrease in the probability of being transcribed as the TSSs and CPA sites of the predicted transcriptional units. We treated the reverse DNA strand identically, only with the direction reversed. We then removed any transcripts that overlap on the same strand with ORFs, annotated transcripts, transposons, sn/snoRNAs, or other known features, leaving us with a set of predicted transcripts that do not correspond to known features. We scaled the data in **Supplementary Figure 7** such that the starts and ends of the predicted transcripts line up and averaged and scaled these so that the minimum and maximum between the two data sets displayed in **Supplementary Figure 7B** are comparable.

Generation and Analysis of 6 kb Randomly-Generated DNA

We designed the four 3 kb fragments that comprised the four 6 kb randomly-designed loci by randomly selecting bases following the *S. cerevisiae* base content (A=T=0.31) and occasionally inserting sites for Abf1, Reb1, Rap1, Rsc3, and Spt15 (TBP). The motifs added were sampled from the position frequency matrix (PFM) for these factors. For instance, if the factor's PFM includes an A at the first position 99% of the time then the inserted motifs included an A at the corresponding position 99% of the time. We chose four fragments for gene-synthesis that when combined were predicted to form several "transcripts". We assembled the four fragments into the four 6 kb combinations and integrated these into the yeast genome with a KanMX selectable marker. We confirmed the integration by PCR across the integration junctions. We grew these four yeast strains in YPD to an OD of ~1.0, at which point we collected mRNA using the hot acid phenol method. We treated the total RNA with DNase I, purified it using an RNeasy mini kit (Qiagen), and isolated the mRNA using a NucleoTrap mRNA mini kit (Clontech). We isolated genomic DNA (gDNA) separately using a YeaStar Genomic DNA kit (Zymo Research) and subsequently sheared it by sonication to an average size of approximately 250 bp. We labelled RNA and DNA samples with G-coupled Cy3 and Cy5 dyes, respectively (Createch). Following dye-coupling, we partly hydrolyzed mRNA samples to an average size of about 250 bp using a NEBNext Magnesium RNA Fragmentation Module (NEB). We performed the microarray hybridization as described elsewhere (Zhang et al. 2004).

The probes on the tiling array were 60 bp long and spanned the region in 1 bp increments, with each probe present on the array at least four times. We normalized the tiling array data using a background model (Huber et al. 2006) which relates, for each probe (k), the signal intensity (y_k) to the background signal (B_k) plus the amount of nucleic acid in the sample (x_k) times the proportionality factor (a_k), or $y_k=B_k+x_k a_k$. Here, we want to determine the amount of nucleic acid in the sample (x_k). We modified this normalization protocol to take advantage of the fact that the target sequences of many probes were absent from some samples (e.g. the A1B1probe target sequences are absent from the A2B2 strain). This allowed us to estimated separate mRNA and gDNA background signals (B_k) for each probe by taking the median signal intensity of the probe in strains lacking the target DNA sequences. We estimated the proportionality factor for each probe (a_k), corresponding to how probe intensity changes with the amount of nucleic acid, by comparing the background-normalized gDNA signal at each probe to the actual amount of gDNA present in the sample, where the target DNA was present in the sample. We estimated the amount of mRNA at each probe using this model, and scaled the amounts to the median intensity across the kanMX gene to make mRNA levels comparable across constructs.

Other Analyses

Analysis of Cooperative Binding for Hrp1 Binding Sites in Terminator regions

We wanted to determine if Hrp1 binding site clustering could be explained by cooperative binding, so we looked at occurrences of the Hrp1 binding motif TAYRTA, where Y=pyrimidine

and R=purine (Irniger and Braus 1994; Ray et al. 2013). Since it had been previously shown that two Hrp1 molecules are capable of simultaneously binding UA(6) but not UA(5) *in vitro* (Perez-Canadillas 2006), we compared the occurrence of motifs that could be bound by two Hrp1 molecules simultaneously (i.e. TAYRTATAYRTA) to those that could be bound by only one Hrp1 molecule at a time (i.e. TAYRTAYRTA), ensuring each motif instance represented only two potential binding sites (e.g. TATATATAYRTA has three overlapping binding sites, so was excluded). In terminator regions, we found 598 10 bp motifs representing two overlapping binding sites and 38 12 bp motifs representing two non-overlapping Hrp1 binding sites. By correcting for the two bases of extra information (TA) encoded in non-overlapping binding sites, we obtained the expected number of adjacent binding sites (57), which is greater than the number observed. Further, we also analyzed the number of TA(6) repeats (to which two Hrp1 molecules could simultaneously bind) and found these to be approximately the number expected given the abundance of TA(5) and TA(7), which were 447, 278 and 181 for 5, 6 and 7 AU repeats, respectively.

Calculating the Number of Convergent Genes That Use the Same CPA Site

We analyzed ORFs on chromosomes 1-8 that are arranged in the convergent orientation and considered only those that are separated by at most 750 bp between their stop codons (592/647 convergently arranged ORFs). Of these, we then counted the number of convergent pairs that the UM predicts have at least one base in the bidirectional CPA state (P(CPA+/-)>50%), of which there were 240 (40.54%). Of these, the UM predicts that the majority (68.33%) overlap by at least 100 bp.

