Annotation of directed genomic states unveils variations in the Pol II transcription cycle

Benedikt Zacher[1,2], Michael Lidschreiber[1,4], Patrick Cramer[1,4], Julien Gagneur[1,**] & Achim Tresch[1,2,3,*]

[1]*Gene Center and Department of Biochemistry, Center for Integrated Protein Science CIPSM, Ludwig-Maximilians-Universität München, Munich, Germany*

[2]*Institute for Genetics, University of Cologne, Cologne, Germany*

[3]*Max Planck Institute for Plant Breeding Research, Cologne, Germany*

[4]*Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany*

** Corresponding author. Tel.: +49-221-5062-161; Fax: +49-221-5062-163; E-mail: tresch@mpipz.mpg.de*

*** Corresponding author. Tel.: +49-89-2180-76742; Fax: +49-89-2180-76797; E-mail: gagneur@genzentrum.lmu.de*

DNA replication, transcription and repair involve the formation of protein complexes that undergo transitions in their composition as they progress along the genome. Average occupancy profiles of DNA-bound proteins over genes have been instrumental to understand these processes. However, averaging requires predefined gene sets, hides gene-specific variations, and obscures transitions that do not occur at fixed distances from the aligned gene feature. To overcome these limitations, we introduce bidirectional HMMs which infer directed genomic states from occupancy profiles de novo. Application to RNA polymerase II-associated factors in yeast and chromatin modifications in human T-cells recover the vast majority of transcribed loci, reveal gene-specific variations in the yeast transcription cycle and indicate the existence of directed chromatin state patterns at

transcribed, but not at repressed regions in the human genome. In yeast, we identify 32 new transcribed loci and the analysis indicates a regulated initiation-elongation transition, the absence of elongation factors Ctk1 and Paf1 from a class of genes, a distinct transcription mechanism for highly expressed genes, and novel DNA sequence motifs associated with transcription termination. We anticipate bidirectional HMMs to significantly improve analysis of genome-associated directed processes.

An important question in molecular systems biology is how the occupancy of a genomic position with protein factors relates to the composition of genome-associated protein complexes at this position. This question is of high relevance to fundamental genome-associated processes such as DNA replication, transcription and repair because these generally involve the formation of functional multiprotein complexes that undergo transitions in their protein composition along the genome. For example, during transcription, RNA polymerase (Pol) II progresses through the initiation, elongation, and termination phases, which are characterized by the presence of distinct Pol II-associated proteins, various post-translational modifications of Pol II and histones, and nucleosome occupancy. Analysis of genome-wide occupancy maps of Pol II-associated factors obtained by chromatin immunoprecipitation (ChIP) in yeast indeed indicated the presence of distinct protein complexes for the initiation, elongation, and termination of transcription at different genomic positions, and suggested a universally conserved mRNA transcription cycle[1–3]. These studies however used metagene analysis, i.e. the averaging of ChIP occupancy signals over many selected genes that are often also scaled to a common length, and have limitations.

Despite its wide use and successful application, metagene analysis has several drawbacks. First, it only reveals the average behavior of genes in a selection, and bears the risk of averaging out variations between genes and hiding accurate spatial information. Second, it is limited to annotated loci, thereby biasing conclusions towards well-described transcription units. Third, genes are aligned with their transcription start

sites (TSSs) during metagene analysis, and therefore neither changes that occur at a variable distance from the TSS nor changes that take place only in a minority of genes are detected. At least in part due to these shortcomings, the nature of the various transient Pol II protein complexes in vivo, the order of their occurrence within the transcription cycle, and the extent of variation in these parameters between different genes is not well understood.

To address these issues, we introduce the concept of 'directed genomic states', position-specific annotations inferred from genomic occupancy data without metagene averaging. These states may reflect distinct genome-associated complexes, but do not necessarily have to correspond to biochemically defined entities. The states are directional because genomic processes such as transcription occur in a directional manner, with opposite directions on the two DNA strands (+ and -). To define directed genomic states, we developed a bidirectional hidden Markov model (bdHMM), a novel method that annotates directed states from genomic data that are not strand-specific (such as ChIP), and optionally strand-specific data (such as RNA expression). Hidden Markov models[4] describe longitudinal observations (here: occupancy along DNA) as a sequence of discrete hidden states (here: genomic states). These states are inferred from the data in an unbiased fashion, without the need for prior genome annotation. HMMs have been used to infer chromatin states, enhancers, promoters, and quiescent regions in the genome of human[5–11] and fly[12,13]. However, current HMM approaches are not able to infer directed states that underlie strand-specific processes such as transcription that occur in forward or reverse direction. We evaluate the performance of bdHMM on simulated data and demonstrate the general applicability of this method to different organisms and datatypes on a tiling array dataset in yeast and a deep-sequencing dataset in human. Our bdHMM analysis of previously defined chromatin states in human T cells[11] *de novo* identified directed chromatin states patterns. Application of our bdHHM method to a set of 22 genomic profiles in the yeast *S. cerevisiae* revealed new transcription units and DNA sequence motifs, and additionally unveiled variations in Pol II transcription that were previously hidden by metagene analysis. The yeast and human

data sets, state annotation, and bdHMMs which generated them, are available from the supplementary website www.treschgroup.de/STAN.html.

1 Results

Annotation of directed genomic states using bdHMMs

Bidirectional HMMs are best understood with the help of a simulated dataset. In the example illustrated in Figure 1, we described a part of the genome where transcription occurs as a sequence of three states associated with different genomic segments. The transcribed segment is flanked by segments that are in an untranscribed state (U), and is split in two distinct states corresponding to early (E) and late (L) transcription activity. The order of the three states U, E, and L along the genome depends on gene orientation (Fig. 1a, grey arrows). ChIP measurements for a single protein were simulated with low, medium and high average occupancy in the different states of transcription. Note that ChIP signals are not strand-specific.

A standard HMM with 3 hidden states can distinguish the three occupancy levels of the protein and recover the three states. However, the transition probabilities in the standard HMM are symmetric (Fig. 1b) because the number of observed transitions between successive states - say E to L - in the forward direction equals the number of transitions in the reverse direction, L to E. Hence, standard HMMs are neither able to capture the strand-specificity of transcription (i.e. the two different directions of transcription along the genome) nor do they infer biologically meaningful transitions along the genome as they occur during transcription.

In order to infer directed transitions and directed genomic states, bdHMMs have 'twin states', one for each strand and genomic state. For instance, the early state E is split up into the twin states E^+ and E^-. Twin states are coupled by two symmetry conditions. First, twin states are required to have identical

emission probabilities, i.e., in our example $\Pr(o_t|s_t = E^+) = \Pr(o_t|s_t = E^-)$, where o_t is the observed data and s_t is the hidden (transcription) state at position t. Second, twin states satisfy transition symmetry, a novel generalization of reversible Markov chains (see Online Methods for details), which requires that state transitions are invariant under reversal of time and direction, i.e. $\Pr(s_t = L^+|s_{t-1} = E^+) = \Pr(s_{t-1} = L^-|s_t = E^-)$. By splitting up E respectively L into E^+ and E^- (respectively L^+ and L^-), the bdHMM learns the transitions for each direction separately, but not independently of each other. In our example, this results in the bdHMM transition probabilities $\Pr(s_t = L^+|s_{t-1} = E^+) > 0$, and $\Pr(s_t = L^-|s_{t-1} = E^-) = 0$, as opposed to $\Pr(s_t = L|s_{t-1} = E) > 0$ and $\Pr(s_t = E|s_{t-1} = L) > 0$ in the HMM (Fig. 1b,c). These two conditions make it possible to recover the directed genomic states (Fig. 1a). Although the formal number of states doubles, the effective number of parameters remains the same.

Parameters are inferred using a constrained Baum-Welch algorithm, the validity of which was assessed by simulations showing that model parameters and states were recovered with high accuracy, even when only few training data was used (Online Methods, Supplementary Fig. 8). The bdHMM is implemented in the R package STAN (STrand-specific ANnotation of genomic data), which is freely available from the supplementary website.

Genomic state annotation results in a global, strand-specific transcription map

We applied the bdHMM to ChIP data in S. cerevesiae, where high-resolution data sets for dozens of proteins of the transcription machinery are available. We compiled genome-wide ChIP-chip experiments for transcription initiation factors (TFIIB, Kin28), elongation factors (Spt5, Spn1, Bur1, Spt16, Ctk1, Paf1), termination factors (Pcf11, Rna15, Nrd1), Pol II and various modifications of its C-terminal domain (CTD) (Tyr1P, Ser2P, Ser5P, Ser7P) and nucleosomes[1,14–16]. The data set was complemented by

strand-specific mRNA expression data[17] (Fig. 2).

The number of bdHMM states needed to be specified in advance. Bearing in mind that our states should distinguish biologically different genomic states, classical model selection criteria (BIC, AIC, MDL) are not useful. Those criteria balance the number of parameters/states against the precision of the data fit. Since our data is very rich, they suggest a very high number of states, which cannot be interpreted. This issue has been reported repeatedly in association with HMMs[9–11] for integrative analysis of ChIP data. We tried several state numbers (data not shown) and found that 20 states yield an appropriate trade-off between model complexity and biological interpretability. The genome-wide state annotation was derived as the most likely state path (Viterbi decoding, Fig. 2), which partitioned the 12 Mb yeast genome into 48,507 directed and 10,760 undirected state segments with distinct bdHMM states. This analysis yielded a strand-specific map partitioning the yeast genome into segments with different directed genomic states. In principle, the strand-specific expression of this dataset could also be used with standard HMMs to learn directed states. However, fitting a standard HMM did not recognize directed genomic states. In particular, there is no obvious pairing between the forward (+) and the reverse (-) states, demonstrating the need for bdHMM (Online Methods, Supplementary Fig. 1).

Transcription cycle phases have a substructure

To understand how the 20 bdHMM states relate to phases of the transcription cycle, we analyzed their average frequencies along annotated, transcribed genes (Fig. 3b, Online Methods). The states showing a single frequency peak (18 out of 20 states) were grouped into six transcription phases, according to the location of their peak on the average gene: Promoter (P, 2 states), Promoter Escape (PE, 2 states), early Elongation (eE, 3 states), mid Elongation (mE, 5 states), late Elongation (lE, 3 states), and Termination (T, 2 states). Two states showed two peaks in frequency, in each case with one peak upstream of the transcription start site and one peak around the polyadenylation (pA) site. We interpreted these two states as

mixed promoter and termination states and labeled them accordingly P/T1 and P/T2 (Fig. 3a,b). Hence, although overlapping transcription is not explicitly modeled by bdHMMs, this phenomenon could be captured by specific states.

The mean factor occupancy defining a particular state is indicative of the composition of the transcription complex and its activity (Fig. 3a). Indeed we found that the enrichment or depletion of protein factors in each state was in accordance with their known roles in transcription (Fig. 3a). For instance, the initiation factors TFIIB and Kin28 were enriched in promoter and promoter escape states (P2, PE1, PE2), and were depleted in states of other transcription phases (Fig. 3a). States related to the same transcription phase often peaked at successive genomic positions. For instance, the mid-elongation phase comprises successive states mE1-mE5 (Fig. 3b,c) that were characterized by a gradual decrease in the occupancy of initiation factors, capping-related factors, and Nrd1 (Fig. 3a).

The association of states to phases of transcription is in accordance with state-specific enrichment of DNA sequence motifs (Online Methods, Supplementary Information). While promoter state P2 shows enrichment of known promoter-associated motifs, termination state T1 is enriched with known termination signals and mixed state P/T2 contains both, promoter- and termination-associated motifs. We also found potentially unknown sequence motifs, which we could not associate to kown functions or binders (Supplementary Information). Overall, these results show that unsupervised bdHMM analysis can define meaningful genomic states that reflect phases of transcription at every single gene.

The transcription cycle shows gene-specific variation

Our bdHMM annotation did not only recapitulate known events during transcription, it also provided unexpected, new insights. For example, the flux diagram (Fig. 3c, showing the most likely transitions between successive states) indicated variability within the transcription cycle. We found different states at

the same position within genes that may reflect alternative functional transcription complexes (Promoter: P1, P2, P/T1, or P/T2; Promoter escape: PE1 or PE2; Fig. 3a,b). These alternative states are located within different branches of the flux diagram (Fig. 3c). A pronounced bifurcation occurs at the transition from P2 to promoter escape, entering either highly productive (PE2) or weak transcription (PE1). These two branches of the transcription cycle converge again during late elongation (lE2, lE3) or termination (T1). Hence, the analysis of state frequency distributions and transition diagrams suggests gene-specific variation of the transcription cycle.

For a systematic investigation of gene-specific variation during the transcription cycle, we clustered genes based on their annotated state path. To that end, the state paths of 4,263 genes were rescaled to a common length and clustered into 55 groups according to their Hamming distance (Fig. 4a,b, Online Methods). The obtained gene clusters show distinct patterns of protein occupancies suggesting mechanistic differences in transcription (Fig. 4, Supplementary Fig. 2 and below). Moreover, the gene clusters corresponded to distinct functional gene groups (Supplementary Table 1) and differed by gene length, expression level, and genomic context (e.g. termination overlaps with a neighboring downstream promoters or bidirectionality of promoters).

Cluster 14, which contains 694 genes (Fig. 4b,c, Supplementary Fig. S3), shows a transcription cycle most similar to the canonical one proposed previously[1]. In this cluster, the promoter escape state PE2 was characterized by peak occupancy of the Pol II core subunit Rpb3 between 100 and 200 bp downstream of the TSS, and phosphorylation of the CTD serine 2 residue reaches maximum levels between 600 and 1,000 bp (Fig. 4d), as observed in previous metagene analysis. The cycle ends with the canonical termination state T1, which is characterized by the presence of elongation factors Spn1, Paf1, Ctk1, Bur1, Spt16, Spt5, and termination factors Pcf11 and Rna15 (Fig. 3a).

