bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Benedikt Zacher[1,*], Margaux Michel[4], Björn Schwalb[4], Patrick Cramer[4], Achim Tresch[2,3,**],Julien Gagneur[1,5,***]

February 24, 2016

1. Feodor-Lynen-Str. 25, Munich, Germany, Gene Center and Department of Biochemistry, Center for Integrated Protein Science CIPSM, Ludwig-Maximilians-Universität München.
2. Zülpicher Str. 47, Cologne, Germany, Department of Biology, University of Cologne.
3. Carl-von-Linne-Weg 10, 24105 Cologne, Germany, Max Planck Institute for Plant Breeding Research.
4. Am Fassberg 11, Göttingen, Germany, Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry
5. Present address: Technische Universität München, Department of Informatics, Boltzmannstr. 3, 85748 Garching, Germany

*Corresponding Author: Benedikt Zacher, Tel.: +49-89-2180-76740; Fax: +49-89-2180-76797; E-mail: zacher@genzentrum.lmu.de

**Corresponding Author: Achim Tresch, Tel.: +49-221-470-8292; Fax: +49-221-5062-163; E-mail: tresch@mpipz.mpg.de

*** Corresponding Author: Julien Gagneur, Tel.: +49-89-2891-19411; Fax: +49-89-2891-19414; E-mail: gagneur@in.tum.de

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Accurate maps of promoters and enhancers are required for understanding transcriptional regulation. Promoters and enhancers are usually mapped by integration of chromatin assays charting histone modifications, DNA accessibility, and transcription factor binding. However, current algorithms are limited by unrealistic data distribution assumptions. Here we propose GenoSTAN (Genomic STate ANnotation), a hidden Markov model overcoming these limitations. We map promoters and enhancers for 127 cell types and tissues from the ENCODE and Roadmap Epigenomics projects, today's largest compendium of chromatin assays. Extensive benchmarks demonstrate that GenoSTAN consistently identifies promoters and enhancers with significantly higher accuracy than previous methods. Moreover, GenoSTAN-derived promoters and enhancers showed significantly higher enrichment of complex trait-associated genetic variants than current annotations. Altogether, GenoSTAN provides an easy-to-use tool to define promoters and enhancers in any system, and our annotation of human transcriptional cis-regulatory elements constitutes a rich resource for future research in biology and medicine.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

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) [3]. 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 [8]. 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 [35, 30, 7].

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 [53, 6, 17, 26, 33]. 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 [13, 61], NIH Roadmap Epigenomics [12], BLUEPRINT [1] and FANTOM [5, 14].

In this context, the computational approaches employed to classify genomic regions as enhancers or promoters play a decisive role [53, 33]. 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. [32, 39, 51, 60]). 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 [17, 20, 27, 26, 64, 29, 12, 13]. 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 [49]. However, these methods differ in the way the distribution of ChIP-seq signals for each chromatin state is modeled. ChromHMM [17, 18, 20], 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 [2]. Segway [26, 27], 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 [43] 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 [43]. 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

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[13], Segway 38,922 enhancers [13], and EpicSeg 53,982 enhancers [43]. 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.

Results

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. Specifically, 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). 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 [62], 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 1A).

We performed an extensive benchmarking of GenoSTAN against alternative methods (Figure 1B). 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 [13]. 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 1B). Benchmark III

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 1B lists all model names and studies that were considered.

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). As pointed out by others [17, 26], 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 [17, 26, 62]. 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). Figure 2A 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 [43, 27, 13, 25].

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 [43, 20, 27]. The median read coverage in state segments and genomic distributions were very similar for both the Poilog-K562 and the nb-K562 models (Figure 2B, Supplementary Figure 1). 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 [2, 45, 55]. Promoter (Prom) states were located close to annotated GENCODE TSSs [24], 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 2B, Supplementary Figure 1). Promoter and enhancer states also differed in their DNA sequence features. 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 2B, Supplementary Figure 1). 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 [16]. 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,

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 2B, Supplementary Figure 1). 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.

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' [13, 27] and 'ChromHMM-Nature' [20]), Segway ('Segway-ENCODE' [13, 27], 'Segway-nmeth' [26] and 'Segway-Reg.Build' [64]) and EpicSeg [43]. 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 (Supplementary Figure 2). 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 [32]. The number of enhancer segments ranged from 10,932 segments in GenoSTAN-Poilog-K562 to 80,043 segments in one Segway annotation [26] (Supplementary Table 1). Therefore, a thorough assessment of these predictions was necessary to provide a robust and accurate prediction of these elements.

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 [16]), of transcription factor binding (ENCODE high occupancy target, or HOT regions [61], and ENCODE TF binding sites [13]), and of cis-regulatory activity (enhancer activity assessed by reporter assays [30]), 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 [31, 5, 14, 16]. To benchmark the predictions using evidence for transcription, we used published data from a protocol called GRO-cap [16], 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 3A). GenoSTAN-Poilog-K562 had the highest recall and the lowest FDR (Methods, Figure 3A). 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 [27, 13], ChromHMM-Nature [20]) 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GRO-cap is a very sensitive method that captures also a large amount of TSSs of unstable transcripts. 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 [61], which were shown to function as enhancers [34] and are enriched in disease- and trait-associated genetic variants [40]. 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 3B). 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 3C). 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 [14] and enhancers [5] 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 2A). 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 [5], whereas FANTOM5 promoters were defined as regions where transcription was biased towards one direction. The FANTOM5 annotation of enhancers and 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 [16]. 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 [13] and GRO-cap TSSs [16]. 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 3D, Supplementary Table 2). 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 21% FANTOM5 promoters. In accordance with this, many promoter regions were erroneously classified as enhancer regions in this segmentation (e.g. TAL1 promoter in Figure 2A). The recall of FANTOM5 enhancers by promoter states was generally higher (17% - 37%). GenoSTAN-Poilog-K562 and

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

-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 3D). 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 2A). 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 [30]. 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 3E). 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 3F), including the original study (p-value < 0.01, ChromHMM-nature enhancers) by Kheradpour et al. [30]. This analyis shows that GenoSTAN has higher succes rate in predicting *in vivo* enhancer activity than previous methods.

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. Both GenoSTAN-Poilog-K562 and GenoSTAN-nb-K562 had a lower FDR at a similar or higher recall than all three other methods (Supplementary Figure 3). 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 (Supplementary Figure 3). 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 (Supplementary Figure 3). 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) 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 (Supplementary Fig. 4A, 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

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (Supplementary Fig. 4C, D and Methods). 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.

Benchmarks 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 [12] (Supplementary Figure 5, Benchmark II, Figure 1B 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 (Supplementary Figure 6, Benchmark III, Figure 1B 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 [13]. 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) [12, 19]. 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 4A, assessed for all 127 cell types and tissues), of GRO-cap TSSs (Figure 4B assessed for the cell types with available GRO-Cap TSSs), and of HOT regions (Figure 4C, assessed for the cell types with available HOT regions). Moreover, both GenoSTAN models distinguished better promoters from enhancers than previous annotations (Figure 4D, Supplementary Table 2). The low accuracy of ChromHMM-15 and ChromHMM-18 promoters might be caused by frequent state switching between the promoter and promoter flanking state (Supplementary Figure 7). 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 (Supplementary Table 1). 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 (Supplementary Table 1). Although GenoSTAN predicted more enhancers than the ChromHMM-25 model, the fraction of putative enhancers bound by individual TFs was greater (Figure 4E). 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 4E). Collectively, these results show that the improved performance of GenoSTAN is not restricted to the K562 dataset.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [12, 20, 58, 57, 54, 42] 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 1B) with phenotype-associated genetic variants from the NHGRI genome-wide assocation studies catalog (NHGRI GWAS Catalog [59]). 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) 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 (Supplementary Figure 8). 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), enhancers of all GenoSTAN models overlaped a higher fraction of GWAS variants at a similar to better per base pair density compared to the current ChromHMM annotations (Figure 5A). The same trend was observed for promoters when restricting to 1% of genomic coverage (Figure 5B). 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 [12, 20] we found that individual variants were strongly enriched in enhancer or promoter states specifically active in the relevant cell types or tissues (Figure 5C, Supplementary Figure 8C). Variants associated with height were significantly associated 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 5C) and promoters (Supplementary Figure 8C). 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.

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 is available as Supplementary Data. All individual chromatin state annotations are available at http://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN. For the combined anno-

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tation 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, [52]). 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 2, Supplementary Figure 6). 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 (Supplementary Figure 9) 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.

Promoters and enhancers have a distinct TF regulatory landscape

The biochemical distinction between enhancers and promoters is a topic of debate [53, 6]. 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 [13]. Nine TF modules were defined by clustering based on binding pattern similarity across enhancers and promoters (Methods, Figure 6). These 9 TF modules were further characterized by the propensity of their TFs to bind promoters, enhancers or both (Figure 6). In accordance with previous studies [4, 22], 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 [5, 6]. In accordance with the recent finding of widespread transcription at enhancers [16], Pol II and multifunctional TFs Myc, Max, and MAZ [56] 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 6).

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 [22, 4]. 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 6). 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

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

elongation (TBP, TAF1, CCNT2, SP1) and other TFs involved in the regulation of specific gene classes (e.g. cell cycle: E2F4) [56]. 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 [22, 47]. It also contained the enhancer-specific transcription factor P300 [55] and transcriptional activators (e.g. ATF1) and repressors (e.g. HDAC2, REST) [56]. 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.

