Analysis and prediction of super-enhancers using sequence and chromatin signatures

Aziz Khan[1,2] and Xuegong Zhang[1,3,]*

[1] MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/Department of Automation, Tsinghua University, Beijing, 100084, China
[2] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0349 Oslo, Norway
[3] School of Life Sciences, Tsinghua University, Beijing, 100084, China

* Corresponding Author

Author's email address: aziz.khan@ncmm.uio.no; zhangxg@tsinghua.edu.cn

Abstract

Background: Super-enhancers are clusters of active enhancers densely occupied by the Mediators, transcription factors and chromatin regulators, control expression of cell identity and disease associated genes. Current studies demonstrated the possibility of multiple factors with important roles in super-enhancer formation; however, a systematic analysis to asses the relative contribution of chromatin and sequence features of super-enhancers and their constituents remain unclear. In addition, a predictive model that integrates various types of data to predict super-enhancers has not been established.

Results: Here, we integrated diverse types of genomic and epigenomic datasets to identify key signatures of super-enhancers and their constituents and to investigate their relative contribution. Through computational modelling, we found that Cdk8, Cdk9 and Smad3 as new key features of super-enhancers along with many known. Comprehensive analysis of these features in embryonic stem cells and pro-B cells revealed their role in the super-enhancer formation and cellular identity. Further, we observed significant correlation and combinatorial predictive ability among many cofactors at the constituents of super-enhancers. By utilizing these features, we developed computational models which can accurately predict super-enhancers and their constituents. We validated these models using cross-validation and also independent datasets in four human cell-types.

Conclusions: Our analysis of these features and prediction models can serve as a resource to further characterize and understand the formation of super-enhancers. Taken together, our results also suggest a possible cooperative and synergistic interactions of numerous factors at super-enhancers and their constituents. We have made available our analysis pipeline as an open-source tool with a command line interface at https://github.com/asntech/improse.

Keywords*:* Gene regulation, epigenomics, enhancer, super-enhancer, prediction, embryonic stem cells

Background

Enhancers are *cis*-regulatory regions in the DNA that not only augment the transcription of associated genes but also play a key role in cell-type-specific gene expression [1, 2]. A myriad of transcription factors (TFs) bind to enhancers and regulate gene expression by recruiting coactivators and RNA polymerase II (RNA Pol II) to target genes [3–7]. A typical mammalian cell is estimated to have thousands of active enhancers, a number which rises to roughly one million in the human genome [1, 8]. It has been more than three decades since the first enhancer was discovered [9], but our understanding of the mechanisms by which enhancers regulate gene expression is still limited. However, development of methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) and DNase I hypersensitivity followed by sequencing (DNase-seq), have helped to discover and characterize enhancers at genome scale. Many factors have been associated with enhancer activity, including mono methylation of histone H3 at lysine 4 (H3K4me1), acetylation of histone H3 at lysine 27 (H3K27ac), binding of the coactivator proteins p300 and CBP, and DNase I hypersensitivity [8, 10, 11]. By exploiting these factors and other genomic features, many computational approaches have been developed to predict enhancers genome-wide [4, 12].

Mediator, a transcriptional coactivator, forms a complex with Cohesin to create cell-type-specific DNA loops and by facilitating enhancer-bound transcription factors, to recruit RNA Pol II to the promoters of target genes [13, 14]. In embryonic stem cells (ESC), the pluripotency transcription factors Oct4, Sox2 and Nanog (OSN) are known to have 100% enhancer activity (25/25) [15]. By using ChIP-seq data for Oct4, Sox2 and Nanog in ESC, 10,227 co-bound regions have been identified and classified into super-enhancers (SEs) and typical enhancers (TEs) by using ChIP-seq signal for Mediator subunit Med1 [16]. Super-enhancers form clusters of active enhancers, are cell-type specific, associated with key cell identity genes, and linked to many biological processes which define the cell identity [16]. These super-enhancers are densely loaded with the Mediator complex, master transcription factors and chromatin regulators [16–19]. Many disease- and trait-associated single nucleotide polymorphisms (SNPs) have been found in these regions [18]. Super-enhancers differ from typical enhancers in terms of size, ChIP-seq density of various cofactors, DNA motif content, DNA methylation level, enhancer RNA (eRNA) abundance, ability to activate transcription and sensitivity to perturbation [16–18, 20–22]. Further, studies have found super-enhancers in multiple cancers and demonstrated their importance in cellular-identity and disease and emphasized their use as potential biomarkers [17, 18, 23–25]. Other parallel studies demonstrated nearly similar patterns by using different approaches and termed them ‘stretch enhancers’ [26] and ‘enhancer clusters’ [27].

Since the discovery of super-enhancers, the research community used ChIP-seq data for different factors to differentiate super-enhancers from typical enhancers in different cell-types. ChIP-seq data for Med1 optimally differentiated super-enhancers and typical enhancers by comparing it with enhancer marks, including H3K27ac, H3K4me1 and DNase I hypersensitivity [16]. BRD4, a member of the BET protein family, was also used to distinguish super-enhancers from typical enhancers as it is highly correlated with MED1 [17]. H3K27ac was extensively used to create a catalogue of super-enhancers across 86 different human cell-types and tissues due to its availability [18]. Other studies used the coactivator protein P300 to define super-enhancers [28, 29]. However, the knowledge about these factors' ability to define a set of super-enhancers in a particular cell-type and their relative and combinatorial importance remains limited. Master transcription factors which might form the super-enhancer domains are largely unknown for most of the cell-types, while performing ChIP-seq for the Mediator complex is difficult and costly. Current studies demonstrated the possibility of multiple cofactors with important roles in super-enhancer formation; however, a predictive model that integrates various types of data to predict super-enhancers and their constituents (enhancers within a super-enhancer) has not been established. In addition, the degree to which the sequence-specific features of constituents by itself explains the differences between super-enhancers and typical enhancers remains unknown.

Herein, to identify key features of super-enhancers and to investigate their relative contribution to predict super-enhancers, we integrated diverse types of publicly available datasets, including ChIP-seq data for histone modifications, chromatin regulators and transcription factors, DNase I hypersensitive sites and genomic data. Using correlation analysis and computational modelling, we found that Med1, Med12, H3K27ac, Brd4, Cdk8, Cdk9, p300 and Smad3 were significantly correlated and had a higher predictive importance. By utilizing these features, we developed imPROSE (integrated methods for prediction of super-enhancers) to predict super-enhancers and their constituents from a list of enhancers. We implemented and compared six different state-of-the art learning models and validated them using 10-fold stratified cross validation as well as by independent datasets in four human cell-types. imPROSE trained on Smad3 and H3K27ac data in mESC predicts more cell-type-specific super-enhancers in pro-B cells as compared with H3K27ac-based method (ROSE). We also performed a genome-wide analysis of Cdk8, Cdk9 and Smad3 binding in mESC and pro-B cells to analyse and assess their relative importance in defining super-enhancers. By using ChIP-seq data, RNA-seq based gene expression data, Gene Ontology (GO) and motif analyses we found that these factors differentiate super-enhancers from typical enhancers. Our prediction model and derived features can be further used as a platform to precisely define super-enhancers in other cell-types.

Results

Analysis of features including chromatin and transcription factors

Studies have shown that super-enhancers are occupied by various cofactors, chromatin regulators, histone modifications, RNA polymerase II and transcriptions factors [16, 18]. An understanding of the occupancy of these factors at the constituents of super-enhancers and typical enhancers is lacking. We extensively analysed 32 publicly available ChIP-seq and DNase-seq datasets to unveil their association with the constituents of super-enhancers and typical enhancers in mouse embryonic stem cells (mESC). We found that most of these factors, which are enriched in super-enhancers, were also highly enriched in the constituents of super-enhancers relative to typical enhancers (Fig. 1a; Figure S1 in Additional file 1). It is understandable to see that Oct4, Sox2 and Nanog were nearly equally enriched across the constituents of super-enhancers and typical enhancers because these constituents are defined by intersecting OSN co-bound regions.

Through correlation analysis, we found that most of the factors were highly correlated at the constituents of super-enhancers compared to typical enhancers (Fig. 1b). This suggests a possible combinatorial interplay among these cofactors at super-enhancers and that could be the reason that super-enhancers are more active and sensitive to perturbation as compared to typical enhancers. Interestingly, features with similar lineage/functionality were clustered together. For example, histone modifications (H3K27ac and H3K4me3), Mediator complex subunits (Med1, Med12 and Cdk8), the pluripotency genes of ESC (Oct4, Sox2, Nanog) were clustered together. It was particularly interesting to observe that Smad3 clustered together with the co-activator protein p300/CBP. Previous studies have shown that p300/CBP interacts with Smad3 [30, 31].

Analysis of sequence specific features

Super-enhancers differ from typical enhancers in terms of size, ChIP-seq density of various cofactors, TF content, ability to activate transcription, and sensitivity to perturbation [16–18]. But, to what extent constituents of super-enhancers differ from the constituents of typical enhancers in terms of sequence composition remains unknown. To gain insights into their biological functions, we sought to identify DNA sequence signatures of constituent enhancers. We tested GC-content, repeat fraction, size, and phastCons across the constituents of super-enhancers and typical enhancers in mouse ESC and pro-B cells.

Previous studies have shown that GC-rich regions have distinct features including frequent TF binding [32], active conformation [33] and nucleosome formation [34]. We found that constituents of super-enhancers are significantly more GC-rich than the constituents of typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 1c). This suggests that GC content has an important role in super-enhancers formation and it can be a defining feature to distinguish them from typical enhancers.

Enhancers, larger than 3 kb have been shown to be cell-type-specific and are known as stretch enhancers [26]. We checked the size (bp) of constituents and found that constituents of super-enhancers are significantly larger than the constituents of typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 1d). A previous study showed that majority of super-enhancers do overlap with stretch enhancers [35]. Taken together, these results suggests that super-enhancers are actually clusters of stretch enhancers.

Enhancers are hardly conserved across mammalian genomes and evolved recently from ancestral DNA exaptation, rather than lineage-specific expansions of repeat elements [36]. We did not find any significant difference in conservation at constituents of super-enhancers and typical enhancers in mESC (p-value = 0.6285, Wilcoxon rank sum test), but in pro-B cells the conservation score was statistically significant (p-value < 1.7e-4, Wilcoxon rank sum test) (Fig. 1e). Similarly, there was no significant difference in repeat fraction at constituents of both super-enhancers and typical enhancers in mESC (p-value = 0.0202, Wilcoxon rank sum test) and pro-B cells (p-value = 0.8976, Wilcoxon rank sum test) (Fig. 1f).

Feature-ranking revealed previously known and new features of super-enhancers

With the increasing discovery of factors associated with super-enhancers, the determination of their relative importance in defining super-enhancers is important. Hence, we ranked chromatin and transcription factors to find a minimal optimal subset, which can be used to optimally distinguish super-enhancers from typical enhancers. We used a random-forest based approach, Boruta [37] to assess the importance of each feature by ranking them based on their predictive importance (Fig. 2a) (Methods). We also used an out-of-bag approach to calculate the relative importance of each feature and achieved almost identical results (Figure S5 in Additional file 1).

After ranking chromatin features, we found Brd4, H3K27ac, Cdk8, Cdk9, Med12 and p300 as the six most important factors with Brd4 and H3K27ac as the top two most informative factors (Fig. 2a). It was particularly interesting to observe that Cdk8 and Cdk9 were ranked as the third and fourth most informative features, respectively. Cdk9, a subunit of the positive transcription elongation factor b (P-TEFb) has been found in enhancers and promoters of active genes along with the Mediator coactivator [17]. Cdk8, a subunit of Mediator complex, positively regulates precise steps in the assembly of transcriptional elongation, including the recruitment of P-TEFb and Brd4 [38].

Previous studies have shown that five ESC transcription factors (Sox2, Oct4, Nanog, Esrrb and Klf4) and other TFs (Smad3, Stat3, Tcf3, Nr5a2, Prdm14 and Tcfcp2l1) were enriched in super-enhancers [16, 18]. As these transcription factors are specific for ESC biology, we ranked them separately from other cofactors to find their relative importance in ESC. Surprisingly, Smad3 turn to be the most informative among other transcription factors including Klf4 and Esrrb which were previously described as key defining features of super-enhancers [16] (Fig. 2b). It will be interesting to further understand the importance of Cdk8, Cdk9 and Smad3 in the formation of super-enhancers.

We also ranked the chromatin and transcription factors together, not surprisingly Med1 was ranked as most informative feature followed by H3K27ac, Brd4, Cdk8, Cdk9. We found that Smad3 was ranked higher than p300 and also ESC specific master TFs (Fig. 2c). To test this in more differentiated cell -types, we have ranked the several factors, including H3K27ac, H3K4me1, H3K4me3, p300, Smad3, PU.1, Foxo1, Ebf1, Pol2 and DNaseI in pro-B cell. Interestingly, we found that Smad3 turn to be the most informative feature followed by PU.1, p300 and H3K27ac (Fig. 2d). This shows that Smad3 might be more informative feature to distinguish super-enhancers in differentiated cells.

Comparison of six different state-of-the-art machine learning models

We compared six different state-of-the-art, supervised, machine learning models, including Random Forest, linear SVM, k-NN, AdaBoost, Naive Bayes and Decision Tree. We used chromatin, transcription factors and sequence-specific features to train these models individually and evaluated their performance using 10-fold cross validation. The parameters used for each model can be found in the methods section.

Using all chromatin, transcription factors and sequence-specific features, Random Forest performed well with AUC=0.98, while Naive Bayes performed poorly with AUC=0.85 (Fig. 3b, Figure S2D in Additional file 1). Linear SVM and k-NN performed equally with AUC=0.95, while Decision

Tree achieved AUC=0.91. We found that AdaBoost performed almost equally with AUC=0.96 as compared to Random Forest with AUC=0.98, but we achieved stable precision and recall for Random Forest.

We further compared these models, using individual features as well as different types of features and found feature-type-specifics for each model (Figure S3; Table S1, S2 in Additional file 1). For example, linear SVM performed poorly on DNA sequence-specific features with AUC=0.69, while Random Forest and AdaBoost performed best with AUC=0.81.

