Enhancer prediction in the human genome by probabilistic modeling of the chromatin feature patterns

Maria Osmala and Harri Lähdesmäki

Department of Computer Science, Aalto University School of Science,
Espoo, 02150, Finland

Abstract

Background: The human genome is far from completely annotated. Specifically, the locations ofgene-distal regulatory enhancers are difficult to locate. Enhancers are binding sites of transcription factors and occupied by nucleosomes with modified histones. The binding sites of transcription factors (TFs) and the localization of histone modifications can be quantified by the chromatin immunoprecipitation assay coupled with next generation sequencing (ChIP-seq). The resulting data has been successfully adopted for genome-wide enhancer identification by several unsupervised and supervised machine learning methods. However, the current methods predict different numbers and different sets of enhancers for the same cell type, and they do not utilize the pattern of the ChIP-seq coverage profiles efficiently. It is also difficult to estimate the accuracy and specificity of the genome-wide enhancer predictions.
Results: We developed PREPRINT, a PRobabilistic Enhancer PRedictIoN Tool. We considered the pattern of, for example, the ChIP-seq coverage profile around the enhancers. The data at the positive and negative examples of enhancers was utilized to probabilistically model the enhancer coverage pattern and to train a kernel-based classifier. We demonstrated the performance of the method using ENCODE data from two cell lines. The predicted enhancers were computationally validated based on the TFs and co-regulatory factor binding sites. We compared our enhancer predictions to the ones obtained by other methods. The effects of different parameter choices during training, testing and validation were studied, and finally, the approach to validate the genome-wide predictions was investigated.
Conclusion: PREPRINT performed comparably to the state-of-the-art methods and provided probabilistic interpretation (i.e. uncertainty) for the predictions. PREPRINT generalized to data from cell type not utilized for training, and often the performance of PREPRINT was superior to RFECS. We observed that the choice of training data and the choice of parameter values at different steps of the enhancer prediction and validation influenced on the final set of predictions. PREPRINT identifed biologically validated enhancers not predicted by the competing methods. The enhancers predicted by PREPRINT can aid the genome interpretation in functional genomics and clinical studies.
Keywords: Enhancer, Probabilistic modeling, Classifier, ChIP-seq, Coverage pattern
Availability: `https://github.com/MariaOsmala/preprint`
Contact: maria.osmala@aalto.fi

Background

Large multinational consortia, such as the Encyclopedia of DNA Elements (ENCODE), the Functional Annotation of the Mammalian Genome (FANTOM) and the reference genome annotation for human (GENCODE), aim to annotate protein-coding genes, functional transcripts, and regulatory regions of the human genome [1, 2, 3]. Of the regulatory regions, enhancers have been ascertained to be the main regulators of the cell type-specific gene expression, and they have an important function in the cell differentiation [4, 5, 6]. The number of enhancers in the human genome is estimated to be hundreds of thousands. However, enhancers are difficult to locate as they are independent in position, distance, and orientation with respect to their target genes [7, 8], and lack general sequence specificities. An enhancer forms the regulatory interaction with its target gene promoter, for example, through chromosomal looping interactions [9, 5]. Enhancers are estimated to be overrepresented (60–80%) in the discoveries of genome-wide association studies (GWAS) aiming to detect single nucleotide polymorphisms (SNPs) associated with both rare and common diseases [10, 11, 12]. Therefore, strategies to locate the enhancers in the human genome in all cell types and patient-derived tissue samples are needed.

Enhancers possess certain genomic features which can be utilized to identify them genome-wide: the location of nucleosomes at enhancers can be quantified applying micrococcal nuclease digestion followed by sequencing

(MNase-seq) [13, 14]. The modifications in the histone tails of nucleosomes and the binding sites of transcription factors (TFs) and co-regulatory factors can be quantified by Chromatin Immunoprecipitation coupled with Sequencing (ChIP-seq) [15, 16]. Due to the binding of regulatory proteins, enhancers exhibit DNase I hypersensitivity (DNase I HS) signal quantified by DNase I HS sequencing (DNase-seq) or Formaldehyde-Assisted Isolation of Regulatory Elements and sequencing (FAIRE-seq) [17, 18]. These data sets have been adopted in several studies to locate enhancers [19, 20, 21, 22, 23, 24, 5, 25, 26]. The features measured by different techniques generate a set of signals along the genome. The signals can be processed with machine learning methods to cluster and classify the genomic loci. Machine learning methods predicting enhancers in different organisms based on different data types and combinations of data have been earlier reviewed and compared [27, 28, 29, 30, 31, 32]. Among the most popular methods is an unsupervised method, ChromHMM, based on a hidden Markov model to learn a small number of chromatin states from the histone modification ChIP-seq data [33, 34]. Some of the learned states display typical features of enhancers. Before learning ChromHMM, the ChIP-seq coverage is converted to binary values (present or absent) depending on a predefined threshold, and the resulting signal is modeled with independent Bernoulli distributions conditioned on the hidden state. The choice of threshold is non-trivial, and due to the binarization, the quantitative information of the ChIP-seq coverage is lost. Moreover, ChromHMM considers the coverage in 200 base pair (bp) bins and does not exploit the special pattern of the coverage profile observed at the regulatory regions. Another drawback of the unsupervised methods is that the correspondence between the identified clusters and regulatory elements is often unknown. In addition, only 30% of the enhancer loci identified by unsupervised approaches of ENCODE Consortium displayed functional activity in massive parallel reporter assays (MPRA) [35].

A supervised random forest classification method called RFECS has been introduced for the enhancer prediction task [36]. RFECS considers the coverage pattern vectors of different histone modification ChIP-seq data as the different features. The data is extracted in a 2 kilo base (kb) window centered at the genomic loci of interest, the window is divided into 20 bins of length 100 bps, hence one feature of one genomic loci is a 20-dimensional vector. When training RFECS, at each node in a tree, a subset of features are randomly selected from the feature set, and the single feature that produces the best separation of classes according to a predetermined criterion is utilized to partition the training data. To reduce the dimension of a feature from 20 to 1, at each node, RFECS applies Fisher Linear Discriminant Analysis, an example of a multi-variate node-splitting technique. The authors of RFECS claim that this approach allows the utilization of the coverage pattern as well as the abundance information. Rajagopal et al.[36] demonstrated that RFECS outperformed the other supervised methods Chromia [37], CSIANN[38], and ChromaGenSVM [39]. RFECS does not make any distributional assumptions of the data, and the algorithm automatically discovers the optimal subset of the features. Another supervised approach, a deep neural network trained on the FANTOM enhancer atlas [40] has been introduced for enhancer prediction [41]. In terms of data, the authors used the mean value of the coverage signal at 200 bp windows along the genome and the predictions were made on the 200 bp windows. This approach did not likely capture the whole coverage pattern at the regulatory regions. In addition, training the classifier with the FANTOM enhancer atlas might bias the results: The FANTOM enhancer atlas contains only around 40000 enhancers across tens of different cell types. The FANTOM enhancers are identified by quantifying the transcription of non-coding enhancer RNA (eRNA) [42, 43] with the Cap Analysis of Gene Expression (CAGE-seq) [44]. Another technique to measure eRNA is the global run-on and sequencing GRO-seq [45]. The enhancer RNA is highly unstable and degrades rapidly; thus sufficient sequencing depth is required for CAGE-seq and GRO-seq to capture all enhancer RNA transcription. The FANTOM enhancers are likely ones representing the strongest eRNA signals, and clearly their number is an underestimate of the enhancers in the human genome.

The different machine learning methods predict different sets of enhancers for the same cell type, they do not generalize well between cell types, and the predicted enhancers may have different properties [31, 32, 46, 47]. Moreover, the lengths of enhancers predicted by different methods vary (from a few hundreds to a few thousands of bps), and the set of enhancers might not be saturated; more enhancers could be identified by lowering the prediction thresholds, or by analyzing more cell types [47]. The inconsistencies between the sets of predicted enhancers are likely due to many factors: First, most methods utilize the ChIP-seq coverages within a large genomic window as features and do not efficiently utilize the pattern of the coverage signal in subsequent small windows within the large window. Regions with a low signal intensity may still have the characteristic enhancer coverage pattern. Second, training data definition and the features used for enhancer prediction likely hae a large impact on the genome-wide predictions. Most enhancer prediction methods exploit only a subset of all possible data that can be utilized for enhancer prediction. In this work, the data contain histone modification ChIP-seq data, MNase-seq data, DNase-seq data, and RNA polymerase II (RNA Pol II) and CCCTC-binding factor (CTCF) ChIP-seq data. Particularly, the utilization of RNA Pol II occupancy and MNase-seq data are less exploited to predict enhancers, although the occupancy of RNA Pol II ChIP-seq is easier to measure than

the native eRNA produced at enhancers. Fourth, the prediction scores of the classifiers should be calibrated, for example to control the false positive rate. Finally, it is difficult to estimate the accuracy and specificity of the genome-wide predictions as there are no large gold standard set of enhancers for the human cells. In this work, we introduce PRobabilistic Enhancer PRedictIoN Tool PREPRINT, which is based on constructing a profile pattern characteristic of an enhancer and a probabilistic classification. If the query genomic region is close to the enhancer pattern, the region is characterized as an enhancer. In terms of data, we employ the above-mentioned next-generation sequencing data from ENCODE for the myelogenous leukemia cell line (K562) and the lymphoblastoid cell line (GM12878). We assess the probability of misclassification of whole-genome predictions in advance and build a general tool that once trained can predict enhancers on data originating from any cell type. Moreover, we experiment with different definitions of the training data and present a principled way to validate the genome-wide predictions.

Results

Evaluating the generalization performance of the classifiers

The classification performance of PREPRINT and RFECS was evaluated using the area under the receiver operating characteristics curve (AUC) values. The performance of the methods were first evaluated on small sets of 1000 enhancers, 1000 promoters and 1000 random regions, defined either in cell line K562 or GM12878. The K562 data is denoted as a training data, and the GM12878 data as a test data. The random regions were defined in two ways: either purely random regions were considered (pure random) or random regions with the coverage values above a certain threshold (random regions with signal). For more details, see Methods. The training data from cell line K562 was divided into cross-validation (CV) sets to evaluate the classification performance of the methods on data from a single cell line. The test data from cell line GM12878 was used to test the generalization performance of the methods between the cell lines. The AUC values for the different methods and data sets are shown in Table 1. The classification performance on the K562 CV data set was almost perfect (0.99), and the performance decreased only slightly when predicting enhancers on GM12878 data using classifier trained on K562 data. As expected, enhancers were easier to separate from the pure random regions than from the random regions with signal. This is especially the case in the GM12878 cell line. In addition, RFECS separated enhancers from the pure random regions better in cell line GM12878 than the probabilistic maximum likelihood (ML) and Bayesian methods. Nevertheless, among the methods trained on the random regions with signal, the Bayesian approach generalized to the GM12878 data the best. However, classifying the small set of highly significant enhancers and promoters is a rather simple task, and next we present the evaluations of the whole-genome predictions.

Table 1: The classification performance (AUC) of PREPRINT and RFECS in the 5-fold CV data set from cell line K562 and test data from the cell line GM12878. For RFECS, we did not compute the AUC values on the K562 CV data. The method with the best generalization performance on the GM12878 data is indicated with the bold font.

		AUC	
Method	Cell line	Pure random regions	Random regions with signal
Bayesian	K562	0.993	0.988
ML	K562	0.993	0.990
Bayesian	GM12878	0.982	**0.960**
ML	GM12878	0.982	0.938
RFECS	GM12878	**0.987**	0.959

PREPRINT predicted a larger amount and shorter enhancers than RFECS and ChromHMM

PREPRINT and RFECS trained on the whole training data from cell line K562 predicted enhancers genome-wide in both cell lines. Both PREPRINT and RFECS scan the genome in subsequent 2 kb windows advancing in 100

bp shifts along the genome. For each of the genomic windows, PREPRINT and RFECS assign a prediction score. If the prediction score is above a certain threshold, the window is predicted as an enhancer. First, a prediction threshold 0.5 was utilized for both PREPRINT and RFECS: In order for RFECS to predict a genomic region as an enhancer, 50% or more of the trees of the random forest need to vote for an enhancer class. By contrast, in order for PREPRINT to predict a genomic region as an enhancer, the enhancer class probability needs to exceed 0.5. The prediction threshold 0.5 is likely suboptimal and not well calibrated. Hence, for PREPRINT, the best operating point threshold and 1% false positive rate (FPR) threshold were estimated from the performance evaluation measures on the K562 CV data set. In addition, the best operaing point threshold and the 1% FPR threshold for cell line GM12878 data were estimated from the performance evaluation measures on the GM12878 test data. However, the prediction thresholds optimized for the K562 data should be adopted when predicting enhancers in other cell types, since no test regions might be available for the other cell type. With the prediction threshold 0.5, RFECS predicted notably less enhancers than PREPRINT (see Table 2). Therefore, to equalize the number of enhancers predicted by RFECS and PREPRINT, the prediction threshold for RFECS was lowered to a sufficient degree. Furthermore, the training data enhancers and promoters were removed from the final genome-wide predictions. In addition to PREPRINT and RFECS, ChromHMM Strong Enhancer and Weak Enhancer clusters obtained from ENCODE were included in the method comparison [33, 34].

Predicting enhancers by PREPRINT and RFECS results in subsequent windows having a prediction score higher than the chosen threshold. To increase the resolution of predictions, one or several single windows need to be chosen within a wider region. RFECS predicts very wide regions and aims to find multiple local maxima within a region. In turn, PREPRINT chooses only the window with the maximum prediction score. If multiple windows have the same maximum score, one is selected at random. However, instead of predicting an enhancer as a single window, we could also consider the whole region of subsequent enhancer predictions as an enhancer. Consequently, we computed the normalized frequencies of lengths of predictions obtained by the different methods. For RFECS and PREPRINT, we utilized two different thresholds (0.5 and 0.75). Figure 1 plots the normalized frequencies recorded when PREPRINT and RFECS were trained on the pure random regions. Figure 1 shows that PREPRINT and RFECS predicted proportionally shorter enhancers than ChromHMM, which predicted mostly enhancers of length 1 – 10 kb. In addition, RFECS predicted proportionally more enhancers of length 100 bp and of length larger than 1 kb, whereas PREPRINT predicted proportionally more enhancers of length 200 bp – 1 kb. From now on, the enhancers of length 100–1000 bp are denoted as short enhancers, and the enhancers of length larger than 1 kb are denoted as long enhancers. The proportions of short and long enhancers predicted by RFECS are increased and decreased, respectively, when adopting the more stringent threshold (0.75). This behaviour is expected, because with the more stringent threshold, large prediction regions were divided into separate smaller regions, or they became shorter or both. By contrast, when adopting the more stringent threshold for PREPRINT, the proportional frequencies remained almost the same. This suggests that the prediction scores of PREPRINT advance from a low value to a high value and back within a short region (within a low number of the window shifts), whereas the prediction scores of RFECS increase and decrease smoothly within a large genomic window. To conclude, the lengths of the PREPRINT enhancers are less sensitive to changes in the prediction threshold.

