Dissertation zur Erlangung des Doktorgrades

der Fakultät für Chemie und Pharmazie

der Ludwig-Maximilians-Universität München

Genomic data integration with hidden Markov models to understand transcription regulation

Benedikt Zacher

aus

München, Deutschland

2016

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Patrick Cramer betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 11.03.2016

Benedikt Zacher

Dissertation eingereicht am 11.03.2016

1. Gutachter: Prof. Dr. Patrick Cramer
2. Gutachter: Prof. Dr. Julien Gagneur

Mündliche Prüfung am 02.05.2016

Acknowledgments

I am very grateful to Prof. Dr. Julien Gagneur for his supervision and guidance during this project. He created a working atmosphere in his group where I could thrive. He gave suggestions and help whenever needed and enough freedom to explore and follow my own ideas. He trusted me with planning own projects which gave me the unique opportunity to visit Stanford University during my work on this thesis.

Prof. Dr. Achim Tresch has been an important person already during my very early career as a researcher and since then I have been enjoying working with him. I am thankful for his support and encouragment not only during this thesis. It is also thanks to him that I chose to follow a scientific career. His casual manner made it very easy and fun to work with him and taking things not too serious once in a while.

I also want to thank Prof. Dr. Patrick Cramer for supervising this thesis and his support during collaborations ever since I started working at the Gene Center. It is admirable to see him manage his full schedule while still keeping his interest and positive attitude, which you can feel in his whole group.

I would like to thank the rest of my thesis committee: Dr. Dietmar Martin, Prof. Dr. Klaus Förstemann and Prof. Dr. Karl-Peter Hopfner.

I owe my gratitude to all present and former members of the Gagneur, Tresch, Cramer and Söding group for an extraordinary working atmosphere. Special thanks to Dr. Björn 'Poony' Schwalb, Daniel Bader and Chris Mertes for withstanding my stupid jokes in the office (and all the rest who were not sitting as close but had to deal with it once in a while). Many thanks to Michael Lidschreiber for sharing data way before publication, for introducing me to 'Extrem Bosseln' and just having a nice time. Many many thanks again to Björn Schwalb and Margaux Michel for working together and much beyond that! Many thanks to Carina Demel, Daniel Schulz, Matthias Siebert, Phillipp Torkler, Saskia Gressel, Jürgen Niesser, Katja Frühauf and Wolfgang Mühlbacher for fun times within and beyond science.

I would be nothing without my friends. Thank you for being in my life: Jannis, Zissy, Magda, Doris, Joni, Jan, Biddy, Paul, Sara, Veri, Micha, Jimmy, Domi, Flo, Clemens, Anna, Simon, Tisy, X, Laia, Julia N., Martin, Jesse, Joubi, Simone, Amrei, Peti, Julia I., Alex and Miri. And everybody I forgot, you know who you are.

Further I would like to thank Siegfried, Christine and Marvin Maschke. I am very lucky to have you in my family!

I owe my deepest gratitude to my parents, Silvia and Werner, who made this (and me) possible with their support, love and trust. I also want to thank my sister Vroni for being my sister.

Above all and everything I want to thank my wife Jasmin. Writing down everything that I owe you would take another 3,5 years, so I'll make it short: You are the love of my life!

Summary

Transcription is a tightly controlled process that involves the recruitment and prost-translational modification of DNA-associated protein complexes, which can be mapped to the genome using high-throughput experimental assays. An accurate annotation of genomic elements such as transcription units or cis-regulatory elements such as promoters or enhancers is crucial for the use and interpretation of data generated by these assays. Thus, integrative genomic data analysis of high-throughput assays with hidden Markov models (HMMs) has become a popular tool for genome annotation. However, current algorithms are limited by unrealistic data distribution assumptions and variance models. Moreover, they are not able to assign forward or reverse direction to states or properly integrate strand-specific (e.g., RNA expression) with non-strand-specific (e.g., ChIP) data, which is essential to characterize directed processes such as transcription.

In this thesis new HMM-based methods are proposed to overcome these limitations. These include (i) bidirectional HMMs (bdHMMs) which integrate strand-specific with non-strand-specific data to infer directed genomic states *de novo* and (ii) GenoSTAN (***Geno***mic ***ST***ate ***AN***notation), a HMM using discrete probability distributions to model count data, for genome annotation from Next-Generation-Sequencing data. Both approaches are made available in the R/Bioconductor package STAN (***ST***ate ***AN***notation) which provides an efficient implementation that can be run on large genomes such as human.

STAN is used to derive new and improved annotations of transcription in yeast and human and to generate a map of promoters and enhancers in 127 human cell types and tissues.

Integration of transcription factor binding and RNA expression data in yeast recovers the majority of transcribed loci, reveals gene-specific variations in the yeast transcription cycle, identifies 32 new transcribed loci, 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.

Moreover, promoters and enhancers are predicted in 127 human cell types and tissues are mapped by integrating sequencing data from the ENCODE and Roadmap Epigenomics projects, today's largest compendium of chromatin assays. Promoters and enhancers are identified with consistently higher accuracy and show significantly higher enrichment of complex trait-associated genetic variants than current annotations. Investigation of binding of 101 transcription factors in human K562 cells reveals common and distinctive TF binding properties of enhancers and promoters.

Application of STAN to transient transcriptome sequencing (TT-Seq) data in human K562 cells recovers stable mRNAs, long intergenic non-coding RNAs, and additionally maps over 10,000 transient RNAs, including enhancer RNAs, antisense RNAs, and promoter-associated RNAs. Further analyses reveal that transient RNAs such as enhancer RNAs are short and lack U1 motifs and secondary structure.

Taken together, the annotations inferred in this thesis gave new insights into transcription and its regulation and will be an important resource for future research in genomics. STAN is a valuable tool to create such annotations also in other organisms and as more data becomes available improve the existing ones.

Publications

Parts of this work have been published or are in the process of publication:

2016 **Accurate promoter and enhancer identification in 127 ENCODE and Roadmap Epigenomics cell types and tissues by GenoSTAN**
B. Zacher*, M. Michel, B. Schwalb, P. Cramer, A. Tresch*, J. Gagneur*
(* corresponding author)
submitted, preprint available on biorxiv (doi: http://dx.doi.org/10.1101/041020)

Author contributions: BZ, JG and AT developed the statistical methods and computational workflow of the study. BZ developed and implemented all software and scripts and carried out all computational analyses. BS helped with preprocessing of the K562 data set. MM and PC helped with interpretation of the biological results. BZ, JG and AT wrote the manuscript with input from all authors. All authors read and approved the final version of the manuscript.

2016 **TT-Seq captures the human transient transcriptome**
B. Schwalb*, M. Michel*, **B. Zacher***, K.Frühauf, C. Demel, A. Tresch, J. Gagneur, P. Cramer (* joint first authorship)
accepted for publication in Science

Author contributions: MM carried out all experiments. BS carried out all bioinformatics analysis except transcript calling, RNA classification, analysis of U1 sequence motifs, and prediction of RNA secondary structure, which were carried out by BZ. BZ, JG and AT developed the chromatin state annotation. KF designed RNA spike-in probes. CD established the spike-in normalization method. BS, JG and PC designed research. JG and PC supervised research. BS, MM, JG, and PC prepared the manuscript, with input from all authors.

2014 **Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle**
B. Zacher, M. Lidschreiber, P. Cramer, J. Gagneur, A. Tresch
Molecular Systems Biology, 10(12):768–768
http://msb.embopress.org/content/10/12/768.long

Author contributions: BZ, AT and JG developed the statistical methods and computational workflow of the study. BZ developed and implemented all software and scripts and carried out all computational analyses. AT initiated the study. AT and JG supervised research. ML and PC helped with preprocessing of the data and interpretation of the biological results. BZ, AT, JG and PC wrote the manuscript. All authors read and approved the final version of the manuscript.

Contents

Part I
Introduction

Transcription is a fundamental process of life and crucial for the development and function of living organisms. Transcription of a DNA template into messenger RNA (mRNA) is the first step of the central dogma of molecular biology [1]. mRNA synthesis is catalyzed in the cell's nucleus by RNA Polymerase II (Pol II), exported into the cytoplasm and translated into proteins by a multiprtoein complex called the ribosome [2]. Besides mRNA, Pol II produces several other species of non-coding RNAs, including small nuclear RNA (snRNA), small nucleolar RNA (snoRNA) and other stable and instable RNAs such as cryptic unstable transcripts (CUTs), stable unannotated transcripts (SUTs) or enhancer RNAs (eRNA) [3, 4, 5]. There exist two other Polymerases in eurkaryotes, RNA Polymeras I (Pol I) and III (Pol III) [6]. Pol I transcribes the ribosomal RNA (rRNA), which is part of the ribosome and neccessary for translation of mRNA into protein. Pol III transcribed transfer RNA (tRNA), which carry amino acids to the ribosome for protein synthesis. It is of great importance to understand the underlying mechanisms of transcription and its regulation since transcription defects have been shown to be implicated in a variety of diseases [7, 8].

.

1 Transcription regulation by cis-regulatory elements

Whether a gene is transcribed or not is controlled by cis-regulatory elements in the DNA. In yeast, transcription regulation is mostly carried out by two cis-regulatory elements: promoters, upstream activating sequences (UASs) and silencer sequences (Figure 1) [9]. Each of these elements harbours binding sites of transcription factors (TFs) which control gene transcription upon binding. Promoters are the most fundamental cis-regulatory elements required for gene transcription. These elements are regions of DNA that initiate transcription of a particular gene. They are located near the transcription start sites of genes, on the same strand and upstream. The UAS and silencer sequences are usually located within several hundred base pairs from the promoter.
In metazoans, the cis-regulatory architecture is more complex (Figure 1). Additionally to promoters, there are enhancer regions, which were originally defined as DNA elements that can increase expression of a gene over a long distance in an orientaton-independent fashion relative to the gene [10]. In mammals these elements can be located megabases from their target gene [11]. However, experimental results showing that yeast UASs act as enhancers when expressed in human HeLa cells indicate that UASs and enhancers are functioanlly analogue [12]. The recent discovery of pervasive transcription at enhancers [3, 13] challenged the original enhancer definition and there is an ongoing debate about the differences between enhancers and promoters [14]. Metazoan silencers suppress transcription and can be located within kilobases upstream or downstream of their target gene [9].

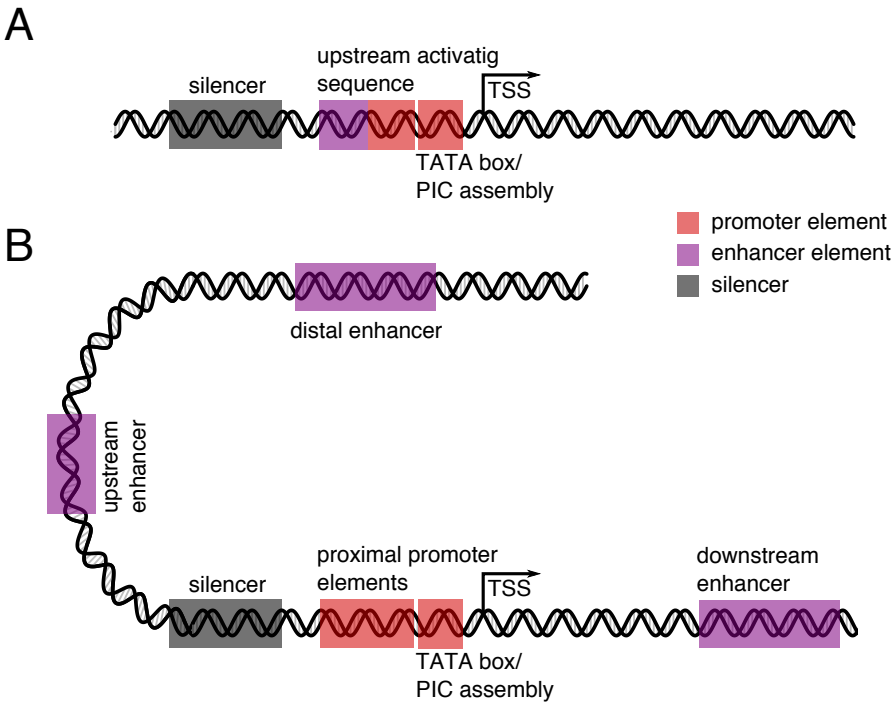


Figure 1: Comparison of a simple eukaryotic promoter and diversified metazoan regulatory modules adopted from [9]. (A) Simple eukaryotic cis-regulatory elements include promoters, upstream activating sequences and silencer elements, which are loctated within several hundred base pairs upstream of the transcription start site (TSS). Cis-regulatory elements harbour binding sites of transcription factors which control gene transcription. For instance, the TATA-binding protein (TBP) binds to the TATA-box in the promoter during formation of the pre-initiation complex (PIC). (B) The regulatory architecture is more complex in metazoans. Enhancer and silencer elements can be located within megabases upstream or downstream of the TSS.

1.1 Transcription regulation by UASs and enhancers

The first step in transcription regulation is the sequence-specific binding of transcription factors to UASs or metazoan enhancers or silencers to control gene activation or repression [9, 15]. In contrast to lower eukaryotes, transcription of metazoan genes is usually controlled by multiple enhancer regions, each of which contributes to the expression profile of the gene [9]. Moreover, inactive enhancers are known to be bound by pioneer factors such as FOXA1 during development to drive cell-type specification and therefore mediate transcription in a tissue- or cell-type-specific manner [16, 17].

1.2 The pre-initiation complex assembles at the promoter

A specific class of TFs - the general transcription factors (GTFs) - assembles at the promoter into the pre-initiation complex (PIC) before transcription is initiated (Figure 2). *In vitro* experiments have shown that six general transcription factors (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) assemble into the PIC [18]. According to this model, TFIID is the first protein to bind the promoter. One of its subunits, the TATA-binding protein (TBP) - recognizes and binds the TATA-box. This is followed by binding of TFIIB, which stabilizes TFIID binding to the promoter and aids in the recruitment of the TFIIF-Pol II complex [18, 19]. PIC assembly is completed by binding of TFIIE and TFIIH, which are required for transcription initiation. However, *in vivo* PIC assembly might be more variable with more transcription factors involved since the heterogeneity of promoters suggests different paths to promoter recognition [19].

1.3 The mediator complex

The mediator complex is a highly conserved protein complex consisting of 21 subunits in yeast and 26 subunits in mammals [20]. It is involved in global transcription control by associating with the sequence-specific TFs bound to the UAS or enhancer, Polymerase and the promoter [21]. To this end, the mediator complex acts as a bridge between enhancer and promoter regions to communicate regulatory signals by protein-protein-interactions, but the exact mechanisms remain to be elucidated.

2 The transcription cycle

Transcription folows a series of steps, the transcription cycle. Each step in the transcription cycle is tightly controlled and requires specific proteins or protein modifications. The following sections focus on studies and transcription factors in *S. cerevisiae*. If not noted otherwise, the mentioned TFs are yeast TFs. A list of all factors analyzed in this work and their role in the transcription cycle is shown in Table 1.

2.1 Initiation and promoter escape

Initiation starts with the formation of the open complex [15], where double-stranded DNA is inserted into the jaw and downstream cleft of Pol II, which is assisted by TFIIB [37]. Stimulated by TFIIE, the DNA-Helicase TFIIH separates both strands, which is followed by the insertion of the single-stranded DNA into the active site of Pol II [37]. Then Pol II scans the downstream nucleotides for a TSS with the help of the TFIIB-reader domain and initiates transcription, leading to the 'initially transcribing complex' [37,38]. This inital phase of transcription is abortive.

factor	function
Nucleosome	Most basic structural unit of chromatin [2]
Pcf11	3´-RNA processing factor interacts with phosphorylated serine 2 at the CTD [22, 23]
Rna15	Component of the cleavage and polyadenylation factor I [22, 23]
Nrd1	Binds phosphorylated serine 5 and a tetramer in the RNA to mediate 3'-end formation of some mRNAs, snRNAs, snoRNAs, and CUTs [24, 25, 26]
Ctk1	Essential protein that phosphorylates serine 2 at the CTD [27]
Paf1	Subunit of the Paf complex which recruits histone modifiers [28]
Spt16	Part of the FACT complex which reorganizes nucleosomes to facilitate access to DNA during elongation [29]
Bur1	Non-essential protein that can phosphorylate serine 2 at the CTD [30]
Spt5	Regulation of transcription elongation and involved in capping by binding guanylyltransferases [31, 32]
Spn1	Involved in transcription elongation. Binding correlates with serine 2 phosphorylation at the CTD [31, 33]
Cbp20	Part of the cap-binding complex that binds co-transcriptionally to the 5' cap of pre-mRNAs [31]
Cet1	RNA 5'-triphosphatase involved in mRNA 5' capping [31]
Abd1	Catalyzes the transfer of a methyl group to the 5' end of capped mRNA [31]
TFIIB	General transcription factor which is part of the pre-initiation complex and required for transcription initiation [18]
Kin28	Subunit of the general transcription factor TFIIH, which phosphorylates serine 5 at the CTD [34]
Rpb3	Subunit of RNA polymerase II [35]
Ser2P	Recruits Spt6 and RNA 3'-processing and termination factors Pcf11 and Rtt103 [34]
Tyr1P	Impairs recruitment of termination factors during elongation [36]
Ser5P	Recruits the mRNA capping enzymes and early termination factor Nrd1 [34]
Ser7P	Co-occurs with Ser5P at the 5' end of genes [33]. The exact function of Ser7P remains to be elucidated [34]

Table 1: Transcription factors and their functions are shown for factors analyzed in this work.

Many rounds of initiation may be needed until the RNA–DNA hybrid reaches a length of 8 nucleotides, which leads to a considerable reduction of abortive initiation [39]. Promoter escape is complete after synthesis of roughly 23 nucleotides downstream of the TSSs [39].

2.2 Elongation

Pol II undergoes a transition from intiation to elongation at about 150 nucleotides downstream of the TSS [33]. During this transition initiation factors are exchanged with elongation factors and Pol II is phoshporylated in Ser5 and Ser7 residues at its C-terminal domain (CTD), which is required for 5' capping of the nascent mRNA [31, 32, 33]. The CTD is a heptapeptide reapeat (YSPTSPS) in the largest subunit of Pol II (Rpb1) which undergoes a specific sequence of post-translational modifications during transcription also known as the CTD cycle [40]. The serine 5 phosphorylation at the CTD (introduced by the kinase Kin28, Figure 2) and the elongation factor Spt5 help to recruit the mRNA-capping enzymes Cet1, Ceg1, and Abd1 [31, 32]. These proteins catalyse three reactions which result in an addition of a 7-methyl-guanosine (m7G) cap to the 5' end of the nascent mRNA [32], which protects it from degradation [41]. The cap is then bound by the cap-binding complex which is important for recruitment of elongation factors Ctk1 and Bur1 to promote elongation and capping enzyme release (Figure 2) [31]. After the exchange of initiation with elongation factors at the 5' transition Pol II undergoes Ser2 phophorylation at the CTD by kinases Ctk1 and Bur1 [27,30,33]. Moreover Tyr1 CTD phosphorylation is introduced (in human by c-Abl), which stimulates binding of elongation factor Spt6 and suppresses transcription termination by impairing the recruitment of termination factors [36].

2.3 Termination

The first step of Pol II transcription termination takes place at the polyadenylation (pA) site, which is a conserved hexamer with the consensus AAUAAA [42]. The pA site is recognized by specific factors, which process the RNA by endonucleolytic cleavage and polyadenylation [23, 43, 44]. In yeast, this is carried out by cleavage factor IA (CFIA), cleavage factor IB (CFIB), and cleavage and polyadenylation factor (CPF) [23]. CFIA consists of Rna14, Rna15, Clp1 and Pcf11 [23], which interacts with the serine 2 phosphorylation of the CTD [22] and the CPF [44]. The RNA is cleaved by the CPF endonuclease Ysh1 and the polyA-tail (roughly 70 nt in yeast, 200 nt in mammals) is added to the mRNA to protect it from 3' degradation [22, 44].
Transcription continues several hundred (yeast) or thousand (human) nt downstream of the pA site [44]. Two major conditions are thought to contribute to destbilization and release of Pol II from the DNA [44]: the speed of the Pol II elongation complex and the stability of the RNA:DNA hybrid. There are two models for Pol II release from the DNA, but the exact mechanisms remain unclear. The 'torpedo model' assumes that an exonuclease degrades the RNA in $5' \rightarrow 3'$ direction after cleavage until it reaches and destabilizes the Pol II complex [45,46]. In the 'allosteric model', transcription of the pA site induces a destabilizing conformational change in the Pol II complex, which causes Pol II realease and transcription termination [46, 47].
There is another transcription termination pathway in yeast, which is dependent on Nrd1. This pathway terminates transcription of snRNAs, snoRNAs, SUTs, CUTs, XUTs and several ORFs [22, 24, 26, 48, 49]. Nrd1 binds the serine 5 phosphorylated CTD of Pol II [25, 50] and a tetramer motif in the RNA [26, 49]. It interacts with the helicase Sen1 and Nab1 to promote termination [51]. Unlike the pA-dependent termination pathway, transcripts terminated by the Nrd1-dependent pathway are rapidly degraded by the nuclear exosome [22].

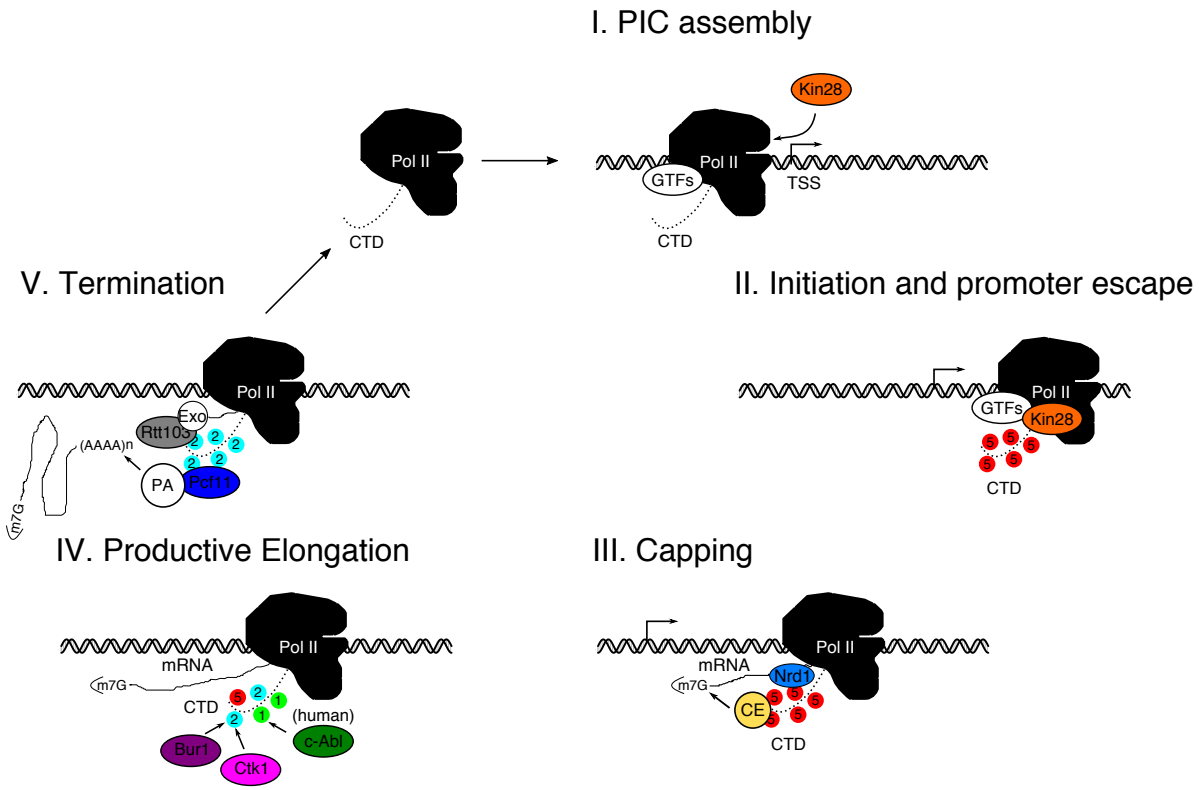


Figure 2: Overview of the transcription cycle adopted from [34]. Pol II is recruited during the assembly of the pre-initiation complex (PIC). PIC assembly is completed by binding of TFIIH whose subunit Kin28 phosphorylates the serine 5 residue (red, Ser5P) at the c-terminal domain (CTD) of Pol II during initiation. Ser5P helps to recruit capping encymes (CE) which introduce a 7-methyl-guanosine (m7G) cap to the 5' end of the nascent mRNA to protect it from degradation. Nrd1 binds the serine 5 phosphorylated CTD of Pol II and a tetramer motif in the RNA. It interacts with the helicase Sen1 and Nab1 to promote transcription termination of snRNAs, snoRNAs, SUTs, CUTs, XUTs and several ORFs. Productive elongation involves phosphorylation of the serine 2 (cyan, Ser2P, by kinases Bur1 and Ctk1) and tyrosine 1 (green, Tyr1P, by human c-Abl) residues, while Ser5P is dephosphorylated. Tyr1P impairs recruitment of termination factors during elongation. Ser2P recruits the polyadenylation (PA) machinery and 3'-RNA processing and termination factors Pcf11 and Rtt103. The exonuclease complex (Exo) degrades cleaved RNA in $5' \rightarrow 3'$ direction, which triggers destabilization and release of the Pol II complex in the torpedo model. The allosteric model assumes that a destabilizing conformational change is induced by transcription of the poly-adenylation site.

3 The histone code of transcription

Nucleosomes are the most basic unit of structural organization of the chromosomes in the nucleus [2]. The core nucleosome consists of a protein octamer - formed by H2A, H2B, H3 and H4 - and 146 nucleotides (nt) of double-stranded DNA, which is wrapped around it. These occur in an average distance of about 80 nt along the DNA, but distances can vary. The positioning of nucleosomes is determined by DNA properties and DNA-binding proteins [2,52]. For instance the SWI/SNF complex is a chromatin remodeler that can dislocate nucleosomes from the promoter to make it accessible for transcription factors and polymerase [53]. Nucleosome depletion is a hallmark of cis-regulatory regions and a strong periodic pattern of nucleosome positioning has been observed downstream of the transcription start site [54, 55].
Each core histone contains an N-terminal tail, which can be post-translationally modified. A plethora of these histone tail modifications has been described [56]. For instance the H3 lysine 36 trimethylation (H3K36me3), H4 lysine 20 monomethylation (H4K20me1) have been shown to occur in gene bodys and are predictive for gene expression [57,58,59,60]. The H3 lysine 9 trimethylation is a frequent modification in inactive heterochromatin and is implicated in chromatin structure by binding of the heterochromatin-forming protein HP1 [61]. Another heterochromatic modification is H3 lysine 27 trimethylation which is associated with transcriptionally repressed regions [59]. H3 mono- and trimethylation at lysine residue 4 (H3K4me1 and H3K4me3) are present at promoters and enhancers [62], which can be further distinguished into active and poised by presence of repressive marks H3K27me3, H3K9me3 and active marks H3K9ac, H3K27ac [63, 64]. Despite the fact that many histone modifications could be linked to different functions, the exact mechanisms of how the 'histone code' affects the associated biological processes remain largely unclear [55].

4 High-throughput experimental assays

In this section a short overview of high-throughput assays and experimental techniques which are analyzed in this work is given.

4.1 Microarrays

In order to understand genomic processes like transcription, high-throughput experimental techniques are needed to measure genomic features and phenotypes like protein binding or RNA expression in a genome-wide manner. The first assays to accomplish this were based on microarray technology. Microarrys harbour thousands to millions of oligonucletides which are anchored on their surface [65, 66]. The abundance of free DNA or RNA in a sample is measured by hybridization to the immobilized oligonucleotides on its surface.

4.1.1 Transcriptome profiling using microarrys

The first mircroarray for gene expression measurement was developed in 1995 [67]. It was designed to detect in parallel the expression of 45 genes in *Arabidopsis thaliana*. To this end, *Arabidopsis* mRNA was reverse transcribed into cDNA which was labeled using a fluorescent dye. The labeled cDNA was hybridized to the array and gene expression measurements were obtained by a scanner. Two years later a microarray was developed to measure expression of 2,479 genes in *S. cerevisiae* [68]. These early microarrays represented only a small fraction of oligonucleotides of

the genomes of interest and therefore only provided a biased view of a genome. This limitation was overcome by the develpment of whole genome microarrays (or tiling arrays), which covered the complete genomic sequence of an organism [65].
In 2003, the first whole genome tiling array design was developed for *Arabidopsis thaliana* [69]. It covered 94% of the *Arabidopsis* genome on 12 arrays, each of which contained approximately 834,000 25-mer oligonucleotides. For the first time, the transcriptome could be measured in an unbiased and genome-wide manner and 5,817 novel transcription units were discovered. In 2006, David et al. [70] developed a tiling array for *S. cerevisiae* that contained 6.5 million 25-mer oligonucleotides which represented the full genomic sequence tiled at an average of eight nucleotide intervals on each strand. This enabled genome-wide measurements in a four nucleotide resolution for double-stranded and eight nucleotide resolution for strand-specific targets. Using this array the authors quantified RNA expression of the complete yeast genome in rich media and found a total of 85% of the genome to be expressed. Later this array was used to annotate thousands of stable (stable unannotated transcripts, SUTs) and rapidly degraded RNAs (cryptic unstable transcripts, CUTs) in the yeast genome and it was shown that pervasive transcription is generated by bidirectional promoters [4].

4.1.2 ChIP-chip

ChIP-chip combines chromatin immunoprecipitation (ChIP) with hybridization on microarrays. ChIP is a widely used method to detect and quantify protein-DNA binding *in vivo* [71,72]. First protein complexes are cross-linked to DNA using formaldehyde. Then cells are harvested, lysed and the chromatin is isolated. After cell lysis, the DNA is sheared into fragments. Immunoprecipitation with an antibody which is specific for the protein of interest is used to enrich DNA fragments that are bound by it. Cross-links are then reversed, free DNA samples are isolated, amplified and labeled with a dye. In the case of ChIP-chip, the DNA is then hybridized to the microarray. Binding intensities to the oligonucleotides on the chip are detected by a scanner.
In 2000 the first for ChIP-chip studies measured binding of indiviual transcription factors in yeast intergenic regions [73,74,75]. Later, this protocol was applied on a larger scale to infer the yeast transcription regulatory network [76,77], and to create a genome-wide map of nucleosomes [54] and the pre-initiation complex in yeast [78]. The same yeast tiling array used for transcriptome profiling developed in [70] was later applied with ChIP-chip to measure genome-wide occupancies of transcription-related proteins in yeast [4,31,33,36,54], which are re-analyzed in this work together with the expression data from Xu et al. [4] (see section 14).

4.2 Next-Generation Sequencing

One important limitation of whole genome microarrays was that their application was only feasible on relatively small genomes. For larger genomes a high number of microarrays would be needed to tile the entire genome, which made their use impractical for organisms like human [65]. Roughly at the same time as tiling arrays were extensively used for whole genome analyses, massively parrallel sequencing (so-called 'Next-Generation-Sequencing') technologies began to emerge. In 2005-2007, the first systems were released and applied to Whole Genome Sequencing (WGS) [79,80]. It did not take long until Next-Generation-Sequencing was combined with other experimental assays.

4.2.1 ChIP-Seq

In analogy to the ChIP-chip protocol, ChIP-Seq combines ChIP with Next-Generation-Sequencing. Instead of hybridization with a microarray, the immunoprecipitated and amplified DNA fragments are sequenced in a massively parallel manner. In 2007 three independent groups published protocols combining ChIP with Next-Generation-Sequencing (ChIP-Seq) using the Illumina's Solexa system [59, 60, 81]. Barski et al. [59] and Mikkelsen et al. [60] used ChIP-Seq to measure covalent histone modifications in human and mouse, while Johnson et al. [81] determined binding sites of the neuron-restrictive silencer factor (NRSF) in human. While ChIP-chip and ChIP-Seq produce highly reproducible results, ChIP-seq generally exhibits a better signal-to-noise ratio and detects more and narrower peaks than ChIP-chip [82].

4.2.2 RNA-Seq and Dynamic Transcriptome Analysis

In 2008 two studies presented sequencing based methods for quantifying RNA expression (RNA-Seq) in yeast [83] and mouse [84]. RNA-Seq measures the total RNA levels in a cell. However these total RNA levels are the result of synthesis and degradation, which cannot be inferred from standard RNA-Seq protocols. To address this problem, Dynamic Transcriptome Analysis was developed (DTA) [85, 86]. During a DTA experiment, newly synthesized RNAs are labeled with the nucleoside analog 4-thiouridine (4sU), which is incorporated into RNA during transcription. The labeled RNAs are then biotinylated and purified using streptavidin-coated magnetic beads. Labeled and unlabeled RNA fractions are quantified using microarrays [86] or Next-Generation-Sequencing [26, 87], which is then used to estimate RNA synthesis and decay rates with kinetic modeling. A clear advantage of DTA compared to standard transcriptomics is that the cellular response to external stimuli can be observed at higher sensitivity and temporal resolution [86]. In this work a comprehensive map of the transcriptome in human K562 cells is derived from transient transcriptome sequencing (TT-Seq), a new and improved protocol based on labeling newly synthesized RNA using 4sU (see secion 16).

4.2.3 Detecting DNase hypersensitivity sites with DNase-Seq

Endonucleolyitc cleavage of DNA by Deoxyribonuclease I (DNase I) can be used to determine accessible regions (i.e. not bound by nucleosomes) in the genome. The mapping of DNase I hypersensitive sites (DHS) - i.e. sites that are sensitive to DNase I cleavage - is a tool to determine potential regulatory regions [88, 89]. Combined with Next-Generation sequencing it can be used to map these regions genome-wide [90, 91]. For instance the ENCODE Consortium applied DNase-Seq to 125 human cell and tissue types to identify $\sim$2.9 million DHSs [92, 93]. In the first steps of DNase-Seq, cells are lysed, nuclei are isolated and the DNA is digested with DNase I [90]. This leaves single-stranded overhangs in the DNA which are blunt ended by T4 DNA polymerase before ligation to a linker DNA. DNA fragments are attached to Dynal beads, amplified by PCR and then sequenced. A computational analysis then identifies regions with high density of DNase I cleavege sites [91]. However, a substantial amount of sequence specificity of DNase I was observed in DNase-Seq data sets and thus care must be taken when interpreting the results [94].

5 Large-scale genome projects

Since the release of the human genome sequence several large-scale projects were initiated to annotate functional elements in the DNA and understand their implications in human development and disease [95, 96, 97, 98]. To accomplish this task these projects generated and integrated massive amounts of genome-wide experimental data. The next sections introduce two of these large-scale projects from which data is used in this work to annotate regulatory elements in the human genome (see section 15).

5.1 The ENCODE project

The Encyclopedia of DNA Elements (ENCODE) project was started in 2003 as a follow up to the Human Genome Project to annotate all functional elements in the DNA [96]. The first phase (pilot phase) focused on the annotation of a subset of 30 megabases (1%) of the human genome [96, 99]. This subset (ENCODE pilot regions) consisted of 44 regions in human genome. Half of the ENCODE pilot regions were selected manually and had to contain either well characterized genes or functional elements or were required to have a high amount of comparative sequence data available. During the pilot phase, experimental approaches were tested and then implemented in the 'technology development phase'. After the release of the results of the ENCODE pilot phase in 2007 [99], the third 'data production phase' started to apply the established protocols and measure a variety of biological features in the human genome. These included various techniques for the analysis of genes and their transripts (e.g. RNA-Seq), transcription factors (e.g. ChIP-Seq), chromatin features (e.g. ChIP-Seq or DNase-Seq), DNA methylation (e.g. Methyl-Seq) and chromatin interactions (e.g. CHIA-PET) [100]. These were used for instance to annotate regulatory elements (such as promoters and enhancers) [93, 101, 102, 103] or transcripts [104]. Overall the ENCODE project assigned a biological function to 80.4% of the human genome [93]. However, this conclusion lead to one of the major critiques of ENCODE and was attributed to the loose use of the term 'function' and other methodological flaws [105, 106].

5.2 The Roadmap Epigenomics project

In 2008, the NIH Roadmap Epigenomics Mapping Consortium was launched with the goal to define epigenomic maps in stem cells and primary *ex vivo* tissues, which include DNA methylation, histone modifications and RNA transcripts [95]. This effort was intended to generate a resource for studies investigating human development, diversity and disease [107, 108, 109]. Recently Roadmap Epigenomics data was combined with ENCODE data to annotate regulatory elements in 127 cell types and tissues [110], revealing significant differences of chromatin features between various group-wise comparison such as sex, tissue or cell type indicating the need for cell-type/tissue-specific epigenomic maps [111].

6 Genomic data analysis with hidden Markov models

Hidden Markov models (HMMs) are a powerful tool for the analysis of longitudinal data and therefore ideally suited for genomic data analysis. Here I quickly want to introduce the concept and idea of a HMM.

A formal definition of a HMM is given by [112, 113]:

Definition. A **hidden Markov mode**l (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 $\mathcal{K} \times \mathcal{K}$ (row) 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*.

