UNIVERSITY OF CALIFORNIA, SAN DIEGO

# Navigating the Human Epigenome through Random Forests

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor in Philosophy

in

Bioinformatics and System Biology

by

Nisha Rajagopal

Committee in charge:

Professor Bing Ren, Chair
Professor Wei Wang, Co-Chair
Professor Vineet Bafna
Professor Trey Ideker
Professor Gene Yeo

2013

`

`

The Dissertation of Nisha Rajagopal is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

________________________________________

________________________________________

________________________________________

________________________________________
Co-Chair

________________________________________
Chair

University of California, San Diego

2013

# DEDICATION

This thesis is dedicated to my parents Chitra Rajagopal and V Rajagopal. They have always encouraged me in my pursuit of higher education and also been unconditionally supportive of the decisions I have taken towards my career goals. Even though having both their daughters far away from them has been difficult, they have never said a word of this to either of us and always tried to shield me and my sister from anything that might distress us or cause us to be distracted from our life goals. I also dedicate this thesis to my little sister, Aruna, who as in all aspects of my life, has been my confidante and friend through the last 6 years of my PhD.

My best friend and roommate, Josue, has been an inseparable part of my life in San Diego. Together, we have gone through all the ups and downs of the PhD life and helped each other grow as people and as scientists. He has always believed in my potential even when I didn't believe in myself and has kept me strong through my moments of weakness. My life outside of the lab would not have been anywhere as memorable and meaningful if he had not been a part of it.

When I met my boyfriend, Demetrius halfway through my PhD, I realized that this country far away from home could feel like home too. His sweet, unassuming support through the past 3 years, even when he moved to the other coast, has kept me grounded and happy through the most trying times.

Especially at such times, I was not always the easiest person, but he has always been infinitely patient and for this, I am very grateful.

I would like to thank my very good friend Indrajeet Singh, whom I have known for over 8 years and is like a brother to me. He was always ready to drive down from Riverside to San Diego at the drop of a hat if ever I needed him. Despite his own busy schedule, he has never said “No” to any favor I ever asked from him. And for this unique friendship I am truly grateful.

I have been fortunate to have some amazing classmates in the Bioinformatics program at UCSD - Harish, Nitin and Roger. They have always been there for me, be it at a professional or personal level. Harish and Nitin have been like family and I look forward to this connection even as we all go on to do different things.

During my years in San diego, I have made some really good friends, Sowparnika, Gaby, Marta, Nancy, Niki, Barbara and Ruben, who have made my life outside lab memorable and fun. Among these, especially, I would like to thank Marta, who has been very generous with her help during my thesis days.

My cousin Anuradha, has kept a watchful eye on me, from her apartment in Manhattan, since the day I landed on the coast of California. She called regularly during my first two years, to make sure I was happy and comfortable and did not want for anything. I have had some really fun times in New York city thanks to her and she has helped me feel at home in the United states. My uncle Ravi, also on the east coast, kept in touch during my first days here and was always most generous in offering any help that I might need.

When I first arrived in California, friends of my family, put me up at their houses and helped me settle into my apartment in San Diego. Moving in would have been a lot harder if I did not have these people helping me. Thanks to Sudha and Krishnan as well as Deepa and Srinivas, for their warm hospitality.

Last but definitely not the least, I want to thank my co-workers and friends, the Ren lab members whom I have known for the past 4 years. I want to thank them for always taking the time to help me with my research and improve on various aspects of my development as a scientist including public speaking and writing. In particular, I am grateful to my running buddies, Gary, Andrea, Inkyung and Feng, for providing a stimulating diversion from work. In particular, I would like to thank Gary Hon, who has acted as a bioinformatics mentor to me in lab. He has been selfless and generous in sharing any of his work that could help me. Finally, I thank my mentor Bing Ren for giving me the opportunities to travel, learn and grow.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

Chapter 2, in full, has been published as Nisha Rajagopal, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, Bing Ren. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. Plos Computational biology 2013. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Nisha Rajagopal, Jason Ernst, Pradipta Ray, Jie Wu, Michael Zhang, Manolis Kellis, Bing Ren. Distinct and Predictive histone lysine acetylation patterns at promoters, enhancers and gene bodies. The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, is currently being prepared for submission for publication of the material. Nisha Rajagopal, Danny Leung, Wei Xie, Pradipta Ray, Ah Young Lee, Chia-An Yen, Michael Q.Zhang, Bing Ren. Dynamic epigenetic signatures at regulatory elements in human tissues. The dissertation author was one of two primary investigators and author of this material.

# VITA

2007 Bachelor of Technology, Indian Institute of Technology Guwahati, Biotechnology

2013 Doctor of Philosophy, University of California, San Diego
Bioinformatics and Systems Biology
Dissertation title: Navigating the Human Epigenome through Random Forests

# PUBLICATIONS

**1.** Rajagopal N, Xie W, Li Y, Wagner U, Wang W, et al. (2013) RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. PloS Comput Biol 9(3): e1002968. Doi:10.1371/journal.pcbi.1002968

2. Minjia Tan, Hao Luo, Sangkyu Lee, Fulai Jin, Jeong Soo Yang, Emilie Montellier, Thierry Buchou, Zhongyi Cheng, Sophie Rousseaux, Nisha Rajagopal et al. Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification. *Cell* 146(2011), pp1016-1028

3. Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal et al. Epigenomic Analysis of multi-lineage Differentiation of Human Embryonic Stem Cells. *Cell* 09(2013)

# FIELDS OF STUDY

Major Field: Bioinformatics and Systems Biology

# ABSTRACT OF DISSERTATION

Navigating the Human Epigenome through Random Forests

by

Nisha Rajagopal

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2013

Professor Bing Ren, Chair

Professor Wei Wang, Co-Chair

With the recent identification of over 100 histone modifications in mammalian cell-types, there is an urgent need to discover the minimal set of modifications that can completely characterize a genomic element. Of particular interest are transcriptional enhancers that play critical roles in cell-type specific gene expression but are difficult to characterize because they often act in a distal manner to the gene they regulate. We developed a Random-Forest based algorithm, RFECS (Random Forest based Enhancer identification from Chromatin State) for genome-wide prediction of enhancers which allowed us to identify the most informative and robust set of three chromatin marks for enhancer prediction. In addition, RFECS was seen to have improved accuracy of prediction over previous methods.

Applying this method to other genomic elements, we identified the minimal set of histone modifications required for prediction of promoters and gene bodies. Further, we elucidated the distinctive localization of histone lysine acetylations at enhancers, promoters and gene bodies, and obtained novel insights into the association of chromatin modification patterns with splicing. Using our algorithm, we predicted enhancers and promoters in 26 human primary tissues and 6 cell-lines, including 5 early developmental lineages. This lead us to the discovery of a novel class of cis-regulatory elements that can behave as enhancers in one cell-type and promoters in another. Further, we were able to associate the evolutionary conservation of regulatory sequences with properties such as tissue-specificity.

RFECS is a powerful algorithm with two-fold advantage. First, we can identify the most informative set of modifications characterizing or distinguishing particular genomic elements, thus enabling an insight into the biological mechanism of function at these regions. Second, we can make accurate predictions of enhancers and promoters in a genome-wide fashion, enabling the comparison of regulatory mechanisms across various human tissues or cellular conditions. Variations of histone modification patterns at the predicted tissue-specific *cis*-regulatory elements may substantially influence gene expression, which could potentially explain the distinct phenotypes of genotypically identical tissues.

# Chapter1.Introduction

For many centuries now, people have debated the relative importance of the innate qualities of the individual versus the impact of personal experiences. With the sequencing of the human genome and discovery of many genes associated with several traits, the nature versus nurture debate has been brought to the limelight once again. While the genetic encoding in the DNA comprises of the “nature” side of the debate, the field of epigenetics can be said to take the side of “nurture”.

Epigenetics has been defined as any heritable change in gene function that cannot be explained by changes in DNA sequence alone[1]. While the mechanism and extent of heritability is still under study[2,3], it is well-accepted that the epigenetic state of a cell can be dynamically regulated under the impact of environmental stimuli[4,5]. DNA is not found naked in the cell but is wrapped around histone octamers comprising of duplicate copies of 4 different types of histones, H2A, H2B, H3 and H4[6]. One of the major kinds of epigenetic changes in the mammalian cell are post-translational modifications of the tails of these various histones[7]. It is these patterns of histone modification patterns that form the basis of my thesis-work.

As of now, over 130 histone post-translational modifications(PTMs) have been discovered using a modified mass-spectrometry based approach[8]. These include methylation, acetylation, propionylation, butyrylation, formylation, phosphorylation, ubiquitylation, sumoylation, citrullination, proline isomerization, and ADP ribosylation at over 60 amino acid residues in histones[8]. The “histone code” hypothesis has been proposed stating that combinations of histone modifications may dictate function rather than single histone modifications[9]. Such a histone code may be implicated in the regulation of multiple cellular functions such as replication, recombination, transcription etc. Histone PTMs are thought to contribute to the regulation of such chromatin-templated processes in 2 main ways[10,11]. First the histone PTMs may change the net charge of histone molecules or alter inter-nucleosomal interactions, thereby regulating chromatin structure and 2lternative2y to other transcriptional factors[10,11]. Second, the combination of histone modifications maybe read by a set of “chromatin readers”, or PTM-specific binding proteins, which allow translation into a particular function[12,13].

In order to understand further the localization of histone modifications on a genome-wide scale, the ChIP-seq technology[7] has been used to map these modifications. The first large-scale study of this kind was carried out in CD4+ T-cells, which included 20 histone methylations and 18 histone acetylations[14]. Later efforts such as the Roadmap Epigenomics consortium and ENCODE have expanded the list of cell-types and tissues under investigation[15,16,17].

Unsupervised methods have been applied to understand all patterns of combinations occurring throughout the genome or at particular genomic elements[18,19,20]. Even with over 10 modifications, only a small number of chromatin states have been observed[21] emphasizing the redundancy in modifications and the tendency of certain modifications to co-occur. This has further lead to the question of reducing the set of modifications to a minimal non-redundant number that could fully characterize the genome. The observation that some genomic elements are characterized by particular combinations of histone modifications have been exploited to build predictive models for genome-wide discovery of genomic elements such as enhancers, promoters and gene bodies[22,23,24,25,26,27]. While promoters and gene bodies may be identified based on markers of gene activity such as CAGE[28] or RNA-seq[29] respectively, enhancers are particular difficult to identify since they maybe many kilobases away from the gene they regulate. The first chapter of my thesis involved developing a random-forest based algorithm for genome-wide prediction of enhancers from chromatin modifications or RFECS[23] that out-performed other existing enhancer prediction algorithms[25,26,27] in terms of accuracy. The ability of the algorithm to rank the different variables (histone modifications) by their importance in the prediction task, was used to find a minimal set of 3 modifications that could accurately predict genome-wide enhancers. In the second chapter of my thesis, I used this variable ranking feature of the RFECS algorithm to address the question of finding informative, non-redundant subsets of histone modifications, especially histone acetylations, associated with various genomic elements such as promoters, gene bodies, and splice site junctions.

Enhancers have been shown to be marked by highly cell-type specific histone modification patterns that strongly relate to cell-type specific gene expression[30]. As a consequence, understanding the changes in enhancer activity across multiple cell-types or conditions could lead to a better understanding of the differences between these cellular states. For instance, diseased and normal cell-types might vary considerably in their enhancer activity[31,32] and many disease-associated mutations are found to be located at such distal tissue-specific regulatory regions[33] potentially disrupting the enhancer activity in the diseased state. Changes in enhancer activity have been well-characterized during development in different organisms as well indicating their importance in lineage-specification[32,34]. The next step would be to understand the enhancer dynamics across multiple human tissues so as to understand better the determinants of cellular identity in the human body. The last chapter of my thesis focuses on predicting genome-wide enhancers and promoters across 6 cell-lines, at varying stages of development, as well as 26 primary human tissues. This enabled a comparison of the epigenetic changes at tissue-specific regulatory elements across early and late lineages, as well as a comparison across various tissue-types and germ layers leading to several novel insights into mammalian regulatory dynamics.

# Chapter 2.RFECS: A Random-Forest based algorithm for Enhancer Identification from Chromatin State

## Abstract

Transcriptional enhancers play critical roles in regulation of gene expression, but their identification in the eukaryotic genome has been challenging. Recently, it was shown that enhancers in the mammalian genome are associated with characteristic histone modification patterns, which have been increasingly exploited for enhancer identification. However, only a limited number of cell types or chromatin marks have previously been investigated for this purpose, leaving the question unanswered whether there exists an optimal set of histone modifications for enhancer prediction in different cell types. Here, we address this issue by exploring genome-wide profiles of 24 histone modifications in two distinct human cell types, embryonic stem cells and lung fibroblasts. We developed a Random-Forest based algorithm, RFECS (Random Forest based Enhancer identification from Chromatin States) to integrate histone modification profiles for identification of enhancers, and used it to identify enhancers in a number of cell-types. We show that RFECS not only leads to more accurate and precise prediction of enhancers than previous methods, but also helps identify the most informative and robust set of three chromatin marks for enhancer prediction.

## Introduction

Enhancers are distal regulatory elements with key roles in the regulation of gene expression. In higher eukaryotes, a diverse repertoire of transcription factors bind to enhancers to orchestrate critical cellular events including differentiation[35,36], maintenance of cell-identity[30,37] and response to stimuli[24,38,39]. While enhancers have long been recognized for their regulatory importance, the fact that they lack common sequence features and often reside far away from their target genes has made them difficult to identify. Computational techniques relying on transcription factor motif clustering or comparative analyses have had some success in identifying enhancers, but these predictions are neither comprehensive nor tissue-specific[40,41,42,43,44,45].

Recently, several high-throughput experimental approaches have been developed to identify enhancers in an unbiased, genome-wide manner. The first is mapping the binding sites of specific transcription factors by ChIP-seq[46]. Because this approach requires the knowledge of a subset of transcription factors (TFs) that are not only expressed but also occupy all active enhancer regions in the cell-type of interest, identification of all enhancers using this approach is not a trivial task. The second approach involves mapping the binding sites of transcriptional co-activators such as p300 and CBP[24,37,47], which are recruited by sequence-specific transcription factors to a large number of enhancers[38,48,49]. Since not all enhancers are marked by a given set of co-activators[50,51], and ChIP-grade antibodies against these proteins may not always be available, systematic identification of enhancers by mapping the locations of co-activators is not generally feasible. A third approach relies on identifying open chromatin with techniques such as DNase I hypersensitivity mapping[52]. However, since open chromatin regions can correspond to not only enhancers, but also silencers/repressors, insulators, promoters[53,54] or other functionally unknown sequences occupied by nuclear proteins, this approach lacks specificity in enhancer identification. Finally, a fourth approach interrogates covalent modifications of histones[21,24,25,26,27] as it was observed that certain histone modifications form a consistent signature of enhancers. It is on this approach that the present work is focused.

Previously, we and others observed that distinct chromatin modification patterns were associated with transcriptional enhancers[24,54,55]. Specifically, active promoters are marked by trimethylation of Lys4 of histone H3 (H3K4me3), whereas enhancers are marked by monomethylation, but not trimethylation, of H3K4 (H3K4me1). This chromatin signature has been used to develop a profile-based method for enhancer discovery[24]. Both unsupervised[21,56] and supervised learning approaches have also been employed to exploit chromatin modification-based differences to identify enhancers. The supervised machine learning techniques include HMM[25,40], neural networks[26] and genetic algorithm-optimized SVM[27] based approaches, and have proved to be improvements over the profile-based method. While these methods have led to identification of a great number of enhancers in the human and mouse genomes

[21,30,57], the current computational techniques have thus far been limited by the small number of the training set samples and limited number of chromatin modifications examined. Thus, it is possible that these approaches may not fully capture the entire range of chromatin modification patterns at enhancer elements. With the discovery of ever more histone modifications, it is likely that additional chromatin modifications may distinguish enhancers from other functional elements in the genome. This additional data should in principle allow us to answer the key question: what is the optimal set of modifications required for enhancer prediction?