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) [43, 17, 26]. GenoSTAN is released as part of the open-source R/Bioconductor package STAN [62, 28, 21], which provides a fast, multiprocessing implementation that can process data from 127 human cell types in less 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 [13, 12]. Binding of enhancer-associated co-activator CBP and histone acetyltransferase P300 was used by several studies for the genome-wide prediction of enhancers [55, 45, 2]. From these predictions a distinctive chromatin signature for promoters and enhancers was derived based on H3K4me1 and H3K4me3 [2]. 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 [53, 6]. Among other characteristics, a shared architecture of promoters and enhancers was proposed based on the recent discovery of widespread bidirectional transcription at enhancers [31, 6, 16]. This was supported by the observation that enhancers, which are depleted in CpG islands have similar transcription factor (TF) motif enrichments as CpG poor promoters [5]. However, another study showed that TF co-occurrence differed between gene-proximal and gene-distal sites [22, 4]. 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

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [5]. 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 [48, 46] or nascent transcriptome profiling [11], the chromatin state data undoubtedly add valuable information for the indentification 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 [62].

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 [7]. 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* [7]. Other experimental assays for the validation of predicted ENCODE enhancers lead to different results [35, 30]. 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 [50]. 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

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] The blueprint project. http://www.blueprint-epigenome.eu/.

[2] Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–8, 2007.

[3] Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*, 13(4):233–245, 2012.

[4] Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Research*, 23:1142–1154, 2013.

[5] R. Andersson, C. Gebhard, and I. et al. Miguel-Escalada. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, Mar 2014.

[6] Robin Andersson. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, 37(3):314–323, 2015.

[7] Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Lukasz M Boryn, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, 339(6123):1074–1077, 2013.

[8] J Banerji, S Rusconi, and W Schaffner. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308, 1981.

[9] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi,

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. Haidar, Y. Yu, Y. Ruan, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[10] M. G. Bulmer. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics*, 30(1):101–110, 2011.

[11] L. S. Churchman and J. S. Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373, Jan 2011.

[12] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[13] The ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[14] The Fantom Consortium. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, 2014.

[15] John D Cook. Notes on the Negative Binomial Distribution, 2009.

[16] Leighton J. Core, André L. Martins, Charles G. Danko, Colin T Waters, Adam Siepel, and John T. Lis. Analysis of transcription start sites from nascent RNA supports a unified architecture of mammalian promoters and enhancers. *Submitted*, 46(12):1311–1320, 2014.

[17] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8):817–825, 2010.

[18] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[19] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, 2015.

[20] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[21] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, 2004.

[22] Mark B. Gerstein, Anshul Kundaje, Manoj Hariharan, Sherman M. Weissman, and Michael Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.

[23] Vidar GrÃžtan and Steinar Engen. *poilog: Poisson lognormal and bivariate Poisson lognormal distribution*, 2008. R package version 0.4.

[24] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, 22(9):1760–1774, Sep 2012.

[25] Sven Heinz, Casey E Romanoski, Christopher Benner, and Christopher K Glass. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3):144–154, 2015.

[26] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff a Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012.

[27] Michael M. Hoffman, Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, Paul M. Ellenbogen, Jeffrey a. Bilmes, Ewan Birney, Ross C. Hardison, Ian Dunham, Manolis Kellis, and William Stafford Noble. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–841, 2013.

[28] Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *J. Comp. Graph. Stat.*, 5(3):299–314, 1996.

[29] Peter V Kharchenko, Artyom a Alekseyenko, Yuri B Schwartz, Aki Minoda, Nicole C Riddle, Jason Ernst, Peter J Sabo, Erica Larschan, Andrey a Gorchakov, Tingting Gu, Daniela Linder-Basso, Annette Plachetka, Gregory Shanower, Michael Y Tolstorukov, Lovelace J Luquette, Ruibin Xi, Youngsook L Jung, Richard W Park, Eric P Bishop, Theresa K Canfield, Richard Sandstrom, Robert E Thurman, David M MacAlpine, John a Stamatoyannopoulos, Manolis Kellis, Sarah C R Elgin, Mitzi I Kuroda, Vincenzo Pirrotta, Gary H Karpen,

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and Peter J Park. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, 471(7339):480–485, 2011.

[30] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S. Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23:800–811, 2013.

[31] Tae-Kyung Kim, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F. Worley, Gabriel Kreiman, and Michael E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 2010.

[32] D. Kleftogiannis, P. Kalnis, and V. B. Bajic. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Research*, 43(1):e6–e6, 2015.

[33] Dimitrios Kleftogiannis, Panos Kalnis, and Vladimir B Bajic. Progress and challenges in bioinformatics approaches for enhancer identification. *Briefings in Bioinformatics*, (August):1–13, 2015.

[34] E. Z. Kvon, G. Stampfel, J. O. Yanez-Cuna, B. J. Dickson, and A. Stark. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.*, 26(9):908–913, May 2012.

[35] J. C. Kwasnieski, C. Fiore, H. G. Chaudhari, and B. a. Cohen. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, pages gr.173518.114–, 2014.

[36] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Apr 2012.

[37] Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842, 2009.

[38] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin Morgan, and Vincent Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.

[39] Dongwon Lee, Rachel Karchin, and Michael A Beer. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome research*, 21(12):2167–80, 2011.

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

[41] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[42] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alfoldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke, Brian J. Raney,

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Andrew Cree, Huyen H. Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R. Lewis, Lynne V. Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, Erica Sodergren, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[43] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, 16(1):151, 2015.

[44] Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Nicolas Servant, Luc Jouneau, Denis Laloe, Caroline Le Gall, and Brigitte Schae. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. 14(6), 2012.

[45] Dalit May, Matthew J Blow, Tommy Kaplan, David J McCulley, Brian C Jensen, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Veena Afzal, Paul C Simpson, Edward M Rubin, Brian L Black, James Bristow, Len A Pennacchio, and Axel Visel. Large-scale discovery of enhancers from human heart tissue. *Nature genetics*, 44(1):89–93, 2012.

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

[47] Tõnis Org, Dan Duan, Roberto Ferrari, Amelie Montel-hagen, Ben Van Handel, A Marc, Rajkumar Sasidharan, Liudmilla Rubbi, Yuko Fujiwara, Matteo Pellegrini, Stuart H Orkin, Siavash K Kurdistani, and Hanna K A Mikkola. Scl binds to primed enhancers in mesoderm to regulate hematopoietic and cardiac fate divergence. (January):1–20, 2015.

[48] M. Rabani, R. Raychowdhury, M. Jovanovic, M. Rooney, D. J. Stumpo, A. Pauli, N. Hacohen, A. F. Schier, P. J. Blackshear, N. Friedman, I. Amit, and A. Regev. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, 159(7):1698–1710, Dec 2014.

[49] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.

[50] N. Rajagopal, S. Srinivasan, K. Kooshesh, Y. Guo, M. D. Edwards, B. Banerjee, T. Syed, B. J. Emons, D. K. Gifford, and R. I. Sherwood. High-throughput mapping of regulatory DNA. *Nat. Biotechnol.*, 34(2):167–174, Feb 2016.

[51] Nisha Rajagopal, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, and Bing Ren. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS computational biology*, 9(3):e1002968, 2013.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[52] D. Ramskold, E. T. Wang, C. B. Burge, and R. Sandberg. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, 5(12):e1000598, Dec 2009.

[53] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4):272–86, 2014.

[54] Gosia Trynka, Cynthia Sandor, Buhm Han, Han Xu, Barbara E Stranger, X Shirley Liu, and Soumya Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130, 2012.

[55] Axel Visel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M. Rubin, and Len A. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, 2009.

[56] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, 22(9):1798–1812, Sep 2012.

[57] Lucas D. Ward and Manolis Kellis. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Research*, page gkv1340, 2015.

[58] Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics*, 46(11):1160–5, 2014.

[59] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006, 2014.

[60] Kyoung-Jae Won, Xian Zhang, Tao Wang, Bo Ding, Debasish Raha, Michael Snyder, Bing Ren, and Wei Wang. Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic acids research*, 41(8):4423–32, 2013.

[61] Kevin Y Yip, Chao Cheng, Nitin Bhardwaj, James B Brown, Jing Leng, Anshul Kundaje, Joel Rozowsky, Ewan Birney, Peter Bickel, Michael Snyder, and Mark Gerstein. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 13(9):R48, 2012.

[62] B. Zacher, M. Lidschreiber, P. Cramer, J. Gagneur, and A. Tresch. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Molecular Systems Biology*, 10(12):768–768, 2014.

[63] Gabriel E Zentner, Paul J Tesar, and Peter C Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–83, 2011.

[64] Daniel R Zerbino, Steven P Wilder, Nathan Johnson, Thomas Juettemann, and Paul R Flicek. The Ensembl Regulatory Build. *Genome Biology*, 16(1):1–8, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Availability of GenoSTAN and chromatin state annotations

GenoSTAN is freely available from http://bioconductor.org/ as part of our prevsiouy published R/Bioconductor package STAN [62]. Chromatin state annotations for benchmark I, II and III can be downloaded from http://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN. The combined promoter and enhancer annotation is available as Supplementary Data.

Motivation of 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 [15]:

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

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 [10]:

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

A closed form solution for this distribution does not exist. Thus numerical integration is needed to calculate probabilities, which is done in GenoSTAN by using the R package poilog [23, 28].