A HMM defines a probability distribution on a sequence of observations $\mathcal{O} = (o_0, ..., o_T)$ of length $T+1$. 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 $\Pr(s_0 = i) = \pi_i$, $i \in \mathcal{K}$.

The following toy example illustrates how HMMs can be used to model transcription (Figure 3). Let's assume that we measured the binding signal of several transcription initiation, elongation and termination factors at a protein coding genes. Initiation factors have a high binding signal in promoter regions, but a low signal in the gene body. In contrast, elongation factors are high in the gene body and low in promoter regions, etc.. Therefore each phase of transcription is represented in the data by a specific combination of protein binding signals.
Now we relate above formal definition and parameters of the HMM to our example. A HMM assumes that some observable variable (protein binding signal in our example) is generated by an unobserved or 'hidden process' (transcription in our example). This hidden process is assumed to have a discrete number of states, the 'hidden states'. In our example the hidden states are represented by the different phases of transription, e.g. initiation, elongation, termination. These hidden states generate (or emit) the observations with a certain probability distribution (the emission distribution). For instance in the 'initiation state' we observe a different composition of the transcription machinery (from the binding data) than in the elongation or termination states. This is modeled by different probability distributions. Moreover, the hidden states occur in a sequential order along a gene, usually in the order initiation $\Rightarrow$ elongation $\Rightarrow$ termination. This sequential order is modeled by transition probabilities in the HMM. For instance in our example the probability to transition from initiation to elongation is high but the probability to transition from intiation directly to termination is zero (because it is not observed in the data). In summary, a HMM allows us to explain and characterize the process of transcription as a series of 'hidden (transcription) states' directly from the protein binding data.
The observations in HMMs are not limited to binding data, but could also be the letters of a DNA or protein sequence or gene expression data. The following section illustrates the wide applicability of HMMs in genomics by giving a short overview of HMM-based methods in the field.

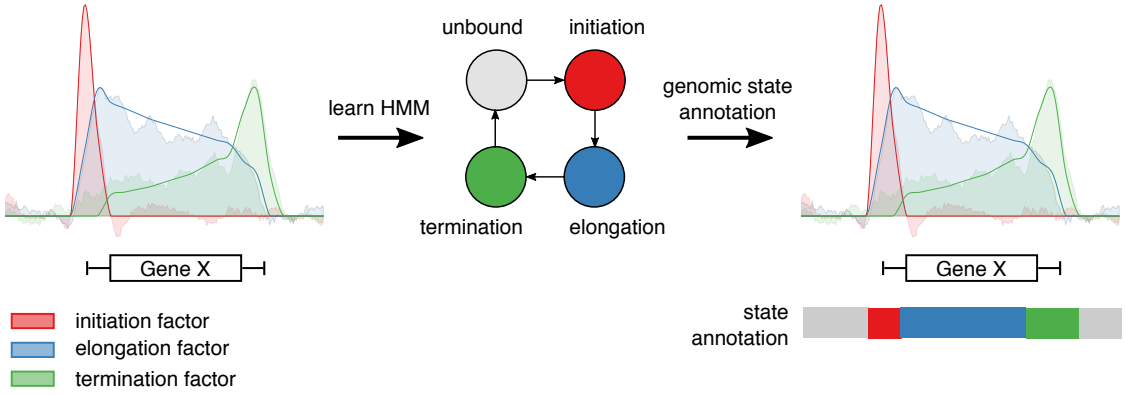


Figure 3: Example use of hidden Markov models to model transcription. (Left) The (toy) data was measured for three transcription factor at gene X. The initiation factor exhibits high signals at the 5' end of the gene, the elongation factor has high signals in the gene body and the signal of the termination factor increases towards the 3' end of the gene. Up- and downstream of the gene, signals are low for all three factors. (Middle) The signal distribution of all three factors can be modeled (in $5' \Rightarrow 3'$ direction) using a hidden Markov model (HMM) with four states. The depicted graph represents the transition probabilities (without self transitions) between the states learned from the toy example. Starting in an unbound state (upstream of the gene), the model transition to the initiation, elongation and termination state before returning to the unbound state (downstream of the gene). (Right) The model can then be used to generate an annotation of the these states along the genomic sequence (depicted as colored boxes below the signal tracks).

6.1 HMMs to detect functional elements in nucleic acid sequences

Early applications of HMMs in genomics were mainly used to find functional elements in the DNA. In the 1990s the first HMM-based gene finding programs were developed [114, 115]. Thereby, the structure of genes was encoded by the transitions between hidden states of the model. For instance the gene finder for *E. Coli* by Krogh et al. [114] used a sequence of three hidden states to encode the start codon ATG (one state for each nucleotide). The last G in the start codon then transititioned into a sequence of codon triplets modeling an exon, followed by states for the stop codon. Over the past two decades new and improved HMM-based gene finding algorithms were developed [116, 117, 118, 119]. Recently also RNA-Seq data was integrated into the models [120, 121].

One way to find potential functional elements (such as promoters, enhancers or genes) in the DNA is to look for genomic regions which are conserved between species [122]. To this end the so called phyloHMMs were developed which combine hidden Markov models with phylogenetic models [122, 123]. These methods learn a HMM with states for 'conserved' and 'not conserved' regions in the genome. The emission distributions of the states are based on phylogenetic models from a multiple sequence alignment [122]. The model is then used to compute a conservation score for each genomic position.

The detection of cis-regulatory elements (CREs) or transcription factor bindig sites in DNA sequences also is a problem that can be addressed with HMMs. The discovery of these elements was previously done by exploiting phylogenetic information from sequence alignments [124, 125, 126] or by scanning the DNA with individual or combinations of transcription factor binding sites modeled by a HMM [127,128,129]. The method by Wong and Nielsen [124] for instance used a very

simple approach to represent CRE structure in the HMM [124]. The model consists of two types of states: the background states and the CRE states. The background states model the nucleotide distribution of genomic background regions using relative nucleotide frequencies. The CRE states consist of a set of transcription factor binding motifs, which are modeled using multiple subsequent nucleotides (one state per nucleotide). More recently HMMs were also applied to *de novo* motif discovery to infer transcription factor binding motifs from nucleic acid sequences [130].

6.2 HMMs and protein sequences

Another class of HMMs that should be shortly mentioned here are profile-HMMs, whch have been applied extensively with protein sequences. For instance, they were used to classify proteins into families [131], for domain classification of transcription factors [132] or the detection of remotely homologuous proteins in sequence alignments [133]. Other HMM-based approaches that work with protein sequences are secondary structure [134] and transmembrane topology prediction [135].

6.3 Applications of HMMs to genome-wide high-throughput experiments

High-throughput experiments such as ChIP-chip or ChIP-Seq are often aimed at finding 'active' and 'inactive' (or 'bound' and 'unbound') regions. The prediction of these regions thorughout the genome can be done with a simple two-state HMM, where one state models the active and the other state the inactive regions. These can be for instance TF-bound [136,137,138,139], transcribed [136,140] or methylated regions [141,142]. A more complex example is microMUMMIE for binding site identification of microRNAs [143], which combines binding data with RNA sequence. Another example application is the identification of copy number variations using multiple states in the HMM for duplicated and depleted chromosomal regions [144, 145, 146].

6.4 Large-scale prediction of cis-regulatory elements

The discovery that genomic elements such as promoters and enhancers carry specific combinations of histone modifications (see section 3) made it possible to use these features in an integrative manner with HMMs to predict these elements genome-wide. Currently, the most popular algorithms used for this task are based on HMMs. The ENCODE and Roadmap Epigenomics projects used HMM-based methods to *de novo* infer 'chromatin states' from ChIP-Seq of TFs and histone modifications and DNase-Seq. Thereby a 'chromatin states' defines a recurring combination of histone modifications which are modeled with a probability distribution (the emission distribution). The crucial difference between these methods lies in the choice of the emission distribution. The first HMM-based methods to do this modeled combinations of histone modifications using multivariate normal distributions or nonparametric histograms [147, 148, 149]. The following sections will shortly present three state-of-the-art methods which are widely used in current genomics research.

6.4.1 ChromHMM

ChromHMM was applied by the ENCODE and Roadmap Epigenomics projects and various other studies to generate chromatin state annotations in various human cell types and tissues [93,103,110,150,151,152]. ChromHMM assumes multivariate independent Bernoulli distributions as emissions to model the presence or absence of chromatin marks (i.e. histone modification, DHS,

etc.) in 200bp bins along the genome. The emission probability of an observation vector o_t given hidden state s_t at position t is

$$\Pr(o_t|s_t) = \prod_{d\in\mathcal{D}} p_{s_{t,d}}^{o_{t,d}} \left(1 - p_{s_{t,d}}\right)^{o_{t,d}}$$

Thus, the data needs to be binarized for this model. The default approach used by ChromHMM defines presence or absence of chromatin marks in a 200bp bin as the smallest discrete number $n_d > 0$ such that $\Pr(X > n_d) < 10^{-4}$ where X is a Poisson random variable with mean $\lambda_d = \frac{\sum_{t=0}^{T} o_{t,d}}{T+1}$ [150]. But in principle any other binarization or peak-finding method can be used. Model learning in ChomHMM is done on the complete genome of an organism using an incremental version of the Expectation-Maximization (EM) algorithm [150, 153]. When applied to multiple cell types or tissues, the data is concatenated into one 'artifical genome' to learn a joint model across cell types or tissues [93, 102, 103].
An obvious limitation of ChromHMM is that the quantitative information in the data is lost due to the binariaztion. This might restrict the discrimination of features that exhibit presence of a set of marks, but at different ratios. This is the for instance the case for promoters and enhancers which have been shown to be marked by both H3K4me1 and H3K4me3, but at different ratios [62]. Moreover, all genomic positions with read counts below the binarization threshold are collapsed into a 'silent state', and the choice of the binarization cutoff is arbitrary, but crucial for the output of the model [154].

6.4.2 Segway

Segway was applied in the ENCODE project to annotate chromatin states in the human genome [93, 101, 102]. It is also the method of choice of the Ensembl Regulatory Build, which integrates epigenomic data to annotate regulatory elements in a robust manner [155]. Formally Segway is defined as a dynamic bayesian network. It is not exactly a HMM but has some similarity to it [101]. Segway uses independent Gaussian distributions to model the sequencing data (each track independently of each other) and therefore also relies on signal transformtation. This is done by applying the inverse hyperbolic sine function $o_t = \ln\left(o_t + \sqrt{o_t^2 + 1}\right)$, normalization and smoothing of the data [101, 102]. However the resulting data is still zero-inflated which can cause singularity of the variance in low coverage states during model learning. This problem is addressed in Segway by assuming a shared variance over states for a data track. Therefore the emission probability of an observation vector o_t given hidden state s_t at position t is

$$\Pr(o_t|s_t) = \prod_{d\in\mathcal{D}} \frac{1}{\sigma_d\sqrt{2\pi}} e^{-\frac{\left(o_t - \mu_{s_t,d}\right)^2}{2\sigma_d^2}}$$

Segway was originally desinged to model the data in 1 bp resolution [101], but binning is also possible [155]. In contrast to ChromHMM, Segway learns separate models when applied to multiple cell types or tissues [102]. Limitations of Segway concern its speed, data preprocessing and choice of emission function. In particular when applied in 1 bp resolution Segway is order of magnitudes slower than ChromHMM and requires much more computational resources [154]. As for ChromHMM, the method used for signal transformation is arbitrary. Moreover the zero-inflated distribution of the transformed data can be very different from a Gaussian distribution [154] and the shared variance over states is a poteintially flawed and inflexible assumption about the variance in the data.

6.4.3 EpicSeg

More recently, EpicSeg (Epigenome Count-based Segmentation) was proposed to overcome limitations of Segway and ChromHMM [154]. EpicSeg models the raw read counts in 200 bp bins along the genome using a negative multinomial distribution and runs at a comparable speed as ChromHMM [154]. Therefore - except for read mapping - it does not rely on preprocessing of the data. The negative multinomial distribution can be defined as hierarchical model where read counts o_t follow a multinomial distribution with parameters p_d, $d \in \mathcal{D}$ and $O_t^+ = \sum_{d\in\mathcal{D}} o_{t,d}$ follows a negative binomial distribution with parameters mean $\mu_{o_t,d}$, $d \in \mathcal{D}$ and rate parameter r. The default mode of EpicSeg assumes a shared rate parameter r for all states, since free r_{s_t} parameters for all states might lead to 'unrealistic models where different states have wildly different dispersion parameters' [154]. The emission probability of an observation vector o_t given hidden state s_t at position t is

$$\begin{aligned}
\Pr(o_t|s_t) &= NegativeBinomial\left(O_t^+|s_t, r, \mu_{s_t,d}\right) Multinomial\left(o_t|s_t, p_{s_t,d}\right) \\
&= \frac{\Gamma\left(r+O_t^+\right)}{O_t^+!\Gamma(r)}\left(\frac{r}{r+\mu_{s_t,d}}\right)^{O_t^+}\left(1-\frac{r}{r+\mu_{s_t,d}}\right)^{r}\frac{\left(\sum_{i\in\mathcal{D}} o_{t,d}\right)!}{\prod_{d\in\mathcal{D}}\left(o_{t,d}!\right)}\prod_{d\in\mathcal{D}} p_d^{o_{t,d}}
\end{aligned}$$

While having the advantage of modeling raw read counts and therefore omitting arbitrary preprocessing steps, also EpicSeg makes a rigid assumption about the variance by assuming only one shared rate parameter r for all states and data tracks.

7 Aims and scope of this thesis

As pointed out above (see section 6.4) current methods for genome segmentation have important shortcomings. They are limited by unrealistic data distribution assumptions and variance models. Moreover, they are not able to assign forward or reverse direction to states or properly integrate strand-specific (e.g., RNA expression) with non-strand-specific (e.g., ChIP) data, which is indispensable to accurately characterize directed processes such as transcription.
These issues are addressed in this thesis by developing the theory of bidirectional HMMs (bdHMMs) which can integrate strand-specific with non-strand-specific data to infer *directed* genomic states from genomic data *de novo*. Moreover, a more realistic emission model for count data from sequencing experiments is proposed with the GenoSTAN (Genomic STate ANnotation) algorithm.
This thesis can be divided into three parts with different applications of the developed methods that address various reasearch questions related to transcription and its regulation.
In the first application, bdHMMs are applied to a set of 22 previously published ChIP-chip and RNA expression measurements with tiling arrays to investigate the following questions (see section 14):

- What are the different states of the transcription cycle?
- What is the composition of the transcription machinery in each state?
- Is the sequence of states universal for all genes?

The second application uses GenoSTAN with today's largest compendium of chromatin assays to identify promoters and enhancers in 127 human cell types and tissues (see section 15). This

application aims at providing a reference map of promoters and enhancers with significantly higher accuracy than previous maps, which is then used to address the following question:

- What biochemical and regulatory features characterize promoters and enhancers?

In the third application, GenoSTAN is used to study transcription in human K562 cells by mapping transcription units (TUs) from TT-Seq (a new and sensitive variant of RNA-Seq) data. TUs are classified using the improved promoter and enhancer annotation, resulting in an annotation of mRNAs, lincRNAs, enhancer RNAs, antisense RNAs, and promoter-associated RNAs. This comprehensive map of stable and transient RNAs in human K562 cells is used to investigate the following questions (see section 16):

- Are there differences in transcription of promoters and enhancers?
- What are potential determinants of transcript stability?

Part II
Methods

8 Hidden Markov models

This section presents the theoretical basis for parameter learning and genomic state annotation in GenoSTAN (***Geno***mic ***ST***ate ***AN***notation), a hidden Markov model with Poisson-lognormal and negative binomial emission distributions. GenoSTAN is implemented in the software package STAN [113], which is available from Bioconductor [156]. For a formal definiton of HMMs see section 6.

8.1 The Baum-Welch algorithm

The learning problem for HMMs 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 [153], an expectation-maximization (EM) algorithm [157]. 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 $P(\mathcal{O};\, \theta)$.
Before deriving parmeter updates we define some auxiliary terms [112]. The (full) likelihood of a HMM is

$$\begin{aligned}
\Pr(\mathcal{O}, \mathcal{S};\, \theta) &= \Pr(\mathcal{O} \mid \mathcal{S};\, \theta) \cdot \Pr(\mathcal{S};\, \theta) \\
&= \prod_{t=0}^{T} Pr(o_t \mid s_t;\, \Psi) \cdot \prod_{t=1}^{T} Pr(s_t \mid s_{t-1};\, A) \cdot Pr(s_0;\, \pi) \\
&= \prod_{t=0}^{T} \psi_{s_t}(o_t) \cdot \prod_{t=1}^{T} a_{s_{t-1} s_t} \cdot \pi_{s_0} \qquad (1)
\end{aligned}$$

Let $\theta^{old} = (\mathcal{K}, \pi, A, \mathcal{D}, \Psi)$ be a HMM. Let $\mathcal{O} = (o_0, ..., o_T)$ be a sequence of observations. For $i, j \in \mathcal{K}$, $t = 1, ..., T$, we define the posterior probabilities

$$\zeta_t(i, j) = \Pr(s_{t-1} = i, s_t = j \mid \mathcal{O};\, \theta^{old}) \qquad (2)$$

$$\gamma_t(i) = \Pr\left(s_t = i | \mathcal{O}; \theta^{old}\right) \qquad (3)$$

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.

$$
\begin{aligned}
\alpha_t(i) &= \Pr(s_t = i, o_1, ..., o_t;\, \theta^{old}) \\
&= \sum_{k \in \mathcal{K}} \Pr(s_{t-1} = k, s_t = i, o_1, ..., o_t;\, \theta^{old}) \\
&= \sum_{k \in \mathcal{K}} \Pr(o_t \mid s_t = i;\, \theta^{old}) \cdot \Pr(s_t = i \mid s_{t-1} = k;\, \theta^{old}) \cdot \Pr(s_{t-1} = k, o_1, ..., o_{t-1};\, \theta^{old}) \\
&= \psi_i^{old}(o_t) \sum_{k \in \mathcal{K}} a_{ki}^{old} \alpha_{t-1}(k) \qquad (4)
\end{aligned}
$$

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

$$
\begin{aligned}
\beta_t(j) &= \Pr(o_{t+1}, ..., o_T \mid s_t = j;\, \theta^{old}) \\
&= \sum_{k \in \mathcal{K}} \Pr(o_{t+1} \mid s_{t+1} = j;\, \theta^{old}) \cdot \Pr(s_{t+2} = k \mid s_{t+1} = j;\, \theta^{old}) \\
&\quad \cdot \Pr(o_{t+2}, ..., o_T \mid s_{t+1} = k;\, \theta^{old}) \\
&= \psi_j^{old}(o_t) \sum_{k \in \mathcal{K}} a_{jk}^{old} \beta_{t+1}(k) \qquad (5)
\end{aligned}
$$

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

$$
\zeta_t(i,j) = \frac{\alpha_t(i) a_{ij}^{old} \beta_{t+1}(j) \psi_j^{old}(o_{t+1})}{\sum_{k \in \mathcal{K}} \alpha_t(k) \beta_t(k)} \qquad (6)
$$

and

$$
\gamma_t(i) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{k \in \mathcal{K}} \alpha_t(k) \beta_t(k)} \qquad (7)
$$

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 $\Pr(\mathcal{S} \mid \mathcal{O};\, \theta^{old})$,

$$
\begin{aligned}
Q\left(\theta;\, \theta^{old}\right) &= \sum_{\mathcal{S}} \Pr\left(\mathcal{S} | \mathcal{O};\, \theta^{old}\right) \log \Pr(\mathcal{O}, \mathcal{S};\, \theta) \\
&= \sum_{\mathcal{S}} \Pr\left(\mathcal{S} | \mathcal{O};\, \theta^{old}\right) \left\{ \log \pi_{s_0} + \sum_{t=1}^{T} \log a_{s_{t-1} s_t} + \sum_{t=0}^{T} \log \psi_k(o_t) \right\} \qquad (8)
\end{aligned}
$$

8.2 Updates of the initial state and transition probabilities

To get updates for the initial state probabilities π, we need to maximize $Q\left(\theta;\, \theta^{old}\right)$ under the constraint $\sum_{i \in \mathcal{K}} \pi_i = 1$. The Lagrange multiplier $\lambda\left(1 - \sum_{k \in \mathcal{K}} \pi_k\right)$ is introduced and the partial derivatives of $\frac{\partial}{\partial \pi_i} Q\left(\theta;\, \theta^{old}\right),\ i \in \mathcal{K}$ are set to zero with respect to π_i. This leads to the parameter updates [112]:

$$
\pi_i = \gamma_0(i)
$$

We introduce Lagrange multipliers $\lambda_k \left(1 - \sum_{l \in \mathcal{K}} a_{kl}\right)$, $k \in \mathcal{K}$ to derive estimates for the trainsition probabilities a_{ij}. Setting the partial derivatives of $\frac{\partial}{\partial a_{ij}} Q\left(\theta;\, \theta^{old}\right)$ with respect to a_{ij} to zero, leads to parameter updates [112]:

$$a_{ij} = \frac{\sum_{t=1}^{T} \xi_t\left(i,j\right)}{\sum_{t=1}^{T} \gamma_{t-1}\left(i\right)}$$

8.3 Poisson-lognormal and negative binomial emissions

The Poisson-lognormal and the negative binomial distribution can be thought of as extensions of the Poisson distribution that allow for greater variance. We will now motivate both distributions from a Poisson distribution with a prior on the mean of the Poisson.
Suppose that $X \sim Poisson(x|\Lambda)$ is a Poisson random variable and $\Lambda \sim Gamma\left(\lambda|\alpha,\beta\right)$. From this we can derive the negative binomial with success rate p and size r [158]:

$$\begin{aligned}
\Pr\left(X = x|\alpha,\beta\right) &= \int_0^\infty Poisson\left(x|\lambda\right) Gamma\left(\lambda|\alpha = r, \beta = \frac{p}{1-p}\right) d\lambda \\
&= \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} \lambda^{r-1} \frac{e^{-\lambda\frac{1-p}{p}}}{\left(\frac{p}{1-p}\right)^r \Gamma\left(r\right)} d\lambda \\
&= \frac{\Gamma\left(r+x\right)}{x!\Gamma\left(r\right)} p^x \left(1-p\right)^r \quad \text{where } r > 0, p \in [0,1]
\end{aligned}$$

In order to increase interpretability in the context of read counts, we re-parameterize this with mean $\mu = \frac{r(1-p)}{p}$:

$$\Pr\left(X = x|\mu,r\right) = \frac{\Gamma\left(r+x\right)}{x!\Gamma\left(r\right)} \left(\frac{r}{r+\mu}\right)^x \left(1 - \frac{r}{r+\mu}\right)^r \quad \text{where } \mu > 0$$

The Poisson-lognormal distribution can be motivated likewise. Assume that $X \sim Poisson(x|\Lambda)$ is a Poisson random variable and $\Lambda \sim \mathcal{N}\left(\log\left(\lambda\right)|\mu,\sigma\right)$. Then the Poisson-lognormal is given by [159]:

$$\begin{aligned}
\Pr\left(X = x|\mu,\sigma\right) &= \int_0^\infty Poisson\left(x|\lambda\right) \mathcal{N}\left(\log\left(\lambda\right)|\mu,\sigma\right) d\lambda \\
&= \frac{\sqrt{2\pi\sigma^2}}{x!} \int_0^\infty \lambda^{x-1} e^{-\lambda} e^{-\frac{(\log(\lambda)-\mu)^2}{2\sigma^2}} d\lambda
\end{aligned}$$

A closed form solution for this distribution does not exist. Thus numerical integration is needed to calculate probabilities, which is done using the R package poilog [160, 161].

8.3.1 Parameter updates

Assuming that the components $o_{t,d}$, $d \in \mathcal{D}$, of a single observation o_t are independent, and hence $\psi_k(o_t) = \prod_{d \in \mathcal{D}} \psi_{k,d}(o_{t,d})$. The value of s_t determines the probability of observing o_t by $\Pr(o_t \mid s_t) = \psi_{s_t}(o_t)$. HMM learning is carried out using the Baum-Welch algorithm [112]. The optimization problem for the paramters of a single emission distribution $\psi_{i,d}$ can be written as

$$\arg\max_{\psi_{i,d}} \sum_{t=0}^{T} \Pr\left(s_t = i|\mathcal{O}; \theta^{old}\right) \log \psi_{i,d}\left(o_{t,d}\right),$$

where $\Pr(s_t = i \mid \mathcal{O}; \theta^{old})$ is calculated efficiently by the Forward-Backward algorithm, and $\psi_{i,d}$ is maximized within the class of negative binomial or Poisson-lognormal distributions. An analytical solution for this problem does not exist. Thus, we resort to numerical optimzation. As indicated by [154], above formula can be very costly to compute, since the function needs to evaluate a sum over the complete observation sequence (i.e. the complete binned genome) in each iteration. However, computations are greatly simplified by grouping together observations $o_{t,d}$ with the same count number. Let $\mathcal{C}_d$ be the set of unique counts in dimension d. Then the following terms can be precomputed for all $c \in \mathcal{C}_d$ before optimization:

$$f\left(c\right) = \sum_{t;\, o_{t,d}=c} \Pr\left(s_t = i|\mathcal{O}; \theta^{old}\right)$$

The objective function becomes

$$\arg\max_{\psi_{i,d}} \sum_{c \in \mathcal{C}_d} f\left(c\right) \log \psi_{i,d}\left(c\right)$$

which avoids redundant calculations of $\psi_{i,d}(o_t)$, $t = 0, ..., T$, and greatly reduces complexity since $|\mathcal{C}_d| \ll T$.

8.3.2 Correction for library size

The sequencing depth can be very different between experiments. To address this problem precomputed scaling factors were used to correct for varying sequencing depths for a data track between cell types. In this work, the 'total count' method is used [162]. Let $\mathcal{L}$ be the set of cell types and $r_{d,l}$ the number of reads of data track $d \in \mathcal{D}$ in cell line $l \in \mathcal{L}$. The scaling factor is then computed as

$$s_{d,l} = \frac{r_{d,l}}{\sum_{k \in \mathcal{L}} r_{d,k}} \cdot \frac{\sum_{k \in \mathcal{C}} r_{d,k}}{|\mathcal{L}|}$$

The probability of an observation $o_{t,l}$ is calculated as $\Pr\left(o_{t,l}|\frac{\mu}{s_{d,l}}, r\right)$ in the case of negative binomial and $\Pr\left(o_{t,l}|\log\left(\frac{\mu}{s_{d,l}}\right), \sigma\right)$ in the case of Poisson-lognormal emissions.

8.3.3 Initialization

Initialization of model parameters is crucial for HMMs since the EM algorithm is a gradient method which converges to a local maximum. K-means is a widely used approach to derive an initial clustering to estimate model parameters [112]. In order to make this approach applicable to

sequencing data, we added a pseudocount and log-transformed the data before k-means clustering. However, without further processing k-means rarely converged and the procedure was slow on the complete data set. To address these issues, we further processed and filtered the data. First, a threshold for signal enrichment for each data track is calculated using the default binarization approach of ChromHMM [150]. The threshold is the smallest discrete number $n_d > 0$ such that $\Pr(X > n_d) < 10^{-4}$ where X is a Poisson random variable with mean $\lambda_d = \frac{\sum_{t=0}^{T} o_{t,d}}{T+1}$. All $o_{t,d} < n_d$ were set to 0, which improved convergence of k-means. To improve the speed, all genomic bins $o_{t,d}$ where $\forall d \in \mathcal{D} : \; o_{t,d} = 0$ were removed and defined as a 'background cluster'. K-means was then run on the rest of the data with $|\mathcal{K}| - 1$ clusters. This clustering (the 'background' and k-means clusters) was then used to derive an initial estimate of emission function parameters. Initial state and transition probabilities were initialized uniform.

8.4 The optimal hidden state sequence

There are two popular ways to infer the optimal hidden state sequence $\mathcal{S}$ of a HMM given an observation sequence $\mathcal{O}$. One approach simply takes the state which maximizes the posterior probability at each position [112]:

$$s_t = \arg\max_i \gamma_t(i)$$

Alternatively one can calculate the optimal hidden state sequence $\mathcal{S}$ such that $\Pr(\mathcal{S}, \mathcal{O}|\theta)$ is maximized. This is also known as the Viterbi algorithm (the optimal $\mathcal{S}$ is then referred to as the viterbi path) [112, 163]. To this end, $\delta_t(i)$ recursively computes the maximum probability of a state sequence that ends in $s_t = i$:

$$\begin{aligned} \delta_{t+1}(j) &= \max_i \delta_t(i)\, a_{ij} \psi_j(o_{t+1}) \\ \delta_t(i) &= \max_{s_1, s_2, ..., s_{t-1}} \Pr(s_1, s_2, ..., s_t = i|\theta) \\ \delta_0(i) &= \pi_i \psi_i(o_0) \end{aligned}$$

The optimal hidden state path is then obtained via backtracking [112, 163].

9 Bidirectional hidden Markov models

A bdHMM is a HMM which satisfies three additional conditions. The first two conditions deal with the structure of the underlying hidden Markov chain, and the last condition considers the nature of observations. As will be shown in the subsequent paragraph on the semantic of bdHMMs, these conditions are by no means ad hoc.

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 a 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*: The transition matrix A and the initial state distribution π satisfy

$$\pi_i a_{ij} = \pi_{\bar{j}} a_{\bar{j}\bar{i}} \quad , \; i, j \in \mathcal{K} \tag{9}$$

2. *Initiation symmetry*: The initial state distributionπ satisfies

$$\pi_i = \pi_{\bar{i}} \quad , \; i \in \mathcal{K} \tag{10}$$

3. *Observation symmetry*: Ψ satisfies

$$\psi_i(o) = \psi_{\bar{i}}(\bar{o}) \quad , \; i \in \mathcal{K}, \, o \in \mathcal{D} \tag{11}$$

9.1 The semantic of bdHMMs

Why did we choose (9), (10), and (11) as the defining properties of a bdHMM? In order to motivate our choice, let $\theta = ((\mathcal{K}, i_{\mathcal{K}}), \pi, A, (\mathcal{D}, i_{\mathcal{D}}), \Psi)$ be a bdHMM. By initiation 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 distribution π of a bdHMM is always a stationary state distribution of A. It might be surprising that the initital state distribution has to match the steady-state probabilities. This is however an uncritical constraint in practical applications, for two reasons. First, low complexity regions (unassembled regions, repeat regions, telomeres, centromeres, etc.) lead to frequent large stretches of missing values. Hence, the model is not run on complete chromosomes, but on the remaining regions with complete data. Therefore, taking the steady-state probability as initial probability is a reasonable modeling assumption. Second, these regions are typically long enough so that the initial state distribution has minimal influence on genomic state annotation.
Moreover, generalized detailed balance and initiaion symmetry together imply that the relation

$$\begin{aligned} P(s_{t-1} = i, \, s_t = j) &= P(s_{t-1} = i) \cdot P(s_t = j \mid s_{t-1} = i) = \pi_i a_{ij} \\ &= \pi_{\bar{j}} a_{\bar{j}\bar{i}} = P(s_{t-1} = \bar{j}) \cdot P(s_t = \bar{i} \mid s_{t-1} = \bar{j}) \\ &= P(s_{t-1} = \bar{j}, \, s_t = \bar{i}) \end{aligned} \tag{12}$$

holds for all states $i, j \in \mathcal{K}$ and all positions $t = 1, ..., T$. This is a most natural condition as it says that at any position of the state sequence, the probability of consecutively observing states i and j equals that of observing the respective conjugate states in reversed order. Vice versa, (12) obviously implies generalized detailed balance. Under the mild assumption that $\lim_{t \to \infty}(\pi A^t)$ always exists (this is the case, e.g., if the matrix A is ergodic, see [164]), it can be shown that (12) also implies initiation symmetry: By induction, using

$$P(s_t = j) = \sum_i P(s_{t-1} = i, s_t = j) = \sum_i P(s_{t-1} = \bar{j}, s_t = \bar{i}) = P(s_{t-1} = \bar{j}) \tag{13}$$

it follows that $P(s_t = j) = \begin{cases} \pi_j & \text{if } t \text{ is even} \\ \pi_{\bar{j}} & \text{if } t \text{ is odd} \end{cases}$. Therefore,

$$\pi_j = \lim_{t\to\infty} P(s_{2t} = j) = \lim_{t\to\infty} (\pi A^{2t})_j = \lim_{t\to\infty} (\pi A^{2t+1})_j = \pi_{\bar{j}} \tag{14}$$

which is exactly condition (10). Hence the natural condition (12) is essentially equivalent to (9) and (10). The reason for using the latter two conditions for the definition of a bdHMM is that they are simple relations in terms of the model parameters π and A.

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 undirected states (denoted by $\mathcal{K}^0$) - the fixed points $k = \bar{k}$ of i_k - and directed 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 denoted by $\mathcal{K}^+$, the reverse states by $\mathcal{K}^-$. The observation symmetry condition (11) merely ensures that conjugate directed 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 (standard) 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 $\mathcal{O} = (o_t)_{t=0,\ldots,T}$, let $\mathcal{O}^{rev} = (o_t^{rev} = \bar{o}_{T-t})_{t=0,\ldots,T}$ denote the 'reversed' observation sequence obtained by taking conjugates of all observations and reversing their order. Similarly, given a hidden state sequence $\mathcal{S} = (s_0, ..., s_T)$, let $\mathcal{S}^{rev} = (s_t^{rev} = \bar{s}_{T-t})_{t=0,\ldots,T}$ denote the 'reversed' hidden state sequence. Verify that

$$\begin{aligned} \Pr(\mathcal{S};\theta) &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \overset{(9)}{=} \pi_{s_T} \prod_{t=1}^{T} a_{\bar{s}_t \bar{s}_{t-1}} \\ &\overset{(10,9)}{=} \pi_{\bar{s}_T} \prod_{t=1}^{T} a_{\bar{s}_{T-(t-1)}\bar{s}_{T-t}} = \pi_{s_0^{rev}} \prod_{t=1}^{T} a_{s_{t-1}^{rev} s_t^{rev}} \\ &= \Pr(\mathcal{S}^{rev};\theta) \end{aligned} \tag{15}$$

Moreover,

$$\begin{aligned}\Pr(\mathcal{O} \mid \mathcal{S}; \theta) &= \prod_{t=0}^{T} \psi_{s_t}(o_t) \overset{(11)}{=} \prod_{t=0}^{T} \psi_{\bar{s}_t}(\bar{o}_t) \\ &= \prod_{t=0}^{T} \psi_{\bar{s}_{T-t}}(\bar{o}_{T-t}) = \prod_{t=0}^{T} \psi_{s_t^{rev}}(o_t^{rev}) \\ &= \Pr(\mathcal{O}^{rev} \mid \mathcal{S}^{rev}; \theta) \end{aligned} \tag{16}$$

Equations (15) and (16) imply

$$\Pr(\mathcal{O}, \mathcal{S};\, \theta) = \Pr(\mathcal{O} \mid S;\, \theta) \cdot \Pr(\mathcal{S};\, \theta) = \Pr(\mathcal{O}^{rev} \mid S^{rev};\, \theta) \cdot \Pr(\mathcal{S}^{rev};\, \theta) = \Pr(\mathcal{O}^{rev}, \mathcal{S}^{rev};\, \theta) \tag{17}$$

and

$$\Pr(\mathcal{S} \mid \mathcal{O};\, \theta) = \Pr(\mathcal{S}^{rev} \mid \mathcal{O}^{rev};\, \theta) \tag{18}$$

Finally, a bdHMM is reversible in the generalized sense,

$$\begin{aligned}\Pr(\mathcal{O};\, \theta) &= \sum_{\mathcal{S}} \Pr(\mathcal{O}, \mathcal{S};\, \theta) = \sum_{\mathcal{S}} \Pr(\mathcal{O}^{rev}, \mathcal{S}^{rev};\, \theta) \\ &= \sum_{\mathcal{S}^{rev}} \Pr(\mathcal{O}^{rev}, \mathcal{S}^{rev};\, \theta) = \Pr(\mathcal{O}^{rev};\, \theta) \end{aligned} \tag{19}$$

The second-last equality in (19) 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 (17) motivated the development of bdHMMs, and indeed condition (17) is almost their defining property: We mention without proof that under very mild assumptions on the probability distributions Ψ, any HMM satisfying (17) is a bdHMM.