Supplementary Figure S4, Additional File 1, plots the normalized frequencies of enhancer lengths when the PREPRINT and RFECS were trained on the K562 data and on the random regions with signal. In comparison to Figure 1, the differences in the proportional frequencies of the long RFECS enhancers predicted by the two thresholds were even more evident. In general, proportional frequencies of the short enhancers were higher for the methods trained with the random regions with signal compared to the methods trained the pure random regions. Moreover, the normalized frequencies of very short (< 400 bp) PREPRINT enhancers are higher than the corresponding frequencies for RFECS. Similar results are obtained for the data from cell line GM12878 (Supplementary Figures S5 and S6, Additional File 1). Finally, a large proportion of predicted enhancers consist of only one window (100 bp), or of short enhancers (< 2 kb), suggesting that defining the location of the PREPRINT enhancers by choosing the window with the maximal score within a larger region is an adequate approach.

The number of genome-wide enhancer predictions for each method and threshold are provided in Table 2. The numbers were recorded before and after TSS removal. When predicting enhancers in cell line GM12878, we used either the best operating point threshold or the 1% FPR threshold estimated from the K562 CV, or the thresholds estimated from the GM12878 test data. The best operating point thresholds for the K562 cell line were all close to 0.5, whereas for the GM12878 cell line, when PREPRINT was trained on the random regions with signal, the best operating point thresholds were close to 0.3, resulting in a very high number of enhancers. In general, RFECS predicted less enhancers than ChromHMM and PREPRINT with the 0.5 prediction threshold. Specifically, the numbers of predictions obtained by PREPRINT with the 1% FPR threshold are still higher than the numbers of predictions obtained by RFECS with the prediction threshold 0.5. Furthermore, with the prediction

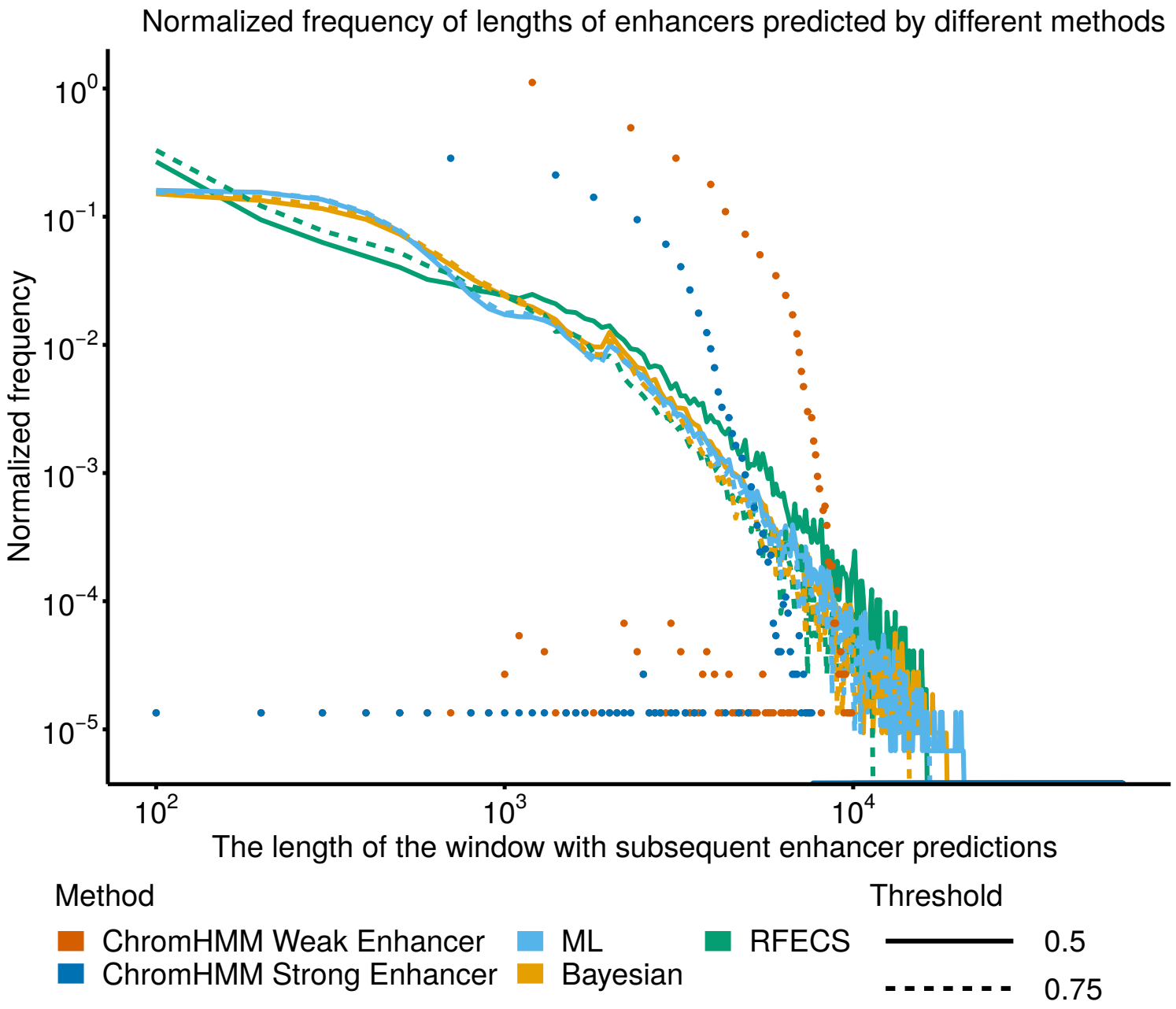

Figure 1: The normalized frequencies of varying lengths of enhancers predicted in cell line K562 by different methods using two prediction thresholds (0.5 and 0.75). PREPRINT and RFECS were trained on the pure random regions. For each method and threshold, the frequencies were divided by the total number of regions predicted as enhancers by each method. The regions were formed by combining the subsequent enhancer predictions into one region.

threshold 0.5, RFECS trained on the random regions with signal predicted much less enhancers compared to when trained on the pure random regions. When using a lower prediction threshold for RFECS (0.25), the numbers of enhancers were comparable between the different random data definitions. By contrast, with the prediction threshold 0.5 and the best operating point threshold, PREPRINT predicted more enhancers when trained on the random regions with signal compared to the pure random regions. With the 1% FPR thresholds, the number of PREPRINT enhancers were comparable between the different random data definitions. The Bayesian approach predicted a lower number of enhancers than the ML approach, except in cell line GM12878 when PREPRINT was trained on the random regions with signal. Finally, the 1% FPR thresholds estimated from the GM12878 test data were notably higher compared to the 1% FPR thresholds estimated from the K562 CV data. Therefore, one should be cautious when generalizing the prediction thresholds between data from different cell lines. To conclude, the larger number of enhancer predictions for PREPRINT may result from PREPRINT predicting proportionally more short enhancers than RFECS, as seen, for example, in Figures 1 and S4, Additional File 1. In addition, the random region definition affected the estimated thresholds and the number of enhancer predictions, depending on the cell line and the threshold setting (0.5, the best operating point, or the 1% FPR threshold).

Table 2: The number of genome-wide enhancers predicted by different methods and thresholds.

			Threshold 0.5		The best operating point or threshold 0.25 for RFECS			FPR 1% Threshold		
Method	Cell	random	all	without TSS	threshold	all	without TSS	threshold	all	without TSS
ML	K562	pure	145208	127857	0.442	157057	138760	0.735	106161	92509
Bayesian	K562	pure	105386	89859	0.526	101986	86768	0.827	62210	51838
RFECS	K562	pure	37072	30593	0.250	76412	63216			
ML	K562	with signal	310477	288243	0.579	225970	206912	0.747	108428	94998
Bayesian	K562	with signal	227902	208035	0.591	162421	145801	0.771	80755	69210
RFECS	K562	with signal	18655	15622	0.250	63850	53773			
ChromHMM Weak Enhancer	K562		180471	176912						
ChromHMM Strong Enhancer	K562		69019	66888						
ML	GM12878	pure	151732	127438	0.442	173101	147286	0.735	129013	106737
Bayesian	GM12878	pure	64594	55818	0.526	130019	109373	0.827	96939	80762
RFECS	GM12878	pure	37287	33227	0.250	113662	101117			
ML	GM12878	with signal	266968	415960	0.579	200436	179475	0.747	102566	87379
Bayesian	GM12878	with signal	287042	265754	0.591	210085	192613	0.771	101388	90428
RFECS	GM12878	with signal	20670	18359	0.250	113609	103588			
ChromHMM Weak Enhancer	GM12878		178474	175487						
ChromHMM Strong Enhancer	GM12878		64052	62599						
Thresholds estimated for the GM12878 cell line										
ML	GM12878	pure			0.572	151732	127438	0.927	96283	78287
Bayesian	GM12878	pure			0.656	116236	97127	0.940	74434	62508
ML	GM12878	with signal			0.359	446790	415960	0.962	23925	16296
Bayesian	GM12878	with signal			0.303	562483	530530	0.843	67722	59307

Enhancers uniquely predicted by PREPRINT validated with a small number of overlapping transcription factor binding sites

The genome-wide enhancer predictions were validated by inspecting the overlap between the predicted enhancers and the histone acetyltransferase (p300) binding sites (SydhK562P300Iggrab and SydhGm12878P300Iggmus ChIP-seq peak sets). In addition, a large set of TF and other co-regulatory protein binding sites from the Transcription factor ChIP-seq Uniform peaks from ENCODE were utilized for validation. The peaks for RNA Pol II, CTCF, CREB-binding protein (CBP) and p300 were removed from the Uniform peak set resulting in peaks for 111 and 76 individual DNA binding proteins for cell lines K562 and GM12878, respectively. For more details about the validation data, see Additional File 2. However, using the ChIP-seq peaks for validation can be problematic: first, not all TSS-distal protein binding sites are enhancers; the binding sites can be some other functional genomic regions, such as silencers or insulators. Second, the DNA binding proteins may contain both activating and repressing factors. Hence, the enhancers expressing repressing chromatin features are validated as enhancers, not only the active ones. The prediction of the repressed enhancers is a relevant task itself, but is not supported by the choice of the enhancer coverage patterns used in this work.

An enhancer was validated if the 2 kb prediction window overlaps at least 1 bp of at least 1 peak in the validation peak sets. The ChromHMM enhancer clusters contain regions with varying sizes; these were validated similarly. Instead of requiring at least 1 overlapping Uniform ChIP-seq peak to validate a prediction, a more rigorous requirement for the number of overlapping peaks could be adopted. However, to chooce the threshold for the required number of overlapping peaks might not be straightforward. Figure 2 shows the proportions of enhancer predictions having an overlap with varying numbers of TF or co-regulatory protein binding sites. The predictions were obtained in the K562 cell line by PRERINT and RFECS, and in each comparison (a,b,c or d), an equal number of enhancers were predicted by the methods; in comparisons a and b, the number of enhancers were 30593 and 15622, respectively. These were the numbers of enhancers predicted by RFECS with threshold 0.5. In comparisons c and d, the numbers 51838 and 69210 corresponded to the number of enhancers predicted by PREPRINT with the Bayesian approach using the 1% FPR threshold. In comparisons a and c, the pure random regions, and in comparisons b and d, the random regions with signal, respectively, were utilized when training the model. In comparisons a and b, smaller numbers of top enhancer predictions were considered, and the proportions of predictions having 0 TF or co-regulatory binding site were higher for PREPRINT than for RFECS. In addition, the proportions of predictions overlapping 5–20 TF or co-regulatory binding sites were slightly higher for RFECS. Conversely, the proportions of enhancers overlapping a small number (1–2) of TF or co-regulatory binding sites were higher for PREPRINT. The proportions became comparable between the different methods when increasing the number of predictions (comparisons c and d).

Similar results were obtained for the predictions in cell line GM12878 (see Supplementary Figure S7, Additional File 1). The comparisons a–d were the same as in Figure 2, and the comparisons e and f corresponded to

the 1% FPR thresholds estimated from the GM12878 test data. By contrast to the results seen in comparisons c and d in Figure2, the proportions of predictions having zero TF or co-regulatory binding sites are lower for PREPRINT than for RFECS (see comparisons c, d and e in Supplementary Figure S7, Additional File 1). This is especially seen for the both ML and Bayesian approaches when the methods were trained on the pure random regions. Moreover, the proportions of enhancers having a small number (1–3) of TF or co-regulatory binding sites were higher for PREPRINT than for RFECS. In comparison f, the number of enhancers predicted by PREPRINT with the 1% FPR threshold (16295) was similar to the number of enhancers predicted by RFECS with threshold 0.5 (18359); hence, the comparisons b and f resulted in similar graphs.

To conclude, first, it would be preferable that the proportion of predictions having zero TF or co-regulatory binding sites were low for a set of predictions. RFECS predictions in cell line K562 contain less predictions with zero peaks compared to PREPRINT predictions, especially among the predictions with the largest prediction scores (comparisons a and b). Second, in comparisons a and b in both cell lines, the RFECS predictions contained proportionally more enhancers overlapping 5–20 different validation peaks than the predictions obtained by PREPRINT. By contrast, PREPRINT predicted proportionally more enhancers with a small number (1–3) of overlapping peaks; these enhancers may display weaker chromatin feature signals and may be missed by RFECS, while they are still weakly validated. Third, the frequency distributions between RFECS and PREPRINT became comparable when the number of predictions increases (comparisons c and d). Finally, in cell line GM12878, the frequency of predictions having zero TF or co-regulatory binding sites, and the frequency of enhancers validating with a small number of (1–3) of overlapping peaks are lower and higher, respectively, for PRERPRINT methods compared to RFECS. This might reflect a good generalization performance of PREPRINT to the data from the GM12878 cell line. Based on these results, it is still challenging to define the threshold for the required number of overlapping validation peaks. Therefore, in the following sections, the validation was still founded on the requirement of at least 1 overlapping ChIP-seq peak.

To study the performance of the methods to predict enhancers in the whole human genome, we selected an equal amount of enhancers and non-enhancers predicted by PREPRINT and RFECS. The non-enhancers were chosen randomly among all regions having the prediction scores less or equal to 0.5. The predicted enhancers and non-enhancers were labelled either as true positives, false positives, true negatives or false negatives considering the overlap between the regions and the validation data ChIP-seq peaks. For both RFECS and PREPRINT, the number of enhancers and non-enhancers was set to the number of enhancers predicted by RFECS with prediction threshold 0.5. Table 3 provides the AUC values for the genome-wide predictions. RFECS obtained the highest AUC scores in almost all four settings (p300 or TF, K562 or GM12878). Nevertheless, PREPRINT with the ML approach reached the best AUC value 0.837 in cell line K562 when the method was trained on the pure random regions, and the predictions were validated using the p300 peaks.