Evidence for regulated promoter escape

We next analyzed clusters with variations compared to the canonical transcription cycle. Cluster 32 (43 genes) differs from the canonical cluster 14 in the transition from promoter escape to elongation. State frequency and gene-averaged ChIP signals suggest that transcription is attenuated after promoter escape in cluster 32 (Fig. 4b,c). In this cluster, a strong promoter escape (PE2) is followed by the weak elongation state eE3, which is characterized by low levels of Pol II and elongation factors (Fig. 4c). Moreover, elongation factors Ctk1 and Paf1 appear to be absent from those genes (Fig. 4c, Supplementary Fig. 2c). In contrast, cluster 14 exhibits similarly strong promoter escape yet transitions into the highly productive elongation states eE2 and mE1, which are characterized by high occupancies of all measured elongation factors (Fig. 3b,c). This comparison supports the existence of a regulatory checkpoint for transcription elongation after promoter escape. This is likely related to transcription attenuation with the help of the early termination factor Nrd1[18] (Fig. 4c, Supplementary Fig. 3). The individual occupancy profiles (Fig. 4c, Supplementary Fig. 2c) indicate that this checkpoint separates the binding events of Spt5, Spn1, Bur1, Spt16 from the binding of Ctk1 and Paf1. Thus, it appears that attenuated genes recruit early elongation factors including Spt5 and Spt16, but not the later factors Paf1 and Ctk1.

Evidence for distinct transcription mechanisms for highly expressed genes

Cluster 38 differs strikingly from the canonical transcription cycle during early elongation and termination (Fig. 4b, Supplementary Fig. 2d, 147 genes enriched for genes involved in translation, Supplementary Table 1, Online Methods). Cluster 38 is characterized by the high occupancy promoter state P/T1 (Fig. 3a) and by the early elongation state eE1 (for 58% of all cluster 38 genes, and in turn 48% of genes with eE1 state are in cluster 38). During early elongation, serine 2 phosphorylation levels increase more steeply than in cluster 14, indicating that productive elongation is reached earlier at those genes (Fig. 4d). Moreover, Pol II does not exhibit the typical occupancy peak 150 bp downstream of the TSS but

immediately reaches a stable high level (Fig. 4d). This profile could be the consequence of a lower drop-off rate at this position[1], a more constant elongation rate along the gene, or a high and uniform coverage by elongating polymerases. Specifically to cluster 38, a sharp decrease of the occupancy of essentially all factors is observed well-positioned at the stop codon. The data indicates that most factors (Cbp20, Nrd1, Ctk1, Paf1, S5P, S7P, Spt16 and Bur1) are then released, as their occupancy remains low after the stop codon. Moreover, the Pol II subunit Rpb3, the serine 2 phosphorylation, and the elongation factors Spt5 and Spn1 recover their occupancy levels at the pA site, suggesting a higher elongation rate for Pol II and that these factors stay bound to the transcription machinery within the 3' UTR. This indicates that the previously reported early release of elongation factors for ribosomal genes[1] is sharply positioned at the stop codon and also involves release of the cap-binding protein Cbp20, the early termination factor Nrd1, and dephosphorylation of the CTD residues Ser5 and Ser7. Taken together, cluster 38 suggests that highly expressed genes exhibit distinct transcription mechanisms, characterized by efficient factor recruitments during early elongation and specific processes of factor release around the stop codon.

Not all termination regions are depleted of nucleosomes

Nucleosome depletion has been reported at the 3'-end of genes[19]. However, cluster 19, whose 634 genes terminate in state T1, does not show nucleosome depletion in this region. In contrast, nucleosome depletion is a hallmark of all our promoter states. We therefore hypothesized that the termination of genes in clusters other than cluster 19 overlaps with promoters of downstream genes. Genes in clusters 1, 5, 6, 12, 32, 33, and 38 showed nucleosome-depleted termination states P/T1 and P/T2. Their termination regions indeed overlap with a downstream promoter, as indicated by TFIIB enrichment downstream of their pA site (Supplementary Fig. 2). This supports previous reports that nucleosome depletion is not an intrinsic mark of transcription termination[20]. Thus, bdHMM analysis of the genomic context allows distinguishing canonical binding patterns from spurious ones caused by spill-over effects from

neighboring genes.

Genome-wide prediction of bidirectional promoters and transcripts

The Viterbi sequence provided by the bdHMM can be searched by regular expressions (RegEx) to identify patterns in the state sequence that correspond to genomic features such as bidirectional promoters or coding transcripts (Fig. 5). To search for bidirectional promoters, we use a RegEx consisting of a promoter state flanked by an upstream transcript on the Crick strand and a downstream transcript on the Watson strand (Fig. 5a,b). Using this RegEx, we detected 1,076 bidirectional promoters in yeast (Fig. 5b), which agrees well with a previous estimate of 1,049 bidirectional promoters[17]. In order to re-annotate transcription throughout the yeast genome, we applied a RegEx approach (Fig. 5a), which predicted 6068 transcripts with a minimal length of 80 bp on both strands. Matching transcript boundaries to previously published ones[17], 3690 (72%) of all annotated protein-coding transcripts were recovered (best reciprocal hits, Online Methods). Most predicted TSS (72%) and pA sites (58%) were within 100 bp of the published ones[17] (Supplementary Fig. 4). Moreover, we found 32 novel transcripts (4 overlapping a coding region, 28 non-coding, Fig. 5c,d, Online Methods). This demonstrates that RegEx is a powerful tool to search and de novo annotate specific genomic features from the bdHMM state sequence. This result is of particular significance because the S. cerevisiae transcriptome has been thoroughly studied and annotated.

Comparison to standard HMM on chromatin states of human T-cells

We evaluated the performance of bdHMM on sequencing data and large genomes, by applying bdHMM to a dataset of 41 chromatin marks in human T-cells[11]. This dataset had been analyzed with a standard HMM approach (chromHMM). To handle the binarized chromatin marks data defined by Ernst and Kellis[11], we extended bdHMM and included binary emission distributions. We fixed the emission

distributions during bdHMM learning, allowing a direct comparison of bdHMM states to HMM states. Moreover, this ensured that differences in the result are only due to differences in the modeling of state transitions. We developed a directionality score (Online Methods) to decide that in the bdHMM, 36 out of a total of 51 chromHMM states are modeled as directed state pairs and 15 chromHMM states are modeled as undirected states. Consistently, we identified directed chromatin states around transcribed, but not at repressed or repetitive regions (Supplementary Figs. 6,7). Up to state directionality, 88% of state annotations agreed between the two methods at non-repressive genomic regions (Online Methods). Comparison of the chromHMM with the bdHMM transitions revealed that in chromHMM, transition probabilities between two states are similar in both directions (Supplementary Fig. 7b), whereas the bdHMM can resolve the true order of chromatin states (Fig. 6b, Supplementary Figure 7a). For example, transcribed states are accessed almost exclusively from TSS downstrem promoter states. Promoter states in turn are followed by a sequence of 5' proximal, spliced exon, 5' and distal transcription states, ending in states associated with transcription termination (Figure 6b). Analysis of promoter and transcribed state frequencies at the TSS showed that state annotations matched the reading (sense) direction of the transcribed loci (Figure 6c). Promoter states showed pronounced peaks in sense direction at the TSS, which are further downstream followed by high frequencies of (sense) 5' proximal transcribed states. We conclude that bdHMM significantly improves the annotation of the human epigenome, because it correctly recovers the flow of chromatin states as they occur during transcription.

2 Discussion

We introduced bidirectional Hidden Markov Models (bdHMMs), a method for *de novo* and unbiased inference of directed genomic states from genome-wide profiling data. In contrast to previously described HMM-based approaches, bdHMM explicitly models directed genomic processes and allows for

the integration of strand-specific experimental data such as RNA expression profiles together with non-strand-specific data, such as ChIP occupancy data. The open-source package STAN provides a fast, multiprocessing implementation that can process the human chromatin data set in less than one day.

Application of bdHMM analysis significantly improved insights into previously defined combinatorial chromatin marks[11], indicating the presence of directed chromatin state patterns around the transcribed, but not the repressed portion of the human genome. Our analysis of gene transcription in the budding yeast enabled us to automatically recover the majority of known and even new Pol II transcription units. We could assign different directed genomic states that are characterized by the presence of different transcription factors and Pol II CTD modification marks.

The most significant advance of bdHMM analysis over previous methods is its potential to *de novo* identify characteristic sequences (patterns) of directed states on the genome. These patterns identify gene-specific variation in transcription - or other directed processes - that were previously hidden by metagene analysis of experimental data. Metagene analysis derives only average profiles for groups of genes defined beforehand, and is thus biased towards annotated genes. In contrast, bdHMM allows investigating variations in the sequence of genomic states associated with transcription. This is done by first identifying distinct genomic states de novo and then clustering genes based on the succession of these genomic states. This analysis was consistent with a general transcription cycle and uniform transitions of a core Pol II transcription complex that occurs at all genes[1–3]. On the other hand, it also indicated gene-specific variations to the general transcription cycle, because the resulting clusters differed markedly in the sequence of their genomic states. First, a few dozen genes that apparently show Nrd1-mediated transcription attenuation are shown here to lack elongation factors Ctk1 and Paf1, suggesting that transcription attenuation occurs before Ctk1 and Paf1 are recruited. Second, we provide evidence for a distinct mechanism for highly expressed genes leading to the immediate recruitment of a full complement of Pol II-associated factors downstream of the transcription start site. Third, we found that nucleosome depletion is not a

necessary feature of transcription termination.

Altogether, we foresee bdHMM to be instrumental for studying gene transcription and other directional genomic processes, such as DNA replication, recombination or DNA repair by integrative analysis of genome-wide data.

3 Online Methods

Experimental data and preprocessing

The experimental yeast dataset was compiled from public data [1,14–17]. All measurements were done using the high density custom-made Affymetrix tiling array (PN 520055) which tiles each strands of genomic DNA in yeast at a resolution of 8bp. ChIP experiments were normalized using the R/Bioconductor [21,22] package Starr [23] as previously described [24]. Expression data was normalized using the tilingArray package [25].

The human chromatin modification dataset was downloaded from the supplemental website of Ernst and Kellis[11], where they provided the preprocessed Sequencing and binary data..

The bidirectional hidden Markov model

Bidirectional hidden Markov models belong to the class of hidden Markov models (HMMs). It is therefore beneficial to introduce HMMs first, along with some notation.

Definition. A **hidden Markov model** (HMM) is a tuple $\theta = (\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ such that

1. $\mathcal{K}$ is a finite set, the elements of which are called *states*.

2. The *initial state distribution* $\pi = (\pi_i)_{i \in \mathcal{K}}$ is a probability (row) vector, i.e., $0 \leq \pi_i \leq 1$, $i \in \mathcal{K}$, and $\sum_{i \in \mathcal{K}} \pi_i = 1$.

3. The *transition matrix* $A = (a_{ij})_{i,j \in \mathcal{K}}$ is a $\mathcal{K} \times \mathcal{K}$ stochastic matrix, i.e., each row of A is a probability vector.

4. The *emission distributions* $\Psi = \{\psi_i;\ i \in \mathcal{K}\}$ form a set of probability distributions on a space $\mathcal{D}$, the *space of observations*.

An HMM defines a probability distribution on a sequence of observations $\mathcal{O} = (o_0, ..., o_T)$ of length T. It assumes that each observation o_t is *emitted* by a corresponding hidden (unobserved) state variable s_t which can assume values in $\mathcal{K}$. The value of s_t determines the probability of observing o_t by $\Pr(o_t \mid s_t) = \psi_{s_t}(o_t)$. The hidden variables are assumed to form a homogenous Markov chain $\mathcal{S} = (s_0, ..., s_T)$, with (time-independent) transition probabilities $\Pr(s_t = j \mid s_{t-1} = i) = a_{ij}$, $i, j \in \mathcal{K}$, $t = 1, ..., T$, and with initial state distribution $\Pr(s_1 = i) = \pi_i$, $i \in \mathcal{K}$. The (full) likelihood of an HMM is

$$\begin{aligned} \Pr(\mathcal{O}, \mathcal{S};\ \theta) &= \Pr(\mathcal{O} \mid \mathcal{S};\ \theta) \cdot \Pr(\mathcal{S};\ \theta) \\ &= \prod_{t=0}^{T} Pr(o_t \mid s_t;\ \Psi) \cdot \prod_{t=1}^{T} Pr(s_t \mid s_{t-1};\ A) \cdot Pr(s_1;\ \pi) \\ &= \prod_{t=0}^{T} \psi_{s_t}(o_t) \cdot \prod_{t=1}^{T} a_{s_{t-1} s_t} \cdot \pi_{s_1} \end{aligned} \tag{3.1}$$

A bdHMM is an HMM which satisfies three additional conditions. The first two conditions deal with the structure of the underlying hidden Markov chain, and the last condition consider the nature of observations.