9.2 Learning of the transition matrix and the initial state distribution

In this paragraph, 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. 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., $Q(\theta^{old};\, \theta^{old}) = \Pr(\mathcal{O};\, \theta^{old})$ [157]. 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 (2) and (3), 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,

$$
\begin{aligned}
Q\left(\theta;\, \theta^{old}\right) &= \sum_{\mathcal{S}} \Pr\left(\mathcal{S}|\mathcal{O};\, \theta^{old}\right) \left\{\sum_{t=1}^{T} \log a_{s_{t-1}s_t}\right\} + \sum_{\mathcal{S}} \Pr\left(\mathcal{S}|\mathcal{O};\, \theta^{old}\right) \{\log \pi_{s_0}\} + c \\
&= \sum_{s_{t-1}\in\mathcal{K}} \sum_{s_t\in\mathcal{K}} \Pr\left(s_{t-1}, s_t|\mathcal{O}, \theta^{old}\right) \left\{\sum_{t=1}^{T} \log a_{s_{t-1}s_t}\right\} \\
&\quad + \sum_{s_{t-1}\in\mathcal{K}} \Pr\left(s_1|\mathcal{O}, \theta^{old}\right) \{\log \pi_{s_0}\} + c \\
&= \sum_{k\in\mathcal{K}} \sum_{l\in\mathcal{K}} \sum_{t=1}^{T} \zeta_t\,(k,l) \log a_{kl} + \sum_{k\in\mathcal{K}} \gamma_1(k) \log \pi_{s_0} + c \qquad (20)
\end{aligned}
$$

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

$$
\frac{\partial}{\partial a_{ij}} \tilde{Q}\left(\theta;\, \theta^{old}\right) = \frac{1}{a_{ij}} \sum_{t=1}^{T} \zeta_t\,(i,j) \qquad (21)
$$

$$
\frac{\partial}{\partial a_{kl}} \frac{\partial}{\partial a_{ij}} \tilde{Q}\left(\theta;\, \theta^{old}\right) = \begin{cases} -\frac{1}{a_{ij}^2} \sum_{t=1}^{T} \zeta_t\,(i,j) \leq 0 & \text{if } (k,l) = (i,j) \\ 0 & \text{else} \end{cases} \qquad (22)
$$

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 [165] and Rsolnp [166]. 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 (section 10 and 14) with $|\mathcal{K}| = 20$ states, we achieved an acceleration by a factor of about 25.
Given a bdHMM parameter set $\theta = ((\mathcal{K}, i_{\mathcal{K}}), \pi, A, (\mathcal{D}, i_{\mathcal{D}}), \Psi)$, we define the bdHMM parameter set $\bar{\theta} = ((\mathcal{K}, i_{\mathcal{K}}), \bar{\pi}, \bar{A}, (\mathcal{D}, i_{\mathcal{D}}), \bar{\Psi})$, where $\bar{\pi}_i = \pi_{\bar{i}}$, $\bar{a}_{ij} = a_{\bar{i}\bar{j}}$, $\bar{\psi}_i(o) = \psi_{\bar{i}}(o)$, $i, j \in \mathcal{K}$, $o \in \mathcal{D}$. The modified target function is defined as

$$
\tilde{Q}\left(\theta;\, \theta^{old}\right) = Q(\theta;\, \theta^{old}) + Q(\theta;\, \bar{\theta}^{old}) \qquad (23)
$$

where Q is defined as in (8). Since both Q terms in the sum in (23) are, up to some additive constant, lower bounds of the marginal likelihood function $\Pr(\mathcal{O};\,\theta)$, so is $\tilde{Q}(\theta;\,\theta^{old})$.
For $\mathcal{S}=(s_0,...,s_T)$ let $\bar{\mathcal{S}}=(\bar{s}_0,...,\bar{s}_T)$. It is elementary to verify that

$$\begin{aligned}\Pr(\mathcal{O},\mathcal{S};\,\theta) &= \pi_{s_0}\prod_{t=1}^{T}a_{s_{t-1}s_t}\prod_{t=0}^{T}\psi_{s_t}(o_t)\\ &= \bar{\pi}_{\bar{s}_0}\prod_{t=1}^{T}\bar{a}_{\bar{s}_{t-1}\bar{s}_t}\prod_{t=0}^{T}\bar{\psi}_{\bar{s}_t}(o_t) \;=\; \Pr(\mathcal{O},\bar{\mathcal{S}};\,\bar{\theta})\end{aligned} \tag{24}$$

From (24) we deduce that

$$\Pr(s_{t-1}=i,s_t=j\mid\mathcal{O};\,\bar{\theta}^{old})=\Pr(s_{t-1}=\bar{i},s_t=\bar{j}\mid\mathcal{O};\,\theta^{old})=\zeta_t(\bar{i},\bar{j}) \tag{25}$$

and

$$\begin{aligned}Q(\theta;\,\bar{\theta}^{old}) &= \sum_{s_{t-1}\in\mathcal{K}}\sum_{s_t\in\mathcal{K}}\Pr\left(s_{t-1},s_t|\mathcal{O},\bar{\theta}^{old}\right)\left\{\sum_{t=1}^{T}\log a_{s_{t-1}s_t}\right\}+c\\ &= \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{K}}\left(\sum_{t=1}^{T}\zeta_t\left(\bar{k},\bar{l}\right)\right)\log a_{kl}+c\end{aligned} \tag{26}$$

Equations (20) and (26) 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

$$\tilde{Q}\left(\theta;\,\theta^{old}\right) \;=\; \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{K}}\sum_{t=1}^{T}\left(\zeta_t\left(k,l\right)+\zeta_t\left(\bar{l},\bar{k}\right)\right)\log\left(a_{kl}\right)+\sum_{k\in\mathcal{K}}\lambda_k\left(1-\sum_{l\in\mathcal{K}}a_{kl}\right)+c \tag{27}$$

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

$$0=\frac{\partial}{\partial a_{ij}}\tilde{Q}\left(\theta;\,\theta^{old}\right)=\frac{1}{a_{ij}}\sum_{t=1}^{T}\left(\zeta_t\left(i,j\right)+\zeta_t\left(\bar{j},\bar{i}\right)\right)\;-\lambda_i \tag{28}$$

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

$$\begin{aligned}\lambda_i\underbrace{\sum_{j\in\mathcal{K}}a_{ij}}_{1} &= \sum_{t=1}^{T}\Bigg(\underbrace{\sum_{j\in\mathcal{K}}\zeta_t\left(i,j\right)}_{\gamma_{t-1}(i)}+\underbrace{\sum_{j\in\mathcal{K}}\zeta_t\left(\bar{j},\bar{i}\right)}_{\gamma_t(\bar{i})}\Bigg)\\ \lambda_i &= \sum_{t=1}^{T}\left(\gamma_{t-1}\left(i\right)+\gamma_t\left(\bar{i}\right)\right)\end{aligned} \tag{29}$$

After substitution of (29) into (28), we solve for a_{ij}.

$$a_{ij} = \frac{1}{\lambda_i} \sum_{t=1}^{T} \left(\zeta_t \left(i, j \right) + \zeta_t \left(\bar{j}, \bar{i} \right) \right) = \frac{\sum_{t=1}^{T} \left(\zeta_t \left(i, j \right) + \zeta_t \left(\bar{j}, \bar{i} \right) \right)}{\sum_{t=1}^{T} \left(\gamma_{t-1} \left(i \right) + \gamma_t \left(\bar{i} \right) \right)} \quad , \; i, j \in \mathcal{K} \tag{30}$$

Let

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

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

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

Further, π almost satisfies initiation symmetry:

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

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 [113].

9.3 Parameter updates for multivariate gaussian emissions

The emissison distributions Ψ are also updated by maximizing the original target function Q in Equation (8). Summarizing irrelevant terms in a constant c, we have

$$Q\left(\theta, \theta^{old}\right) \quad = \quad \sum_{k \in \mathcal{K}} \sum_{t=0}^{T} \gamma_t \left(k \right) \log \left(\psi_k \left(o_t \right) \right) + c$$

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 [167]. 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 [101, 147].
We choose to model the emission probabilities $\psi_i(o_t) = P(o_t \mid s_t = i)$, $i \in [1; \mathcal{K}]$, as multivariate Gaussians, specified by the parameters μ^i, Σ^i with mean $\mu_i \in \mathbb{R}^D$ and covariance matrix $\Sigma_i \in \mathbb{R}^{D \times D}$,

$$\psi_i(o_t) = \mathcal{N}(o_t | \mu^i, \Sigma^i) = \frac{1}{(2\pi)^{\frac{D}{2}} \cdot |\Sigma^i|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} \left(o_t - \mu^i \right)^T \cdot \left(\Sigma^i \right)^{-1} \cdot \left(o_t - \mu^i \right) \right)$$

Partial derivatives $\frac{\partial}{\partial \mu^i} \log\left(\psi_i\left(o_t\right)\right)$ and $\frac{\partial}{\partial \left(\Sigma^i\right)^{-1}} \log\left(\psi_i\left(o_t\right)\right)$ are:

$$\begin{aligned}
\frac{\partial}{\partial \mu^i} \log\left(\psi_i\left(o_t\right)\right) &= \left(o_t - \mu^i\right)^T \left(\Sigma^i\right)^{-1} \\
\frac{\partial}{\partial \left(\Sigma^i\right)^{-1}} \log\left(\psi_i\left(o_t\right)\right) &= \frac{\Sigma^i}{2} - \frac{\left(o_t - \mu^i\right)\left(o_t - \mu^i\right)^T}{2}
\end{aligned}$$

Making use of emission symmetry $\psi_i(o) = \psi_{\bar{i}}(\bar{o})$, we calculate partial derivatives $\frac{\partial}{\partial \mu^i} Q\left(\theta, \theta^{old}\right)$:

$$\begin{aligned}
\frac{\partial}{\partial \mu^i} Q\left(\theta, \theta^{old}\right) &= \frac{\partial}{\partial \mu^i} \left[\sum_{k \in \mathcal{K}} \sum_{t=1}^{T} \gamma_t\left(k\right) \log\left(\psi_k\left(o_t\right)\right)\right] \\
&= \sum_{t=1}^{T} \gamma_t\left(i\right) \frac{\partial}{\partial \mu^i} \left[\log\left(\psi_i\left(o_t\right)\right)\right] + \\
&\quad \sum_{t=1}^{T} \gamma_t\left(\bar{i}\right) \frac{\partial}{\partial \mu^i} \left[\log\left(\psi_i\left(\bar{o}_t\right)\right)\right] \\
&= \sum_{t=1}^{T} \gamma_t\left(i\right) \left(o_t - \mu^i\right)^T \left(\Sigma^i\right)^{-1} + \\
&\quad \sum_{t=1}^{T} \gamma_t\left(\bar{i}\right) \left(\bar{o}_t - \mu^i\right)^T \left(\Sigma^i\right)^{-1}
\end{aligned}$$

We set this to 0 and solve for μ^i:

$$\begin{aligned}
0 &= \sum_{t=1}^{T} \gamma_t\left(i\right) \left(o_t - \mu^i\right)^T \left(\Sigma^i\right)^{-1} + \\
&\quad \sum_{t=1}^{T} \gamma_t\left(\bar{i}\right) \left(\bar{o}_t - \mu^i\right)^T \left(\Sigma^i\right)^{-1} \\
0 &= \sum_{t=1}^{T} \left[\gamma_t\left(i\right) \left(o_t - \mu^i\right) + \gamma_t\left(\bar{i}\right) \left(\bar{o}_t - \mu^i\right)\right] \\
0 &= \sum_{t=1}^{T} \left[\gamma_t\left(i\right) o_t + \gamma_t\left(\bar{i}\right) \bar{o}_t\right] - \\
&\quad \mu^i \left(\sum_{t=1}^{T} \left[\gamma_t\left(i\right) + \gamma_t\left(\bar{i}\right)\right]\right) \\
\hat{\mu}^i &= \frac{\sum_{t=1}^{T} \left[\gamma_t\left(i\right) o_t + \gamma_t\left(\bar{i}\right) \bar{o}_t\right]}{\sum_{t=1}^{T} \left[\gamma_t\left(i\right) + \gamma_t\left(\bar{i}\right)\right]}
\end{aligned}$$

And thus:

$$\hat{\mu}^i = \begin{cases} \frac{\sum_{t=0}^{T}[\gamma_t(i)o_t+\gamma_t(\bar{i})\bar{o}_t]}{\sum_{t=0}^{T}[\gamma_t(i)+\gamma_t(\bar{i})]} & \textbf{if } i \textbf{ is directed} \\ \frac{\sum_{t=0}^{T}\gamma_t(i)o_t}{\sum_{t=0}^{T}\gamma_t(i)} & \textbf{if } i \textbf{ is undirected} \end{cases}$$

Next, we calculate partial derivatives $\frac{\partial}{\partial(\Sigma^i)^{-1}}Q\left(\theta,\theta^{old}\right)$:

$$\begin{aligned} \frac{\partial}{\partial\left(\Sigma^i\right)^{-1}}Q\left(\theta,\theta^{old}\right) &= \frac{\partial}{\partial\left(\Sigma^i\right)^{-1}}\left[\sum_{k\in\mathcal{K}}\sum_{t=1}^{T}\gamma_t\left(k\right)\log\left(\psi_k\left(o_t\right)\right)\right] \\ &= \sum_{t=1}^{T}\gamma_t\left(i\right)\left(\frac{\Sigma^i}{2}-\frac{\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T}{2}\right)+ \\ &\quad \sum_{t=1}^{T}\gamma_t\left(\bar{i}\right)\left(\frac{\Sigma^i}{2}-\frac{\left(\bar{o}_t-\mu^i\right)\left(\bar{o}_t-\mu^i\right)^T}{2}\right) \end{aligned}$$

Setting this to 0 and solving for Σ^i yields:

$$\begin{aligned} 0 &= \sum_{t=1}^{T}\gamma_t\left(i\right)\left(\frac{\Sigma^i}{2}-\frac{\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T}{2}\right) \\ &\quad \sum_{t=1}^{T}\gamma_t\left(\bar{i}\right)\left(\frac{\Sigma^i}{2}-\frac{\left(\bar{o}_t-\mu^i\right)\left(\bar{o}_t-\mu^i\right)^T}{2}\right) \\ 0 &= \Sigma^i\sum_{t=1}^{T}\left[\gamma_t\left(i\right)+\gamma_t\left(\bar{i}\right)\right]- \\ &\quad \sum_{t=1}^{T}\left[\gamma_t\left(i\right)\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T+\gamma_t\left(\bar{i}\right)\left(\bar{o}_t-\mu^i\right)\left(\bar{o}_t-\mu^i\right)^T\right] \\ \hat{\Sigma}^i &= \frac{\sum_{t=1}^{T}\left[\gamma_t\left(i\right)\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T+\gamma_t\left(\bar{i}\right)\left(\bar{o}_t-\mu^i\right)\left(\bar{o}_t-\mu^i\right)^T\right]}{\sum_{t=1}^{T}\left[\gamma_t\left(i\right)+\gamma_t\left(\bar{i}\right)\right]} \end{aligned}$$

Therefore,

$$\hat{\Sigma}^i = \begin{cases} \frac{\sum_{t=0}^{T}\left[\gamma_t(i)\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T+\gamma_t(\bar{i})\left(\bar{o}_t-\mu^{\bar{i}}\right)\left(\bar{o}_t-\mu^{\bar{i}}\right)^T\right]}{\sum_{t=0}^{T}[\gamma_t(i)+\gamma_t(\bar{i})]} & \text{if } i \text{ is directed} \\ \frac{\sum_{t=0}^{T}\gamma_t(i)\left(o_t-\mu^i\right)\left(o_t-\mu^i\right)^T}{\sum_{t=0}^{T}\gamma_t(i)} & \text{if } i \text{ is undirected} \end{cases}$$

9.4 De novo inference of state direction

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

$$dir_k = \frac{\sum_{t=0}^{T} \left| \Pr\left(s_t = k|\mathcal{O},\theta\right) - \Pr\left(s_t = \bar{k}|\mathcal{O},\theta\right) \right|}{\sum_{t=0}^{T} \left(\Pr\left(s_t = k|\mathcal{O},\theta\right) + \Pr\left(s_t = \bar{k}|\mathcal{O},\theta\right) \right)} \tag{31}$$

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 direction 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 (31) 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 section 14 and Supplementary Figure 5 in [113]).

9.5 Initialization of bdHMMs

If strand-specific data is available, the number of directed and undirected states can be set in an intuitive manner in advance. For the yeast data, the strand-specific expression data was first split into regions expressed on either the + or - strand and unexpressed regions. Directed state means were initialized as a k-means clustering from the expressed regions while undirected states were initialized using k-means on the unexpressed regions. We found that initialization by k-means works very well and generally converges to a higher likelihood than multiple random starts, in agreement with [112]. To not introduce further biases towards the k-means initialization and allow the EM to explore solutions which are further from it, covariance matrices were initially set to the covariance of the whole data and transition and initial state probabilities were initialized uniform.
In the absence of strand-specific data and without directionality annotation, we suggest to apply the directionality score that can be used as a posterior criterion to merge twin states into one undirected state, as we demonstrate for the CD4 T-cell chromatin modification data.

9.6 Simulations

The performance of bdHMM regarding parameter inference and state annotation on data not used for training was assessed using simulated datat sets. 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 $\pi^* = (\pi^*_i)_{i\in\mathcal{K}}$, $\pi^* A^* = \pi^*$.

$$a_{ij} \quad = \quad \frac{\pi^*_i a^*_{ij} + \pi^*_{\bar{j}} a^*_{\bar{j}\bar{i}}}{\pi^*_i + \pi^*_{\bar{i}}}$$

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

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,

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

and A together with π satisfy generalized detailed balance,

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

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 directed 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 Figure 4. 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 were generated. Model 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 (Figure 4). 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 (Figure 4).

10 Analysis of directed genomic states in yeast

The following describes analyses that were carried out to annotate and analyze directed genomic states in yeast using bdHMMs. The results are presented in section 14.

10.1 Experimental data and preprocessing

The experimental yeast dataset was compiled from public data [4,31,33,36,54]. All measurements were done using the high density custom-made Affymetrix tiling array (PN 520055) which tiles each strand of genomic DNA in yeast at a resolution of 8bp. ChIP experiments were normalized using the R/Bioconductor [156,161] package Starr [168] as previously described [169]. Expression data was normalized using the tilingArray package [170].
The human chromatin modification dataset was downloaded from the supplemental website of [150], where they provided the preprocessed sequencing and binary data.

10.2 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 [4]. 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 a ± 250 bp flanking region. We further selected 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 [171, 172]. A GO group was considered active if the posterior probability was > 0.5.

10.3 Targeted identification of genomic features

Let $S_+ = \{PE1_+, PE2_+, eE1_+, eE2_+, eE3_+, mE1_+, mE2_+, mE3_+, mE4_+, mE5_+, lE1_+, lE2_+lE3_+, T1_+\}$ be a set containing all forward states, excluding state $P/T2_+$. Let S_- be defined likewise for reverse states. 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. Transcripts were constrained to have a minimal length of 80bp. We uniquely assigned the 6,068 predictions to previously annotated transcripts [4], 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 [173] and expression data from [4].

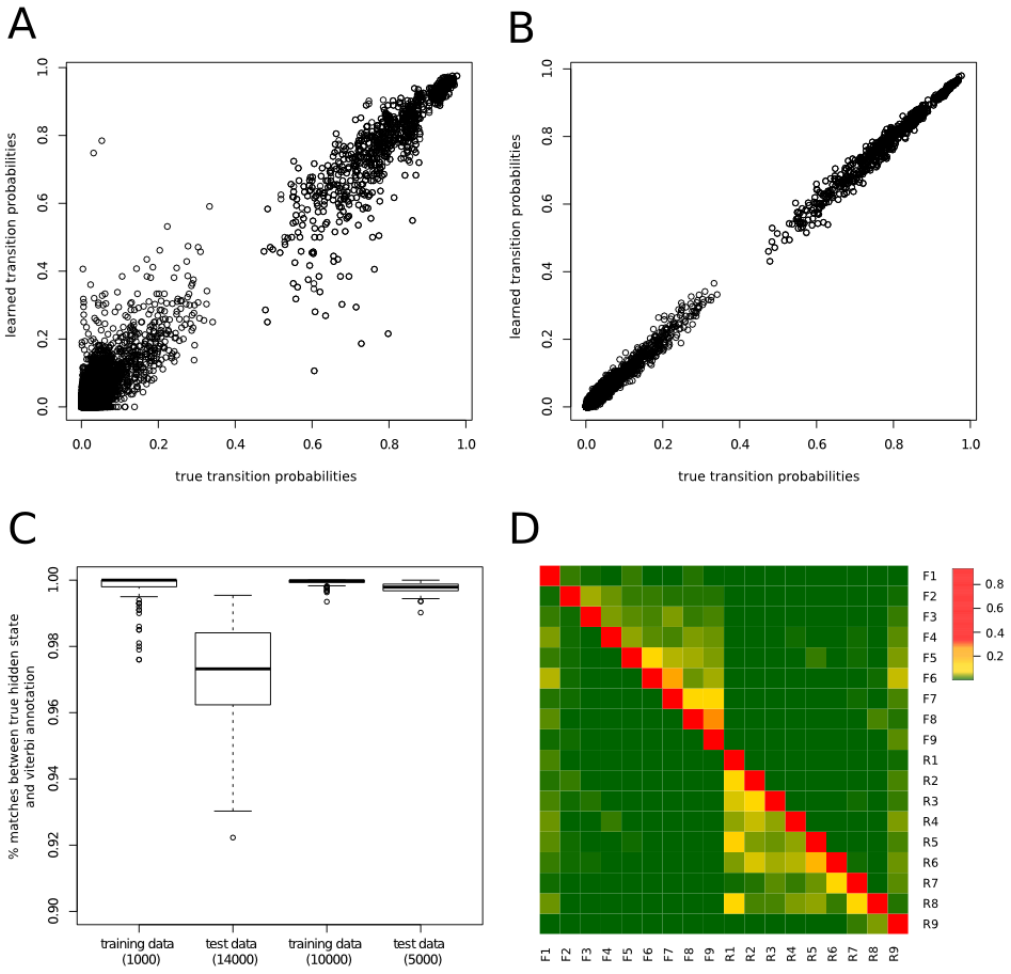


Figure 4: Simulations show good performance of bdHMM parameter inference. 100 simulations (15000 observations in each round) were carried out to assess performance of recovery of bdHMM transitions when the model was learned on 1000 (A) or 10000 (B) observations. (C) shows the respective recovery of state annotation on the training data and data not used for learning (test data). (D) shows an example of simulated bdHMM transitions.

10.4 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 lengths for a state). Motif search was carried out using XXmotif [174], 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 [175] was used to search databases for similar known motifs. Functional descriptions of transcription factors were obtained from SGD [173].

11 Analysis of chromatin modifications in human CD4 T-cells

The following describes analyses that were carried out to annotate and analyze directed chromatin states in human CD4 T-cells. The results are presented in section 14.

11.1 Fitting a standard HMM and a bdHMM to human chromatin modifications

We fitted a bdHMM to binary chromatin modification data from Ernst and Kellis [150] 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 [104] transcribed units (version 3c) to set directionality 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 a median 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.

11.2 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 distributions. This means that conjugate forward and reverse bdHMM states are mapped to the same ChromHMM state. 83% of state annotations matched between bdHMM and ChromHMM. To account for differences in the implementation and model fitting (ChromHMM for instance uses a non-deterministic version of the online EM while our implementation uses the standard EM algorithm) of ChromHMM and bdHMM, we also re-fitted the transitions of a standard HMM using the STAN package, which was initialized with the

parameters reported by Ernst and Kellis [150], keeping the emission distributions fixed. The agreement between the bdHMM and re-fitted HMM annotation was 97%, showing that bdHMMs essentially add directionality to chromatin states.

12 Chromatin state annotation and benchmark of GenoSTAN in 127 ENCODE cell types and tissues

The following describes analyses that were carried out to annotate and analyse chromatin states in 127 human cell types and tissues from the ENCODE [93] and Roadmap Epigenomics projects [110] using GenoSTAN. The results are presented in section 15.

12.1 Data preprocessing

Three data sets, benchmark I, II and III (Figure 13B), where compiled from the ENCODE and Roadmap Epigenomics projects [93, 110]. Benchmark I (K562 ENCODE) sequencing data was mapped to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium) using Bowtie 2.1.0 [176]. Samtools [177] was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped. To estimate midpoint positions of the ChIP-Seq fragments, the (single end) reads were shifted in the appropriate direction by half the average fragment length as estimated by strand coverage cross-correlation using the R/Bioconductor package chipseq [156]. Next, ChIP-Seq tracks were summarized by the number of fragment midpoints in consecutive bins of 200 bp width. The data for the 127 ENCODE and Roadmap Epigenomics cell types (benchmark II and III) was downloaded as preprocessed tagAlign files from the Roadmap Epigenomics supplementary website [110]. Fragment length was again estimated using the R/Bioconductor package chipseq and reads were shifted by the fragment half size to the average fragment midpoint [156]. The genome was partitioned into 200bp bins and reads were counted within each bin.

12.2 Model fitting of GenoSTAN

GenoSTAN was fitted on the complete data of benchmark data set I. The signal used for GenoSTAN model training on Benchmark data set II and III was extracted from ENCODE pilot regions (1% of the human genome analyzed in the ENCODE pilot phase [99]) for each cell type, which together covered 20% and 127% of the human genome. The GenoSTAN-nb-20 model was learned in one day, the GenoSTAN-Poilog-20 model in two days using 10 cores. Model learning on Benchmark set II using 10 cores took three (GenoSTAN-nb-127) and six days (GenoSTAN-Poilog-127). Precomputed library size factors were used to correct for variation in read coverage.

12.3 Model fitting of ChromHMM, Segway and EpicSeg

The data was binarized as described in [150] and ChromHMM was fitted with default parameters. Before applying Segway, the data was transformed using the hyperbolic sine function [101] and a runing mean over a 1kb sliding window was computed to smooth the data. Segway was fitted on ENCODE pilot regions using a 200bp resolution. EpicSeg was fitted on the untransformed count data with default parameters.

12.4 Processing of chromatin state annotations and external data

All state annotations and external data were lifted to the hg20/hg38 (GRCh38) genome assembly using the liftOver function from the R/Bioconductor package rtracklayer [178]. Overlap of state annotations with external data was calculated with GenomicRanges [179].

12.5 Computation of area under curve

AUC values were calculated on benchmark I for GenoSTAN, ChromHMM, Segway and EpicSeg. To this end, a segmentation was transformed into a binary classifier and evaluated as follows. Each 200bp bin in the genome overlapping with HOT (TSS) regions was considered as 'true condition', the rest as 'false'. For each state S the precision for recalling HOT (TSS) regions was calculated as the fraction of all segments annotated with S that overlapped with a HOT (TSS) region. States were then sorted by decreasing precision. The rank of each state was used as score in the prediction of HOT (TSS) regions on each 200bp bin in the genome, which was then used to calculate AUC values.

12.6 Analysis of transcription factor (co-)binding

TF enrichment in chromatin states was calculated as described earlier [180]. Let TF^{nt} be the total number of nucleotides in the binding sites (peaks) of a TF and TF_s^{nt} the number of nucleotides in the binding sites that overlap with state s. Further let s^{nt} be the total number of nucleotides in the genome covered by state s and let l be the length of the genome. TF enrichment is then calculated as $\frac{TF_s^{nt}/TF^{nt}}{s^{nt}/l}$. For each TF, enrichments were normalized to sum up to 1 across all 18 chromatin states (GenoSTAN-Poilog-K562). The co-binding rate was calculated as the frequency of binding sites of two TFs that co-occur in a chromatin state divided by the number of all binding sites of the two TFs (Jaccard index).

12.7 Tissue-specific enrichment of disease- and complex trait-associated variants in regulatory regions

The GWAS catalog was obtained from the gwascat package from Bioconductor [156, 181]. Statistical testing was carried out in a similar manner as described in [110]. The enrichment of SNPs from individual genome-wide association studies was calculated for traits with at least 20 variants. SNPs for each trait were overlapped with promoter and enhancer regions and tested against the rest of the GWAS catalogue as background using Fisher's exact test. P-values were adjusted for multiple testing using the Benjamini-Hochberg correction. In order to calculate the recall and frequency of SNPs, promoter and enhancer states were randomly sampled until a genomic coverage of 2% for enhancers and 1% of promoters was reached. This was done to control for the fact that methods can differ among each other regarding the length of the promoters and enhancers they predict. This procedure was repeated 100 times enabling the calculation of 95% confidence intervals.

12.8 Availability of GenoSTAN and chromatin state annotations

GenoSTAN is part of the R/Bioconductor package STAN [113]. The combined promoter and enhancer annotation and all chromatin state annotations for benchmark I, II and III can be downloaded from http://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN.

13 Mapping the human transient transcriptome from TT-Seq data

The following describes analyses that were carried out to annotate transcription units that were derived from transient transcriptome sequencing (TT-Seq). Additional experimental protocols (carried out by Margaux Michel) and statistical methods for calculation of RNA half lives and synthesis rates (carried out by Björn Schwalb) can be found in Appendix section 18. The results are presented in section 16.

13.1 Transcription Unit (TU) annotation

Genome-wide coverage was calculated from TT-Seq fragment midpoints in consecutive 200 bp bins throughout the genome. Binning reduced the number of uncovered positions within expressed transcripts and increased the sensitivity for detection of lowly synthesized transcripts. A two-state hidden Markov model with a Poisson-lognormal emission distributions was learned using GenoSTAN in order to segment the genome into "transcribed" and "untranscribed" states, which yielded an initial prediction of 86,676 TUs. In order to filter out spurious predictions, we defined a threshold for minimal expression (RPK) based on TUs overlapping with annotated GRO-cap TSSs [13]. The threshold was optimized based on the Jaccard-Index, which resulted in 39,811 TUs with a minimal RPK of 15.5 (Figure 5A). To further filter these, we required each TU to overlap with an annotated GRO-cap TSS, an annotated GENCODE transcript (version 22, [104]), or that the TSS of the TU overlaps with a prediction of an active promoter state (PromW.5 or Prom.11 from GenoSTAN-Poilog-K562, see Methods section 12, Figure 14) or enhancer state (EnhW.2 or Enh.15 from GenoSTAN-Poilog-K562, see Methods section 12, Figure 14) from our chromatin state segmentation. 21,874 TUs were supported by at least one of these external data sets (Figure 5B). Subsequently, TU start and end sites were refined to nucleotide precision by finding borders of abrupt coverage increase or decrease between two consecutive segments in the two 200 bp bins located around the initially assigned start and stop sites via fitting a piecewise constant curve to the coverage profiles (whole fragments) for both replicates using the segmentation method from the R/Bioconductor package tilingArray [170].

13.2 Transcript sorting

We sorted each TU into one of the following seven classes: enhancer RNA (eRNA), short intergenic non-coding RNA (sincRNA), antisense RNA (asRNA), convergent RNA (conRNA), upstream antisense RNA (uaRNA), long intergenic non-coding RNA (lincRNA) and messenger RNA (mRNA). First, TUs reciprocally overlapping by at least 50% in the same strand a GENCODE annotation (version 22, [104]) were classified as the respective GENCODE transcript type (e.g. mRNAs and lincRNAs). Next, TUs located on the opposite strand of either a mRNA or lincRNA were classified as asRNA – if the TSS was located $>$ 1 kb downstream of the sense TSS – as uaRNA if its TSS was located $<$ 1 kb upstream of the sense TSS – and as conRNA if its TSS was located $<$ 1 kb downstream of the TSS. Each of the remaining TUs did not overlap with GENCODE annotation and were classified as eRNA – if its TSS fell into an enhancer state – or as sincRNA – if its TSS fell into a promoter states. This resulted in 19,219 non-ambiguously classified RNAs on which the rest of the analysis was focused.

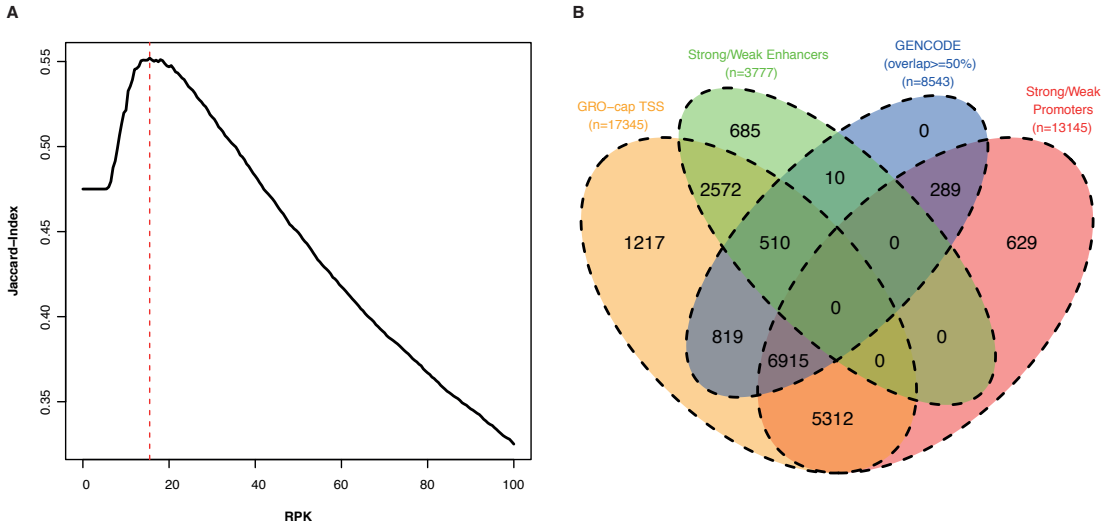


Figure 5: Accurate annotation of transcripts based on TT-Seq data using GenoSTAN (A) Jaccard index (overlap with GRO-cap TSSs) for different choices of thresholds (RPK, x-axis). (B) Venn diagram showing the overlap of the predicted and filtered 21,874 TUs with external data sets.

13.3 RNA structure and U1 motifs

The first 1000 nt from the RNA 5'-end (TSS) of each transcript was divided into 100 nt bins, where successive bins were shifted by 50 nt. The free folding energy of each of these bins was calculated using RNAfold from the ViennaRNA package [182] and the bin with the minimal free energy was selected for plotting as a measure for the most stable local structure within the region. Predicted structured RNAs in the human genome were selected from [183], overlapped with the TT-Seq transcript annotation and half-lives were plotted (Appendix Figure A11B, C). To analyze RNAs for the occurrence of U1 motifs, the 5'-most 1000 nt of each transcript were screened for occurrences of the consensus sequence of the U1 binding site (GGUAAG) and for those of the 5'-splice site GGUGAG and GUGAGU. Transcripts were then divided into 'zero' or 'one or more' occurrences and transcript lengths and half-lives were plotted.

Part III
Results & Discussion

14 Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle

The results presented in this section were previously published in [113]. For detailed author contributions see page ix.

An important question in molecular 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 multi-protein 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 and various post-translational modifications of Pol II and histones. Analysis of genome-wide occupancy maps of Pol II-associated factors obtained by chromatin immunoprecipitation (ChIP) in yeast indicates the presence of distinct protein complexes for the initiation, elongation, and termination of transcription, which are formed during a universally conserved mRNA transcription cycle [33, 78, 184]. These conclusions were deduced from metagene analysis, i.e., the averaging of occupancy profiles over a pre-selected set of representative genes. In the present work, we check this hypothesis on the single-gene level.

To systematically investigate occupancy profiles in an unbiased, position-specific manner, hidden Markov models [112] were used to describe longitudinal observations as a sequence of discrete states (here: genomic states, which model the genome-associated complexes). HMMs have been used to infer chromatin states and annotate enhancers, promoters, transcribed and quiescent regions in the genome of human [101, 102, 103, 147, 148, 150, 151] and fly [185, 186]. For instance, Ernst and Kellis [150] infer promoter and transcribed chromatin states in human T-cells, which occur in a typical order upstream and downstream of annotated transcription start sites (TSSs). However, these state-of-the-art HMM approaches infer genomic states in a non-strand-specific (or undirected) manner. For example, they cannot decide whether a bona fide "TSS-upstream" state generally precedes or follows a bona fide "TSS-downstream" state. Directionality information needs to be included in a post-processing step. Moreover, these models lack a sound way to integrate strand-specific (e.g. expression) with non-strand-specific (e.g. ChIP) data, which is indispensable to appropriately characterize strand-specific genomic processes.

To address these issues, we develop the theory of bidirectional hidden Markov models (bdHMMs), a novel probabilistic model that annotates directed states from non-strand-specific data (such as ChIP), and optionally strand-specific data (such as RNA expression). We introduce the concept of 'directed genomic states', which encode directionality information and thus provide a more realistic model of the underlying genome-associated complexes and their transitions. We present a very efficient algorithm for the learning of the bdHMM, available with the R/Bioconductor package STAN (http://www.bioconductor.org/packages/devel/bioc/html/STAN.html).

The broad applicability of our method is demonstrated on two entirely different data sets, namely on a tiling array transcription factor dataset in yeast and a deep-sequencing histone dataset in

human. We show that bdHMM produces more accurate genome annotations than standard HMM. Our bdHMM analysis of previously defined chromatin states in human T-cells [150] *de novo* identifies directed chromatin state patterns and provides an improved annotation of the human 'histone code'. Application of the bdHHM method to a set of 22 genomic profiles in the *S. cerevisiae* finds new transcription units and DNA sequence motifs, and unveils so far unknown variations in the Pol II transcription cycle. The yeast and human data sets, their state annotation, and bdHMMs which generated them, are available from the supplementary website www.treschgroup.de/STAN.html. Using essentially the same set of parameters, the bdHMM is as easy to learn as standard HMM while extracting more information. We therefore anticipate bdHMM to replace standard HMM in a wide range of genomics analyses.