Moreover, in cell line GM12878, PREPRINT trained on the pure random regions with the ML approach resulted in the AUC value 0.91 when the predictions were validated with the Uniform TF peak set, reaching comparable performance to RFECS (0.936). Of the PREPRINT methods, the ML approach is always better than the Bayesian approach, and the genome-wide enhancers predicted by PREPRINT trained on the pure random regions validate better than enhancers predicted by PREPRINT trained on the random regions with signal. Conversely, enhancers predicted by RFECS trained on the random regions with signal validate better; this likely resulted from RFECS predicting much less enhancers when trained on the random regions with signal compared to trained on the pure random regions. In addition, in cell line GM12878, there was no difference between the performance of RFECS trained on different random data definitions, whereas in cell line K562 there were more differences between the AUC values: 0.821 vs. 0.916 when validation was based on the p300 binding sites, and 0.907 vs. 0.929 when the validation was based on TF and co-regulatory binding sites. By constrast, the differences in the AUC values of PREPRINT trained on the different random definitions are larger in cell line GM12878 than in cell line K562. To conclude, the validation performance of the genome-wide predictions were comparable across methods, and the ML approach reached a good generalization performance between data from different cell lines. However, the different settings, e.g. the cell line used for training and prediction, the type of validation data, and the definition of the random regions, lead to divergent results.

To further investigate the performance of the methods to predict the enhancers genome wide, we computed the proportion of validated enhancers, e.g. the validation rate, for a varying number of the top genome-wide predictions. Figure 3 illustrates the validation rates for the predictions obtained in cell line K562 by the different methods trained on the the pure random regions. In Figure 3, the numbers on x-axis correspond to number of enhancers predicted by: RFECS with the prediction threshold 0.5 (30593), the Bayesian approach with the 1% FPR threshold (51838), the Bayesian approach with the threshold 0.5 (89859), the ML approach with the 1%FPR threshold (92509), and the ML approach with the threshold 0.5 (127857). In addition to computing the validation rates of the top predictions, the validation rates for the random regions of equal size were also computed. For comparison, the validation rates of the ChromHMM Weak and Strong Enhancers clusters were

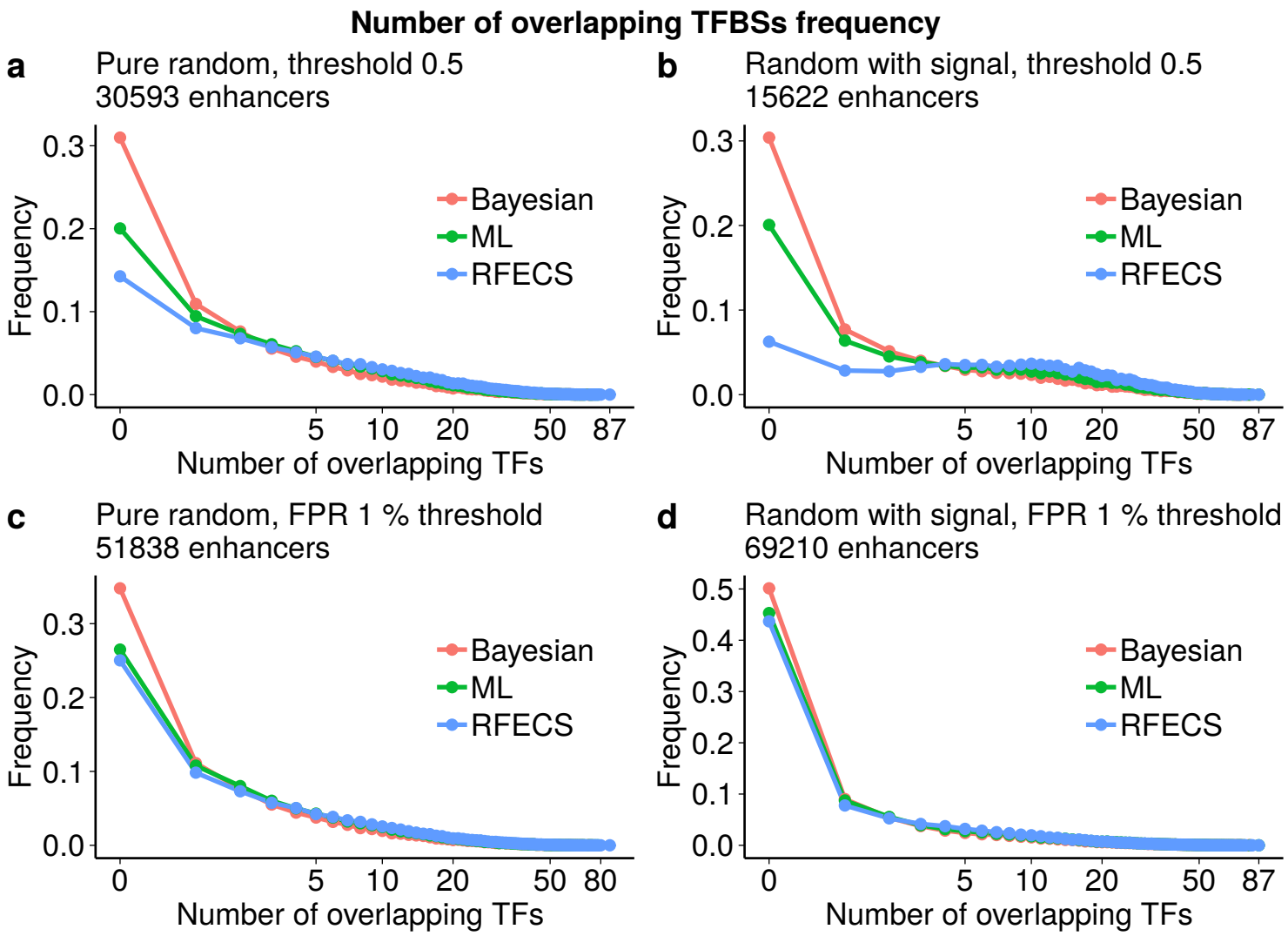

Figure 2: The proportions of the genome-wide enhancer predictions having an overlap with the varying number of ChIP-seq peaks in cell line K562. The proportions are shown for the different random region definitions and for the different thresholds. In **a** and **c**, the methods were trained on the pure random regions, and in **b** and **d**, the methods were trained on the random regions with signal. The number of enhancers in each comparison are shown above the figure. In **a** and **b**, the number of enhancers was the minimum number of enhancers predicted by any of the methods with the threshold 0.5, and in **c** and **d**, the number of enhancers was the minimum number of enhancers predicted by PREPRINT methods with their 1% FPR thresholds.

provided. To conclude, the validation rates of enhancers predicted by any of the methods were clearly higher than the validation rates of the random regions. When comparing the different methods, the validation rates were higher for RFECS than for PREPRINT when the number of enhancer predictions were low (30593 and 51838), but when considering a high number of enhancers (89859 and higher), PREPRINT reached comparable or even higher validation rates. Notably, for the high number of enhancers, the predictions obtained by the ML approach have higher validation rate than than the predictions obtained by RFECS. Similar results were obtained in cell line K562 when the methods were trained on the random regions with signal (See Supplementary Figure S8, Additional File 1). When trained on the random regions with signal, the validation rates were in general lower than for the methods trained on the pure random regions. In addition, there were less differences in the validation rates between the PREPRINT methods. Supplementary Figures S9 and S10, Additional File 1, shows the validation rates in cell line GM12878. In cell line GM12878, the validation rates were higher for PREPRINT than for RFECS even on a modest number of enhancers (80000–90000), especially for the ML approach. These results suggest that RFECS may predict a restricted set of enhancers with the strongest chromatin feature signals, whereas PREPRINT can predict a larger number of enhancers, containing enhancers with both strong and weak feature signals. When requiring the methods to predict a larger number of enhancers, the PREPRINT enhancers have a higher validation rate.

Table 3: The AUC values for the genome-wide predictions. The true labels of the predictions were based on the overlap between the predictions and the validation data ChIP-seq peaks. An equal number of enhancers predicted by PREPRINT and RFECS were chosen; the number was the minimum number predicted with threshold 0.5 over all methods. In each setting of the validation data, method, and the random data definition, the AUC value of the best method was highlighted with the bold font. In addition, the AUC value of the TF validation data, the GM12878 cell line, the PREPRINT ML approach, and the pure random set was highlighted due to comparable generalization performance (AUC = 0.91) to RFECS.

		AUC p300		AUC TF	
Cell line	method	pure	with signal	pure	with signal
K562	ML	**0.837**	0.826	0.884	0.854
	Bayesian	0.809	0.797	0.839	0.811
	RFECS	0.821	**0.916**	**0.907**	**0.929**
GM12878	ML	0.831	0.792	**0.910**	0.841
	Bayesian	0.821	0.750	0.879	0.765
	RFECS	**0.875**	**0.876**	**0.936**	**0.943**

The overlap between predictions made by different methods

We investigated the overlap of predictions obtained by the different methods. PREPRINT and RFECS predictions were considered as 2 kb windows. For an overlap of any two enhancers predicted by any two methods,

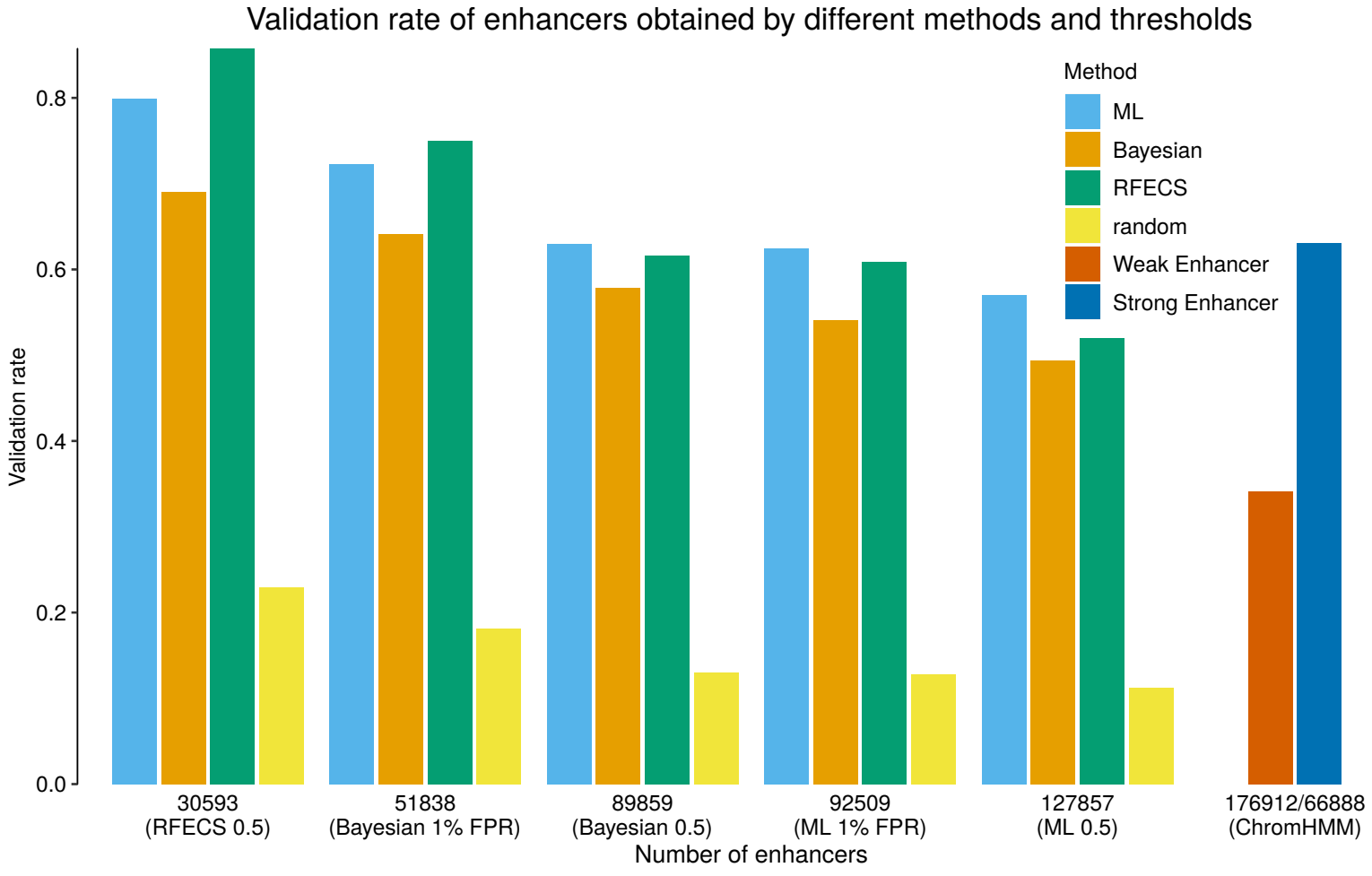

Figure 3: The validation rate of the genome-wide enhancer predictions obtained by the different methods and thresholds in cell line K562. The methods were trained on the pure random regions. An enhancer prediction was validated if it overlapped at least 1 bp of at least one TF or co-regulatory ChIP-seq peak.

the minimum required overlap was 1 bp. As an enhancer predicted by the one method might overlap with two enhancers predicted by the other method, the overlaps between enhancers predicted by different methods are not symmetric for each pair of methods. Hence, the overlap was computed in both directions. The numbers of unique and overlapping genome-wide predictions obtained by the different methods were illustrated as Venn diagrams. In each are of the Venn diagram circles, the percentages of the validated enhancers were provided. The validation was again performed as described above. Figure 4 shows a Venn diagram of the predictions obtained in cell line K562. In the Venn diagrams, the numbers of PREPRINT and RFECS enhancers were the same, and the number was chosen to be the minimum number of enhancers predicted by PREPRINT or RFECS with the 0.5 prediction threshold. In Figure 4, around half of the enhancers predicted by PREPRINT or RFECS were predicted by the all three methods, and this set had the highest validation rate (around 90%). In the set of predictions shared by all methods, the number of enhancers predicted by ChromHMM was much larger. Hence, their validation rate was lower (around 70%), likely reflecting the fact that ChromHMM enhancers were not very precise, or the cluster labels along the genome were altered quite often between the enhancer state and the other states, or both. Furthermore, the enhancer predictions shared by two methods had rather high validation rates (60–90%), and the enhancers predicted uniquely by only one method had the validation rate range of 40–90%. Of the enhancers uniquely predicted by the different methods, the RFECS enhancers tended to have the highest validation rate, although the number of unique RFECS enhancers was smaller than the number of unique enhancers for PREPRINT or ChromHMM.