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Optimization of Poisson-lognormal and negative binomial emissions

Let $\mathcal{O} = (o_0, ..., o_T)$, $o_t = (o_{t,d})_{d \in \mathcal{D}} \in \mathbb{N}_0^{\mathcal{D}}$ be an obsevational sequence of $|\mathcal{D}|$-dimensional count vectors o_t. An HMM assumes that each observation o_t is *emitted* by a corresponding hidden (unobserved) variable s_t, $t = 0, ..., T$. A hidden variable can assume values from a finite set of states $\mathcal{K}$. Each state $k \in \mathcal{K}$ is associated to an emission distribution ψ_k, which defines the probability of making a certain observation, $\psi_k(o_t)$. GenoSTAN assumes 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 [49]. The optimization problem for the paramters of a single emission distribution $\psi_{i,d}$ can be written as

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

where $\Pr(s_t = i \mid \mathcal{O})$ 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 [43], 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:

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

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 $|\mathcal{C}_d| \ll T$.

Correction for library size

The sequencing depth can be very different between experiments. GenoSTAN addresses this problem by using pre-computed scaling factors to correct for varying sequencing depths for a data track between cell types. In this work, the 'total count' method is used [44]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Model 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 [49]. 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 [17]. 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.

Data preprocessing

Benchmark I (K562 ENCODE) sequencing data was mapped to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium) using Bowtie 2.1.0 [36]. Samtools [41] was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped. To obtain 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 [21]. 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 [12]. 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 [21]. The genome was partitioned into 200bp bins and reads were counted within each bin.

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 [9]) 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.

Model fitting of ChromHMM, Segway and EpicSeg

The data was binarized as described in [17] and ChromHMM was fitted with default parameters. Before applying Segway, the data was transformed using the hyperbolic sine function [26] 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 [37]. Overlap of state annotations with external data was calculated with GenomicRanges [38].

Computation of area under curve

AUC values were calculated on Benchmark set 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 (TSSs) 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.

Analysis of transcription factor (co-)binding

Enrichment of TFs in chromatin states was calculated as described earlier [4]. The TF co-binding rate was calculated as the Jaccard-Index: $\frac{|S_{TF} \cap S_i|}{|S_{TF} \cup S_i|}$, where S_{TF} is the set of TF binding sites and S_i is the set genomic regions that are annotated with state i.

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

The GWAS catalog was obtained from the gwascat package from Bioconductor [21, 59]. Statistical testing was carried out in a similar manner as described in [12]. 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.

Acknowledgments

We thank Lars Steinmetz, Judith Zaugg, Aino Jarvelin, Wu Wei and Philip Brennecke for fruitful discussions and hosting BZ during the early phase of the project. BZ was supported by a DAAD short term research grant.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

Conflict of interest

The authors declare that they have no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

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 1: 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' [13, 27], 'ChromHMM-Nature' [20], 'ChromHMM-15', '-18' and '-25' [12]), Segway ('Segway-ENCODE' [13, 27], 'Segway-nmeth' [26] and 'Segway-Reg.Build' [64]) and EpicSeg [43].

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

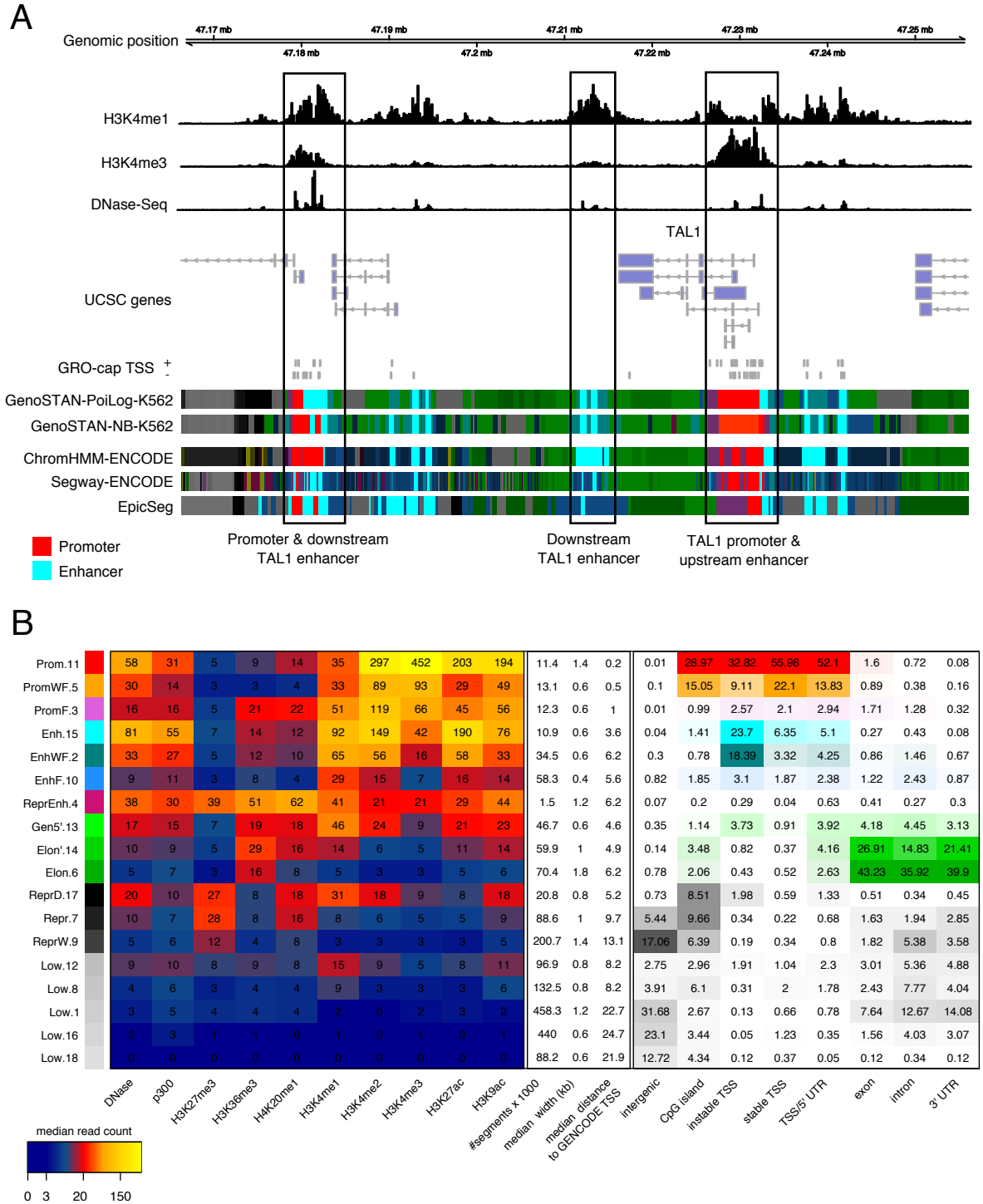


Figure 2: Chromatin states fitted on Benchmark I using GenoSTAN. (A) GenoSTAN segmentations are shown with published segmentations using ChromHMM-ENCODE [13], Segway-ENCODE [13] and EpicSeg [43] at the TAL1 gene and three known enhancers. GenoSTAN-Poilog-K562 correctly recalls all known promoter and 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

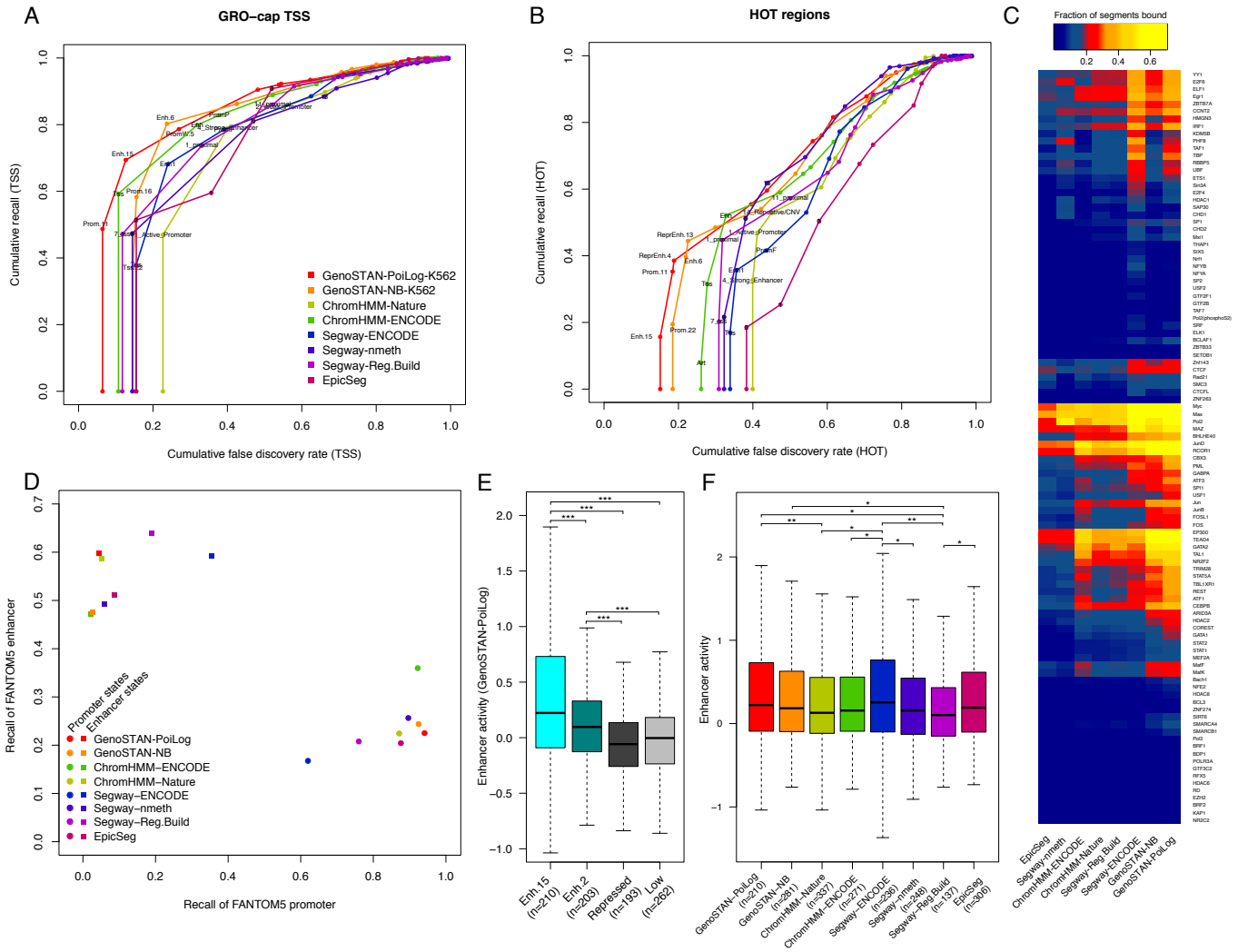


Figure 3: Comparison of GenoSTAN to other published segmentations on benchmark set I. (A) Performance of chromatin states in recovering GRO-cap transcription start sites. Cumulative FDR and recall are calculated by subsequently adding states (in order of increasing FDR). (B) The same as in (A) for ENCODE 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