Definition. A **bidirectional hidden Markov mode**l (bdHMM) is a tuple $\theta = ((\mathcal{K}, i_{\mathcal{K}}), \pi, A, (\mathcal{D}, i_{\mathcal{D}}), \Psi)$ such that $(\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ is an HMM, $i_{\mathcal{K}} : \mathcal{K} \to \mathcal{K}, k \mapsto \bar{k}$ and $i_{\mathcal{D}} : \mathcal{D} \to \mathcal{D}$, $o \mapsto \bar{o}$ are involutions ($i_{\mathcal{K}}^2 = \mathrm{id}$, $i_{\mathcal{D}}^2 = \mathrm{id}$), and the following symmetry conditions hold:

1. Generalized detailed balance relation: The transition matrix A and the initial state distribution π satisfy

$$\pi_i a_{ij} = \pi_{\bar{j}} a_{\bar{j}\bar{i}} \quad , \, i, j \in \mathcal{K} \tag{3.2}$$

2. Initiation symmetry: The initial state distributionπ satisfies

$$\pi_i = \pi_{\bar{i}} \quad , \, i \in \mathcal{K} \tag{3.3}$$

3. Observation symmetry: Ψ satisfies

$$\psi_i(o) = \psi_{\bar{i}}(\bar{o}) \quad , \, i \in \mathcal{K}, \, o \in \mathcal{D} \tag{3.4}$$

Let $\theta = ((\mathcal{K}, i_{\mathcal{K}}), \pi, A, (\mathcal{D}, i_{\mathcal{D}}), \Psi)$ a bdHMM. Then by initiaion symmetry and generalized detailed balance,

$$(\pi A)_j = \sum_{i \in \mathcal{K}} \pi_i a_{ij} = \sum_{i \in \mathcal{K}} \pi_{\bar{j}} a_{\bar{j}\bar{i}} = \pi_{\bar{j}} = \pi_j \quad , \, j \in \mathcal{K},$$

which proves $\pi A = \pi$. In other words, the initial state distributionπ of a bdHMM is always a stationary state distribution of A.

The semantic of bdHMMs Bidirectional HMMs model directional processes in a sequence of observations. It is reasonable to expect that an observation contains information about the directionality of the underlying process that generated it. The involution $i_{\mathcal{D}}$ is meant to map an observation $o \in \mathcal{D}$ to its so-called conjugate observation $\bar{o} = i_{\mathcal{D}}(o)$, which denotes the corresponding observation that one would make if the observation sequence were viewed from the opposite direction. E.g., in the case of genomic measurements, $\mathcal{D}$ is modeled as $\mathcal{D} = \mathcal{D}^0 \times \mathcal{D}^+ \times \mathcal{D}^-$, the Cartesian product of a space $\mathcal{D}^0$ of non strand-specific observations (e.g. ChIP measurements of protein binding), a space $\mathcal{D}^+$of forward strand-specific observations (like RNA transcription originating from the forward strand), and a corresponding set $\mathcal{D}^-$ of reverse strand-specific observations. The forward and reverse strand-specific observations are paired in the sense that $\mathcal{D}^+ = \mathcal{D}^-$. The involution $i_{\mathcal{D}}$ acts as the identity on $\mathcal{D}^0$ and it swaps the strand-specific observations, $i_{\mathcal{D}} : o = (o^0, o^+, o^-) \mapsto \bar{o} = (o^0, o^-, o^+)$. In hidden Markov models, observations will be emitted from hidden states that may indicate typical processes occurring in forward or in reverse direction, or undirectional processes. The involution $i_{\mathcal{K}}$ splits the states $\mathcal{K}$ of the HMM into undirectional states (denoted by $\mathcal{K}^0$) - the fixed points $k = \bar{k}$ of i_k - and directional states which occur in pairs $(k, \bar{k})$, $k \neq \bar{k}$ of 'conjugate' or 'twin' states. One member of such a pair is deemed to be involved in forward, the other in reverse directional processes (note that at this point we do not specify which of the two does what). The forward states are typically denoted by $\mathcal{K}^+$, the reverse states by $\mathcal{K}^-$. Observation symmetry states that conjugate directional states encode essentially the same probability distribution, up to reversal of the observations.

Note that if $i_{\mathcal{K}} = \mathrm{id}$ is the identity map, condition (3) is void, and condition (2) reduces to the common detailed balance relation for reversible HMMs. If additionally the involution $i_{\mathcal{D}}$ is the identity map, condition (5) is also void. Thus, a bdHMM $\theta = ((\mathcal{K}, \mathrm{id}), \pi, A, (\mathcal{D}, \mathrm{id}), \Psi)$ is nothing but a reversible HMM, i.e., an HMM which additionally satisfies the detailed balance relation $\pi_i a_{ij} = \pi_j a_{ji}$, $i, j \in \mathcal{K}$. It

follows that our algorithms for bdHMM learning will immediately apply to reversible HMMs.

Given an observation sequence $\mathcal{O} = (o_t)_{t=1,...,T}$, let $\mathcal{O}^{rev} = (o_t^{rev} = \bar{o}_{T+1-t})_{t=1,...,T}$ denote the 'reversed' observation sequence obtained by taking conjugates of all observations and reversing their order. Similarly, given a hidden state sequence $\mathcal{S} = (s_1, ..., s_T)$, let $\mathcal{S}^{rev} = (s_t^{rev} = \bar{s}_{T+1-t})_{t=1,...,T}$ denote the 'reversed' hidden state sequence. Verify that

$$\begin{aligned}
\Pr(\mathcal{S};\theta) &= \pi_{s_1}\prod_{t=1}^{T} a_{s_t s_{t-1}} = \pi_{s_T}\prod_{t=1}^{T} a_{\bar{s}_{t-1}\bar{s}_t} \\
&\overset{(3.3,3.2)}{=} = \pi_{\bar{s}_T}\prod_{t=1}^{T} a_{\bar{s}_{T+1-(t-1)}\bar{s}_{T+1-t}} = \pi_{s_1^{rev}}\prod_{t=1}^{T} a_{s_{t-t}^{rev}s_t^{rev}} \\
&= \Pr(\mathcal{S}^{rev};\theta)
\end{aligned} \tag{3.5}$$

Moreover,

$$\begin{aligned}
\Pr(\mathcal{O}\mid\mathcal{S};\theta) &= \prod_{t=0}^{T}\psi_{s_t}(o_t) \overset{(3.4)}{=} \prod_{t=0}^{T}\psi_{\bar{s}_t}(\bar{o}_t) \\
&= \prod_{t=0}^{T}\psi_{\bar{s}_{T+1-t}}(\bar{o}_{T+1-t}) = \prod_{t=0}^{T}\psi_{s_t^{rev}}(o_t^{rev}) \\
&= \Pr(\mathcal{O}^{rev}\mid\mathcal{S}^{rev};\theta)
\end{aligned} \tag{3.6}$$

Equations (3.5) and 3.6) imply

$$\Pr(\mathcal{O},\mathcal{S};\,\theta) = \Pr(\mathcal{O}\mid S;\,\theta)\cdot\Pr(\mathcal{S};\,\theta) = \Pr(\mathcal{O}^{rev}\mid S^{rev};\,\theta)\cdot\Pr(\mathcal{S}^{rev};\,\theta) = \Pr(\mathcal{O}^{rev},\mathcal{S}^{rev};\,\theta) \tag{3.7}$$

and

$$\Pr(\mathcal{S}\mid\mathcal{O};\,\theta) = \Pr(\mathcal{S}^{rev}\mid\mathcal{O}^{rev};\,\theta) \tag{3.8}$$

Finally, a bdHMM is reversible in the generalized sense,

$$\begin{aligned} \Pr(\mathcal{O};\,\theta) &= \sum_{\mathcal{S}} \Pr(\mathcal{O},\mathcal{S};\,\theta) = \sum_{\mathcal{S}} \Pr(\mathcal{O}^{rev},\mathcal{S}^{rev};\,\theta) \\ &= \sum_{\mathcal{S}^{rev}} \Pr(\mathcal{O}^{rev},\mathcal{S}^{rev};\,\theta) = \Pr(\mathcal{O}^{rev};\,\theta) \end{aligned} \tag{3.9}$$

The second-last equality in (3.9) holds because if $\mathcal{S}$ runs over all possible state sequences, then so does $\mathcal{S}^{rev}$. The need for a model satisfying the natural condition (3.7) motivated the development of bdHMMs, and indeed condition (3.7) is almost their defining property: We mention without proof that under very mild assumptions on the probability distributions Ψ, any HMM satisfying (3.7) is a bdHMM.

Learning of the transition matrix and the initial state distribution The learning problem for bdHMMs consists in maximizing the marginal likelihood of the model,

$$\hat{\theta} = \underset{\theta}{argmax}\, \Pr(\mathcal{O};\,\theta)$$

Parameter estimation in an HMM is commonly done using the Baum-Welch algorithm[26], an expectation-maximization (EM) algorithm[27]. The EM algorithm is an iterative procedure in which a target function $Q\left(\theta;\,\theta^{old}\right)$ is maximized with respect to the parameters θ, given a previous parameter guess θ^{old} . This algorithm will converge to a local maximum of the marginal likelihood $P(\mathcal{O};\,\theta)$. We will derive an EM algorithm for the learning of the bdHMM parameters A, π.

Let $\theta^{old} = ((\mathcal{K}, i_{\mathcal{K}}), \pi^{old}, A^{old}, (\mathcal{D}, i_{\mathcal{D}}), \Psi^{old})$ be a bdHMM. Let $\mathcal{O} = (o_1, ..., o_T)$ be a sequence of

observations. For $i, j \in \mathcal{K}$, $t = 1, ..., T$, we define the posterior probabilities

$$\zeta_t(i, j,) = \Pr(s_{t-1} = i, s_t = j \mid \mathcal{O};\, \theta^{old}) \tag{3.10}$$

$$\gamma_t\,(i) = \Pr\left(s_t = i | \mathcal{O}, \theta^{old}\right) \tag{3.11}$$

These posterior probabilities can be calculated efficiently using the forward probabilities $\alpha_t(i) = \Pr(s_t = i, o_1, ..., o_t;\, \theta^{old})$ and the backward probabilities $\beta_t(j) = \Pr(o_{t+1}, ..., o_T \mid s_t = j;\, \theta^{old})$, $i, j \in \mathcal{K}$. Forward and backward probabilities are calculated recursively.

$$\begin{aligned}
\alpha_t(i) &= \Pr(s_t = i, o_1, ..., o_t;\, \theta^{old}) \\
&= \sum_{k \in \mathcal{K}} \Pr(s_{t-1} = k, s_t = i, o_1, ..., o_t;\, \theta^{old}) \\
&= \sum_{k \in \mathcal{K}} \Pr(o_t \mid s_t = i;\, \theta^{old}) \cdot \Pr(s_t = i \mid s_{t-1} = k;\, \theta^{old}) \cdot \Pr(s_{t-1} = k, o_1, ..., o_{t-1};\, \theta^{old}) \\
&= \psi_i^{old}(o_t) \sum_{k \in \mathcal{K}} a_{ki}^{old} \alpha_{t-1}(k)
\end{aligned} \tag{3.12}$$

for $t = 1, ..., T$, and $\alpha_0(i) = \pi_i^{old} \psi_i^{old}(o_0)$. Similarly for the backward probabilities,

$$\begin{aligned}\beta_t(j) &= \Pr(o_{t+1}, ..., o_T \mid s_t = j;\, \theta^{old}) \\ &= \sum_{k\in\mathcal{K}} \Pr(o_{t+1} \mid s_{t+1} = j;\, \theta^{old}) \cdot \Pr(s_{t+2} = k \mid s_{t+1} = j;\, \theta^{old}) \cdot \Pr(o_{t+2}, ..., o_T \mid s_{t+1} = k;\, \theta^{old}) \\ &= \psi_j^{old}(o_t) \sum_{k\in\mathcal{K}} a_{jk}^{old} \beta_{t+1}(k) \end{aligned} \tag{3.13}$$

for $t = T-1, ..., 0$, and $\beta_T(j) = \psi_j^{old}(o_T)$. It follows that

$$\zeta_t(i,j) = \frac{\alpha_t(i) a_{ij}^{old} \beta_{t+1}(j) \psi_j^{old}(o_{t+1})}{\sum_{k\in\mathcal{K}} \alpha_t(k)\beta_t(k)} \tag{3.14}$$

and

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{k\in\mathcal{K}} \alpha_t(k)\beta_t(k)} \tag{3.15}$$

Note that the quantities $\zeta_t(i,j)$ and $\gamma_t(i)$ are always non-negative. The target function $Q\left(\theta;\, \theta^{old}\right)$ is defined as the expectation of the log likelihood $\Pr(\mathcal{O}, \mathcal{S};\, \theta)$, where expectation is taken with respect to the unknown hidden state sequence $\mathcal{S}$ and its posterior distribution $\Pr(\mathcal{S} \mid \mathcal{O};\, \theta^{old})$,

$$\begin{aligned} Q\left(\theta;\, \theta^{old}\right) &= \sum_{\mathcal{S}} \Pr\left(\mathcal{S}|\mathcal{O};\, \theta^{old}\right) \log \Pr(\mathcal{O}, \mathcal{S};\, \theta) \\ &= \sum_{\mathcal{S}} \Pr\left(\mathcal{S}|\mathcal{O};\, \theta^{old}\right) \left\{ \log \pi_{s_1} + \sum_{t=1}^{T} \log a_{s_{t-1}s_t} + \sum_{t=0}^{T} \log \psi_k(o_t) \right\} \end{aligned} \tag{3.16}$$

It can be shown that $Q\left(\theta;\, \theta^{old}\right)$ is a lower bound of the marginal likelihood function $\Pr(\mathcal{O};\, \theta)$ which touches the likelihood function at $\theta = \theta^{old}$, i.e., $Q(\theta^{old};\, \theta^{old}) = \Pr(\mathcal{O};\, \theta^{old})$[27]. These properties guarantee that the iterative maximization of Q leads to a local maximum of $\Pr(\mathcal{O};\, \theta)$. We want to maximize Q with respect to A and π under the constraints of a bdHMM. Using the posterior probabilities