Some researchers have tried to tackle this issue by using algorithms such as simulated annealing[25] or genetic-algorithm optimization[27]. We sought to develop a method in which the selection of the optimal set is automatically built into the training-process and is easily adapted to a large number of features.

As part of the NIH Epigenome Roadmap project, we have generated genome-wide profiles for 24 chromatin modifications and DNase-I hypersensitivity sites in 2 distinct cell types- human embryonic stem cell (H1) and a primary lung fibroblast cell line (IMR90)[17]. Additionally, we have experimentally determined a large number of promoter-distal p300 binding sites in each cell type, providing a rich training set for development of accurate and robust enhancer prediction algorithms. We now describe a random-forest[58] based method for integrative analysis of diverse histone modifications to predict enhancers. We show that this new algorithm outperforms the existing methods and leads to the automatic discovery of an optimal set of chromatin modifications for enhancer predictions.

## Results

### Prediction of enhancers using random forest and multiple chromatin marks

Random forests have recently become a popular machine learning technique in biology[59] due to their ability to run efficiently on large datasets without over-fitting, and their inherently non-parametric structure. Since random forests use a single variable at a time, they can give an automatic measure of feature importance [60]. Hence, we developed an algorithm based on this random forest technique for the purpose of enhancer prediction. Conventional random forests utilize a single scalar value associated with each feature at each node of the tree. In order to train a random-forest for enhancer prediction we wanted to use histone modification profiles at p300 binding sites. Because the spatial organization of histone modifications along a linear chromosome can be as informative as their actual levels, they are better represented as vectors of binned reads. Inspired by recent modifications to the random-forest approach such as discriminant random forests[61] or oblique random forests[62] that utilize a linear classifier at each node, we developed a new vector-based random forest algorithm RFECS or Random Forest for Enhancer Identification using Chromatin States (see Methods).

Genome-wide distal p300 binding sites were found using ChIP-seq in H1 and IMR90 cell-lines. We selected p300 binding sites overlapping DNase-I hypersensitive sites and distal to annotated TSS as active p300 binding sites representative of enhancers. We found 5899 such p300 binding sites in H1 and 25109 such sites in IMR90, and observed several distinct and diverse chromatin states using an unsupervised clustering technique, ChromaSig(fig.1.1A,B). All clusters showed enrichment of H3K4me1 and depletion of H3K4me3 as previously observed[24]. However, different clusters were characterized by varying levels of histone acetylation, H3K4me2 or H3K27me3. Clusters with presence or absence of H3K36me3 may represent genic and intergenic enhancers respectively. In order to ensure we represented all these different chromatin states at active p300 binding sites, we selected a relatively large number of these sites(>5000) for training as compared to previous methods.

To train the forest, active and distal p300-binding sites(BS) were selected as representative of the enhancer class. As non-enhancer classes, we considered annotated transcription start sites (TSS) that overlap DNase-I, and random 100 bp bins that are distal to known p300 or TSS (see Methods). The confidence of each enhancer prediction is given by the percentage of trees that predict this site to be an enhancer. In general, a genomic region is predicted as an enhancer if it has a background cutoff greater than 0.5 (>50% trees vote in it's favor). At higher cutoffs, confidence of prediction is higher, but fewer enhancers are predicted.

We used Receiver Operating Characteristic (ROC) curves to determine optimal parameters for our classification algorithm[63]. In the case of enhancer predictions, we can only obtain an approximate measure of specificity since we can never be certain that the randomly selected elements of the non-p300 class are all true negatives. Hence, in addition to the ROC curves generated using 5-fold cross-validation, we also verified parameter selection by comparing the percentage of predicted enhancers at each cutoff that overlap markers of active enhancers (validation rate) or TSS (misclassification rate). The markers of active enhancers include distal DNase-I hypersensitivity sites (HS), p300 binding sites(excluding those used in training), occupancy by CBP or sequence-specific transcription factors known to act at embryonic stem cell enhancers such as NANOG, OCT4 and SOX2.

In the case of Random forests, the main parameter to be determined is the number of trees. Since the non-enhancer class is assumed to be several times enriched compared to the enhancer class in the genome, we select a greater number of non-p300 training sites as compared to p300 sites and this proportion is also adjusted using the above-described methods. Previous algorithms[25] as well as empirical observations showed a width of -1kb to +1kb around the p300 binding site as optimal but we further verified this selection by cross-validation in the H1 cell-type (fig.1.2A). The difference in cross-validation curves using a width of 0.5kb or 1kb is not obvious on the cross-validation curve while a width of 1.5kb clearly shows a sharp drop in the area under the ROC curve(fig.1.2A). When we further made enhancer predictions using all three widths(fig.1.2B,C), it can be seen that a width of 1kb on either side shows best validation and misclassification rates as compared to 0.5 or 1.5 kb widths.

## Enhancer predictions in H1 and IMR90 cells

To determine the optimal number of trees for the random-forest, we examined the area under the ROC curve in H1 and IMR90 and found both to be stable beyond 45 trees (fig.1.3A,B). This was further verified by the lack of change in the validation and misclassification curves upon increasing the number of trees (fig.1.4A-D). In the end, we selected 65 trees for training the random forest to obtain a sufficient number of cutoffs. This is also provided as a default parameter for training and prediction of our algorithm. The training-set ratio of p300 to non-p300 was set at 1:7 since the ROC curve did not appear to change much beyond this ratio.(fig.1.4E,F)

In order to estimate the accuracy of the enhancer prediction by RFECS, we applied this algorithm to chromatin profiles of 24 marks obtained in H1 and IMR90. We then calculated the validation rate as the percentage of predicted enhancers overlapping with DNase-I hypersensitivity sites and binding sites of p300 and a few sequence specific transcription factors known to function in each cell type(true positive markers). We also computed the misclassification rate as the percentage of predicted enhancers overlapping with known promoters. These overlaps were computed using a window of -2.5 to +2.5kb. Incase, both a true positive marker as well as promoter lay within this window, the criteria used to decide if the enhancer was ”validated” or “misclassified” is discussed in detail in the Methods section. In H1 cells, we obtained a total of 55382 predicted enhancers at the lowest voting cutoff of 0.5. Over 80% of these predicted enhancers overlap with distal DNase-I hypersensitive sites and the binding sites of p300,NANOG, OCT4 and SOX2. Upon randomly generating enhancer predictions in the H1 genome 100 times, we found the average validation rate to be 18.43% and the actual validation rate of 80% to be highly significant with a one-sided t-test p-value of 10^-256. Additionally, we found that 5% of them overlap with UCSC TSS, indicating a low misclassification rate of 5%(fig.2C,E, in red). A similar high level of validation rate and low misclassification rate were observed when RFECS was applied to IMR90 cells, where 83581 enhancers were predicted with a validation rate of 85%(average random validation rate=16.13%, pvalue=2X10^-279), and misclassification rate of 4% (fig.2D,F). Thus, RFECS appears to accurately predict putative enhancer sequences based on chromatin modification state of the genome.

We next tried to assess the linear resolution of RFECS predictions. We calculated the distance between the predicted enhancers and locations of enhancer markers such as DNase-I hypersensitive sites, or p300 binding sites in each cell type, and found that the majority of predicted enhancers are within 200bp of these sites(fig.1.5A,B). In H1, nearly 62% of enhancers lie within 200bp of an enhancer marker site (fig.1.5A), while in IMR90 this value is around 70%

(fig.1.5B). Thus, the majority of enhancer predictions also show a high distance resolution in terms of proximity to the validation marker.

We also confirmed that our enhancer predictions showed an activation of gene expression in the proximal TSS. In order to do this, we compared RNA-seq datasets(manuscript under revision, Cell) in H1 and IMR90 using 14Iter[64] to identify H1-specific and IMR90-specific TSS. Then we identified enhancer predictions specific to either H1 or IMR90 using a filter distance of 2.5kb.When we look at the average distribution of H1-specific enhancers they are clearly enriched in the vicinity of H1-specific TSS as compared to IMR90-specific TSS (fig.1.5C) and this enrichment is found to significant at distances upto atleast 500kb using a Wilcoxon test(p-vaue<10^-6). Similarly, in the case of IMR90-specific enhancers, we observe them to be more enriched in the proxiumity of IMR90-specific TSS as compared to H1-specific TSS(fig.1.5D,p-value<10^-23).

As further evidence that RFECS accurately predicts enhancers, chromatin modifications at the predicted enhancers showed presence of all chromatin states observed in the training sets comprised of a subset of distal p300 binding sites (fig.1.1). In H1, clusters 1,2 and 8 of enhancer predictions (fig.1.6) are similar to clusters 1-3 of the p300 binding sites (fig.1.1A), clusters 3-4 appear to correspond to cluster 5 of p300 BS, while clusters 5-6 look like cluster 4 of p300 BS. In IMR90, similar trends could be observed when comparing chromatin states at enhancer predictions (fig.1.7) to those of p300 binding sites (fig.1.1B). Further, it can be observed that clusters 3-6 of the enhancer predictions in

H1(fig.1.6) that have weaker acetylation and/or enrichment of H3K27me3 also tend to have lower voting percentage of trees.

In summary, we showed that RFECS accurately predicted enhancers in the two cell lines H1 and IMR90 using a set of 24 chromatin modifications. These enhancers showed high validation rates, low misclassification rates and sharp resolution.

### Random forest trained on one cell-type can accurately predict enhancers in other cell-types

To make enhancer predictions, our approach requires a construction of a random forest trained on promoter-distal p300 binding sites. It is time-consuming and expensive to create a new training set for enhancer prediction in each new cell type, so it is desirable to use a random forest developed in one cell type to predict enhancers in another. To evaluate the feasibility of such approach, we first trained a random-forest using chromatin modification profiles obtained in H1, and then applied it to the IMR90 cells. Compared to RFECS predictions using IMR90 chromatin profiles as training set, RFECS predictions using H1 training dataset reduces the validation rate by ~5-8% and increases the misclassification rate by ~2% (fig.1.3C,E black vs red). Similarly, we also developed a random forest using the IMR90 data as the training set and then applied it to H1. This led to an average reduction of 2-3% in validation rate (fig.1.3D,black vs red).

Therefore, RFECS trained using one cell type may be applied to a different cell type, albeit with slightly lower accuracy.

We sought to examine if this moderate decrease in performance was largely due to cell-type specific differences or was within the limits of technical or biological variability between replicates. To this end, we trained a random forest on one replicate of a cell-type, and made predictions on the other replicate of the same cell type. RFECS trained on IMR90 and then applied to the replicate 1 of the H1 profiles (blue dot vs asterisk) actually showed a higher validation rate and lower misclassification rate than RFECS trained using replicate 2 of H1 (fig.1.3C,E), while similar performance was observed with enhancer predictions on replicate 2 of H1 independent of whether the random-forest was trained on H1 replicate 1 or IMR90 (green dot vs asterisk). Similar trends were observed when comparing predictions made on individual replicates of IMR90 using either H1-training or training on the other replicate (fig.1.3D,F). In conclusion, predicting enhancers using the random forest built from a different cell type exhibits a modest decrease in performance compared to a same-cell training set. However, this decrease in performance is comparable to the decrease that can arise due to variability between two replicates of the same cell-type.

### Optimal set of chromatin marks required for enhancer prediction

With the increasing number of histone modifications being discovered and mapped, determination of the relative importance of each mark in defining genomic elements is important. An out-of-bag measure of variable importance is a natural by-product of random forest classification scheme[60] wherein the relative importance of each feature is assessed as the increase in classification error upon permutation of feature values across classes. In both H1 and IMR90, the variable importance was assessed for random forests trained on 5 cross-sections of data for each of the 2 sets of replicates individually as well as the set of averaged replicates. Upon ranking histone modifications by variable importance, it is apparent that H3K4me1 and H3K4me3 are the top 2 most robust modifications across replicates and cross-sectional samples in both cell types, followed by H3K4me2(fig.1.8A, B). This indicates that these 3 modifications maybe the most informative in the prediction of enhancers in any unknown cell type as well.

Beyond the top 3 modifications, there is variability among the cell types. In IMR90, the other modifications appear to contribute almost equally, while in H1 there is a much clearer difference in variable importance. These differences are supported by correlation analyses in H1 and IMR90 (fig.1.8C,D). In H1, several modifications are highly correlated, which could explain the larger differences in variable importance, as only a few variables maybe needed to form a non-redundant set. In IMR90, the correlations are lower and hence each of the modifications may contribute non-redundant information and thus contribute equally to the variable importance. Modifications that cluster together in both H1 and IMR90 (shown in the same non-black colors, fig.1.8C,D) suggest cell-type independent redundancy.

Having established the relative importance of each histone modification in predicting enhancers, we next examined the accuracy of predictions using different sets of modifications. Validation rates obtained by using the minimal set of H3K4me1-3 is within 2% of that for all 24 modifications in H1 (fig.1.9A). Furthermore, this minimal set performs considerably better than the more conventionally selected set of H3K4me1 and H3K4me3[24,30] and at times, H3K27ac[65,66] (fig.1.9A,B, in black and blue). The set of H3K4me1-2-3 is more comparable to H3K4me1-H3K4me3-H3K27ac in IMR90 but does have a slightly lower misclassification rate (fig.1.9D). In both cases the use of the minimal set of 3 modifications shows a much closer resemblance in performance to all 24 modifications than to the set of 2 marks H3K4me1 and H3K4me3(fig.1.9A-D).

It can also be observed that in conjunction with H3K4me1 and H3K4me3, using H3K4me2 picks up a larger proportion of enhancers with weaker acetylation enrichment as compared to H3K27ac (fig.1.6,fig.1.7), supporting our prediction of the minimal set.

We also made enhancer predictions using all possible combinations of 3 modifications in chromosome 1 for replicate 1 and replicate 2 of H1. The average validation rate for a fixed range of enhancers was compared across replicates and it can be seen the set corresponding to H3K4me1, H3K4me2 and H3K4me3, is the highest performing combination common to both replicates (fig.1.9E). We also found the performance of the combination of H3K27ac with H3K4me1 and H3K4me3 appears to be comparable in this case (3, fig.1.9E), validating the use of H3K27ac as a feature for enhancer prediction when H3K4me2 is not available.

Some of the worst performing combinations include H3K9me3 and H4K20me1 (4 and 5, fig.1.9E), which also show up as variables with least importance in fig.1.8A.

In many currently existing datasets, H3K27ac is the more commonly sequenced histone modification as compared to H3K4me2 due to it's perception as a marker of active enhancers. While using H3K4me2 may improve enhancer prediction in some cell-types, use of H3K27ac in addition to H3K4me1 and H3K4me3 marks does show considerable improvement over using just the top 2 marks H3K4me1 and H3K4me3 (fig.1.9A-D). Hence, for many of the currently existing datasets, we could use H3K4me1, H3K4me3 and H3K27ac as the features in our random-forest with satisfactory performance.

Overall, these comparisons indicate the suitability of selecting H3K4me1,H3K4me2 and H3K4me3 as three minimal chromatin marks for purposes of enhancer prediction. Additional chromatin modifications required for improving upon enhancer predictions may depend on cell-type specific characteristics, as indicated by the differences in variable importance between H1 and IMR90 (fig.1.8A,B).