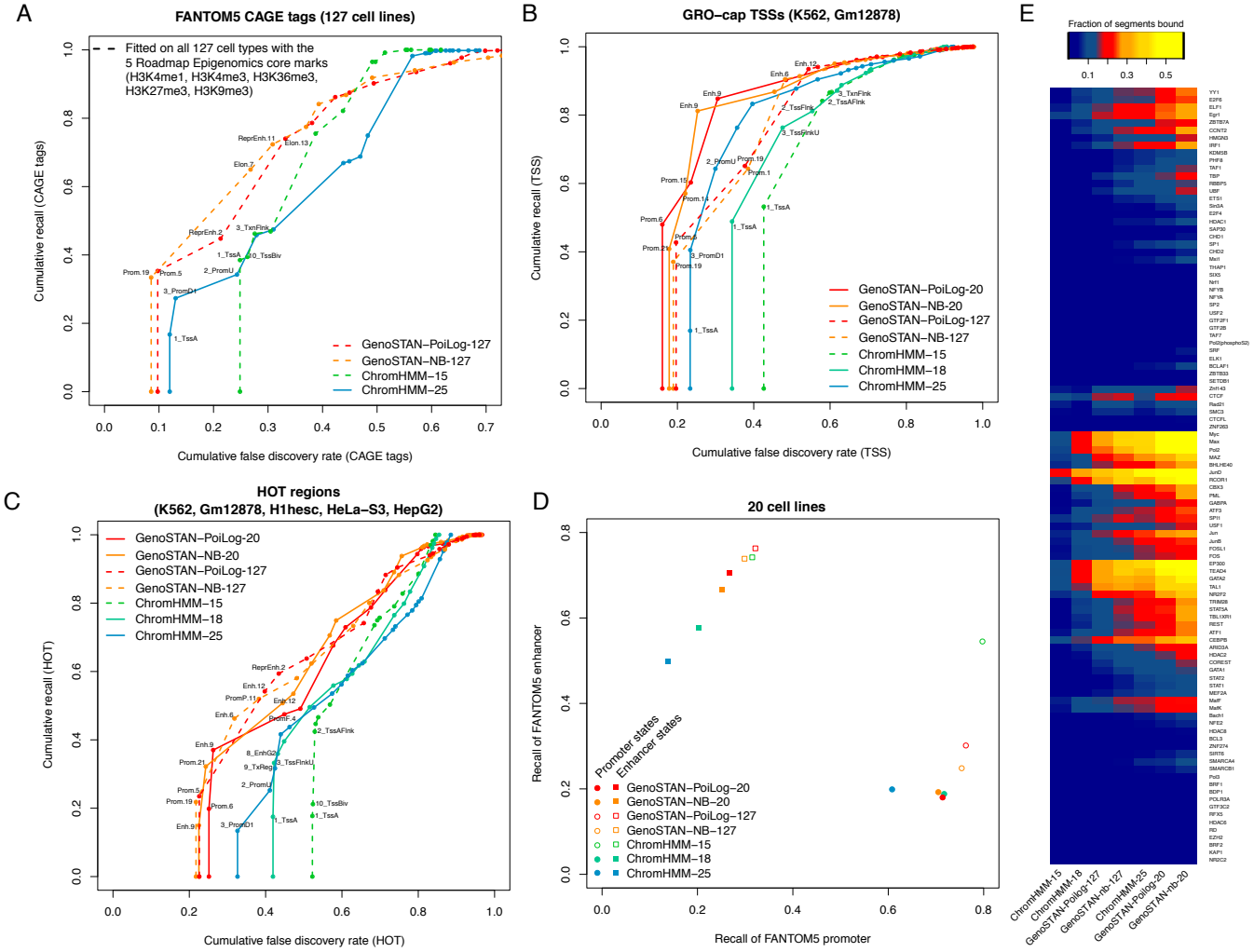


Figure 4: Comparison of GenoSTAN to other published segmentations on benchmark II and III. (A) Performance of chromatin states in recovering FANTOM5 CAGE tags in 127 cell types. CAGE tags were verlapped with chromatin states wihout the use of cell type information. Cumulative FDR and recall are calculated 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 enhancersis 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

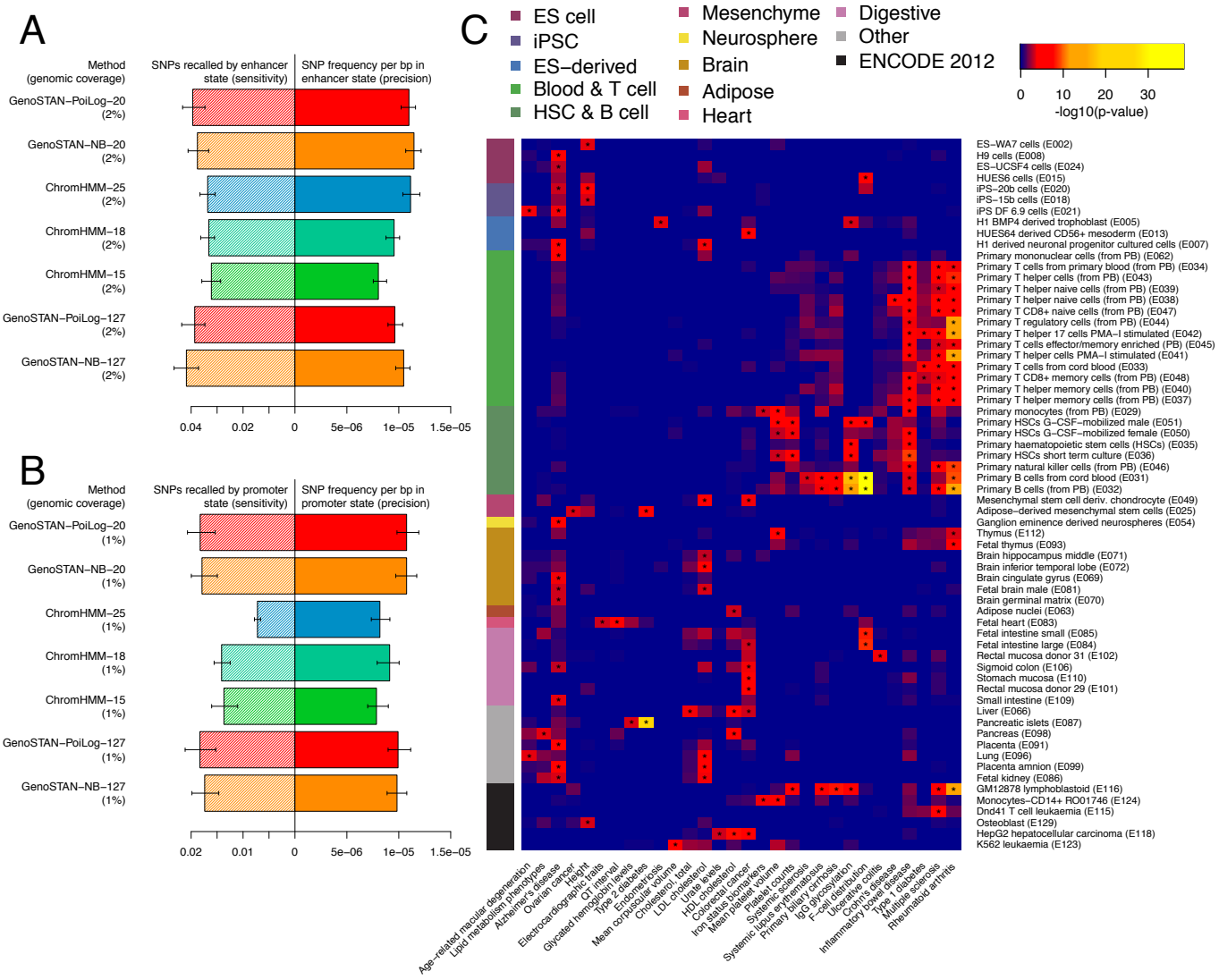


Figure 5: Enrichments of genetic variants associated with diverse traits in enhancers and promoters are specific to the relevant cell types or tissues. (A) Median SNP recall and frequency was calculated for enhancer states in different segmentations by restricting it to a total genomic coverage of 2% (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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