Random Forest performed optimally across different combinations of features and data sampling approaches. Hence, we chose Random Forest for further analysis due to its performance and flexibility, though any of these models can be used to predict super-enhancers to some extend depending on the type of features used. This comparative analysis of models using various types of features, provides a guide to select a model based on the type of features available.

Prediction using chromatin and transcription factors

Random Forests are ensemble and non-parametric models, which run efficiently on large datasets without over-fitting. Here, we developed imPROSE, a Random Forest based model, to predict super-enhancers and their constituents. A detailed workflow of imPROSE is illustrated in (Fig. 3a). We investigated six state-of-the-art machine learning models, including Random Forest, linear SVM, k-NN, AdaBoost, Naive Bayes and Decision Tree. In order to train the models, we used normalized ChIP-seq profiles for all chromatin, cofactors and transcription factors, and other genomic features at the constituents of super-enhancers and typical enhancers defined by Med1 signal at OSN (Oct4, Sox2 and Nanog) co-bound sites. To avoid over-fitting, we used a hybrid-sampling approach to balance the training and test data (Additional file 1). The prediction accuracy was assessed using 10-fold stratified cross-validation (Methods).

We investigated the individual predictive power of features using all the six models (Table S2 in Additional file 1). The AUC (Area Under the Curve) scores reported in this manuscript are based on the Random Forest model, unless stated otherwise. The features which were ranked as more important predictors in (Fig. 2a, b), achieved higher AUC and PRC (Precision Recall Curve) scores. Among the top ranked chromatin features, including H3K27ac, Brd4, Cdk8, Cdk9, Med12 and p300, we observed Brd4 performed slightly better than H3K27ac, with AUC=0.85 and 0.84, respectively. The model achieved AUC=0.84 for Cdk8, 0.83 for Cdk9, 0.83 for Med12 and 0.76 for p300 (Fig. 3c). After combining the top three chromatin features (H3K27ac, Brd4 and Cdk8) the model performed well with AUC=0.93. This shows that the combination of Brd4, H3K27ac and Cdk8 is more effective at predicting super-enhancers than either alone. Among the top ranked transcription factors, including Smad3, Esrrb, Klf4, Tcfcp2l1, Nr5f2a and Stat3, we found that Smad3 achieved the highest predictive power with AUC=0.73, followed by Esrrb with AUC=0.69 and Klf4 with AUC=0.65 (Fig. 3d). The model achieved AUC=0.65, 0.65, 0.64 for Tcfcp2l1, Nr5f2a and Stat3, respectively. After combining the top three ranked transcription factors, including Smad3, Esrrb and Klf4, the model performed better with AUC=0.84 than either alone.

This shows that combinatorial information of cofactors is more effective at predicting super-enhancers than the individual features alone. We performed the combinatorial analysis based on feature types in the following section. We also noticed that the top ranked features, including Cdk8, Cdk9, p300, CBP, Smad3 and Brd4 are highly correlated with Med1, at the constituents of super-enhancers and typical enhancers (Fig. 1b).

Prediction using sequence-specific features

We used genomic features, including conservation score (phastCons), GC content and repeat fraction, and investigated their individual and combinatorial predictive power. The model achieved AUC=0.58 for phastCons, AUC=0.63 for GC content and AUC=0.64 for repeat fraction (Fig. 3e). By combining GC content and phastCons the model achieved AUC=0.71 and by combining GC content and repeat fraction it achieved AUC=0.76. By using three of the sequence-specific features, including GC content, phastCons and repeat fraction together, the model performed significantly higher with AUC=0.81. This shows that only genomic features could be enough to predict super-enhancers where high throughput sequencing data is not available.

Further, we used the DNA motifs for the 11 transcription factors, including Oct4, Sox2, Nanog, Esrrb, Klf4, Tcfcp2l1, Prdm14, Nr5a2, Smad3, Stat3 and Tcf3 to train the model. Using only the motifs information, the model achieved AUC=0.72. By using the ChIP-seq signal for these TFs, it achieved AUC=0.93 (Fig. 3g).

We also tested prediction accuracy using a sequence specific *k*-mer based approach[39]. We achieved AUC=0.74 for stitched sequences of super-enhancers and typical enhancers and AUC=0.75 for the constituents of super-enhancers and typical enhancers (Figures S2G, S2H in Additional file 1).

Combinatorial predictive ability of chromatin and genomic features

Previous studies have shown a combinatorial interplay between multiple histone modifications, transcriptions factors and DNA motifs, and this can be functionally informative [40, 41]. We noticed significant correlation among many chromatin regulators and TFs at the constituents of super-enhancers. This suggests the existence of a combinatorial relationship among these factors, which might dictate an accurate explanation for their role in super-enhancer formation. This combinatorial information could also be more predictive than the individual information of each factor.

We therefore investigated the combinatorial predictive power of chromatin, TFs and sequence-specific features. We tested 22 different combinations of these features and reported precision, recall, F1-score, AUC, and PRC (Fig. 3f). By training the model with the top six chromatin features, including Brd4, H3K27ac, Cdk8, Cdk9, Med12 and p300, it achieved AUC=0.95. By training the model with top two features (Brd4 and H3K27ac), it achieved AUC=0.91. By further adding the sequence specific features, the predictive ability of model greatly increased with AUC=0.95. By combining histone modifications, including H3K27ac, H3K4me1 and H3K4me3, with sequence specific features, the model achieved AUC=0.94. By training the model on known enhancer features data, including H3K4me1, H3K27ac, p300 and DNaseI, it achieved AUC=0.92. It is well known that Mediator forms a complex with Cohesin to create cell-type-specific DNA loops, and it facilitates enhancer-bound transcription factors to recruit RNA Pol II to the promoters of target genes [13, 14]. Hence, when we tested the combinatorial predictive power of Mediator sub-unit Med12, Cohesin sub-unit Smc1 and RNA Pol II, the model achieved AUC=0.91. Further, by adding the sequence specific features, the model performance considerably improved to AUC=0.95.

Furthermore, we tested the predictive ability of features that were grouped based on their type and functionality, and found that transcription factors, histone modifications and chromatin regulators have higher predictive power with AUC=0.93, 0.92 and 0.92, respectively. For models trained on Mediator complex, Cohesin, coactivators and Pol II data, achieved AUC=0.89, 0.79, 0.83 and 0.81, respectively (Fig. 3g). It was particularly interesting to see that a model trained on genomic features, including GC content, phastCons and repeat fraction, achieved AUC=0.81. We also examined the combinatorial importance of these grouped features based on their functionality and type (Figure S2E in Additional file 1). The increase in the model AUC is statistically significant (p-value = 0.001, Wilcoxon rank sum statistic) (Figure S6 in Additional file 1).

Our analysis shows that the combinatorial information greatly increased the predictive power of the models. Further, the sequence-specific features alone are reliable predictors, and the addition of sequence-specific features to other features greatly enhanced their predictive power.

Model validation using independent datasets

The above described models performed well on mESC data using 10-fold cross validation as a validation strategy. To further validate, we used independent datasets, which were not seen by the models during the training. We used publicly available data in four human tumor cell-types, including B-cell lymphoma (P493-6), multiple myeloma (MM1.S), small cell lung carcinoma (H2171) and glioblastoma (U87). We chose these cell-types because ChIP-seq data for MED1 and the two top-ranked chromatin features, BRD4 and H3K27ac, were publicly available (Table S2 in Additional file 1). Initially, we used H3K27ac ChIP-seq peaks 2 kb upstream and downstream of the transcription start site (TSS), to define constituent enhancers and ranked those constituents based on MED1 ChIP-seq signal to define super-enhancers as described in [16–18].

First, we trained the model on ESC data and tested it on each of the four human cell-type data and achieved AUC=0.92, 0.90, 0.90 and 0.86 for P493-6, MM1.S and U87 cells, respectively (Fig. 3h). We also checked the classification accuracy after combining five cell-types data by training the model on four cell-types data and testing it on one of the remaining cell-type data and repeated this for all the combinations (Figure S2H in Additional file 1). We achieved the highest AUC=0.95 for the model tested on P493-6 cell-type data and the lowest AUC=0.88 for the model tested U87 cell-type data (Figure S2F in Additional file 1). The classification measures, including precision, recall, F1-score and AUC for each tested cell-type after training the model on remaining four cell-types can be found in (Table S3 in Additional file 1).

We next trained the model on one genome data and tested on another. We used four human cell-types (P493-6, Plasma cell, H2171 and U87), and one mouse cell-type (mESC) data. The model trained on mouse cell-type data tested it on human cell-type data, achieved AUC=0.90 (Fig. 3i). The model trained on human cell-type data and tested on mouse cell-type data, achieved AUC=0.85.

We also tested whether a model trained on constituent data can predict the stitched regions and vice versa. The model, trained with H3K27ac data, accurately predicted the stitched super-enhancers with AUC=0.92 (Fig. 3j). The same model when trained on stitched data and tested on constituent data, performed poorly with AUC=0.68.

Genome-wide profiles of Cdk8, Cdk9 and Smad3 at super-enhancers

Through the ranking of chromatin and transcription factors, we found that Cdk8, Cdk9 and Smad3 were important features along many known signatures of super-enhancers, including H3K27ac, Brd4, Med12 and p300 which are well characterized at super-enhancers [16–18, 28]. However, the genome-wide profiles of Cdk8, Cdk9 and Smad3 are not well characterized at super-enhancers. Hence, we investigated the genome-wide profiles of Cdk8, Cdk9 and Smad3 at super-enhancers, identified by using Med1 in mESC. We found that, ChIP-seq binding sites of Cdk8, Cdk9 and Smad3 were highly co-localized with Med1, Brd4, H3K27ac, p300 and DNaseI (Fig. 4a, b). Like Med1, the ChIP-seq density for Cdk8, Cdk9 and Smad3 is exceptionally higher at super-enhancers compared to typical enhancers (Fig. 4c). The ChIP-seq density of Med1, Cdk8, Cdk9 and Smad3 at super-enhancers is significantly higher compared to typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 4d). Similarly, Med1, Cdk8, Cdk9 and Smad3 are also enriched at ESC super-enhancers identified using Med1 (Fig. 4f). The ChIP-seq binding for Med1 and master TFs, including Sox2, Oct4 and Nanog, is exceptionally higher and forms clusters at super-enhancers, which are associated with cell-type-specific genes [16]. We found that the ChIP-seq binding sites for Cdk8, Cdk9 and Smad3 also form clusters at super-enhancer regions and associated with cell-type-specific genes. For example, the super-enhancer (mSE_00038) is associated with ESC pluripotency gene Sox2 (Fig. 3e). In another example, the super-enhancer (mSE_00085) is associated with the ESC pluripotency gene Nanog and the super-enhancer (mSE_00084) is associated with Dppa3 (developmental pluripotency associated 3) gene, which plays a key role in cell division and maintenance of cell pluripotency (Figure S4A in Additional file 1).

Furthermore, we calculated the Pearson's correlation of Med1 with Cdk8, Cdk9 and Smad3 and found a high and significant correlation (p-value < 2.2e-16) (Fig. 4g). The correlation between Med1/Cdk8 (Pearson's r = 0.90, p-value < 2.2e-16), Med1/Cdk9 (Pearson's r = 0.86, p-value < 2.2e-16), and Med1/Smad3 (Pearson's r = 0.76, p-value < 2.2e-16). Previous studies have used the ChIP-seq binding sites of the co-activator protein p300 to find enhancers [4, 10]. We found that Smad3 is significantly correlated with p300 (Pearson's r = 0.85, p-value < 2.2e-16) (Figure 4g).

Identification and characterization of super-enhancers by using Cdk8, Cdk9 and Smad3 in mESC

Since we found that Cdk8, Cdk9 and Smad3 are highly correlated with Med1 and co-occupy super-enhancers genome-wide (Fig. 4), we investigated the importance of Cdk8, Cdk9 and Smad3 in super-enhancer formation and also compared them with super-enhancers identified by Med1. We used ChIP-seq data and RNA-seq data to identify and characterize super-enhancers by using Cdk8, Cdk9 and Smad3 in mESC. We found 400, 494 and 435 super-enhancers by using Cdk8, Cdk9 and Smad3, respectively (Fig. 5a). A list of all the super-enhancers and typical enhancers can be found in

(Additional file 2). Further, Cdk8, Cdk9 and Smad3 successfully identified 88%, 84% and 73% of the Med1 super-enhancers, respectively (Fig. 5a). After Med1, we can see more clear distinction of super-enhancers and typical enhancers by using Cdk8 as compared with H3K27ac (Figure 5b). The majority of super-enhancers identified using Cdk8, Cdk9 and Smad3 do overlap with super-enhancers identified using Med1, which is 66% of the super-enhancers identified using Med1 (Fig. 5c).

The ChIP-seq density at super-enhancers, identified using Cdk8, Cdk9 and Smad3, is significantly higher compared with typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 5d). Our analysis showed that Cdk8 could separate most of the super-enhancers (88%) defined using Mediator complex (Med1) ChIP-seq signal.

In ESC, the DNA motifs Klf4 and Esrrb were particularly enriched at the constituents of super-enhancers, compared to typical enhancers [16]. Hence, we tested the frequency of these two motifs at the constituents of super-enhancers and typical enhancers; we defined using Cdk8, Cdk9 and Smad3. We found that the frequency of binding motifs Klf4 and Esrrb is significantly higher at constituents of super-enhancers than typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Figure S4G in Additional file 1). Further, when we compared the frequency of known ESC specific motifs (Oct4, Sox2, Nanog, Esrrb and Klf4) at the constituents of super-enhancers and typical enhancers defined by Med1, Cdk8, Cdk9 and Smad3. We found a higher frequency of these motifs at super-enhancers defined by Cdk9, Cdk8 and Smad3 as compared with Med1 (Fig. 5e). Further, the frequency of these motifs was slightly higher at the typical enhancers identified by Med1 as compared with Cdk8, Cdk9 and Smad3.

The genes associated with super-enhancers are significantly expressed, compared to genes associated with typical enhancers [16–18]. To test this we associated genes with the super-enhancers and typical enhancers as described in [16, 18]. We found that 65% of the Med1 super-enhancers associated genes were also associated with super-enhancers identified by Cdk8, Cdk9 and Smad3 (Fig. 5f). Further, these genes associated with super-enhancers were significantly expressed, compared to genes associated with typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 5g).