### Comparison of RFECS with other enhancer prediction methods

We next asked if our enhancer prediction algorithm performed better than several other current techniques for enhancer prediction – CSIANN, ChromaGenSVM and Chromia [25,26,27,66]. In previous studies, CSIANN and

ChromaGenSVM were applied on the histone modification dataset in CD4 T-cells[26,27,66]. In order to make a comparison of performance of our method with previous approaches, we applied RFECS to the CD4+ T cell dataset as well and determined parameters using cross-validation (fig.1.10). Using H3K4me1, H3K4me3, and H3K27ac, CSIANN made 21832 predictions[66] and ChromaGenSVM method made 23574 predictions [27] . We made enhancer predictions using H3K4me1, H3K4me3 and H3K27ac with RFECS as well as Chromia[25]. Cutoffs were selected that yielded a similar number of enhancer predictions for both Chromia (21895) and RFECS (22947)(fig.1.11A), so as to make a fair comparison across methods.

To compare these different sets of enhancer predictions, we computed validation rates by comparing them to TSS-distal DNase-I hypersensitive sites, p300 binding sites, and CBP binding sites and misclassification rates by comparing to known UCSC TSS using a window of -2.5kb to +2.5kb as described in the methods. (fig.1.11A). The validation rate of RFECS predictions is around 70%, which is considerably higher than the other three methods (57% ChromaGenSVM, 51% CSIANN,60% Chromia). Further, the misclassification rates of RFECS is less than 7%, much lower than the 27%, 35% and 15% rates of ChromaGenSVM, CSIANN and Chromia, respectively. These results suggested that overall procedure for RFECS, including selection of training set as well as training and prediction using the vector-random-forest, performs better than currently available techniques for enhancer prediction.

In the above comparison, we selected our enhancer-representative training set as p300 peaks called using MACS[67] that were distal to known UCSC TSS and overlapped DNase-I locations while CSIANN and ChromaGenSVM used a training-set of p300 peaks called using SICER previously[68].We also wanted to compare the performance of the different algorithms on our own datasets using the same training-set to evaluate the performance of the random-forest based part of the algorithm. To achieve this, we ran the various enhancer prediction methods on H3K4me1, H3K4me2 and H3K4me3 datasets of H1, with help from the author of ChromaGenSVM[27] (fig.1.11B). We tried to make the pre-processing stages of the various algorithms as consistent as possible by merging several replicates of each histone modification files and input files into single bed files and randomly selecting a smaller subset of p300 peaks for training, since these were the requirements of the other algorithms such as CSIANN and ChromaGenSVM. Incase of CSIANN, the selection of background was hard-coded in the software but in all other cases we used our own background training set as well. In fig.1.11B, it can be observed that RFECS shows a maximum validation rate of around 82.8% as compared to 54.6%, 66.8%, 57.7% and 63.3% for ChromaGenSVM (with default background selection), ChromaGenSVM (background selected by RFECS), CSIANN and chromia respectively. Further, RFECS showed the lowest misclassification rate of 4.9% as compared to 5.3%, 8.3%, 36.7% and 10.1% rates for the above-mentioned cases. It is worthy to note that using our background training set for ChromaGenSVM considerably improves performance of the algorithm, indicating that the improvement of performance is a combination of various stages of the RFECS procedure including selection of training-set. In summary, RFECS show considerably improved performance over existing enhancer-prediction algorithms.

## Prediction of enhancers in multiple human cell-types

Comparing enhancer predictions across diverse cell-types can contribute to understanding differences in regulatory mechanisms between cell-types. The ENCODE dataset is an example of a collection of high-throughput datasets such as histone modifications and transcription factor binding data that are available for multiple cell-types[16]. Having a set of high-confidence enhancer predictions in these cell-types would be a valuable resource.

We trained our random forest on the p300 ENCODE data in H1 and made enhancer predictions in 12 ENCODE cell-types using the three marks H3K4me1, H3K4me3 and H3K27ac since these were available for all the cell-types. Validation rates were assessed based on overlap with existing DNAse-I hypersensitivity data while misclassification rates were calculated based on overlap with UCSC TSS. It can be seen that the majority of cell-types show high validation rates between 80 and 95%, while the misclassification rates lie within acceptable levels of 2-7% (fig.1.12A,B).

In order to compare enhancers across cell-types, it is preferable to have enhancer predictions with the same level of confidence. To determine the appropriate cutoff for multiple number of cell-types, we calculate a False Discovery rate by randomly permuting 100 bp bins across the genome and computing the ratio of enhancers predicted in permuted data/enhancers predicted in real data for various cutoffs of voting percentages. In fig.1.12C, it can be seen that different cell-types show a different relationship with FDR. For example, at an FDR of 5%, the voting percentage for GM12878(solid dark blue ) is 0.74, for Nhek(dashed cyan) 0.64 and for Hsmm(solid yellow) it is 0.56.

Using an FDR of 5%, we obtained a consistent set of high-confidence enhancer predictions in the 12 ENCODE cell-types. In fig.1.12D, the numbers of enhancer predictions in each cell type is shown above the bar. The validation rates (in red) are above 90% for all cell-types except H1, Hepg2 and GM12878. In H1 and Hepg2, the numbers of DNase-I hypersensitivity sites are relatively less, i.e. ~150 to 177K as compared to ~230 to 380K in the other cell-lines. This may explain the somewhat lower validation rate in these two cell-types. GM12878 appears to be an outlier and we suspect that enhancer predictions may potentially be improved in this cell line by using a different training set.

In summary, we obtained a high-confidence set of enhancer predictions in multiple ENCODE cell-lines with the same level of confidence. This will enable more rigorous comparisons of regulatory characteristics of these cell-types in the future.

## Discussion

We describe here a novel machine-learning algorithm to accurately predict enhancers in a genome-wide manner based on chromatin modifications. We trained this algorithm using novel p300 training sets in H1 and IMR90 and 24 chromatin modifications in each cell-type. We showed that models trained on one cell-type could be effectively applied on another cell-type. Random forests enable detection of the most informative features required for a classification task. In the case of enhancer prediction, we identified a set of 3 histone modifications that appeared to be the most informative and robust across cell-types and replicates. Such an approach can once again be applied when the number of genome-wide modification maps is expanded in various different cell types and the most informative set of modifications can be further refined. We show that RFECS outperforms other machine-learning based prediction tools in CD4+ T cells, and can be applied in the future to multiple cell types. We successfully applied our enhancer prediction tool to 12 cell-lines in the publicly available ENCODE database and obtained a set of enhancers with a consistently high level of confidence across the cell-types.

In the future, we could potentially adapt the RFECS method to detect other regulatory genomic elements that can be observed to have a distinct chromatin signature and find the minimal set of chromatin marks for this purpose. The ability to detect diverse patterns of features within the training set indicates that the RFECS approach could be used to train on a composite training set comprised of different transcription factors. Combining information from different enhancer-binding proteins may improve prediction of regulatory elements. Random forests are non-parametric hence they can integrate a large number of diverse features. This could suggest the addition of other discrete and continuous data types such as sequence or motif based information or DNA methylation to the prediction of genomic elements.

## Methods

### Datasets used

The H1 and IMR90 datasets were generated as part of the NIH Roadmap Epigenome Project and have been released to the public prior to publication (http://www.genboree.org/epigenomeatlas/multiGridViewerPublic.rhtml). Briefly, 24 chromatin modifications in human embryonic stem cell (H1) and primary lung fibroblast cells (IMR90) were generated by the Ren lab and deposited under the NCBI Geo accession number GSE16256. Additionally, two replicates of H3K9me3 datasets deposited under Geo accession numbers GSM818057 and GSM42829 were used. Genome-wide binding data for p300 in H1 and IMR90, and transcription factors NANOG, SOX2 and OCT4 in H1 were generated in the Ren lab using ChIP-seq and deposited under accession numbers GSE37858, GSE18292 and GSE17917 respectively. Any data mapped to hg18 was converted to hg19 using liftover tools[69]. The DNase-I hypersensitivity datasets for H1 and IMR90 were produced by the Stammatoyanopoulos group at UW[70].

IMR90 DNase-I raw data may be accessed using GSM468792 and narrow peak calls are attached as supplemental information. Narrow DNase-I peaks in H1 were downloaded from UCSC ENCODE page (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).For CD4, previously generated datasets for p300[68], CBP[68] and DNase-I[53] data as well as histone modifications[14,71] were used. Histone modification data and DNase-I hypersensitivity data for the 12 ENCODE cell-lines was downloaded from http://genome.ucsc.edu/ENCODE/downloads.html.

### Data normalization for histone modifications

The ChIP-seq reads for the histone modification as well as corresponding input were binned into 100bp intervals. The binned modification file was normalized against the binned input file using an RPKM(Reads per kilobase per million) measure[72].In the case of 2 or more replicates, the RPKM- level for each bin is averaged to get a single histone modification file, in order to minimize batch-related differences.

### Determination of binding sites for p300 and other transcription factors

MACS[67] software was used to call peaks for p300,CBP and any other TF such as NANOG, SOX2 and OCT4. ChIP-seq input files were used as background and parameters of mfold=20 and default p-value cutoffs were used. Peak calls are available as supplemental files. In case of the p300 and CBP binding sites used to validate enhancer predictions in CD4, we included the regions of enrichment that were previously published as well[68].

**Construction of random forest**

We constructed the forest using the concept of binary classification trees, with each feature being a 20-dimensional vector of 100 bp bins from -1 to +1kb along the genomic element. At each node in the tree, a linear classifier was constructed using the Fischer Discriminant approach using the histone modification vector, allowing for utilization of shape as well as abundance information(fig.1.13A). The utilization of the linear discriminant at each node was inspired by the recent development of methods such as the discriminant random-forests[61] and oblique random forests[62]. The Vector-Random forest algorithm was implemented in MATLAB (MATLAB 7.14.0.739, The Mathworks Inc., Natick, MA, 2012a) as the function “multiclasstree” and utilizes functions from the “classregtree” and “classify” functions of MATLAB, implementing decision trees and linear discriminants respectively. The MATLAB code used for RFECS can be downloaded from: http://enhancer.ucsd.edu/renlab/RFECS_enhancer_prediction/

**Training the random forest for enhancer predictions**

Enhancer prediction involved two stages, which are classification of p300 vs non-p300 and peak-calling.

1) Classification of p300 vs non-p300 for enhancer prediction purposes

i. Training

In the first stage, a forest was constructed with two classes – a class containing p300 binding sites and a second class with an equal number of TSS and x times the number in random background sequences, where x=9 for CD4 and x=7 for H1 and IMR90.

ii. Prediction

In order to make predictions, each 100bp bin along a chromosome is assigned either enhancer or non-enhancer status. The output from the forest is in the form of percentage of trees predicting a 100bp bin to be one element or another. Only bins that have >50% trees voting for the enhancer class, are considered for further analysis.

2) Peak-calling

Using the random forest previously trained to predict whether a 100bp bin along a chromosome is an enhancer or not often yields values >50% for regions on either side of the exact location of a p300-binding site. However, the percentage of trees voting in favor of p300 decreases symmetrically on either side of the actual peak(fig.1.13.B).This property is used to select the bin with maximum voting percentage within a certain peak-filtering distance as the enhancer peak based on the assumption that the flanking regions are part of this same enhancer.

**Computation of variable importance**

A major advantage of the random forest is the inherent ability to select more important variables versus less important ones. In order to compute the order of variable importance, in this case, the importance of individual histone modifications for making enhancer predictions, we use an out-of-bag measure of variable importance[60] implemented in Matlab as the function oobVarImp.

### Application of variable importance to determine the minimal set of modifications required to predict enhancers

Based on the ordering of the variable importance across 5 different cross-sections of the training dataset of multiple replicates and cell types, certain modifications may always be observed to have priority. Due to the non-redundant nature of the ordering of variables as well as their robustness across replicates and samples, these modifications maybe selected as the most informative ones that are required to make enhancer predictions.

### Validation of enhancer predictions

Cross-validated ROC curves were used to estimate parameters for use within the same algorithm. However, comparisons across different algorithms may be biased depending upon the composition of the training set, so we validated enhancer predictions as described below.

Enhancer Predictions outputted from the random forest predictor have background enrichment scores of “voting percentage” ranging from 0.5 to 1 to enable detection of enhancers at different levels of confidence. At higher cutoffs, confidence of prediction is higher, but fewer enhancers are detected. The availability of large-scale datasets such as DNase-I hypersensitive sites, p300 binding sites, CBP binding sites and transcription factor binding sites enabled an estimate of the number of true positives at every cutoff. Further, the number of enhancers misclassified as TSS at each cutoff was also determined. Within the same cell type, an enhancer prediction method that performs better, should pick up more true positive validation markers and fewer TSS, given the number of predictions are the same.

Predicted enhancers are classified as “validated”, “misclassified” or “unknown” based on the criteria below. True Positive Markers (TPM) refer to DNase-I hypsersensitivity site, p300, CBP and Transcription factor binding sites.

1. If the nearest TPM lies within 2.5 kb of the enhancer and the nearest TSS is greater than 1 kb away from the TPM, the enhancer is “validated”
2. If a TSS lies within 2.5 kb of the enhancer, and the nearest TPM is either greater than 2.5 kb away from the enhancer or within 1kb of the TSS, the enhancer is “misclassified”
3. If there is no TPM or TSS within 2.5kb of the enhancer, it is “unknown”.

**Correlation graphs**

The Pearson correlation coefficient between any two modifications was computed for RPKM-normalized histone modification reads between -1 to +1 kb for all elements within the selected training set. The correlation patterns of each histone modification was used to cluster the modifications and order them using MATLAB tools.

This enabled visualization of which modifications are the most similar in their correlation patterns. In the ordering of variable importance, if certain variables showed up as important in two different cell types, the redundancy based on their correlation plots could be used to explain away this variability.

### Visualization of chromatin modification patterns

ChromaSig[19] was used to cluster histone modification patterns along p300 binding sites and predicted enhancers using modification width as 4kb.The resulting clusters were then visualized using Java TreeView[73].