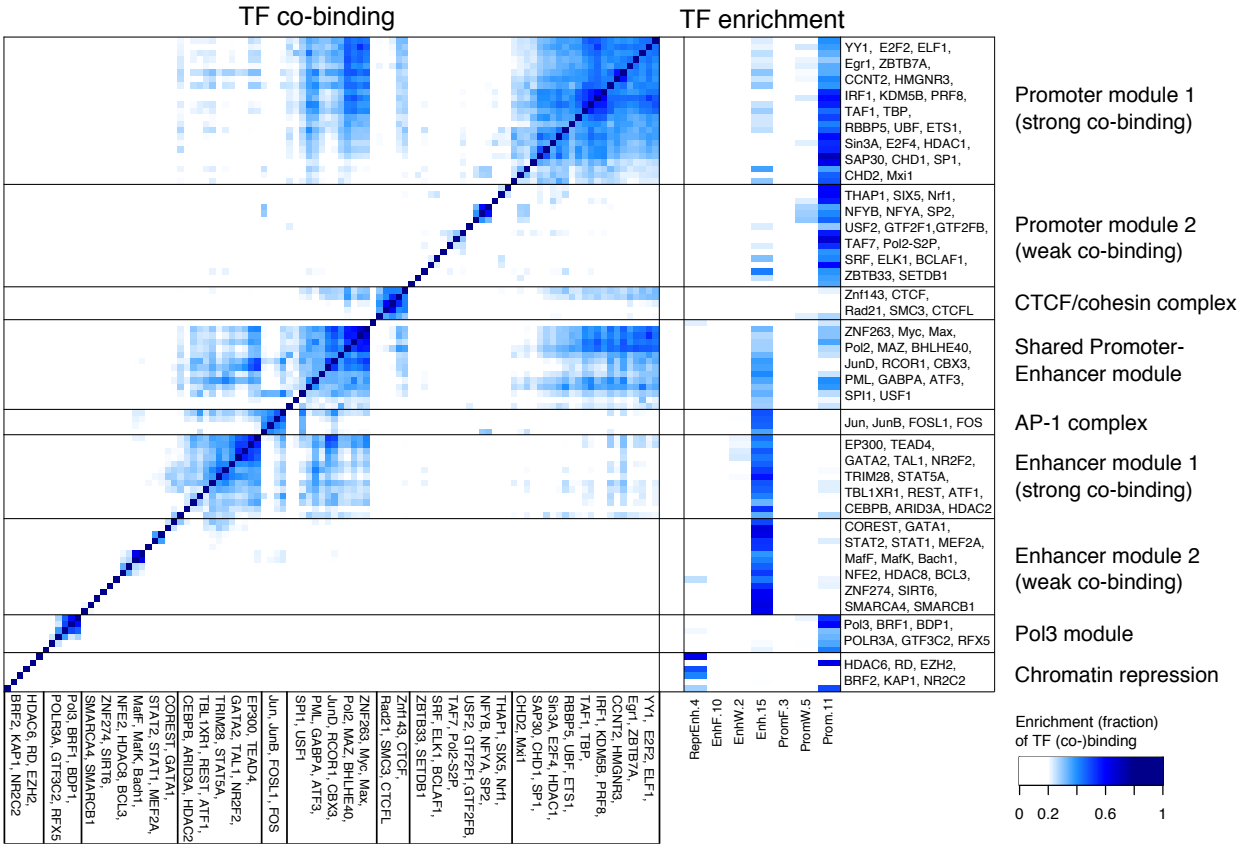
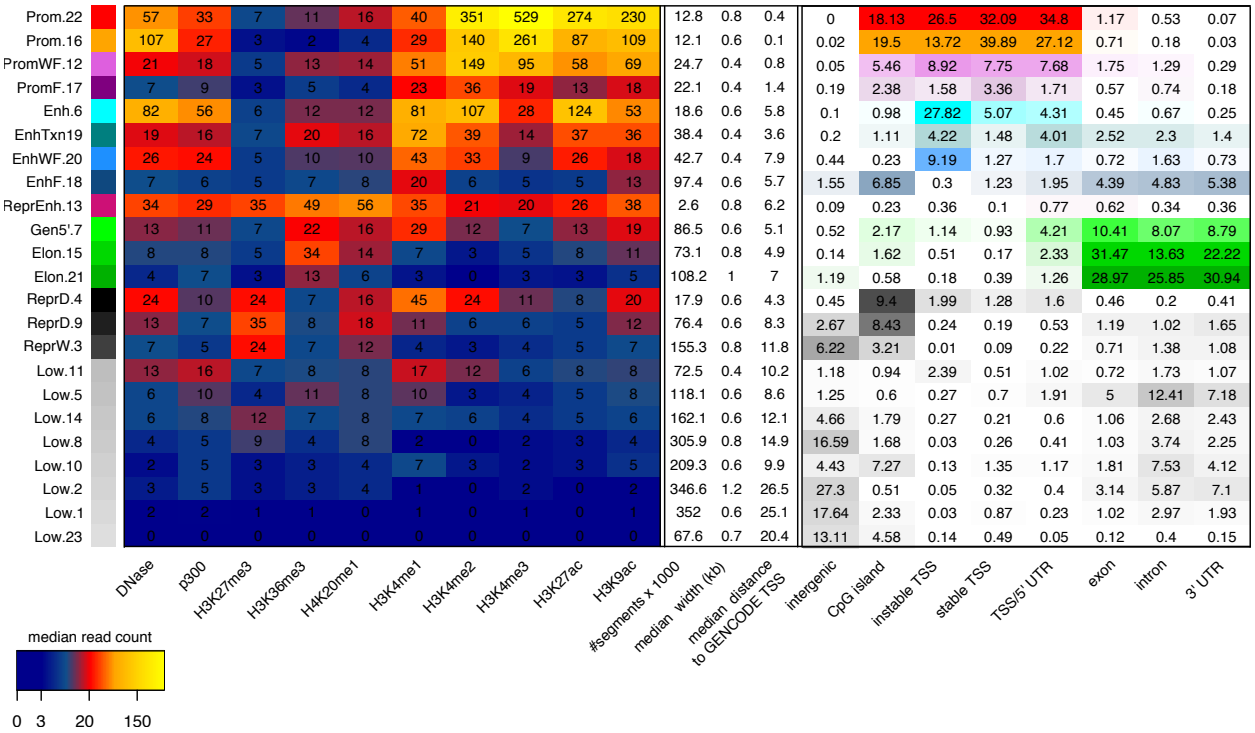


Figure 6: Promoters and enhancers have a distinctive TF regulatory landscape. Co-binding (left) and enrichment of transcription factor binding sites (right) in chromatin states for 101 transcription factor in K562 reveals TF regulatory modules with distinct binding preferences for promoters, enhancers and repressed regions.

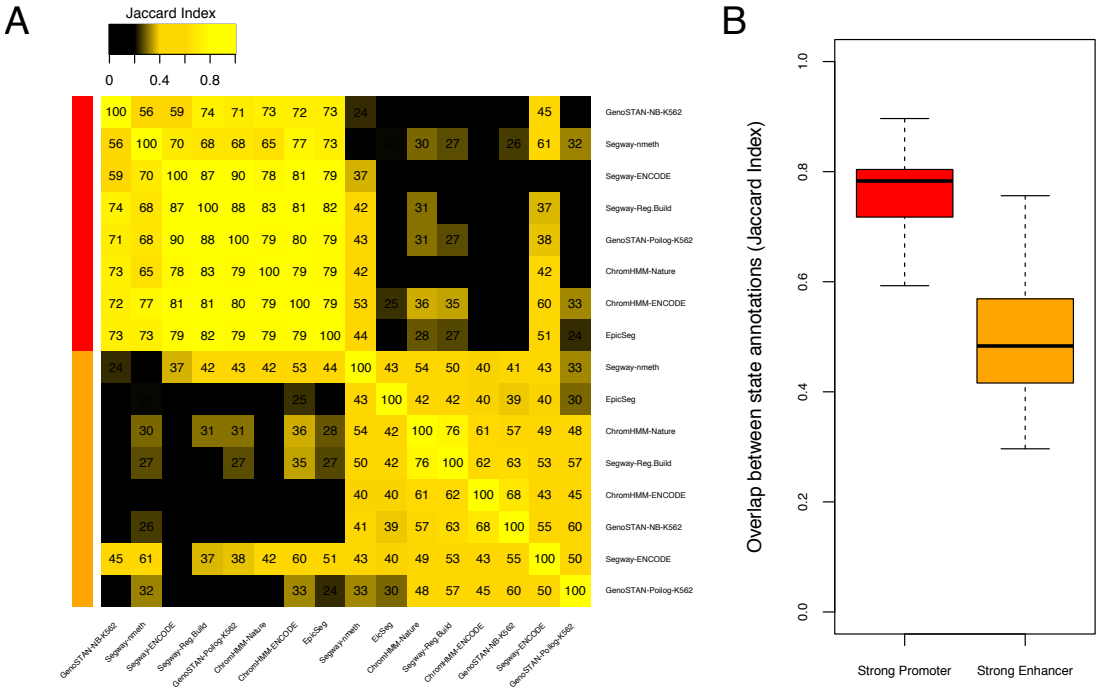
bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figures