14.1 Annotation of directed genomic states using bdHMMs

Standard and bidirectional HMMs are best understood with the help of a simulated dataset. A precise definition of the HMM and a bdHMM is given in the Methods (section 8 and 9). The example in Figure 6 considers a part of the genome where transcription occurs as a sequence of three different genomic segments. The transcribed regions split into segments of early (E) and late (L) transcription activity, and they are flanked by untranscribed (U) segments.

The order of the three segments U, E, and L along the genome depends on the orientation of the respective gene (Figure 6A, grey arrows). ChIP measurements $o_0, o_1, ..., o_T$ for a single protein at genomic positions $t = 0, 1, ..., T$ were simulated with low (U), medium (E) and high (L) average occupancy in the different segments. Note that these ChIP signals do not contain strand-specific information. An HMM defines a probability distribution on a sequence of observations $o_0, ..., o_T$. It assumes that each observation o_t is *emitted* by a corresponding (unobserved) state variable s_t which can assume values from a finite set of hidden states. The value of s_t determines the probability of observing o_t, $\Pr(o_t \mid s_t)$. The hidden variables form a first order Markov chain, which means that the probability for observing s_t depends only on s_{t-1}, the transition probability $\Pr(s_t \mid s_{t-1})$. After the learning of these probabilities, the HMM outputs the so-called Viterbi path, which is the most likely state sequence $s_0, s_1, ..., s_T$ that generated the observations. In our example, the Viterbi path provides a genome annotation.
A standard HMM with 3 hidden states can distinguish the three protein occupancy levels; the three states correspond to the three genomic segments (Figure 6B) and are therefore also called U, E, and L. However, the transition probabilities in the standard HMM are symmetric because the number of observed transitions between successive segments, 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 Methods section 9 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

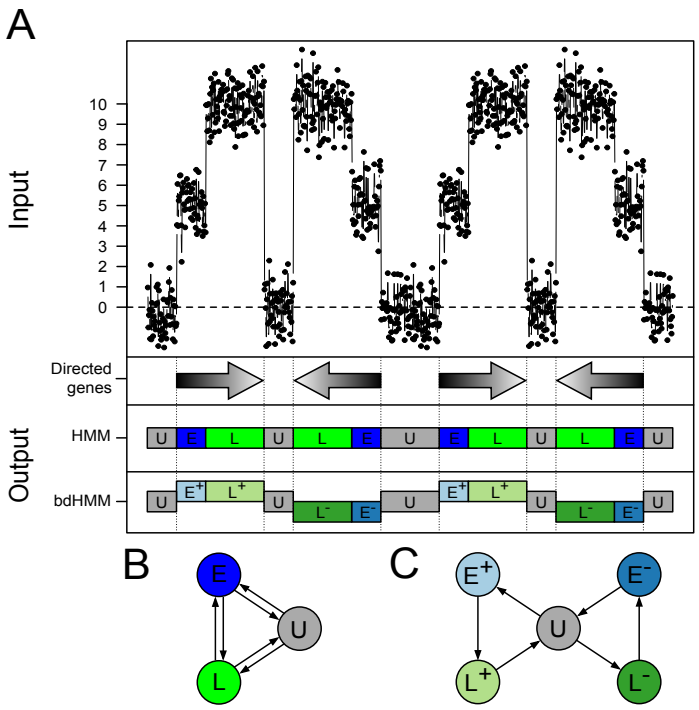


Figure 6: Principle of bidirectional HMM (bdHMM). (A) Simulated occupancy signal (1st track from the top) for a putative factor with a low level (centered at 0) in untranscribed regions (state U), an intermediate level in the 5' part of genes (state E), and a high level in the 3' part of genes (state L). Arrows (2nd track) depict boundaries and orientation of transcription. Unlike standard HMM (3rd track) bdHMM (4th track) infers strands (+ or -) to expressed states (E, L). (B) HMM transition graph. Because orientation of transcription is not modeled by standard HMM, 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 HMM, bdHMM, which has explicit strand-specific expressed states (E^+/E^- and L^+/L^-), allows inferring only the correctly oriented transitions.

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 (Figure 6B,C). These two conditions enable the recovery of the direction of genomic states (Figure 6A). Although the formal number of states doubles, the effective number of parameters does not increase due to the bdHMM constraints.
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 (Methods section 9, Supplementary Figure 8 and 9 in [113]). The bdHMM is implemented in the R package STAN (Genomic ***ST***ate ***AN***notation), which is freely available on Bioconductor (http://www.bioconductor.org/).

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

We applied the bdHMM to ChIP data in *S. cerevisiae*, 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 . The data set was complemented by strand-specific mRNA expression data [4] (Figure 7).

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 [101, 102, 150] for integrative analysis of ChIP data. We tried several state numbers (data not shown) and found that 20 states yielded an appropriate trade-off between model complexity and biological interpretability. Simulations from the inferred bdHMM recovered model parameters with high accuracy and further confirmed the validity and stability of the model (Supplementary Figure 9 in [113]).
The genome-wide state annotation was derived as the most likely state path (Viterbi decoding, Figure 7), which partitioned the 12 Mb yeast genome into 48,507 directed and 10,760 undirected state segments with distinct bdHMM states. This yields a strand-specific partitioning of the yeast genome into segments of directed genomic states. Alternative to Viterbi decoding, posterior decoding or mixed approaches (Posterior-Viterbi decoding, [187]) could be used. Generally, Viterbi decoding is less subject to state flipping compared to posterior decoding. However, we did not see relevant differences between both approaches in this application (97% of genomic positions are annotated with the same state when comparing Viterbi and posterior decoded state paths).

14.3 bdHMM state annotation recovers annotated genomic features with high accuracy

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

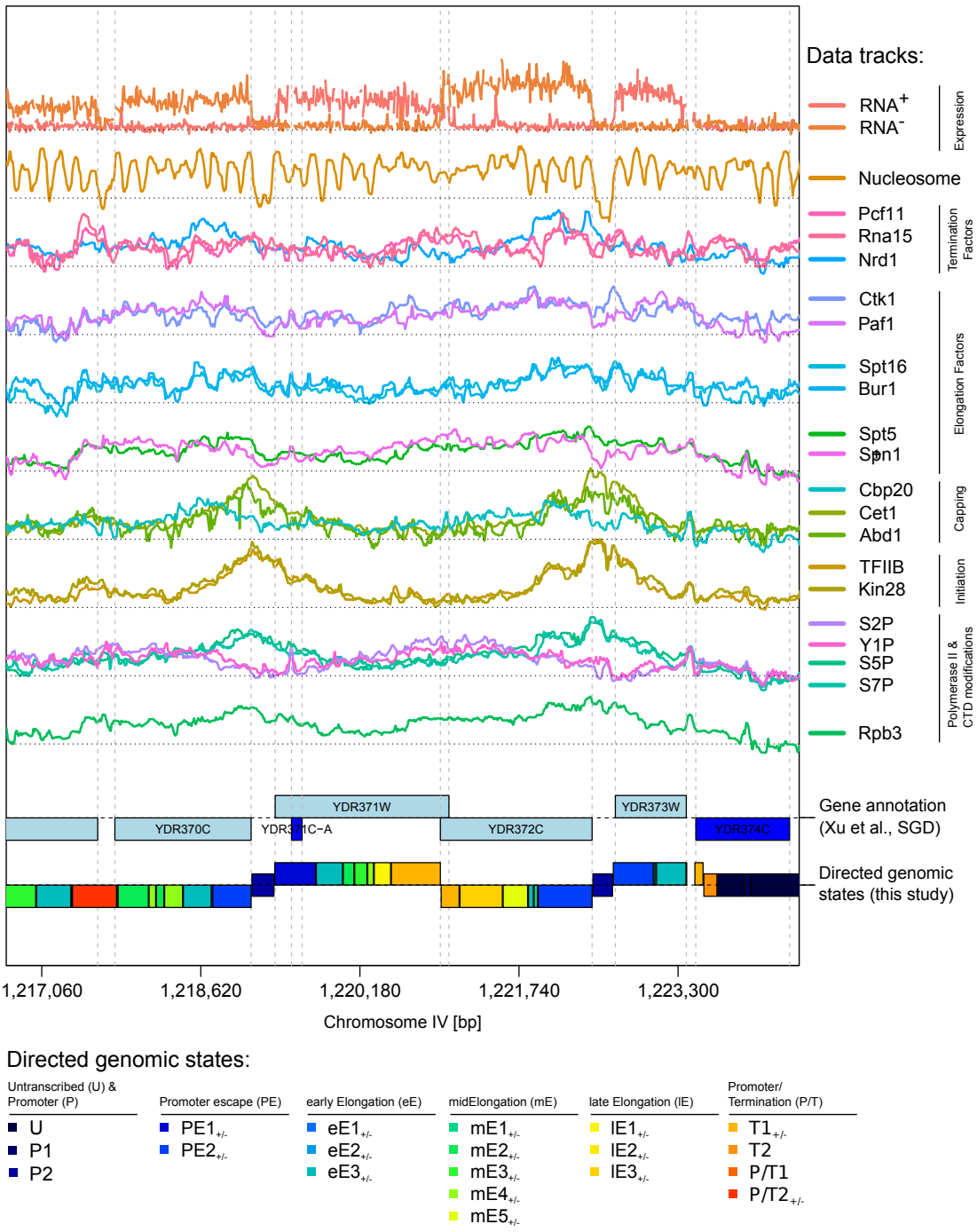


Figure 7: De novo annotation of directed genomic states from genome-wide transcription data in yeast using bdHMM. Input for the bdHMM are, from top to bottom: strand-specific wild-type RNA levels, occupancy maps of nucleosomes, 3 termination factors, 6 elongation factors, 3 capping factors, 2 initiation factors, 4 CTD modifcations, and 1 core Pol II 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.

states. In particular - since the HMM is learned without symmetry constraints for twin states - there is no obvious pairing between the forward (+) and the reverse (-) states, demonstrating the need for bdHMM (Supplementary Figure 1 in [113]).
In order to re-annotate transcription throughout the yeast genome and compare the performance of bdHMM and HMM, we applied a regular expression (RegEx) approach (Figure 8A), to predict transcribed units as continuous stretches of directed transcribed states with a minimal length of 80 bp on both strands from the bdHMM and HMM annotation. Matching predicted transcript boundaries to previously published ones [4], 4186 (82%) of all annotated protein-coding transcripts were recovered from the bdHMM predictions, 11% more than the HMM predicts using the same criteria (3639 transcripts) (best reciprocal hits, Methods section 10). Moreover, the predicted transcription start sites (TSS) were consitently closer to the annotated ones (Figure 8D). In particular, 60% of the preditcted TSSs by the bdHMM were within 50 bp, whereas the best 60% of the HMM TSS predictions were within 100bp of the published ones. Accuracy of pA site prediction was lower, but comparable between bdHMM and HMM, where approximatley 60% of the predicted pA sites were within 100bp of the annotated ones for both methods. Moreover, 32 novel transcripts were predicted from the bdHMM annotation (4 overlapping a coding region, 28 non-coding, Figure 8C, Methods section 10), which is of particular significance because the *S. cerevisiae* transcriptome has been thoroughly studied and annotated.
As another illustration of genomic features that can be extracted from a bdHMM annotation, we searched for bidirectional promoters using a RegEx consisting of a promoter state flanked by an upstream transcript on the Crick strand and a downstream transcript on the Watson strand (Figure 8A,B). We detected 1,076 bidirectional promoters in yeast, which agrees well with a previous estimate of 1,049 bidirectional promoters [4]. Altogether, these results demonstrate the high accuracy of the bdHMM for genome annotation and its advance over the standard HMM.

14.4 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 (Figure 9B, Methods section 10). 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 (Figure 9A,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 (Figure 9A). Indeed we found that the enrichment or depletion of protein factors in each state was in accordance with their known roles in transcription (Figure 9A). 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 (Figure 9A,B). States related to the same transcription phase often peaked at successive genomic positions. For instance, the mid-elongation phase comprises successive states mE1-mE5 (Figure

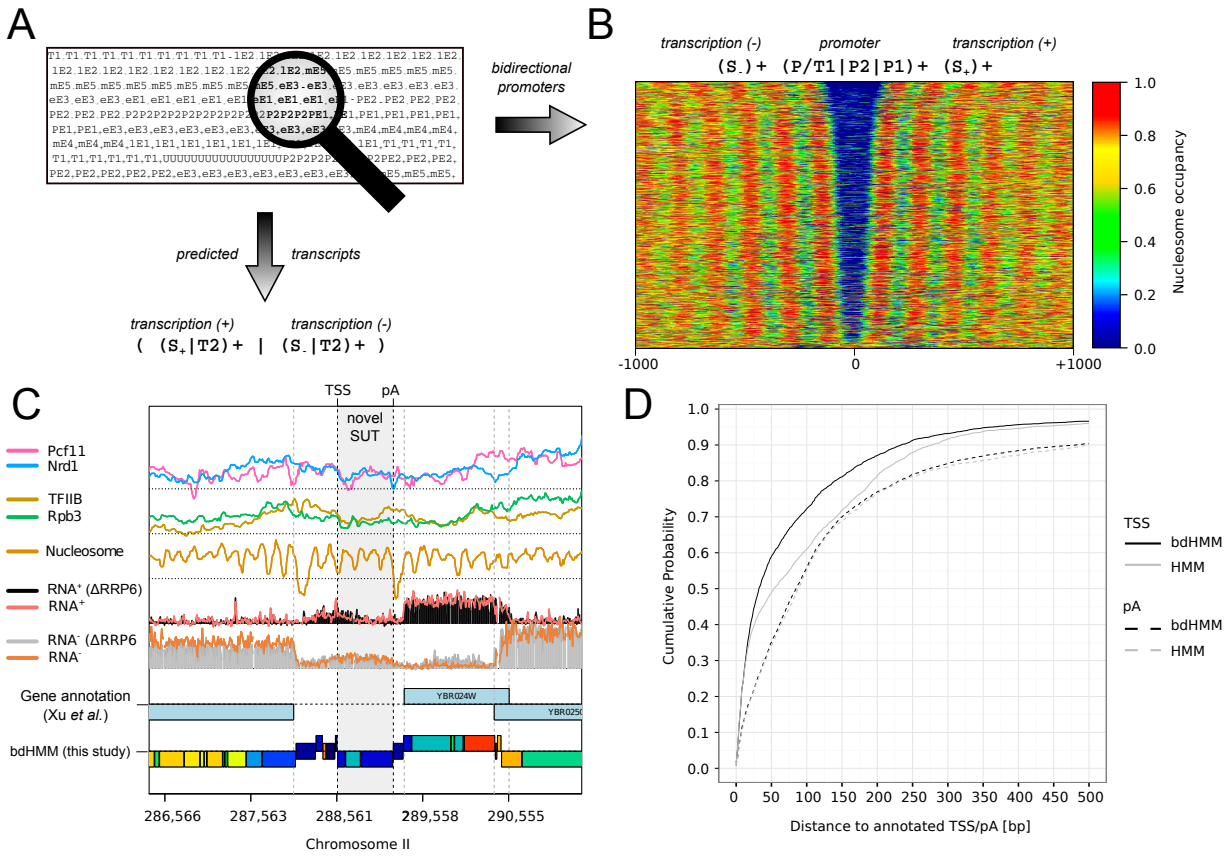


Figure 8: Genomic state annotation predicts bidirectional promoters and (novel) transcripts. (A) The genomic state annotation (viterbi path) was searched with Regular Expressions (RegEx) defining bidirectional promoters (right) and transcripts (bottom). (B) Nucleosome binding patterns centered at 1,076 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) A novel SUT (Stable unannotated transcript, a stable non-coding RNA, grey area) is identified on the - strand by the bdHMM. The locus shows detectable expression but was too low for the criterias used by [4] (D) Estimated cumulative probability of TSS and pA site predictions show higher accuracy of bdHMM in recovering TSSs. pA site prediction has similar accuracy for both models.

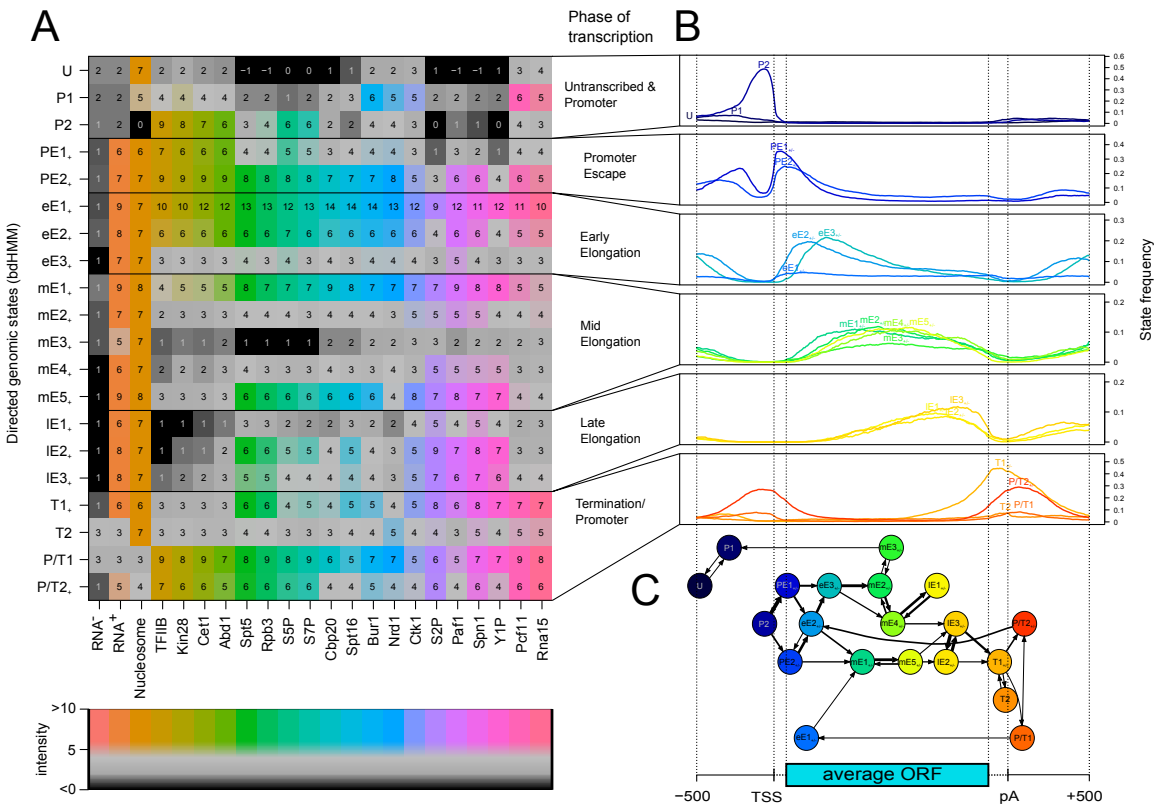


Figure 9: Roles of directed genomic states in the transcription cycle. (A) Mean ChIP enrichment of factors (horizontal axis) indicates the composition of the transcription machinery in each state (vertical axis). Factors were ordered by hierarchical clustering and states were ordered by position of their 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 sequences (viterbi paths) 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.

9B, C) that were characterized by a gradual decrease in the occupancy of initiation factors, capping-related factors, and Nrd1 (Figure 9A).
Overall, these results show that unsupervised bdHMM analysis can define meaningful genomic states that reflect phases of transcription at every single gene.

14.5 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 (Figure 9C, 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; Figure 9A,B). These alternative states are located within different branches of the flux diagram (Figure 9C). 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 (Figure 10A,B, Methods section 10). The obtained gene clusters show distinct patterns of protein occupancies suggesting mechanistic differences in transcription (Figure 10, Supplementary Figure 2 in [113] and below). Moreover, the gene clusters differed by gene length, expression level, and genomic context (e.g. termination overlaps with a neighboring downstream promoters or bidirectionality of promoters). Gene set enrichment analysis showed that clusters also corresponded to distinct functional gene groups (Supplementary Table 1 in [113]). The functional categories range from house-keeping (e.g. cluster 14, 38), cell cycle (e.g. cluster 17) to stress response (e.g. cluster 39). For instance, the high expression of cluster 38 and 14 is in accordance with their associated functions including ribosome biogenesis, positive regulation of transcription, translation or nucleosome assembly. More strikingly, we found the DNA binding motif of SFP1 - a regulator of ribosomal protein and ribosome biogenesis genes - to be enriched in promoter state P/T1 (which is a frequent promoter state of cluster 14 and 38 genes, Supplementary Figure 4 in [113]). In contrast, stress- and autophagy-related genes in cluster 39 show very low expression and protein binding (Supplementary Figure 2B in [113]). Altogether, this suggests that different transcription cycles as they are modeled by the bdHMM correspond to different co-regulated gene sets.

Cluster 14, which contains 694 genes (Figure 10B,C, Supplementary Figure 2B in [113]), shows a transcription cycle most similar to the canonical one proposed previously [33]. 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 (Figure 10D), 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 (Figure 9A).

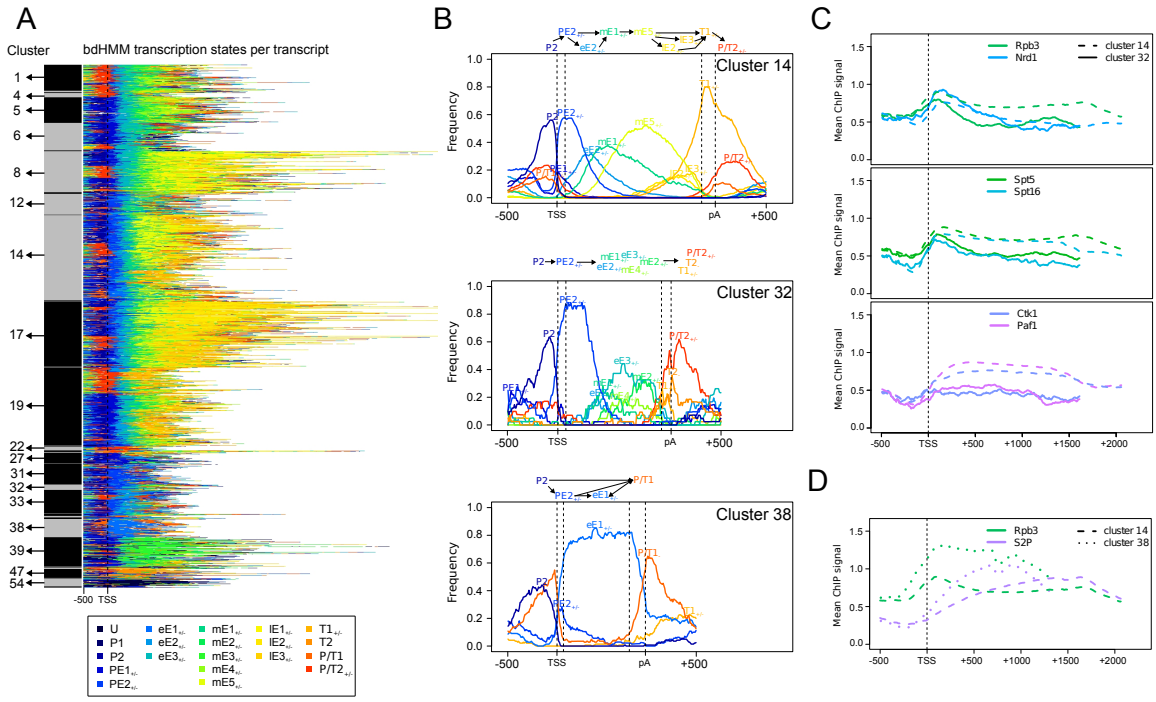


Figure 10: Clustering of state paths reveals gene-specific variations in the transcription cycle. (A) Genomic state sequences of 4,632 genes were clustered into 55 groups (left, only clusters containing at least 20 genes are labeled). Each line corresponds to the state sequence of a single gene. States 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 [33]. 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 Pol II (Rpb3) peak in the 5' region of genes, but Pol II reaches a stable, high level downstream of the TSS in cluster 38. 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.

14.6 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 (Figure 10B,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 (Figure 9B,C, Figure 10C). Moreover, elongation factors Ctk1 and Paf1 appear to be absent from those genes (Figure 10C, Supplementary Figure 2C in [113]). 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 (Figure 9B,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, since cluster 32 genes show significant upregulation after Nrd1 depletion from the nucleus [26] (Figure 10C, Supplementary Figure 3 in [113]). The individual occupancy profiles (Figure 10C, Supplementary Figure 2C in [113]) 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.

14.7 Evidence for distinct transcription mechanisms for highly expressed genes

Cluster 38 differs strikingly from the canonical transcription cycle during early elongation and termination (Figure 10B, Supplementary Figure 2D in [113], 147 genes enriched for genes involved in translation, Supplementary Table 1 in [113], Methods section 10). Cluster 38 is characterized by the high occupancy promoter state P/T1 (Figure 9A) 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 (Figure 10D). Moreover, Pol II does not exhibit the typical occupancy peak 150 bp downstream of the TSS but immediately reaches a stable high level (Figure 10D). This profile could be the consequence of a lower drop-off rate at this position [33], 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 (Supplementary Figure 2D in [113]). 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 [33] 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.

14.8 Not all termination regions are depleted of nucleosomes

Nucleosome depletion has been reported at the 3'-end of genes [188]. 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 Figure 2 in [113]). This supports previous reports that nucleosome depletion is not an intrinsic mark of transcription termination [189]. Thus, bdHMM analysis of the genomic context of transcription allows distinguishing canonical binding patterns from spurious ones caused by spill-over effects from neighboring genes.

14.9 Promoter and termination states are enriched in known and new DNA motifs

To detect putative functional DNA sequence elements associated with certain genomic states, we performed de novo motif discovery on the nucleotide sequences underlying the bdHMM state annotation using XXmotif [174]. In order to correct for local sequence properties like codon bias we chose as negative sequence sets (not containing any motifs) the sequences with a length of 150 bp and a distance of 50 bp upstream of each state. The use of negative control sets strongly improved the sensitivity during motif search (Methods section 10).
Promoter and termination states were enriched with sequence motifs (Figure 11). The state P/T1, which is specific for cluster 38 genes, shows enrichment of TF-binding motifs with specific functions such as ribosome biogenesis and general regulatory function as well as previously unknown sequence motifs, in particular the highly abundant TTTTTTTTG motif present in 76% of all P/T1 sequences (Figure 11). Termination state T1 contains motifs that are known to be involved in the 3'-end formation and pA positioning [190] and one novel motif (TTTTTTTTA). These motifs are located within the 3' UTR of genes, in accordance with the state frequency peak that we observe for state T1 (Figure 9B). The mixed state P/T2 also contains a motif associated with 3' end formation and previously unknown ones (Figure 11). For instance, it is enriched in the motif TTTTTTTTC, which is similar to the one we found in P/T1 (Figure 11). Finally, alternative states of the same phase of the transcription cycle were enriched for distinct motifs. For instance, the Abf1 binding motif and the Mbp1-Swi6 binding motif were specifically found in the promoter state P2 and not in the other promoter states. Together, this shows that bdHMM analysis enhances the identification of novel functional DNA elements.

14.10 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 [150]. The chromatin mark data had been binarized into presence/absence of each mark at a resolution of 200bp bins and analyzed with a standard HMM approach (ChromHMM) [151]. To handle the binarized chromatin marks data defined by [150], we extended bdHMM and included binary (Bernoulli) 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 (Methods

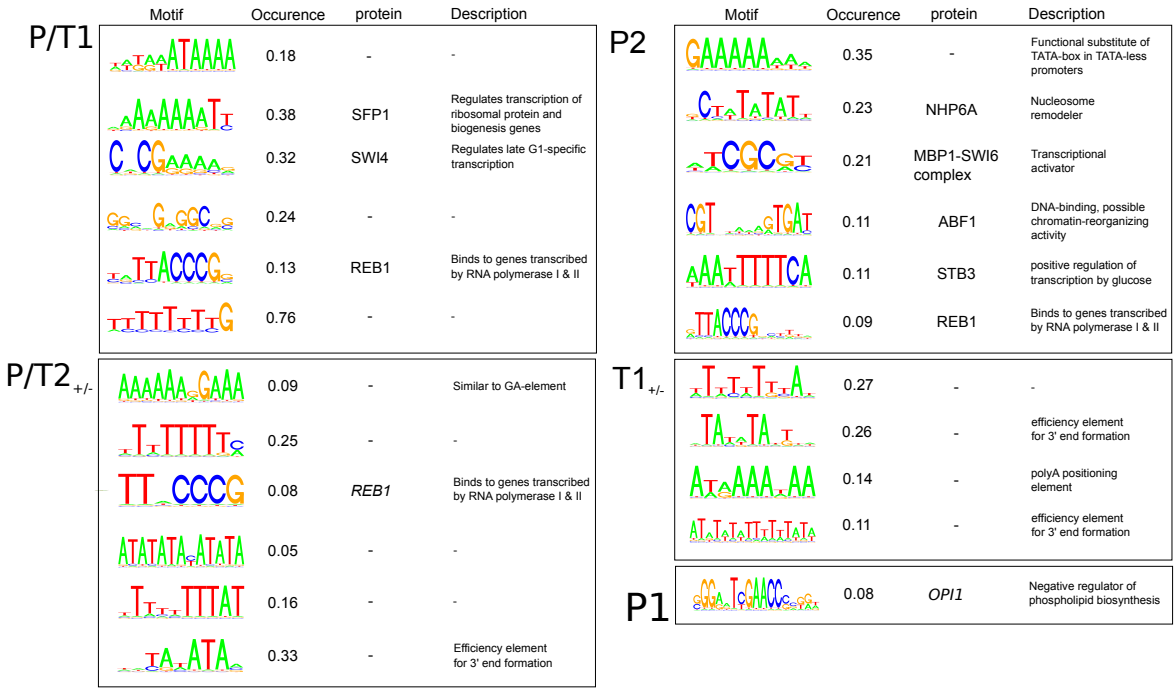


Figure 11: Promoter and termination states are enriched in DNA motifs. De novo motif search in the DNA sequences underlying the genomic state annotation discovered motifs in promoter states and termination states. A short functional description, known binders and the frequency in the sequence set, which was used in the analysis are shown. P2 is enriched in motifs of general transcriptional regulators and chromatin remodelers. T1 contains motifs which are known to be involved in the 3' end formation and polyA positioning. P/T1 is enriched with ribosomal, cell-cycle specific and general transcription factors. P/T2 contains motifs involved in transcription initiation and termination. P1 is enriched with a single motif, that highly resembles OPI1.

section 11) to decide that in the bdHMM, 35 out of a total of 51 ChromHMM states are modeled as directed state pairs and 16 ChromHMM states are modeled as undirected states. Consistently, we identified directed chromatin states around transcribed, but not at repressed or repetitive regions (Supplementary Figure 5 in [113]). Up to state directionality, 83% of state annotations agreed between the two methods (Methods section 11). Comparison of the ChromHMM with the bdHMM transitions revealed that in ChromHMM, transition probabilities between two states are similar in both directions (Figure 12C), whereas the bdHMM can resolve the true order of chromatin states (Figure 12A,B, Supplementary Figure 6 in [113]). For example, transitions from state 6 into states 2 and 3 are high for the forward direction, but low for the reverse drection. In contrast, transitions from states 2 and 3 into state 6 are high in reverse, but low in forward direction (Figure 12B). However, all of these transitions are high in the symmetric ChromHMM model (Figure 12C), demonstrating that bdHMM adds previously unexploited and valuable information to HMM-based analyses by uncoupling the underlying state directionality of genomic processes. Analysis of promoter and transcribed state frequencies at the TSS showed that state annotations matched the reading (sense) direction of the transcribed loci with up to 85% (Supplementary Figure 7 in [113]). 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.

14.11 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. It 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, and outperforms standard HMM in genomic feature annotation. 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 [150], 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 at a higher accuracy than standard HMM. 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

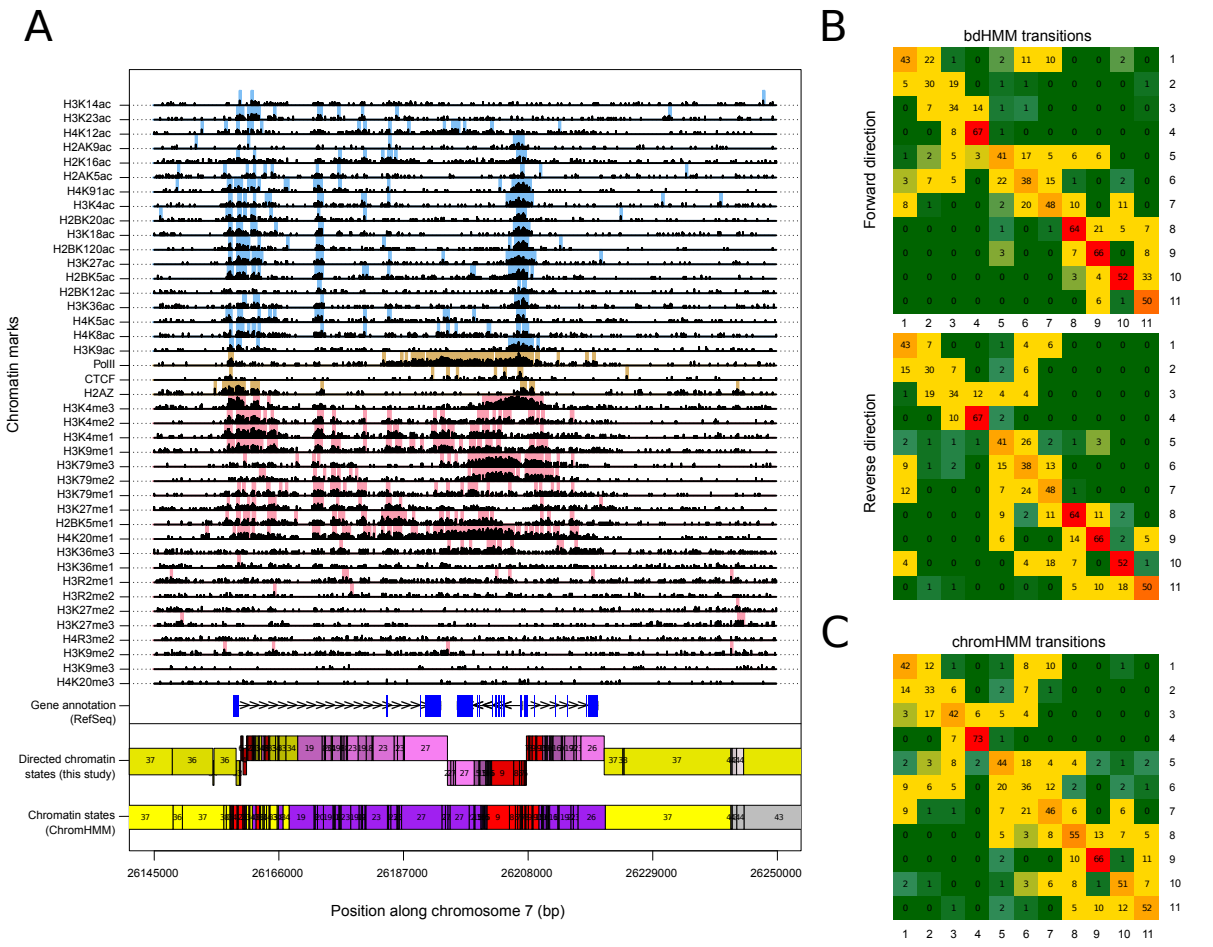


Figure 12: Application of bdHMM to chromatin modifications in human T-cells identifies direction of chromatin states. (A) Example of chromatin state annotation of ChromHMM and bdHMM (bottom tracks) with RefSeq gene annotation and input signal. State direction matches gene orientation of annotated convergent genes and divergent genes. The log-transformed signal [150] 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) bdHMM transitions between promoter-associated states 1-11 are shown for forward and reverse states. While the asymetric, transposed structure of these two submatrices (i.e. transition probabilities favor one direction for pairs a_{ij} anda_{ji}) uncouple the two reading directions, the symmetric ChromHMM transition matrix (C) hides the underlying directed flow of chromatin states.

that occurs at all genes [33, 78, 184]. 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.
Thus, we foresee bdHMMs to be instrumental for studying gene transcription and other directed genomic processes, such as DNA replication, recombination or DNA repair.

15 Accurate promoter and enhancer identification in 127 ENCODE and Roadmap Epigenomics cell types and tissues by GenoSTAN

The results presented in this section are part of the manuscript "Acurate promoter and enhancer identification in 127 ENCODE and Roadmap Epigenomics cell types and tissues by GenoSTAN", which was submitted for publication. For detailed author contributions see page ix.

Transcription is tightly regulated by cis-regulatory DNA elements known as promoters and enhancers. These elements control development, cell fate and may lead to disease if impaired. A promoter is functionally defined as a region that regulates transcription of a gene, located upstream and in close proximity to the transcription start sites (TSSs) [191]. In contrast, an enhancer was originally functionally defined as a DNA element that can increase expression of a gene over a long distance in an orientation-independent fashion relative to the gene [10]. The functional definition of enhancers and promoters leads to practical difficulties for their genome-wide identification because the direct measurement of the regulatory activity of genomic regions is hard, with current approaches leading to contradicting results [192, 193, 194].