Super-enhancers known to be highly enriched for cell-type-specific master regulators and these regulators should have a higher rank. Hence, we checked the rank of super-enhancers, associated with the key cell identity genes, including Oct4, Sox2, Nanog, Esrrb and Klf4 in ESC, and ranked the factors based on the average rank of the super-enhancers that are associated with these genes (Fig. 5h). These genes were selected due to their important roles in the pluripotency and reprogramming of ESC biology [42–44]. The rankings for Med1 super-enhancers were downloaded from [18]. We found that Smad3 achieved a highest rank followed by Med1, Cdk9 and Cdk8. The Smad3 achieved a higher rank for Oct4, compared with other genes. This might be due to the fact that Smad3 co-bonded with the master transcription factor Oct4 genome-wide in ESC [45]. Further, we found almost similar ChIP-seq patterns for factors including Med1, H3K27ac, Brd4, Cdk8, Cdk9 and Smad3 at super-enhancers regions defined by all three factors (Cdk8, Cd9 and Smad3) and are associated with cell-type-specific genes, including Nanog and Dppa3 (Fig. 5i). Like Med1, the genes associated with super-enhancers ranked by Cdk8, Cdk9 and Smad3 were enriched with cell-type-specific GO terms, supporting the notion that super-enhancers regulate cellular identity genes (Figure S7a in Additional file 1). Taken together, our results indicate the role of Cdk8, Cdk9 and Smad3 in defining and formation of super-enhancers.

Identification and characterization of super-enhancers by using Smad3 in pro-B cells

It was particularly interesting to see Smad3 ranked the most informative among transcription factors, including Oct4, Sox2, Nanog, Esrrb, Klf4, Tcfcp2l1, Prdm14, Nr5a2, Stat3 and Tcf3 in mESC (Fig. 2b). In ESC, the highly ranked super-enhancers identified using Smad3 were associated with ES cell-identity genes, including Oct4, Sox2, Nanog, Klf4 and Esrrb compared to Med1 super-enhancers (Fig. 5f). A previous study showed that Smad3 co-occupies the master transcription factors genome-wide [45]. We also observed Smad3 turn to be the highly ranked feature when we ranked Smad3 and several other factors in pro-B cells (Fig. 2d)

Hence, we argued that Smad3 could be used to define super-enhancers instead of Med1. We already showed the ability of Smad3 in defining super-enhancers in ESC. To test this in more differentiated cells, we identified and characterized super-enhancers in pro-B cells using Smad3. We compared the super-enhancers identified using Smad3 with previously identified super-enhancers that use Med1 in pro-B cells [16]. The ChIP-seq density of Smad3 super-enhancers is exceptionally higher compared to typical enhancers (Fig. 6a). Further, Smad3 have strong binding along with Med1 and PU.1 at a super-enhancer(mSE_00293) which is associated with Foxo1 gene (Fig. 6b).

By using Smad3, we identified 694 super-enhancers and among these, 65% were identified by Med1 (Figure S4f in Additional file 1) and with Smad3 we can see a more clear distinction of super-enhancers and typical enhancers as compared with H3K27ac (Fig. 6c). The ChIP-seq density at super-enhancers identified using Smad3 is significantly higher, compared to typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 6d). The genes associated with Smad3 super-enhancers are significantly expressed, compared with typical enhancers (p-value < 2.2e-16, Wilcoxon rank sum test) (Fig. 6e). The GO terms for super-enhancers ranked by Smad3 are highly enriched and cell-type-specific, compared to Med1 (Figure S7b in Additional file 1).

Further, to test the functional importance of super-enhancers identified only by Smad3 or by Med1, we performed GO analysis on these subset of super-enhancers. Interestingly, the super-enhancers identified by Smad3 but not by Med1 turn to be highly enriched for cell-type-specific GO terms such as immune cell development and immune system development. While super-enhancers identified by Med1 but not by Smad3 low enriched for cell-type-specific GO terms (Fig. 6f). These results, taken together with previous studies, demonstrate the importance of Smad3 in super-enhancer formation and Smad3 could be used to define super-enhancers.

imPROSE predicts cell-type-specific super-enhancers

To further test the ability of imPROSE, we compared it with the most commonly used H3K27ac based method ROSE [16]. We have demonstrated above that H3K27ac was ranked as the most informative feature in mESC and Smad3 was ranked as the most informative feature in pro-B cells. Hence, we trained imPROSE using H3K27ac and Smad3 data in mESC and predicted super-enhancers in pro-B cells. Our model predicted ~2,000 super-enhancers in pro-B cells, but when we used ROSE with H3K27ac data, we found 934 super-enhancers (Additional files 3 and 4). To assess the functional importance of super-enhancers identified by these two methods, we performed GO enrichment analysis using GREAT tool [46]. We have listed the top 20 GO terms enriched at super-enhancers Interestingly, we found several GO terms pertinent to the specific biological functions in pro-B cells are highly enriched at the super-enhancers identified by imPROSE and low enriched at super-enhancers identified by ROSE. For example, immune system process, hematopoietic or lymphoid organ development and leukocyte differentiation are significantly enriched for genes associated with super-enhancers predicted by imPROSE (Fig. 7a). Further, cell-type-specific GO terms such as regulation of B cell activation and regulation of B cell proliferation were highly enriched for typical enhancers defined by ROSE as compared with imPROSE (Figure S8 in Additional file 1). This shows that several super-enhancers were possibly labelled as typical enhancers by ROSE.

To further assess the ability of imPROSE, we carried out motif analysis at the constituents of super-enhancers identified by both ROSE and imPROSE methods. We used TRAP tool [47] to find motifs and reported the top ranked motifs. We found that most of the motifs were highly enriched at super-enhancers identified by imPROSE (Fig. 7b). More interestingly, motifs such as PU.1 and Ebf1 were significantly enriched at SEs predicted by imPROSE, which have previously been shown to play important role in the control of B cell identity [48]. Therefore, taken together these results suggest that the integration of multiple factors appears to be a better way to predict super-enhancers than the the current H3K27ac based approach (ROSE).

Discussion

Super-enhancers regulate expression of key genes that are critical for cellular identity, thus, alterations at these regions can lead to several disorders. Hence, exploring these cis-regulatory elements and their features to uncover their molecular mechanisms will help us in designing better, precise, and personalized drugs.

In this study, we first presented a systematic approach to rank and access the importance of different features of super-enhancers. We investigated different features including histone modifications, chromatin regulators, transcription factors, DNA hypersensitive sites and DNA sequence motifs in mESC. We also analysed sequence-specific features including GC content, conversation score and repeat fraction in mESC and pro-B cells. We found new features including Cdk8, Cdk9, Smad3 and GC content as key features of super-enhancers along with many known features, which make super-enhancers distinct from typical enhancers. Further, we developed imPROSE, a supervised machine-learning model, which can accurately predict super-enhancers and their constituents. imPROSE trained on one cell-type data can predict super-enhancers and their constituents in other cell-type. imPROSE trained on only H3K27ac and Smad3 data in mESC can predict super-enhancers in pro-B cells, which are more cell-type specific compared to super-enhancers defined by the current H3K27ac based approach (ROSE).

Among the chromatin features we found that Brd4, H3K27ac, Cdk8, Cdk9, Med12 and p300 were the top six features and by assessing their individual predictive powers we achieved higher AUC for higher ranked features. Previous studies showed the importance of these highly ranked features and their role in transcriptional regulation. Through our ranking of features, H3K27ac achieved a higher ranking, compared to other histone modifications. H3K27ac was found as a mark to separate active enhancers from poised enhancers [11] - this shows that super-enhancers might be the clusters of active enhancers. The Mediator sub-units Med1 and Med12 has been known as master coordinators of cell lineage and development [16, 49]. We did not include Mediator sub-unit Med1 in our training data because a Med1 signal was used to define super-enhancers [16]. Bromodomain-containing protein 4 (Brd4), a member of the BET protein family which functions as an epigenetic reader and transcriptional regulator that binds acetylated lysines in histones [50], was ranked as the second highest important feature. Brd4 has been associated with anti-pause enhancers (A-PEs) which regulate the RNA Polymerase II (Pol II) promoter-proximal pause release [51, 52]. Brd4 regulates the positive transcription elongation factor b (P-TEFb) to allow Pol II phosphorylation and the subsequent elongation of target genes [52, 53]. In ESC it specifically governs the transcriptional elongation by occupying super-enhancers and by recruiting Mediator and Cyclin dependent kinase 9 (Cdk9) to these super-enhancers [54]. Cdk9, a sub-unit of P-TEFb, has been found at enhancers and promoters of active genes along with the Mediator coactivator [17]. Cyclin-dependent kinase 8 (Cdk8), a subunit of Mediator complex, positively regulates precise steps in the assembly of transcriptional elongation, including the recruitment of P-TEFb and BRD4 [38]. During the preparing of this manuscript, a very recent study has demonstrated that Cdk8 regulates the key genes associated with super-enhancers in acute myeloid leukaemia (AML) cells [55]. We identified and characterized super-enhancers in mESC by using Cdk8 and Cdk9, to further validate their importance in super-enhancer formation and cell identity.

Among the transcriptions factors, we found that Smad3, Esrrb, Klf4, Tcfcp2l1, Nr5f2a and Stat3 were the top ranked features and by assessing their individual predictive powers we achieved higher AUC for higher ranked features. It was particularly interesting to see that Smad3 was ranked as the best feature among the transcription factors including Esrrb and Klf4. We know that Smad3 is a target of the TGF-β signaling pathway, and studies have shown that Smad3 is recruited to enhancers formed by master transcription factors [45]. We found significant correlation between Smad3 and coactivators p300/CBP at super-enhancers and previous studies have shown that p300/CBP interacts with Smad3 [30, 31]. The evidence for the enrichment of Smad3 at super-enhancers shows how the transforming growth factor beta (TGF-β) signaling pathway can converge on key genes that control ES cell identity. A very recent study validates our findings by showing that super-enhancers provide a platform for signalling pathways, including TGF-β to regulate genes that control cell identity during development and tumorigenesis [56]. To validate further, we identified and characterized super-enhancers using Smad3 in mESC and pro-B cells. By integrating ChIP-seq and RNA-seq data we showed the importance of Smad3 in super-enhancer formation and cell identity.

By investigating sequence-specific features, we found that the constituents of super-enhancers were significantly GC-rich. The GC-richness of a genomic region is associated with several distinctive features that can affect the *cis*-regulatory potential of a sequence [32, 33]. GC-rich and AT-rich chromatin domains are marked by distinct patterns of histone modifications. GC-rich chromatin domains tend to occur in a more active conformation and histone deacetylase activity represses this propensity throughout the genome [33]. Also GC content and nucleosome occupancy are positively correlated [32] and GC-rich sequences promote nucleosome formation [34]. Transcription factors tend to bind GC-rich regions in the genome, regardless of the distance and orientation [32]. This suggests that there is a role for the GC content in the formation of super-enhancers, which control the cell-type-specific gene expression.

Enhancers function due to cooperative and synergistic interplay of different coactivators and transcription factors [57]. A recent study showed that multiple enhancer variants cooperatively contribute to altered expression of their gene targets [58]. It is not well understood whether constituents of super-enhancers work synergistically or additively. The constituents of super-enhancers make frequent physical contacts with one another [59] and extensive cooperative binding of transcription factors have been found at super-enhancers [60]. A study in ESC demonstrated the functional importance of super-enhancer constituents [56]. Further, two recent studies have suggested additive and functional hierarchy among the constituents of α-globin and Wap super-enhancer locus, respectively [61, 62]. But, a very recent study argues that it is still need to be determined whether the constituents of a super-enhancer functions synergistically or additively [63]. Through computational modelling and correlation analysis, we noticed a combinatorial relationship between chromatin regulators and transcription factors at the constituents of super-enhancers. This advanced our current understanding of the determinants of super-enhancers and led us to hypothesize that these combinatorial patterns may be involved in mediating super-enhancers. Further, the significant correlation of many cofactors at the constituents of super-enhancers suggests cooperative and synergistic interactions. These results, taken together with previous studies suggest a cooperative and synergistic interplay of between constituents of super-enhancers. More sophisticated experiments are needed to validate the functional importance of constituents within a super-enhancer, by utilizing the latest CRISPR-Cas9 system.

Conclusions

We integrated diverse types of genomic and epigenomics datasets to predict super-enhancers and their constituents in a cell-type-specific manner. We investigated the relative importance of each feature in predicting super-enhancers, and also their combinatorial predictive power. We demonstrated that the model trained on one cell-type can be used to predict super-enhancers in other cell-types, and also performed better than the current H3K27ac-based approach. More importantly, we found Cdk8, Cdk9 and Smad3 as new signatures, which can be used to define super-enhancers where Mediator or master transcription data is not available. Taken together with previous studies, our results suggest a possible cooperative and synergistic interactions of numerous factors at super-enhancers. Our feature analysis and prediction models can serve as a resource to further characterize and understand the formation of super-enhancers.

Methods

Data description

We downloaded 32 publicly available ChIP-seq and DNase-seq datasets in mouse embryonic stem cells (mESC) from Gene Expression Ominibus (GEO). These include four histone modifications: H3K27ac, H3K4me1, H3K4me3 and H3K9me3; DNA hypersensitive site (DNaseI); RNA polymerase II (Pol II); transcriptional co-activating proteins (p300, CBP); P-TFEb subunit (Cdk9); sub-units of Mediator complex (Med1, Med12, Cdk8); other chromatin regulators (Brg1, Brd4, Chd7); Cohesin (Smc1, Nipbl); subunits of Lsd1-NuRD complex (Lsd1, Mi2b) and 11 transcription factors (Oct4, Sox2, Nanog, Esrrb, Klf4, Tcfcp2l1, Prdm14, Nr5a2, Smad3, Stat3 and Tcf3). A detailed list of all datasets used in this study is provided in (Table S4 in Additional file 1).

To validate the model using independent data, we downloaded ChIP-seq datasets for (MED1, BRD4 and H3K27ac) in four human tumor cell-types, including B-cell lymphoma (P493-6), Multiple myeloma (MM1.S), Small cell lung carcinoma (H2171) and Glioblastoma (U87), which were not seen by the model during training. A detailed list of ChIP-seq datasets used to validate the model is provided in (Table S5 in Additional file 1).