Supplementary Figure S11, Additional File 1, shows the Venn diagrams for larger sets of enhancer predictions in cell line K562. The number of enhancers were equal to the number of enhancers predicted by PREPRINT with the 1% FPR threshold. In accordance with the results obtained for the smaller set enhancers shown in Figure 4, the unique RFECS enhancers validated better than the unique PREPRINT enhancers, in spite of the fact the number of the unique PREPRINT enhancers was smaller than the unique RFECS enhancers. The validation rates were overall smaller than in Figure 4, except for the unique enhancers predicted by ChromHMM. This was a result of PREPRINT and RFECS beginning to cover come ChromHMM enhancers when lowering their prediction thresholds; the unique enhancer set for ChromHMM became smaller, and their validation rate improved. In addition, the number of overlapping regions between PREPRINT and RFECS was higher for PREPRINT, which is also observed in Figure 4, implying that one RFECS prediction overlapped with multiple neighbouring PREPRINT predictions. Finally, the Supplementary Figure S12, Additional File 1, shows the Venn diagrams between predictions obtained by PREPRINT and RFECS utilizing different thresholds and the random data definitions. The number of enhancers in each comparison was equal for all methods. Around half of the enhancers predicted by any method were found by all methods, and those had the highest validation rate (80–90%). The overlapping enhancers between PREPRINT and RFECS validate better than overlapping enhancers predicted by the ML and Bayesian approaches. There are still significant numbers of enhancers predicted by two methods or by one method only. RFECS predicted the highest number of unique enhancers which also had a high validation rate. Of the unique predictions made by PREPRINT, the ML predictions had the highest validation rate, except in comparison d in Supplementary Figure S12, Additional File 1.

Supplementary Figure S13, Additional File 1, provides the Venn diagram for the predictions in cell line GM12878. The number of enhancers for PREPRINT and RFECS was the minimum number of enhancers predicted by any of the methods with the 0.5 threshold. In addition, Venn diagrams for predictions in cell line GM12878 when using the 1% FPR threshold estimated either from the K562 CV data or from the GM12878 test data are shown in Supplementary Figures S14 and S15, Additional File 1, respectively. The results were comparable to ones obtained for cell line K562. In comparisons a and b in Supplementary Figure S14, Additional File 1, the unique enhancers predicted by PREPRINT reached high validation rates, 50% and 27% for the ML and Bayesian approaches, respectively. In addition, the validation rates of unique PREPRINT enhancers were comparable to the RFECS unique enhancers when the methods were trained on the random regions with signal (comparisons c and d in Supplementary Figure S14, Additional File 1). Moreover, when using the the 1% FPR threshold estimated from the GM12878 test data, the unique predictions of PREPRINT ML approach have a higher validation rate (37%) compared to the unique predictions made by RFECS (26%) (comparion a in Supplementary Figure S15, Additional File 1). However, in the comparison a, the number of the unique PREPRINT predictions were much smaller than the unique RFECS predictions. Furthermore, the Supplementary Figure S16, Additional File 1, shows the Venn diagrams of enhancers predicted between predictions obtained by PREPRINT and RFECS utilizing different thresholds and the random data definitions. Again the results are similar to ones obtained for cell line K562 (Supplementary Figure S12, Additional File 1). Finally, Supplementary Figure S17, Additional File 1, shows the overlapping enhancers predicted by the methods trained on the different random data definitions. In this comparison, about half of the enhancers were shared between the two approaches, but the rest were unique for a random data definition. The unique enhancers predicted by PREPRINT trained on the pure random data had higher validation rate (around 50%)

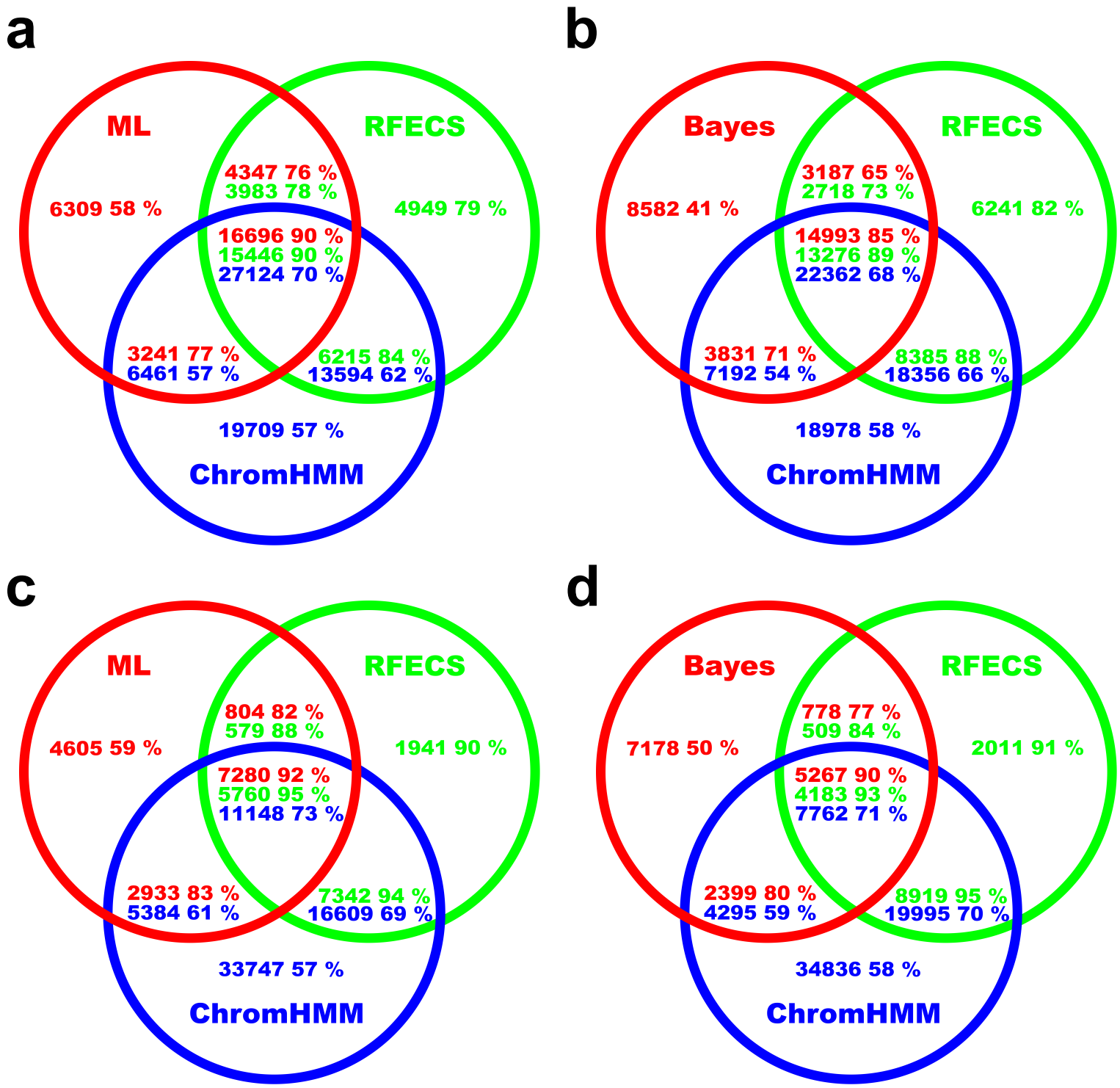

Figure 4: The unique and overlapping genome-wide enhancer predictions made by different methods in cell line K562. In **a** and **c**, the predictions were obtained by the ML approach, and in **b** and **d**, the predictions were obtained by the Bayesian approach. The overlap between the PREPRINT, RFECS and ChromHMM predictions were quantified as the number of enhancers. In figures **a** and **b**, PREPRINT and RFECS were trained on the pure random regions, and in **c** and **d**, PREPRINT and RFECS were trained on the random regions with signal. In each figure, the number of enhancers predicted by PREPRINT and RFECS was the same. The number of enhancers was chosen to be the minimum number of enhancers predicted by either PREPRINT or RFECS with the 0.5 threshold. The numbers were: **a** 30593, **b** 30593, **c** 15622, and **d** 15622. Inside every area, the number of enhancers belonging to the set is shown together with the proportion of validated enhancers in the set. The overlapping areas are not proportional to the number of overlapping regions due to the asymmetry of overlaps.

compared to validation rate of the unique enhancers predicted using the random regions with signal (around 20%). By constrast, the unique enhancers predicted by RFECS trained on the random regions with signal had a higher validation rate (around 70%) compared to RFECS trained on the pure random regions (45%). Finally, RFECS was less sensitive to the random set definition than PREPRINT.

As a conclusion, PREPRINT trained on the K562 data generalized well on the GM12878 data and the ML approach trained on the pure random regions performed generally well and in some comparisons the Bayesian approach obtained similar performance to RFECS. However, it was difficult to compare the validation performance of the overlaps between different methods, for example, as the number of the unique enhancers varied greatly between the methods. Overall, these results again indicated that the choices of the training data, the prediction threshold, the choice of classification method, and the definition for an overlap greatly influenced on the final set of enhancer predictions, their validation rate, and the overlap between enhancers obtained by different approaches. Moreover, multiple PREPRINT predictions tended to overlap with one RFECS prediction, complicating the comparisons further. The asymmetric overlap was likely due to PREPRINT prediction scores along the genome being more stepped than the smoothly changing RFECS prediction scores, and due to RFECS requiring at least 2 kb distance between the individual predictions. PREPRINT tended to predict multiple individual enhancers within a short stretch of DNA, and there were no requirement for the distance between the PREPRINT predicted enhancers.

Some examples of validated enhancers uniquely predicted by PREPRINT

Some examples of validating enhancers uniquely predicted by PREPRINT were visualized in a genome browser together with the chromatin feature data. Figure 5 shows examples of the predicted enhancers in cell line K562 in a 20 kb genomic window in chromosome 1. Starting from top, the validating predictions (red arrows) and a false negative prediction (blue arrow) are shown.Below the arrows, the predictions made by PREPRINT and RFECS are provided, together with the predictions scores and the 0.5 threshold line. In addition, Figure 5 displays the ChromHMM predictions for cell line K562, the GENCODE genes, a subset of the chromatin feature ChIP-seq data tracks, and the Uniform ChIP-seq peaks used for validation. The same figure with all 15 features and validation data Chip-seq peaks is provided in Additional File 3.

The false negative prediction denoted by the blue arrow occupied the binding sites of TFs and co-regulatory proteins, but it was not identified as an enhancer. The false negative prediction was not necessarily an enhancer as it did not show the typical enhancer feature pattern; it was likely some other functional site. Moreover, according to the peaks in the DNase-seq data, there were three strong DNase-seq peaks within this genomic region corresponding to open chromatin. The three strong DNase-seq peaks were likely true enhancers. The false negative finding did not overlap with a strong DNase-seq peak. By contrast,t he validating unique PREPRINT enhancers clearly demonstrated the characteristic enhancer features. However, in addition to the validated predictions, PREPRINT predicted some invalidated enhancers close to the validated ones. It was unclear, whether the invalidated enhancers should be considered the same enhancers as the validated ones, as enhancers do not have clear boundaries. The signals of H3K4me1, H3K27ac and DNase I HS at these sites showed the characteristic enhancer patterns, but the signal intensity was low; hence they might have been just background noise. An example of an invalidated enhancer is shown at the last exon of the CROCC gene (the green version of a gene). The enhancer was a prediction obtained by PREPRINT with the Bayesian approach trained on the random regions with signal. The prediction showed a weak bimodal peak for H3K4me1, histone variant (H2AZ), histone 3 lysine 27 trimethylation (H3K27me3) and MNase-seq peak (see Additional File 3), and therefore, could be a true enhancer.

Figure 6 shows examples of enhancers predicted in cell line GM12878 in a 25 kb window in chromosome 17. The full set of data is provided in Additional File 4. A unique PREPRINT enhancer (red arrow) presented characteristic H3K4me1 pattern, and was validated by the Uniform ChIP-seq peaks. The green arrows indicating validating loci were likely promoters based on their distance to the TSS of the GENCODE genes. The second rightmost green arrow corresponded to a unique PREPRINT enhancer, but it lacked the validating signal and was likely again too close to a promoter. As seen in Figures 5 and 6, the prediction scores of PREPRINT changed gradually from low values to high values in a stepped fashion, whereas the prediction scores of RFECS increased and decreased smoothly. In Figure 6, PREPRINT trained on the pure random regions predicted wide regions (for multiple subsequent windows the prediction score was above 0.5), complicating the identification of enhancers at high resolution. In contrast, the prediction scores of PREPRINT trained on the random regions with signal were more stepped. However, PREPRINT trained on the random regions with signal predicted vary many enhancers within the 25 kb window, probably overstating the number of enhancers. Furthermore, the multiple subsequent enhancers seemed to occur periodically with about the same distance between them. By looking at the different feature signals, such as MNase-seq, H3K4me2, histone 3 lysine 36 trimethylation (H3K36me3) and