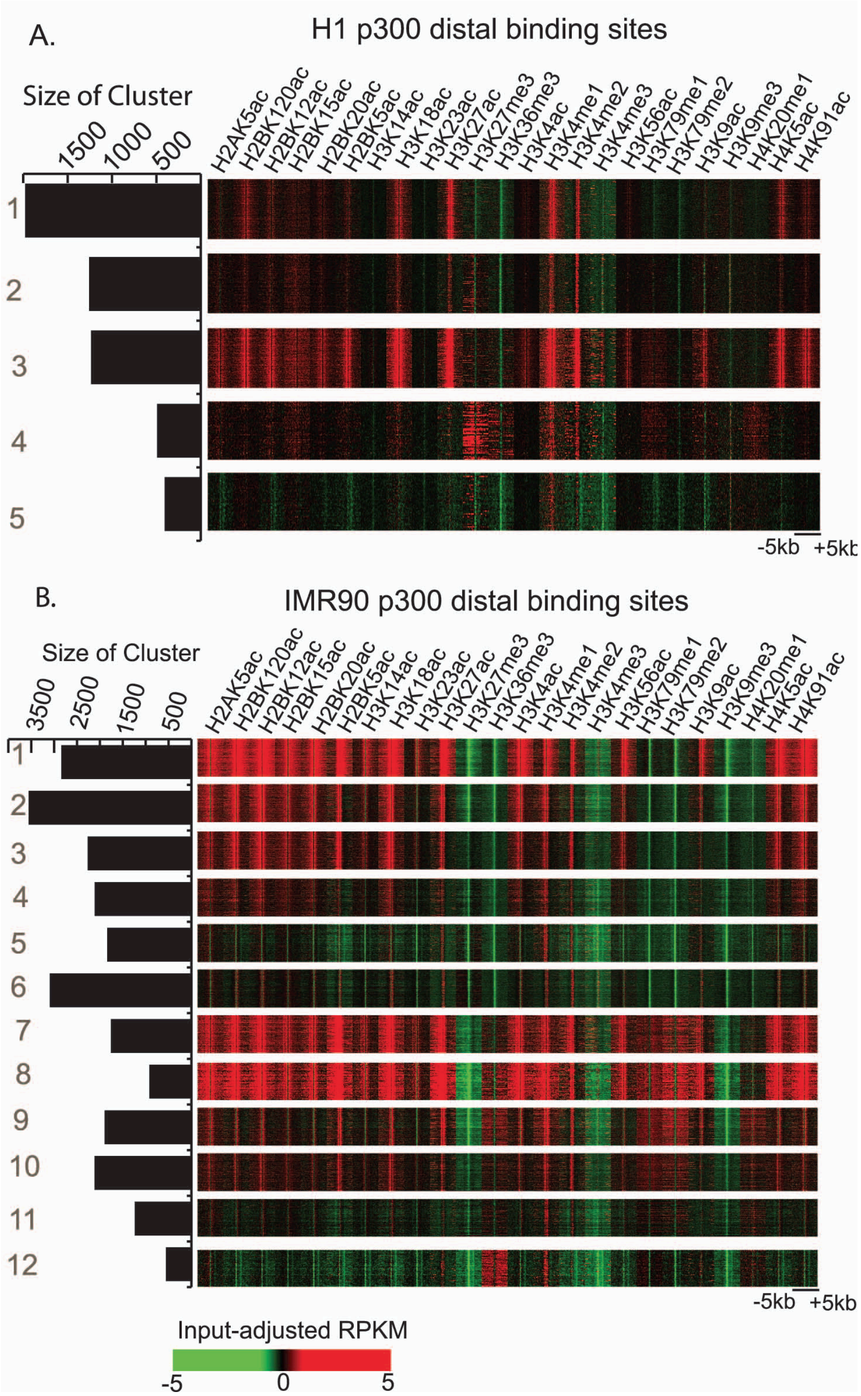

**Figure 2.1. Histone modification patterns at distal p300 binding sites in H1 and IMR90**
A.) Chromatin states for p300 binding sites in H1 cells. B.) Chromatin states for p300 binding sites identified in IMR90 cells , identified by clustering using ChromaSig[19]. The heatmap shows RPKM-normalized histone modification levels in 100bp bins from -5 to +5 kb along p300 binding sites overlapping DHS and distal to known TSS.

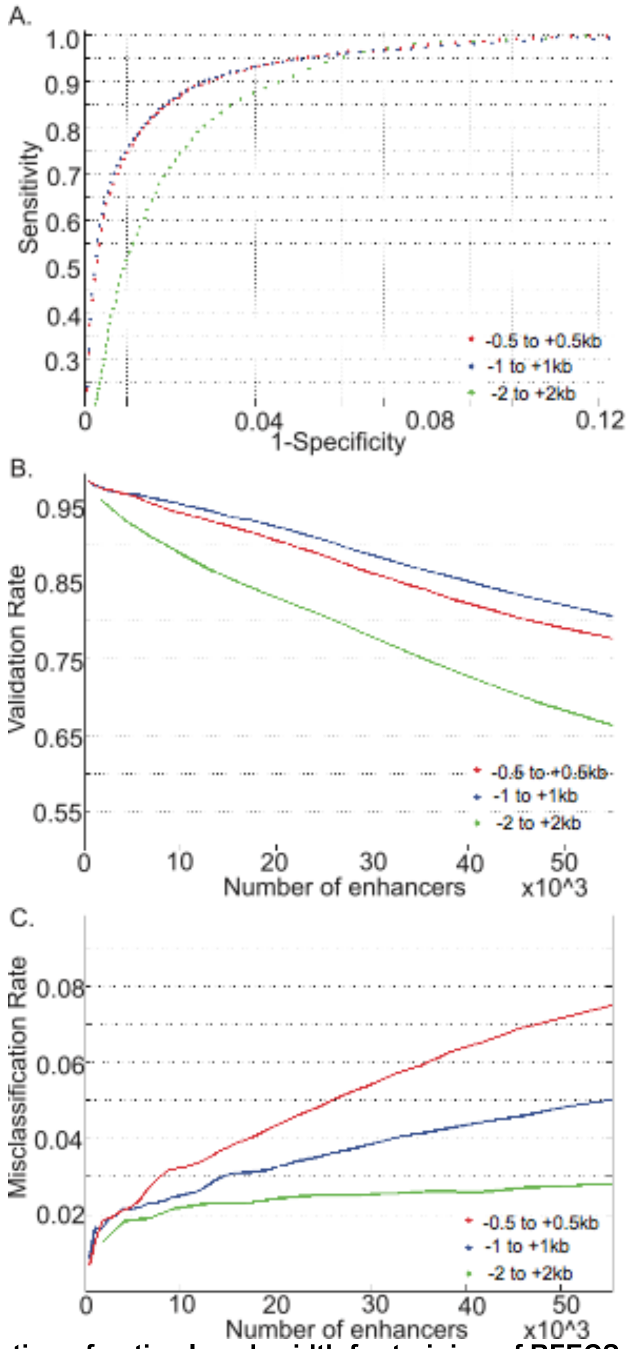

**Figure 2.2. Determination of optimal peak width for training of RFECS predictor in H1 cells** A.)ROC curves for 5-fold cross-validation at different proportions of peak widths of -0.5 to +0.5 kb, -1 to +1 kb and -1.5 to 1.5 kb around training set sites. B.)Percentage of enhancers validated by true positive markers at different numbers of enhancers determined by various cutoffs (Validation rate or VR curve). C.)Percentage of enhancers misclassified as TSS at different numbers of enhancers determined by various cutoffs. (Misclassification rate or MR curve). Overall, the width of 21 to +1 kb appears to show the best performance as expected based on previous observations.

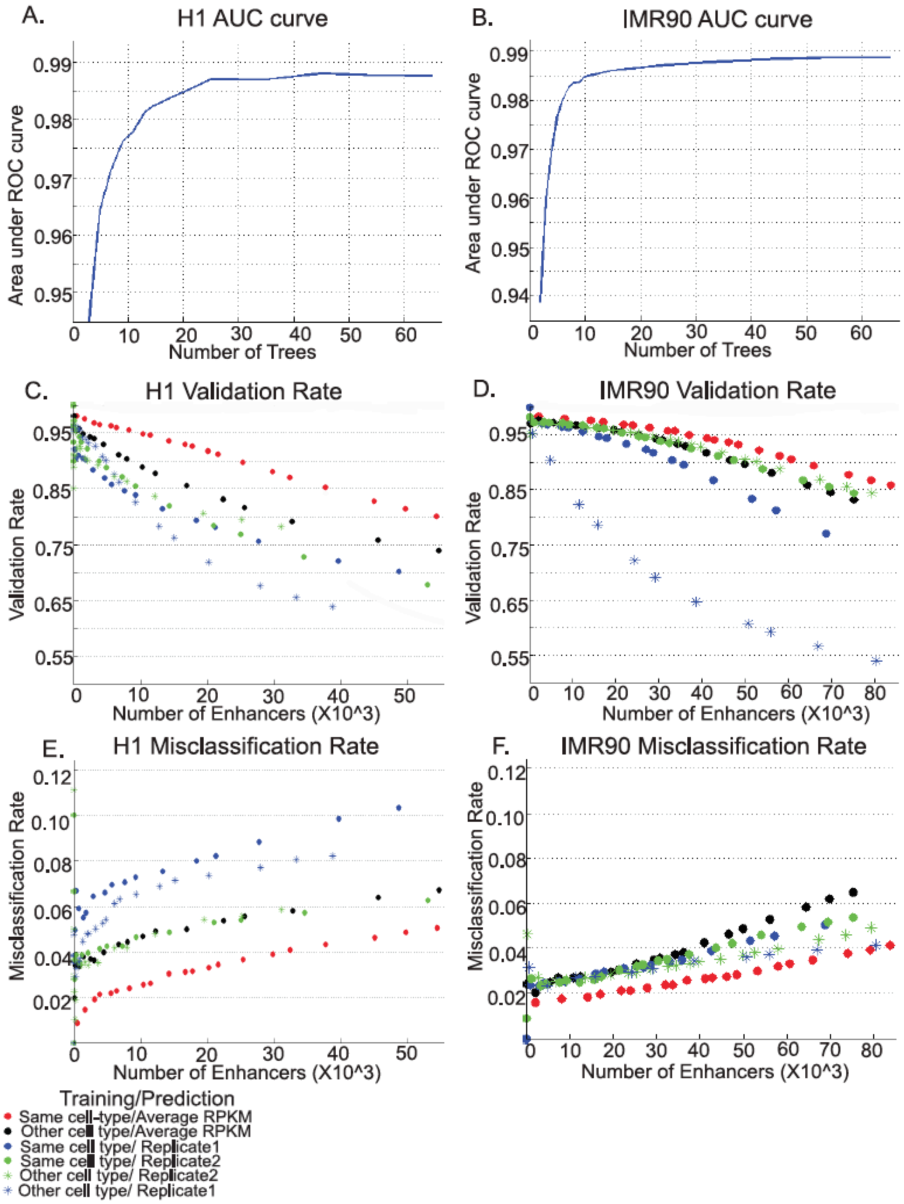

**Figure 2.3. Performance of RFECS for enhancer predictions in H1 and IMR90 cells**

Area under the 5-fold cross-validated ROC curve decreases with increase in number of trees stabilizing gradually in A.)H1 and B.)IMR90 cells. C,D.)Validation Rate of enhancer predicted in A.)H1 cells, as measured by overlap with DNase-I HS and binding sites of p300, NANOG, OCT4 and SOX2.B.) IMR90 as measured by overlap with DNase-I HS or p300 binding sites in the same cells. E,F.) Misclassification Rate of enhancers predicted using RFECS in E.)H1, and F.)IMR90 as measured by overlap with UCSC TSS, versus total number of enhancers determined by taking different voting percentage cutoffs, are shown for forest trained in the same cell type(·red), forest trained in other cell type and predictions made on modifications with averaged RPKM(·black), replicate 1 only(·blue), and replicate 2 only(·green). Training on one replicate and prediction on the other replicate of the same cell-type are indicated by asterisks.

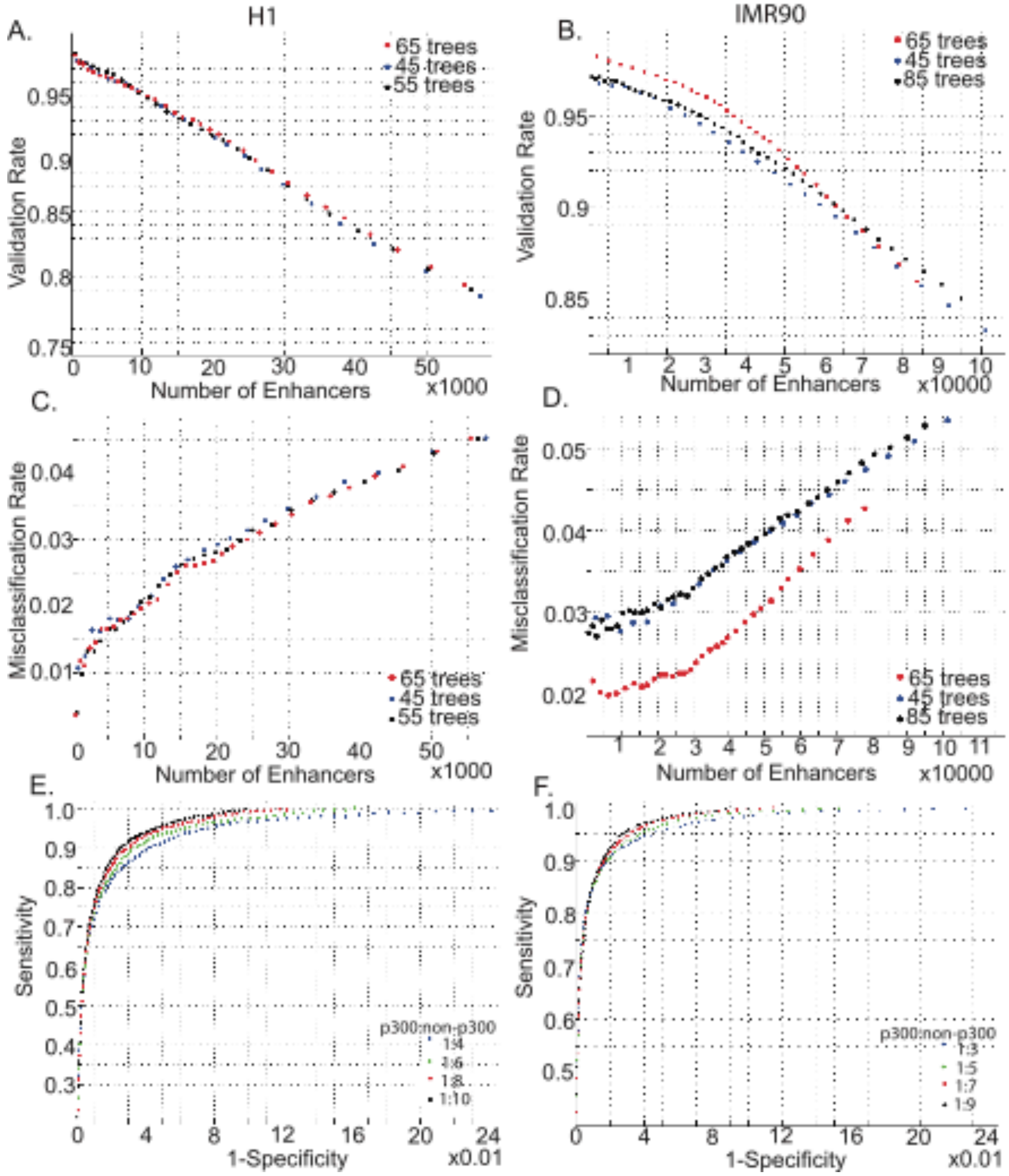

**Figure 2.4**. **Determination of parameters for training of RFECS predictor in H1 and IMR90**
A,B.)Percentage of enhancers validated by true positive markers at different numbers of enhancers determined by various cutoffs (Validation rate or VR curve) in A.)H1 and B.)IMR90, for different number of trees. C,D.)Percentage of enhancers misclassified as TSS at different numbers of enhancers determined by various cutoffs. (Misclassification rate or MR curve) in C.)H1 and D.) IMR90, for different number of trees. . VR and MR curves do not appear to change much beyond 45 trees, confirming the selection of 65 trees as valid. E,F.)ROC curves for 5-fold cross-validation at different proportions of training set ratios of p300:non-p300 in E.) H1 and F.)IMR90. ROC curves appear to be most stable beyond the ratio of 1:7.