To predict super-enhancers and perform features ranking in pro-B cells, we used ChIP-seq data for Med1, PU.1, Foxo1, Smad3, Ebf1, p300, H3K27ac, H3K4me1, H3K4me3 and Pol2 and also DNase-seq (Table S6 in Additional file 1).

We also obtained processed RNA-seq based gene expression data (RPKM) from [64] and [16] for mESC and pro-B cells, respectively. We downloaded super-enhancer regions in mESC and pro-B cells identified using Med1 ChIP-seq occupancy from dbSUPER [65]. We also used other genomic features including GC content, conservation score (phastCons) and repeat fraction downloaded from the UCSC table browser [66].

Data pre-processing and feature extraction

Initially, ChIP-seq reads were aligned to mouse genome-build mm9 using bowtie [67] (Version 0.12.9) with parameters (-k 1, -m 1, -n 2, -e 70, –best). We calculated read densities for 30 ChIP-seq datasets at the constituents of super-enhancers (646) and typical enhancers (9981) and normalized it as described in [16, 68] . Briefly, for each constituent region, reads were extended by 200bp and the density of reads per base pair was calculated using bamToGFF (https://github.com/BradnerLab/pipeline). Next, these densities were normalized in units of reads per million mapped reads per base pair (rpm/bp) with background subtraction. We used the similar approach for DNase-seq data but without background subtraction.

The data for model validation was aligned to hg19 as described above. We used MACS (Model-based Analysis of ChIP-Seq) [69] (Version 1.4.2) to perform the peak calling and to find ChIP-seq-enriched regions over background. We used a p-value (10^{-9}) as the enrichment threshold. To generate wiggle files, we used MACS with parameter -w -S --space=50.

For DNA sequence motif data, we collected DNA binding motif information (PWM) from the transfac professional database version 2014 [70] for all the 11 transcription factors (Fig. 2a). We computed the binding affinity score for the constituents of super-enhancer and typical enhancer sequences using the Transcription factor Affinity Prediction (TRAP) [47] using individual TF's position weight matrix (PWM).

Data sampling

There are two commonly used sampling approaches, over-sampling and under-sampling. In over-sampling we increased the size of the minority class while in under-sampling we through away the samples from the majority class to balance the data. The current data we are dealing with is highly imbalanced. We used Weka implementation of SMOTE [71] to perform over-sampling with parameters (nearest neighbours=5, random seed=1 and oversampling percentage=500). We used a hybrid approach by first applying over-sampling using SMOTE on the minority class and then under-sampling on the majority class. We performed under-sampling by randomly selecting a subset of size similar to the minority class. In this study we used hybrid-sampling approach to perform analysis because it performed better (Additional file 1).

Feature-ranking

To find an optimal feature subset, we used two Random Forest based approaches. First, we used Boruta algorithm [37] to rank important features. Briefly, it finds important features by measuring the relevance of each original feature with respect to a reference attribute using Random Forest. Second, we used Random Forest's out-of-bag approach to calculate the relative importance of each feature. Briefly, this approach takes one feature out and measures its relative importance and contribution to the model.

We divided the features into two groups with the aim to achieve two different goals. The 11 transcription factors in mouse embryonic stem cells, were used to rank the transcription factors and explore their importance in super-enhancer prediction. The other 20 chromatin features including, histone modifications (HMs), RNA polymerase II (Pol II), transcriptional co-activating proteins, chromatin regulators (CRs), Cohesin, and sub-units of Lsd1-NuRD complex, were used to develop a general super-enhancer prediction model which can be used further to predict super-enhancers in other cell-types. We also ranked these features together in mESC and pro-B cells.

Training data

We downloaded 10627 loci of constituents of super-enhancers and typical enhancers in mESC defined based of Med1 ChIP-seq signal [16]. Among these 646 were constituents of super-enhancers and 9981 were typical enhancers. The median size of enhancer constituents is 703bp and super-enhancer constituents is 862bp. After performing hybrid-sampling we have 10,336 instances of data and among these 50% (5,168) are constituents of super-enhancers, and 50% (5,168) are constituents of typical enhancers. We considered constituents of super-enhancers as positive class and constituents of typical enhancers as negative class. In total, we have 45 features, including 20 chromatin features, 11 transcription factors, 11 DNA motifs and three sequence-specific features. We excluded Med1 from our training data because super-enhancers were defined based on Med1 ChIP-seq signal [16]. Not surprisingly, we achieved best classification results by using Med1 as a feature.

Prediction models

We investigated six state-of-art supervised approaches including: Random Forest [72], Support Vector Machines (SVM) [73], K-Nearest Neighbor (k-NN) [74], AdaBoost [75], Decision Tree and Naive Bayes. For all the analysis, we used scikit-learn (version 0.14.1), a Python library for machine learning [76]. We used LibSVM [77] with a linear kernel and regularization parameter C = 1.0. We used Random Forest with the number of trees=20. We calculated an out-of-bag error to find the optimal number of trees to use for Random Forest (Figure S2C in Additional file 1). For other models we used default parameters set in the scikit-learn library.

***K*-mer based prediction**

We used a sequence specific enhancer prediction method (Kmer-SVM) [39] to classify constituents of typical enhancers and super-enhancers. We used the default settings with *k*-mer size=5, spectrum kernel, regularization parameter (C) = 1.0. We used 5-fold cross validation for model validation.

Performance evaluation

We used 10-fold stratified cross-validation (CV) to validate models, which makes the folds by preserving the percentage of samples for each class. Stratified CV is generally consider a better scheme than standard CV in terms of bias and variance [78]. To evaluate the performance of the models, we reported precision, recall, F1-score, area under the ROC curve (AUC) and the precision-recall curve (PRC). The receiver-operating characteristic (ROC) is a graphical representation of true positive rate (sensitivity) v/s false positive rate (1-sensitivity). The true positive rate is also known as sensitivity or recall. The false positive rate is also known as (1-sensitivity). F1 score is an accuracy measure, which considers both the precision and the recall of the test to compute the score. The mathematical representation of these measures is as follows:

- *Precision* = true positive/(true positive + false positive)
- *Recall* = true positive/(true positive + false negative)
- *F1 score* = 2 [(precision x recall)/(precision + recall)]

We also tested if the increase in model AUC is statistically significant by using permutation test (1000 runs). The p-value is calculated using Wilcoxon rank sum statistic.

Super-enhancer identification

We used ROSE (Rank Ordering of Super-Enhancers) with parameters (stitching distance=12.5 kb, TSS exclusive zone= +/- 2 kb) to define super-enhancers as described in [16]. We used the ChIP-seq peaks for H3K27ac as enhancer constituents and MED1 signal to rank them.

Assigning genes to super-enhancers and typical enhancers

We assigned genes to super-enhancers and typical enhancers using a proximity rule as descried in [16, 18]. It is known that enhancers tend to loop and communicate with target genes [7], and most of these enhancer-promoter interactions occur within a distance of ~50kb [79]. This approach identified a large proportion of true enhancer/promoter interactions in ESC [80]. Hence, we assigned all transcriptionally active genes to super-enhancers and typical enhancers within a 50kb window.

Motif analysis

We used FIMO (Find Individual Motif Occurrences) version 4.10.0 [81] for motif analysis with p-value $< 10^{-4}$ with a custom library of TRANSFAC motifs including (Oct4: M01124, Sox2: M01272, Nanog: M01123, Esrrb:M01589, Klf4: M01588). The number of occurrences of each motif were counted for the constituents of super-enhancer and typical enhancer regions. Motif analysis in Fig. 7b was performed with TRAP tool using TRANSFAC vertebrates motif library, mouse promoters as the background, and Benjamini-Hochberg as the correction [47]. A DNA sequence (FASTA format) was extracted from mm9 genome as input for FIMO and TRAP.

Gene ontology analysis

We performed the gene ontology analysis using Genomic Regions Enrichment of Annotations Tool (GREAT) web tool (version 3.0.0) [46] with whole-genome as background and default parameters. We reported the top Gene Ontology (GO) terms with the lowest p-value.

Visualization and statistical analysis

We generated box plots using R programming language by extended the whiskers to 1.5x the interquartile range. The P-values were calculated based on Wilcoxon signed-rank test for box plots, by using wilcox.test function in R. We used ngs.plot [82], to generate heat maps and normalized binding profiles at the constituents of super-enhancers and typical enhancers and their flanking 3kb regions (for example, Fig. 1a, 4a).

imPROSE availability

To foster the reproducible research, we developed our analysis pipeline as an open-source Python package with a command line interface and made it freely available for academic use at https://github.com/asntech/improse. A detailed documentation can be found at http://improsedoc.readthedocs.io/.

Additional files

Additional file 1: A PDF document contains Supplementary Figures S1–S8 and Tables S1-S3. And also the public dataset used in this study in Supplementary Table S4-S6.

Additional file 2: An Excel spreadsheet contains a list of super-enhancers and typical enhancers and their associated genes in mESC identified by using Cdk8, Cdk9 and Smad3.

Additional file 3: An Excel spreadsheet contains a list of super-enhancers and typical enhancers identified by using H3K27ac and Smad3 in pro-B cell.

Additional file 4: An Excel spreadsheet contains a list of super-enhancers and typical enhancers predicted by imPROSE in pro-B cell using Smad3 and H3K27ac data.

Abbreviations

TFs: Transcription Factors; ChIP-seq: Chromatin immune precipitation followed by high-throughput sequencing; ESCs: Embryonic Stem Cells; SEs: Super-enhancers; TEs: Typical enhancers; imPROSE: Integrated methods for prediction of super-enhancer; TSS: Transcription Start Site; SNPs: Single Nucleotide Polymorphisms; ROC: Receiver Operating Characteristic; AUC: Area Under the Curve; GEO: Gene Expression Omnibus; GO: Gene Ontology.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported in part by the National Basic Research Program of China [2012CB316504], Hi-tech Research and Development Program of China [2012AA020401] and NSFC grant [91010016].

Authors’ contributions

AK conceived and designed the experiments. XG reviewed and approved experiment design. AK performed the experiments and analysed the data. AK wrote the manuscript and XG reviewed it. All authors read and approved the final manuscript.

Acknowledgements

We thank Anthony Mathelier and Masood ur Rehman Kayani for their comments and suggestions to improve the manuscript. We would like to thank Lydia Danner for proofreading the manuscript.

Author details

[1] MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/Department of Automation, Tsinghua University, Beijing, 100084, China
[2] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0349 Oslo, Norway
[3] School of Life Sciences, Tsinghua University, Beijing, 100084, China

References

1. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching K a, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov V V, Stewart R, Thomson J a, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459**:108–112.

2. Heinz S, Romanoski CE, Benner C, Glass CK: **The selection and function of cell type-specific enhancers**. *Nat Rev Mol Cell Biol* 2015, **16**:144–154.

3. Levine M, Cattoglio C, Tjian R: **Looping back to leap forward: transcription enters a new era.** *Cell* 2014, **157**:13–25.

4. Shlyueva D, Stampfel G, Stark A: **Transcriptional enhancers: from properties to genome-wide predictions.** *Nat Rev Genet* 2014, **15**:272–86.

5. Wamstad J a, Wang X, Demuren OO, Boyer L a: **Distal enhancers: new insights into heart development and disease.** *Trends Cell Biol* 2014, **24**:294–302.

6. Kolovos P, Knoch T a, Grosveld FG, Cook PR, Papantonis A: **Enhancers and silencers: an**

integrated and simple model for their function. *Epigenetics Chromatin* 2012, **5**:1.

7. Ong C-T, Corces VG: **Enhancer function: new insights into the regulation of tissue-specific gene expression.** *Nat Rev Genet* 2011, **12**:283–93.

8. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, et al.: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489**:75–82.

9. Banerji J, Rusconi S, Schaffner W: **Expression of a β-globin gene is enhanced by remote SV40 DNA sequences**. *Cell* 1981, **27**:299–308.

10. Visel A, Blow MJ, Li Z, Zhang T, Akiyama J a, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio L a: **ChIP-seq accurately predicts tissue-specific activity of enhancers.** *Nature* 2009, **457**:854–8.

11. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato M a, Frampton GM, Sharp P a, Boyer L a, Young R a, Jaenisch R: **Histone H3K27ac separates active from poised enhancers and predicts developmental state.** *Proc Natl Acad Sci U S A* 2010, **107**:21931–21936.

12. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U: **Predicting cell-type-specific gene expression from regions of open chromatin.** *Genome Res* 2012, **22**:1711–1722.

13. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, Taatjes DJ, Dekker J, Young RA: **Mediator and cohesin connect gene expression and chromatin architecture.** *Nature* 2010, **467**:430–435.

14. Allen BL, Taatjes DJ: **The Mediator complex: a central integrator of transcription.** *Nat Rev Mol Cell Biol* 2015, **16**:155–166.

15. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh Y-H, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung W-K, Clarke ND, Wei C-L, Ng H-H: **Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells**. *Cell* 2008, **133**:1106–1117.

16. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA: **Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes**. *Cell* 2013, **153**:307–19.

17. Lovén J, Hoke H a, Lin CY, Lau A, Orlando D a, Vakoc CR, Bradner JE, Lee TI, Young R a: **Selective inhibition of tumor oncogenes by disruption of super-enhancers.** *Cell* 2013, **153**:320–34.

18. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova A a, Hoke H a, Young R a: **Super-enhancers in the control of cell identity and disease.** *Cell* 2013, **155**:934–47.

19. Pott S, Lieb JD: **What are super-enhancers?** *Nat Genet* 2014, **47**:8–12.

20. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A, Lin CY, Royo R, Sanchez-Mut J V., Martinez R, Gut M, Torrents D, Orozco M, Gut I, Young RA, Esteller M: **Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer**. *Genome Biol* 2016, **17**:11.

21. Lin CY, Erkek S, Tong Y, Yin L, Federation AJ, Zapatka M, Haldipur P, Kawauchi D, Risch T, Warnatz H-J, Worst BC, Ju B, Orr BA, Zeid R, Polaski DR, Segura-Wang M, Waszak SM, Jones DTW, Kool M, Hovestadt V, Buchhalter I, Sieber L, Johann P, Chavez L, Gröschel S, Ryzhova M, Korshunov A, Chen W, Chizhikov V V., Millen KJ, et al.: **Active medulloblastoma enhancers reveal subgroup-specific cellular origins**. *Nature* 2016:1–20.