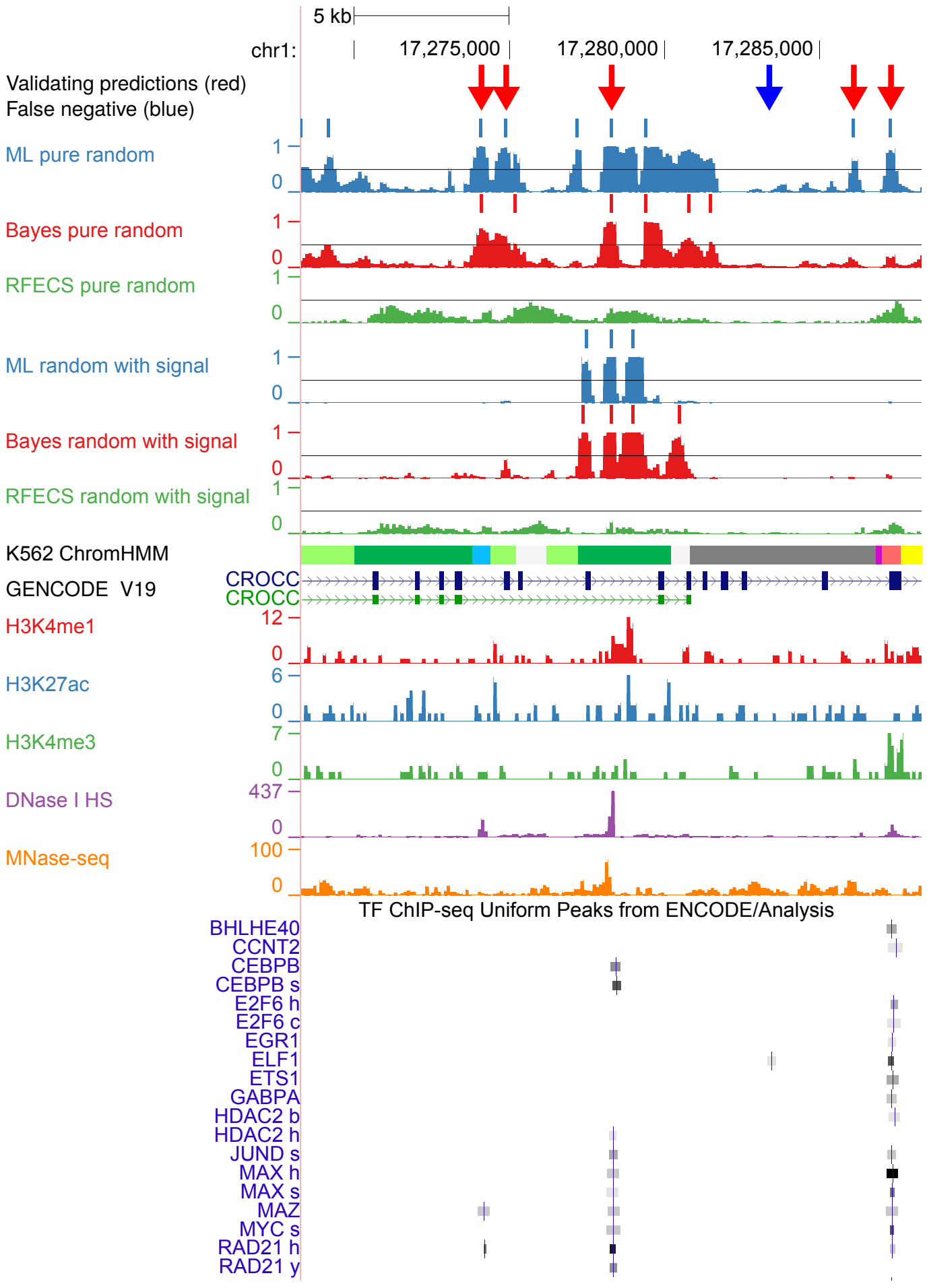

Figure 5: Genome browser visualization of examples of enhancers uniquely predicted by PREPRINT. Data was from cell line K562. Color codes for ChromHMM clusters: light green: Weak transcription, dark green: Transcription elongation/transition, blue: Insulator, gray: Repetitive/CNV or repressed, purple: Poised promoter, light red: Weak promoter, yellow: Weak enhancer.

H3K20me1 (see Additional File 4), there were clear peak-valley-peak patterns occurring subsequently in the genome, and the PREPRINT trained on random regions with signal was sensitive to these and falsely predicted enhancers in the locations of valleys.

To conclude, PREPRINT predicted apparent enhancers not predicted by RFECS or ChromHMM. However, the visual inspection of the properties of the predictions reflected the results and challenges reported in the previous chapters: First, the challenge of defining the length of an enhancer, and predicting the enhancer location with a high accuracy. Second, the challenge of defining the optimal prediction threshold. Third, when validating the predictions, the difficulty to choose the number of the overlapping ChIP-seq. Fourth, the definition of an overlap between two genomic regions or the set of regions, and finally, all the other parameter choices made during the analysis (e.g. for the required distance between enhancers and promoters). In other words, visual inspection of the individual predictions further justifies the concerns of being cautious when drawing conclusions about the enhancers predicted by machine learning methods.

Discussion

The supervised machine learning methods for the enhancer prediction task may not have adopted the full potential of the probabilistic approach. Therefore, we developed a PRobabilistic Enhancer PRedictIoN Tool PREPRINT. Earlier studies and the results presented here indicated that the prediction task is challenging; the different methods predicted diverging sets of enhancers having varying validation rates. The reasons for the inconsistencies are various. First is the choice of the prediction score threshold. Using cross-validation within the training data, the best operating-point threshold (close to 0.5) and the threshold giving the 1% false positive rate were obtained for PREPRINT. The threshold estimated from the training data should generalize to the whole genome data and to the data originating from other cell lines. We did not estimate the corresponding thresholds for RFECS, however, the 0.5 threshold for RFECS might be too stringent resulting in a low number of enhancers, and optimizing the RFECS threshold could affect the results from the method comparisons. Second, the methods compared in this work predicted very wide genomic regions as enhancers, and the exact enhancer location inside a large window needed to be identified. The approaches used to pinpoint the individual enhancers were likely suboptimal. Although RFECS often had the best validation performance, even RFECS was not able to pinpoint enhancers very accurately, due to the smoothly increasing and decreasing prediction scores along the genome. The ChromHMM enhancers were not associated to any prediction scores to reflect their significance. The prediction scores could have been derived from the posterior probabilities of the ChromHMM cluster assignments for genomic windows of length 200 bp, but that was beyond the scope of this paper. In addition, the unsupervised methods like ChromHMM switch unnecessarily often between states. Third, the training and testing of the methods and the performance measure computations could have been done on multiple random subsets of the training and test data to evaluate the uncertainty in the AUC values. Finally, according to the results, ML trained on the pure random regions performed often better than the Bayesian approach or the methods trained on the random regions with signal. This might imply that the individual sample's fit, for example, to the enhancer average coverage profile should be modeled locally, instead of using a global genome-wide distribution of the scaling parameter as in the Bayesian approach.

Although a clustering method, ChromHMM was included into the comparison due to its popularity. RFECS does not make any distributional assumptions and is considered as the state-of-the-art method for supervised enhancer prediction. RFECS is also claimed to utilize the coverage pattern information, hence the relative performance of PREPRINT to RFECS is of interest. PREPRINT assumed a non-negative count data having distribution resembling the Gamma-Poisson mixture or the negative binomial distribution. Even the distributional assumptions made in this work might be suboptimal. Different distributional assumptions could be used: the count data could be corrected by the log-concave Poisson approach [48], or the exact negative binomial distribution could be used. As the subtraction of normalized input signal from the ChIP-seq signal resulted in continuous and negative values, instead of the Gamma-Poisson mixture model, distributions like Skellam [49] or distribution that models the difference between two independent negative binomial random variables [50] could be used. Moreover, the enhancer and non-enhancer pattern averages and the uncertainty inhered in them could be modelled, in contrast to a fixed mean assumed in this work. However, this would lead to a more challenging estimating procedure, such as the expectation-maximization (EM) algorithm. In addition, the usage of continuous and real valued ChIP-seq data likely caused RFECS to perform superior to PREPRINT as the data discretization and especially the conversion to nonnegative values both lose information. In other words, the negative dip between the signals from the two well-positioned nucleosomes at enhancers helped RFECS to separate enhancers from the random background signal. Moreover, the differences in the way the methods utilized the input signal for DNase-seq and MNase-seq signals slightly invalidated the comparison, although the differences likely had only a modest effect on the results. Finally, the ChIP-seq data for histone modifications

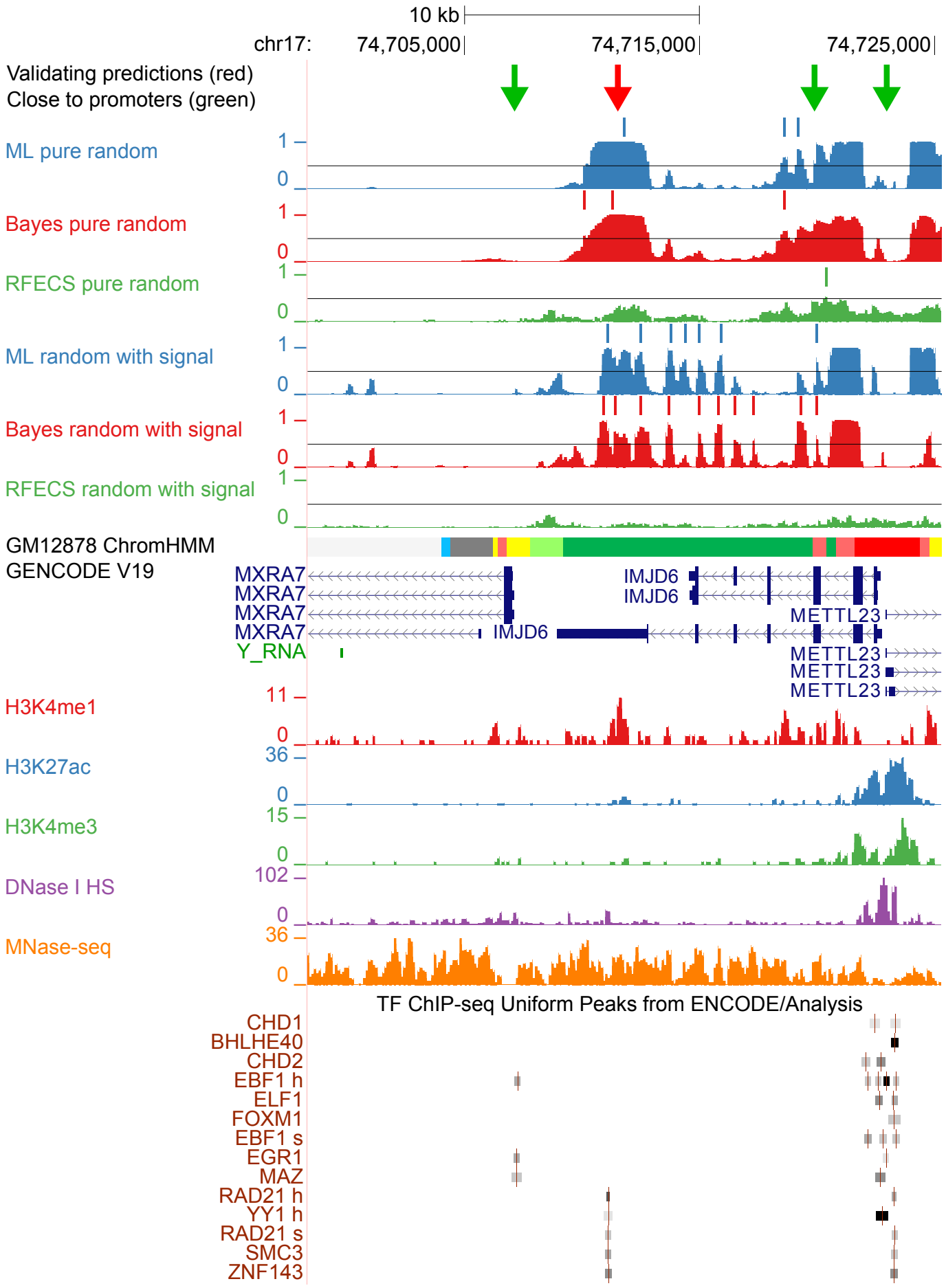

Figure 6: Genome browser visualization of examples of enhancers uniquely predicted by PREPRINT. Data was from cell line GM12878. Color codes for ChromHMM clusters: Orange: Strong enhancer, yellow: Weak enhancer, bright red: Active promoter, light green: Weak transcribed.

should be deeply sequenced for the coverage signals to be saturated. The GENCODE data used in this work is likely outdated in this sense, and more deeply sequenced data could be used. However, deeper sequencing brings additional costs, and thus the probabilistic methods are likely valuable to detect the weak signals from the insufficiently sequenced samples.

As was shown in this work, the choice of the set of trainin gdata random regions affected the performance of the methods. In addition to the random regions and promoters, some other definitions of non-enhancers could be used, such as promoters driving different levels of gene expression, inactive promoters, exons, introns, miRNA and other non-coding genomic sites exhibiting at least some epigenetic signals. The profiles at individual non-enhancer regions could be scrambled to generate a versatile set of non-enhancer examples. Moreover, the enhancer prediction task is an example of a class imbalance problem; the number of non-enhancers greatly outnumbers the number of enhancers genome wide. In previous studies, the ratio of enhancers to non-enhancers has been set to 1:10 for the training or validation set or both [41]. In unbalanced classification, area under precision-recall curve should be computer instead of the area under the ROC curve. Moreover, although not studied in this work, the choice of training data enhancer definition likely affects the results. To study the effect of the training data enhancer definition, the enhancers could be clustered to identify subclasses of among them. The enhancers belonging to separate clusters might display different patterns and intensities of chromatin features [51, 52, 53]. It would be interesting to correlate the obtained clusters to biological and functional properties of the enhancers to identify enhancer subtypes or enhancer states, such as poised, silenced, primed and active enhancers. The clustering could be done prior to classifier training. Furthermore, tn this work, the middle base of an enhancer was defined as the summit of the p300 ChIP-seq peak; this approach has naturally an inhered uncertainty associated with it. The individual ChIP-seq profiles could be shifted to improve the alignment between the enhancers. Moreover, the distribution of nucleosomes might vary between different enhancers, for example, the distance between the two well-positioned nucleosomes varies, and this should be considered in the prediction task.

There are no generally accepted definitions of an enhancer and its common features; the different definitions and features have different strengths and weaknesses when predicting enhancers based on them [54, 31, 47]. The functional role of the histone modifications and the regulatory proteins at enhancers are still largely unknown, and new histone modifications characteristic of enhancers might still to be found. Even the functional role of the chromatin environment at promoters to the expression of their target genes is not clear [55]. In the enhancer prediction, the requirement of the presence of histone 3 lysine 27 acetylation (H3K27ac) with high levels of the histone 3 lysine 4 monomethylation (H3K4me1) and low levels of the histone 3 lysine 4 trimethylation (H3K4me3) has been largely utilized, but this approach has drawbacks and it generates considerable false positives and false negatives [56, 57]. The presence of H3K4me1 is not necessarily required for a functional enhancer [58, 59]. In flies and mouse embryonic stem cells (ESC), the most active enhancers are enriched with H3K4me3 rather than H3K4me1 [56]. The biological validation of enhancers predicted by ENCODE [60] revealed that the H3K27ac and H3K36me3 depleted Weak Enhancer class drove higher gene expression in the Cis-regulatory element analysis by sequencing (CRE-seq) reporter assays than the Strong Enhancer class [61, 35]. The active enhancers possessing high levels of H3K4me3 bears a resemblance to the promoters; indeed, enhancers have been shown to function as promoters by producing enhancer RNA, and some promoters have been found to function as enhancers [62]. Even a continuum of cis-regulatory region spectrum has been proposed; thus, the promoters and enhancers may represent the extreme ends of the spectrum [63].