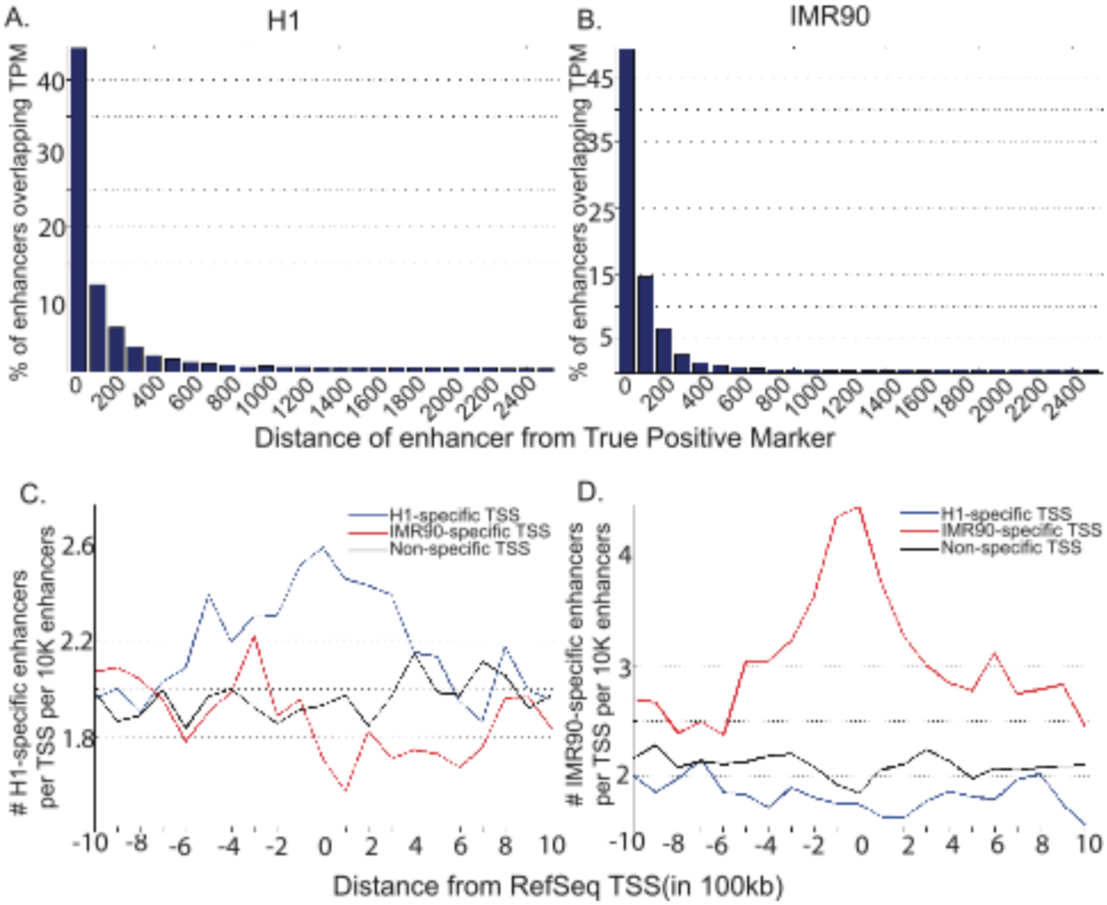

**Figure 2.5. Linear resolution and association with expression of genes for enhancer predictions in H1 and IMR90**

Distribution of distances between predicted enhancers and known markers of active enhancers such as DNase-I hypersensitivity sites, p300 and transcription factor binding sites in A.)H1 and B.)IMR90. Distribution of average number of celltype specific enhancers around the TSS specific to either H1 (blue), IMR90 (red) or non-specific (black) where the cell-type is C.)H1 or D.)IMR90.

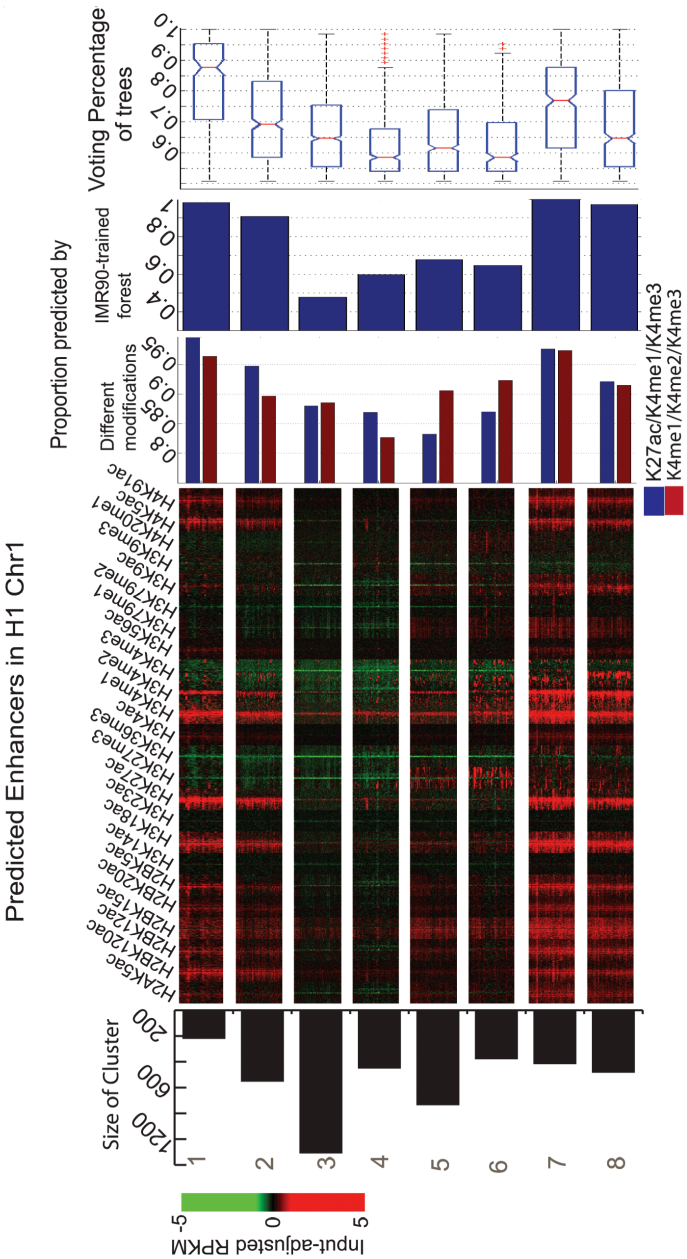

**Figure 2.6. Histone modification patterns at enhancer predictions in H1. Clustering was performed using ChromaSig**

Java treeview-generated Heatmap shows RPKM-normalized histone modification levels in 100 bp bins from -5 to +5 kb along genomic elements overlapping enhancers in Chromosome1 predicted using all 24 modifications. On the left panel, the state number and sizes are indicated. On the right panel, percentage of each state detected by different combinations of histone modifications or H1-trained forest are shown. Also shown are the distribution of background cutoffs associated with each chromatin state.

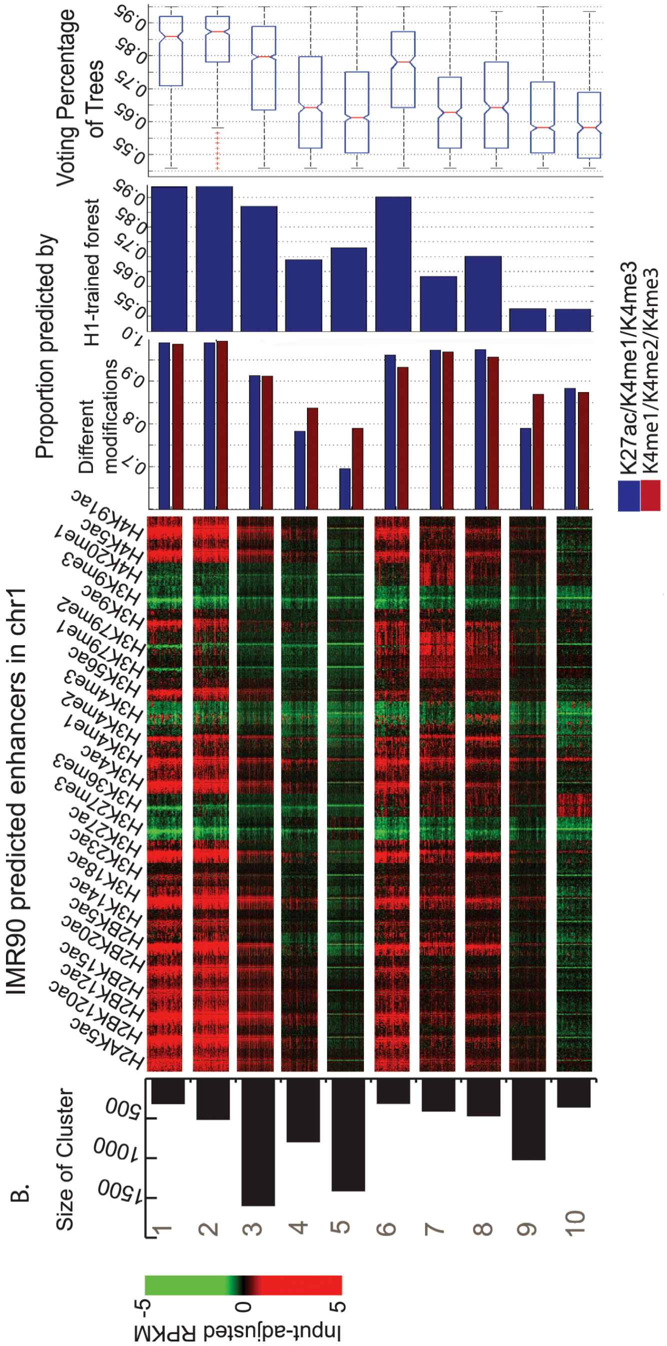

**Figure 2.7. Histone modification patterns at enhancer predictions in IMR90**

Clustering was performed using ChromaSig. Java treeview-generated Heatmap shows RPKM-normalized histone modification levels in 100 bp bins from -5 to+5 kb along genomic elements overlapping enhancers in Chromosome1 predicted using all 24 modifications. On the left panel, the state number and sizes are indicated. On the right panel, percentage of each state detected by different combinations of histone modifications or H1-trained forest are shown. Also shown,are the distribution of background cutoffs associated with each chromatin state.

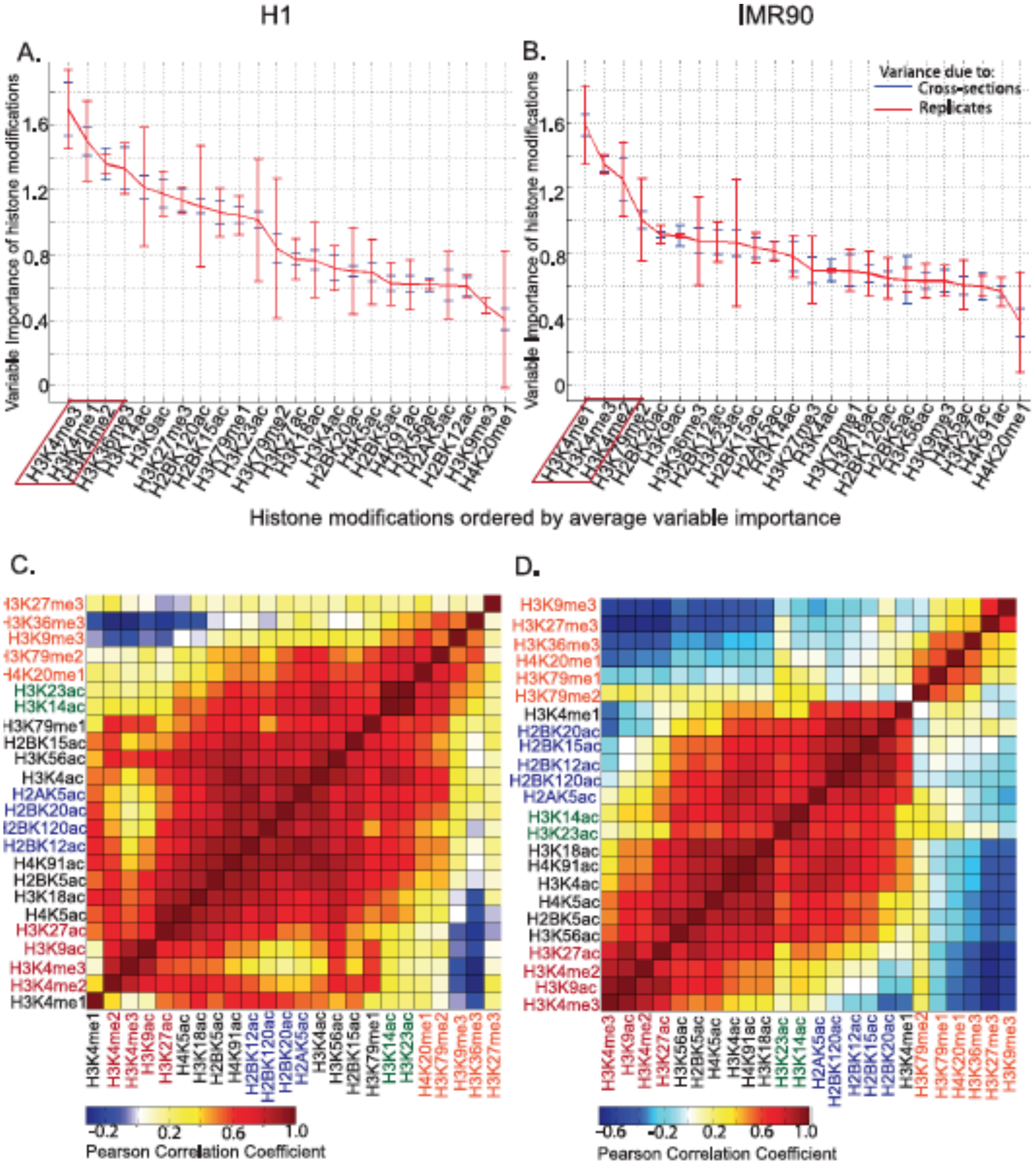

**Figure 2.8. Out-of-bag Variable importance of histone modifications in enhancer prediction** The average variable of histone modifications across 5 cross-sections of data in 2 sets of replicates as well as averaged replicates using all 24 modifications in A.) H1 and B.)IMR90 cells. Out-of-bag variable importance was calculated from the random-forest based classification of p300 binding sites against TSS+genomic background. Robust appearance of H3K4me1, H3K4me3 and H3K4me2 among the most important marks across replicates and cell types, indicates these may form a minimal set for prediction of enhancers. Differences observed in correlation clustering of the same 24 modifications in C.)H1 and D.)IMR90 explain some of the differences in ordering of variables in the two cell types. Same non-black colors of modifications indicate clusters that co-occur in both cell-types.