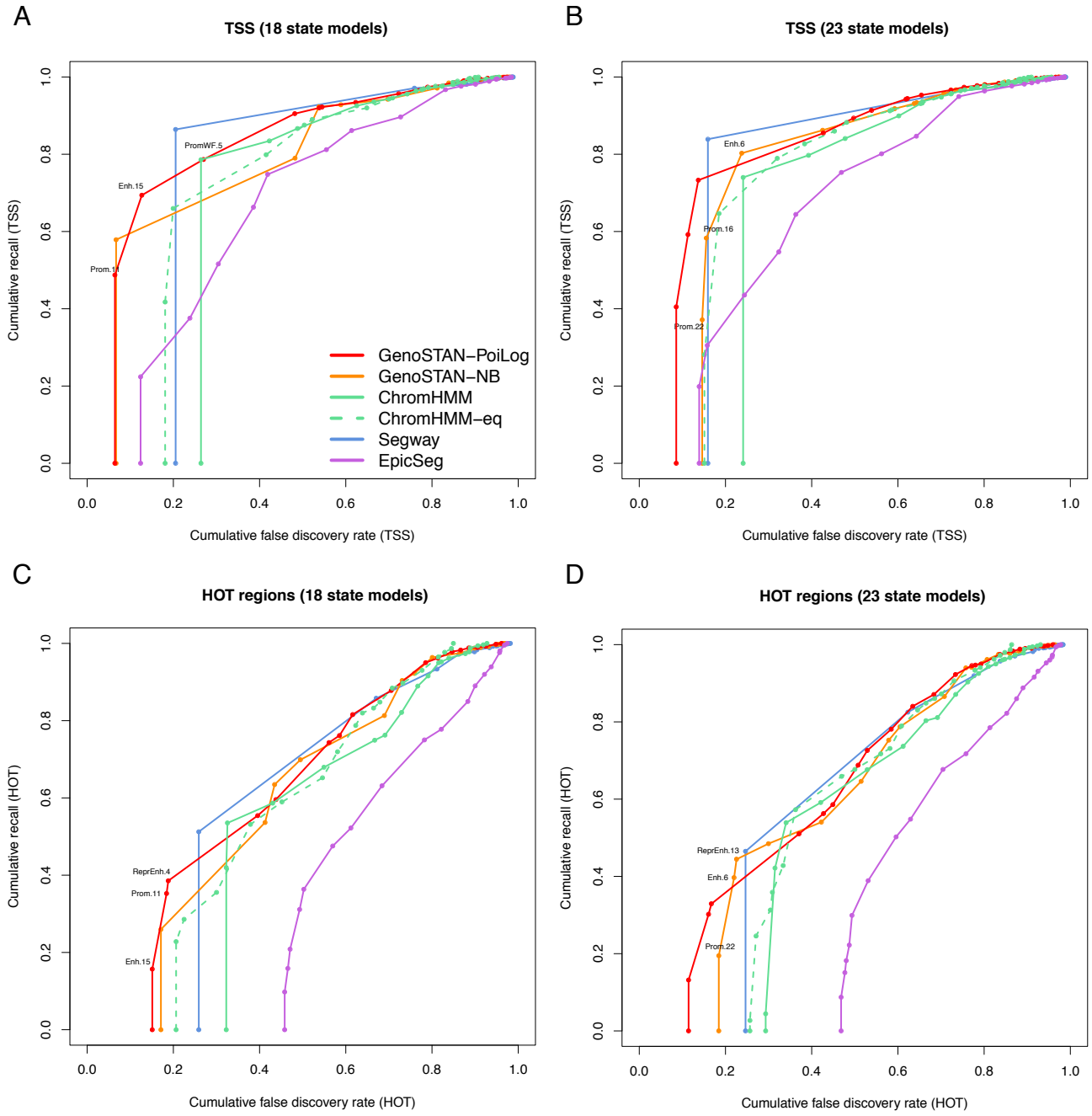
Supplementary Figure 1: Median read coverage of GenoSTAN-nb-K562 chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSS (middle). The right panel shows recall of genomic regions by chromatin states.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



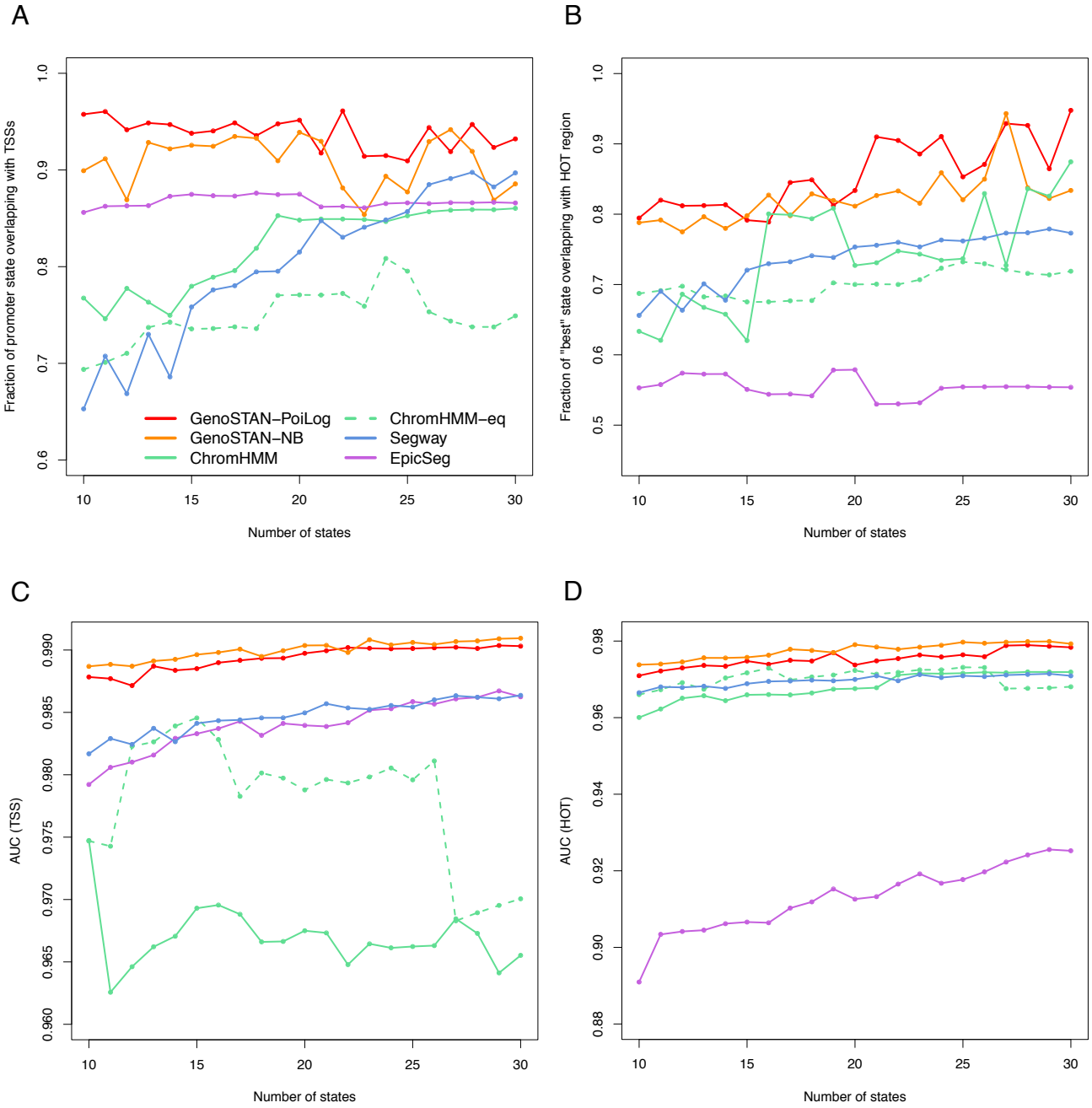
Supplementary Figure 2: (A) Heatmap of pairwise overlap (Jaccard index) of promoter (red) and enhancer (orange) state annotations from different studies on benchmark I. Rows and columns were ordered by separate clustering of promoter and enhancer overlaps. (B) Distribution of pairwise Jaccard indices for strong promoters and enhancers (off-diagonal elements of promoter and enhancer sub-matrices from (A)).