22. Hah N, Benner C, Chong L, Yu RT, Downes M, Evans RM: **Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs**. *Proc Natl Acad Sci U S A* 2014.

23. Chapuy B, McKeown MR, Lin CY, Monti S, Roemer MGM, Qi J, Rahl PB, Sun HH, Yeda KT, Doench JG, Reichert E, Kung AL, Rodig SJ, Young R a, Shipp M a, Bradner JE: **Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma.** *Cancer Cell* 2013, **24**:777–90.

24. Mansour MR, Abraham BJ, Anders L, Gutierrez A, Durbin AD, Lawton L, Sallan SE, Silverman LB, Loh ML, Hunger SP, Sanda T, Richard A: **An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element**. *Science (80-)* 2014, **346**:1373–1377.

25. Ooi WF, Xing M, Xu C, Yao X, Ramlee MK, Lim MC, Cao F, Lim K, Babu D, Poon L-F, Lin Suling J, Qamra A, Irwanto A, Qu Zhengzhong J, Nandi T, Lee-Lim AP, Chan YS, Tay ST, Lee MH, Davies JOJ, Wong WK, Soo KC, Chan WH, Ong HS, Chow P, Wong CY, Rha SY, Liu J, Hillmer AM, Hughes JR, et al.: **Epigenomic profiling of primary gastric adenocarcinoma reveals super-enhancer heterogeneity**. *Nat Commun* 2016, **7**:12983.

26. Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama J a, van Bueren KL, Chines PS, Narisu N, Black BL, Visel A, Pennacchio L a, Collins FS: **Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants.** *Proc Natl Acad Sci U S A* 2013, **110**:17921–6.

27. Pasquali L, Gaulton KJ, Rodríguez-Seguí S a, Mularoni L, Miguel-Escalada I, Akerman I, Tena JJ, Morán I, Gómez-Marín C, van de Bunt M, Ponsa-Cobas J, Castro N, Nammo T, Cebola I, García-Hurtado J, Maestro MA, Pattou F, Piemonti L, Berney T, Gloyn AL, Ravassard P, Gómez-Skarmeta JL, Müller F, McCarthy MI, Ferrer J: **Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants.** *Nat Genet* 2014, **46**:136–43.

28. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SCJ, Erdos MR, Davis SR, Roychoudhuri R, Restifo NP, Gadina M, Tang Z, Ruan Y, Collins FS, Sartorelli V, O'Shea JJ: **Super-enhancers delineate disease-associated regulatory nodes in T cells**. *Nature* 2015, **520**:558–562.

29. Witte S, Bradley A, Enright AJ, Muljo SA: **High-density P300 enhancers control cell state transitions**. *BMC Genomics* 2015, **16**:903.

30. Inoue Y, Itoh Y, Abe K, Okamoto T, Daitoku H, Fukamizu a, Onozaki K, Hayashi H: **Smad3 is acetylated by p300/CBP to regulate its transactivation activity.** *Oncogene* 2007, **26**:500–8.

31. Pouponnot C, Jayaraman L, Massague J: **Physical and Functional Interaction of SMADs and p300/CBP**. *J Biol Chem* 1998, **273**:22865–22869.

32. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res* 2012, **22**:1798–812.

33. Dekker J: **GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p.** *Genome Biol* 2007, **8**:R116.

34. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A: **Determinants of nucleosome organization in primary human cells**. *Nature* 2011, **474**:516–520.

35. Niederriter A, Varshney A, Parker S, Martin D: **Super Enhancers in Cancers, Complex Disease, and Developmental Disorders**. *Genes (Basel)* 2015, **6**:1183–1200.

36. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, Turner JMA, Bertelsen MF, Murchison EP, Flicek P, Odom DT: **Enhancer Evolution across 20 Mammalian Species**. *Cell* 2015, **160**:554–566.

37. Kursa MB, Rudnicki WR: **Feature Selection with the Boruta Package**. *J Stat Softw* 2010, **36**.

38. Donner AJ, Ebmeier CC, Taatjes DJ, Espinosa JM: **CDK8 is a positive regulator of transcriptional elongation within the serum response network**. *Nat Struct Mol Biol* 2010, **17**:194–201.

39. Fletez-Brant C, Lee D, McCallion AS, Beer M a: **kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets.** *Nucleic Acids Res* 2013, **41**(Web Server issue):W544-56.

40. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh T-Y, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40**:897–903.

41. Ernst J, Kellis M: **Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types**. *Genome Res* 2013, **23**:1142–1154.

42. Feng B, Jiang J, Kraus P, Ng J-H, Heng J-CD, Chan Y-S, Yaw L-P, Zhang W, Loh Y-H, Han J, Vega VB, Cacheux-Rataboul V, Lim B, Lufkin T, Ng H-H: **Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb**. *Nat Cell Biol* 2009, **11**:197–203.

43. Young RA: **Control of the embryonic stem cell state.** *Cell* 2011, **144**:940–54.

44. Takahashi K, Yamanaka S: **Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors**. *Cell* 2006, **126**:663–676.

45. Mullen AC, Orlando D a, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young R a: **Master transcription factors determine cell-type-specific responses to TGF-β signaling.** *Cell* 2011, **147**:565–76.

46. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nat Biotechnol* 2010, **28**:495–501.

47. Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, Masri N El, Roider HG, Manke T, Vingron M: **Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs.** *Nat Protoc* 2011, **6**:1860–9.

48. Lin YC, Jhunjhunwala S, Benner C, Heinz S, Welinder E, Mansson R, Sigvardsson M, Hagman J, Espinoza C a, Dutkowski J, Ideker T, Glass CK, Murre C: **A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate.** *Nat Immunol* 2010, **11**:635–43.

49. Yin J-W, Wang G: **The Mediator complex: a master coordinator of transcription and cell lineage development.** *Development* 2014, **141**:977–87.

50. Belkina AC, Denis G V: **BET domain co-regulators in obesity, inflammation and cancer.** *Nat Rev Cancer* 2012, **12**:465–477.

51. Liu W, Ma Q, Wong K, Li W, Ohgi K, Zhang J, Aggarwal AK, Rosenfeld MG: **Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release.** *Cell* 2013, **155**:1581–95.

52. Zhang W, Prakash C, Sum C, Gong Y, Li Y, Kwok JJT, Thiessen N, Pettersson S, Jones SJM, Knapp S, Yang H, Chin K-C: **Bromodomain-containing protein 4 (BRD4) regulates RNA polymerase II serine 2 phosphorylation in human CD4+ T cells.** *J Biol Chem* 2012, **287**:43137–55.

53. Itzen F, Greifenberg AK, Bösken C a., Geyer M: **Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation**. *Nucleic Acids Res* 2014, **42**:7577–7590.

54. Di Micco R, Fontanals-Cirera B, Low V, Ntziachristos P, Yuen SK, Lovell CD, Dolgalev I,

Yonekubo Y, Zhang G, Rusinova E, Gerona-Navarro G, Cañamero M, Ohlmeyer M, Aifantis I, Zhou M-M, Tsirigos A, Hernando E: **Control of Embryonic Stem Cell Identity by BRD4-Dependent Transcriptional Elongation of Super-Enhancer-Associated Pluripotency Genes**. *Cell Rep* 2014:1–14.

55. Pelish HE, Liau BB, Nitulescu II, Tangpeerachaikul A, Poss ZC, Silva DH Da, Caruso BT, Arefolov A, Fadeyi O, Christie AL, Du K, Banka D, Schneider E V, Jestel A, Zou G, Si C, Ebmeier CC, Bronson RT, Krivtsov A V, Myers AG, Kohl NE, Kung AL, Armstrong SA, Madeleine E, Taatjes DJ, Shair MD: **Mediator kinase inhibition further activates super-enhancer-associated genes in AML**. *Nature* 2015, **526**:273–276.

56. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA: **Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers**. *Mol Cell* 2015:1–9.

57. Carey M: **The enhanceosome and transcriptional synergy**. *Cell* 1998, **92**:5–8.

58. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sallari R, Lupien M, Markowitz S, Scacheri PC: **Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits**. *Genome Res* 2014, **24**:1–13.

59. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI, Zhao K, Young RA: **Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes**. *Cell* 2014, **159**:374–387.

60. Siersbæk R, Rabiee A, Nielsen R, Sidoli S, Traynor S, Loft A, Poulsen LLC, Rogowska-Wrzesinska A, Jensen ON, Mandrup S: **Transcription factor cooperativity in early adipogenic hotspots and super-enhancers.** *Cell Rep* 2014, **7**:1443–55.

61. Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen LLP, Kassouf MT, Oudelaar AM, Sharpe JA, Suciu MC, Telenius J, Williams R, Rode C, Li P-S, Pennacchio LA, Sloane-Stanley JA, Ayyub H, Butler S, Sauka-Spengler T, Gibbons RJ, Smith AJH, Wood WG, Higgs DR: **Genetic dissection of the α-globin super-enhancer in vivo.** *Nat Genet* 2016(July):1–12.

62. Shin HY, Willi M, Yoo KH, Zeng X, Wang C, Metser G, Hennighausen L: **Hierarchy within the mammary STAT5-driven Wap super-enhancer.** *Nat Genet* 2016, **48**:904–11.

63. Dukler N, Gulko B, Huang Y, Siepel A: **Is a super-enhancer greater than the sum of its parts□?** 2017, **49**:2–7.

64. Ouyang Z, Zhou Q, Wong WH: **ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells.** *Proc Natl Acad Sci U S A* 2009, **106**:21521–21526.

65. Khan A, Zhang X: **dbSUPER: a database of super-enhancers in mouse and human genome**. *Nucleic Acids Res* 2016, **44**(Database issue):D164–D171.

66. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita P a., Guruvadoo L, Haeussler M, Harte R a., Heitner S, Hickey G, Hinrichs a. S, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit a. F a., Speir ML, Zweig a. S, Haussler D, Kuhn RM, Kent WJ: **The UCSC Genome Browser database: 2015 update**. *Nucleic Acids Res* 2014, **43**:D670–D681.

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

68. Rahl PB, Lin CY, Seila AC, Flynn R a., McCuine S, Burge CB, Sharp P a., Young R a.: **C-Myc regulates transcriptional pause release**. *Cell* 2010, **141**:432–445.

69. Zhang Y, Liu T, Meyer C a, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM,

Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**:R137.

70. Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie a, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel a E, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**(Database issue):D108-10.

71. Hall M, National H, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA Data Mining Software□: An Update**. *SIGKDD Explor* 2009, **11**:10–18.

72. Breiman LEO: **Random Forests**. *Mach Learn* 2001, **45**:5–32.

73. Cortes C, Vapnik V: **Support-vector networks**. *Mach Learn* 1995, **20**:273–297.

74. Fix E, Jr JH: **Discriminatory analysis-nonparametric discrimination: consistency properties**. 1951, **57**:238–247.

75. Freund Y, Schapire RE: **A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting**. *J Comput Syst Sci* 1997, **55**:119–139.

76. Pedregosa F, Weiss R, Brucher M: **Scikit-learn□: Machine Learning in Python**. *J Mach Learn Res* 2011, **12**:2825–2830.

77. Chang C, Lin C: **LIBSVM□: A Library for Support Vector Machines**. *ACM Trans Intell Syst Technol* 2011, **2**.

78. Kohavi R: **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. 1995, **5**.

79. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K: **Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization.** *Cell Res* 2012, **22**:490–503.

80. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485**:376–80.

81. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics* 2011, **27**:1017–8.

82. Shen L, Shao N, Liu X, Nestler E: **ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases**. *BMC Genomics* 2014, **15**:284.

Figure titles and legends

Fig. 1: Analysis of features including chromatin, transcription factors and sequence specific. (a) Average ChIP-seq profile (RPM) of Med1, Brd4, H3K27ac, H3K4me1, DNaseI, p300, Cdk8, Cdk9, Smad3, Oct4, Sox2 and Nanog at the constituents of super-enhancers and typical enhancers, and their flanking 3kb regions (b) Correlation plot using Pearsons' correlation coefficient with hierarchical clustering of normalized ChIP-seq signals (rpm/bp) of 32 factors at the constituents of super-enhancers (646) and typical enhancers (9981). (c) Box plot shows the fraction of GC content, across the constituents of super-enhancers and typical enhancers in mESC and pro-B cells (p-value < 2.2e-16, Wilcoxon rank sum test). (d) Constituent enhancers size (bp) in mESC and pro-B cells (p-value < 2.2e-16, Wilcoxon rank sum test). (e) Conservation score (phastCons) in mESC (p-value = 0.6285, Wilcoxon rank sum test) and in pro-B cells (p-value < 1.7e-4, Wilcoxon rank sum test). (f) Repeat fraction in mESC (p-value = 0.0202, Wilcoxon rank sum test), pro-B cells (p-value = 0.8976, Wilcoxon rank sum test).

Fig. 2: Ranking of features including chromatin and transcription factors in mESC and pro-B cells. (a) Box plot shows the importance of features, including histone modifications, chromatin regulators, coactivators, DNaseI and other features. The feature importance is calculated by using a Random Forest based algorithm, Boruta. The colors represents (Blue= shadow features; Red=negative features; Orange=Important features). (b) Box plot shows the feature importance for 11 transcription factors in mESC. (c) Box plot shows the importance of feature after combining the chromatin, coactivators and transcription factors features in mESC. (d) Box plot shows the importance features, including chromatin, coactivators and master transcription factors in pro-B cells.

Fig. 3: The predictive power of histone modifications, chromatin regulators, co-activators, transcription factors other genomic features. (a) A detailed workflow of our computational prediction pipeline, imPROSE. (b) ROC plot shows the AUCs for six state-of-the-art machine learning models using all features. (c) Predictive power of top ranked 6 features including Brd4, H3k27ac, Cdk8, Cdk9, Med12 and p300. (d) Predictive power of top ranked 6 transcription factors including Smad3, Klf4, Esrrb, Stat3, Tcfcp2l1 and Nr5a2. (e) ROC plot for sequence specific features. (f) Predictive power of models trained on different combinations of features based on their functional importance. (g) ROC plot shows the AUCs of features grouped based on their type. (h) Model validation using independent data in four human cell-types (i) ROC plot for model trained on one genome and test on another. (j) Predicting stitched super-enhancers using model trained on constituents.