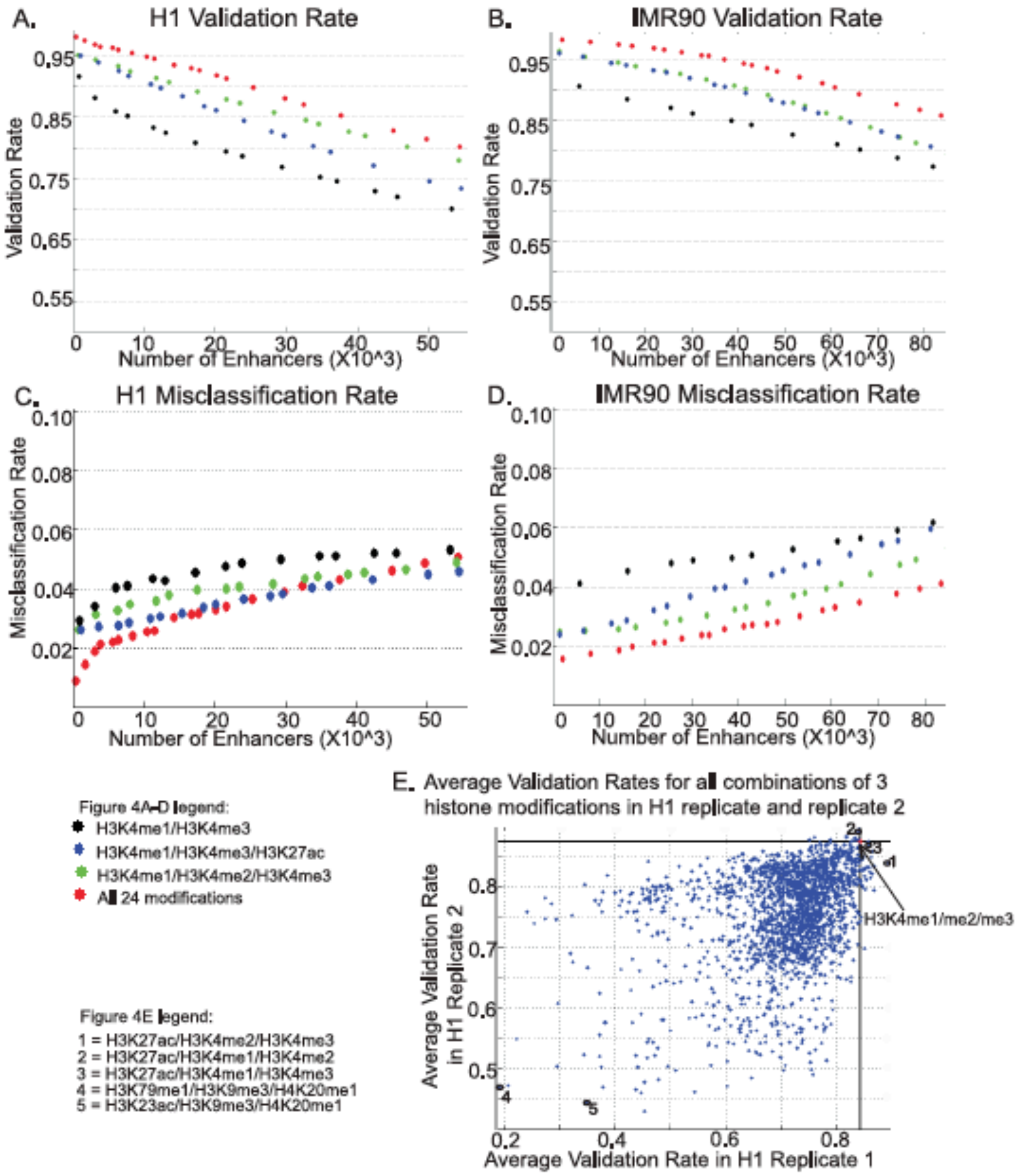

**Figure 2.9. Validation rate and Misclassification rate of enhancers predicted using RFECS in H1 and IMR90**

A.)Validation Rate in H1 measured by overlap with DNase-I HS, p300,NANOG,OCT4 or SOX2 , B.) Misclassification Rate in H1 measured as overlap of UCSC TSS, c.) Validation Rate in IMR90 measured by overlap with DNase-I HS or p300, d.) Misclassification Rate in IMR90 measured as overlap of UCSC TSS , versus total number of enhancers determined by taking different enrichment cutoffs, are shown for all 24 modifications(red), predicted minimal set of H3K4me1/ H3K4me2/ H3K4me3 (green)and conventionally used marks H3K4me1/ H3K4me3(black) or H3K4me1/ H3K4me3/H3K27ac(blue).

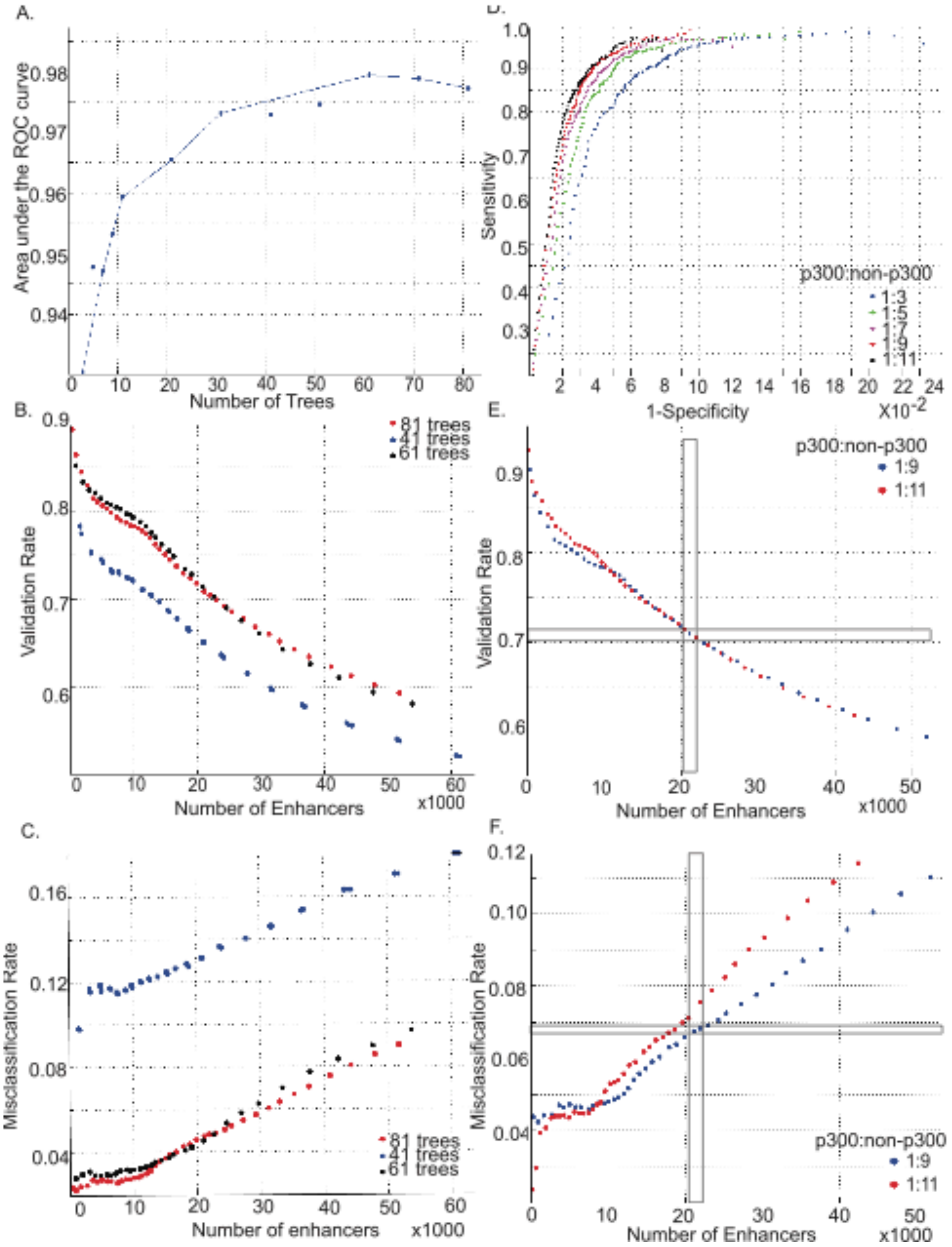

**Figure 2.10. Determination of parameters for training of RFECS predictor in CD4 T-cells**
A.) Area under the 5-fold cross-validated ROC curve decreases with increase in number of trees 41lternative gradually B.) Percentage of enhancers validated by true positive markers at different numbers of enhancers determined by various cutoffs (Validation rate or VR curve) and C.) Percentage of enhancers misclassified as TSS at different numbers of enhancers determined by various cutoffs. (Misclassification rate or MR curve), for 41, 61 and 81 trees. VR and MR curves do not appear to change much beyond 61 trees, confirming the selection of 81 trees as valid. D.) ROC curves for 5-fold cross-validation at different proportions of training set ratios of p300:non-p300. ROC curve does not appear to change much beyond a ratio of 1:9 E.) Validation Rate curve for training set ratios of 1:9 and 1:11. F.) Misclassification Rate curve for training set ratios of 1:9 and 1:11. The VR and MR curves validate the choice of 1:9 as an appropriate training set ratio.

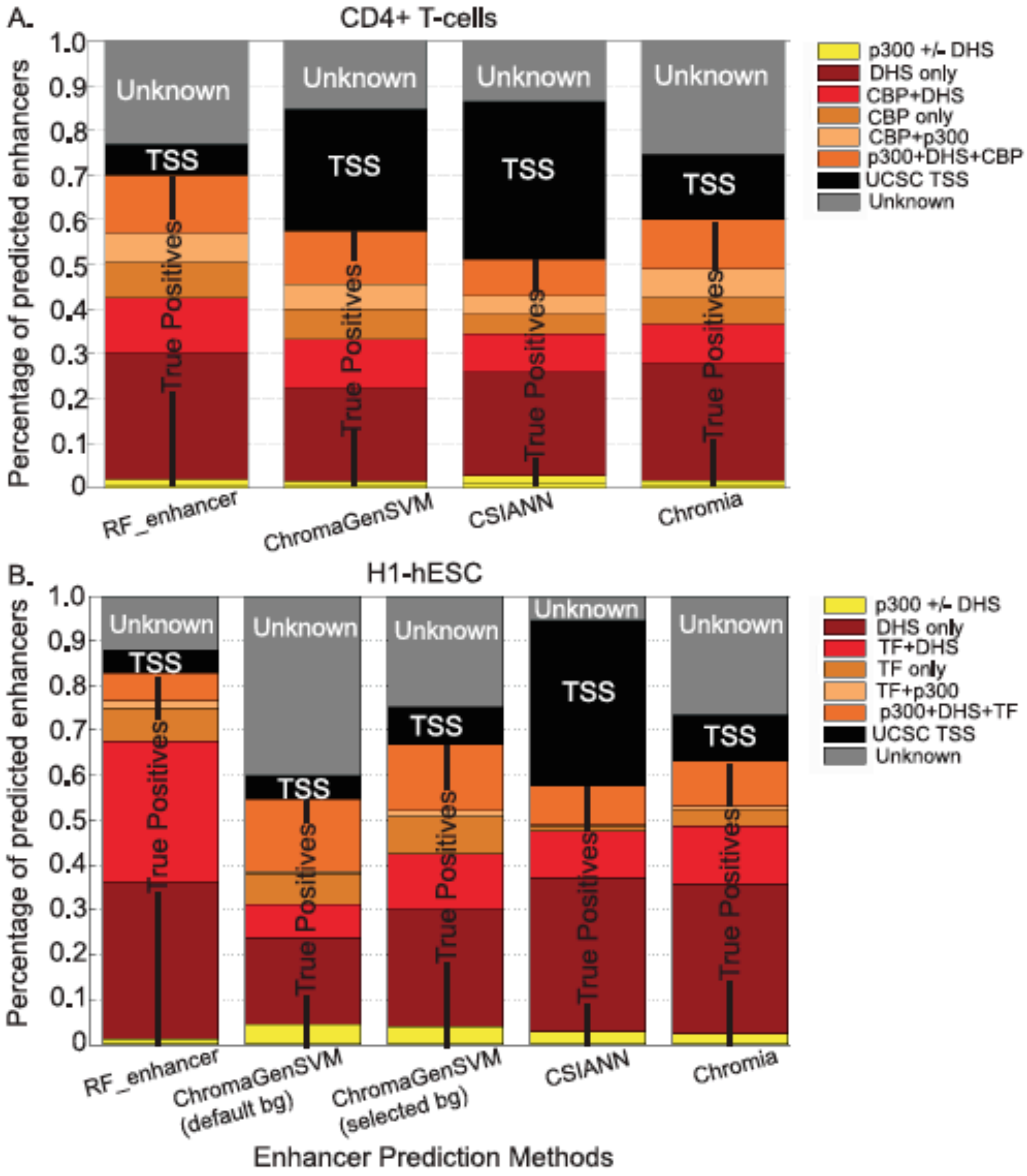

**Figure 2.11. Enhancer Predictions in CD4+ T-cells using RFECS, ChromaGenSVM, CSIANN and Chromia**

True positive rates were measured as overlap with either DNase-I hypersensitive sites(DHS), p300 or CBP binding sites, while false positives were measured as overlap with UCSC TSS.

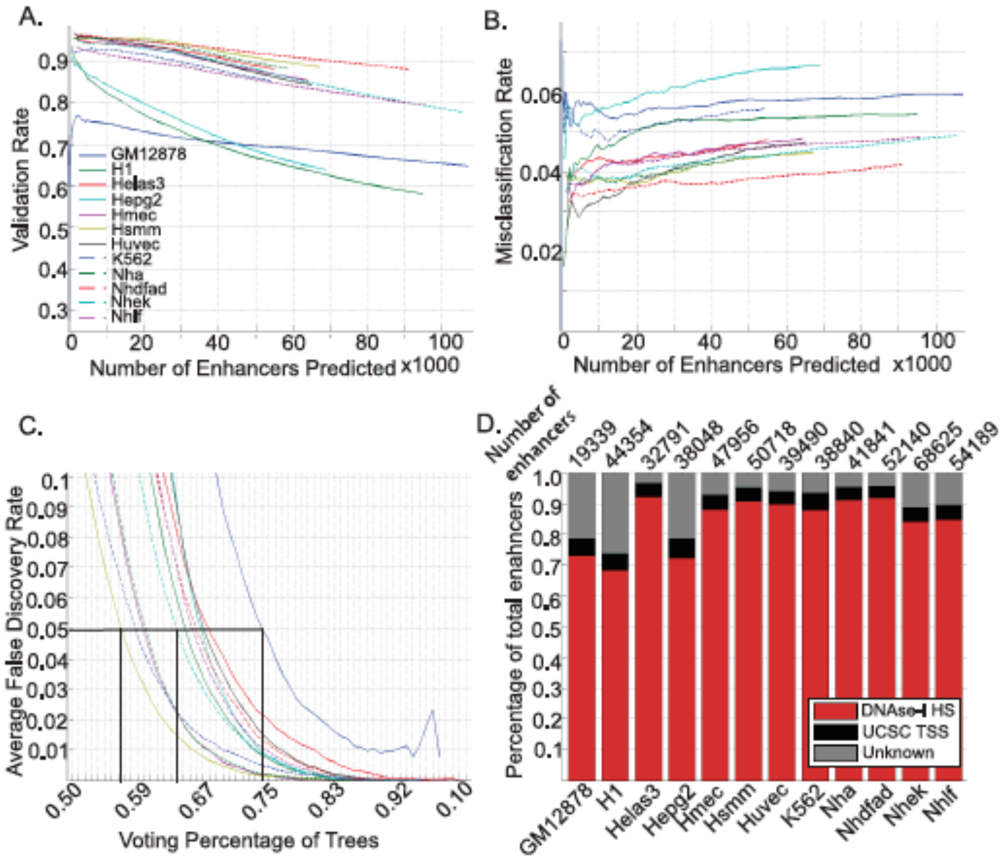

**Figure 2.12. Enhancer predictions in ENCODE cell-lines using RFECS**
A.) Validation Rate in the 12 cell-types measured by overlap with DNase-I HS, B.) Misclassification Rate in the  cell-types measured as overlap of UCSC TSS, C.) Average false discovery rate(FDR) over the 22 autosomal chromosomes for each cell-type plotted as a function of voting percentage of trees, D.) Validation rate and misclassification rate for each cell-type at a FDR of 5% with number of enhancer predictions shown above the bar.