Since the direct measurement of cis-regulatory activity is challenging, a biochemical characterization of the chromatin at these elements based on histone modifications, DNA accessibility, and transcription factor binding has been proposed [14, 101, 150, 195, 196]. This approach leverages extensive genome-wide datasets of chromatin-immunoprecipitation followed by sequencing (ChIP-Seq) of transcription factors (TFs), histone modifications, or Cap analysis gene expression (CAGE) that have been generated by collaborative projects such as ENCODE [93, 197], NIH Roadmap Epigenomics [110], BLUEPRINT [198] and FANTOM [199, 200].

In this context, the computational approaches employed to classify genomic regions as enhancers or promoters play a decisive role [195, 196]. As the experimental data are heterogeneous, we generally refer to them as tracks. Several studies used supervised learning techniques to predict enhancers based on tracks such as histone modifications or P300 binding (e.g. [201, 202, 203, 204]). However, a training set of validated enhancers is needed in this case, which is hard to define since only few enhancers have been validated experimentally so far and these might be biased towards specific enhancer subclasses. Alternatively, unsupervised learning algorithms were developed to identify promoters and enhancers from combinations of histone marks and protein-DNA interactions alone [93, 101, 102, 103, 110, 150, 155, 205]. These unsupervised methods perform genome segmentation, i.e. they model the genome as a succession of segments in different chromatin states defined by characteristic combinations of histone marks and protein-DNA interactions found recurrently throughout the genome. All popular genome segmentations are based on hidden Markov models [112]. However, these methods differ in the way the distribution of ChIP-seq signals for each chromatin state is modeled. ChromHMM [103, 150, 151], one of the two methods applied by the ENCODE consortium, requires binarized ChIP-seq signals that are then modeled with independent Bernoulli distributions. Consequently, the performance of ChromHMM highly depends on the non-trivial choice of a proper binarization cutoff. Moreover, quantitative information is lost with this approach, which is especially important for distinguishing promoters from enhancers since these elements are both marked with H3K4me1 and H3K4me3, but at different ratios [62]. Segway [101, 102], the other method applied by the ENCODE consortium, uses independent Gaussian distributions of log-transformed and smoothed ChIP-seq signal. Although Segway preserves some quantitative information, the transformation of the original count data

leads to variance estimation difficulties for very low counts. Therefore, Segway further makes the strong assumption that all tracks have the same variance. Recently, EpicSeg [154] used a negative multinomial distribution to directly model the read counts without the need for data transformations. However, similar to the variance model of Segway, the EpicSeg model leads to a common dispersion (the parameter adjusting the variance of the negative multinomial) for all tracks. Moreover, EpicSeg does not provide a way to correct for sequencing depth, which makes the application to data sets with multiple cell types with varying library sizes difficult. Also, EpicSeg has been applied only to three cell types so far [154]. These methods not only differ in their modeling assumptions but also lead to very different results. In the K562 cell line for instance, ChromHMM identified 22,323 enhancers [93], Segway 38,922 enhancers [93], and EpicSeg 53,982 enhancers [154]. Altogether, improved methods and detailed benchmarkings are required for a reliable annotation of transcriptional cis-regulatory elements.

Here we propose a new unsupervised genome segmentation algorithm, GenoSTAN (***Geno***mic ***St***ate ***An***notation from sequencing experiments), which overcomes limitations of current state-of-the-art models. GenoSTAN learns chromatin states directly from sequencing data without the need of data transformation, while still having track-specific variance models. We applied GenoSTAN to a total of 127 cell types and tissues covering 16 datasets of ENCODE and all 111 datasets of the Roadmap Epigenomics project as well as another ENCODE ChIP-seq dataset for the K562 cell line. GenoSTAN consistently performed better when benchmarked against Segway, ChromHMM and EpicSeg segmentations using independent evidence for activity of promoter and enhancer regions. Co-binding analysis of TFs reveals that promoters and enhancers both shared the Polymerase II core transcription machinery and general TFs, but they are bound by distinct TF regulatory modules and differ in many biophysical properties. Moreover, GenoSTAN enhancer and promoter annotations had a higher enrichment for complex trait-associated genetic variants than previous annotations, demonstrating the advantage of GenoSTAN and our chromatin state map to understand genotype-phenotype relationships and genetic disease.

15.1 Modeling of sequencing data with Poisson-lognormal and negative binomial distributions

We developed a new genomic segmentation algorithm, GenoSTAN, which implements hidden Markov models with more flexible multivariate count distributions than previously proposed. GenoSTAN supports two multivariate discrete emission functions, the Poisson-lognormal distribution and the negative binomial distribution. For the sake of reducing running time, the components of these multivariate distributions are assumed to be independent. However, the variance is modeled separately for each state and each track, which provides a more realistic variance model than current approaches. To be applicable to data sets with replicate experiments or multiple cell types, GenoSTAN corrects for different library sizes of experiments (Methods section 8). All parameters are learnt directly from the data, leaving the number of chromatin states as the only parameter to be manually set. We provide an efficient implementation of the Baum-Welch algorithm for inference of model parameters, which can be run in a parallelized fashion using multiple cores. The method is implemented as part of our prevsiouy published R/Bioconductor package STAN [113], which is freely available from http://bioconductor.org/. Altogether, GenoSTAN uniquely combines flexible count distributions, short running times, and minimal number of manually entered parameters (Figure 13A).
We performed an extensive benchmarking of GenoSTAN against alternative methods (Figure

A

	GenoSTAN	ChromHMM	Segway	EpicSeg
Count distribution	yes	no	no	yes
Library size correction	yes	no	no	no
Track-specific variance	yes	no	no	no
Running time (one cell line)	minutes-hours	minutes-hours	hours-days	minutes-hours
Manually entered parameters	Number of states	Number of states, binarization cutoff	Number of states, data transformation and smoothing	Number of states

B

Benchmark	Cell/tissue types	GenoSTAN model (ID)	Chromatin marks	Benchmarked models (ID)
I	K562 (ENCODE)	Poilog-K562 nb-K562	H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K9ac, H3K27ac, H3K27me3, H4K20me1, P300, DNase-Seq	ChromHMM-ENCODE ChromHMM-Nature Segway-ENCODE Segway-nmeth Segway-Reg.Build EpicSeg
II	127 cell/tissue types (ENCODE, Roadmap Epigenomics)	Poilog-127 nb-127	H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3	ChromHMM-15 ChromHMM-25
III	20 cell/tissue types (ENCODE, Roadmap Epigenomics)	Poilog-20 nb-20	H3K4me1, H3K4me3, H3K36me3, H3K9ac, H3K27ac, H3K27me3, H3K9me3, DNase-Seq	ChromHMM-15 ChromHMM-18 ChromHMM-25

Figure 13: Overview of chromatin state annotation methods and study design. (A) Comparison of features of GenoSTAN against three previous chromatin state annotation algorithms. (B) Description of the three benchmark sets used in this study. GenoSTAN is benchmarked against published chromatin state annotations using ChromHMM ('ChromHMM-ENCODE' [93, 102], 'ChromHMM-Nature' [103], 'ChromHMM-15', '-18' and '-25' [110]), Segway ('Segway-ENCODE' [93, 102], 'Segway-nmeth' [101] and 'Segway-Reg.Build' [155]) and EpicSeg [154].

13B). Benchmark I compares GenoSTAN and the three alternative methods for the K562 cell line. K562 is a major model system to study human transcription and the ENCODE cell line with the largest number of experiments [93]. Moreover, all three other methods (ChromHMM, Segway, and EpicSeg) had been run by their own authors for K562, so that the algorithm parameters for these annotations can be assumed to have been set at best expert knowledge. Benchmark II compares methods for all 127 cell types and tissues provided by ENCODE and Roadmap Epigenomics. This is the largest benchmark dataset of all three we considered but also the one with the least number of common tracks (5 chromatin marks: H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K9me3, Figure 13B). Benchmark III compares results for a subset of 20 ENCODE and Roadmap Epigenomics cell types and tissues which had moreover H3K27ac, H3K9ac ChIP-seq and DNAse I hypersensitivity data (DNAse-Seq). This distinction allowed to provide more accurate annotations for the better characterized cell types and tissues. For benchmark II and benchmark III only annotations for the method ChomHMM were available to compare against GenoSTAN annotation. Figure 13B lists all model names and studies that were considered.

15.2 Benchmark I: Improved chromatin state annotation in human K562 cells

We first fitted two GenoSTAN models, one with Poisson-lognormal emissions (henceforth referred to as GenoSTAN-Poilog-K562 model) and one with negative binomial emissions (GenoSTAN-nb-K562 model) to a dataset of ChIP-seq data of 9 histone modifications, of the histone acetyltransferase P300, and DNA accessibility (by DNase-Seq) data for the K562 cell line at 200 bp binning resolution (Methods section 12). As pointed out by others [101, 150], there is no purely statistical criterion for choosing the number of states from the data of practical usage in such a setting. In practice, the number of states is manually defined by trading off goodness of fit against interpretability of the model [101, 113, 150]. For GenoSTAN-Poilog-K562, we used 18 chromatin states. For GenoSTAN-nb-K562, we used 23 states, since lower state numbers did not provide enough resolution to give a fine-grained map of chromatin states on this data set (Methods section 12). Figure 14A compares the two GenoSTAN segmentations to segmentations from other studies using ChromHMM, Segway and EpicSeg on a region containing the TAL1 gene together with three known enhancers [16, 93, 102, 154].

15.2.1 Chromatin states recover biologically meaningful features

In order to assign biologically meaningful labels to each state of the Poilog-K562 and of the nb-K562 GenoSTAN models, we investigated their read coverage distributions and overlapped the occurrence of a state in the genome with known genomic features. In line with previous studies, this led to the definition of promoter, enhancer, repressed, actively transcribed and low coverage states [102, 103, 154]. The median read coverage in state segments and genomic distributions were very similar for both the Poilog-K562 and the nb-K562 models (Figure 14B, Appendix Figure A1). Promoter states were characterized by a low (< 1) H3K4me1/H3K4me3 ratio, in contrast to enhancer states which showed a high ratio (> 1). Further, P300 levels were roughly two-fold higher in the enhancer state, which is in accordance with previous observations [62, 206, 207]. Promoter (Prom) states were located close to annotated GENCODE TSSs [104], with a median distance of 220 bp for GenoSTAN-Poilog-K562 and 400 bp for GenoSTAN-nb-K562 model. Enhancer states (Enh) on the other hand were located further away from TSSs, with a median distance of 3.6 kb (Poilog-K562) (respectively 5.8 kb for nb-K562, Figure 14B, Appendix Figure A1). Promoter and enhancer states also differed in their DNA sequence features.

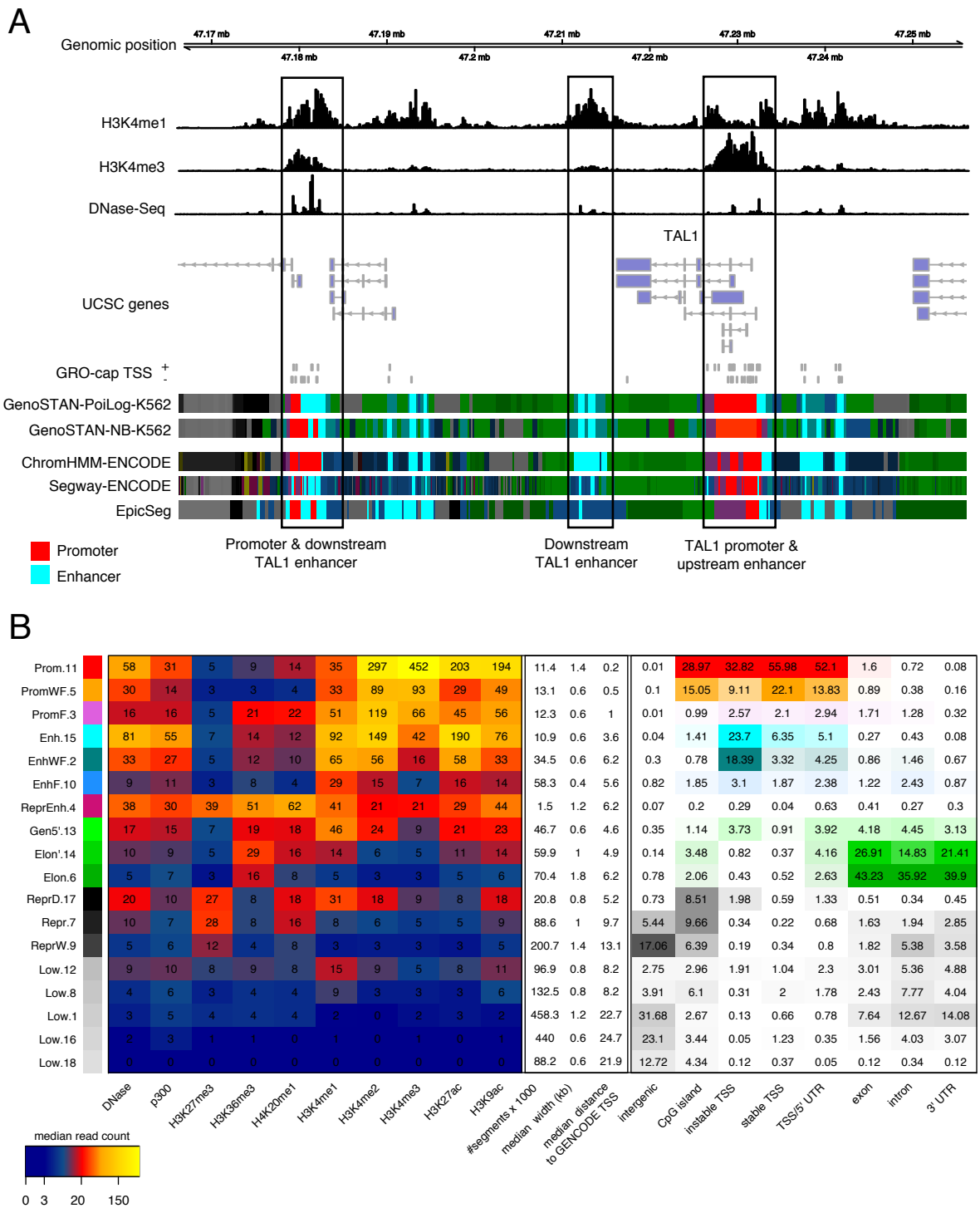


Figure 14: Chromatin states fitted on Benchmark I using GenoSTAN. (A) GenoSTAN segmentations are shown with published segmentations using ChromHMM-ENCODE [93], Segway-ENCODE [93] and EpicSeg [154] at the TAL1 gene and three known enhancers. GenoSTAN-Poilog-K562 correctly recalls all known promoter and enhancer regions. GenoSTAN-nb-K562 misses the upstream enhancer. ChromHMM-ENCODE misclassifies most of the downstream enhancer region as promoter. (B) Median read coverage of GenoSTAN-Poilog-K562 chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSS (middle). The right panel shows recall of genomic regions by chromatin states.

45% of CpG islands were located within promoter states (strong, weak and promoter flanking states) in both models, but only 3% in enhancer states (strong, weak, enhancer flanking states, Figure 14B, Appendix Figure A1). While promoter states mostly recovered stable TSSs, enhancer states were located at unstable TSSs (GRO-cap TSSs that are not recovered by the GENCODE annotation), which supports previous findings [13]. Furthermore, both models (GenoSTAN-nb-K562 and GenoSTAN-Poilog-K562) contained 3 states, that we classified as "actively transcribed states", which were characterized by high values of H3K36me3 and overlap with UTRs, introns and exons. Two out of three "transcribed" states were also enriched in promoter associated marks (H3K4me1-3, H3K27ac, H3K9ac) and H4K20me1 and thus represented 5' transitions in transcription. Moreover, both models fitted four repressed states showing high read coverage of H3K27me3. Two of these states also exhibited high DNase-Seq and promoter/enhancer associated histone modification signals, suggesting that these states might reflect repressed regulatory regions (ReprEnh, ReprD). These elements were distal to annotated GENCODE TSSs (median distance: 5.2-11.8 kb). ReprEnh states were also enriched in P300 and recovered 0.2% of CpG islands, while ReprD states had lower P300 levels and recovered 8-9% of CpG islands in the genome (Figure 14B, Appendix Figure A1). The remaining states exhibited low coverage in chromatin marks and therefore were labeled as "low" states. Altogether, GenoSTAN accurately recovered many features of known chromatin states and provided a high resolution map of these in K562.

15.2.2 High variation of enhancer predictions between chromatin state annotations of different studies

To assess the consistency of promoter and enhancer predictions across studies, we compared the GenoSTAN segmentations to other published segmentations in K562 by ChromHMM ('ChromHMM-ENCODE' [93,102] and 'ChromHMM-Nature' [103]), Segway ('Segway-ENCODE' [93,102], 'Segway-nmeth' [101] and 'Segway-Reg.Build' [155]) and EpicSeg [154]. We computed pairwise Jaccard indices (the ratio of the number of common elements over all elements predicted by two methods) of promoter and enhancer states to quantify the agreeement between the predictions of the different studies (Appendix Figure A2). Promoter state annotations generally agreed well (median Jaccard-Index: 0.78). However, enhancer prediction varied more (median Jaccard Index: 0.48), suggesting that enhancers are more difficult to annotate. This variation of enhancer calls was also reflected in the different numbers of annotated enhancer segments, which had been shown to vary greatly between different prediction methods [201]. The number of enhancer segments ranged from 10,932 segments in GenoSTAN-Poilog-K562 to 80,043 segments in one Segway annotation [101] (Appendix Table A1). Therefore, a thorough assessment of these predictions was necessary to provide a robust and accurate prediction of these elements.

15.2.3 Comparison of GenoSTAN with published chromatin state annotations

In order to benchmark the different segmentations, we used independent data including evidence of transcriptional activity (GRO-cap TSSs [13]), of transcription factor binding (ENCODE high occupancy target, or HOT regions [197], and ENCODE TF binding sites [93]), and of cis-regulatory activity (enhancer activity assessed by reporter assays [193]), which are all expected to be characteristics of promoters and enhancers. Transcription initiation activity is not only the hallmark of promoters, but also of enhancers [3, 13, 199, 200]. To benchmark the predictions using evidence for transcription, we used published data from a protocol called GRO-cap [13], a nuclear run-on protocol, which very sensitively maps transcription start sites genome-wide. To

this end, we sorted for each method chromatin states by their overlap with GRO-cap TSSs by decreasing precision. Starting with the most precise state (i.e. highest overlap with TSSs) we calculated cumulative recall and false discovery rate (FDR) by subsequently adding states with decreasing precision (Figure 15A). GenoSTAN-Poilog-K562 had the highest recall and the lowest FDR (Methods section 12, Figure 15A). GenoSTAN-nb-K562 performed similar to other segmentations (Segway-Reg. Build, ChromHMM-ENCODE). In particular, 94% of GenoSTAN-Poilog-K562 promoters (Prom.11) and 81% of its enhancer regions (Enh.15) overlapped with GRO-cap TSSs. This compares to 85% (Prom.16) and 65% (Enh.6) of GenoSTAN-nb-K562 and 89% (Tss) and 52% (Enh) of ChromHMM-ENCODE promoter and enhancer regions. Interestingly, the two ChromHMM segmentations (ChromHMM-ENCODE [93, 102], ChromHMM-Nature [103]) had very different accuracies for TSSs, which might be due to different data sets or cutoffs used for ChIP-seq binarization in the two studies. In contrast, the overall accuracy of the Segway annotations was comparable across studies. This comparison shows that GenoSTAN chromatin state annotation identifies putative promoters and enhancers which show transcriptional activity more frequently than previous annotations.

GRO-cap is a very sensitive method that captures also a large amount of TSSs of unstable transcripts [13]. However it is limited to capped RNA species, misses RNAs below the detection threshold and cannot be used to validate repressed (i.e. transcriptionally inactive) regulatory elements. To address these shortcomings we used two additional independent features, TF binding and HOT regions. The binding of TFs to a region of DNA is a pre-requisite for potential regulatory function and transcriptional activity. High Occupancy of Target (HOT) regions are genomic regions which are bound by a large number of different transcription-related factors [197], which were shown to function as enhancers [208] and are enriched in disease- and trait-associated genetic variants [209]. As for the benchmark with TSSs, we sorted chromatin states by overlap with HOT regions by decreasing precision and calculated cumulative recall and FDR (Figure 15B). The best performing segmentations for HOT regions were GenoSTAN-Poilog-K562 and GenoSTAN-nb-K562, followed by ChromHMM-ENCODE. The ordering of states with HOT regions was indeed different from the GRO-cap TSSs benchmark. Additionally to GenoSTAN promoter and enhancer states, the repressed enhancer state frequently overlapped with HOT regions with an overall precision of 81% (GenoSTAN-Poilog-K562) and 77% (GenoSTAN-nb-K562). In comparison, the top three ChromHMM-ENCODE states had together a precision of 67%. All other segmentation methods showed a lower precision and recall for HOT regions. This was also reflected in the frequency of individual TF binding sites at enhancer regions, which were generally higher in GenoSTAN enhancer states than in other segmentations (Figure 15C). In particular, only a very small fraction of EpicSeg and Segway-nmeth enhancers were found to be bound by TFs. EpicSeg and Segway-nmeth segmentations were also those with the highest number of predicted enhancers, suggesting that many of these predictions are spurious.

Next, we calculated the recall of FANTOM5 promoters [200] and enhancers [199] to assess how well the models distinguish promoters from enhancers, as it was evident from inspection of specific examples that this distinction was difficult to be established by current methods (Figure 15D). The FANTOM5 consortium have performed extensive mapping of capped transcripts 5' ends using CAGE and defined enhancers and promoters based on transcriptional activity pattern. FANTOM5 enhancers were defined as regions showing balanced bidirectional capped transcripts, a hallmark of enhancer RNAs [199], whereas FANTOM5 promoters were defined as regions where transcription was biased towards one direction. The FANTOM5 annotation of enhancers and

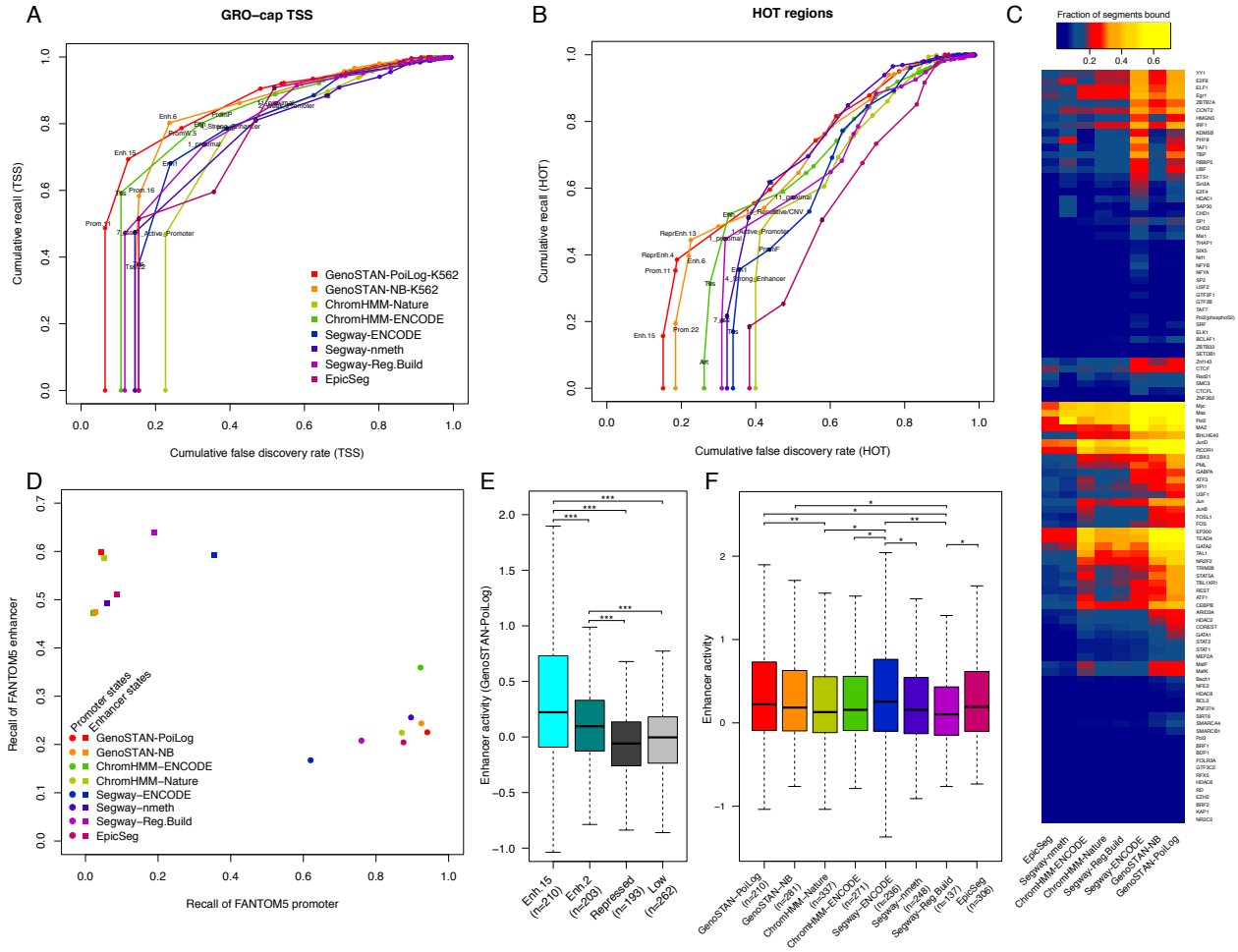


Figure 15: Comparison of GenoSTAN to other published segmentations on benchmark set I. (A) Performance of chromatin states in recovering GRO-cap transcription start sites. Cumulative FDR and recall are calculated by subsequently adding states (in order of increasing FDR). (B) The same as in (A) for ENCODE HOT regions. (C) The fraction of predicted enhancer segments bound by individual TFs is shown for different studies. GenoSTAN enhancers are more frequently bound by TFs than those from other studies. (D) Recall of FANTOM5 promoters and enhancers which are active in K562 (i.e. overlapping with a GRO-cap TSS and an ENCODE DNase hypersnesitivity site) by predicted promoters and enhancers is plotted to assess how well models distinguish promoters from enhancers. (E) Predicted enhancers show significantly higher activity than repressed and low coverage regions as measured by a reporter assay ('*', '**' and '***' indicate p-values <0.05, 0,01 and 0.001). (F) Comparison of experimetal measures of enhancer activity between different studies.

promoters could not entirely replace a chromatin state based approach because (i) the use of expression data in FANTOM5 limits the identified regulatory regions to transcriptionally active elements and (ii) CAGE was shown to be not as sensitive to rapidly degraded transcripts as GRO-cap and therefore might miss regulatory enhancers with unstable transcripts [13]. Nonetheless, FANTOM5 provides an annotation of enhancers and promoters based on independent data that is well suited to assess how well the models distinguish promoters from enhancers. We filtered the FANTOM5 annotation to promoters and enhancers for activity in K562 by overlapping them with DHS [93] and GRO-cap TSSs [13]. We considered that a promoter state performed well, when the recall of FANTOM5 promoters was high and the recall of FANTOM5 enhancers was low and vice versa for enhancer states. GenoSTAN-Poilog-K562 and ChromHMM-nature enhancer states recall most FANTOM5 enhancers (60%, Figure 15D, Appendix Table A2). For enhancer states, the recall of FANTOM5 promoters was around 10% except for those of Segway-ENCODE, which recalls almost 35% of FANTOM5 promoters and EpicSeg which recalls 21% of FANTOM5 promoters. In accordance with this, many promoter regions were erroneously classified as enhancer regions in this segmentation (e.g. TAL1 promoter in Figure 14A). The recall of FANTOM5 enhancers by promoter states was generally higher (17% - 37%). GenoSTAN-Poilog-K562 and -nb-K562 recalled more than 90% of FANTOM5 promoters and around 20% of FANTOM5 enhancers which is comparable to other studies (Segway-nmeth, ChromHMM-nature, EpicSeg). ChromHMM-ENCODE promoter states had a comparable recall of FANTOM5 promoters (92%), but higher recall of FANTOM5 enhancers (37%) (Figure 15D). This strong overlap of ChromHMM-ENCODE promoters with FANTOM5-labeled enhancers is in accordance with our observation that some enhancer regions were errenuously classified as promoters in ChromHMM-ENCODE (Figure 14A). These results show that GenoSTAN segmentations distinguish promoters from enhancers at similar or better accuracy than other segmentations.

So far we only used indirect evidence (TSSs, HOT regions, TF binding, FANTOM5 enhancer) to draw conclusions about the cis-regulatory activity of a candidate enhancer. As additional and direct evidence for the cis-regulatory activity of enhancer regions inferred by GenoSTAN, we overlapped our enhancers to genomic sequences that were previously tested for cis-regulatory activity in a reporter assay, where candidate elements had been cloned into a plasmid upstream of the promoter of a reporter gene [193]. Enhancers from GenoSTAN segmentations showed significantly higher activity than repressed or low coverage regions (GenoSTAN-Poilog-K562 & GenoSTAN-nb-K562: p-value < 0.001 wilcoxon-test, Figure 15E). Interestingly, repressed regions (marked by H3K27me3) showed lower activity than low coverage regions. Moreover, GenoSTAN-Poilog-K562 enhancers showed significantly higher enhancer activity than those of three other studies (Figure 15F), including the original study (p-value < 0.01, ChromHMM-nature enhancers) by Kheradpour et al. [193]. This analyis shows that GenoSTAN has higher succes rate in predicting *in vivo* enhancer activity than previous methods.

15.2.4 Comparison of the GenoSTAN, ChromHMM, Segway and EpicSeg algorithms on a common dataset

The K562 genome segmentations of ChromHMM, Segway and EpicSeg used so far were derived from different combinations of data tracks. To verify that the favorable performance of GenoSTAN is mainly due to an improved modeling and not due to different data, we also ran ChromHMM, Segway and EpicSeg on the same data as GenoSTAN-Poilog-K562 and GenoSTAN-nb-K562 (Methods section 12). Both GenoSTAN-Poilog-K562 and GenoSTAN-nb-K562 had a

lower FDR at a similar or higher recall than all three other methods (Appendix Figure A3). Moreover, we found that changing the binarization for ChromHMM dramatically affected its outcome. Without further manual processing of the data, ChromHMM fitted only one transcriptionally active state, which modeled both promoters and enhancers, regardless of state number (Appendix Figure A3). We suspected that the high read coverage in the H3K4me1 and H3K4me3 signal tracks made promoters and enhancers indistinguishable after binarization (H3K4me1 and H3K4me3 were called present at both, promoters and enhancers, and they were both called absent elsewhere). When all data tracks were subsampled to the same (and lower) library size, this problem was solved and ChromHMM fitted multiple transcriptionally active states thereby distinguishing promoters from enhancers and, at the same time, increased in accuracy (Appendix Figure A3). The same problem occurred for Segway, but changing Segway's parameters did not help distinguish different transcriptionally active chromatin states.
To make sure that these results did not depend on the arbitrary choice of the number of states, we ran each method using 10 to 30 states (Methods section 12) and calculated the precision of each state S for recalling HOT (respectively TSS) regions as the fraction of all segments annotated with S that overlapped with a HOT (respectively TSS) region. For each number of states and each segmentation algorithm, we determined the state with highest precision (Appendix Figure A4A, B). Independently of the number of states, GenoSTAN-Poilog and GenoSTAN-nb consistently performed best. Even at low state numbers, precision remained constantly high, while it decreased considerably for other methods. We also derived an area under curve (AUC) score for each model, to assess the spatial accuracy in calling TSSs or HOT regions (Appendix Figure A4C, D and Methods section 12). Again, AUC scores were consistently highest for the GenoSTAN segmentations.
Altogether this extensive benchmark in the K562 cell line demonstrates that GenoSTAN-Poilog and to a slightly lesser extent GenoSTAN-nb, outperforms current chromatin state annotation algorithms for identifying enhancers and promoters.

15.3 Benchmark II and III: Chromatin state annotation for ENCODE and Roadmap Epigenomics cell types and tissues

We next applied GenoSTAN to 127 cell types and tissues from ENCODE and Roadmap Epigenomics, the largest compendium of chromatin-related data, using genomic input and the five chromatin marks H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K9me3 that have been profiled across the whole compendium [110] (Appendix Figure A5, Benchmark II, Figure 13B for data tracks and model names). Moreover, we performed a dedicated analysis to 20 of these cell types and tissues which had three further important data tracks: H3K27ac, H3K9ac and DNase-Seq (Appendix Figure A6, Benchmark III, Figure 13B for data tracks and model names). These further three tracks are important features of active promoters and enhancers, which can lead to more precisely mapped enhancer boundaries [93]. We performed similar comparisons as decribed above to the three available segmentations from the Roadmap Epigenomics project with 15, 18 and 25 states (ChromHMM-15, -18, and -25) [110,152]. All methods were less performant than in Benchmark I, possibly due to lower read coverage or to less rich data. Nonetheless, the GenoSTAN annotations consistently outperformed the existing ones. Specifically, this held when assessing the recovery of FANTOM5 CAGE tags (Figure 16A, assessed for all 127 cell types and tissues), of GRO-cap TSSs (Figure 16B assessed for the cell types with available GRO-Cap TSSs), and of HOT regions (Figure 16C, assessed for the cell types with available HOT regions). Moreover, both GenoSTAN models distinguished better promoters from enhancers than pre-

vious annotations (Figure 16D, Appendix Table A2). The low accuracy of ChromHMM-15 and ChromHMM-18 promoters might be caused by frequent state switching between the promoter and promoter flanking state (Appendix Figure A7). Consequently, the number of promoter regions in K562 was up to 30% higher in the ChromHMM-15 and -18 segmentations than in the GenoSTAN or ChromHMM-25 segmentations (Appendix Table A1). The number of predicted enhancers (in K562) also differed greatly. The ChromHMM-15 and -18 state models predict 92,824 (7_Enh) and 22,678 (9_EnhA1) enhancers, while ChromHMM-25 predicts 12,706 and GenoSTAN-Poilog-20 and -127 predict 15,655 and 45,955 enhancers (Appendix Table A1). Although GenoSTAN predicted more enhancers than the ChromHMM-25 model, the fraction of putative enhancers bound by individual TFs was greater (Figure 16E). For instance 46% (25%) of enhancers were bound by Pol II in the GenoSTAN-Poilog-20 (-127) model, compared to 8%, 18% and 36% in the ChromHMM 15, 18 and 25 state models. Also, the lineage-specific enhancer-binding transcription factor TAL1 binds at 37% (GenoSTAN-Poilog-20) and 27% (GenoSTAN-Poilog-127) of predicted enhancers. Conversely, 13%, 16% and 27% of putative enhancers were bound by TAL1 in the respective 15, 18 and 25 state ChromHMM models (Figure 16E). Collectively, these results show that the improved performance of GenoSTAN is not restricted to the K562 dataset.

15.4 Cell-type specific enrichment of disease- and other complex trait-associated genetic variants at promoters and enhancers

Previous studies showed that disease-associated genetic variants are enriched in potential regulatory regions [103, 110, 210, 211, 212, 213] demonstrating the need for accurate maps of these elements to understand genotype-phenotype relationships and genetic disease. To study the potential impact of variants in regulatory regions on various traits and diseases, we overlapped our enhancer and promoter annotations from 127 cell types and tissues (Benchmark II, Figure 13B) with phenotype-associated genetic variants from the NHGRI genome-wide assocation studies catalog (NHGRI GWAS Catalog [181]). First, we intersected trait-associated variants with enhancer and promoter states (GenoSTAN-Poilog-127). Overall, 37% of all trait-associated SNPs were located in potential enhancers and 7% in potential promoters. The number of traits significantly enriched (at FDR <0.05) with enhancers or promoters in at least one cell type or tissue was larger for GenoSTAN-Poilog-127 (69 traits for GenoSTAN-Poilog-127 for enhancers and 20 traits for promoters, Methods section 12) than for the best performing ChromHMM-model (ChromHMM-15, 64 traits for enhancers and 18 traits for promoters). The better performance of GenoSTAN-Poilog-127 was found at all FDR cutoffs (Appendix Figure A8). To control for the fact that methods can differ among each other regarding the length of the promoters and enhancers they predict, we furthermore computed the recalls of GWAS variants for a fixed genomic coverage. Restricting to a total genomic coverage of 2% (random subsetting, also allowing confidence interval computation, Methods section 12), enhancers of all GenoSTAN models overlapped a higher fraction of GWAS variants at a similar to better per base pair density compared to the current ChromHMM annotations (Figure 17A). The same trend was observed for promoters when restricting to 1% of genomic coverage (Figure 17B). The improved overlap with trait-associated variants indicates that GenoSTAN annotation has a higher enrichment for functional elements than the current annotation.