(3.10) and (3.11), and summarizing the ψ_k terms into one constant c which does not depend on A or π, the modified target function Q assumes a convenient form.The quantity Q is calculated in the E-step,

$$\begin{aligned} Q\left(\theta;\, \theta^{old}\right) &= \sum_{\mathcal{S}} \Pr\left(\mathcal{S}|\mathcal{O};\, \theta^{old}\right) \left\{\sum_{t=1}^{T} \log a_{s_{t-1}s_t}\right\} + \sum_{\mathcal{S}} \Pr\left(\mathcal{S}|\mathcal{O};\, \theta^{old}\right) \{\log \pi_{s_1}\} + c \\ &= \sum_{s_{t-1}\in\mathcal{K}} \sum_{s_t\in\mathcal{K}} \Pr\left(s_{t-1}, s_t|\mathcal{O}, \theta^{old}\right) \left\{\sum_{t=1}^{T} \log a_{s_{t-1}s_t}\right\} + \sum_{s_{t-1}\in\mathcal{K}} \Pr\left(s_1|\mathcal{O}, \theta^{old}\right) \{\log \pi_{s_1}\} + c \\ &= \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{K}} \sum_{t=1}^{T} \zeta_t\left(k,l\right) \log a_{kl} + \sum_{k\in\mathcal{K}} \gamma_1(k) \log \pi_{s_1} + c \end{aligned} \tag{3.17}$$

We calculate the Jacobian matrix and the Hessian matrix of Q and show that Q is a convex function.

$$\frac{\partial}{\partial a_{ij}} \tilde{Q}\left(\theta;\, \theta^{old}\right) = \frac{1}{a_{ij}} \sum_{t=1}^{T} \zeta_t\left(i,j\right) \tag{3.18}$$

$$\frac{\partial}{\partial a_{kl}} \frac{\partial}{\partial a_{ij}} \tilde{Q}\left(\theta;\, \theta^{old}\right) = \begin{cases} -\frac{1}{a_{ij}^2} \sum_{t=1}^{T} \zeta_t\left(i,j\right) \leq 0 & \text{if } (k,l) = (i,j) \\ 0 & \text{else} \end{cases} \tag{3.19}$$

The Hessian matrix is a diagonal matrix with non-positive diagonal entries, hence it is negative semidefinite. This means that Q is concave. The maximization of Q is performed under the constraints that π is a probability vector, A is a stochastic matrix, and that initiation symmetry and generalized detailed balance holds. Unfortunately, these constraints define a non-convex optimization domain. Still, powerful numerical solvers for concave functions exist. In our case, we used the *ipopt* solver[28] and Rsolnp[29]. Transition probabilities might become very small or even 0, which may cause problems for the optimization since the lower boundary for the parameters is 0. Numerical optimizers tend to become very slow or even fail to converge at the boundary of the solution space. To ensure numerical stability and proper convergence,

we set state transitions $a_{ij} = 0$ that drop below a certain cutoff $\sum_{t=1}^{T} \xi_t(i,j) < c$. When the algorithm approximates a point of convergence it becomes less and less likely for a transition to be removed. The EM-algorithm will find an optimal point with the additional constraints that some transitions are 0. The numerical optimization approach becomes slow for very large data sets and for a high number of hidden states. In our second approach, we therefore introduce a modified lower bound function $\tilde{Q}(\theta;\, \theta^{old})$ which can be maximized analytically and hence very efficiently. We iterate this maximization process in the same fashion as in the EM algorithm. Although we were not able to prove convergence of the parameter sequence, this was always the case in practice. Moreover, the results obtained by our heuristic were always identical to those obtained by the numerical solver. Our heuristic is substantially faster, for our yeast data with $|\mathcal{K}| = 20$ states, we achieved an acceleration by a factor of about 25.

Given a bdHMM parameter set $\theta = ((\mathcal{K}, i_{\mathcal{K}}), \pi, A,\ (\mathcal{D}, i_{\mathcal{D}}), \Psi)$, denote by $\bar{\theta} = ((\mathcal{K}, i_{\mathcal{K}}), \bar{\pi}, \bar{A},\ (\mathcal{D}, i_{\mathcal{D}}), \bar{\Psi})$ the bdHMM parameter set defined by $\bar{\pi}_i = \pi_{\bar{i}}$, $\bar{a}_{ij} = a_{\bar{i}\bar{j}}$, $\bar{\psi}_i(o) = \psi_{\bar{i}}(o)$, $i, j \in \mathcal{K}$, $o \in \mathcal{D}$. The modified target function is defined as

$$\tilde{Q}\left(\theta;\, \theta^{old}\right) = Q(\theta;\, \theta^{old}) + Q(\theta;\, \bar{\theta}^{old}) \tag{3.20}$$

where Q is defined in the original sense (3.16). Since both Q terms in the sum in (3.20) are, up to some additive constant, lower bounds of the marginal likelihood function $\Pr(\mathcal{O};\, \theta)$, so is $\frac{1}{2}\tilde{Q}(\theta;\, \theta^{old})$.

For $\mathcal{S} = (s_1, ..., s_T)$ let $\bar{\mathcal{S}} = (\bar{s}_1, ..., \bar{s}_T)$. It is elementary to verify that

$$\begin{aligned} \Pr(\mathcal{O}, \mathcal{S};\, \theta) &= \pi_{s_1} \prod_{t=1}^{T} a_{s_{t-1}s_t} \prod_{t=0}^{T} \psi_{s_t}(o_t) \\ &= \bar{\pi}_{\bar{s}_1} \prod_{t=1}^{T} \bar{a}_{\bar{s}_{t-1}\bar{s}_t} \prod_{t=0}^{T} \bar{\psi}_{\bar{s}_t}(o_t) \;=\; \Pr(\mathcal{O}, \bar{\mathcal{S}};\, \bar{\theta}) \end{aligned} \tag{3.21}$$

From (3.21) we deduce that

$$\Pr(s_{t-1} = i, s_t = j \mid \mathcal{O};\, \bar{\theta}^{old}) = \Pr(s_{t-1} = \bar{i}, s_t = \bar{j} \mid \mathcal{O};\, \theta^{old}) = \zeta_t(\bar{i}, \bar{j}) \tag{3.22}$$

and

$$\begin{aligned} Q(\theta;\, \bar{\theta}^{old}) &= \sum_{s_{t-1}\in\mathcal{K}} \sum_{s_t\in\mathcal{K}} \Pr\left(s_{t-1}, s_t|\mathcal{O}, \bar{\theta}^{old}\right) \left\{\sum_{t=1}^{T} \log a_{s_{t-1}s_t}\right\} + c \\ &= \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{K}} \left(\sum_{t=1}^{T} \zeta_t\left(\bar{k}, \bar{l}\right)\right) \log a_{kl} + c \end{aligned} \tag{3.23}$$

Equations (3.17) and (3.23) imply

$$\begin{aligned} \tilde{Q}\left(\theta;\, \theta^{old}\right) &= Q(\theta;\, \theta^{old}) + Q(\theta;\, \bar{\theta}^{old}) + c \\ &= \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{K}} \sum_{t=1}^{T} \left(\zeta_t\left(k, l\right) + \zeta_t\left(\bar{l}, \bar{k}\right)\right) \log a_{kl} + c \end{aligned}$$

To maximize $\tilde{Q}$ under the constraint(s) that A is a stochastic matrix, we introduce Lagrange multipliers $\lambda_k \left(1 - \sum_{l\in\mathcal{K}} a_{kl}\right)$, $k \in \mathcal{K}$, and rewrite $\tilde{Q}$ as

$$\tilde{Q}\left(\theta;\, \theta^{old}\right) = \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{K}} \sum_{t=1}^{T} \left(\zeta_t\left(k, l\right) + \zeta_t\left(\bar{l}, \bar{k}\right)\right) \log\left(a_{kl}\right) + \sum_{k\in\mathcal{K}} \lambda_k \left(1 - \sum_{l\in\mathcal{K}} a_{kl}\right) + c \tag{3.24}$$

For $i, j \in \mathcal{K}$, we set the partial derivatives of $\tilde{Q}$ with respect to a_{ij} to zero,

$$0 = \frac{\partial}{\partial a_{ij}} \tilde{Q}\left(\theta;\, \theta^{old}\right) = \frac{1}{a_{ij}} \sum_{t=1}^{T} \left(\zeta_t\left(i, j\right) + \zeta_t\left(\bar{j}, \bar{i}\right)\right) - \lambda_i \tag{3.25}$$

Multiplication by a_{ij} and summation over all equations $j \in \mathcal{K}$ leads to

$$\begin{aligned}
\lambda_i \underbrace{\sum_{j\in\mathcal{K}} a_{ij}}_{1} &= \sum_{t=2}^{T}\Bigg(\underbrace{\sum_{j\in\mathcal{K}} \zeta_t(i,j)}_{\gamma_{t-1}(i)} + \underbrace{\sum_{j\in\mathcal{K}} \zeta_t(\bar{j},\bar{i})}_{\gamma_t(\bar{i})}\Bigg) \\
\lambda_i &= \sum_{t=2}^{T}\left(\gamma_{t-1}(i) + \gamma_t(\bar{i})\right) \qquad (3.26)
\end{aligned}$$

After substitution of (3.26) into (3.25), we solve for a_{ij}.

$$a_{ij} = \frac{1}{\lambda_i}\sum_{t=2}^{T}\sum_{t=2}^{T}\left(\zeta_t(i,j) + \zeta_t(\bar{j},\bar{i})\right) = \frac{\sum_{t=1}^{T}\left(\zeta_t(i,j) + \zeta_t(\bar{j},\bar{i})\right)}{\sum_{t=1}^{T}\left(\gamma_{t-1}(i) + \gamma_t(\bar{i})\right)} \quad , \; i,j \in \mathcal{K} \qquad (3.27)$$

Let

$$\pi_i = \frac{1}{2T}\sum_{t=1}^{T}\left(\gamma_{t-1}(i) + \gamma_t(\bar{i})\right) \quad , \; i \in \mathcal{K}$$

Then π is a probability vector which together with A satisfies detailed balance,

$$\pi_i a_{ij} = \frac{1}{2T}\sum_{t=1}^{T}\left(\zeta_t(i,j) + \zeta_t(\bar{j},\bar{i})\right) = \pi_{\bar{j}} a_{\bar{j}\bar{i}} \quad , \; i,j \in \mathcal{K}$$

Further,

$$|\pi_i - \pi_{\bar{i}}| = \frac{1}{2T}\left\|\gamma_T(\bar{i}) - \gamma_T(i) + \gamma_0(i) - \gamma_0(\bar{i})\right\| \leq \frac{1}{T} \quad , \; i \in \mathcal{K}$$

Although the vector π does not exactly satisfy initiation symmetry, the amount by which this symmetry is violated is generally substantially smaller than $\frac{1}{T}$. This difference is negligible for large T, i.e., for long observation sequences.

We have developed two strategies: The first, computer-intensive strategy is to do numerical optimization using standard solvers; the second strategy is a fast heuristic. Both methods in practice lead to the same results, and they are implemented in our R/Bioconductor software package STAN.

Estimation of the emission probabilities The emissison distributions Ψ are also updated by maximizing the original target function Q in Equation (3.16) Summarizing irrelevant termis in a constant c, we have

$$Q\left(\theta, \theta^{old}\right) = \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} \gamma_t\left(k\right) \log\left(\psi_k\left(o_t\right)\right) + c$$

We assume multivariate Gaussian emission probabilities, $\psi_i(o_t) = \mathcal{N}(o_t; \mu^i, \Sigma^{(i)})$, $i \in \mathcal{K}$, with mean $\mu^i \in \mathbb{R}^D$ and covariance matrix $\Sigma^{(i)} \in \mathbb{R}^{D \times D}$. We have implemented bdHMM with multivariate Gaussian emission probabilities, since they are appropriate distributions for microarray data on a log or quasi-log scale [30]. Moreover, the covariance matrix of multivariate Gaussians allows modeling correlations between factors in each state. This is important because factor occupancies tend to scale with the gene expression level. Such dependencies are captured by the covariance matrix, Application to sequencing-based datasets can be done by transforming the data such that it approximately follows a normal distribution [5,10].

Setting the partial derivatives $\frac{\partial Q\left(\theta, \theta^{old}\right)}{\partial \mu_d^i}$, $i \in \mathcal{K}$, $d \in D$, to zero and solving this equation system for μ^i leads to (see Supplementary Information):

$$\hat{\mu}^i = \begin{cases} \frac{\sum_{t=1}^{T}[\gamma_t(i) o_t + \gamma_t(\bar{i}) \bar{o}_t]}{\sum_{t=1}^{T}[\gamma_t(i) + \gamma_t(\bar{i})]} & \textbf{if } i\textbf{is directed} \\ \frac{\sum_{t=1}^{T} \gamma_t(i) o_t}{\sum_{t=1}^{T} \gamma_t(i)} & \textbf{if } i\textbf{is undirected} \end{cases}$$

Analoguously, Setting the partial derivatives $\frac{\partial Q\left(\theta,\theta^{old}\right)}{\partial \Sigma^i}$, $i \in \mathcal{K}$, $c, d \in D$, to zero and solving this equation system for Σ^i leads to (see Supplementary Information):

$$\hat{\Sigma}^i = \begin{cases} \frac{\sum_{t=1}^{T}\left[\gamma_t(i)\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T+\gamma_t(\bar{i})\left(\bar{o}_t-\mu^{\bar{i}}\right)\left(\bar{o}_t-\mu^{\bar{i}}\right)^T\right]}{\sum_{t=1}^{T}[\gamma_t(i)+\gamma_t(\bar{i})]} & \text{if } i\text{is directed} \\ \frac{\sum_{t=1}^{T}\gamma_t(i)\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T}{\sum_{t=1}^{T}\gamma_t(i)} & \text{if } i\text{is undirected} \end{cases}$$

bdHMM learning without strand-specific observations A bdHMM can even be learned from entirely strand-unspecific data ($i_{\mathcal{D}} = \text{id}$). However forward and reverse states are unidentifiable under these conditions, because $\Pr(\mathcal{O};\theta) = \Pr(\mathcal{O};\bar{\theta})$, . It is necessary to a priori annotate some positions with proper directions. We introduce the flag sequence $\mathcal{F} = (f_1, ..., f_T)$, $f_t \subset \mathcal{K}$, which lists the states f_t that are allowed at a position t. We then set

$$\Pr\left(o_t|s_t = i;\ \Psi, \mathcal{F}\right) \quad = \quad \begin{cases} \psi_i\left(o_t\right) & \text{if } i \in f_t \\ 0 & \text{else} \end{cases},$$

ignoring this does not define a probability function for $i \notin f_t$.