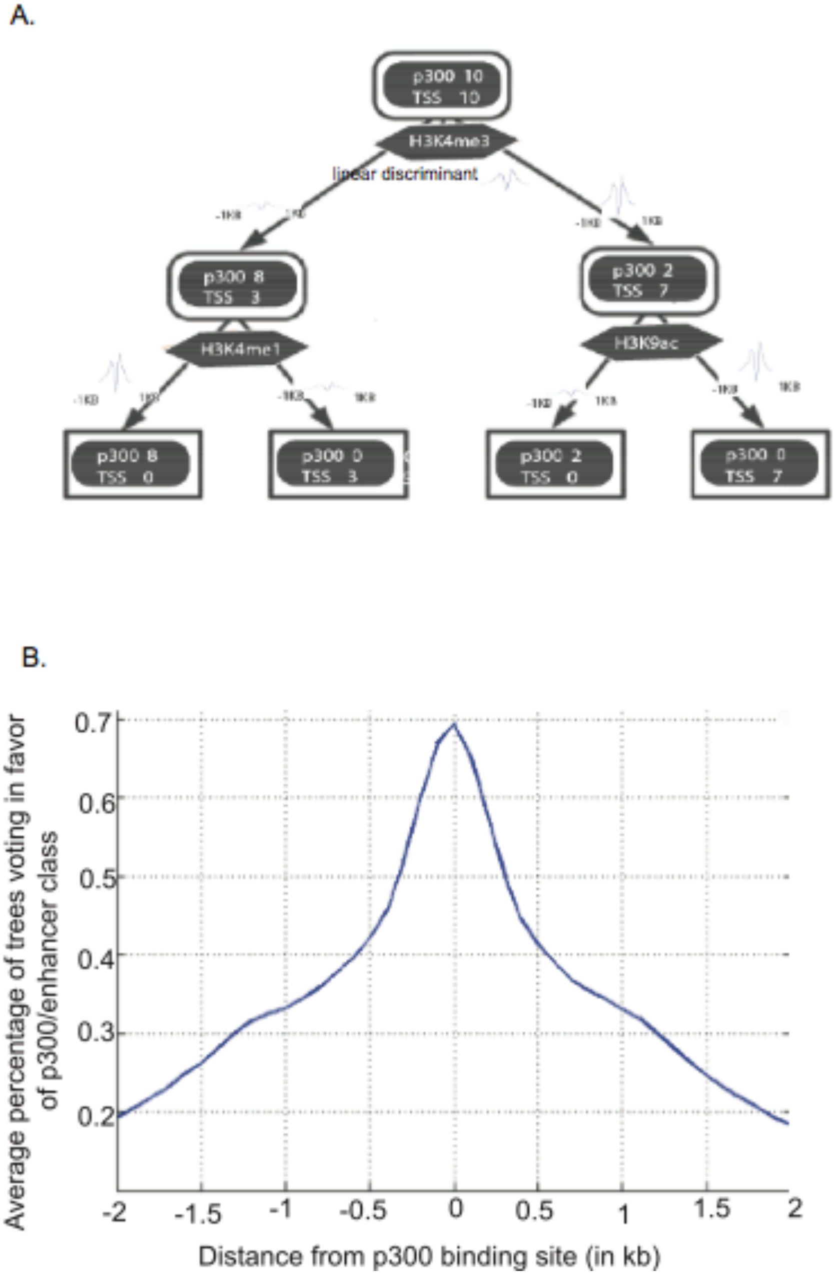

**Figure 2.13. Training of RFECS for enhancer prediction**
A.) Example of the vector-based random-forest classifying p300 binding sites and TSS using histone modifications. B.) Average percentage of trees voting in favor of the enhancer class around a p300-binding site. Percentage of trees in the random forest predictor that vote in favor of the enhancer class decrease symmetrically with increasing distance from the p300-binding peak. This property is used to develop a peak-calling method that can predict the most probable location of the enhancer.

ACKNOWLEDGEMENTS

Chapter 2, in full, has been published as Nisha Rajagopal, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, Bing Ren. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. Plos Computational biology 2013. The dissertation author was the primary investigator and author of this material.

# Chapter 3.Distinct and Predictive histone lysine acetylation patterns at promoters, enhancers and gene bodies

## Abstract

In eukaryotic cells, histone tail lysines are frequently acetylated. However, unlike modifications such as methylation, histone acetylation is often considered redundant. As such, the functional roles of distinct histone acetylates is largely unexplored. We previously developed an algorithm RFECS to discover the most informative modifications associated with the classification or prediction of genome-wide enhancers. Here, we use this tool to identify the modifications most predictive of promoters, enhancers, and gene bodies. Surprisingly, we find that histone acetylation alone performs well in distinguishing these unique genomic regions. Further, we find the association of characteristic acetylation patterns with genic regions and provide novel insights into the association of chromatin state with splicing. Taken together, our work underscores the diverse functional roles of histone acetylation in gene regulation, and provides several testable hypotheses to dissect these roles.

## Introduction

In eukaryotes, DNA is packaged into nucleosomes, each consisting of an octamer of proteins called histones that can undergo a variety of post-translational modifications[7]. Recent studies have shown that various genomic elements exhibit characteristic patterns of histone modifications, which have been exploited to predict the genomic location of cell-type specific regulatory elements. For instance, the combination of H3K4me1 and H3K4me3 is distinct at enhancers and promoters[24], which led to the development of several algorithms to predict them[23,25,26,27]. Similarly, H3K36me3, H3K79me1/2 and H4K20me1 are enriched within gene bodies[14] and the combination of H3K4me3 and H3K36me3 has frequently been used to discover novel genes[22].

In recent years, many more histone modifications have been identified using mass spectrometry based approaches[8] and mapped using ChIP-seq technology. Large-scale epigenomic maps such as those generated in CD4+ T-cells [14,71] showed combinatorial patterns of histone modifications at various genomic elements. Further, several sophisticated supervised and unsupervised machine-learning tools have been developed to annotate the epigenome that could provide a deeper understanding into the combinations of modifications that are characteristic of certain elements[18,19,23]. But given the large number of histone modifications, the question remains what the minimal set of modifications are to identify genomic elements such as enhancers, promoters or gene bodies.

Histone acetylations are largely considered markers of activity at regulatory elements such as promoters and enhancers, but due to their tendency to co-occur it has been difficult to elucidate the non-redundant roles of these acetylations[31]. Histone acetylations are indirectly targeted in the treatment of diseases such as Cancer and HIV by the use of HDAC(histone deacetylase) inhibitors[74]. Understanding the specific role of histone acetylations at different genomic elements has the potential to improve such therapies by increasing the specificity of targeting. Certain lines of evidence have suggested non-redundant roles of histone acetylation such as the fact that HDACs as well as histone acetyl-transferases (HATs) have unique genomic distributions [68,75]. Indeed, a previous study found certain acetylations such as H3K9ac to be present at promoters and H4K16ac along gene bodies [71]. However, the extent to which these acetylations are predictive of particular elements is still unknown.

While chromatin has been clearly associated with enhancers, promoters, and gene bodies, the discovery of co-transcriptional splicing, the finding that pre-mRNA can be spliced during the process of transcription itself, [76,77] suggested that chromatin could also be indicative of alternative splicing. Subsequently, it was found that exons are marked by elevated levels of H3K36me3[78,79]. Further supporting this notion, changes in acetylation levels could lead to changes in alternative splicing [80,81,82]. Here, we explore this subject on a genome-wide scale, describing the extent of association of chromatin with alternative splicing and also the comparison of such associations across two mammalian cell-lines.

In a previous study, we developed a random-forest based method of learning (RFECS) that could effectively identify genome-wide enhancers as well as the most informative set of modifications required for this task[23]. Here, we expand the application of this tool to determine the optimal set of discriminative histone acetylations for accurately predicting various genomic elements. Using this approach, we find distinctive patterns of acetylations that are associated with promoters, enhancers, gene bodies and splice junctions. Also, by comparing the epigenomes of two very distinct cell types (H1 human embryonic stem cells and IMR90 fetal lung fibroblasts), we expect our findings to generally hold for other cell types.

## Results

### Differential histone acetylation patterns at promoters and enhancers

We previously observed that H3K4me1 and H3K4me3 are the most distinctive marks between promoters and enhancers among a limited set of 6 histone modifications focused on 1% of the human genome [24]. To further define the marks that distinguish these two regulatory elements in genome-wide maps of 24 histone modifications[23], we compared active TSSs (TSSs overlapping DNase-I HS sites) with an equal number of enhancers defined by TSS-distal p300 binding. After z-score normalization(Methods), we observe that the mean histone modification profile of either class separates clearly into TSS-specific and enhancer-specific groups(Fig.2.1A, positive vs negative axes). We confirmed that the deviations of most of the histone modifications as compared to a set of elements with randomly shuffled labels is statistically significant(Fig.2.1A, p-value <10^-5 using Wilcoxon test , except for bars marked by black dots).In both H1 and IMR90,we consistently find that H3K4me1, H2BK20ac and H2BK120ac are significantly enhancer-specific while H3K4me3, H3K4me2, H3K9ac, H3K56ac, H4K5ac and H3K27ac are TSS-specific (Fig.2.1B). The histone modification profiles between -1 to +1kb along these elements are also observed to be different from the random set (Fig.2.2A,B, blue vs red).

To assess the importance of each modification in classifying promoters and enhancers, we constructed classifiers using each mark individually. H3K4me3, followed by H3K4me2 and H3K9ac, showed the highest classification accuracy in both H1 and IMR90 (Fig.2.1B, blue, red). Nearly all modifications showed a classification accuracy of atleast 55% (in H1) and 75% (in IMR90),which is above the classification accuracy of 50% expected at chance (we verified that classification accuracy upon randomly shuffling labels was found to be ~50%). Clearly, the most significantly TSS-specific modifications are H3K4me3, H3K4me2 and H3K4me1. For enhancers, H3K4me1 is the most distinctive, followed by H2BK20ac. In addition, we also observed cell-type specific contributions. To verify if the modifications specific to H1 are due to the distinct biology of stem cells, we repeated our analysis in H9 human embryonic stem cells, and observed trends resembling H1 (Fig.2.1B, green vs blue).

We next classified p300 and TSSs using all 24 marks. Interestingly, H3K4me3 alone achieved the same accuracy as all 24 modifications. Next, we examined whether histone acetylation alone could classify these two elements (Fig.2.1C). The difference in classification accuracy using all 15 acetylations is <1% different from using all 24 marks. Clearly, acetylations are quite distinctive between the enhancers and promoters.

To identify the specific histone acetylations contributing most to the accurate classification of promoters and enhancers, we computed the out-of-bag variable importance[23,60] for each acetylation. For both H1 and IMR90, the top acetylation was H3K9ac (Fig.2.1D, Fig.2.2C), achieving 85 and 89% classification accuracy, respectively (Fig.2.1C,D). The next mark in ordering of variable importance of H1 was H2BK120ac, while in the case of IMR90, several marks including H2BK20ac shared the same position (Fig.2.1D, Fig.2.2C). However, correlation clustering indicates that H2BK20ac and H2BK120ac are highly correlated in both H1 and IMR90 (Fig.2.1E,F), suggesting that these are redundant modifications. Hence we selected the top two marks as H3K9ac and H2BK120ac, and found that this combination achieved a classification accuracy of within 1% of using all 15 acetylations in IMR90, while in H1, this fell short by ~3%. Including the next mark in the ordering of H1, H3K14ac improved this accuracy by ~2%(Fig.2.1C).

In summary, we observed that H3K4me3 alone could achieve the same accuracy of classification as all 24 modifications in H1 (~94%) and in IMR90(~95%). Using all acetylations could accurately separate these two classes nearly as well as using all 24 modifications indicating differential enrichment of certain acetylations at enhancers and TSS. In particular, H3K9ac, H2BK120ac and H3K14ac appear to be most informative in combination, of which H2BK120ac is enhancer-specific while the other two are TSS-specific (Fig.2.1A).

### Histone Acetylation patterns accurately predict genome-wide enhancers and promoters

Our analysis suggests that histone acetylations are distinct at promoters and enhancers. Next, we wondered if these acetylations could predict promoters and enhancers genome-wide. As a first step, we extended the application of the RFECS methodology, previously used to predict enhancers[23], to the prediction of genome-wide promoters (Methods).

Using all 24 histone modifications, our approach can accurately predict promoters with ~92% true positive (TP) rate and ~1.6% false positive (FP) rate in H1, while in IMR90 we observed even better performance (TP ~95%, FP ~ 0.3%) (Fig.2.3A,B). Using the out-of-bag variable measure, we identified H3K4me3 as the most informative mark required to predict promoters, followed by H3K4me2 and H3K4me1 (Fig.2.4A, B). In terms of the area under the curve (AUC), this minimal set performs comparably with the set of all 24 modifications in both H1 and IMR90 ($AUC_{min}/AUC_{all}$=0.99,Fig.2.3A,red vs blue) .While in H1, this set is comparable to using just H3K4me3 (Fig.2.3A,black vs red), in IMR90, the addition of the two marks leads to an improvement of ~10% in TP rate as compared to H3K4me3 (Fig.2.3B,black vs red).

Next, to assess if acetylation can accurately predict promoters, we repeated our analysis on all 15 histone acetylations. In IMR90, overall performance was comparable to using all 24 modifications ($AUC_{ac}/AUC_{all}$ =0.99,Fig.2.3B, green vs blue) while in H1, the TP rate was the same for FP rates beyond 1.3% (Fig.2.3A, green vs blue). To determine which acetylations are the most informative and whether these are robust across cell-types, we computed out-of-bag variable importance for acetylations (Fig.2.3C,D). H3K9ac is clearly the most informative, while the next few marks that are comparable across the two cell-types appear to be H2BK120ac, H2AK5ac and H3K18ac. Several other H2BK-ac also occur among the top ranks in IMR90(Fig.2.3D), but are redundant with H2BK120ac (Fig.2.1E).

We then made predictions using just H3K9ac, the top 2 marks in variable importance for H1 and IMR90 and also the predicted minimal set of 4 acetylations. In H1, there is a significant difference in the ROC(Receiver operating characteristic) curve between H3K9ac and the top 2 marks and an equivalent increase upon including the next two marks, H2AK5ac and H3K18ac (~8% increase in TP rate for values of FP>1%, Fig.2.3E,black vs green vs red). Even though the performance is not as accurate as using all 15 acetylations, including more marks appears to contribute incrementally to the curves, such as using the top 6 marks (<2% change in TP for FP>1%, Fig.2.3E, magenta vs red). In IMR90, there is a significant improvement from using H3K9ac compared to the top 2 modifications, with difference in TP ranging between 5 to 20% at the same FP (Fig.2.3F, black vs cyan). Beyond this, improvements appear to be more incremental(<2%) such as in using the predicted minimal set of 4 modifications (Fig.2.3F,red dotted) or even upon including top 8 marks (Fig.2.3F, magenta).

Applying the RFECS algorithm (Rajagopal et al, 2013) to enhancers, we compared validation and misclassification rates of prediction using just acetylations to that using all 24 marks or the minimal set of H3K4me1, H3K4me2 (or H3K27ac) and H3K4me3 [5]. In H1, the validation rate using just acetylations appears to be comparable to the set of 3 marks, H3K4me1, H3K4me3 and H3K27ac (Fig.2.4C) while the misclassification rate appears to be within 1% of that using all 24 modifications(Fig.2.4E). In IMR90, the validation rate using just acetylations is within 3% of that using all 24 modifications (Fig.2.4D,green vs blue) and a misclassification rate that is within 1% using all 24 modifications (Fig.2.4F,green vs blue).