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



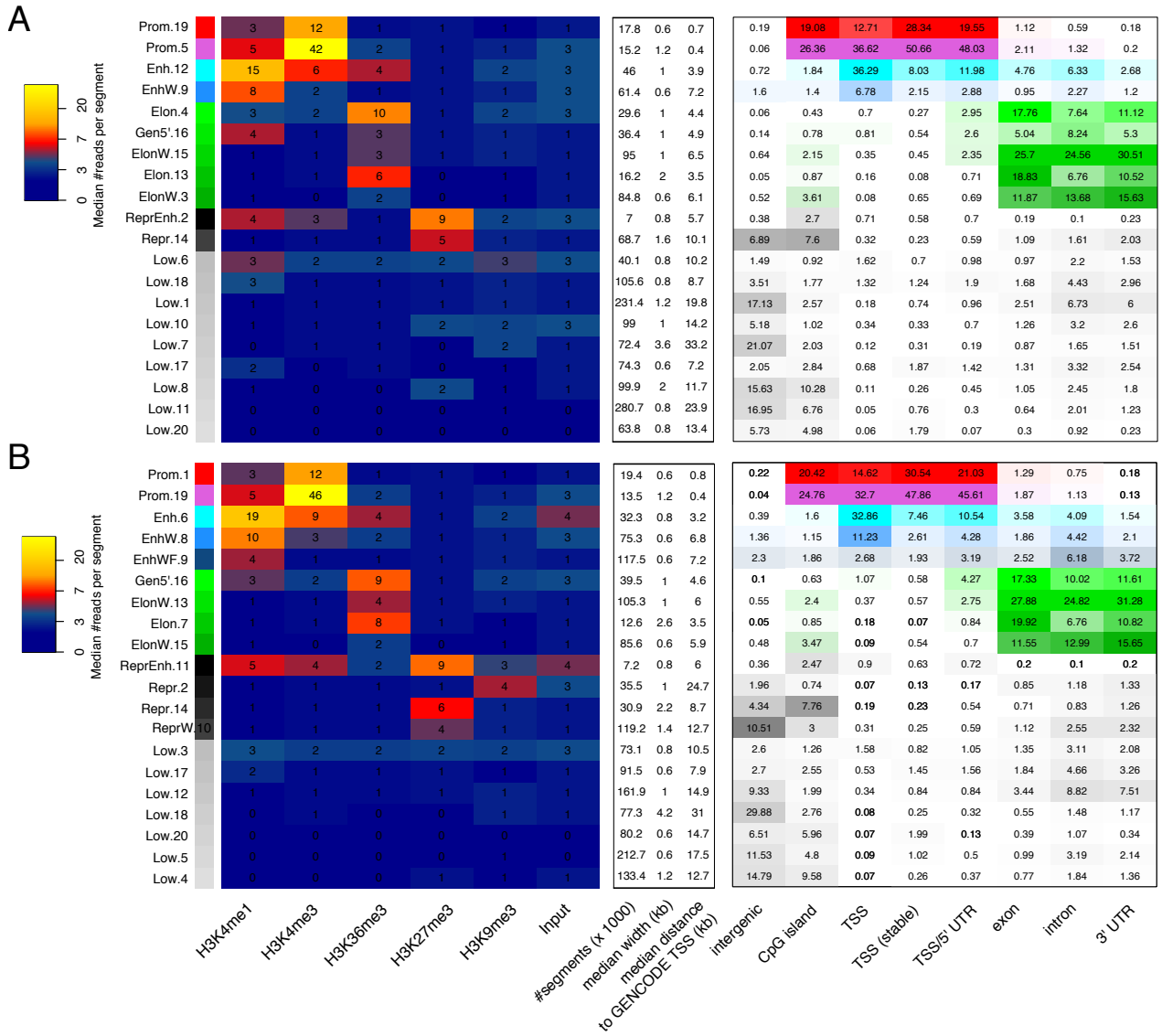
Supplementary Figure 3: Comparison of GenoSTAN to other methods using 18 and 23 states on the same data set in K562 (Benchmark I). (A-B) Performance of chromatin states in recovering GRO-cap transcription start sites for the 18 and 23 state models. Cumulative FDR and recall are calculated by subsequently adding states (sorted by decreasing precision). (C-D) The same as in (A-B) for ENCODE HOT regions.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



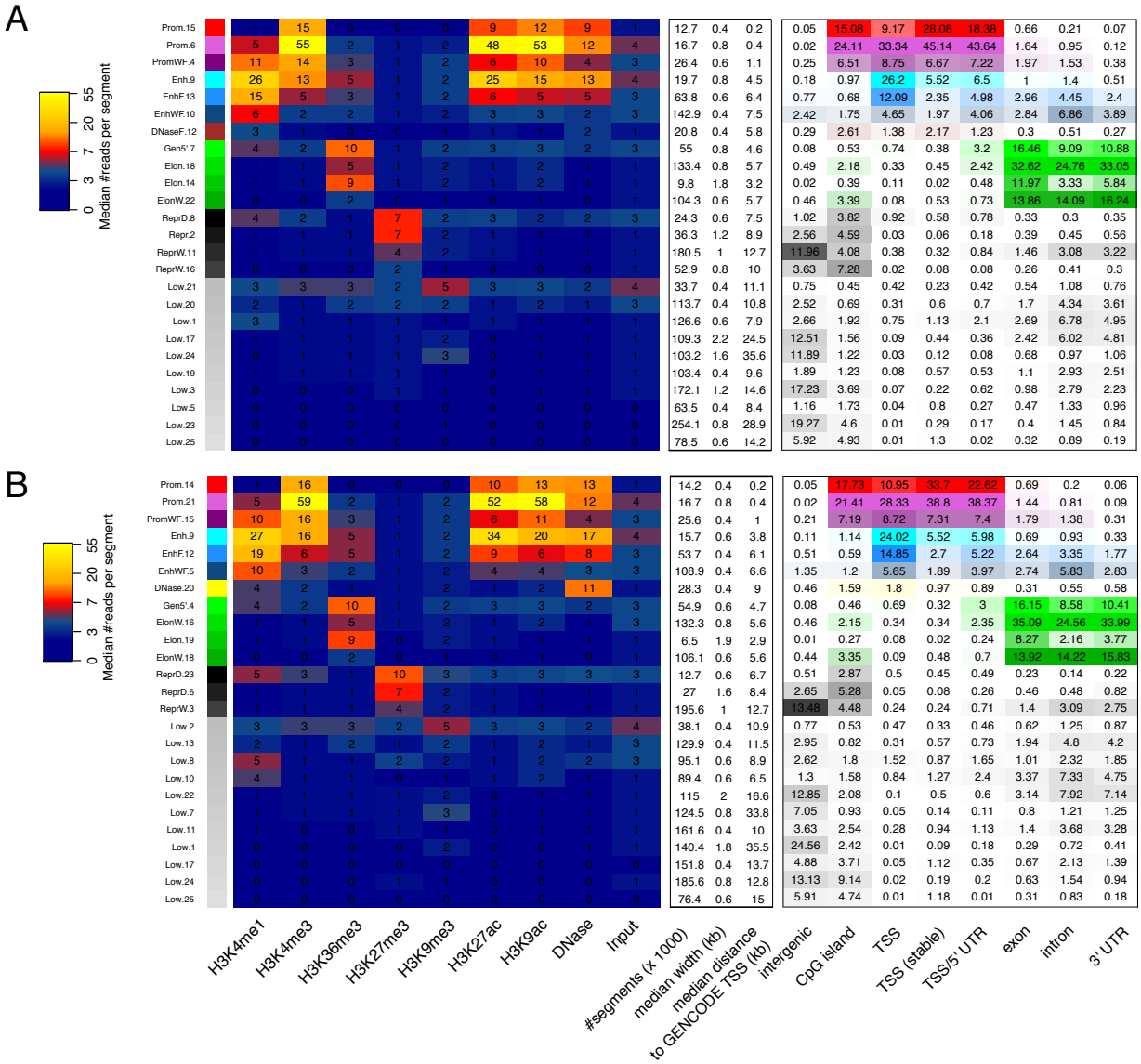
Supplementary Figure 4: Comparison of chromatin segmentation algorithms with respect to their ability to call GRO-cap transcription start sites (left panels) and ENCODE HOT regions (right panels), as a function of the state number used in the respective algorithm (x-axes). All models were learned on 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) is plotted to asses the spatial accuracy of a genome segmentation.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



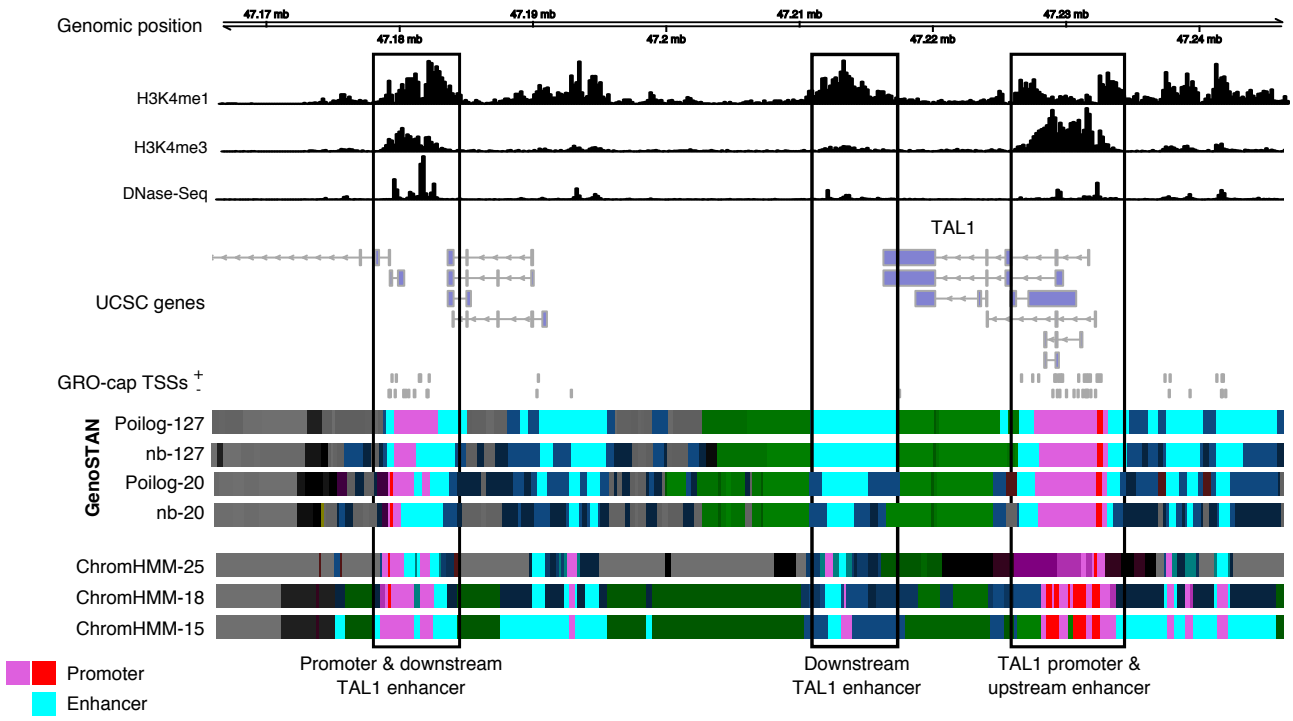
Supplementary Figure 5: GenoSTAN models for benchmark II. (A) Median read coverage of GenoSTAN-Poilog-127 (Benchmark set II, fitted on the 127 ENCODE and Roadmap Epigenomics cell types and tissues) chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSSs of segments (middle). The right panel shows recall of genomic regions by chromatin states. (B) The same as (A) for GenoSTAN-nb-127.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