In the context of transcription data, non-overlapping genes can be used to set flags allowing only forward (respectively reverse) and undirectional states.

De novo inference of state directionality

Let k be a directed state in a bdHMM. We introduce dir_k, a measure for the directionality of state k which is based on the posterior probabilities for observing k respectively its conjugate $\bar{k}$ at positions $t = 0, ..., T$.

$$dir_k = \frac{\sum_{t=0}^{T} \left| \Pr\left(s_t = k | \mathcal{O}, \theta\right) - \Pr\left(s_t = \bar{k} | \mathcal{O}, \theta\right) \right|}{\sum_{t=0}^{T} \left(\Pr\left(s_t = k | \mathcal{O}, \theta\right) + \Pr\left(s_t = \bar{k} | \mathcal{O}, \theta\right) \right)} \tag{3.28}$$

The score will be low if the differences in the probability for observing the forward twin state and the probability for observing the respective reverse twin state is low. It, will be high if this differences is large and thus the directionality of twin states is well distinguishable. In order to account for the overall probability of state k, the sum of absolute differences in the nominator in (3.28) is normalized by the sum over all positions t of the posterior probabilities for observing k or $\bar{k}$. The direcitionality score is used to infer whether a directed state pair $(k, \bar{k})$ of a bdHMM truely contains directional information, or whether it should be collapsed into one undirected state of a new bdHMM. Our rule of thumb is to collapse a directed state pair if $dir_k < 0.5$ (see also Results and Supplementary Fig. 6).

Simulations

The performance of bdHMM regarding parameter inference and state annotation on data not used for training was assessed using simulated datat set. For this purpose, we construct a transition matrix $A = (a_{ij})_{i,j \in \mathcal{K}}$ and an initial state distribution $\pi = (\pi_i)_{i \in \mathcal{K}}$ which satisfy generalized detailed balance and initiation symmetry. Choose an arbitrary transition matrix $A^* = \left(a^*_{ij}\right)_{i,j \in \mathcal{K}}$ and a stationary distribution $\pi^* = (\pi^*_i)_{i \in \mathcal{K}}$, $\pi^* A^* = \pi^*$.

$$a_{ij} = \frac{\pi_i^* a_{ij}^* + \pi_{\bar{j}}^* a_{\bar{j}i}^*}{\pi_i^* + \pi_{\bar{i}}^*}$$

$$\pi_i = \frac{1}{2}(\pi_i^* + \pi_{\bar{i}}^*)$$

Verify that π is a probability vector that satisfies initiation symmetry:

$$\sum_{i \in \mathcal{K}} \pi_i = \frac{1}{2} \sum_{i \in \mathcal{K}} \pi_i^* + \frac{1}{2} \sum_{i \in \mathcal{K}} \pi_{\bar{i}}^* = \frac{1}{2} + \frac{1}{2} = 1$$

$$\pi_{\bar{i}} = \frac{1}{2}(\pi_{\bar{i}}^* + \pi_{\bar{\bar{i}}}^*) = \frac{1}{2}(\pi_i^* + \pi_{\bar{i}}^*) = \pi_I$$

Furter, A is a stochastic matrix,

$$\begin{aligned}
\sum_{j \in \mathcal{K}} a_{ij} &= \frac{1}{\pi_i^* + \pi_{\bar{i}}^*} \left(\sum_{j \in \mathcal{K}} \pi_i^* a_{ij}^* + \sum_{j \in \mathcal{K}} \pi_{\bar{j}}^* a_{\bar{j}i}^* \right) \\
&= \frac{1}{\pi_i^* + \pi_{\bar{i}}^*} \left(\pi_i^* + (\pi^* A^*)_{\bar{i}} \right) \\
&= \frac{1}{\pi_i^* + \pi_{\bar{i}}^*} \left(\pi_i^* + \pi_{\bar{i}}^* \right) \; = \; 1
\end{aligned}$$

and A together with π satisfy generalized detailed balance,

$$\pi_i a_{ij} = \frac{1}{2}\pi_i^* a_{ij}^* + \pi_{\bar{j}}^* a_{\bar{j}\bar{i}}^* = \pi_{\bar{j}} a_{\bar{j}\bar{i}}$$

We mention that A is ergodic if A^* is ergodic.

To make our simulations realistic, we sample A^* as follows: Introduce an arbitrary linear order '$\leq$' on $\mathcal{K}^+$ (this order is meant to describe the preferential order of events for the directional states). Then,

$$a_{ij}^* \sim \begin{cases} \mathcal{U}(0.95, 0.99) & \text{if } i = j \\ \mathcal{U}(0.1, 0.7) & \text{if } (i, j \in \mathcal{K}^+ \wedge j > i) \vee (i, j \in \mathcal{K}^- \wedge j < i) \\ \mathcal{U}(0.01, 0.05) & \text{if } (i, j \in \mathcal{K}^+ \wedge j < i) \vee (i, j \in \mathcal{K}^- \wedge j > i) \\ \mathcal{U}(0.001, 0.02) & \text{if } i = \bar{j} \\ \mathcal{U}(0.001, 0.005) & \text{else} \end{cases}$$

where $\mathcal{U}(a, b)$ is the uniform distribution with lower bound a and upper bound b. Rows of A^* are then normalized to sum up to 1. An example of a simulated transition matrix is shown in Supplementary Figure 8. To get realistic simulations, emission distributions were simulated from fitted emissions of the yeast data set, using five non-strand-specific (ChIP) and two strand-specific (expression) observation tracks.

We did 100 simulation runs. The state numbers were randomly chosen from $\mathcal{U}(5, 10)$ in each single run and sequences with 15000 observations was generated. Models parameters were initialized as

follows

$$
\begin{aligned}
a_{ij} &= 1/\mathcal{K} \\
\pi_i &= 1/\mathcal{K} \\
\mu_i &= \mu_{true} + \epsilon_i \quad , \epsilon_i \sim \mathcal{N}(0, 0.01) \\
\Sigma_i &= 0.01 \cdot E
\end{aligned}
$$

where E is the identity matrix. In each simulation run, models were learned on simulated observation sequences of length 1,000 (respectively 10,000). The fitted values $\hat{a}_{ij}$ showed a good agreement with the true parameter values a_{ij}, even when the model was only trained on 1000 observations (Supplementary Fig. 8). The state annotation recovered a median of 97% respectively 99.5% of the true underlying hidden states on sequences not used for training, when the model was trained on an observation sequence of length 1,000 respectively 10,000 (Supplementary Fig. 8).

Clustering of state sequences

A set of valid coding genes was selected from initally 6,603 ORFs from SGD. 5,088 of them had an annotation of transcript boundaries provided by [17] (overall 5498). Next, we selected transcripts where the TSS was located upstream and the pA site downstream of the coding region, yielding 4,687 genes. Then, state paths were extracted from the bdHMM annotation with ± 250 bp flanking region. We filtered out transcripts where more than 80% of positions were annotated to the proper strand. This resulted in 4,263 genes, which were rescaled to a common length. Pairwise Hamming distances were computed and the sequences were hierarchically clustered. The dendrogram was cut off to yield 55 clusters. Gene-set enrichment analysis was carried out using mgsa [31,32]. A GO group was considered active if the posterior probability was > 0.5.

Targeted identification of genomic features

We defined regular expressions $((S_+|T2) + |(S_-|T2)+)$ and $(S_-) + (P/T1|P2|P1) + (S_+)+$ to search for transcripts and bidirectional promoters throughout the yeast genome, where $S_+ = \{PE1_+, PE2_+, eE1_+, eE2_+, eE3_+, mE1_+, mE2_+, mE3_+, mE4_+, mE5_+, lE1_+, lE2_+lE3_+, T1_+\}$ defines a set containing all forward states, excluding state $P/T2_+$. S_- is defined likewise. Transcripts were constrained to have a minimal length of 80bp, which yielded . We uniquely assigned the 6,068 predictions, using the best reciprocal hit with respect to transcript boundary distance. This yielded 4,186 uniquely assigned transcript predictions. Estimated cumulative distribution functions were computed to assess the accuracy of the predictions. The predictions of bidirectional promoters were not subsequently filtered. The newly identified transcription units were assigned a class (coding, SUT or CUT) using the SGD ORF annotation and expression data from [17].

De novo motif discovery

DNA sequences were extracted for each genomic state. To increase sensitivity of the motif search we excluded very long and very short sequences (min. length: 150bp, max length: 90% quantile of sequence length for current state). Motif search was carried out using XXmotif [33], which uses a negative sequence set to calculate p-values for motif enrichment. The choice of this negative set can be crucial, since it corrects for general sequence features. We chose as negative sets, upstream sequences starting at -50bp relative to the current genomic state. A sequence motif was considered to be enriched if it had an e-value $< 10^{-6}$ and occured in at least 5% of all sequences. The TOMTOM software [34] was used to search databases for similar known motifs.

Fitting a standard HMM and a bdHMM to human chromatin modifications

We fitted a bdHMM to binary chromatin modification data from Ernst and Kellis[11] which previously

had been analysed by the chromHMM algorithm. The Bernoulli emission probabilities learned by chromHMM were fixed and only transitions were updated during the learning of the bdHMM. This was done to ensure that the improvements over chromHMM are only due to the altered modeling of the transitions. First, an HMM transition matrix was fitted using chromHMM transitions (51 states) as initialization, whereby 10^{-3} was added to each transition probability. The bdHMM transition matrix was generated by inflating the transition matrix learned by the standard HMM to a 102×102 matrix. Thus our model initially did not contain any undirected states. A flag sequence was generated from annotated GENCODE[35] transcribed units (version 3c) to set directional constraints at actively transcribed regions. The 39,447 GENCODE annotations were filtered for non-overlapping transcripts with a minimal length of 1000 bp and minimal distance of 5,000 bp to neighbouring transcripts on both strands (6,385). This set was filtered for expressed transcripts showing median a Pol II signal greater than the 25% quantile. This yielded 1,637 actively transcribed regions, which were used to generate a flag sequence, covering approximately 6% of genomic positions. After EM-learning of the bdHMM transitions, the most likely state path was calculated using Viterbi decoding. Running time for bdHMM learning was 22h using the multiprocessing version of STAN with 30 cores.

Comparison of bdHMM and chromHMM

The bdHMM annotation (i.e., the Viterbi path) was compared to the chromHMM annotation. The comparison was carried out by identifying bdHMM states with their chromHMM counterpart having identical emission distribution. This means that conjugate forward and reverse bdHMM states are mapped to the same chromHMM state. 88% of state annotations matched between bdHMM and chromHMM at non-repressive genomic regions. When all states were considered, the annotation agreed only on 43% of the genomic positions. This is mainly caused by the different occurrence of the 'unbound' chromHMM state (state 40). While the bdHMM annotates 50% of the genomic positions with this state, chromHMM anno-

tates only 13% of all positions as 'unbound'. We find the latter fact surprising since 75% of all positions in the data of Ernst and Kellis[11] are zero for all considered chromatin marks, which is best modeled by the 'unbound' state. We thus re-fitted the transitions of a standard HMM initialized with the parameters reported by Ernst and Kellis[11], keeping the emission distributions fixed. In theory, this should not increase the likelihood substantially. However, the likelihood increased by a factor of 3, affecting mostly the transitions into the 'unbound' state, indicating that the lack of unbound states in the chromHMM annotation is due to a fitting artifact. Agreement between the bdHMM and re-fitted HMM annotation was 97%, showing that bdHMMs essentially add directionality to chromatin states.

References

1. Mayer, A. *et al.* Uniform transitions of the general RNA polymerase II transcription complex. *Nat. Struct. Mol. Biol.* **17**, 1272–1278 (2010).

2. Venters, B. J. & Pugh, B. F. A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. *Genome Res.* **19**, 360–371 (2009).

3. Bataille, A. R. *et al.* A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. *Mol. Cell* **45**, 158–170 (2012).

4. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).

5. Day, N., Hemmaplardh, A., Thurman, R. E., Stamatoyannopoulos, J. A. & Noble, W. S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424–1426 (2007).

6. Thurman, R. E., Day, N., Noble, W. S. & Stamatoyannopoulos, J. A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17**, 917–927 (2007).

7. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

8. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

9. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).

10. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).

11. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

12. modENCODE Consortium, T. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787–1797 (2010).

13. Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143**, 212–224 (2010).

14. Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **39**, 1235–1244 (2007).

15. Mayer, A. *et al.* CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* **336**, 1723–1725 (2012).

16. Lidschreiber, M., Leike, K. & Cramer, P. Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. *Mol. Cell. Biol.* **33**, 3805–3816 (2013).

17. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).

18. Schulz, D. *et al.* Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. *Cell* (2013).

19. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).

20. Fan, X. *et al.* Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17945–17950 (2010).

21. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comp. Graph. Stat.* **5**, 299–314 (1996).

22. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).

23. Zacher, B., Kuan, P. F. & Tresch, A. Starr: Simple Tiling ARRay analysis of Affymetrix ChIP-chip data. *BMC Bioinformatics* **11**, 194 (2010).

24. Zacher, B., Torkler, P. & Tresch, A. Analysis of Affymetrix ChIP-chip data using starr and R/Bioconductor. *Cold Spring Harb Protoc* **2011**, pdb.top110 (2011).

25. Huber, W., Toedling, J. & Steinmetz, L. M. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963–1970 (2006).

26. Baum, L. E., Petrie, T., Soules, G. & Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* **41**, 164–171 (1970).

27. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* **39**, 1–38 (1977).