In this work, to date the largest collection of TF and co-regulatory binding sites quantified by ChIP-seq was used to validate the genome-wide predictions. For an enhancer to be validated, at least 1 bp overlap was required between an enhancer prediction of size 2 kb and at least 1 TF or co-regulatory protein ChIP-seq peak. This definition of validation is not without problems: First, the width and the uncertainty of the ChIP-seq peaks vary depending on the antibody specificity in the ChIP experiment, the ChIP-seq data quality filtering, preprocessing of the raw reads ,and the peak-calling methods. Some peaks might have been missed as they reside just below the selected significance threshold, often selected arbitrarily. The human genome is estimated to encode around 1700 transcription factors [64] whose binding sites in more than 200 different cell lines are largely unknown, for example, due to lack of antibodies. In addition to TFs, there are also a large number of co-regulatory binding proteins, of which many are also unknown. Second, not all TSS-distal TF or co-regulatory protein binding sites are enhancers; they might be some other type of regulatory regions or repressed enhancers. However, our training data does not support identification of the latter. Third, there is an inherent uncertainty in the exact location of the predicted enhancers, as the subsequent enhancer prediction windows can be very wide (e.g. Figure 1). The 2 kb window centered at an enhancer was chosen to investigate the overlap between an enhancer and any Uniform ChIP-seq peak, although narrower window, such as 500 bp, could have been adopted to reveal the method's performance on predicting the exact locations of enhancers. Fourth, it was difficult to choose the requirement for the number of different TF and co-regulatory protein binding sites for,

example, based on Figure 2. Finally, for the AUC values to be comparable in Tables 3, the numbers of enhancers and non-enhancers should be the same in each comparison when comparing the different random data sets or different validation (p300 or TF) data sets.

With the advancements in the genomic data generation, big data analysis methods and machine learning, the race towards the ultimate genome annotation will accelerate with the emergence of new types of next-generation sequencing data sets. In addition to the 15 enhancer features used in this study, some other features could be used. Of the co-regulatory factors, cohesin could be used as a mark of an enhancer [65, 4]. DNA methylation data has also been used to predict enhancers [66]. Other potential useful features could be the information about the CpG islands, phastCons evolutionary conservation scores, TF motifs or DNA sequence itself. The larger collections of data containing more cell lines and larger sets of histone modifications, such as the Epigenetic Roadmap data or ENCODE Phase 3 and 4 data, could be used for enhancer prediction. Still today, not all possible genomic features are measured for all possible cell lines. Therefore, feature selection should be used of reveal the most important data types for the enhancer prediction. Using only a subset of highly discriminative features would decrease the costs and aid the biological interpretation of the findings. There is clearly a need for a large-scale open contest for benchmarking computational and experimental methods for enhancer prediction, also suggested by others [47, 41]. This is still hindered by the lack of sufficient data for different cell types and the lack of a gold-standard set of positive and negative enhancers. The set of massive parallel reporter assay (MPRA) validated enhancers could be used as the training data [35, 67], although these sets are still rather small in the human cells. Moreover, the MPRA approaches have their limitations: they include only a small segment of the the predicted regulatory DNA, and the plasmid-based assays do not consider the chromatin environment of the predicted region. The choice of training data determines whether to predict only active enhancers driving currently the expression of their target genes, or enhancers which are primed to activate in response to some intracellular or extracellular signals. The data used in this work originate from a population of cells. With the recent advancements of single cell or small cell population ChIP-seq techniques [68, 69], methods to identify enhancer and their chromatin landscape within single cells could benefit from probabilistic approaches, as the single cell data is very noisy.

This work demonstrated that the length of an enhancer and the enhancer boundaries are hard to define. The length of an enhancer can be defined as: the distance between the two well-positioned nucleosomes, the length of the stretch of the DNA containing motifs for enhancer binding TFs, or the shorterst possible DNA sequence driving, for example the reporter gene expression [47]. The length and boundaries are on the one hand related to the resolution of the enhancer prediction. In this and many other works, the ChIP-seq data was presented in 100 bp bins along the genome. The bin size could be decreased down to 1 bp. However, this would increase the memory and time requirements of the computational methods. Using data with a higher resolution has not been studied or exploited in its full potential. On the othder hand, the resolution of the enhancer prediction refers to the distance between the prediction center and the center of the true enhancer. The smaller the distance, the better the prediction resolution. This work demonstrated that the exact pinpointing of an enhancer is challenging: The centers of the training data enhancers were likely uncertain, and the methods predicted multiple subsequent genomic windows with large prediction scores as enhancers, and it was difficult to pinpoint an individual enhancer centers within a large window. When validating the predictions, a large 2 kb window centered at the prediction was considered, not the enhancer center. The enhancer center as a concept is also no properly defined, for example, it might refer to the p300 ChIP-seq peak summit, the p300 peak center, or the middle base between two well-positioned nucleosomes. Furthermore, recently a concept super enhancer has gained a lot of interest [70, 71, 72]. Super enhancers are large genomic (several kb) regions with broad and strong enhancer feature signals. The super enhancers are clusters of multiple enhancers, the well-known beta-globin locus control region is an example of super enhancer containing 5 individual enhancers [73, 74]. Super enhancers are found across diverse cell types [70, 71, 75], they contain high levels of H3K27ac, are occupied by enhancer-associated proteins, such as Mediator and RNA Pol II, they are cell type-specific, and regulate the genes controlling cell state. Tumor cells often acquire super enhancers near oncogenes [76, 77], and super enhancers are also remarkably enriched for GWAS identified SNPs that associate to several common diseases [78]. Some of the training data enhancers defined in this work might reside at the super enhancers. This might have resulted to mixing of the chromatin feature signals at nearby training data enhancers belonging to the same super enhancer cluster. A solution could be to filter out the super enhancers from the training enhancer set. Nevertheless, it would be very useful for the enhancer prediction method to able to pinpoint the individual enhancers within the super enhancer cluster. In Figure 6, PREPRINT trained on the random regions with signal tended to predict multiple subsequent enhancers within a larger region, which would be desirable when predicting the super enhancers. However, the behaviour was likely a result of PREPRINT being overly sensitive to peak-valley-peak patterns in the feature signal, or PREPRINT having difficulties to generalize to the GM12878 data. Similar behaviour was not observed in the K562 genome browser example.

In addition to locating enhancers in the genome, their target genes should be identified. This is especially important when inferring the GWAS SNPs and estimating the functional effects of the non-coding genetic variants. The enhancer-promoter interactions can be quantified using various high-throughput chromatin conformation capture methods [79, 80, 81, 82, 83]. Especially, the individual enhancers at super enhancers physically interact with one another, and these interactions together with the interactions of target promoters are likely essential for the super enhancer function [83]. However, the resolution of the chromating conformation capture methods has not been sufficient enough to reveal the interactions between individual regulatory elements, such as enhancers and promoters. In addition, recentlya concept called hub enhancer was defined, the knockout of a hub enhancer within a super enhancer cluster resulted in a significant decrease in the super enhancer activity [84]. The hub enhancers are occupied by CTCF and cohesin, the first utilized as one feature for enhancer prediction in this work, the latter recognized as a potential feature mark for the future studies.

Conclusion

Despite the development of several unsupervised and supervised enhancer prediction methods, the vast data generation in various cell lines, the full list of enhancers in the human genome is still incomplete. When predicting enhancers, the following questions rise: First, which genomic and chromatin features should be used for enhancer prediction? In this work, the common features of the ENCODE first data production phase Tier 1 cell lines K562 and GM12878 were used. Especially, the RNA Pol II and CTCF Chip-seq data and MNase-seq data were employed, the types of data that are rarely included as features when predicting enhancers. Second, how many functional enhancers there are in each individual cell type or across all human cell types? Why do different methods predict different number and different sets of enhancers, and why their validation rates differ? How the generalization of the classifiers between cell lines could be improved? We developed a PRobabilistic Enhancer PRedictIoN tool PREPRINT that utilized the pattern of, for example, the histone modification ChIP-seq coverage profiles at the genomic region of interest. We believe that the probabilistic approach and the modeling the profile pattern are not utilized to their full potential. We studied the performance of PREPRINT and the competing methods in predicting enhancers in the two cell lines. The selected prediction threshold influenced the final number of predicted enhancers. Our method performed comparably to the state-of-the-art methods, and provided uncertainty estimates for the predictions. In addition, PREPRINT was shown to generalize well between the cell lines. Third, how the training data should be defined? We experimented with different definitions of the non-enhancer examples in the training data and showed that the choice of the training data had a notable effect to the method performance. Fourth, how to evaluate the genome-wide predictions? To validate the genome-wide predictions we used the the large set of TF and co-regulatory protein binding sites quantified by ChIP-seq. This approach had several limitations and included choosing the parameter values from a set of options. Nevertheless, PREPRINT predicted unique enhancers not predicted by the competing methods, and some of the unique enhancers validated based on the ChIP-seq peaks. the enhancers predicted by PREPRINT tended to overlap a smaller number ChIP-seq peaks. Finally, the set of enhancers predicted by different methods overlapped significantly.

Accurate annotation of the regulatory regions across the human genome is a prerequisite for the interpretation of the findings of regulatory genomics studies. In this work, a machine learning tool is developed to obtain a set of genome-wide enhancers in two cell lines. The predicted enhancers can be utilized in the genome interpretation in functional genomics studies as well as in clinical studies. In the future, the next-generation sequencing methodologies will likely transition to standard clinical tests, and in clinical diagnostics, the analysis will be broadened from genotyping the protein-coding genes towards profiling the non-coding DNA.

There are many general problems related to the computational genome-wide enhancer prediction, and future studies are necessary. There is a need to generate a golden standard set of enhancers to benchmark the computational methods and to find the most relevant enhancer features. In the supervised setting, the non-enhancer set definition should be optimized. Moreover, there are likely enhancer subsets possessing different features and feature patterns. Therefore, clustering as a preprocessing step or semi-supervised approaches should be adopted. The resolution of the used data and the prediction precision and accuracy should be increased, for example, to pinpoint the individual enhancers at super enhancers.

Methods

Overall approach

In many earlier studies, the binding sites of co-regulatory proteins and histone acetylases, such as CBP orp300, have been used to identify enhancer locations [20, 19, 22, 21, 42]. The transcription start site (TSS)-distal p300 binding sites overlapping DNase I HS peaks were considered as examples of true enhancers. The data at 1000 enhancers is shown in Figure 7. The data was presented at a window of 4 kb centered at each enhancer location, the window was divided into 40 bins of length 100 bp, and for each bin, the coverages of different chromatin feature signals awere computed. In addition to the heatmap showing data at the individual enhancers, the average profiles of chromatin features are provided. According to MNase-seq data, there were two well positioned nucleosomes flanking the enhancer, and the nucleosomes were more mobile when moving further from the enhancer center. Many histone modifications, such as H3K4me1, formed bimodal peaks that colocalize with the two flanking, well-positioned nucleosomes. In comparison, the signal profiles at unoriented promoters are shown in Supplementary Figure S1, Additional File 1. On the one hand, the profiles of enhancers and promoters were similar, on the other hand, they were different. The enhancer prediction tool should be able to distinguish the enhancers from the promoters, as well as from the genomic background. Therefore, in addition to the promoters, two versions of random genomic regions were created: first, pure random regions sampled uniformly from the whole genome; second, random genomic regions displaying coverage above certain threshold. The latter were denoted as the random regions with signal. The data from two definitions of random regions are shown in Figures S2 and S3, respectively, in Additional File 1.

Different distributional assumptions about the next-generation sequencing data including ChIP-seq data have been made, such as Poisson [85], and negative binomial [86, 87], the log-concave Poisson approach [48], and Poisson log-normal [88]. The most common distributional assumption for the ChIP-seq coverage is the Poisson distribution. However, the coverage data shows widespread and consistent overdispersion, i.e. there is a large number of high-count bases, much more than expected from the Poisson distribution [89, 48]. The negative binomial distribution is equivalent to a Poisson distribution with the Gamma conjugate prior for the mean parameter. The single parameter of the Poisson distribution models both the mean and variance of the number of reads aligning to a genomic location. In constrast, the negative binomial has separate parameters for mean and variance, hence accounting partly for the inherent overdispersion in ChIP-seq data. When adopting the conjugate Gamma prior for the Poisson mean parameter, the posterior distribution of the mean parameter is still a Gamma distribution, whereas the posterior predictive distribution is negative binomial distribution, i.e. the mixture of Gamma distributions.

We believe that the coverage pattern at enhancers is important for the enhancer prediction, and therefore, we developed distance measures that quantified the similarity between the average enhancer signal profiles (seen in Figure 7) and the sample region to be tested. The distance measure can be as simple as correlation as in [20, 21], but the the distance measures were based on probabilistic modeling of the ChIP-seq signal. The patterns of the signals at individual enhancers are similar, but the signal intensity might vary. This was considered by modeling the read counts with a Poisson distribution and setting a conjugate Gamma prior for the Poisson mean parameter. The mean parameter was learned separately for each chromatin feature. Finally, computed the the posterior predictive value for each sample; these values were interpreted as the distance measures between the average enhancer signals and the sample signals. The variation in the signal intensity were modeled with both a maximum likelihood (ML) approach and a Bayesian approach, and the performance of the different approaches were evaluated. The distance measures were used to train a support vector machine (SVM) classifier. New enhancers were predicted genome-wide in both cell lines K562 and GM12878, and the generalization of the method to GM12878 data when trained on K562 data was studied. The performance of the new probabilistic method was computationally compared with the state-of-the-art methods, unsupervised ChromHMM and supervised RFECS.