In accordance with previous studies [103, 110] we found that individual variants were strongly enriched in enhancer or promoter states specifically active in the relevant cell types or tissues (Figure 17C, Appendix Figure A8C). Variants associated with height were significantly associated

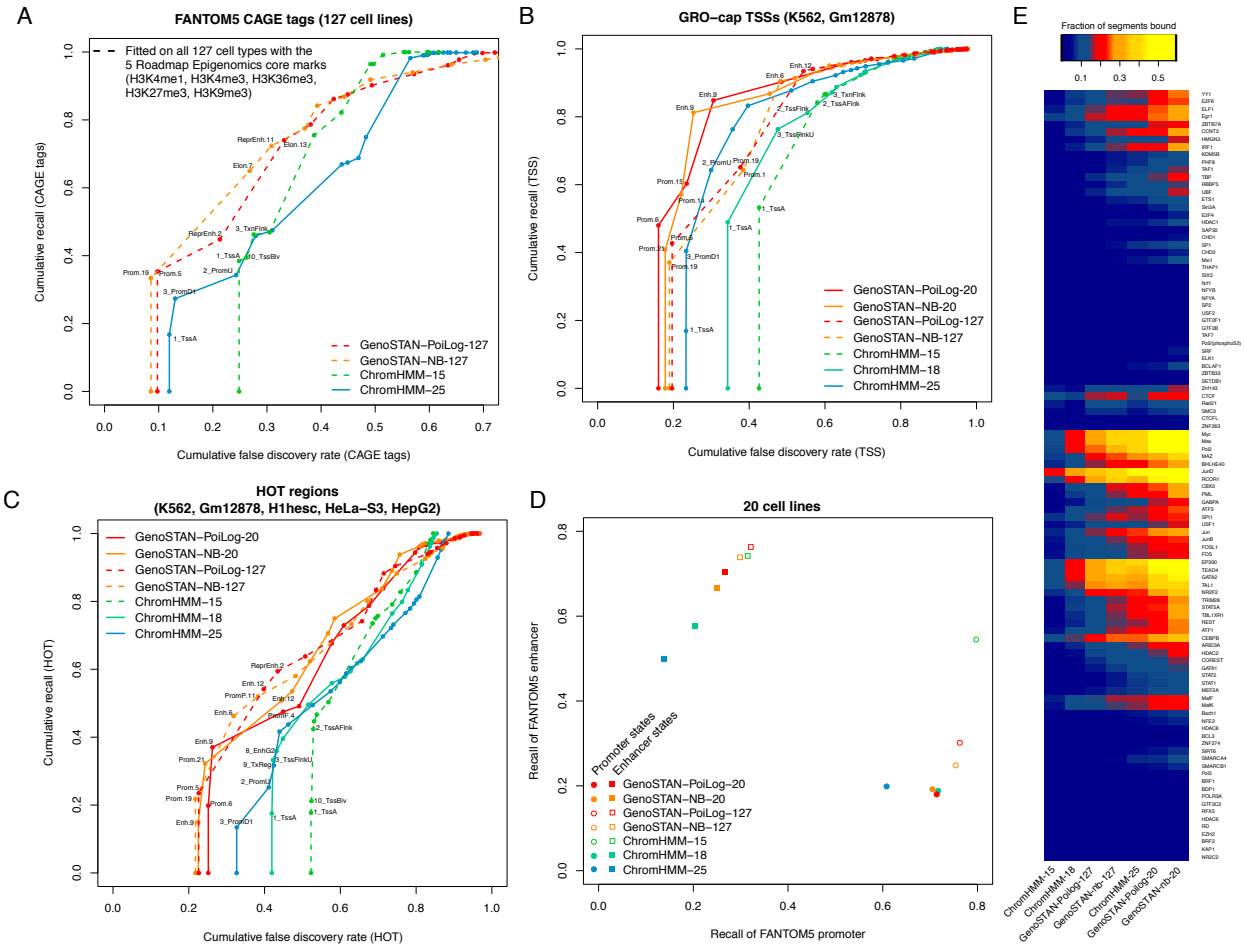


Figure 16: Comparison of GenoSTAN to other published segmentations on benchmark II and III. (A) Performance of chromatin states in recovering FANTOM5 CAGE tags in 127 cell types. CAGE tags were overlapped with chromatin states wihout the use of cell type information. Cumulative FDR and recall are calculated by subsequently adding states (in order of increasing FDR). (B) Performance of chromatin states in recovering GRO-cap transcription start sites in two cell types where GRO-cap data was available. (C) The same as in (B) for ENCODE HOT regions for five cell types where annotation of HOT regions was available. (D) Recall of FANTOM5 promoters and enhancers by predicted promoters and enhancers is plotted to assess how well models distinguish promoters from enhancers. (E) The fraction of predicted enhancer segments bound by individual TFs is shown for different studies. GenoSTAN enhancers are more frequently bound by TFs than those from other studies.

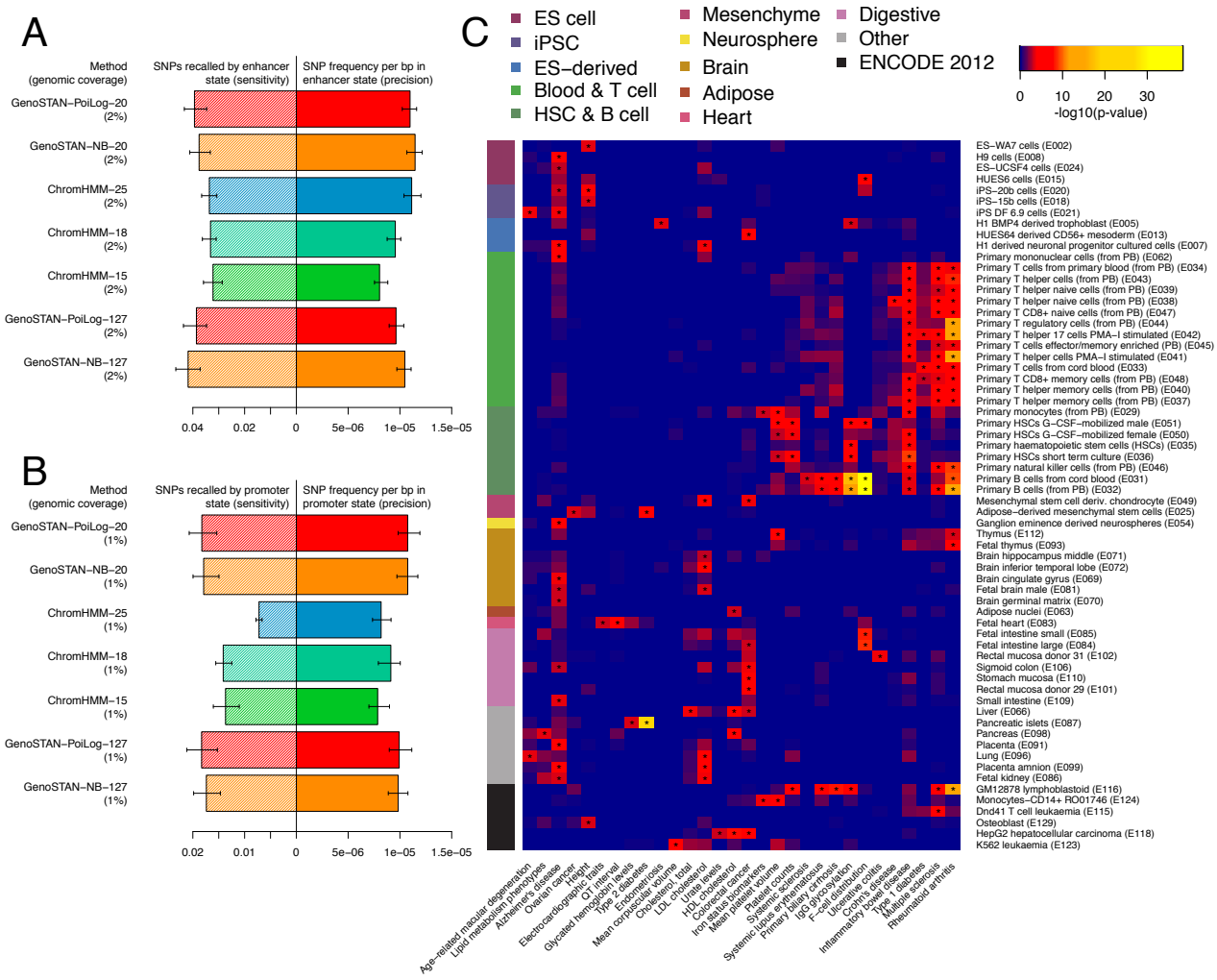


Figure 17: Enrichments of genetic variants associated with diverse traits in enhancers are specific to the relevant cell types or tissues. (A) Median SNP recall and frequency was calculated for enhancer states in different segmentations by restricting it to a total genomic coverage of 2% (100 samples of random subsetting) to control for different number of enhancer calls between the segmentations. Error bars show the 95% confidence interval. (B) The same as in (A) but for promoters. (C) The heatmap shows the -log10(p-value) of significantly enriched traits in enhancer states (GenoSTAN-Poilog-127, p-value < 0.001, marked by '*'). Only cell types and tissues where at least one trait was significantly enriched are shown. P-values were adjusted for multiple testing using the Benjamin-Hochberg correction.

with osteoblasts (at FDR <0.001 here and after, performed on Benchmark II for consistency across cell types and tissues). Variants associated with immune response or autoimmune disorders were enriched in B- and T-cell enhancers (Figure 17C) and promoters (Appendix Figure A8C). These incude for instance HIV-1 control, autoimmune disease associated SNPs for systemic lupus erythematosus, inflammatory bowel disease, Ulcerative colitis, Rheumatoid arthritis, Primary biliary cirrhosis and Multiple sclerosis. Variants associated with electrocardiographic traits and QT interval were enriched in fetal heart enhancers. SNPs associated with colorectal cancer were enriched in enhancers specific to the digestive system. These results illustrate that the annotation of potential promoters and enhancers generated in this study can be of great use for interpreting genetic variants associated, and underscore the importance of cell-type or tissue specific annotations.

15.5 A novel annotation of enhancers and promoters in human cell types and tissues

We then compiled the results from the best performing annotations for each cell type and tissue into a single annotation file. The combined annotation file and all individual chromatin state annotations are available at http://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN. For the combined annotation file, we chose GenoSTAN with Poisson-lognormal in every instance, as it performed best in almost every comparison we conducted. We used the results from benchmark I for K562, from benchmark III for the 20 cell types and tissues, and from benchmark II for all the remaining Roadmap Epigenomics cell types and tissues. Overall, our annotation reports typically between 8,945 and 16,750 (10% and 90% quantiles of number of promoters across all 127 cell types and tissues) active promoters per cell type or tissue. This number is consistent with the typical number of expressed genes per tissue (in 11,953 to 16,869 range, [214]). However, the median width of these elements depends on the data on which the annation was based. For the benchmark III dataset, promoters are much narrower (800bp median) than for the K562 annotations (1.4 kb, Benchmark I data sst), suggesting that promoter regions in the 20 cell types more accurately recover DNase hypersensitivity sites (DHS) of the core promoter (Figure 14, Appendix Figure A6). The number of enhancers per cell type or tissue varied more greatly (between 8,208 and 33,596 for the 10% and 90% quantiles). The large variation of the number of enhancers might be partly due to differences of sensitivity in complex biological samples. Consistent with this hypothesis, much fewer enhancers were identified in tissues than in primary cells and cell lines (Appendix Figure A9) likely because enhancers that are active only in a small subsets of all cell types present of a tissue may be not detected. As more cell-type specific data will be available, improved maps can be generated. The GenoSTAN software, which is publicly available, will be instrumental to update these genomic annotations.

15.6 Promoters and enhancers have a distinct TF regulatory landscape

The biochemical distinction between enhancers and promoters is a topic of debate [14, 195]. We explored to which extent enhancers and promoters are differentially bound by TFs using the K562 cell line dataset because i) we obtained the most accurate annotation for this cell line (GenoSTAN-Poilog-K562, Benchmark I) and ii) ChIP-seq data was available for as many as 101 TFs in this cell line [93]. Nine TF modules were defined by clustering based on binding pattern similarity across enhancers and promoters (Methods section 12, Figure 18). These 9 TF modules were further characterized by the propensity of their TFs to bind promoters, enhancers

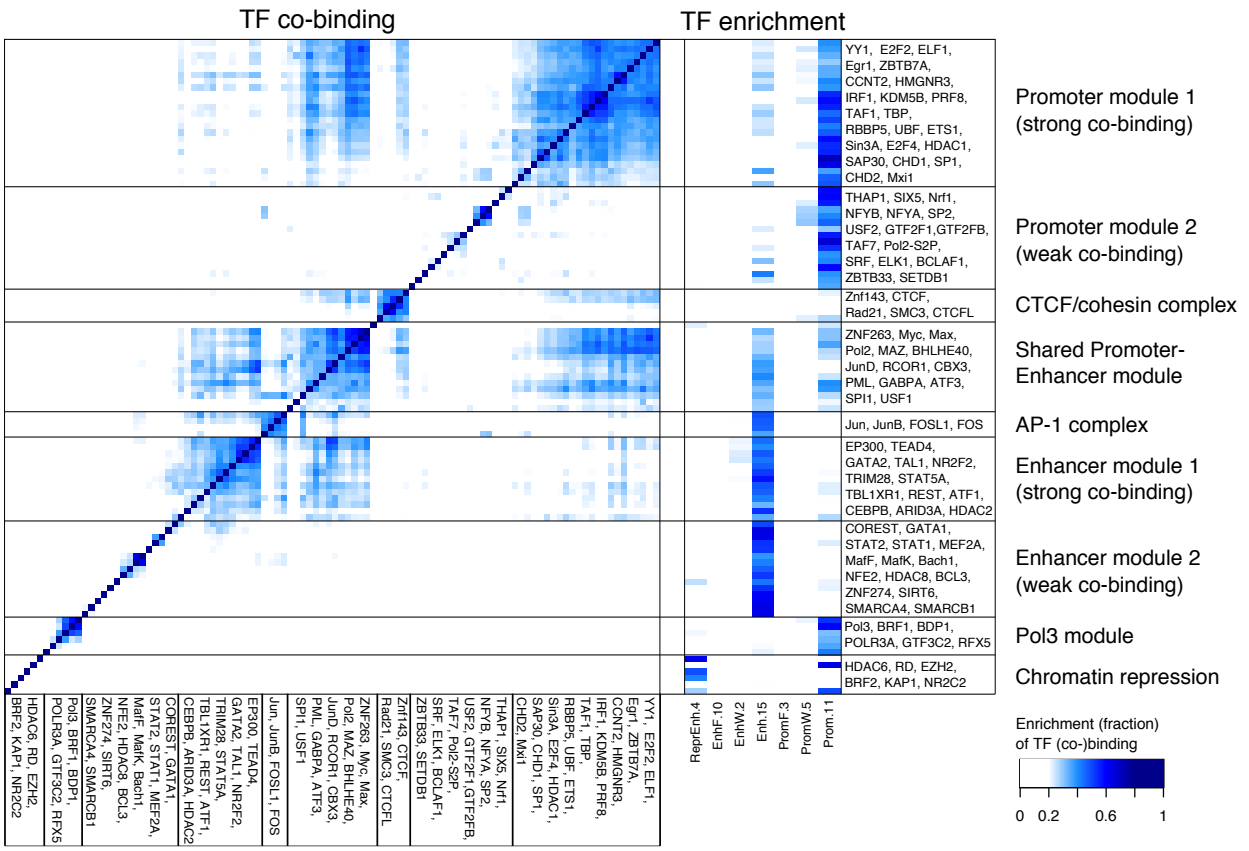


Figure 18: Promoters and enhancers have a distinctive TF regulatory landscape. Co-binding (left) and enrichment of transcription factor binding sites (right) in chromatin states (GenoSTAN-Poilog-K562) for 101 transcription factors in K562 reveals TF regulatory modules with distinct binding preferences for promoters, enhancers and repressed regions. The co-binding is depicted as the frequency of binding sites of two TFs that co-occur in a chromatin state divided by the number of all binding sites of the two TFs (Jaccard index). For each TF, enrichments were normalized to sum up to 1 across all 18 chromatin states of GenoSTAN-Poilog-K562.

or both (Figure 18). In accordance with previous studies [180, 215], this recovered many complexes and promoter-associated and enhancer-associated proteins, including the CTCF/cohesin complex (CTCF, Rad21, SMC3, Znf143), the AP-1 complex (Jun, JunB, FOSL1, FOS), Pol3, promoter and enhancer associated modules, and factors associated with chromatin repression (EZH2, HDAC6).

Moreover, the modules identified provided insights into the distinction of promoters and enhancers. On the one hand, some TFs are common to both enhancers and promoters, which supports previous reports [14, 199]. In accordance with the recent finding of widespread transcription at enhancers [13], Pol II and multifunctional TFs Myc, Max, and MAZ [216] are part of a TF module - which we called the Promoter-Enhancer-Module (PEM) - which had approximately equal binding preferences for promoter and enhancer states, but also co-localized with other TFs specifically binding enhancers or promoters (Figure 18).
On the other hand enhancers and promoters were also bound by distinct TFs, which is consistent with previously reported TF co-occurrence patterns at gene-proximal and gene-distal

sites [180, 215]. Among the promoter and enhancer-associated proteins we defined Promoter module 1 and 2 (PM1, PM2), Enhancer module 1 and 2 (EM1, EM2), which had a strong preference for binding either a promoter or an enhancer, but exhibited different co-binding rates (Figure 18). Promoter module 1 contained TFs which were specifically enriched in promoter states and associated with basic promoter functions, such as chromatin remodeling (CHD1, CHD2), transcription initiation or elongation (TBP, TAF1, CCNT2, SP1) and other TFs involved in the regulation of specific gene classes (e.g. cell cycle: E2F4) [216]. However, it also included TFs known as transcriptional repressors (e.g. Mxi1, a potential tumor suppressor, which negatively regulates Myc). While TFs in PM1 showed a high co-binding rate, PM2 factors exhibited low co-binding. This might be partially explained by lower efficiency of the ChIP, since PM2 also contained general TFs such as TFIIB, TFIIF or the Serine 2 phospho-isoform of Pol II, which are expected to co-localize with other general TFs from PM1.
EM1 contained TFs with high co-binding rate, which included TAL1, an important lineage-specific regulator for erythroid development (K562 are erythroleukemia cells) and which had been shown to interact with CEBPB, GATA1 and GATA2 at gene-distal loci [215, 217]. It also contained the enhancer-specific transcription factor P300 [207] and transcriptional activators (e.g. ATF1) and repressors (e.g. HDAC2, REST) [216]. Analogously to PM2, EM2 contained enhancer-specific transcriptional activators and repressors with a low co-binding rate.
Altogether this analysis highlights the common and distinctive TF binding properties of enhancers and promoters.

15.7 Discussion

We introduced GenoSTAN, a method for *de novo* and unbiased inference of chromatin states from genome-wide profiling data. In contrast to previously described methods for chromatin state annotation, GenoSTAN directly models read counts, thus avoiding data transformation and the manual tuning of thresholds (as in ChromHMM and Segway), and variance is not shared between data tracks or states (as in EpicSeg and Segway) [101, 150, 154]. GenoSTAN is released as part of the open-source R/Bioconductor package STAN [113, 156, 161], which provides a fast, multiprocessing implementation that can process data from 127 human cell types in 3-6 days (GenoSTAN-Poilog-127: 6 days, -nb: 3 days).
Application of GenoSTAN significantly improved chromatin state maps of 127 cell types and tissues from the ENCODE and Roadmap Epigenomics projects [93, 110]. Binding of enhancer-associated co-activator CBP and histone acetyltransferase P300 was used by several studies for the genome-wide prediction of enhancers [62, 206, 207]. From these predictions a distinctive chromatin signature for promoters and enhancers was derived based on H3K4me1 and H3K4me3 [62]. In particular, the ratio H3K4me1/H3K4me3 was found to be low at promoters, in comparison to enhancers. Active and poised enhancers could also be distinguished by presence or absence of H3K27me3 and H3K9me3 [63]. All these features could be confirmed by GenoSTAN, making it a promising tool for the biochemical characterization of enhancers and promoters. Moreover, extensive benchmarks based on independent data including transcriptional activity, TF binding, cis-regulatory activity, and enrichment for complex trait-associated variants showed the highest accuracy of GenoSTAN annotations over former genome segmentation methods.
The GenoSTAN annotation sheds light on the common and distinctive features of promoters and enhancers, which currently are an intense subject of debate [14,195]. Among other characteristics, a shared architecture of promoters and enhancers was proposed based on the recent discovery of widespread bidirectional transcription at enhancers [3, 13, 14]. This was supported by the obser-

vation that enhancers, which are depleted in CpG islands have similar transcription factor (TF) motif enrichments as CpG poor promoters [199]. However, another study showed that TF co-occurrence differed between gene-proximal and gene-distal sites [180, 215]. GenoSTAN chromatin states revealed a very distinct TF regulatory landscape of these elements and therefore suggest that promoters and enhancers are fundamentally different regulatory elements, both sharing the binding of the core transcriptional machinery. Our annotation of enhancers and promoters will be a valuable resource to help characterizing the genomic context of the binding of further TFs.
Indirectly, our analysis showed that chromatin state annotations are better predictors of enhancers than the transcription-based definition provided by the FANTOM5 consortium [199]. While FANTOM5 enhancers are an accurate predictor for transcriptionally active enhancers, the sensitivity remains poor (only 4,263 enhancers were called by overlap with GRO-cap TSSs and DHS, which is less than the estimated number of transcribed genes, for K562 cells compared to about 20,000-30,000 for ChromHMM and 10,000-20,000 for GenoSTAN). Although, the sensitivity of the transcription-based approach can increase with transient transcriptome profiling [86, 218] or nascent transcriptome profiling [219], the chromatin state data undoubtedly add valuable information for the identification of promoters and enhancers. Because it models count data, GenoSTAN analysis can in principle also integrate RNA-seq profiles, for instance using it in a strand-specific fashion [113].
Systematic identification of cis-regulatory active elements by direct activity assays is notoriously difficult. STARR-Seq for instance is a high-throughput reporter assay for the *de novo* identification of enhancers [194]. It was previously used to identify thousands of cell-type specific enhancers in *Drosophila*, but has not been applied to human yet. Moreover, STARR-Seq makes rigid assumptions about the location of the enhancer element with respect to the promoter, and it does not account for the native chromatin structure. This might identify regions that are inactive *in situ* [194]. Other experimental assays for the validation of predicted ENCODE enhancers lead to different results [192, 193]. Complementary to these approaches, the systematic evaluation of cis-regulatory activity based on candidate regions in human cells have made progress with the advent of high-throughput CRISPR perturbation assays [220]. Because it requires candidate cis-regulatory regions in a first place, such approach will benefit from improved annotation maps as the one we are providing.
Thus, we foresee GenoSTAN to be instrumental in future efforts to generate robust, genome-wide maps of functional genomic regions like promoters and enhancers.

16 TT-Seq captures the human transient transcriptome

All results presented in this section were obtained in collaboration with Margaux Michel and Björn Schwalb and are part of the manuscript "TT-Seq captures the human transient transcriptome", which was accepted for publication in Science. For detailed author contributions see page ix.

Pervasive transcription at cis-regulatory elements generates many different RNA species including protein-coding mRNAs and non-coding RNAs (ncRNAs), such as enhancer RNAs (eRNAs) [13,199,221]. Most ncRNAs are rapidly degraded and therefore difficult to detect. Sensitive protocols to detect these short-lived RNAs such as GRO-cap [13] are limited to detection of the transcription start sites. However, a comprehensive map of these transient RNAs in their full length is required to understand regulation of RNA transcription and degradation.
To this end, transient transcriptome sequencing (TT-Seq) was developed, which maps transcribed genomic regions with high sensitivity and enables the computation of RNA synthesis and degradation rates with kinetic modeling (Appendix section 18 for details). The TT-Seq protocol is a modfication of 4sU-Seq [222], where newly synthesized RNAs are labeled with the nucleoside analog 4-thiouridine (4sU), which is incorporated into RNA during transcription. The labeled RNAs are then purified and sequenced (Appendix Figure A10A, left panel). 4sU-Seq is more sensitivity in detecting short lived RNAs than standard RNA-Seq [222]. However, the extracted RNA in 4sU-Seq contains unlabeled 5' parts, which were synthesized before 4sU labeling leading to a 5' bias of in the sequenced fragments (Appendix Figure A10B). TT-Seq modifies the 4sU-Seq protocol by fragmenting the RNAs prior to isolation of labeled RNA (Appendix Figure A10A, right panel). Thus only the labeled parts of newly synthesized RNA are extracted. This results in an almost uniform coverage of sequenced RNA fragments (Appendix Figure A10B). Moreover, kinetic modeling with TT-Seq allows quantification of RNA synthesis and degradation rates (see Appendix section 18).
Here we apply GenoSTAN (see section 8 and 15) with TT-Seq in human K562 cells to generate a comprehensive and sensitive map of the transcriptome, including thousands of transient RNAs such as eRNAs. This map reveals differences between promoter and enhancer transcritption. Moreover, we show that RNA sequence features correlate with degradation rates and transcript lengths. Altogether, we anticipate the transcriptomic map derived using TT-Seq and GenoSTAN to be a useful resource for studying transcription in human.

16.1 A comprehensive map of the transcriptome in human K562 cells

Using TT-Seq data we identified 21,874 genomic intervals of apparently uninterrupted transcription (Transcriptional Units, TUs) by applying GenoSTAN with Poisson-lognormal emission distributions (Figure 19A, Methods section 13). TT-Seq shows high sensitivity, recovering 65% of the transcription start sites (TSSs) obtained by GRO-cap (overlapping segments in a window of ± 400 bp) [13]. A total of 8,543 TUs overlapped a GENCODE annotation [104] in sense direction of transcription (50% reciprocal overlap of segment size, Figure 5, Methods section 13), revealing 7,810 mRNAs, 302 long intergenic non-coding RNAs (lincRNAs), and 431 antisense RNAs (asRNA). 2,916 TUs overlapped with GENCODE annotations by less than 50% segment size and were not classified. The remaining 10,415 (48%) TUs were newly detected ncRNAs that we characterized further.
The remaining 10,415 non-annotated TUs were now classified based on the GenoSTAN-Poilog-K562 (Figure 14, see section 15) chromatin state annotation and their positions relative to known

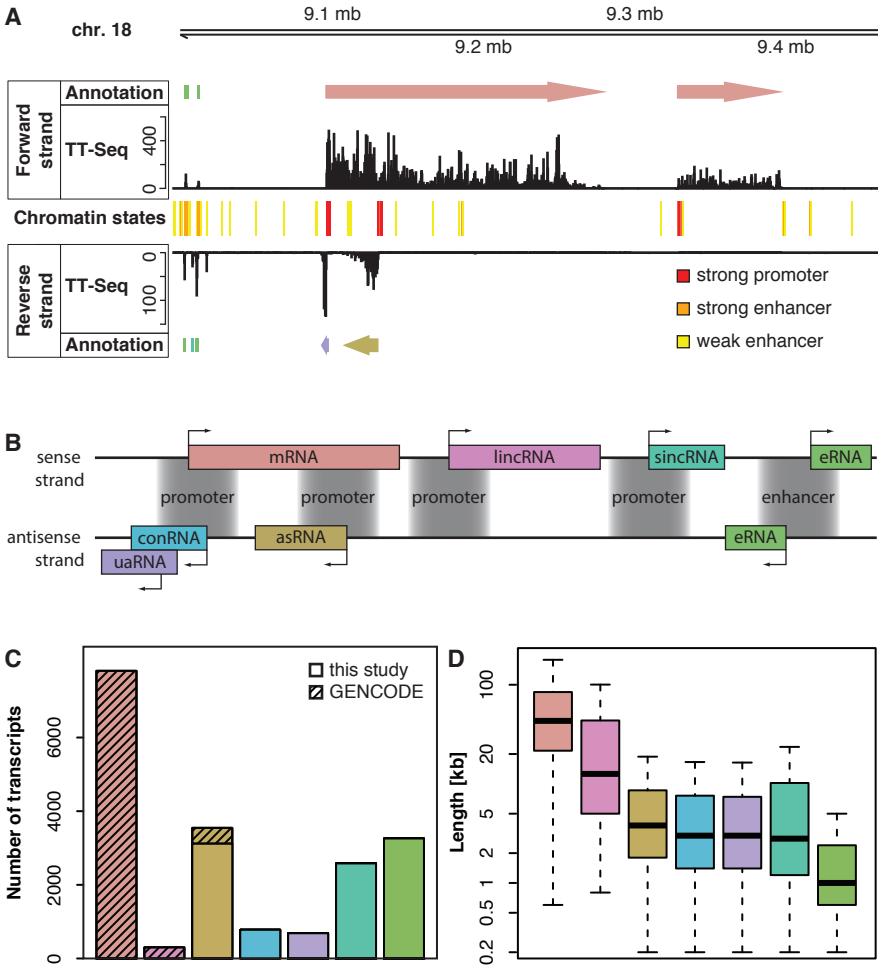


Figure 19: Annotation of transient RNAs mapped by TT-Seq. (A) Genome browser view of an exemplary region on chromosome 18. The depicted region shows transcripts from five of a total of seven transcript classes. It also shows three of a total of 18 chromatin states (GenoSTAN-Poilog-K562, see section 15). (B) Schematic giving the definition of transcript classes and color-code used for different RNA types. (C) Number of transcripts in different classes (portions covered by GENCODE are hatched). (D) Distribution of transcript lengths.

GENCODE annotations (Figure 19B). TUs within 1 kb of an GENCODE mRNA TSS included 685 upstream antisense RNAs (uaRNAs) [223] and 778 convergent RNAs (conRNAs) [224]. The 3,115 TUs on the strand opposite of a GENCODE mRNA were classified as antisense RNAs (asRNAs) when they were more than 1 kb away from the GENCODE mRNA TSS [225]. Remaining TUs were grouped according to their underlying GenoSTAN-Poilog-K562 chromatin state at their TSS. 2,580 TUs originate from promoter state regions (Methods section 13) and were called short intergenic ncRNAs (sincRNAs) (Figure 19C). Most sincRNAs (67%) were located within 10 kb of an GENCODE mRNA TSS. The remaining 3,257 TUs originate from enhancer state regions (Methods section 13) and were classified as enhancer RNAs (eRNAs) [3, 226]. New ncRNAs are short, and as sincRNAs are on average five times shorter than lincRNAs, and eRNAs have a median length of only 974 nucleotides (Figure 19D).

16.2 Transcript half-lives correlate with RNA sequence features

Kinetic modeling of TT-Seq and RNA-Seq data enabled us to estimate local RNA synthesis rates (average nucleotide bond formation) and half-lives (average nucleotide bond breakage) genome-wide (Appendix section 18) (Figure 20A, Appendix Figure A11A). We found that mRNAs and lincRNAs have the highest synthesis rates and longest median half-lives, 50 min and 38 min, respectively [218]. Other transcript classes show low synthesis and short mean half-lives, explaining why these ncRNAs are generally not detected. Short RNA half-lives, as short as ~ 2 minutes for eRNAs, correlate with a lack of secondary structure [183] (Appendix Figure A11B). In eRNAs, only 10% of the sequence is predicted to be structured, compared to 52% in mRNAs (Appendix Figure A11C). Indeed the folding energy [183] of eRNAs compares to the genomic background level (Figure 20B).

16.3 Differences in the transcription of promoters and enhancers

Our analysis further reveals differences in the transcription of promoters and enhancers [227]. With respect to initiation, enhancers show lower occupancy with the initiation factor TBP than mRNA promoters (12-fold less, p-value $< 10^{-16}$), whereas occupancies with factors involved in polymerase pausing such as NELF-E and the P-TEFb subunit cyclin T2 [228] are similar (Figure 20C). eRNA synthesis terminates early (Figure 19D), likely because eRNAs are not enriched in U1 snRNP-binding sites (U1 signals; GGUAAG, GUGAGU, or GGUGAG) that are known to protect mRNA transcription from early termination [229, 230]. eRNAs contained U1 signals at the level of genomic background (47%), whereas mRNAs were enriched (69%, p-value $< 10^{-16}$) (Figure 20D). Generally, an enrichment of U1 signals in the first 1,000 nucleotides positively associated with RNA length for all transcript classes (Figure 20E). These results suggest that evolution of stable RNAs generally involves acquisition of U1 signals, as shown previously for uaRNAs [231].

16.4 Discussion

The use of genomic data or the interpretation of genetic variants highly depend on the reference annotation of a genome. The transcriptome annotation using GenoSTAN with TT-Seq provides an improved reference map of the human transcriptome in K562 cells. TT-Seq also allows the calculation of RNA synthesis rates and half lives which we used with our annotation to derive new insights into human genome transcription. This revealed differences between transcription of promoters and enhancers. In particular, transcription factor binding differed between promoters and enhancer and the of lack secondary structure correlated with short RNA half lives. The occurence of U1 sites was found to be positively associated with transcript length, which were present in enhancer RNAs at genomic background frequency but were enriched in mRNAs.
Differences and similarities between promoters and enhancers are currently under debate [14, 195] (see section 15.7 for a detailed discussion). Moreover, enhancer RNAs have been shown to be required for proper enhancer function [232] or activition of the chromatin at the promoter of the target gene [233], but the exact mechnisms of enhancer RNA function remain unclear. Transcriptome mapping from TT-Seq data is a promosing tool for elucidating further features and functions of transcripts from promoters and enhancers by providing higher sensitivity than current RNA-Seq approaches to detect transient RNAs.

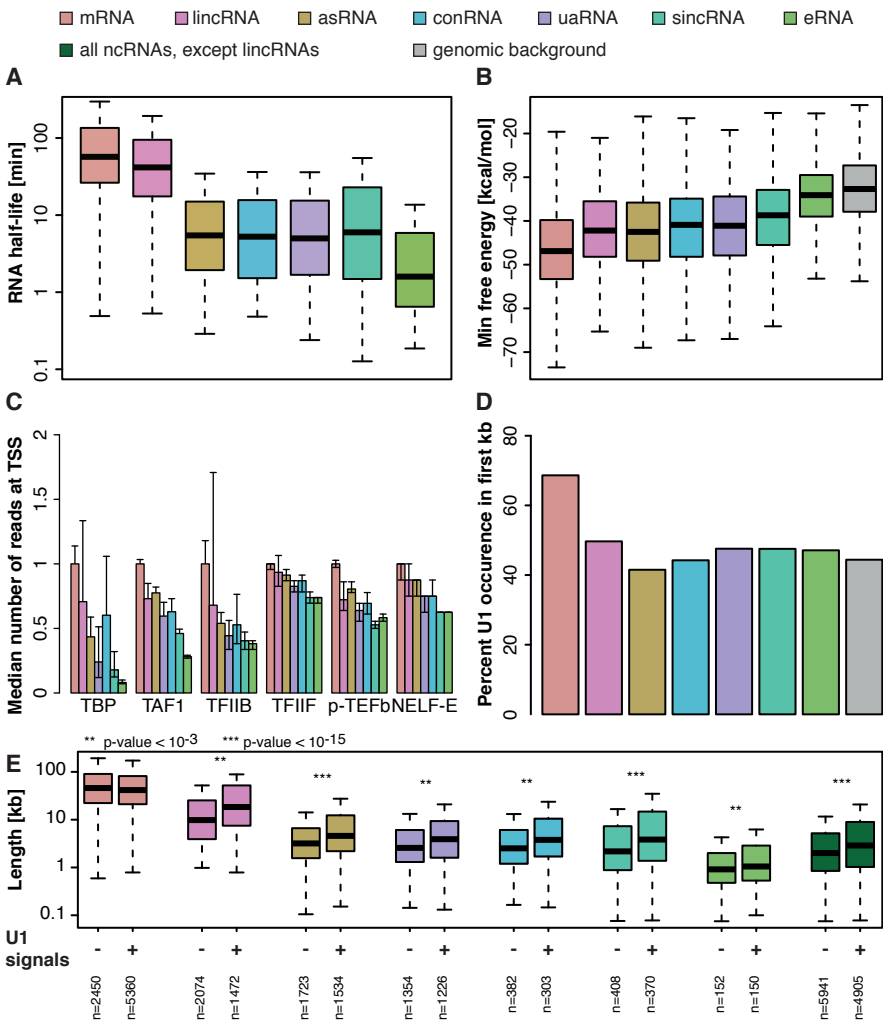


Figure 20: Transcript half-lives correlate with RNA sequence features. (A) Distribution of half-lives in different transcript classes. (B) Distribution of the minimum free energy in the first 1000 nucleotides (nt) per transcript class. (C) Distribution of relative peak occupancies with factors binding promoters (+/- 100 bp from TSS) for transcript classes. (D) Occurrence of U1 signal in the first 1000 nt for different transcript classes. (E) Distribution of transcript lengths in transcript classes depends on the presence of U1 signals in the first 1000 nt.

Part IV
Conclusion

Recent technological advances (such as microarrays and Next-Generation-Sequencing) have made it possible to measure genomic features and phenotypes in a genome-wide and cost-effective manner. Consortia like ENCODE [93] and Roadmap Epigenomics [110] have generated today's largest data sets measuring a variety of genomic features (for instance histone modifications, transcription factor binding, RNA expression, DNA accessibility or DNA methylation) in hundreds of cell types and tissues. This huge amount of data requires new and efficient computational tools to integrate, analyse and annotate these data. Accurate methods to accomplish this are therefore crucial for current and future biological research.