28. Wächter, A. & Biegler, L. T. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming* **106**, 25–57 (2006).

29. Ghalanos, A. & Theussl, S. Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method., howpublished = R package version 1.14, year = 2012.

30. Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl. 1**, S96–S104 (2002).

31. Bauer, S., Gagneur, J. & Robinson, P. N. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.* **38**, 3523–3532 (2010).

32. Bauer, S., Robinson, P. N. & Gagneur, J. Model-based gene set analysis for Bioconductor. *Bioinformatics* **27**, 1882–1883 (2011).

33. Hartmann, H., Guthohrlein, E. W., Siebert, M., Luehr, S. & Soding, J. P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.* **23**, 181–194 (2013).

34. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

35. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

Figure 1 Principle of bidirectional HMM (bdHMM). (a) Simulated occupancy signal (1st track from the top) for a putative factor with a low level (centered at 0) in untranscribed regions (state U), an intermediate level in 5' part of genes (state E), and a high level in 3' part of genes (state L). Arrows (2nd track) depict boundaries and orientation of transcription. Unlike standard HMMs (3rd track) bdHMM (4th track) infer strands (+ or -) to expressed states (E, L). (b) HMM transition graph. Because orientation of transcription is not modeled by standard HMMs, the spurious reverse transitions ($E \Rightarrow U$, $L \Rightarrow E$, and $U \Rightarrow L$) are as likely as the correctly oriented transitions ($U \Rightarrow E$, $E \Rightarrow L$, and $L \Rightarrow U$). (c) bdHMM transition graph. In contrast to HMMs, bdHMMs, which have explicit strand-specific expressed states (E^+/E^- and L^+/L^-), allow inferring only the correctly oriented transitions.

Figure 2 De novo annotation of directed genomic states from genome-wide transcription data in yeast using bdHMM. Input for the bdHMM are, from top to bottom: strand-specific wild-type RNA levels (salmon for the + strand and orange for the -strand), occupancy maps of nucleosomes (ocher), of 3 termination factors, 6 elongation factors, 3 capping factors, 2 initiation factors, 4 CTD modifcations, and 1 core PolII member (Rpb3). Inferred directed genomic states are shown as colored boxes in the lowest track (see color legend beneath) where expressed states on the + (respectively -) strand are positioned above (respectively under) the axis, and not expressed states are centered on the axis. Previous transcriptome annotation is shown in the 2nd track from the bottom.

Figure 3 Roles of directed genomic states in the transcription cycle. (a) Mean ChIP enrichment of factors (horizontal axis) indicates the composition of the transcription machinery in each state (vertical axis). Factors were ordered by hierarchical clustering and states were ordered

by position of their most frequent occurrence along the average gene. (b) Each state was assigned to a phase in the transcription cycle by investigating the frequency (y-axis) of each state at an average transcript. This spatial state distribution was calculated from the genomic sate sequence (viterbi path) of 4,362 genes. (c) The flux diagram shows probabilities of state transitions calculated from the viterbi paths. Branches mark alternative successions of states at individual genes and thus reveal extensive variation in the transcription cycle as it is modeled by the genomic states. Each node (state) is positioned according to the most frequent position on a metagene. The diagram contains at least one incoming and one outgoing transition for each state as well as transitions observed with a frequency > 0.01 on the metagene.

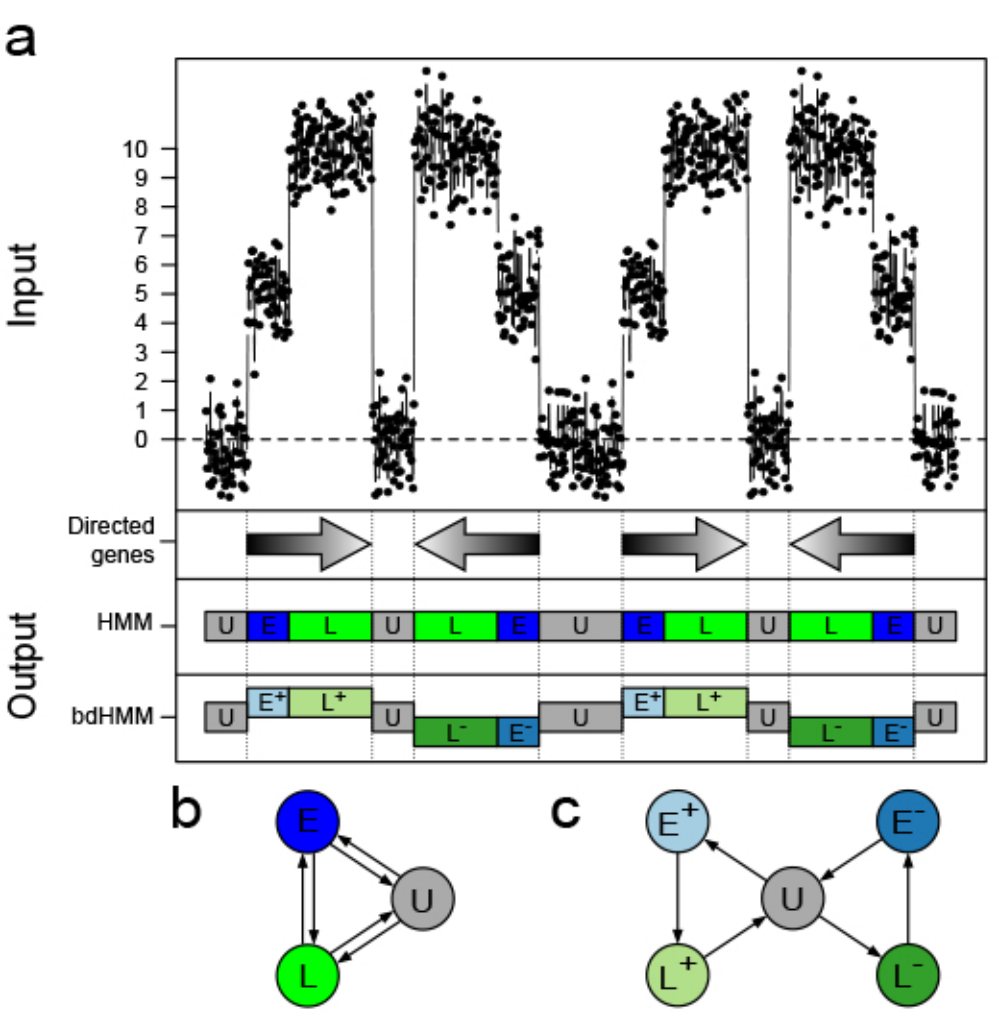
Figure 4 Clustering of state paths reveals gene-specific variations in the transcription cycle. (a) Genomic state sequences of 4632 genes were clustered into 55 groups (left, only clusters containing at least 20 genes are labeled). Each line corresponds to the state sequence of a single gene. States are colored as shown in the legend. (b) Clusters exhibit distinct state frequency distributions and transition patterns (shown as schematic flux diagrams on top of panels). Cluster 14 shows a transcription cycle closest to the canonical one proposed by [22]. Genomic state sequences of cluster 32 and 38 differ from the canonical one, indicating variations in the transcription cycle. (c) Cluster 14 and 32 exhibit distinct recruitment of factors to genes. PolII subunit Rpb3, Nrd1, Spt5 and Spt16 binding is very similar in the beginning of genes, but decreases much stronger in cluster 32 throughout the transcripts. Ctk1 and Paf1 are depleted at cluster 32, but not at cluster 14 genes. (d) Cluster 14 shows the canonical PolII (Rpb3) peak in the 5' region of genes, but PolII reaches a stable, high level downstream of the TSS. This may suggest a lack of the mechanism for Pol II peaking observed in cluster 14. The

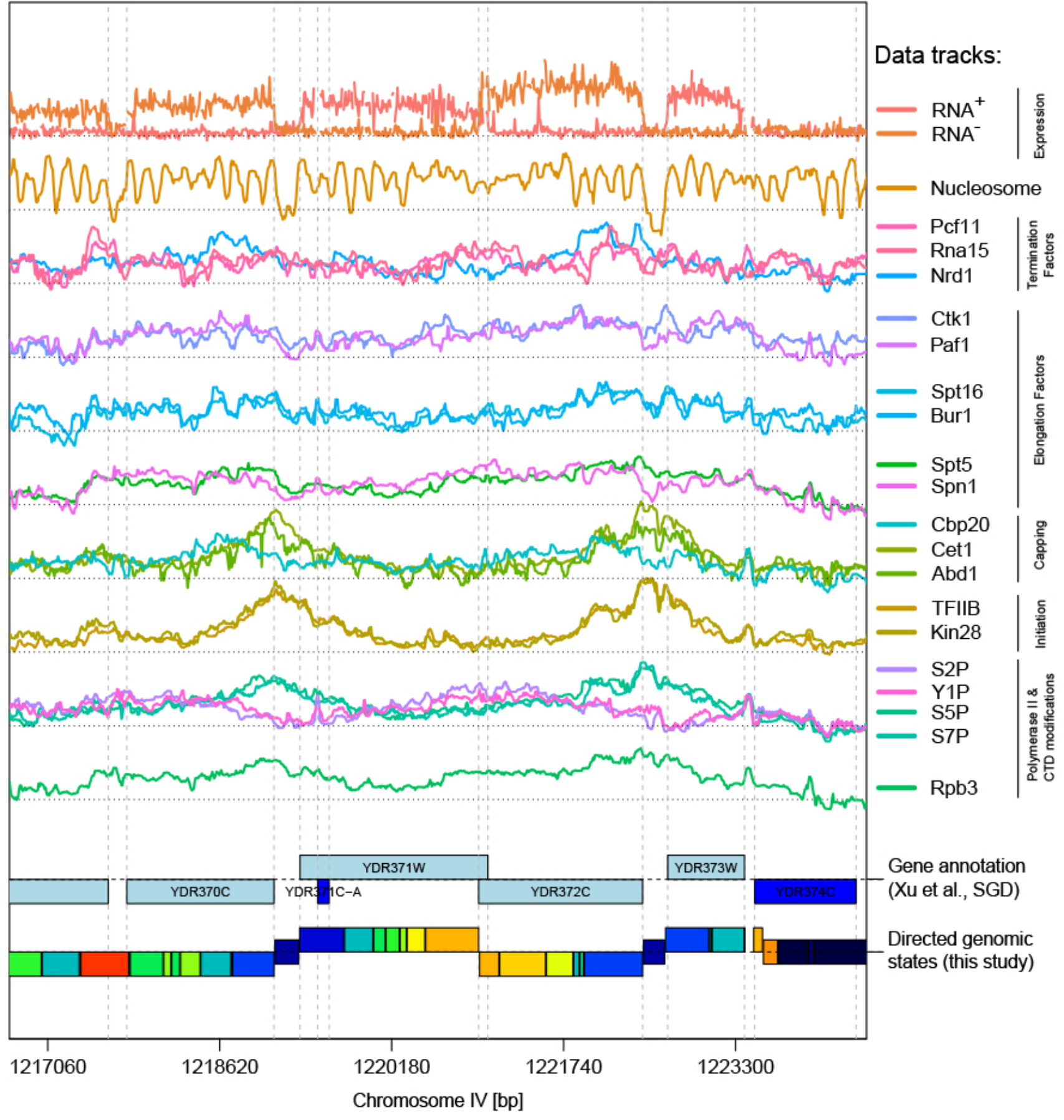
steep increase of Serine 2 phophorylation in cluster 38 might indicate that productive elongation is reached earlier at those genes.

Figure 5 Genomic state annotation redicts bidirectional promoters and novel transcripts. (a) The denomic state annotation (viterbi path) was searched with Regular Expressions (RegEx) defining bidirectional promoters (right) and transcripts (bottom). See Online Methods for a detailed description. (b) Nucleosome binding patterns centered at 1,076 identifed bidirectional promoters found with the RegEx. Each line in the heatmap corresponds to one pair of transcripts. Binding signal is color-coded (right). (c) The grey area highlights a region predicted to be expressed on the + strand by the bdHMM yet not captured by former annotations based on the wild-type RNA levels alone. Data obtained from a mutant of an RNA degradation pathway ($\Delta RRP6$, black area for the + strand, grey areas for the - strand) shows presence of a CUT (cryptic untranslated transcript, a transcript rapidly degraded in an RRP6-dependent fashion) on the + strand, i.e. validates that transcription occurs at this location. (d) A novel SUT (Stable unannotated transcript, a stable non-coding RNA, grey area) is identified on the -strand by the bdHMM. The locus is expressed and shows binding of PolII and transcription factors.