Data

In this work, publicly available data from ENCODE Consortium was used [1]. The ChIP-seq raw reads from ENCODE/Broad Institute data set for 10 histone modifications, histone variant H2AZ and a protein CTCF were downloaded for myelogenous leukemia cell line K562 and lymphoblastoid cell line GM12878. RNA polymerase II data was downloaded from Transcription Factor Binding Sites by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard data set. DNase-seq data from Open Chromatin by DNaseI HS from ENCODE /OpenChrom (Duke University) data set and MNase-seq data from Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU were downloaded as already aligned (bam-format). Data for chromosomes chrY and chrM were excluded from all data. The paths to the downloaded files, their ENCODE Data Coor-

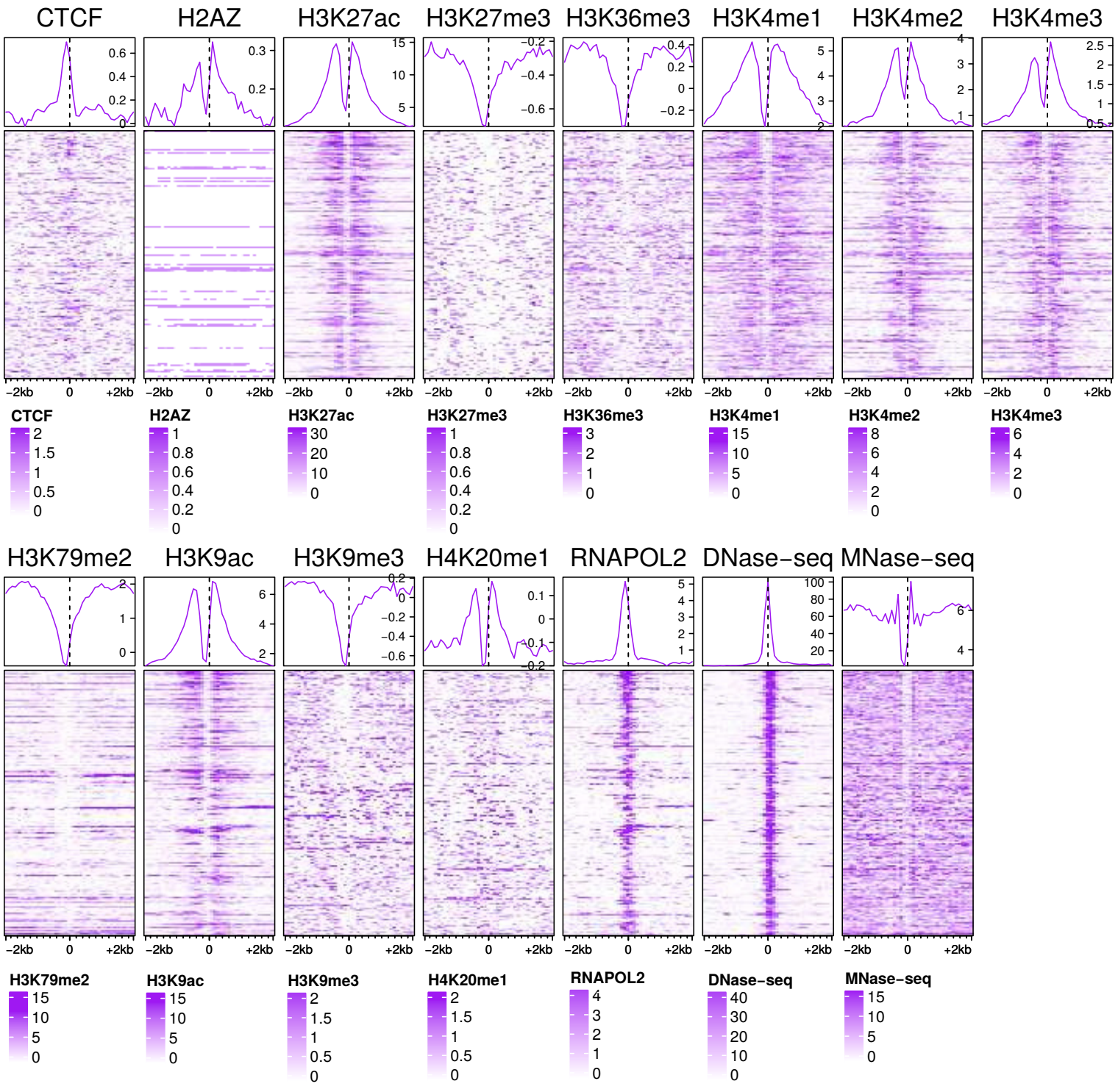

Figure 7: The coverage of different chromatin features at 1000 individual enhancers, and their average profile over all regions. The data is from cell type K562, features are presented in 4 kb window with bin size 100 bp.

dination Center (DCC) Accession names, Data Coordination Centers (ENCODE DCC), and Gene Expression Omnibus (GEO) sample accession number are provided in Additional file 2.

The ChIP-seq data were processed using the following steps:

1. Raw reads were aligned to the human genome version hg19 using Bowtie 2 [90] (bowtie2-2.3.3.1) with the default options

2. Reads mapping to exactly same location were considered as polymerase chain Reaction (PCR) duplicates [91] and only one of the duplicate reads was retained for the analysis

3. Possible isogenic replicates were pooled

4. The fragment lengths for ChIP-sed reads were estimated from cross-correlation profiles using phantompeakqualtools (spp version 1.14) [92, 93] and R version R-3.3.1.

5. The ChIP-seq reads were shifted by half of the fragment length using combination of bedtools2 and samtools, and the genome-wide coverage signals were generated. The MNase-seq reads were shifted by 149/2, the half of the length of DNA wrapping around a nucleosome (149 bps). The DNase I HS data and Input/Control data were not shifted. When creating the coverage signal, the Input coverage was normalized wrt. the ChIP coverage to equalize the library sizes, and the Input coverage was subtracted from the ChIP coverage. Data for GM12878 were normalized wrt. K562 library size. The normalization was done as previously described [94, 92, 95].

6. For PREPRINT, the input signal was not subtracted from the DNase-seq and MNase-seq signal. RFECS requires the input signal for all data types, as it is designed to use histone modification ChIP-seq data. Hence, RFECS normalizes and subtracts the input from the DNase-seq and MNase-seq signals.

7. For PREPRINT, the coverage was computed in 100 bp bins, the coverage values were rounded to the nearest integer, and the negative values were converted to zero. RFECS also utilized data in 100 bp bins.

Definition of the training data

The binding sites of histone acetylase P300 from Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Analysis were used to define the training enhancers in cell line K562 (SydhK562P300Iggrab) and test enhancers in cell line GM12878 (SydhGm12878P300Iggmus). In both cell lines, the training data enhancers were defined as the 1000 most significant (based on q-value) p300 binding sites, whose summit overlapped a DNase I hypersensitivity peak from Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University and). We required the distance between the training data enhancers and any protein coding TSS from Gencode v27 [96] to be larger or equal to 2 kb. The training data enhancers were centered at the p300 peak summits. See Additional File 2 for more details about the origin of the data.

The promoters in both cell lines were defined as the 1000 GENCODE v27 TSS that overlap a DNAse I HS peak and whose distance to any other TSS nearby was more than 2 kb. For each cell line, we selected the 1000 promoters overlapping the most significant DNase I HS peaks based on the p-values of the DNase-seq peaks. The training data enhancers and promoters, defined as the 5 kb regions centered at the p300 peaks or TSS, which overlapped ENCODE blacklist regions were excluded. The ENCODE blacklist regions contained repetitive elements, such as α- and β-satellite repeats, ribosomal and mitochondrial DNA, and some other regions, that are listed in the Mappability or Uniqueness of Reference Genome data set [97]. For more details about the ENCODE blacklist regions, see Additional File 2. At each training and test data location, a 2 kb window centered at the location was defined, and the coverage profiles of different features in 100 bp bins were computed. Figure 7 plots the individual profiles and mean coverage profiles of different data types at the training data enhancers. By contrast, Supplementary Figure S1, Additional File 1, plots the training data promoters. The promoters were not oriented according to the transcription direction.

In addition to promoters, random genomic regions were included in the non-enhancer training set. Random genomic regions not overlapping the ENCODE blacklist regions, had a distance more than 2.5 kb to any p300 peak, and had distance more than 2 kb to TSS of protein coding genes were sampled in both cell lines. These were denoted as pure random regions, and the data and the average profiles for 1000 of such regions in cell line K562 are shown in Supplementary Figure S2, Additional File 1. At the pure random regions, the signals were close to zero, and it is likely an easy task for a classifier to separate enhancers from these. Therefore, another set of random locations was defined as follows: First, we computed the sum of signals excluding MNase-seq signal along the whole genome in 100 bp bins. We selected the bins having the sum equal or larger than 5. Again ENCODE blacklists, the p300 binding sites and TSS were removed from the selected regions. These regions

comprised around 4% of the whole genome, and the random regions were sampled within these regions so that the sum in all 100 bp bins in the 2 kb window centered at the sampled random location was equal or larger than 5. These were denoted as the random regions with signal. Supplementary Figure S3, Additional File 1, visualizes the 1000 random regions with signal in cell line K562.

Probabilistic modeling

A Bayesian model was constructed, and the training data was utilized to estimate the hyperparameters of the prior distribution. The trained model was applied to quantify the resemblance of the individual sample coverage profile to the training data enhancer average coverage profile. The objective was not to develop a full generative model that accounts for all uncertainties, e.g. in the model parameters, but instead to build a simpler approximate model.

The different types of ChIP-seq data sets were indexed by k, $k = 1, \ldots, K$ where K was the number of data types, in this work $K = 15$. The training data for histone modification k was represented as a matrix $\mathbf{Y}_k$ of size $n \times d$ where n was the number of samples and d was the number of bins, i.e. the length of the coverage profile. This matrix was further divided into the training data enhancers $\mathbf{Y}_k^{\text{enh}}$ of size $n_{\text{enh}} \times d$, and the training data non-enhancers $\mathbf{Y}_k^{\text{neg}}$ of size $n_{\text{neg}} \times d$. The coverage profile for histone modification k for sample i was $\mathbf{y}_{ik}$. The coverage of the ChIP-seq data for histone modification k of the ith enhancer profile in bin j was assumed to follow a Poisson distribution. Further assuming the conditional independence of the coverage values in adjacent bins, the likelihood of $\mathbf{y}_{ik}$ was

$$p(\mathbf{y}_{ik} \mid \alpha_k, \mathbf{x}_k) = \prod_{j=1}^{d} \text{Poisson}(y_{ijk} \mid \lambda_{jk} = \alpha_{ik} x_{jk}) \tag{1}$$

where λ_{jk} was the rate parameter of the coverage. The parameter λ_{jk} was a dummy or an auxiliary variable and was composed of two parts, the overall mean of the coverage $\mathbf{x}_k$ and the scaling parameter α_{ik}. The mean $\mathbf{x}_k$ captured the pattern of the histone modification signal at enhancers, and the scaling parameter α_{ik} was shared among all bins along the coverage profile, and it modeled the variation in the coverage that resulted from mapping biases and local chromatin properties, for example. The variations in the coverage were assumed to originate from a local source. In toher words, the variation was shared by bins along the coverage profile. The Gamma distribution with hyperparameters a_{0k} and b_{0k} is a natural choice to model the scaling parameter α_{ik} as it is the conjugate distribution of the Poisson distribution

$$\alpha_k \sim \text{Gamma}(a_{0k}, b_{0k}). \tag{2}$$

The variation due to random sampling of DNA segments during sequencing was captured by the Poisson distribution and the variation in the coverage among enhancers was captured by the Gamma distribution. The parameters of the model were estimated from the data as follows: First, variable $\mathbf{x}_k$ was estimated as the training data enhancer average coverage

$$x_{jk} = \frac{\sum_{i=1}^{n_{\text{enh}}} y_{ijk}^{\text{enh}}}{n_{\text{enh}}}. \tag{3}$$

Second, the distribution of scaling parameter α_{ik} was estimated by learning them individually for each training data enhancer sample and fitting a Gamma distribution for the obtained values. The individual α_{ik} were estimated for each training enhancer sample by maximizing the likelihood in Equation 1 to obtain

$$\hat{\alpha}_{ik} = \frac{\sum_{j=1}^{d} y_{ijk}^{\text{enh}}}{\sum_{j=1}^{d} x_{jk}} \quad \text{for } i = 1, \ldots, n^{\text{enh}}. \tag{4}$$

By fitting Gamma distribution to the estimated $\hat{\alpha}_{ik}$, the estimates for the hyperparameters a_{0k} and b_{0k} were obtained.

To obtain a value that described the fit between an individual sample $\mathbf{y}_{ik}$ and the enhancer coverage profile, a probabilistic score for each training data sample was computed by assuming that the mean $\mathbf{x}_k$ and the distribution of α_k were estimated as described above for all $k = 1, \ldots, K$.

$$p(\mathbf{y}_{ik}) = \int \text{Gamma}(\alpha_k \mid a_{0k}, b_{0k}) \prod_{j=1}^{d} \text{Poisson}(y_{ijk} \mid \alpha_k x_{jk}) \text{d}\alpha_k \tag{5}$$

$$= \frac{\Gamma(a_{0k} + \sum_{j=1}^{d} y_{ijk}) b_{0k}^{a_{0k}} \prod_{j=1}^{d} x_{jk}^{y_{ijk}}}{\Gamma(a_{0k})(b_{0k} + \sum_{i=1}^{d} x_{jk})^{a_{0k} + \sum_{j=1}^{d} y_{ijk}} \prod_{j=1}^{d} y_{ijk}!}.$$

The individual training data enhancers were considered independent. Therefore, the probabilistic scores of individual enhancers provided in Equation 5 were utilized without the product over all samples. Equation 5 resembled the Gamma-Poisson mixture distribution being equivalent to the negative binomial distribution. However, the Equation 5 could not be simplified into the negative binomial distribution due to the product over d adjacent bins and the product $\lambda_{jk} = \alpha_k x_{jk}$; this removed the conjugacy between Gamma and Poisson distributions. The probabilistic score for each training data sample and for each data type were computed to get a matrix of size $n \times K$, the number of features being K for each sample.

In addition to modeling the individual sample's fit to the enhancer average coverage profile, we modeled the fit of the sample to the non-enhancer coverage profile. In other words, when predicting enhancers genome-wide, the desired outcome would be regions that resembled enhancer average coverage profiles but which did not resemble random regions or promoters. The mean coverage values for non-enhancer regions were computed as

$$x_{jk}^{\text{neg}} = \frac{\sum_{i=1}^{n_{\text{neg}}} y_{ijk}^{\text{neg}}}{n_{\text{neg}}} \tag{6}$$

and the ML estimates of α_{ik}^{neg} using data for training data enhancers were computed as

$$\hat{\alpha}_{ik}^{\text{neg}} = \frac{\sum_{j=1}^{d} y_{ijk}^{\text{enh}}}{\sum_{j=1}^{d} x_{jk}^{\text{neg}}} \quad \text{for } i = 1, ..., n_{\text{enh}}. \tag{7}$$

A Gamma distribution was fitted to the α_{ik}^{neg} values to obtain the estimates for hyperparameters a_{0k}^{neg} and b_{0k}^{neg}. Then the probabilistic scores of samples were computed using Equation 5, but instead of integrating over α_k, the integral was over α_k^{neg}. The estimation for α_k^{neg} was done separately for the random regions and the promoters. The probabilistic score values were appended to the previous vector of length K to obtain feature vectors of length $3K$.