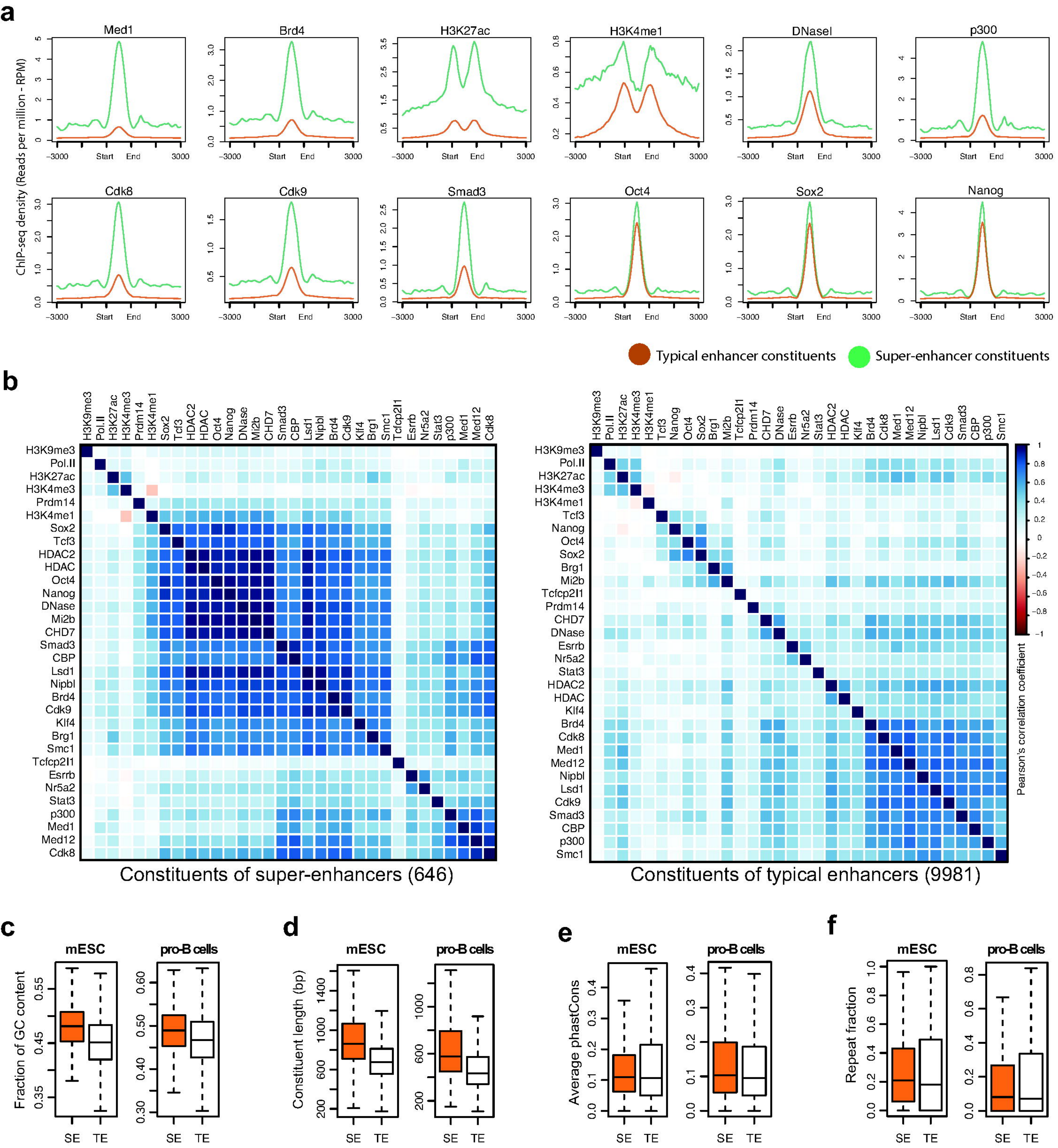
Fig. 4: Genome-wide profiles of Cdk8, Cdk9 and Smad3 across super-enhancers and typical enhancers. (a) The heatmap shows the genome-wide ChIP-seq binding profile of Cdk8, Cdk9 and Smad3 across factors including Med1, Brd4, H3K27ac, Crdk8, Cdk9, Smad3, p300 and DNaseI. (b) The average read count profile of factors including Med1, Brd4, H3K27ac, Crdk8, Cdk9, Smad3, p300 and DNaseI across ChIP-seq sumits of Cdk8, Cdk9 and Smad3. (c) ChIP-seq density plots centred around super-enhancers and typical enhancers defined by Med1. Flanking regions are 3 kb. (d) Box plot shows the ChIP-seq density (rpm/bp) for Med1, Cdk8, Cdk9 and Smad3 in super-enhancers and typical enhancers defined by Med1. Box plot whiskers extend to 1.5x the interquartile range. (e) ChIP-seq binding profiles of Med1, Crdk8, Cdk9, Smad3, Oct4, Sox2 and Nanog at super-enhancer (mSE_00038) at the locus of Sox2 gene. (f) The heatmap of Med1, Crdk8, Cdk9 and Smad3 intensity at 231 mESC super-enhancers. (g) The left three scatter plot shows the Pearson's

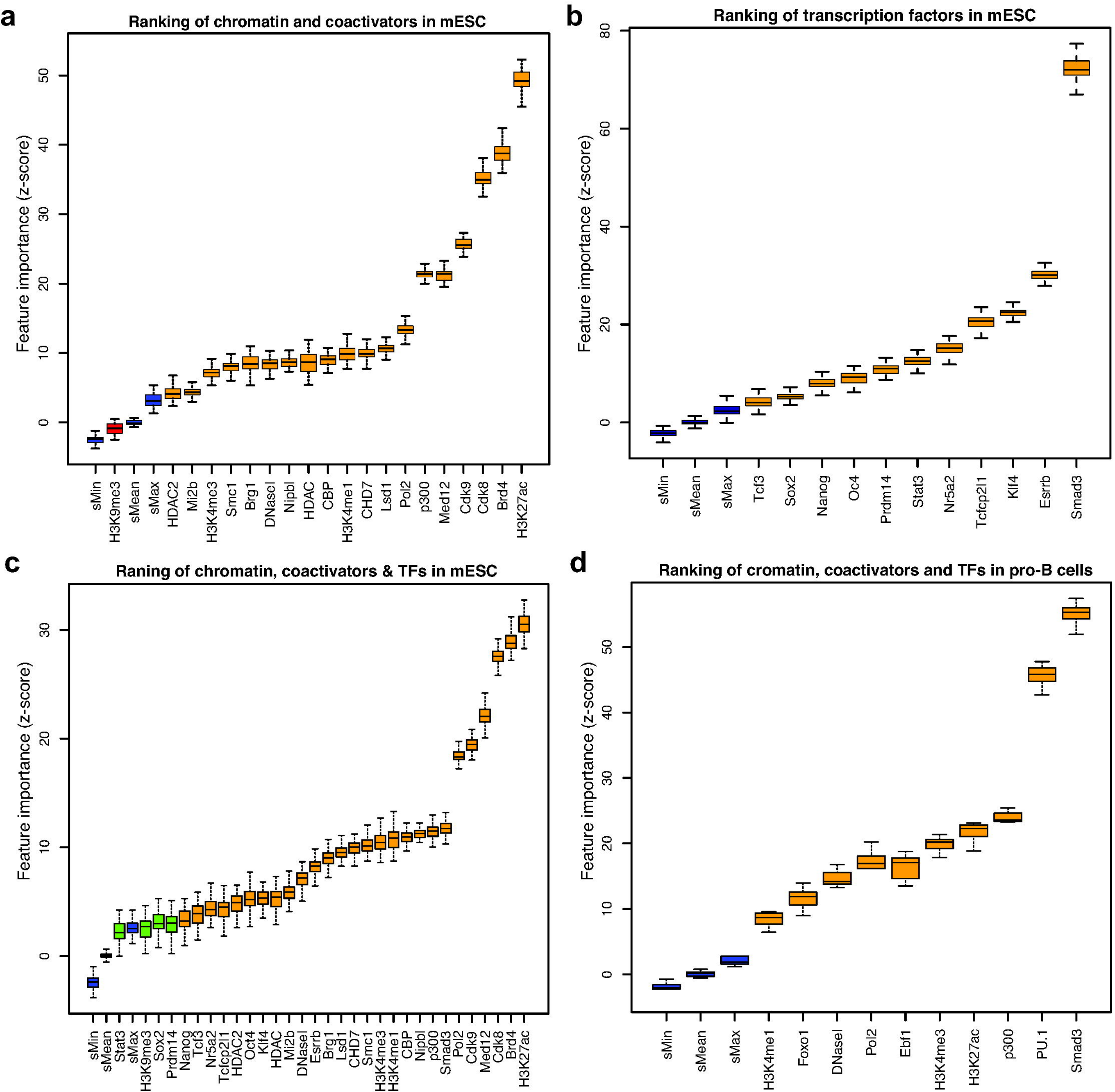
correlation of Med1 with Cdk8, Cdk9 and Smad3 respectively at the OSN regions. The right most scatter plot shows the Pearson's correlation between p300 and Smad3 at enhancer regions.

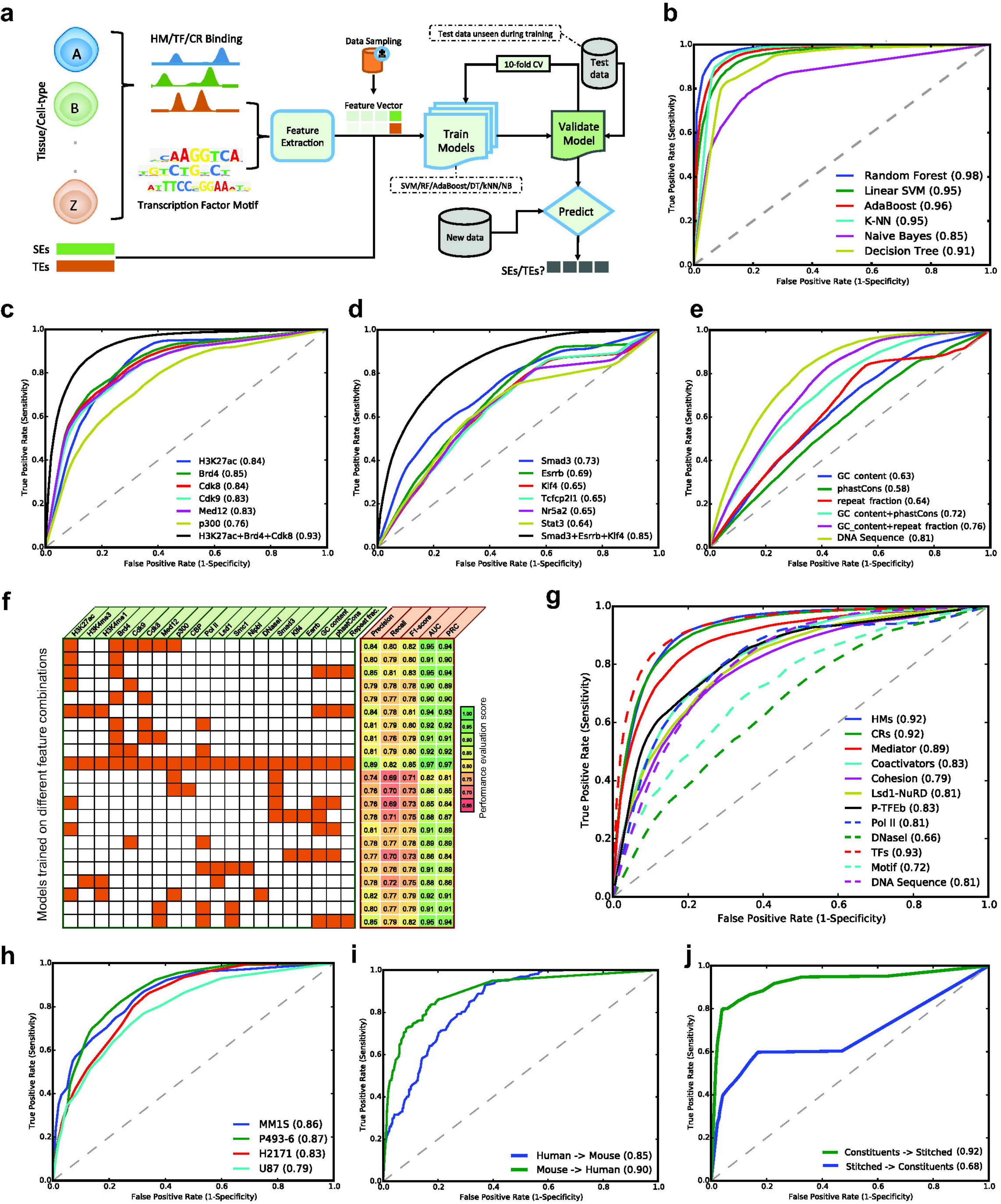
Fig. 5: Super-enhancers identified by using Cdk8, Cdk9 and Smad3. (a) The hokcyplot shows the cut-off used to separate super-enhancers from co-OSN (Oct4, Sox2, and Nanog) regions by using Cdk8, Cdk9 and Smad3. (b) The distribution of normalized ChIP-seq signal of Med1, H3K27ac, Cdk8, Cdk9 and Smad3 at mESC enhancers. For each factor, the values were normalized by dividing the ChIP-seq signal at each enhancer by the maximum signal. The rank of enhancer at each factor was measured independently. The figure zoomed at the cut-off so it can be visualized. (c) Venn diagram shows the number of super-enhancers overlapped, ranked using Med1, Cdk8, Cdk9 and Smad3. (d) Boxplot shows the ChIP-seq density (rpm/bp) in super-enhancers and typical enhancers defined by Cdk8, Cdk9 and Smad3 (p-value < 2.2e-16, Wilcoxon rank sum test). (e) Bar-plot shows the frequency of motifs (Oct4, Sox2, Nanog, Klf4 and Essrb) found at the constituents of super-enhancers and typical enhancers defined by Med1, Cdk8, Cdk9 and Smad3. (f) Venn diagram of genes associated with super-enhancers identified using Med1, Cdk8, Cdk9 and Smad3. The genes associated to Med1 super-enhancers are downloaded from dbSUPER. (g) Boxplot shows the gene expression (RPKM) in super-enhancers and typical enhancers defined by Cdk8, Cdk9 and Smad3. (h) Rank of factors based on the rank of super-enhancers associated with of ESC identity genes, including Sox2, Oct4, Nanog, Esrrb and Klf4. The table is sorted based on the average rank. (i) ChIP-seq binding profiles of different factors at the typical enhancer and super-enhancers at Dppa3 and Nanog gene locus in mESC.