Figure 6 Application of bdHMM to chromatin modifications in human T-cells identifies directionality of chromatin states. (a) Example of chromatin state annotation of chromHMM and bdHMM (bottom tracks) with RefSeq gene annotation and input signal. State directionality matches gene orientation of annotated convergent genes and divergent genes. The log-transformed signal[11] of all 41 data tracks is shown in black on top. Binarized input signal is shown for 18 acetylation marks in blue, 20 methylation marks in red and CTCF/PolII/H2A.Z in brown. (b) Transitions between state groups (detailed graph in Supplementary Fig. 7) of

directed chromatin states (bdHMM) are mostly unidirectional. An edge is drawn between two groups when at least one transition between the groups is > 0.04 in any direction. The graph reveals characteristic directed state sequences. Promoter states either transition to transcribed states (purple colorscale), active intergenic (yellow colorscale), repressive (grey colorscale) or the unbound state (grey colorscale). Transitions between transcribed chromatin states follow the phases of the transcription cycle with proper directionality (from 5' proximal to 5' distal and end of transcription (txn)). (c) State frequencies of promoter and 5' proximal transcription states at RefSeq TSSs. States are shown in sense (match of state and transript directionality) direction of the respective TSS. Sense state frequencies are annotated downstream of TSSs, indicating proper state directionality.


a
Input
10
9
8
7
6
5
4
3
2
1
0
Directed genes
Output
HMM
U
E
L
U
L
E
U
E
L
U
L
E
U
bdHMM
U
E+
L+
U
L-
E-
U
E+
L+
U
L-
E-
U
b
E
U
L
c
E+
E-
U
L+
L-



Directed genomic states:

Untranscribed (U) & Promoter (P)
- U
- P1
- P2

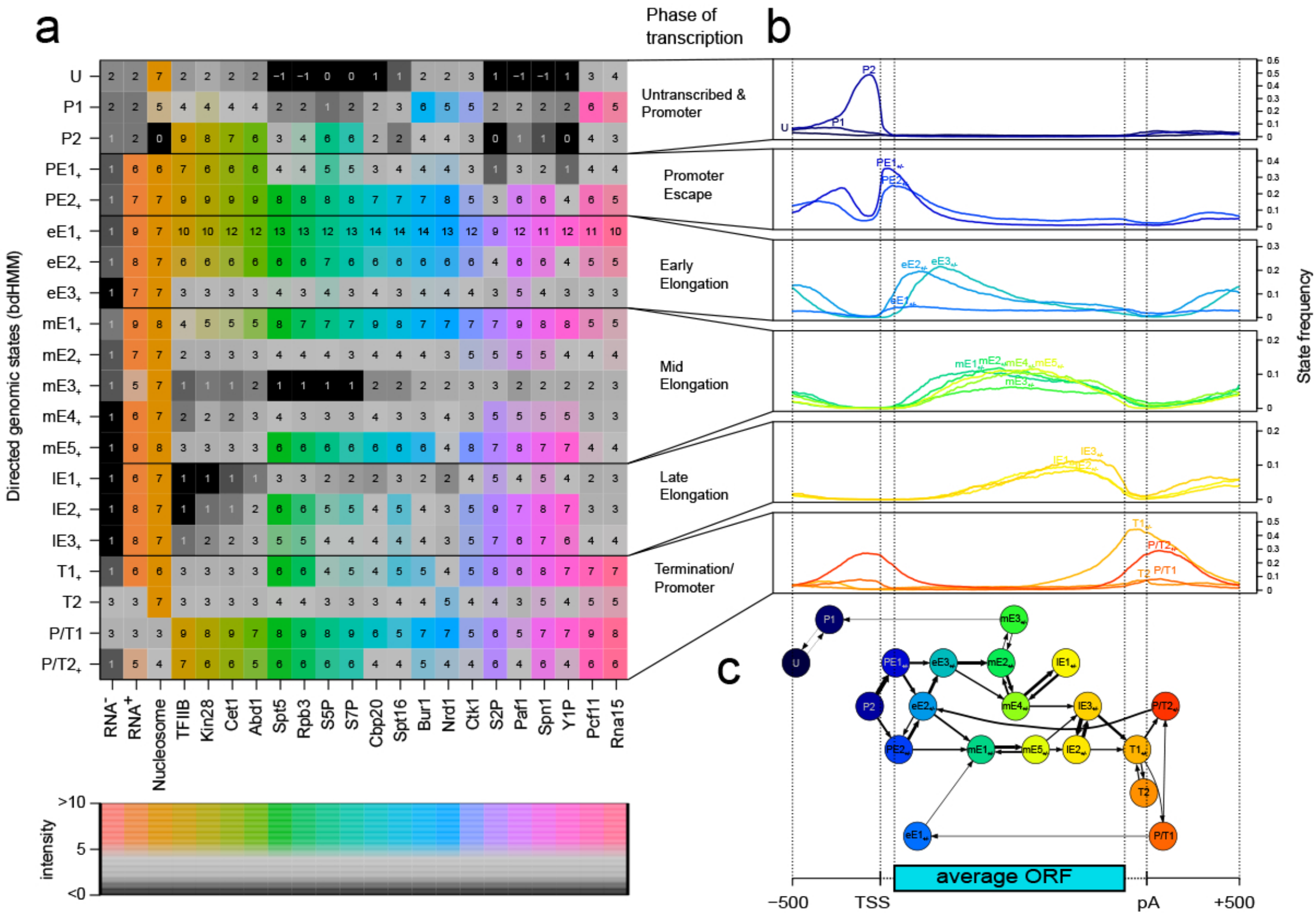
Promoter escape (PE)
- $PE1_{+/-}$
- $PE2_{+/-}$

early Elongation (eE)
- $eE1_{+/-}$
- $eE2_{+/-}$
- $eE3_{+/-}$

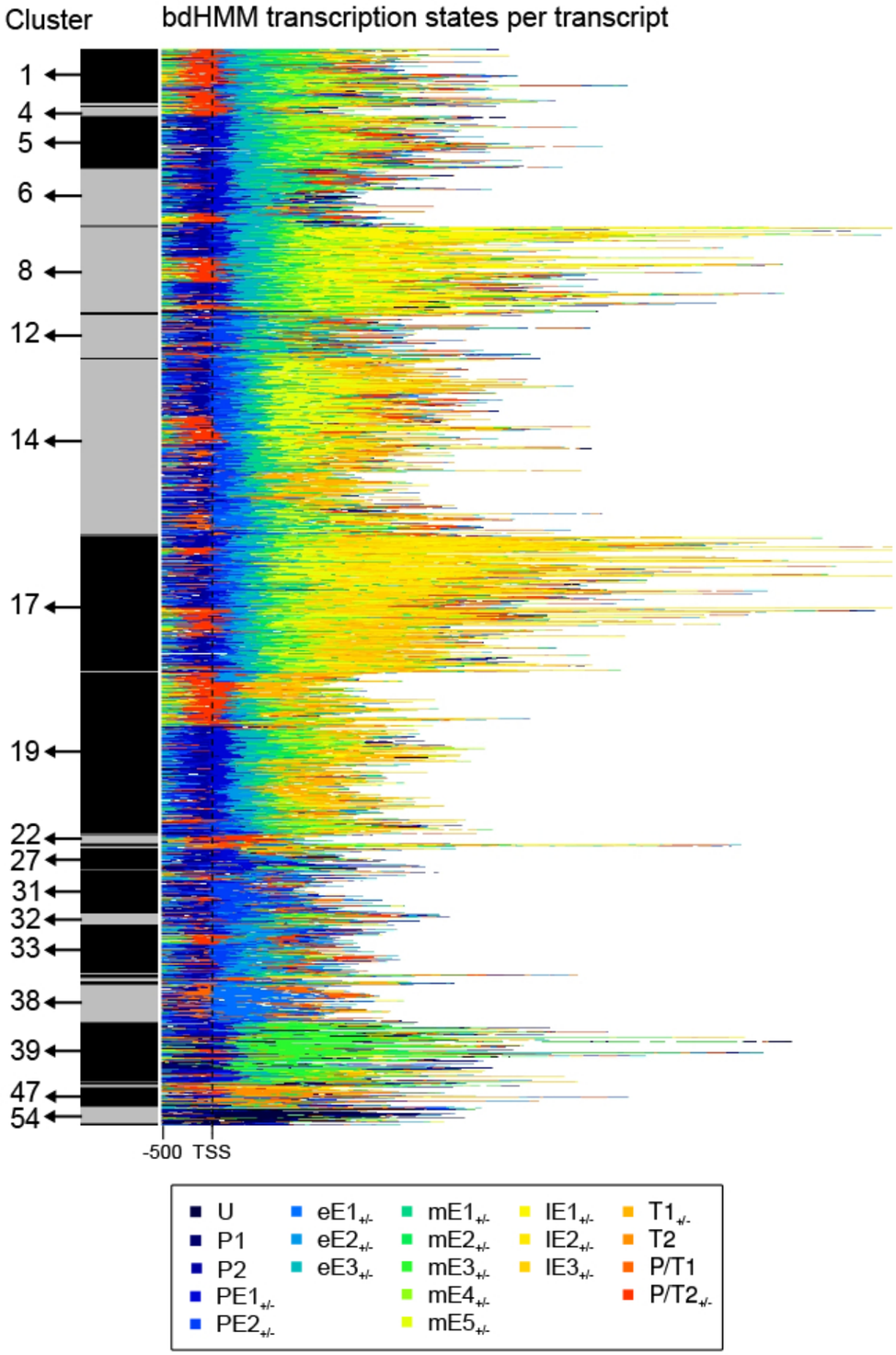
midElongation (mE)
- $mE1_{+/-}$
- $mE2_{+/-}$
- $mE3_{+/-}$
- $mE4_{+/-}$
- $mE5_{+/-}$

late Elongation (lE)
- $lE1_{+/-}$
- $lE2_{+/-}$
- $lE3_{+/-}$

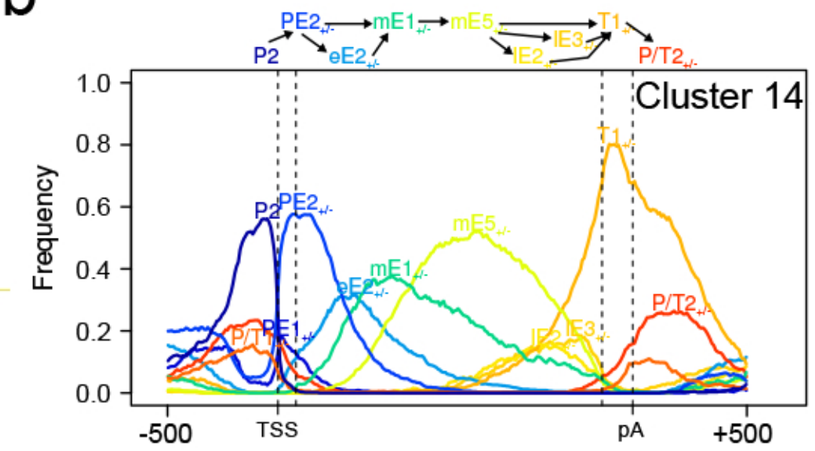
Promoter/ Termination (P/T)
- $T1_{+/-}$
- T2
- P/T1
- $P/T2_{+/-}$

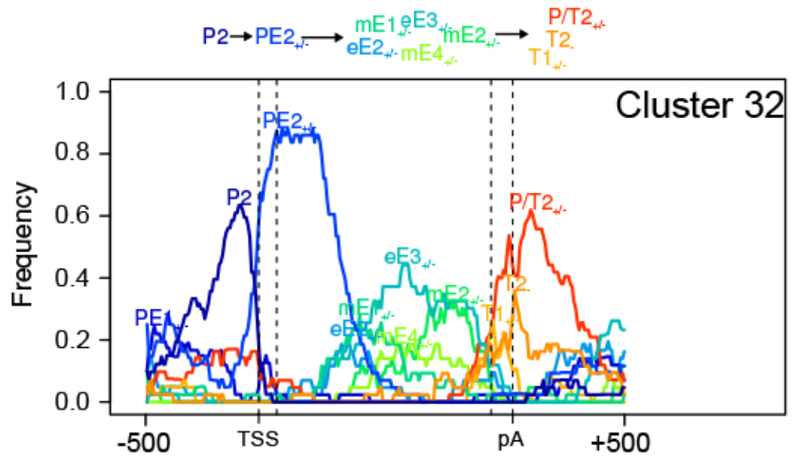

a
b
c
Phase of transcription
Directed genomic states (bdHMM)
State frequency
Untranscribed & Promoter
Promoter Escape
Early Elongation
Mid Elongation
Late Elongation
Termination/ Promoter
U
P1
P2
PE1+
PE2+
eE1+
eE2+
eE3+
mE1+
mE2+
mE3+
mE4+
mE5+
IE1+
IE2+
IE3+
T1+
T2
P/T1
P/T2+
RNA-
RNA+
Nucleosome
TFIIB
Kin28
Cet1
Abd1
Spt5
Rpb3
S5P
S7P
Cbp20
Spt16
Bur1
Nrd1
Ctk1
S2P
Paf1
Spn1
Y1P
Pcf11
Rna15
intensity
>10
5
<0
average ORF
−500
TSS
pA
+500

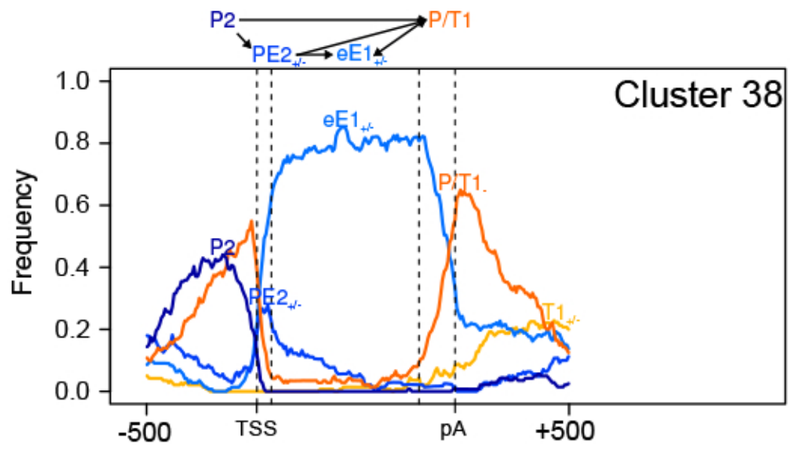
a


Cluster
bdHMM transcription states per transcript
1
4
5
6
8
12
14
17
19
22
27
31
32
33
38
39
47
54
-500
TSS
U
P1
P2
PE1+/-
PE2+/-
eE1+/-
eE2+/-
eE3+/-
mE1+/-
mE2+/-
mE3+/-
mE4+/-
mE5+/-
IE1+/-
IE2+/-
IE3+/-
T1+/-
T2
P/T1
P/T2+/-