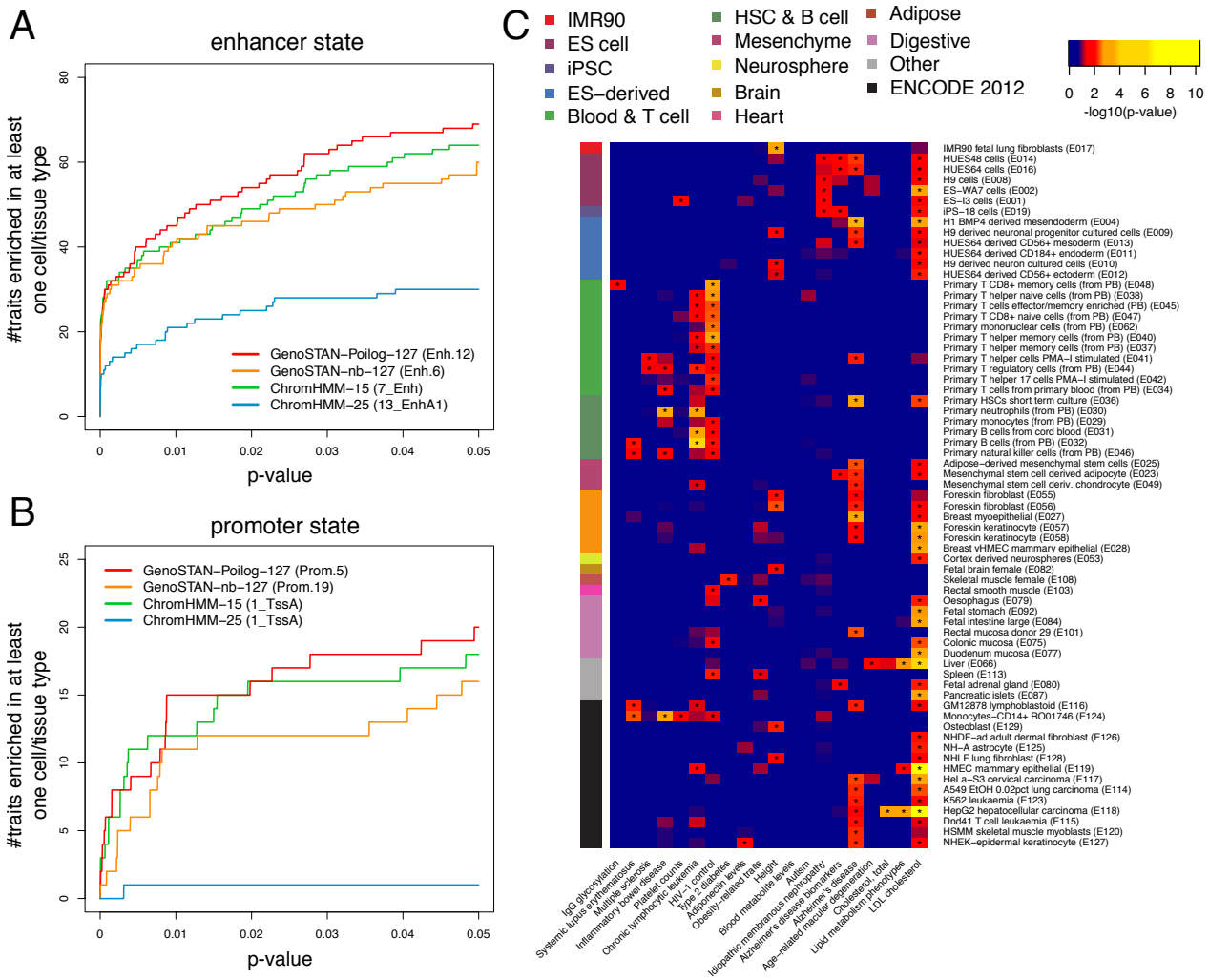
Supplementary Figure 6: GenoSTAN models for benchmark III. (A) Median read coverage of GenoSTAN-Poilog-20 (Benchmark set III, fitted on the 20 ENCODE and Roadmap Epigenomics cell types and tissues) chromatin states (left), their number of annotated segments in the genome, their median width and distance to the closest GENCODE TSSs of segments (middle). The right panel shows recall of genomic regions by chromatin states. (B) The same as (A) for GenoSTAN-nb-20.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



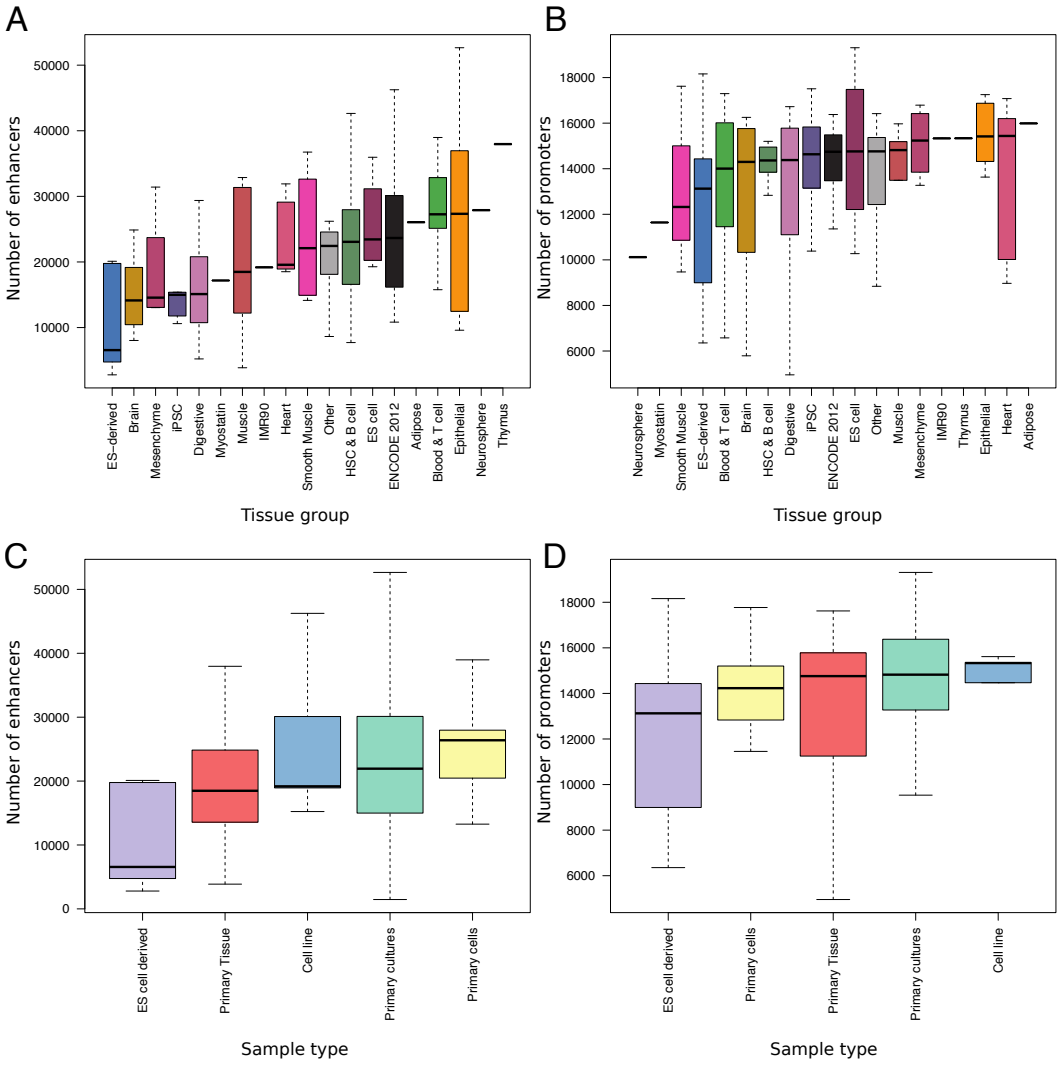
Supplementary Figure 7: Chromatin states annotations for benchmark II and III GenoSTAN segmentations are shown with the three Roadmap Epigenomics ChromHMM segmentations with 15 (ChromHMM-15), 18 (ChromHMM-18) and 25 (ChromHMM-25) states. Shown is the TAL1 locus with segmentations and data from the K562 cell line.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 8: Enrichments of genetic variants associated with diverse traits in enhancers and 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplementary Figure 9: 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.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Tables

	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)

Supplementary Table 1: Number of promoter and enhancer states for the chromatin state annotations analyzed in this study. The original state name is given in brackets.

bioRxiv preprint doi: https://doi.org/10.1101/041020; this version posted February 24, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

	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-Nature	Tss, TssF	Enh, EnhW
ChromHMM-ENCODE	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

Supplementary Table 2: Table showing promoter and enhancers 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.