Calculating the Number of Predicted Promoters and Transcribed Bases in Random DNA Sequence

In order to estimate how often promoter-like sequences will originate in DNA with randomly-generated sequences, we created ten independent 10 Mb DNA sequences in which the base frequency matches that of the yeast genome (A=T=0.31). We then scanned these sequences with the initiation classifier. We identified peaks meeting a score threshold corresponding to the median score of true promoters (0.335), and called predicted promoters (which were required to be at least 100 bp apart and selected in a greedy fashion with the highest scoring peaks selected first). The average number of bases between predicted promoters was in good agreement between the ten sequences (940 bp, SD=6 bp among the ten 10 Mb sequences).

To estimate the fraction of bases that are predicted to be transcribed, we scanned these same 10 Mb sequences with the termination classifier and used the initiation and termination classifier scores to predict gene structure using the UM. We then counted the fraction of bases that were predicted to be transcribed (probability >0.5) on each strand. The percent transcribed for each strand ranged from 32.7% to 33.4% for the ten sequences. An overall average of 33.1% of each strand is predicted to be transcribed (so 66.2% of the sequences are predicted to be transcribed on one of the two strands).

SUPPLEMENTARY REFERENCES

Alper H, Moxley J, Nevoigt E, Fink GR, Stephanopoulos G. 2006. Engineering yeast transcription machinery for improved ethanol tolerance and production. *Science (New York, NY)* **314**: 1565-1568.

Badis G, Chan ET, van Bakel H, Pena-Castillo L, Tillo D, Tsui K, Carlson CD, Gossett AJ, Hasinoff MJ, Warren CL et al. 2008. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **32**(6): 878-887.

Beer Ma, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* **117**: 185-198.

Breiman T, Cutler A, Classification D. 2008. The randomForest Package.

Chen X, Hughes TR, Morris Q. 2007. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics* **23**(13): i72-79.

Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**(7330): 368-373.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**(14): 5320-5325.

de Boer CG, Hughes TR. 2012. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res* **40**(Database issue): D169-179.

Deka P, Bucheli ME, Moore C, Buratowski S, Varani G. 2008. Structure of the yeast SR protein Npl3 and Interaction with mRNA 3'-end processing signals. *J Mol Biol* **375**(1): 136-150.

Dichtl B, Keller W. 2001. Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor. *The EMBO journal* **20**: 3197-3209.

Erb I, van Nimwegen E. 2011. Transcription factor binding site positioning in yeast: proximal promoter motifs characterize TATA-less promoters. *PloS one* **6**(9): e24279.

Guo Z, Sherman F. 1996. 3'-end-forming signals of yeast mRNA. *Trends Biochem Sci* **21**(12): 477-481.

Huber W, Toedling J, Steinmetz LM. 2006. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**(16): 1963-1970.

Irniger S, Braus GH. 1994. Saturation mutagenesis of a polyadenylation signal reveals a hexanucleotide element essential for mRNA 3' end formation in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **91**(1): 257-261.

Kainth P, Sassi HE, Pena-Castillo L, Chua G, Hughes TR, Andrews B. 2009. Comprehensive genetic analysis of transcription factor pathways using a dual reporter gene system in budding yeast. *Methods* **48**(3): 258-264.

Kessler MM, Henry MF, Shen E, Zhao J, Gross S, Silver PA, Moore CL. 1997. Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3'-end formation in yeast. *Genes Dev* **11**(19): 2545-2556.

Kim Guisbert K, Duncan K, Li H, Guthrie C. 2005. Functional specificity of shuttling hnRNPs revealed by genome-wide analysis of their RNA binding profiles. *Rna* **11**(4): 383-393.

Kim Guisbert KS, Li H, Guthrie C. 2007. Alternative 3' pre-mRNA processing in Saccharomyces cerevisiae is modulated by Nab4/Hrp1 in vivo. *PLoS biology* **5**: e6.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39**(10): 1235-1244.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**(9): 709-715.

Li X, Quon G, Lipshitz HD, Morris Q. 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *Rna* **16**(6): 1096-1107.

Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M. 2009. Quantification of the yeast transcriptome by single-molecule sequencing. *Nature biotechnology* **27**: 652-658.

Lu XJ, Olson WK. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* **31**(17): 5108-5121.

Millevoi S, Vagner S. 2009. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic acids research* **4242**: 1-18.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan aP, John B, Milos PM. 2010. Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* **143**: 1018-1029.

Pancevac C, Goldstone DC, Ramos A, Taylor Ia. 2010. Structure of the Rna15 RRM-RNA complex reveals the molecular basis of GU specificity in transcriptional 3'-end processing factors. *Nucleic acids research* **38**: 3119-3132.

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**(7447): 127-131.

Perez-Canadillas JM. 2006. Grabbing the message: structural basis of mRNA 3'UTR recognition by Hrp1. *Embo J* **25**(13): 3167-3178.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**(7457): 172-177.

Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191**(4): 659-675.

Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**: 449-479.

Tacahashi Y, Helmling S, Moore CL. 2003. Functional dissection of the zinc finger and flanking domains of the Yth1 cleavage/polyadenylation factor. *Nucleic Acids Res* **31**(6): 1744-1752.

Takagaki Y, Manley JL. 1997. RNA recognition by the human polyadenylation factor CstF. *Mol Cell Biol* **17**(7): 3907-3914.

van Bakel H, Tsui K, Gebbia M, Mnaimneh S, Hughes TR, Nislow C. 2013. A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. *PLoS Genet* **9**(5): e1003479.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033-1037.

Yassour M, Pfiffner J, Levin JZ, Adiconis X, Gnirke A, Nusbaum C, Thompson D-A, Friedman N, Regev A. 2010. Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome biology* **11**: R87.

Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E et al. 2004. The functional landscape of mouse gene expression. *Journal of biology* **3**(5): 21.