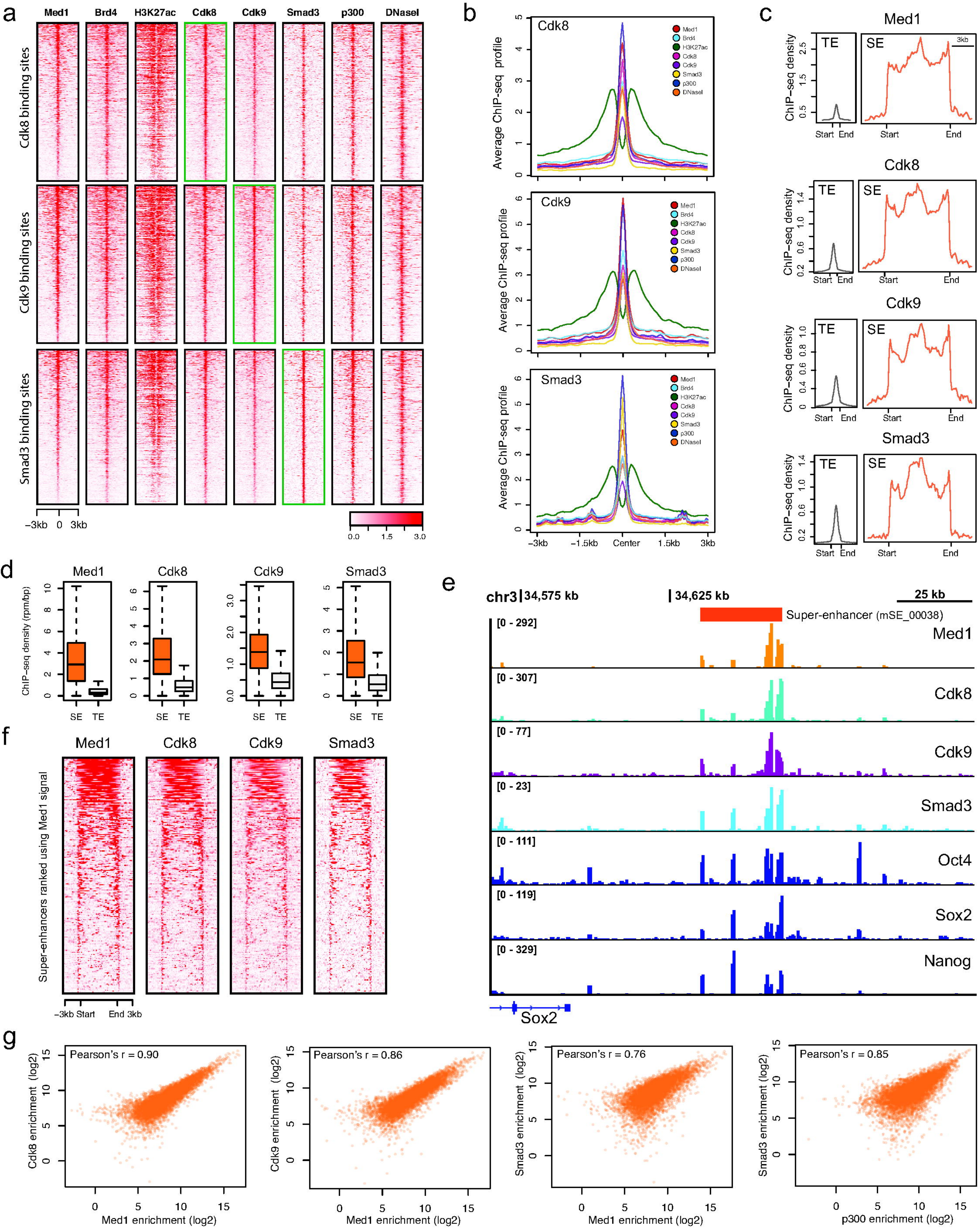
Fig. 6: Comparison of super-enhancers ranked using Med1 and Smad3 in pro-B cells. (a) Average ChIP-seq density of Med1 and Smad3 across 13,814 typical enhancers and 395 super-enhancers identified using Med1. The flanking region is 3kb. (b) ChIP-seq binding profiles for Med1, Smad3 and PU.1 at the locus of Foxo1 gene. The super-enhancer (mSE_00293) is associated with Foxo1 gene. (c) The distribution of normalized ChIP-seq signal of Med1, H3K27ac and Smad3 at pro-B enhancers. For each factor, the values were normalized by dividing the ChIP-seq signal at each enhancer by the maximum signal. The rank of enhancer at each factor was measured independently. The figure zoomed at the cut-off so it can be visualized. (d) Box-plot shows the Smad3 ChIP-seq density (rpm/bp) at super-enhancers and typical enhancers regions defined using Smad3 in pro-B cells. (e) Box plot shows the gene expression (RPKM) for the genes associated with super-enhancers and typical enhancers defined using Smad3 in pro-B cells. (f) The Venn diagrams show the overlap of super-enhancers identified using both Med1 and Smad3 in pro-B cells. Super-enhancer regions identified using Med1 in pro-B were obtained from [16]. Gene Ontology terms (Biological Process) for super-enhancers identified by Med1 only or Smad3 in pro-B cells.

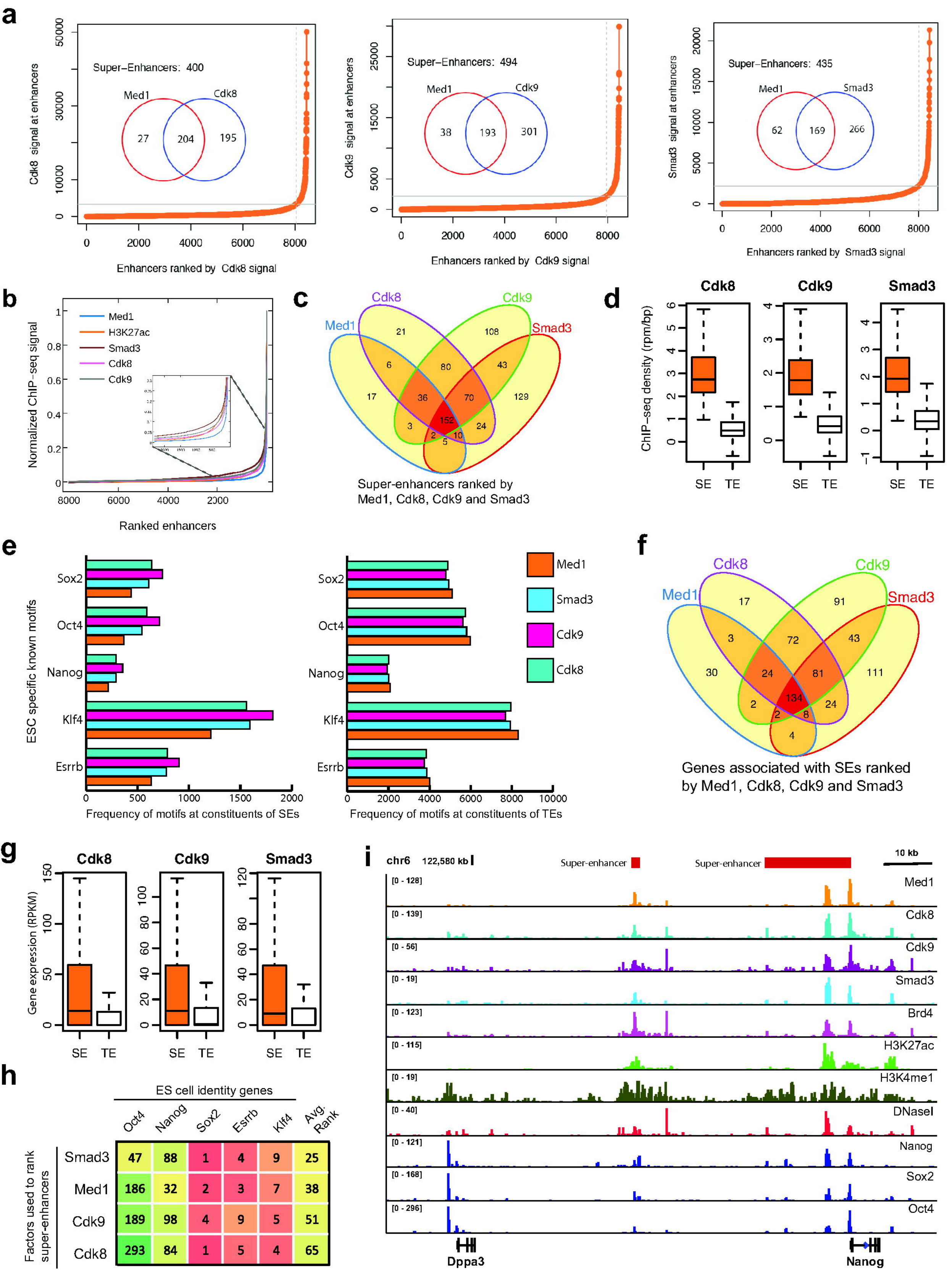
Fig. 7: Comparison of imPROSE and ROSE methods. (a) Top 20 GO terms enriched at the genes associated with super-enhancers predicted by imPROSE using H3K27ac and Smad3, and ROSE using H3K27ac data. (b) Top 20 motifs enriched at the DNA sequence of constituents of super-enhancers predicted by imPROSE using H3K27ac and Smad3, and ROSE using H3K27ac.

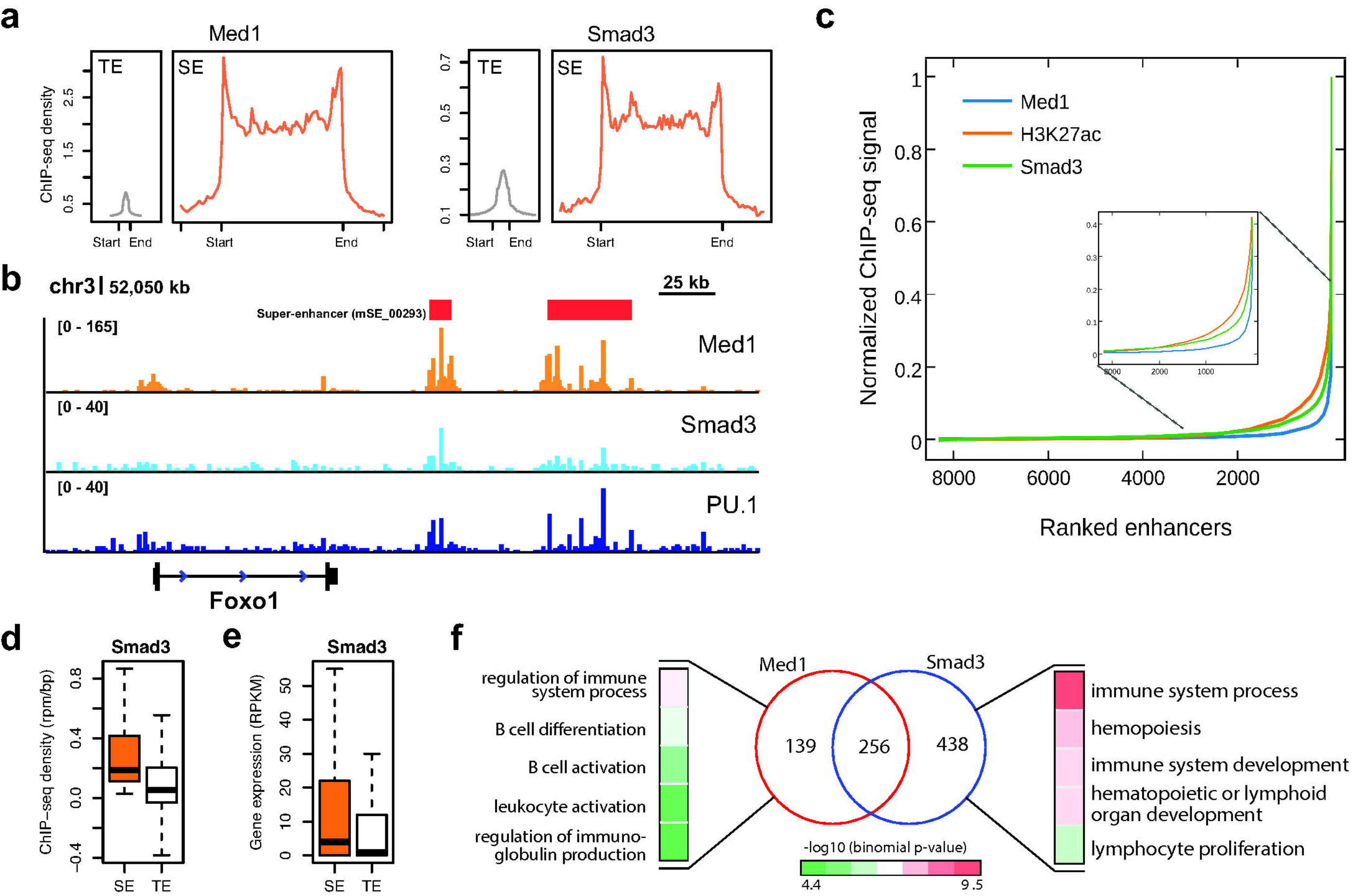

a
ChIP-seq density (Reads per million - RPM)
Med1
Brd4
H3K27ac
H3K4me1
DNaseI
p300
Cdk8
Cdk9
Smad3
Oct4
Sox2
Nanog
−3000
Start
End
3000
● Typical enhancer constituents
● Super-enhancer constituents
b
H3K9me3 Pol.II H3K27ac H3K4me3 Prdm14 H3K4me1 Sox2 Tcf3 HDAC2 HDAC Oct4 Nanog DNase Mi2b CHD7 Smad3 CBP Lsd1 Nipbl Brd4 Cdk9 Klf4 Brg1 Smc1 Tcfcp2l1 Esrrb Nr5a2 Stat3 p300 Med1 Med12 Cdk8
Constituents of super-enhancers (646)
H3K9me3 Pol.II H3K27ac H3K4me3 H3K4me1 Tcf3 Nanog Oct4 Sox2 Brg1 Mi2b Tcfcp2l1 Prdm14 CHD7 DNase Esrrb Nr5a2 Stat3 HDAC2 HDAC Klf4 Brd4 Cdk8 Med1 Med12 Nipbl Lsd1 Cdk9 Smad3 CBP p300 Smc1
Constituents of typical enhancers (9981)
Pearson's correlation coefficient
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
c
mESC
pro-B cells
Fraction of GC content
SE
TE
d
Constituent length (bp)
e
Average phastCons
f
Repeat fraction