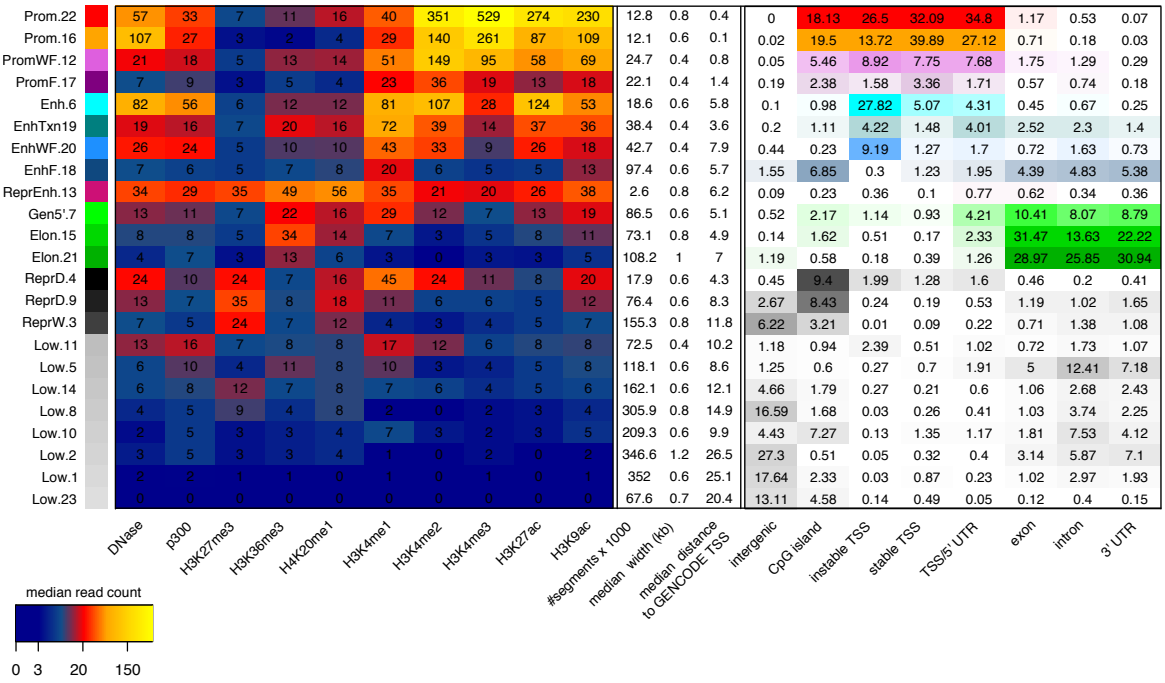
In this thesis bidirectional HMMs (bdHMMs) were proposed for integration of strand-specific with non-strand-specific data as well as a count-based HMM (GenoSTAN) for integration and segmentation of sequencing data. Both methods are made available in the R/Bioconductor package STAN [113]. STAN provides a fast, multiprocessing implementation that can be run efficiently on large genomes such as human.

The STAN package was used to infer accurate annotations of transcription in yeast (section 14) and human K562 cells (section 16) and of cis-regulatory elements in 127 human cell types and tissues (section 15). In particular, bdHMMs were used on a data set of yeast transcription factors to annotate directed genomic states, each of which modeling a specific phase of the transcription cycle. We found that the composition of protein factors in each state matched their known roles in transcription and that the sequence of states in the transcription cycle is depenent on specific gene classes. Application to TT-Seq data in K562 cells provided a sensitive annotation of the human transcriptome that also captured many transient RNAs. Analysis of this annotation revealed differences in transcription of promoters and enhancers and the lack of U1 motifs and RNA secondary structure as potential determinants of transcript length and stability. Prediction of promoters and enhancers with STAN lead to an accurate map of cis-regulatory elements in 127 human cell types and tissues, revealing many biochemical and regulatory features that characterize promoters and enhancers.

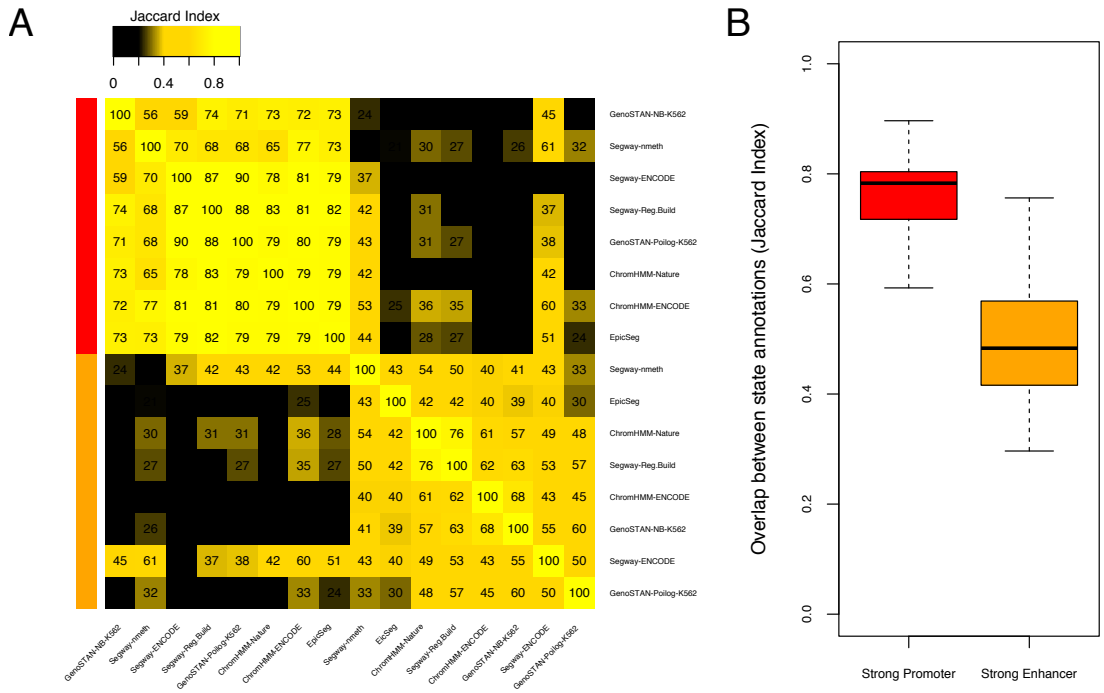
Taken together, we showed that STAN can be used to derive genome annotations, which are more accurate than those of previous methods. Analysis of these improved annotations lead to new biological insights. Thus the annotations derived in this thesis will be an important resource for future genomic data analysis. Moreover, the broad applicability of STAN will make it possible to infer such annotations also in other organisms and improve the current annotations as more data becomes available.

Part V
Appendix

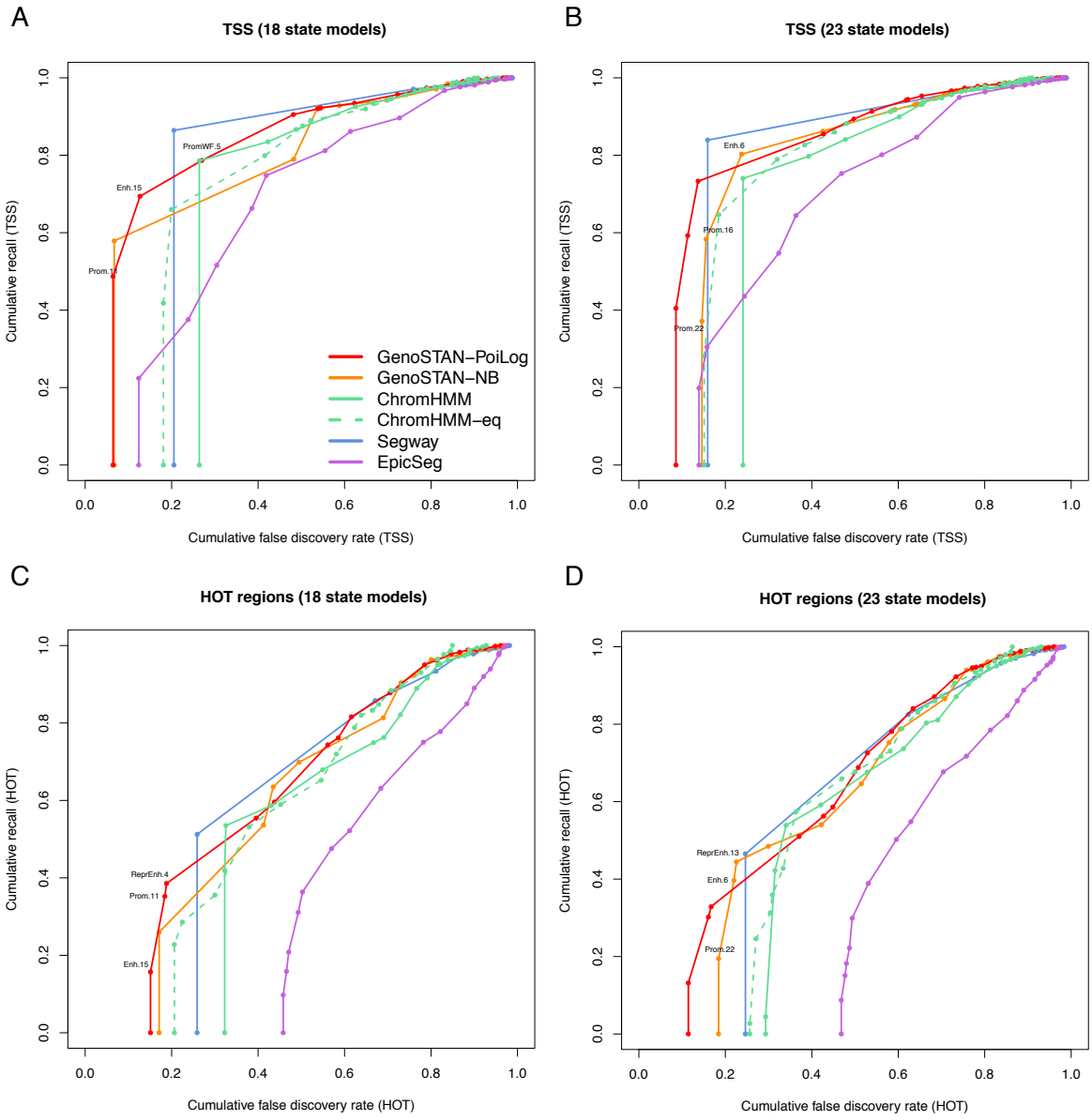
17 Additional information for section 15



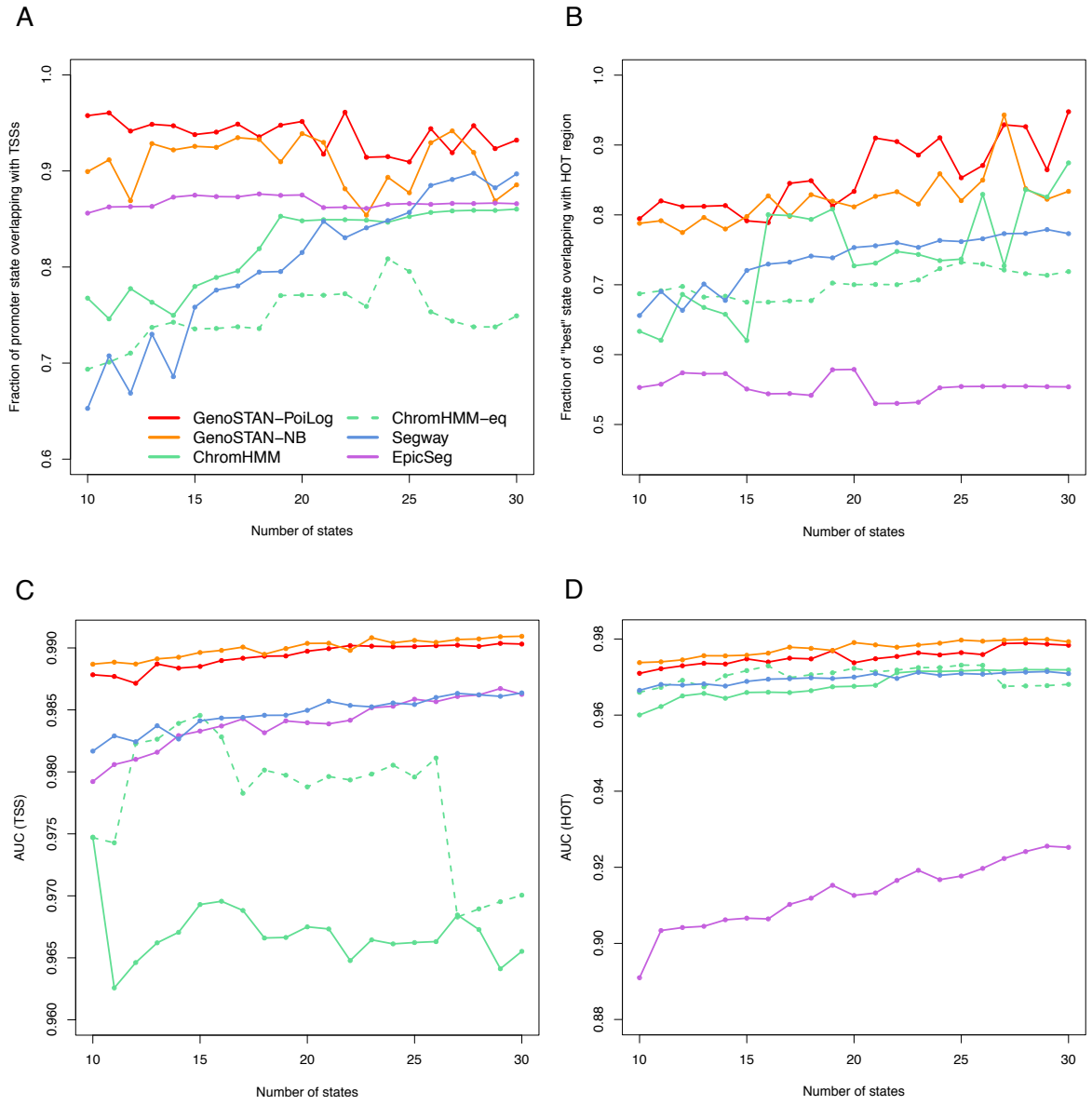
Appendix Figure A1: Median read coverage of GenoSTAN-nb-K562 chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSS (middle). The right panel shows recall of genomic regions by chromatin states.



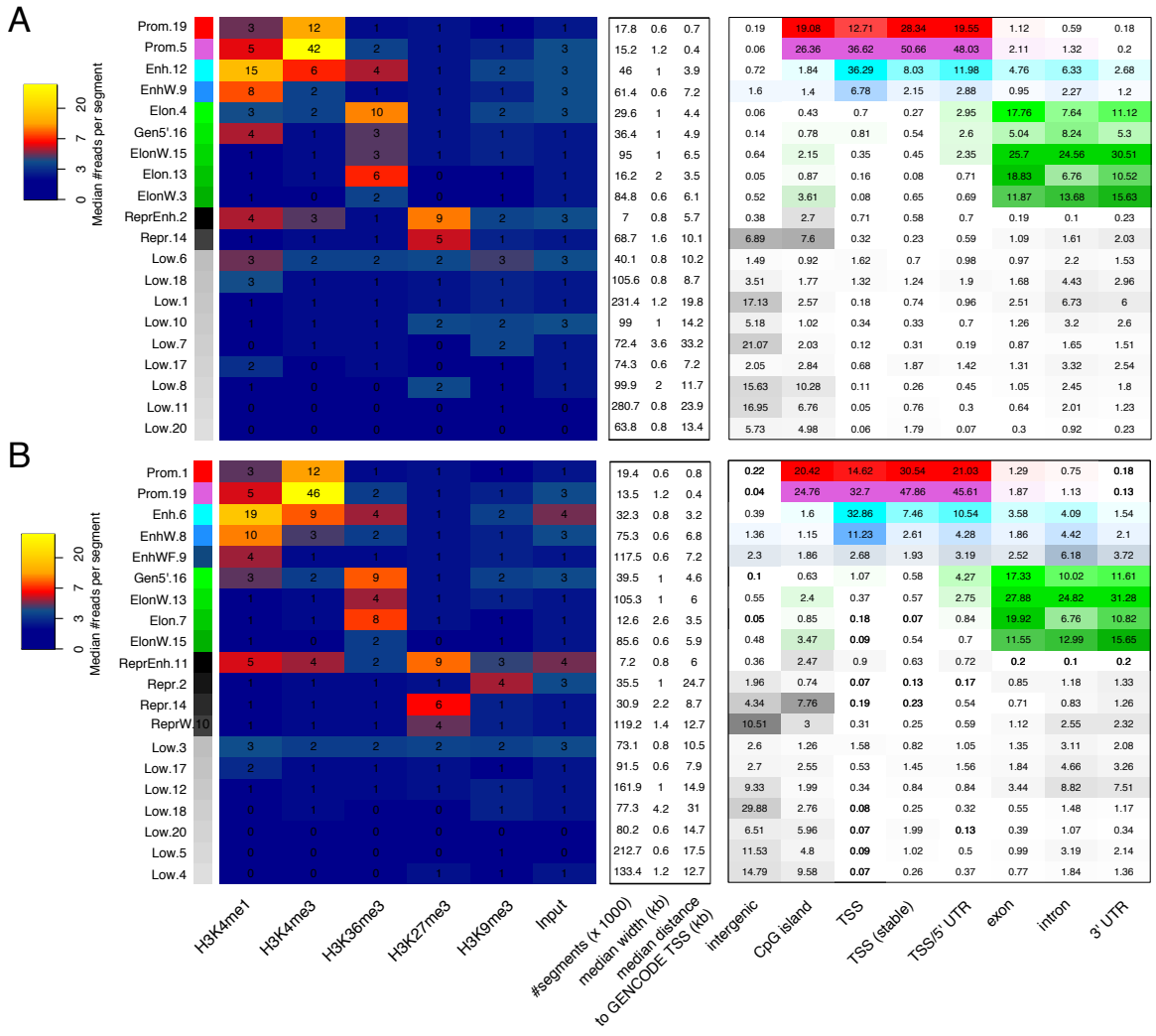
Appendix Figure A2: (A) Heatmap of pairwise overlap (Jaccard index) of promoter (red) and enhancer (orange) state annotations from different studies on benchmark I. Rows and columns were ordered by separate clustering of promoter and enhancer overlaps. (B) Distribution of pairwise Jaccard indices for strong promoters and enhancers (off-diagonal elements of promoter and enhancer sub-matrices from (A)).



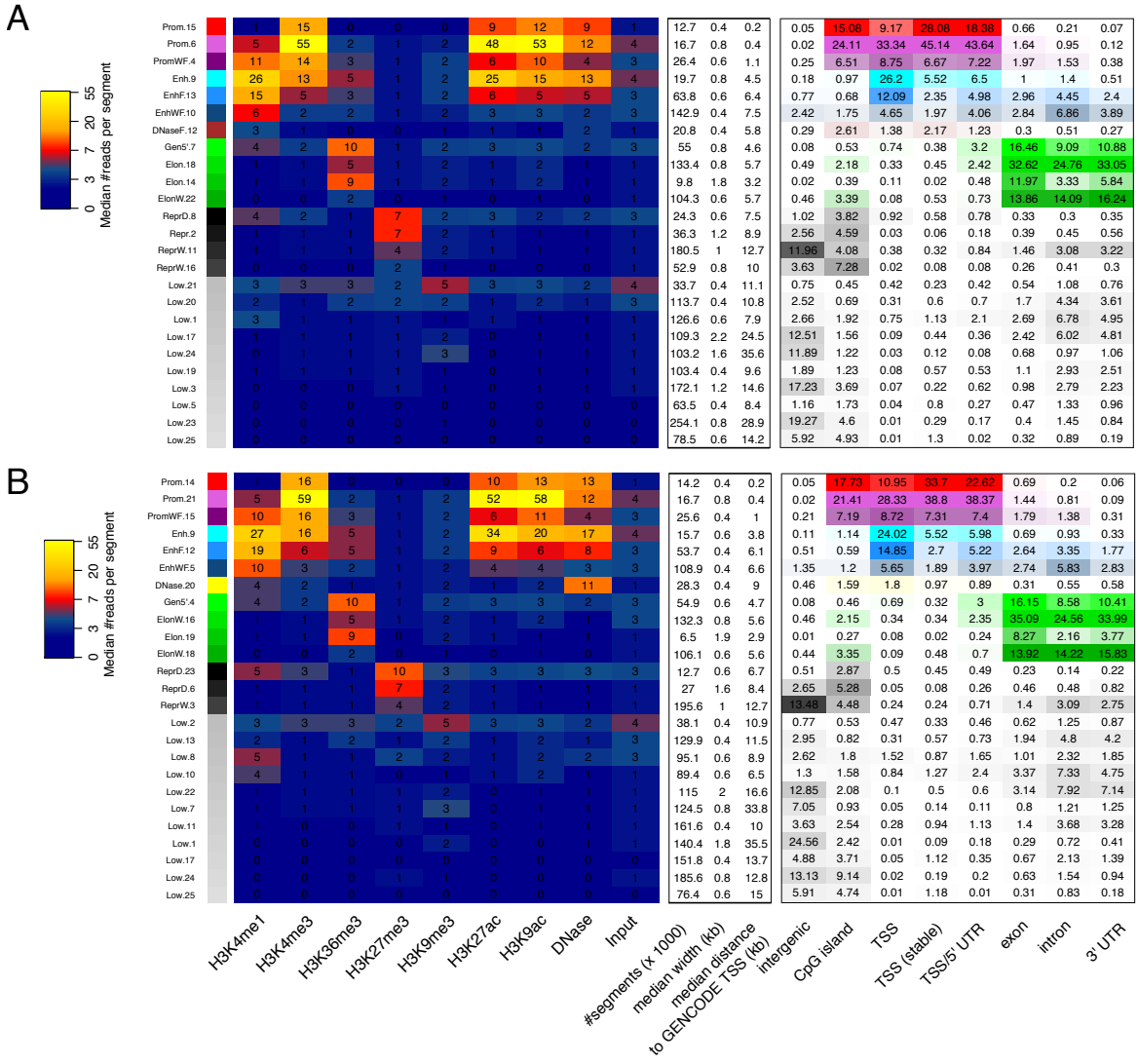
Appendix Figure A3: Comparison of GenoSTAN to other methods using 18 and 23 states on the same data set in K562 (Benchmark I). (A-B) Performance of chromatin states in recovering GRO-cap transcription start sites for the 18 and 23 state models. Cumulative FDR and recall are calculated by subsequently adding states (sorted by decreasing precision). (C-D) The same as in (A-B) for ENCODE HOT regions.



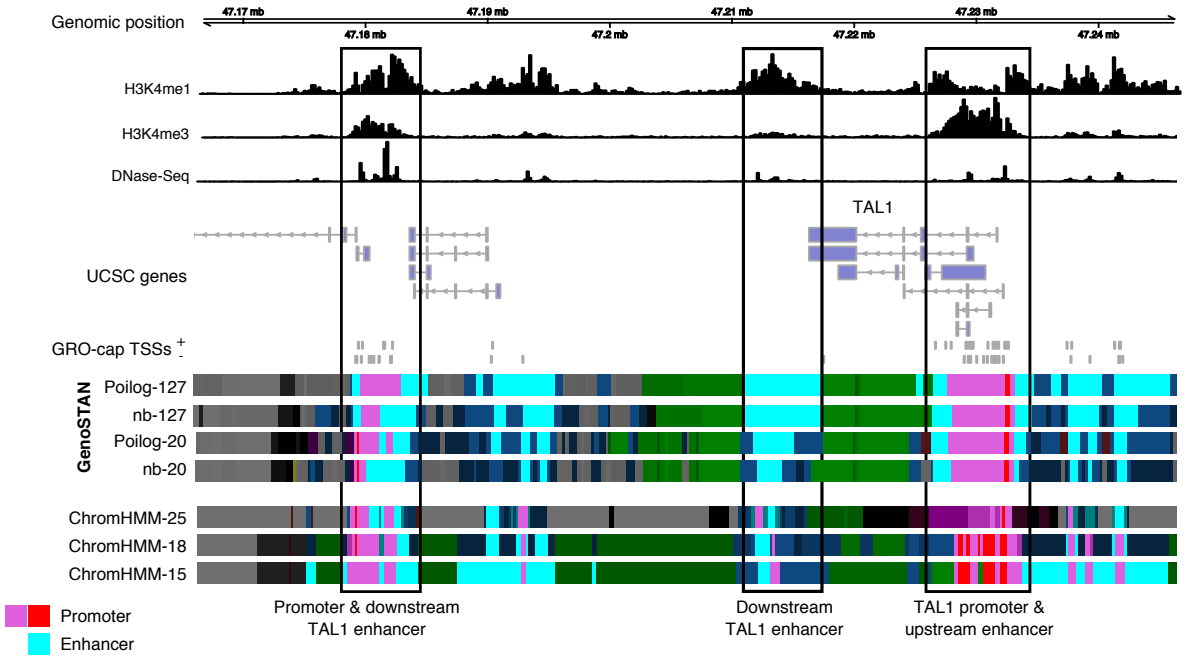
Appendix Figure A4: Comparison of chromatin segmentation algorithms with respect to their ability to call GRO-cap transcription start sites (left panels) and ENCODE HOT regions (right panels), as a function of the state number used in the respective algorithm (x-axes). All models were learned on the data set of benchmark I. (A-B) For each model, the state with highest precision in recalling HOT (respectively TSS) regions is shown. (C-D) For each model, an area under curve (AUC) score (see Methods section 12) is plotted to assess the spatial accuracy of a genome segmentation.



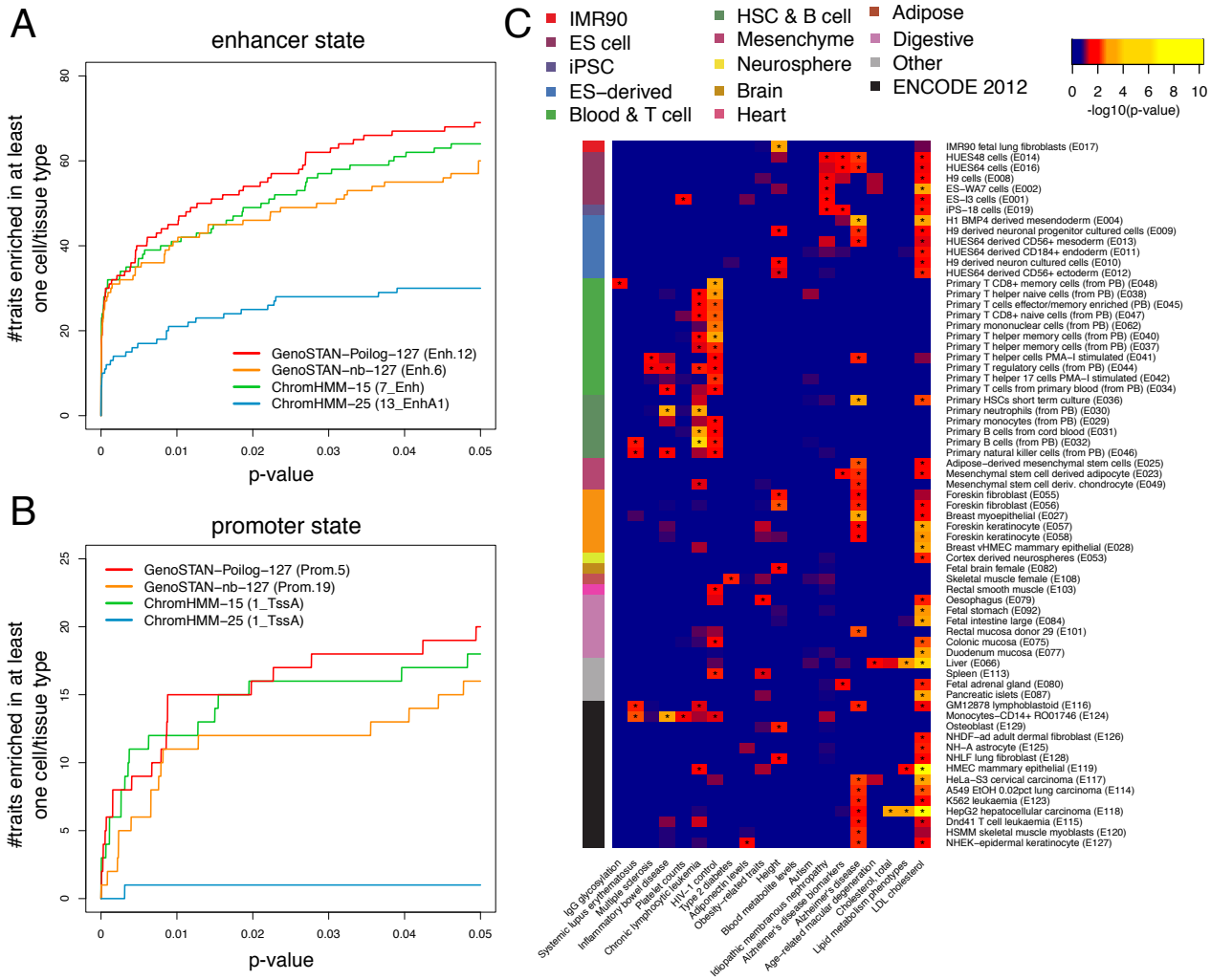
Appendix Figure A5: GenoSTAN models for benchmark II. (A) Median read coverage of GenoSTAN-Poilog-127 (Benchmark set II, fitted on the 127 ENCODE and Roadmap Epigenomics cell types and tissues) chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSS of segments (middle). The right panel shows recall of genomic regions by chromatin states. (B) The same as (A) for GenoSTAN-nb-127.



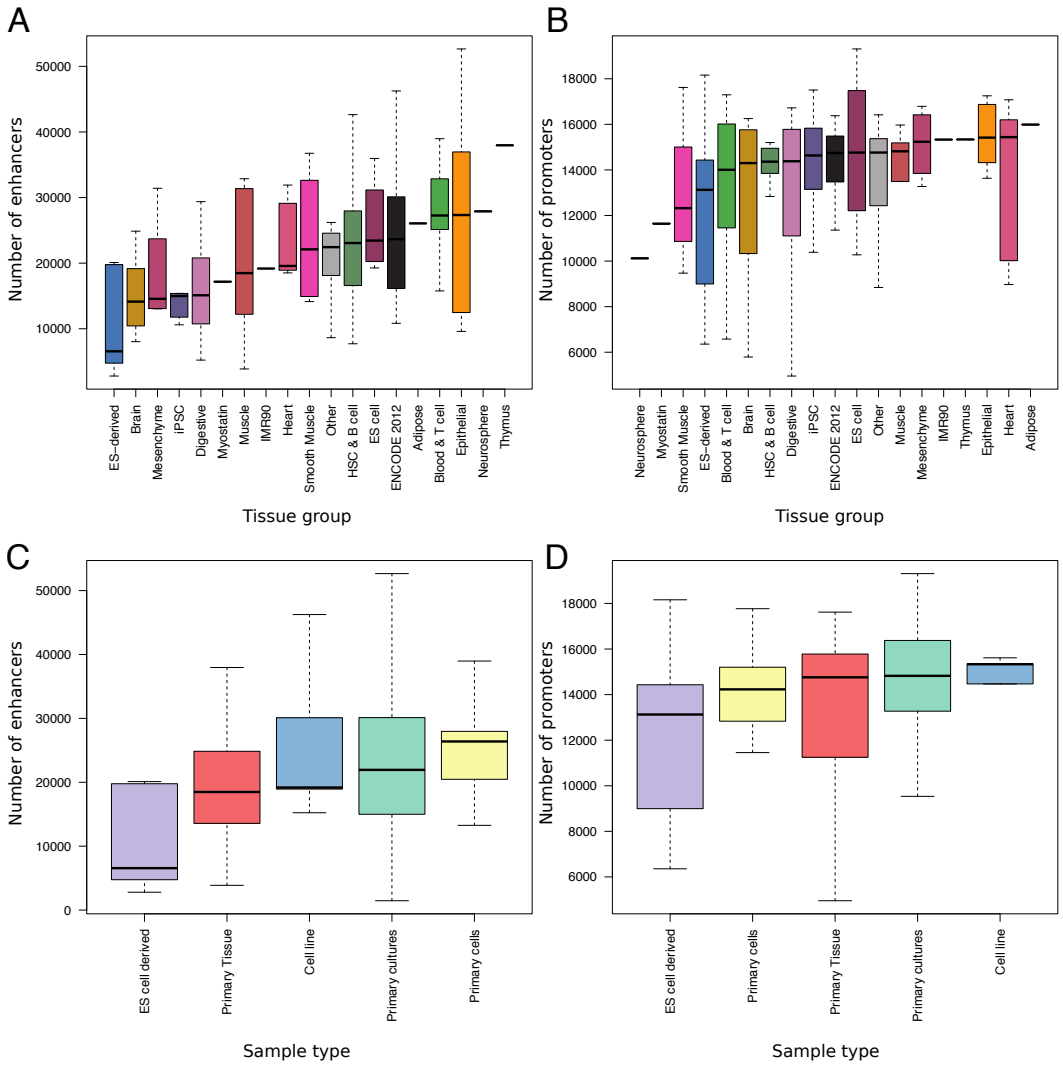
Appendix Figure A6: GenoSTAN models for benchmark III. (A) Median read coverage of GenoSTAN-Poilog-20 (Benchmark set III, fitted on the 20 ENCODE and Roadmap Epigenomics cell types and tissues) chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSS of segments (middle). The right panel shows recall of genomic regions by chromatin states. (B) The same as (A) for GenoSTAN-nb-20.



Appendix Figure A7: Chromatin state annotations for benchmark II and III. GenoSTAN segmentations are shown with the three Roadmap Epigenomics ChromHMM segmentations with 15 (ChromHMM-15), 18 (ChromHMM-18) and 25 (ChromHMM-25) states. Shown is the TAL1 locus with segmentations and data from the K562 cell line.



Appendix Figure A8: Enrichments of genetic variants associated with diverse traits in promoters are specific to the relevant cell types. (A) The number of traits which are enriched in enhancer states in at least one cell type or tissue is plotted for p-values < 0.05. (B) The same as in (A) but for promoters. (C) The heatmap shows the -log10(p-value) of significantly enriched traits in promoter states (GenoSTAN-Poilog-127, p-value < 0.05, marked by '*'). P-values were adjusted for multiple testing using the Benjamin-Hochberg correction.



Appendix Figure A9: Dependency of number of predicted promoters and enhancers on tissue group and sample type. (A) Number of enhancer states per Roadmap Epigenomics cell/tissue group. (B) The same as in (A) for promoters. (C) Number of enhancer states per Roadmap Epigenomics sample type. (D) The same as in (C) for promoters.

	Benchmark I - K562 (one cell type)	
Method/segmentation	#promoters	#enhancers
GenoSTAN-Poilog-K562	11,358 (Prom.11)	10,932 (Enh.15)
GenoSTAN-nb-K562	12,829 (Prom.22)	18,551 (Enh.6)
ChromHMM-Nature	16,118 (1_Active_Promoter)	30,492 (4_Strong_Enhancer)
ChromHMM-ENCODE	16,452 (Tss)	22,323 (Enh)
Segway-ENCODE	19,894 (Tss)	33,518 (Enh1)
Segway-nmeth	25,812 (8)	80,043 (0)
Segway-Reg.Build	13,668 (7_tss)	38,992 (11_proximal)
EpicSeg	16,192 (2)	53,982 (3)

	Benchmark II & III - 127 cell types and tissues	
Method/segmentation	#promoters	#enhancers
GenoSTAN-Poilog-127	15,229 (Prom.5)	45,955 (Enh.12)
GenoSTAN-nb-127	13,547 (Prom.19)	32,280 (Enh.6)
GenoSTAN-Poilog-20	12,710 (Prom.15)	19,730 (Enh.9)
GenoSTAN-nb-20	14,168 (Prom.14)	15,655 (Enh.9)
ChromHMM-15	21,002 (1_TssA)	92,824 (7_Enh)
ChromHMM-18	20,049 (1_TssA)	22,678 (9_EnhA1)
ChromHMM-25	12,525 (1_TssA)	12,706 (13_EnhA1)

Appendix Table A1: Number of promoter and enhancer states for the chromatin state annotations analyzed in this study. The original state name is given in brackets.