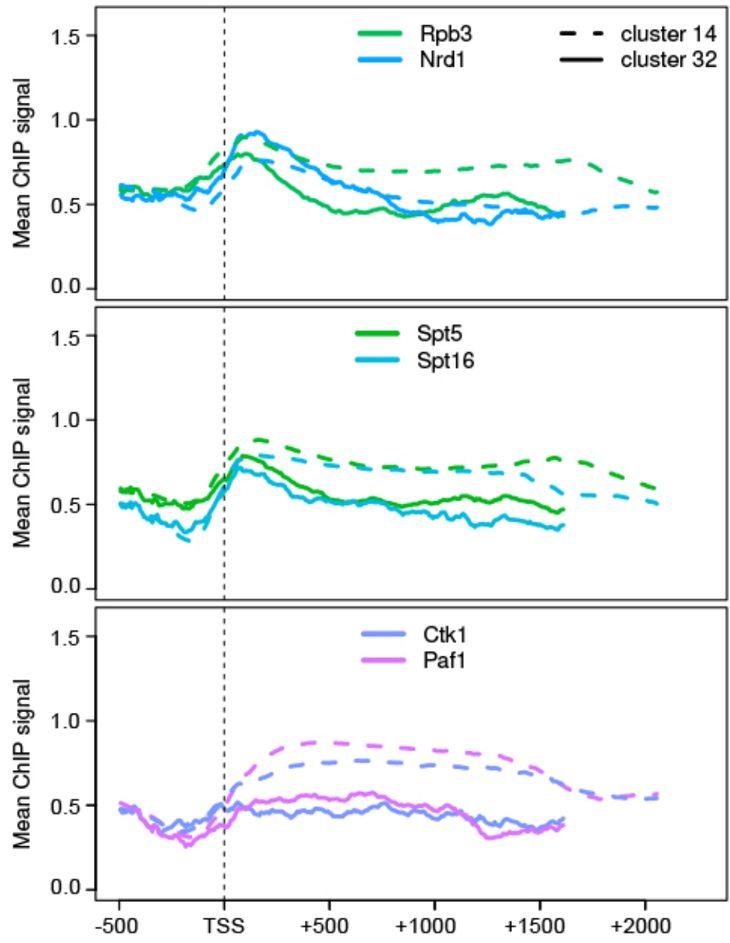
b


Cluster 14
Frequency
-500
TSS
pA
+500

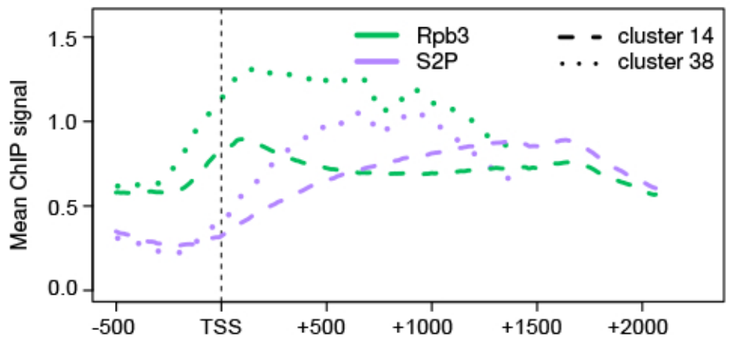

Cluster 32
Frequency
-500
TSS
pA
+500

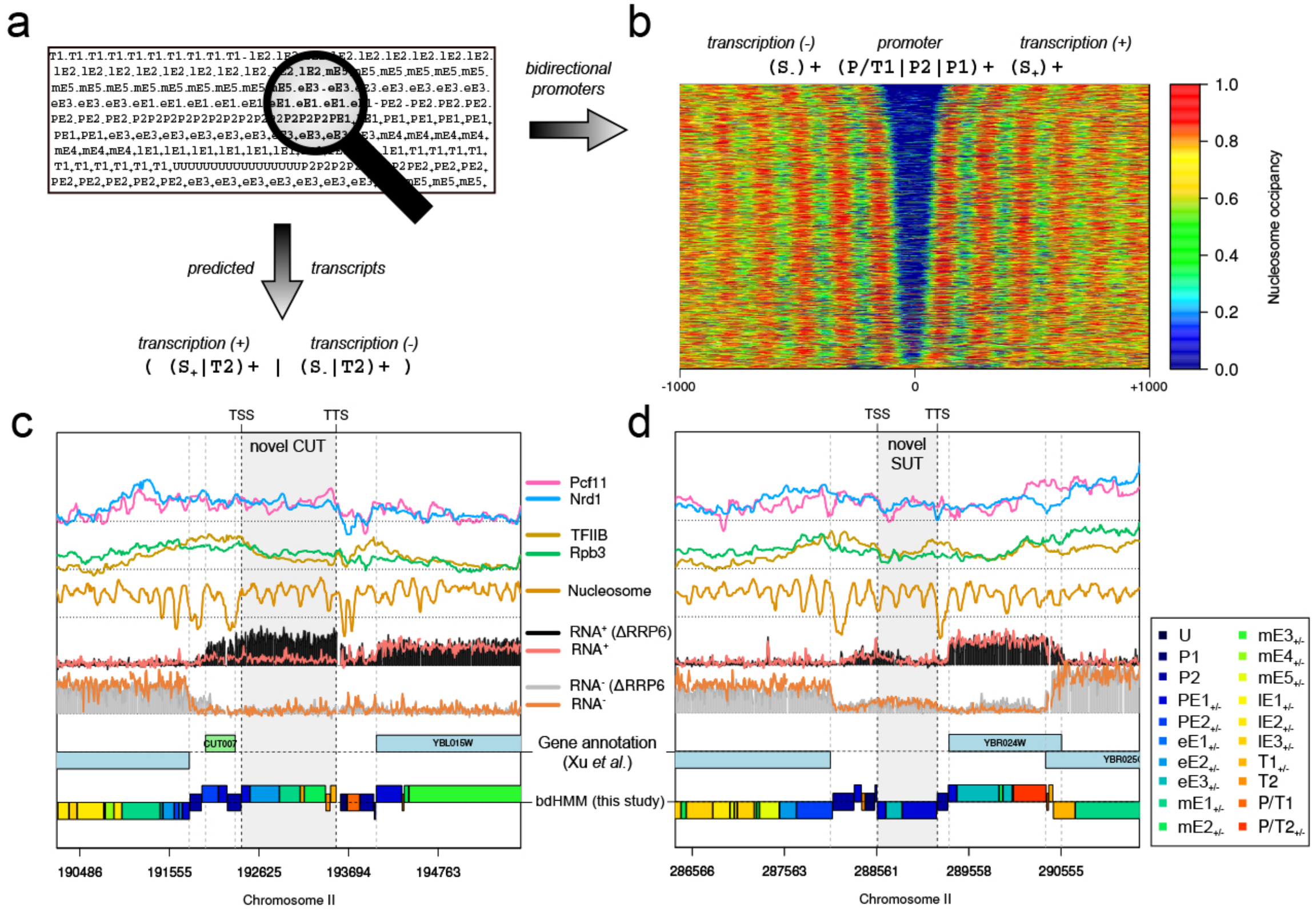

Cluster 38
Frequency
-500
TSS
pA
+500

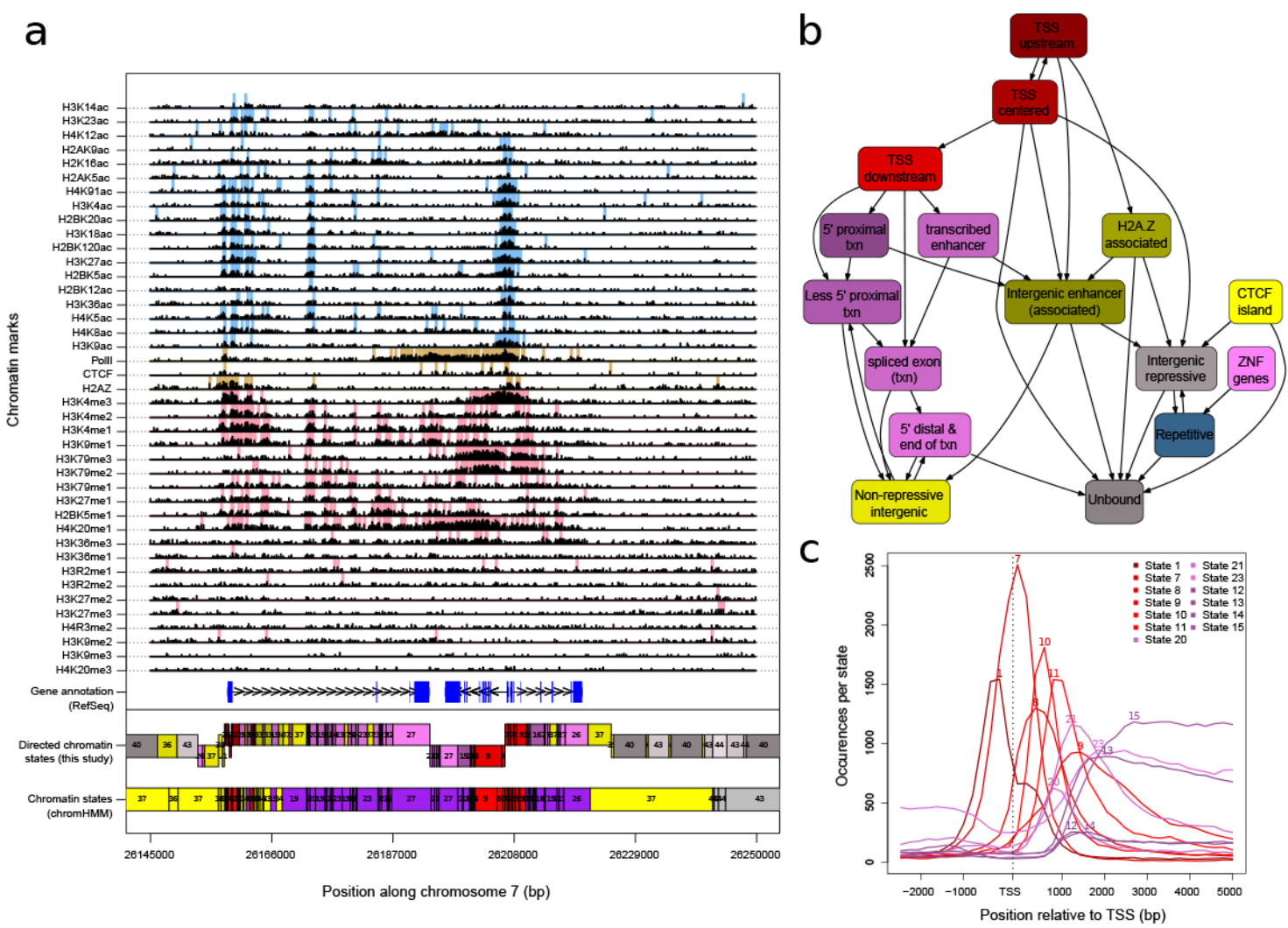
c


Rpb3
Nrd1
cluster 14
cluster 32
Spt5
Spt16
Ctk1
Paf1
Mean ChIP signal
-500
TSS
+500
+1000
+1500
+2000

d


Rpb3
S2P
cluster 14
cluster 38
Mean ChIP signal
-500
TSS
+500
+1000
+1500
+2000


a
bidirectional promoters
predicted transcripts
transcription (+)
transcription (-)
((S+|T2)+ | (S-|T2)+)
b
transcription (-)
promoter
transcription (+)
(S-)+ (P/T1|P2|P1)+ (S+)+
-1000
0
+1000
Nucleosome occupancy
1.0
0.8
0.6
0.4
0.2
0.0
c
TSS
TTS
novel CUT
Pcf11
Nrd1
TFIIB
Rpb3
Nucleosome
RNA+ (ΔRRP6)
RNA+
RNA- (ΔRRP6
RNA-
CUT007
YBL015W
Gene annotation (Xu et al.)
bdHMM (this study)
190486
191555
192625
193694
194763
Chromosome II
d
TSS
TTS
novel SUT
YBR024W
286566
287563
288561
289558
290555
Chromosome II
U
P1
P2
PE1+/-
PE2+/-
eE1+/-
eE2+/-
eE3+/-
mE1+/-
mE2+/-
mE3+/-
mE4+/-
mE5+/-
IE1+/-
IE2+/-
IE3+/-
T1+/-
T2
P/T1
P/T2+/-


a
Chromatin marks
H3K14ac
H3K23ac
H4K12ac
H2AK9ac
H2K16ac
H2AK5ac
H4K91ac
H3K4ac
H2BK20ac
H3K18ac
H2BK120ac
H3K27ac
H2BK5ac
H2BK12ac
H3K36ac
H4K5ac
H4K8ac
H3K9ac
PolII
CTCF
H2AZ
H3K4me3
H3K4me2
H3K4me1
H3K9me1
H3K79me3
H3K79me2
H3K79me1
H3K27me1
H2BK5me1
H4K20me1
H3K36me3
H3K36me1
H3R2me1
H3R2me2
H3K27me2
H3K27me3
H4R3me2
H3K9me2
H3K9me3
H4K20me3
Gene annotation (RefSeq)
Directed chromatin states (this study)
Chromatin states (chromHMM)
26145000
26166000
26187000
26208000
26229000
26250000
Position along chromosome 7 (bp)
b
TSS upstream
TSS centered
TSS downstream
5' proximal txn
transcribed enhancer
H2A.Z associated
Less 5' proximal txn
Intergenic enhancer (associated)
CTCF island
spliced exon (txn)
Intergenic repressive
ZNF genes
5' distal & end of txn
Repetitive
Non-repressive intergenic
Unbound
c
Occurences per state
0
500
1000
1500
2000
2500
State 1
State 7
State 8
State 9
State 10
State 11
State 20
State 21
State 23
State 12
State 13
State 14
State 15
−2000
−1000
TSS
1000
2000
3000
4000
5000
Position relative to TSS (bp)