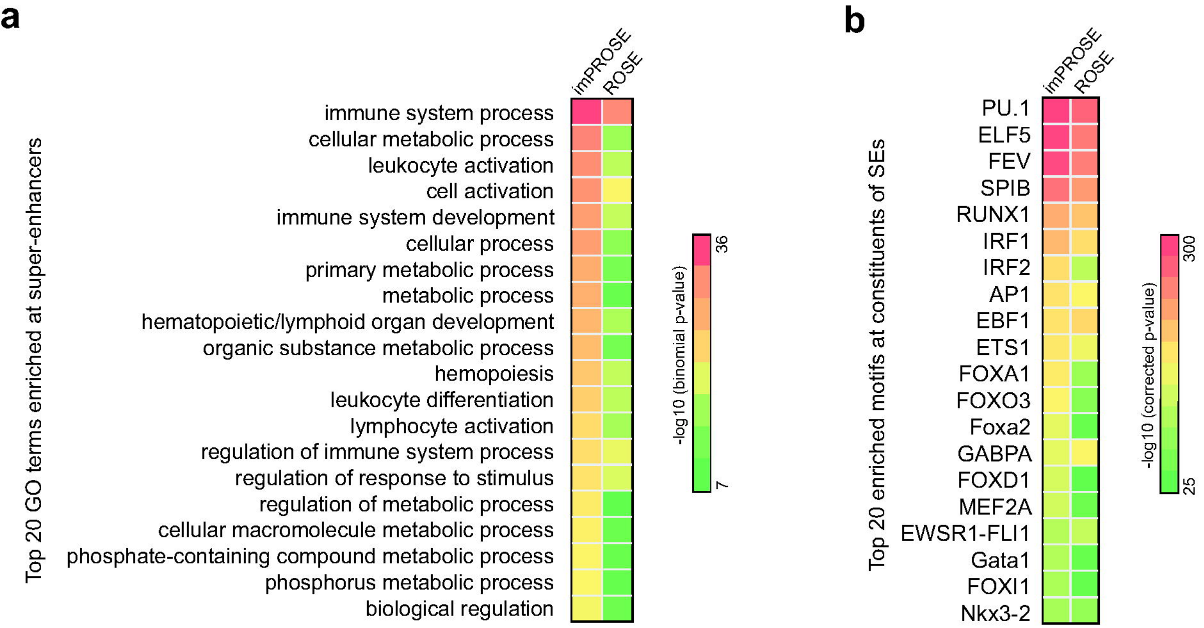

a
Ranking of chromatin and coactivators in mESC
Feature importance (z-score)
0
10
20
30
40
50
sMin
H3K9me3
sMean
sMax
HDAC2
Mi2b
H3K4me3
Smc1
Brg1
DNaseI
Nipbl
HDAC
CBP
H3K4me1
CHD7
Lsd1
Pol2
p300
Med12
Cdk9
Cdk8
Brd4
H3K27ac
b
Ranking of transcription factors in mESC
Feature importance (z-score)
0
20
40
60
80
sMin
sMean
sMax
Tcf3
Sox2
Nanog
Oc4
Prdm14
Stat3
Nr5a2
Tcfcp2l1
Klf4
Esrrb
Smad3
c
Raning of chromatin, coactivators & TFs in mESC
Feature importance (z-score)
0
10
20
30
sMin
sMean
Stat3
sMax
H3K9me3
Sox2
Prdm14
Nanog
Tcf3
Nr5a2
Tcfcp2l1
HDAC2
Oct4
Klf4
HDAC
Mi2b
DNaseI
Esrrb
Brg1
Lsd1
CHD7
Smc1
H3K4me3
H3K4me1
CBP
Nipbl
p300
Smad3
Pol2
Cdk9
Med12
Cdk8
Brd4
H3K27ac
d
Ranking of cromatin, coactivators and TFs in pro-B cells
Feature importance (z-score)
0
10
20
30
40
50
sMin
sMean
sMax
H3K4me1
Foxo1
DNaseI
Pol2
Ebf1
H3K4me3
H3K27ac
p300
PU.1
Smad3


a
Tissue/Cell-type
A
B
Z
HM/TF/CR Binding
Transcription Factor Motif
SEs
TEs
Feature Extraction
Data Sampling
Feature Vector
Test data unseen during training
10-fold CV
Test data
Train Models
Validate Model
SVM/RF/AdaBoost/DT/kNN/NB
Predict
New data
SEs/TEs?
b
Random Forest (0.98)
Linear SVM (0.95)
AdaBoost (0.96)
K-NN (0.95)
Naive Bayes (0.85)
Decision Tree (0.91)
True Positive Rate (Sensitivity)
False Positive Rate (1-Specificity)
c
H3K27ac (0.84)
Brd4 (0.85)
Cdk8 (0.84)
Cdk9 (0.83)
Med12 (0.83)
p300 (0.76)
H3K27ac+Brd4+Cdk8 (0.93)
d
Smad3 (0.73)
Esrrb (0.69)
Klf4 (0.65)
Tcfcp2l1 (0.65)
Nr5a2 (0.65)
Stat3 (0.64)
Smad3+Esrrb+Klf4 (0.85)
e
GC content (0.63)
phastCons (0.58)
repeat fraction (0.64)
GC content+phastCons (0.72)
GC_content+repeat fraction (0.76)
DNA Sequence (0.81)
f
Models trained on different feature combinations
H3K27ac H3K4me3 H3K4me1 Brd4 Cdk8 Cdk9 Med12 p300 CBP Pol II Lsd1 Smc1 Nipbl DNaseI Smad3 Klf4 Esrrb GC content phastCons Repeat frac.
Precision Recall F1-score AUC PRC
0.84 0.80 0.82 0.95 0.94
0.80 0.79 0.80 0.91 0.90
0.85 0.81 0.83 0.95 0.94
0.79 0.78 0.78 0.90 0.89
0.79 0.77 0.78 0.90 0.90
0.84 0.78 0.81 0.94 0.93
0.81 0.79 0.80 0.92 0.92
0.81 0.76 0.79 0.91 0.91
0.81 0.79 0.80 0.92 0.92
0.89 0.82 0.85 0.97 0.97
0.74 0.69 0.71 0.82 0.81
0.76 0.70 0.73 0.86 0.85
0.76 0.69 0.73 0.85 0.84
0.78 0.71 0.75 0.88 0.87
0.81 0.77 0.79 0.91 0.89
0.78 0.77 0.78 0.89 0.89
0.77 0.70 0.73 0.86 0.84
0.79 0.76 0.78 0.91 0.89
0.78 0.72 0.75 0.88 0.86
0.82 0.77 0.79 0.92 0.91
0.80 0.77 0.79 0.91 0.91
0.85 0.79 0.82 0.95 0.94
Perfomance evaluation score
1.00 0.95 0.90 0.85 0.80 0.75 0.70 0.65
g
HMs (0.92)
CRs (0.92)
Mediator (0.89)
Coactivators (0.83)
Cohesion (0.79)
Lsd1-NuRD (0.81)
P-TFEb (0.83)
Pol II (0.81)
DNaseI (0.66)
TFs (0.93)
Motif (0.72)
DNA Sequence (0.81)
h
MM1S (0.86)
P493-6 (0.87)
H2171 (0.83)
U87 (0.79)
i
Human -> Mouse (0.85)
Mouse -> Human (0.90)
j
Constituents -> Stitched (0.92)
Stitched -> Constituents (0.68)
0.0 0.2 0.4 0.6 0.8 1.0


a
Med1
Brd4
H3K27ac
Cdk8
Cdk9
Smad3
p300
DNaseI
Cdk8 binding sites
Cdk9 binding sites
Smad3 binding sites
−3kb 0 3kb
0.0 1.5 3.0
b
Cdk8
Cdk9
Smad3
Average ChIP-seq profile
Med1
Brd4
H3K27ac
Cdk8
Cdk9
Smad3
p300
DNaseI
−3kb −1.5kb Center 1.5kb 3kb
c
Med1
Cdk8
Cdk9
Smad3
TE
SE
3kb
ChIP-seq density
Start End
Start End
d
Med1
Cdk8
Cdk9
Smad3
ChIP-seq density (rpm/bp)
SE TE
e
chr3
34,575 kb
34,625 kb
25 kb
Super-enhancer (mSE_00038)
[0 - 292] Med1
[0 - 307] Cdk8
[0 - 77] Cdk9
[0 - 23] Smad3
[0 - 111] Oct4
[0 - 119] Sox2
[0 - 329] Nanog
Sox2
f
Med1
Cdk8
Cdk9
Smad3
Super-enhancers ranked using Med1 signal
−3kb Start End 3kb
g
Pearson's r = 0.90
Pearson's r = 0.86
Pearson's r = 0.76
Pearson's r = 0.85
Cdk8 enrichment (log2)
Cdk9 enrichment (log2)
Smad3 enrichment (log2)
Smad3 enrichment (log2)
Med1 enrichment (log2)
Med1 enrichment (log2)
Med1 enrichment (log2)
p300 enrichment (log2)
0 5 10 15


a
Super–Enhancers: 400
Med1
Cdk8
27
204
195
Cdk8 signal at enhancers
Enhancers ranked by Cdk8 signal
Super–Enhancers: 494
Med1
Cdk9
38
193
301
Cdk9 signal at enhancers
Enhancers ranked by Cdk9 signal
Super–Enhancers: 435
Med1
Smad3
62
169
266
Smad3 signal at enhancers
Enhancers ranked by Smad3 signal
b
Med1
H3K27ac
Smad3
Cdk8
Cdk9
Normalized ChIP−seq signal
Ranked enhancers
c
Med1
Cdk8
Cdk9
Smad3
21
108
6
80
43
17
36
70
129
152
3
24
2
10
5
Super-enhancers ranked by Med1, Cdk8, Cdk9 and Smad3
d
Cdk8
Cdk9
Smad3
ChIP−seq density (rpm/bp)
SE
TE
e
Sox2
Oct4
Nanog
Klf4
Esrrb
ESC specific known motifs
Frequency of motifs at constituents of SEs
Frequency of motifs at constituents of TEs
Med1
Smad3
Cdk9
Cdk8
f
Med1
Cdk8
Cdk9
Smad3
17
91
3
72
43
30
24
81
111
134
2
24
2
8
4
Genes associated with SEs ranked by Med1, Cdk8, Cdk9 and Smad3
g
Cdk8
Cdk9
Smad3
Gene expression (RPKM)
SE
TE
h
ES cell identity genes
Oct4
Nanog
Sox2
Esrrb
Klf4
Avg. Rank
Factors used to rank super-enhancers
Smad3 47 88 1 4 9 25
Med1 186 32 2 3 7 38
Cdk9 189 98 4 9 5 51
Cdk8 293 84 1 5 4 65
i
chr6
122,580 kb
Super-enhancer
Super-enhancer
10 kb
[0 - 128] Med1
[0 - 139] Cdk8
[0 - 56] Cdk9
[0 - 19] Smad3
[0 - 123] Brd4
[0 - 115] H3K27ac
[0 - 19] H3K4me1
[0 - 40] DNaseI
[0 - 121] Nanog
[0 - 168] Sox2
[0 - 296] Oct4
Dppa3
Nanog


a
Med1
Smad3
ChIP-seq density
0.5
1.5
2.5
TE
SE
Start
End
0.1
0.3
0.5
0.7
c
Normalized ChIP-seq signal
Med1
H3K27ac
Smad3
0
0.2
0.4
0.6
0.8
1
0.1
0.2
0.3
0.4
3000
2000
1000
8000
6000
4000
2000
Ranked enhancers
b
chr3
52,050 kb
25 kb
Super-enhancer (mSE_00293)
[0 - 165]
Med1
[0 - 40]
Smad3
[0 - 40]
PU.1
Foxo1
d
Smad3
ChIP-seq density (rpm/bp)
-0.4
0.0
0.4
0.8
SE
TE
e
Smad3
Gene expression (RPKM)
0
10
20
30
40
50
SE
TE
f
regulation of immune system process
B cell differentiation
B cell activation
leukocyte activation
regulation of immuno-globulin production
Med1
Smad3
139
256
438
-log10 (binomial p-value)
4.4
9.5
immune system process
hemopoiesis
immune system development
hematopoietic or lymphoid organ development
lymphocyte proliferation


a
imPROSE
ROSE
Top 20 GO terms enriched at super-enhancers
immune system process
cellular metabolic process
leukocyte activation
cell activation
immune system development
cellular process
primary metabolic process
metabolic process
hematopoietic/lymphoid organ development
organic substance metabolic process
hemopoiesis
leukocyte differentiation
lymphocyte activation
regulation of immune system process
regulation of response to stimulus
regulation of metabolic process
cellular macromolecule metabolic process
phosphate-containing compound metabolic process
phosphorus metabolic process
biological regulation
-log10 (binomial p-value)
36
7
b
imPROSE
ROSE
Top 20 enriched motifs at constituents of SEs
PU.1
ELF5
FEV
SPIB
RUNX1
IRF1
IRF2
AP1
EBF1
ETS1
FOXA1
FOXO3
Foxa2
GABPA
FOXD1
MEF2A
EWSR1-FLI1
Gata1
FOXI1
Nkx3-2
-log10 (corrected p-value)
300
25