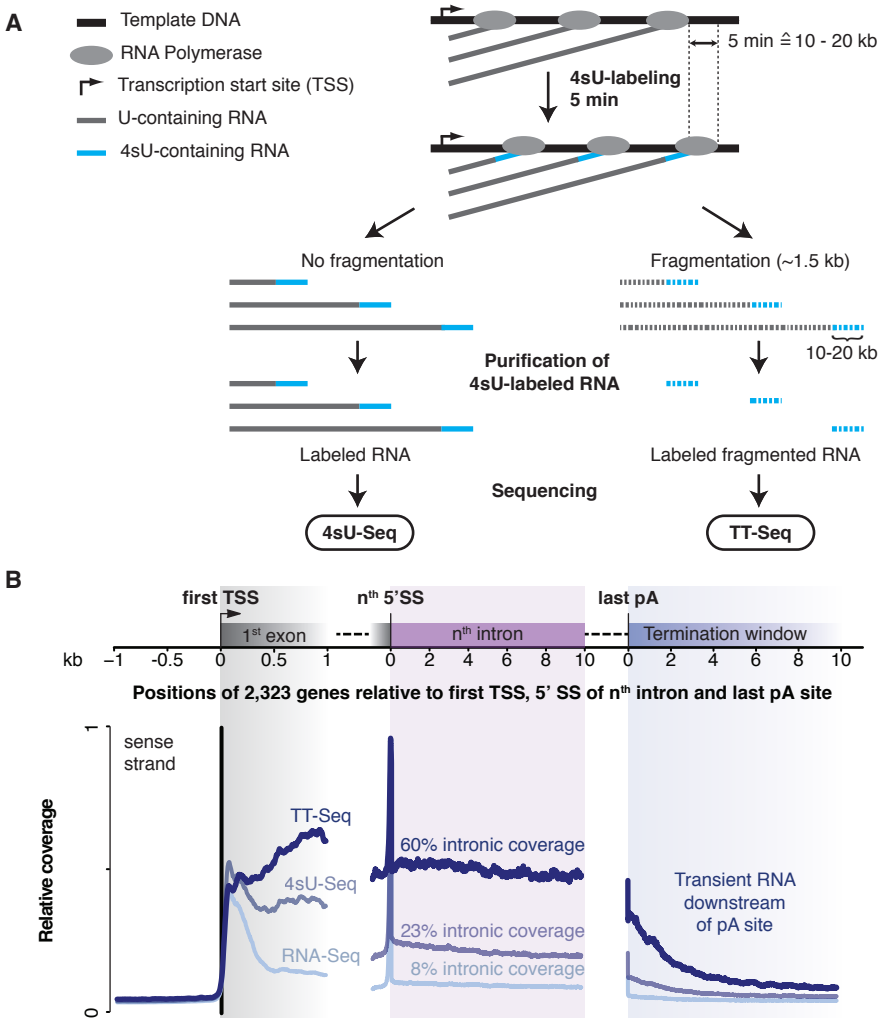
	Benchmark I - K562 (one cell type)	
Method/segmentation	promoter states	enhancer states
GenoSTAN-Poilog-K562	Prom.11, PromW.5	Enh.15, Enh.2
GenoSTAN-nb-K562	Prom.16, Prom.22	Enh.6, Enh.19
ChromHMM-ENCODE	Tss, TssF	Enh, EnhW
ChromHMM-nature	1_Active_Promoter 2_Weak_Promoter	4_Strong_Enhancer, 5_Strong_Enhancer
Segway-ENCODE	Tss, PromF	Enh1, Enh2
Segway-nmeth	8,6	0, 13
Segway-Reg.Build	7_tss, 0_proximal	1_proximal, 11_proximal
EpicSeg	2	3

	Benchmark II & III - 127 cell types and tissues	
Method/segmentation	promoter states	enhancer states
GenoSTAN-Poilog-127	Prom.19, Prom.5	Enh.12, EnhW.9
GenoSTAN-nb-127	Prom.1, Prom.19	Enh.6, EnhW.8
GenoSTAN-Poilog-20	Prom.15, Prom.6	Enh.9, EnhF.13
GenoSTAN-nb-20	Prom.14, Prom.21	Enh.9, EnhF.12
ChromHMM-15	1_TssA, 2_PromU	13_EnhA1, 14_EnhA2
ChromHMM-18	1_TssA, 2_TssFlnk	9_EnhA1, 10_EnhA2
ChromHMM-25	1_TssA, 2_TssAFlnk	7_Enh, 6_EnhG

Appendix Table A2: Table showing promoter and enhancer states used to calculate recall of FANTOM5 promoters and enhancers. Two promoter and enhancer states were used for each segmentation, except for the EpicSeg segmentation, which only fitted one enhancer state.

18 Additional information for section 16

All methods and analyses presented in Appendix Figure A10 and in sections 18.1 and 18.2 were developed and performed by Margaux Michel and Björn Schwalb. They are included in this thesis for the sake clarity and completeness. Results presented in Figure A11 were obtained in collaboration with Margaux Michel and Björn Schwalb. These results are part of the manuscript "TT-Seq captures the human transient transcriptome" which was accepted for publication in Science. For detailed author contributions see page ix.



Appendix Figure A10: TT-Seq enables nearly uniform mapping of the human transient transcriptome. (A) Schematic representation of 4sU-Seq and TT-Seq methods (4sU, 4-thiouridine). (B) Metagene analysis comparing TT-Seq to 4sU-Seq and RNA-Seq. Average coverage in 2,323 TUs (depleted for paused genes as defined in [234]) is shown around the first TSS (left), the 5'-splice site (SS) of an intron of at least 10 kb (middle), and the last pA site (right) relative to the maximum in the first kb from the first TSS.

18.1 Experimental protocol of transient trancriptome sequencing (TT-Seq)

K562 cells were acquired from DSMZ (Braunschweig, Germany). Cells were grown in RPMI 1640 medium (Gibco) supplemented with 10% heat-inactivated FBS (Gibco) and 1% Penicillin/Streptomycin (100x, PAA) at 37°C under 5% CO2. Cells were labeled in media for 5 min with 500 µM 4-thiouridine (4sU, Sigma-Aldrich) and harvested through centrifugation for 2 min at 3,000 rpm. RNA extraction was performed with TRIzol (Life Technologies) following the manufacturers' instructions except for the addition of an RNA spike-in mix together with TRIzol. The purified RNA was split in two samples and one of the two samples was fragmented at 240 ng/µl on a BioRuptor Next Gen (Diagenode) at high power for one cycle of 30"/30" ON/OFF. Fragmented and non-fragmented samples were subjected to labeled RNA purification as previously described [85]. Labeled fragmented (TT-Seq), labeled (4sU-Seq), total (RNA-Seq) and total fragmented (RNA-Seq with fragmentation) RNA were treated with 2 units of DNase Turbo (Life Technologies) and sequencing libraries were prepared with the Ovation Human Blood RNA-Seq library kit (NuGEN) following the manufacturers' instructions. All samples were sequenced on an Illumina HiSeq 1500 sequencer.
Six spike-ins (ERCC-00043, ERCC-00170, ERCC-00136, ERCC-00145, ERCC-00092 and ERCC-00002) from the ERCC RNA spike-in mix (Life Technologies) were chosen as to have the same nucleotide length and U numbers, but with different GC content (40 to 60%). Spike-ins were amplified through PCR with the forward primer containing a T7 promoter sequence. Each spike-in was subjected to in vitro transcription with the Megascript T7 Transcription Kit (Life Technologies) with either 1:10 4sUTP:UTP ratio for spike-ins ERCC-00043, ERCC-00136 and ERCC-00092 or only UTP for spike-ins ERCC-00170, ERCC-00145, ERCC-00002; resulting in labeled and non-labeled spike-ins, respectively. All spike-ins were purified with AMPure XP beads (Beckman-Coulter) and quantified with Nanodrop 2000 (Thermo Scientific), agarose gel and Qubit 3.0 Fluorometer (Life Technologies). Spike-ins were then mixed in equal amount to a final concentration of 6 ng/µl.

18.2 Estimation of RNA synthesis rates and half-lives

Estimation of RNA synthesis rates and half-lives. For all 19,219 classified TUs isoform-independent exonic regions were determined using a model for constitutive exons [235]. Read counts for all features were calculated using HTSeq [236]. To estimate rates of RNA synthesis and degradation, we used a statistical model that describes the read counts k_{ij} for gene i in sample j (TT-Seq or (total cellular) RNA-Seq samples) fragmented or total RNA-Seq samples) by gene-specific amounts of labeled and unlabeled RNA amounts α_i, β_i. The model also includes a parameter L_i for the length of the respective feature, scaling factors σ_j that account for variations in sequencing depth, and cross-contamination rate ϵ_j that models the proportion of unlabeled reads purified in the TT-Seq sample. The expectation of the number of reads k_{ij} was modeled as:

$$E\left(k_{ij}\right) = L_i \cdot \sigma_j \cdot \left(\alpha_i + \epsilon_j \beta_i\right)$$

Note that ϵ_j is set to 1 for the (total cellular) RNA samples. Sequencing depth σ_j and cross-contamination rate ϵ_j were calculated using the spike-ins data, setting $\alpha_i = 0$ and $\beta_i = 1$ for the unlabeled spike-ins and setting $\alpha_i = 1$ and $\beta_i = 0$ for the labeled spike-ins. The model was fitted by maximum likelihood assuming negative binomial distribution with dispersion parameters as calculated by DESeq2 [237]. Having sequencing depth σ_j and cross-contamination rate ϵ_j estimated, the same model was applied to all TUs to provide estimates of the labeled and unlabeled

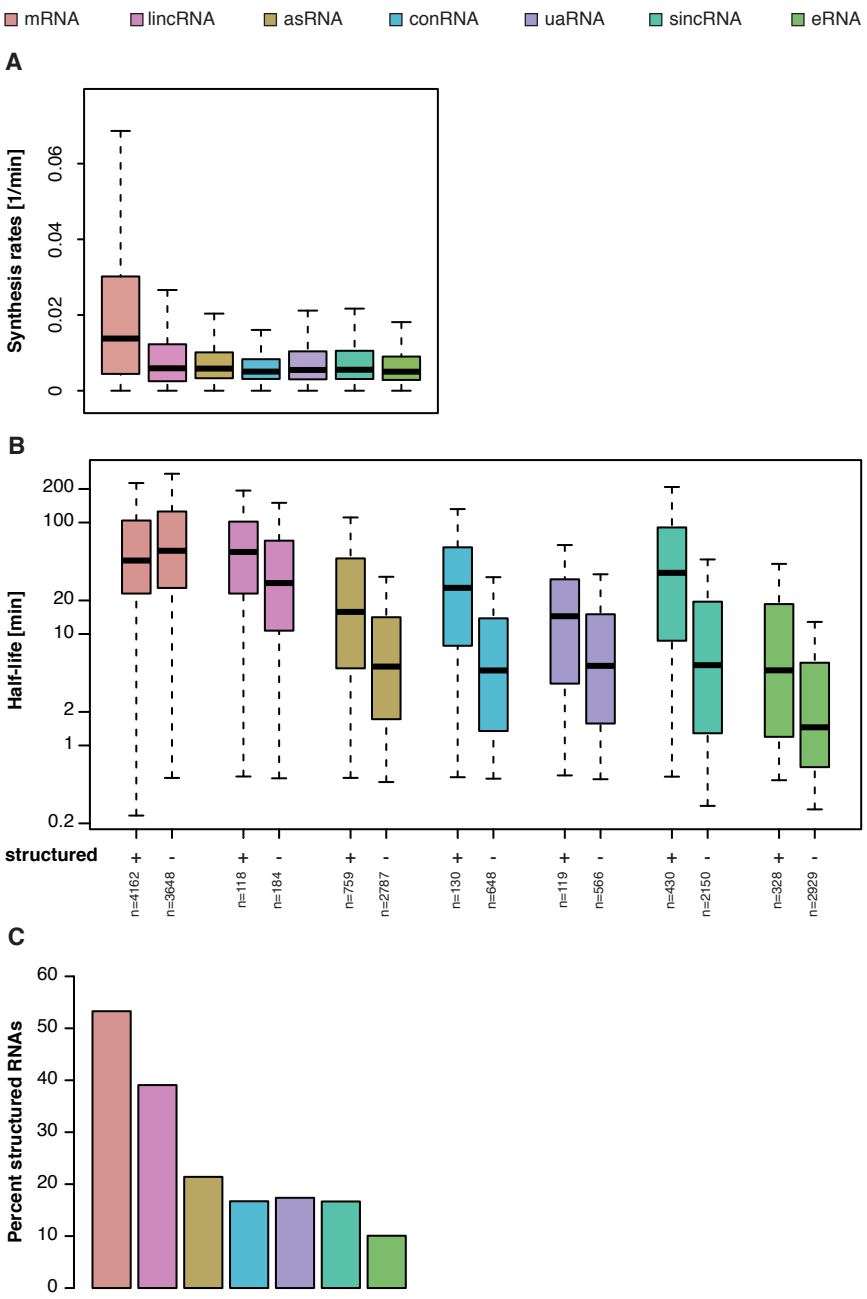
RNA amounts α_i, β_i. These in turn were converted into synthesis and degradation rates (μ_i, λ_i) assuming first-order kinetics as in [86] using the following equations:

$$\begin{aligned} \alpha_i &= \frac{\mu_i}{\lambda_i} \cdot \left(1 - e^{-\lambda_i t}\right) \\ \alpha_i + \beta_i &= \frac{\mu_i}{\lambda_i} \end{aligned}$$

where $t = 5$ minutes. And therefore:

$$\begin{aligned} \lambda_i(t) &= -\frac{1}{t} \log\left(\frac{\beta_i}{\alpha_i + \beta_i}\right) \\ \mu_i(t) &= (\alpha_i + \beta_i) \cdot \lambda_i(t) \end{aligned}$$

Note that this approach is customized for TT-Seq and introduces a conceptually new interpretation of transcript stability because TT-Seq involves the fragmentation of labeled RNAs prior to the purification of their labeled parts. Labeling and modeling approaches that were used so far quantify RNAs as newly synthesized despite the fact that they carry a non-negligible part of pre-existing RNA. This is introducing a bias especially towards longer genes given our short labeling pulse of 5 minutes. Thus this approach can be applied to estimate the local synthesis and degradation rates at any genomic position. When applied to a complete TU, it estimates the typical synthesis rate and half-life of nucleotide bonds within the TU. We think this is necessary given the complexity of the human genes with regard to the vast number of transcript isoforms and the elaborate nature of splicing events that all influence the per gene estimation of synthesis and decay. Note that TT-Seq data did not exhibit the so-called labeling bias.



Appendix Figure A11: Transcript synthesis rates, half-lives, and predicted structure. (A) Distribution of synthesis rates per transcript class. (B) Distribution of half lives of different transcript classes depending on whether they are predicted to be structured or not (+, -). (C) Distribution of percentage of structured RNA in different transcript classes.

References

[1] Crick, F. Central dogma of molecular biology. *Nature* 227, 561–563 (1970).

[2] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. *Molecular Biology of the Cell* (Garland Science, 2002), 4 edn.

[3] Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. & Greenberg, M. E. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182–187 (2010).

[4] Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W. & Steinmetz, L. M. Bidirectional promoters generate pervasive transcription in yeast. *Nature* 457, 1033–1037 (2009).

[5] Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* 10, 833–844 (2009).

[6] Cramer, P., Armache, K. J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., Damsma, G. E., Dengl, S., Geiger, S. R., Jasiak, A. J., Jawhari, A., Jennebach, S., Kamenski, T., Kettenberger, H., Kuhn, C. D., Lehmann, E., Leike, K., Sydow, J. F. & Vannini, A. Structure of eukaryotic RNA polymerases. *Annu Rev Biophys* 37, 337–352 (2008).

[7] Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251 (2013).

[8] Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and disease. *Cell* 136, 777–793 (2009).

[9] Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* 424, 147–151 (2003).

[10] Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27, 299–308 (1981).

[11] Herranz, D., Ambesi-Impiombato, A., Palomero, T., Schnell, S. a., Belver, L., Wendorff, A. a., Xu, L., Castillo-Martin, M., Llobet-Navás, D., Cordon-Cardo, C., Clappier, E., Soulier, J. & Ferrando, A. a. A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nature medicine* 20, 1130–7 (2014).

[12] Webster, N., Jin, J. R., Green, S., Hollis, M. & Chambon, P. The yeast UASG is a transcriptional enhancer in human HeLa cells in the presence of the GAL4 trans-activator. *Cell* 52, 169–178 (1988).

[13] Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. & Lis, J. T. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46, 1311–1320 (2014).

[14] Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* 37, 314–323 (2015).

[15] Hahn, S. & Young, E. T. Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics* 189, 705–736 (2011).

[16] Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* 16, 144–154 (2015).

[17] Serandour, A. A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., Barloy-Hubler, F., Brown, M., Lupien, M., Metivier, R., Salbert, G. & Eeckhoute, J. Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res.* 21, 555–565 (2011).

[18] Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev.* 10, 2657–2683 (1996).

[19] Sikorski, T. W. & Buratowski, S. The basal initiation machinery: beyond the general transcription factors. *Curr. Opin. Cell Biol.* 21, 344–351 (2009).

[20] Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* 16, 155–166 (2015).

[21] Erokhin, M., Vassetzky, Y., Georgiev, P. & Chetverina, D. Eukaryotic enhancers: common features, regulation, and participation in diseases. *Cell. Mol. Life Sci.* 72, 2361–2375 (2015).

[22] Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* 16, 190–202 (2015).

[23] Mandel, C. R., Bai, Y. & Tong, L. Protein factors in pre-mRNA 3'-end processing. *Cell. Mol. Life Sci.* 65, 1099–1122 (2008).

[24] Arigo, J. T., Carroll, K. L., Ames, J. M. & Corden, J. L. Regulation of yeast NRD1 expression by premature transcription termination. *Mol. Cell* 21, 641–651 (2006).

[25] Steinmetz, E. J. & Brow, D. A. Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. *Mol. Cell. Biol.* 16, 6993–7003 (1996).

[26] Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J. & Cramer, P. Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* 155, 1075–1087 (2013).

[27] Hampsey, M. & Kinzy, T. G. Synchronicity: policing multiple aspects of gene expression by Ctk1. *Genes Dev.* 21, 1288–1291 (2007).

[28] Jaehning, J. A. The Paf1 complex: platform or player in RNA polymerase II transcription? *Biochim. Biophys. Acta* 1799, 379–388 (2010).

[29] Stuwe, T., Hothorn, M., Lejeune, E., Rybin, V., Bortfeld, M., Scheffzek, K. & Ladurner, A. G. The FACT Spt16 "peptidase" domain is a histone H3-H4 binding module. *Proc. Natl. Acad. Sci. U.S.A.* 105, 8884–8889 (2008).

[30] Murray, S., Udupa, R., Yao, S., Hartzog, G. & Prelich, G. Phosphorylation of the RNA polymerase II carboxy-terminal domain by the Bur1 cyclin-dependent kinase. *Mol. Cell. Biol.* 21, 4089–4096 (2001).

[31] 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).

[32] Cowling, V. H. Regulation of mRNA cap methylation. *Biochem. J.* 425, 295–302 (2010).

[33] Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Soding, J. & Cramer, P. Uniform transitions of the general RNA polymerase II transcription complex. *Nat. Struct. Mol. Biol.* 17, 1272–1278 (2010).

[34] Zhang, D. W., Rodriguez-Molina, J. B., Tietjen, J. R., Nemec, C. M. & Ansari, A. Z. Emerging Views on the CTD Code. *Genet Res Int* 2012, 347214 (2012).

[35] Cramer, P. Multisubunit RNA polymerases. *Curr. Opin. Struct. Biol.* 12, 89–97 (2002).

[36] Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D. & Cramer, P. CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* 336, 1723–1725 (2012).

[37] Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature* 493, 437–440 (2013).

[38] Kostrewa, D., Zeller, M. E., Armache, K. J., Seizl, M., Leike, K., Thomm, M. & Cramer, P. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* 462, 323–330 (2009).

[39] Luse, D. S. Promoter clearance by RNA polymerase II. *Biochim. Biophys. Acta* 1829, 63–68 (2013).

[40] Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Mol. Cell* 36, 541–546 (2009).

[41] Schwer, B., Mao, X. & Shuman, S. Accelerated mRNA decay in conditional mutants of yeast mRNA capping enzyme. *Nucleic Acids Res.* 26, 2050–2057 (1998).

[42] Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev.* 25, 1770–1782 (2011).

[43] Bienroth, S., Keller, W. & Wahle, E. Assembly of a processive messenger RNA polyadenylation complex. *EMBO J.* 12, 585–594 (1993).

[44] Mischo, H. E. & Proudfoot, N. J. Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. *Biochim. Biophys. Acta* 1829, 174–185 (2013).

[45] Connelly, S. & Manley, J. L. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.* 2, 440–452 (1988).

[46] Rosonina, E., Kaneko, S. & Manley, J. L. Terminating the transcript: breaking up is hard to do. *Genes Dev.* 20, 1050–1056 (2006).

[47] Logan, J., Falck-Pedersen, E., Darnell, J. E. & Shenk, T. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc. Natl. Acad. Sci. U.S.A.* 84, 8306–8310 (1987).

[48] Steinmetz, E. J., Warren, C. L., Kuehner, J. N., Panbehi, B., Ansari, A. Z. & Brow, D. A. Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Mol. Cell* 24, 735–746 (2006).

[49] Carroll, K. L., Pradhan, D. A., Granek, J. A., Clarke, N. D. & Corden, J. L. Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. *Mol. Cell. Biol.* 24, 6241–6252 (2004).

[50] Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S. & Meinhart, A. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat. Struct. Mol. Biol.* 15, 795–804 (2008).

[51] Steinmetz, E. J., Conrad, N. K., Brow, D. A. & Corden, J. L. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* 413, 327–331 (2001).

[52] Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* 20, 267–273 (2013).

[53] Chandy, M., Gutierrez, J. L., Prochasson, P. & Workman, J. L. SWI/SNF displaces SAGA-acetylated nucleosomes. *Eukaryotic Cell* 5, 1738–1747 (2006).

[54] Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R. & Nislow, C. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* 39, 1235–1244 (2007).

[55] Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends Genet.* 27, 389–396 (2011).

[56] Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* 21, 381–395 (2011).

[57] Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2926–2931 (2010).

[58] Tippmann, S. C., Ivanek, R., Gaidatzis, D., Scholer, A., Hoerner, L., van Nimwegen, E., Stadler, P. F., Stadler, M. B. & Schubeler, D. Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol. Syst. Biol.* 8, 593 (2012).

[59] Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* 129, 823–837 (2007).

[60] Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560 (2007).

[61] Fischle, W., Tseng, B. S., Dormann, H. L., Ueberheide, B. M., Garcia, B. A., Shabanowitz, J., Hunt, D. F., Funabiki, H. & Allis, C. D. Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* 438, 1116–1122 (2005).

[62] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318 (2007).

[63] Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* 21, 1273–1283 (2011).

[64] Liang, G., Lin, J. C., Wei, V., Yoo, C., Cheng, J. C., Nguyen, C. T., Weisenberger, D. J., Egger, G., Takai, D., Gonzales, F. A. & Jones, P. A. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7357–7362 (2004).

[65] Mockler, T. C. & Ecker, J. R. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85, 1–15 (2005).

[66] Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R. & Childs, G. Making and reading microarrays. *Nat. Genet.* 21, 15–19 (1999).

[67] Schena, M., Shalon, D., Davis, R. & Brown, P. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* 270, 467–470 (1995).

[68] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O. & Davis, R. W. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A.* 94, 13057–13062 (1997).

[69] Yamada, K. *et al.* Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302, 842–846 (2003).

[70] David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. & Steinmetz, L. M. A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5320–5325 (2006).

[71] Gilmour, D. S. & Lis, J. T. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl. Acad. Sci. U.S.A.* 81, 4275–4279 (1984).

[72] Aparicio, O., Geisberg, J. V., Sekinger, E., Yang, A., Moqtaderi, Z. & Struhl, K. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Mol Biol* Chapter 21, Unit 21.3 (2005).

[73] Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. & Brown, P. O. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409, 533–538 (2001).

[74] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. & Young, R. A. Genome-wide location and function of DNA binding proteins. *Science* 290, 2306–2309 (2000).

[75] Brown, P. O., Lieb, J. D., Botstein, D., Liu, X., Botstein, D. & Brown, P. O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics* 28, 327–334 (2001).

[76] Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298, 799–804 (2002).

[77] Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E. & Young, R. A. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104 (2004).

[78] 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).

[79] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–381 (2005).

[80] Bentley, D. R. *et al.* Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* 456, 53–59 (2009).

[81] Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502 (2007).

[82] Ho, J. W., Bishop, E., Karchenko, P. V., Negre, N., White, K. P. & Park, P. J. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* 12, 134 (2011).

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

[84] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628 (2008).

[85] Dölken, L., Ruzsics, Z., Radle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P. & Koszinowski, U. H. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* 14, 1959–1972 (2008).

[86] Miller, C., Schwalb, B., Maier, K., Schulz, D., Dumcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dolken, L., Martin, D. E., Tresch, A. & Cramer, P. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* 7, 458 (2011).

[87] Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. & Regev, A. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436–442 (2011).

[88] Keene, M. A., Corces, V., Lowenhaupt, K. & Elgin, S. C. DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc. Natl. Acad. Sci. U.S.A.* 78, 143–146 (1981).

[89] McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D. & Felsenfeld, G. A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. *Cell* 27, 45–55 (1981).

[90] Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010, pdb.prot5384 (2010).

[91] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. & Crawford, G. E. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322 (2008).

[92] Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* 489, 75–82 (2012).

[93] Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).

[94] Koohy, H., Down, T. A. & Hubbard, T. J. Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS ONE* 8, e69853 (2013).

[95] Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S. & Thomson, J. A. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048 (2010).

[96] The ENCODE Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640 (2004).

[97] Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073 (2010).

[98] Haussler, D. *et al.* Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* 100, 659–674 (2009).

[99] Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816 (2007).

[100] Myers, R. M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9, e1001046 (2011).

[101] Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A. & Noble, W. S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* 9, 473–476 (2012).

[102] Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C., Dunham, I., Kellis, M. & Noble, W. S. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 41, 827–841 (2013).

[103] Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49 (2011).

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

[105] Niu, D. K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* 430, 1340–1343 (2013).

[106] Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A. & Elhaik, E. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5, 578–590 (2013).

[107] Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343 (2015).

[108] Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336 (2015).

[109] Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L. H. & Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 518, 365–369 (2015).

[110] Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).

[111] Yen, A. & Kellis, M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat Commun* 6, 7973 (2015).

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

[113] Zacher, B., Lidschreiber, M., Cramer, P., Gagneur, J. & Tresch, A. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Mol. Syst. Biol.* 10, 768 (2014).

[114] Krogh, A., Mian, I. S. & Haussler, D. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res.* 22, 4768–4778 (1994).

[115] Thomas, A. & Skolnick, M. H. A probabilistic model for detecting coding regions in DNA sequences. *IMA J Math Appl Med Biol* 11, 149–160 (1994).

[116] Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 26, 1107–1115 (1998).

[117] Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie–gene finding in Drosophila melanogaster. *Genome Res.* 10, 529–538 (2000).

[118] Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* 33, 6494–6506 (2005).

[119] Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 Suppl 2, i215–225 (2003).

[120] Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32, 767–769 (2015).

[121] Testa, A. C., Hane, J. K., Ellwood, S. R. & Oliver, R. P. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* 16, 170 (2015).

[122] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W. & Haussler, D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050 (2005).

[123] Felsenstein, J. & Churchill, G. A. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13, 93–104 (1996).

[124] Wong, W. S. & Nielsen, R. Finding cis-regulatory modules in Drosophila using phylogenetic hidden Markov models. *Bioinformatics* 23, 2031–2037 (2007).

[125] Sinha, S. & He, X. MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. *PLoS Comput. Biol.* 3, e216 (2007).

[126] Bailey, T. L. & Noble, W. S. Searching for statistically significant regulatory modules. *Bioinformatics* 19 Suppl 2, 16–25 (2003).

[127] Drawid, A., Gupta, N., Nagaraj, V. H., Gelinas, C. & Sengupta, A. M. OHMM: a Hidden Markov Model accurately predicting the occupancy of a transcription factor with a self-overlapping binding motif. *BMC Bioinformatics* 10, 208 (2009).

[128] Conkright, M. D., Guzman, E., Flechner, L., Su, A. I., Hogenesch, J. B. & Montminy, M. Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol. Cell* 11, 1101–1108 (2003).

[129] Frith, M. C., Hansen, U. & Weng, Z. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* 17, 878–889 (2001).

[130] Maaskola, J. & Rajewsky, N. Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res.* 42, 12995–13011 (2014).

[131] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–285 (2016).

[132] Stegmaier, P., Kel, A. E. & Wingender, E. Systematic DNA-binding domain classification of transcription factors. *Genome Inform* 15, 276–286 (2004).

[133] Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960 (2005).

[134] Asai, K., Hayamizu, S. & Handa, K. Prediction of protein secondary structure by the hidden Markov model. *Comput. Appl. Biosci.* 9, 141–146 (1993).

[135] Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580 (2001).

[136] Du, J., Rozowsky, J. S., Korbel, J. O., Zhang, Z. D., Royce, T. E., Schultz, M. H., Snyder, M. & Gerstein, M. A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics* 22, 3016–3024 (2006).

[137] Humburg, P., Bulger, D. & Stone, G. Parameter estimation for robust HMM analysis of ChIP-chip data. *BMC Bioinformatics* 9, 343 (2008).

[138] Xu, H., Wei, C. L., Lin, F. & Sung, W. K. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 24, 2344–2349 (2008).

[139] Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., Yu, J. & Chinnaiyan, A. M. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* 11, 369 (2010).

[140] Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* 16, 222 (2015).

[141] Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K. & Schubeler, D. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480, 490–495 (2011).

[142] Burger, L., Gaidatzis, D., Schubeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* 41, e155 (2013).

[143] Majoros, W. H., Lekprasert, P., Mukherjee, N., Skalsky, R. L., Corcoran, D. L., Cullen, B. R. & Ohler, U. MicroRNA target site identification by integrating sequence and binding information. *Nat. Methods* 10, 630–633 (2013).

[144] Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H. & Bucan, M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674 (2007).

[145] Love, M. I., Mysickova, A., Sun, R., Kalscheuer, V., Vingron, M. & Haas, S. A. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol* 10, 52 (2011).

[146] Daemen, A., Gevaert, O., Leunen, K., Legius, E., Vergote, I. & De Moor, B. Supervised classification of array CGH data with HMM-based feature selection. *Pac Symp Biocomput* 468–479 (2009).

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

[148] 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).

[149] Jaschek, R. & Tanay, A. *Research in Computational Molecular Biology: 13th Annual International Conference, RECOMB 2009, Tucson, AZ, USA, May 18-21, 2009. Proceedings*, chap. Spatial Clustering of Multivariate Genomic and Epigenomic Information, 170–183 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009).

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

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

[152] Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* 33, 364–376 (2015).

[153] 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).

[154] Mammana, A. & Chung, H. R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.* 16, 151 (2015).

[155] Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol.* 16, 56 (2015).

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

[157] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the The Royal Statistical Society, Series B* 39, 1–38 (1977).

[158] Cook, J. D. Notes on the negative binomial distribution http://www.johndcook.com/negative_binomial.pdf [Accessed: 2016-03-04].

[159] Bulmer, M. G. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics* 30, 101–110 (2011).

[160] Grotan, V. & Engen, S. *poilog: Poisson lognormal and bivariate Poisson lognormal distribution* (2008). R package version 0.4.

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

[162] Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M. & Jaffrezic, F. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics* 14, 671–683 (2013).

[163] Forney, G. D. The viterbi algorithm. *Proceedings of the IEEE* 61, 268–278 (1973).

[164] Seneta, E. *Non-negative matrices and Markov chains; rev. version.* Springer series in statistics (Springer, New York, NY, 2006).

[165] 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).

[166] Ghalanos, A. & Theussl, S. Rsolnp: General non-linear optimization using augmented lagrange multiplier method. R package version 1.14 (2012).

[167] 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).

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

[169] 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).

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

[171] 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).

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

[173] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. & Botstein, D. SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* 26, 73–79 (1998).

[174] 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).

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

[176] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359 (2012).

[177] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).

[178] Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics* 25, 1841–1842 (2009).

[179] Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9, e1003118 (2013).

[180] Ernst, J. & Kellis, M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.* 23, 1142–1154 (2013).

[181] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–1006 (2014).

[182] Lorenz, R., Bernhart, S. H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F. & Hofacker, I. L. ViennaRNA Package 2.0. *Algorithms Mol Biol* 6, 26 (2011).

[183] Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A. & Stadler, P. F. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23, 1383–1390 (2005).

[184] Bataille, A. R., Jeronimo, C., Jacques, P. E., Laramee, L., Fortin, M. E., Forest, A., Bergeron, M., Hanes, S. D. & Robert, F. 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).

[185] Roy, S. *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330, 1787–1797 (2010).

[186] Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J. & van Steensel, B. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* 143, 212–224 (2010).

[187] Fariselli, P., Martelli, P. L. & Casadio, R. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics* 6 Suppl 4, S12 (2005).

[188] Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I. & Pugh, B. F. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* 18, 1073–1083 (2008).

[189] Fan, X., Moqtaderi, Z., Jin, Y., Zhang, Y., Liu, X. S. & Struhl, K. 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).

[190] Guo, Z. & Sherman, F. 3'-end-forming signals of yeast mRNA. *Trends Biochem. Sci.* 21, 477–481 (1996).

[191] Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* 13, 233–245 (2012).

[192] Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* 24, 1595–1602 (2014).

[193] Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S. & Kellis, M. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* 23, 800–811 (2013).

[194] Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077 (2013).

[195] Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286 (2014).

[196] Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinformatics* [Epub ahead of print] (2015).

[197] Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M. & Gerstein, M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* 13, R48 (2012).

[198] The blueprint project. http://www.blueprint-epigenome.eu/ [Accessed: 2016-03-04].

[199] Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461 (2014).

[200] Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* 507, 462–470 (2014).

[201] Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* 43, e6 (2015).

[202] Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* 21, 2167–2180 (2011).

[203] Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M. & Ren, B. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.* 9, e1002968 (2013).

[204] Won, K. J., Zhang, X., Wang, T., Ding, B., Raha, D., Snyder, M., Ren, B. & Wang, W. Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res.* 41, 4423–4432 (2013).

[205] Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* 471, 480–485 (2011).

[206] May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Afzal, V., Simpson, P. C., Rubin, E. M., Black, B. L., Bristow, J., Pennacchio, L. A. & Visel, A. Large-scale discovery of enhancers from human heart tissue. *Nat. Genet.* 44, 89–93 (2012).

[207] Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M. & Pennacchio, L. A. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854–858 (2009).

[208] Kvon, E. Z., Stampfel, G., Yanez-Cuna, J. O., Dickson, B. J. & Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.* 26, 908–913 (2012).

[209] Li, H., Chen, H., Liu, F., Ren, C., Wang, S., Bo, X. & Shu, W. Functional annotation of HOT regions in the human genome: implications for human disease and cancer. *Sci Rep* 5, 11633 (2015).

[210] Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* 46, 1160–1165 (2014).

[211] Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 44, D877–881 (2016).

[212] Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S. & Raychaudhuri, S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130 (2013).

[213] Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482 (2011).

[214] Ramskold, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* 5, e1000598 (2009).

[215] Gerstein, M. B., Kundaje, A., Hariharan, M., Weissman, S. M. & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100 (2012).

[216] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M. & Weng, Z. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22, 1798–1812 (2012).

[217] Org, T., Duan, D., Ferrari, R., Montel-Hagen, A., Van Handel, B., Kerenyi, M. A., Sasidharan, R., Rubbi, L., Fujiwara, Y., Pellegrini, M., Orkin, S. H., Kurdistani, S. K. & Mikkola, H. K. Scl binds to primed enhancers in mesoderm to regulate hematopoietic and cardiac fate divergence. *EMBO J.* 34, 759–777 (2015).

[218] Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., Hacohen, N., Schier, A. F., Blackshear, P. J., Friedman, N., Amit, I. & Regev, A. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* 159, 1698–1710 (2014).

[219] Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469, 368–373 (2011).

[220] Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J., Gifford, D. K. & Sherwood, R. I. High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* 34, 167–174 (2016).

[221] Jensen, T. H., Jacquier, A. & Libri, D. Dealing with pervasive transcription. *Mol. Cell* 52, 473–484 (2013).

[222] Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R. & Boothroyd, J. C. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.* 23, 232–237 (2005).

[223] Flynn, R. A., Almada, A. E., Zamudio, J. R. & Sharp, P. A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10460–10465 (2011).

[224] Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A. & Churchman, L. S. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* 161, 541–554 (2015).

[225] Khorkova, O., Myers, A. J., Hsiao, J. & Wahlestedt, C. Natural antisense transcripts. *Hum. Mol. Genet.* 23, 54–63 (2014).

[226] Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* 489, 101–108 (2012).

[227] Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* 46, 1–19 (2012).

[228] Ping, Y. H. & Rana, T. M. DSIF and NELF interact with RNA polymerase II elongation complex and HIV-1 Tat stimulates P-TEFb-mediated phosphorylation of RNA polymerase II and DSIF during transcription elongation. *J. Biol. Chem.* 276, 12951–12958 (2001).

[229] Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L. & Dreyfuss, G. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 468, 664–668 (2010).

[230] Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L. & Dreyfuss, G. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 150, 53–64 (2012).

[231] Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* 499, 360–363 (2013).

[232] Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* 39, 170–182 (2014).

[233] Pnueli, L., Rudnizky, S., Yosefzon, Y. & Melamed, P. RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin α-subunit gene. *Proc. Natl. Acad. Sci. U.S.A.* 112, 4369–4374 (2015).

[234] Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322, 1845–1848 (2008).

[235] Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11, 94 (2010).

[236] Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169 (2015).

[237] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).

List of Figures

List of Tables