For comparison, we considered a model that estimated the $\hat{\alpha}_{ik}$ and $\hat{\alpha}_{ik}^{\text{neg}}$ values for each individual sample in the training and test data. The $\hat{\alpha}_{ik}$ and $\hat{\alpha}_{ik}^{\text{neg}}$ were estimated using Equations 4 and 7, respectively. The likelihood values for an individual sample were computed by Equations 8 and 9.

$$p(\mathbf{y}_{ik} | \mathbf{x}_k, \hat{\alpha}_{ik}) = \prod_{j=1}^{d} \text{Poisson}(y_{ijk} | \hat{\alpha}_{ik} x_{jk}) \tag{8}$$

$$p(\mathbf{y}_{ik} | \mathbf{x}_k^{\text{neg}}, \hat{\alpha}_{ik}^{\text{neg}}) = \prod_{j=1}^{d} \text{Poisson}(y_{ijk} | \hat{\alpha}_{ik}^{\text{neg}} x_{jk}^{\text{neg}}). \tag{9}$$

In this model, for each genomic region, we estiamted a fixed scaling parameter that best matched the sample coverage profile to the enhancer or negative average profile. This approach was denoted as the maximum likelihood (ML) approach. In the Bayesian approach, in contrast, the scaling parameter had a probability distribution, and in Equation 5 the likelihood was integrated over α_k.

Classifier training and cross-validation, performance on the test set

A support vector (SVM) classifier implemented in libSVM version 3.22 [98] was trained. The SVM classifier utilized a Gaussian kernel. Using the training data from cell line K562, a nested cross-validation was performed. In the nested CV, the outer cross-validation assessed the performance of the model and the inner optimized the hyperparameters, namely the SVM misclassification penalty C and the Gaussian kernel width

γ. The hyperparameter optimization was performed by a grid-search with values $C = 2^{-5}, 2^{-4.5}, ..., 2^{25}$ and $\gamma = 2^{-25}, 2^{-24.5}, ..., 2^{10}$. For both the outer and inner cross-validation, 5-fold CV was adopted. The method performance for cell line K562 was evaluated by concatenating the predictions obtained from separate cross-validation rounds and computing the area under receiver operation characteristics curve (AUC). For RFECS, the CV within K562 training data was not performed; the AUC values would have likely been close to 1. The final model was trained using all training data from K562, and again a 5-fold CV was performed to optimize the SVM parameters. The final model was used to predict enhancers on the GM12878 test data. To obtain the performance measure for RFECS on the small GM12878 test set, the genome-wide RFECS predictions closest to the test set regions were selected. The performance on test set was evaluated again by AUC.

Genome-wide enhancer predictions and their validation

After predicting enhancers genome-wide by PREPRINT and RFECS, any obscure genomic regions were removed from the predictions: Predictions (defined as 2 kb genomic windows) that overlapped at least 1 bp with the ENCODE blacklist regions were removed from the predictions. In addition, we removed the K562 cell line predictions that had a distance equal or smaller than 1 kb to any training data enhancer. To remove promoters from the set of enhancer prediction, we excluded the predictions whose middle base had a distance equal or smaller than 2 kb to any GENCODE v27 transcription start site (TSS).

List of abbreviations

ATAC-seq: Assay for Transposase-Accessible Chromatin using sequencing, AUC: Area Under the receiver operating characteristics Curve, bp: Base pair, CAGE-seq: Cap Analysis of Gene Expression followed by sequencing, CBP: CREB-binding-protein, ChIP-seq: Chromatin Immunoprecipitation coupled with Sequencing, CRE-seq: Cis-regulatory element analysis by sequencing reporter assay, CTCF: CCCTC-binding factor, CV: Cross-validation, DNase I HS: DNase I hypersensitivity, DNase-seq: DNase I hypersensitive sites sequencing, EM: Expectation-Maximization, ENCODE: Encyclopedia of DNA Elements, eRNA: Enhancer RNA, ESC: Embryonic stem cells, FAIRE-seq: Formaldehyde-Assisted Isolation of Regulatory Elements and sequencing, FANTOM: Functional Annotation of the Mammalian Genome, FPR: False positive rate, GENCODE: The reference genome annotation for human and other species, GM12878: Lymphoblastoid cell line, GRO-seq: Global run-on and sequencing, GWAS: Genome-wide association study, H2AZ: Histone variant, H3K27ac: Histone 3 lysine 27 acetylation, H3K27me3: Histone 3 lysine 27 trimethylation, H3K36me3: Histone 3 lysine 36 trimethylation, H3K4me3: Histone 3 lysine 4 trimethylation, H3K79me2: Histone 3 lysine 79 dimethylation, H3K9ac: Histone 3 lysine 9 acetylation, H4K20me1: Histone 4 lysine 20 monomethylation, K562: Myelogenous leukemia cell line, kb: Kilobase, ML: Maximum likelihood, MNase-seq: Micrococcal nuclease digestion followed by sequencing, MPRA: Massive parallel reporter assay, NRSF: Neuron-Restrictive Silencer Factor, p300: Histone acetyltransferase, PCR: Polymerase Chain Reaction, PREPRINT: Probabilistic enhancer prediction tool, RFECS: Random Forest based Enhancer Identification from Chromatin States, RNA Pol II: RNA polymerase II, ROC: Receiver operating characteristics, SNP: Single nucleotide polymorphism, SVM: Support vector machine, TF: Transcription factor, TSS: Transcription start site

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The download links to the original data analysed in this study are included in the supplementary information files of this published article. The data used in this work and the links to download the data are listed in the Additional File 2. The PREPRINT package and the codes for the data preprocessing steps are available in GitHub `https://github.com/MariaOsmala/preprint`. The data and enhancer predictions are stored as a UCSC Genome Browser track hubs, links to the track hubs are provided in GitHub.

Funding

This work has been supported by the the Academy of Finland [grant numbers 292660 and 314445], and the Finnish Cultural Foundation.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

MO and HL developed the machine learning methods. MO compiled the data sets, implemented the method, performed the computational analysis and wrote the manuscript. HL participated in the evaluations of the findings and revision of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank our colleagues in the Lähdesmäki laboratory for discussions and feedback regarding this manuscript. We thank Dr. Gökçen Eraslan and Dr. Nisha Rajagopal providing support to execute the RFECS experiments. The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project.

References

[1] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

[2] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Research. 2012;22(9):1760–1774.

[3] Forrest ARR, Kawaji H, Rehli M, Baillie JK, De Hoon MJL, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462–470.

[4] Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: From properties to genome-wide predictions. Nature Reviews Genetics. 2014;15(4):272–286.

[5] Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell. 2016;167(5):1170–1187.

[6] Rickels R, Shilatifard A. Enhancer Logic and Mechanics in Development and Disease. Trends in Cell Biology. 2018;28(8):608–630.

[7] Banerji J, Rusconi S, Schaffner W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981;27(2):299–308.

[8] Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. Nucleic Acids Research. 1981;9(22):6047–6068.

[9] Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144(3):327–339.

[10] Karnuta JM, Scacheri PC. Enhancers: bridging the gap between gene control and human disease. Human molecular genetics. 2018;27(R2):R219–R227.

[11] Corradin O, Scacheri PC. Enhancer variants: Evaluating functions in common disease. Genome Medicine. 2014;6(10):85.

[12] Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nature Structural and Molecular Biology. 2014;21(3):210–219.

[13] Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. In: Morse RH, editor. Methods in Molecular Biology. vol. 833. Humana Press; 2012. p. 413–419.

[14] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell. 2008;132(5):887–898.

[15] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497–1502.

[16] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods. 2007;4(8):651–657.

[17] Thurman RE, Rynes E. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75.

[18] McKay DJ. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to identify functional regulatory DNA in insect genomes. Methods in Molecular Biology. 2019;1858(2):89–97.

[19] Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799–816.

[20] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics. 2007;39(3):311.

[21] Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459(7243):108–112.

[22] Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457(7231):854–858.

[23] Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 2011;470(7333):279–285.

[24] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(50):21931–21936.

[25] Spitz F, Furlong EEM. Transcription factors: From enhancer binding to developmental control. Nature Reviews Genetics. 2012;13(9):613–626.

[26] Zabidi MA, Stark A. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. Trends in Genetics. 2016;32(12):801–814.

[27] Elnitski L, Jin VX, Farnham PJ, Jones SJM. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. Genome Research. 2006;16(12):1455–1464.

[28] Su J, Teichmann SA, Down TA. Assessing computational methods of cis-regulatory module prediction. PLoS Computational Biology. 2010;6(12):e1001020.

[29] Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nature Reviews Genetics. 2012;13(7):469–483.

[30] Sheffield NC, Furey TS. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. Genes. 2012;3(4):651–670.

[31] Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. Briefings in Bioinformatics. 2016;17(6):967–979.

[32] Lim LWK, Chung HH, Chong YL, Lee NK. A survey of recently emerged genome-wide computational enhancer predictor tools. Computational Biology and Chemistry. 2018;74:132–141.

[33] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature Biotech. 2010;28:817–825.

[34] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–49.

[35] Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. Genome Research. 2014;24(10):1595–1602.

[36] Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. PLoS Computational Biology. 2013;9(3):e1002968.

[37] Won KJ, Chepelev I, Ren B, Wang W. Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC Bioinformatics. 2008;9:547.

[38] Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics. 2010;26(13):1579–1586.

[39] Fernández M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Research. 2012;40(10):e77.

[40] Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455–461.

[41] Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics. 2018;19(1).

[42] Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465(7295):182–187.

[43] de Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. PLoS Biology. 2010;8(5):e1000384.

[44] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(26):15776–15781.

[45] Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322(5909):1845–1848.

[46] Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017 Jan 1;2017.

[47] Ho EYK, Cao Q, Gu M, Chan RWL, Wu Q, Gerstein M, et al. Shaping the nebulous enhancer in the era of high-throughput assays and genome editing. Briefings in Bioinformatics. 2019 Mar 20;2019, bbz030.

[48] Hashimoto TB, Edwards MD, Gifford DK. Universal Count Correction for High-Throughput Sequencing. PLoS Computational Biology. 2014;10(3):e1003494.

[49] Strackee J, van der Gon JJD. The frequency distribution of the difference between two Poisson variates. Statistica Neerlandica. 1962;16(1):17–23.

[50] Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics. 2011;27(6):870–871.

[51] Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, et al. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. Genome Research. 2012;22(9):1735–1747.

[52] Nielsen FGG, Markus KG, Friborg RM, Favrholdt LM, Stunnenberg HG, Huynen M. CATCHprofiles: Clustering and alignment tool for chip profiles. PLoS ONE. 2012;7(1):e28272.

[53] Nair NU, Kumar S, Moret BME, Bucher P. Probabilistic partitioning methods to find significant patterns in ChIP-Seq data. Bioinformatics. 2014;30(17):2406–2413.

[54] Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. Epigenetics and Chromatin. 2015;8(1):16.

[55] Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nature Reviews Molecular Cell Biology. 2018;19(10):621–637.

[56] Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, et al. Widespread transcriptional pausing and elongation control at enhancers. Genes and Development. 2018;32(1):26–41.

[57] Kim TK, Shiekhattar R. Architectural and Functional Commonalities between Enhancers and Promoters. Cell. 2015;162(5):948–959.

[58] Dorighi KM, Swigut T, Henriques T, Bhanu NV, Scruggs BS, Nady N, et al. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Molecular Cell. 2017;66(4):568–576.

[59] Rickels R, Herz HM, Sze CC, Cao K, Morgan MA, Collings CK, et al. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. Nature Genetics. 2017;49(11):1647–1653.

[60] Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Research. 2013;41(2):827–841.

[61] Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(47):19498–19503.

[62] Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339(6123):1074–1077.

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

[64] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172(4):650–665.

[65] Calo E, Wysocka J. Modification of Enhancer Chromatin: What, How, and Why? Molecular Cell. 2013;49(5):825–837.

[66] Fleischer T, Tekpli X, Mathelier A, Wang S, Nebdal D, Dhakal HP, et al. DNA methylation at enhancers identifies distinct breast cancer lineages. Nature Communications. 2017;8(1):1379.

[67] Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Research. 2013;23(5):800–811.

[68] Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nature Communications. 2019;10(1):568915.

[69] Carter B, Ku WL, Kang JY, Hu G, Perrie J, Tang Q, et al. Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). Nature Communications. 2019;10(1):571208.

[70] Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. XSuper-enhancers in the control of cell identity and disease. Cell. 2013;155(4):934.

[71] Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153(2):307–319.

[72] Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature Structural and Molecular Biology. 2011;18(8):956–963.

[73] Kioussis D, Vanin E, Delange T, Flavell RA, Grosveld FG. β-Globin gene inactivation by DNA translocation in $\gamma\beta$-thalassaemia. Nature. 1983;306(5944):662–666.

[74] van der Ploeg LHT, Flavell RA. DNA methylation in the human $\gamma\delta\beta$-globin locus in erythroid and non-erythroid tissues. Cell. 1980;19(4):947–958.

[75] Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(44):17921–17926.

[76] Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153(2):320–334.

[77] Cohen AJ, Saiakhova A, Corradin O, Luppino JM, Lovrenert K, Bartels CF, et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. Nature Communications. 2017;8:14400.

[78] Corradin O, Cohen AJ, Luppino JM, Bayles IM, Schumacher FR, Scacheri PC. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. Nature Genetics. 2016;48(11):1313–1320.

[79] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–380.

[80] Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–293.

[81] Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458–472.

[82] Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–1680.

[83] Beagrie RA, Scialdone A, Schueler M, Kraemer DCA, Chotalia M, Xie SQ, et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 2017;543(7646):519–524.

[84] Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nature Communications. 2018;9(1):943.

[85] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research. 2008;18(9):1509–1517.

[86] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;11(10):R106–R106.

[87] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11(3):R25.

[88] Zacher B, Michel M, Schwalb B, Cramer P, Tresch A, Gagneur J. Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. PLoS ONE. 2017;12(1):e0169249.

[89] Spyrou C, Stark R, Lynch A, Tavare S. BayesPeak: Bayesian analysis of ChIP-seq data. BMC Bioinformatics. 2009;10(1):299.

[90] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357–359.

[91] Marx V. How to deduplicate PCR. Nature Methods. 2017;14(5):473–476.

[92] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research. 2012;22(9):1813–1831.

[93] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology. 2008;26(12):1351–1359.

[94] Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Annals of Applied Statistics. 2011;5(3):1752–1779.

[95] Le Martelot G, Canella D, Symul L, Migliavacca E, Gilardi F, Liechti R, et al. Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles. PLoS Biology. 2012;10(11):e1001442.

[96] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2019;47(D1):D766–D773.

[97] Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in Genetics. 2014;5:75.

[98] Chang CC, Lin CJ. LIBSVM: A Library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2(3):1–27.

Additional Files

Additional file 1 — Supplementary Figures

This file contains supplemental **Figures S1–S17**. PDF of size 31000 kB.

Additional file 2 — The origin of data

The links to the downloaded files, their ENCODE Data Coordination Center (DCC) Accession names, Data Coordination Centers (ENCODE DCC), and Gene Expression Omnibus (GEO) sample accession numbers. Excel (.xlsx) file of size 34.4 kB.

Additional file 3

The full genome browser example figure of cell line K562 data. PDF of size 186.1 kB.

Additional file 4

The full genome browser example figure of cell line GM12878 data. PDF of size 199.3 kB.