Hence, enhancers can also be accurately predicted using just acetylations. We computed variable importance for the prediction of genome-wide enhancers using acetylations and discovered H3K9ac, H2BK120/20ac and H3K14/23ac as the minimal set of acetylations for the prediction of enhancers(Fig.2.2A,B), which was further confirmed by comparisons of validation and misclassification rates with performance using all acetylations (Fig.2.2C-F).

In summary, we found acetylations alone to predict genome-wide enhancers as well as promoters quite accurately, indicating that acetylations are not only distinct between the two elements but also predictive. The most informative acetylations in the prediction of promoters were H3K9ac, H2BK120ac, H3K18ac and H2AK5ac while in the case of enhancers this set was composed of H3K9ac, H2BK120/20ac and H3K14/23ac.

### Minimal set of modifications to identify active genes

Several histone modifications have been identified as being enriched in the body of active genes[14]. However, it is still an unsolved problem what is the minimum number of modifications required to achieve an accurate prediction of the active gene body. To this end, we identified active refseq genes in the H1 and IMR90 genomes based on the overlap of their TSS with DNase-I HS sites and RNA-seq above log-value of 2 FPKM. Further, we only considered genic regions lying 2.5 kb away from an annotated TSS. As a true negative set, we identified intergenic regions as all those regions not lying within any annotated UCSC, GENCODE or Refseq TSS. We constructed a random-forest based classifier to distinguish these two sets using all 24 histone modifications and observed high sensitivity and specificity at the point of maximum accuracy in both H1 (sens = 89.56%, spec = 94.54%, AUC=0.97) and IMR90 (sens = 96.34%, 1-spec = 97.09%, AUC=0.99) (Fig.2.5A,B).

In both H1 and IMR90, the top 2 informative marks are H3K36me3 and H3K79me1, which rank well above all other marks (Fig.2.5C,D). By area-under-curve (AUC) analysis, the performance of these two marks alone is equivalent to that of all 24 marks in IMR90 ( $AUC_{K36me3,K79me1}$ / $AUC_{all}$=100%) although it seemed somewhat lower in H1 ($AUC_{K36me3,K79me1}$ / $AUC_{all}$ = 96%) (Fig.2.5A-B, green). We found that the 2 marks ranked next that were common to both cell-types were H3K27me3 and H3K9me3(Fig.2.5C,D). These modifications maybe important because of their relative depletion in genic regions and enrichment in larger intergenic regions (Fig.2.6D). By including these marks, our classifier achieved almost the same accuracy as all 24 marks in H1 (H1: $AUC_{top\ 4}$/$AUC_{all}$ = 99%) (Fig.2.5A, magenta vs blue). Thus, we conclude that the minimal set of modifications required to predict genes, within 1% accuracy of the set of all modifications, is between 2 to 4, with H3K36me3 and H3K79me1 being the most informative modifications.

**Acetylations at the gene body**

Next, to assess if gene body acetylation can distinguish genic from non-genic regions, we constructed a supervised classifier using only histone acetylations. Supporting this notion, acetylations show an ROC curve that is well above the line of no discrimination in both H1 and IMR90 (Fig.2.5A,B). However, the performance of acetylations is lower (H1:$AUC_{ac}$/$AUC_{all}$=0.85,IMR90: $AUC_{ac}$/$AUC_{all}$ =0.92) than that achieved using all 24 marks or even the top 4 non-acetylation marks (Fig.2.5A,B, green vs blue). For instance, in IMR90 the sensitivity and specificity are 81.24% and 84.94% respectively, as compared to 95.27% and 97.5% for all 24 marks, at default parameters.

Given the lower proportion of genic regions predicted with acetylations, we wanted to ask if this was because of the lower fractions of gene bodies recovered by acetylations or the existence of distinct categories of genes that are either completely acetylated or not. To this end, we examined the distribution of fractions of genes recovered by either case and that using all 24 marks leads to 90-100% recovery of most genes, while the fractions recovered by just acetylations appear to be more evenly distributed (Fig.2.6A,B). The partial recovery of certain genes using acetylations may indicate a bias towards certain elements within the gene. Since previous studies have found associations of acetylations with the splicing of certain genes[81], we tested the hypothesis that acetylations might have a preference for exonic regions or exon-intron boundaries, and found this to be true in both H1 and IMR90 (Supplementary Text, Fig.2.6).

While acetylations clearly show a bias towards exonic boundaries, there still exist a sizeable fraction of genes (12.7% in H1; 16.11% in IMR90), that can be recovered upto >90% using acetyations alone (Fig.2.6A,B). Since, distal regulatory elements lying within intronic regions may contribute to the acetylation signal as well, we filtered genic regions lying within 2.5kb of a distal DNase-I HS or an exon-intron boundary. Now, we calculated the classification rate of these filtered genic versus non-genic regions using all 24 modifications and just acetylations (Fig.2.7A, Fig.2.8A). It can be seen that the recovery using just acetylations is still well above the line of no-discrimination (significance stats), with a maximum classification accuracy of ~70% in H1 and ~80% in IMR90 (Fig.2.8A, Fig.2.7A).

Since gene body acetylations appeared to be quite discriminative in the case of IMR90, we further examined which acetylations are most enriched within the gene body. H2AK5ac, H3K23ac, H3K14ac, H4K5ac and H2BK5ac were found to be among the top acetylations in order of variable importance(Fig.4B) and also showed enrichment in a majority of genic regions upon normalization to intergenic background (Fig.2.7C). We selected long genes, such as TEAD1 (Fig.2.7D),CHRM2(Fig.2.7E) and CALD1(Fig.2.7F), that could be classified to over 90% against an intergenic background. It can be seen that several modifications such as H2AK5ac, H3K14ac, H3K23ac and H2BK5ac seem to cover a large proportion of the gene as compared to the neighbouring intergenic region. While some of this maybe accounted for by the presence of punctate regulatory elements, there are also regions that show diffuse enrichment of the above-mentioned acetylations, emphasized in Fig.2.7E n the black boxes.

In H1, similar analysis yielded a different set of acetylations that were seen to be among the most enriched at gene bodies, H3K27ac being the top-most in terms of variable importance(Fig.2.8B). Upon visualizing the enrichment of various histone modifications at genic regions versus intergenic ones, it does appear that H3K27ac has a ubiquitous but low presence (Fig.2.8C). The enrichment of several acetylations within the gene body can also been at the active gene PTPRJ, which is in sharp contrast to a neighbouring intergenic block with H3K9me3 enrichment (Fig.2.8D).

Finally, we examined if acetylations have any functional significance in gene bodies. Gene expression levels were slightly higher at acetylated genes (Fig.2.8E), showing a low but significant Pearson correlation coefficient of 0.2 in H1 and 0.14 in IMR90. Further, we examined if the genes with higher acetylation had specific associations with functional annotations. In H1 as well as IMR90, mRNA processing and RNA-binding were among the significantly enriched terms (Table 2.1). In addition, each cell-type showed different categories that were enriched such as that of genes involved in regulation of intracellular protein transport in IMR90 (Supplementary Table 2) or genes involved in mRNA splicing in H1 (Supplementary Table 1).

### Histone Modification Signatures at exon-intron boundaries

Previous observations of co-transcriptional splicing suggest that specific chromatin signatures maybe associated with splicing[78]. As a preliminary investigation, we chose to analyze the predictive power of the histone modifications under study in predicting exon-intron boundaries from the genic background. Using histone modification profiles(in 100bp bins) between -2 to +2 kb around the exon-intron boundaries, we were able to classify all known boundaries from genic background with an accuracy of 87% in H1 ($AUC_{all}$=0.94) or 85.5% in IMR90 ($AUC_{all}$=0.93) We then investigated the contribution of each histone modification under study to the prediction. Upon computing variable importance for each of the histone modifications with respect to the aforementioned classification, we found H3K36me3 followed by H3K79me1 to be the most informative and H3K36me3 alone could classify the boundaries within 3% of the accuracy achieved using all 24 modifications($AUC_{k36me3}/AUC_{all}$~96%).

To further investigate the association of histone modifications at exon-introns with function, we identified various splicing events from paired-end RNA-seq in both H1 and IMR90[34] using SpliceTrap[83]. The algorithm classified each local splicing decision as being one of constitutively spliced exon(CS), alternative donor site (AD), alternative acceptor site (AA), intronic retention (IR) or alternatively spliced exon (CA) with respect to its flanking exons. Based on the diversity of isoforms of a particular gene, this can cause one exon to be part of multiple different such splicing events. In each such splicing event, we may characterize the splicing decision in terms of the inclusion ratio, defined as the ratio of quantified expression level of the inclusion isoform divided by the sum of quantified expression levels of both inclusion and skipped isoforms. Further, each exon can also be quantified in terms of the exonic activity measured as FPKM (fragments per kilobase per million mapped reads). We aim to use these two quantifications at the exonic level to tease out correlations between histone modification signals and splicing activity.

Since there is a wide diversity of splicing activity in the transcriptome, the multiple signals associated with an exon-intron boundary may lead to the observation of a convoluted histone modification signal. As a first step towards deconvolving such putative chromatin modification signals, we discover all possible chromatin modification patterns at exon-intron junctions using a fast k-means++ algorithm[84](see Methods). Six distinct clusters are observed in H1 (Fig.2.9A), with varying levels of acetylations as well as other gene body marks such as H3K36me3, H3K79me1 and H4K20me1.Each of these clusters were characterized in terms of their distinctiveness from the genic background, by classifying the exons assigned to the cluster against the genic background using either all 24 modifications or just acetylations (Fig.2.10A,C) Overall, state 2 is unclassifiable against background using just acetylations indicating that the weak acetylation signature is comparable to the gene body while other states were found  to be either over-enriched (states 1,5,6) or under-enriched(states 3,4) for acetylations as compared to the rest of the gene (Fig.2.10A,C).It is worth noting that only those states with enrichment of acetylations appear to have presence of H3K79me1 as well.

In IMR90, on the other hand, we observe 4 distinct chromatin modification patterns(fig.2.11A). In common with H1 there is an “enhancer-like” cluster, cluster 1 (cluster 1 in H1) and “promoter-like” cluster, cluster 2 (cluster 5&6,H1), based on enrichment of H3K4me1 and me3 respectively. As in H1, these two are significantly enriched in acetylations with respect to genic background, while state 4 is significantly depleted (Fig.2.10B,D).

The learnt histone modification states in H1 cells are ranked in decreasing order of exonic activity based on calculations of statistical significance of the difference of mean RNA-seq FPKM (fragments per kilobase per million) levels between clusters using a Student’s t-test(Fig.2.9A,panel2). In H1, there appears to be a positive correlation with the level of H3K36me3 which is apparent as clusters 2> 3>4 that show significantly decreasing trends of activity also have correspondingly decreasing H3K36me3 (spearman correlation for clusters 1 to 4 =0.59,p-value<2.2X10-308). On the other hand, “TSS”-like signatures(clusters 5 and 6) appear to be even more highly active, irrespective of H3K36me3 enrichment. The same trend maybe observed in IMR90, where cluster 3 with the lowest enrichment of H3K36me3 also has the lowest activity(spearman correlation for clusters 1,3 and 4=0.47,p-value<2.2X10-308), and “TSS-like” state 2, has the maximum exonic activity (Fig.2.11A,panel 2).

In summary, H3K36me3 can accurately classify most exon-intron junctions from genic background. We identified multiple distinct chromatin states at both H1 and IMR90 that are associated with varying levels of exonic activity. We found that there was considerable variation in the levels of acetylations at exon-intron boundaries, many of which were either highly enriched or highly depleted in acetylations with respect to the rest of the gene.

### Chromatin modification patterns predict splice site usage

As described in the section above, an exon can be part of multiple different splicing events such as constitutively spliced exon(CS), alternative donor site (AD), alternative acceptor site (AA), intronic retention (IR) or alternatively spliced exon (CA) with respect to its flanking exons. A single exon-intron junction can have multiple assignments of inclusion values based on the transcript under consideration. Hence, we further developed a metric to characterize the overall splice site usage for every exon-intron boundary based on an expression-weighted average of its inclusion ratio in all transcripts (Methods).

Chromatin modification clusters are ranked in decreasing order of retention or increasing order of splice site usage in H1 using a Wilcoxon test with a p-value cutoff of 10^-5 (Fig.2.9A,panel 3). A clear trend is observed where greater the enrichment of acetylations, stronger the tendency for retention, with clusters 6, 5 and 1 having the maximum tendency for retention. We asked if we could build a predictor for splice site usage based on chromatin modifications, as input features. We defined retained and constitutively spliced exons based on a splice site usage cutoff of <-0.9 and >=0.999 , and filtered any exons that were proximal to the other category. This gives an average accuracy of classification of ~71% (Fig.2.9C,blue). Upon filtering the constitutive background for alternative exons from IMR90, we find a improvement in classification to ~74%(Fig.2.9C,red). This indicates that some constitutive exons in H1 may be pre-marked for alternative splicing in IMR90. If we use just acetylations, we achieve a comparable accuracy of 74% (Fig.2.9C,green) indicating that these are sufficiently distinctive between retained and constitutively spliced exons.

In IMR90 as well, the highly acetylated clusters 2 and 1 showed significantly higher retention of the boundary (ranked I and II based on a p-value cutoff of 10^-5). Prediction of splice sites usage from chromatin state showed a similar accuracy of ~74%, which only improved slightly upon eliminating retained junctions in H1 from the constitutively spliced background(Fig.2.11C,blue vs red). Further, acetylations had sufficient information content to achieve the same accuracy of classification of retained junctions to constitutively spliced ones (Fig.2.11C, green vs red).

Previous studies had shown H3K36me3 to be distinctive between alternatively-spliced exons and constitutively spliced ones[79]. We found that H3K36me3 was able to achieve a maximum classification accuracy of about 66% in both H1 and IMR90. This was ~8% less than that achieved using just acetylations, indicating the stronger association of alternative splicing with acetylation signatures, rather than H3K36me3.

Patterns in both cell-types were also associated with specific splice variants to see if there were significant associations with these (Fig.2.9B,Fig.2.11B). Alternative donor sites or 5' splice sites were enriched in the promoter-like clusters in both cell-types as compared to any other state. However, surprisingly all other splice variants also have a greater tendency to occur proximal to such promoter-like signatures. An example of a series of retained exon-intron boundaries in H1 and constitutively spliced in IMR90, can be seen in the gene PLEKH3 (Fig.2.12A) while the reverse can be seen in the gene VIM (Fig.2.12B). In both cases, the set of exons undergoing various types of retention, excluding alternative 5' site usage, are indicated by a black box and can be seen to be covered by the expansion of H3K4me3 signal in the cell-type with alternate usage. Another observation to note was that state 4 in H1 appeared to be preferential for exons with both ends constitutively spliced while states 1,5 and 6 show preference for other events such as alternative acceptor sites or intronic retention (Fig.2.9A,B).

In conclusion, acetylation-rich exon-intron junctions appear to be more enriched for retained boundaries in both H1 and IMR90. “Enhancer-like” and “Promoter-like” chromatin states are common to both cell-types that appear to be associated with splice site retention, of which the latter is the most strongly associated with a variety of splice site variants, not just alternative 5’ sites. Using just histone acetylation information, we were able to classify splice sites that are highly retained from those that are purely constitutive, upto an accuracy of about 74% in either cell-type.

### Dynamics of Chromatin modification states at splice sites

Certain chromatin modification clusters in H1 appear to be analogous to ones in IMR90 based on the patterns of modifications, such as the “enhancer-like” state 1(H1) with state 1(IMR90), and the “promoter-like” state 5 and 6 (H1) with state 2(IMR90)(Fig.2.9A,2.11A). However, the other clusters are not so easily comparable in terms of chromatin modifications. In this regard, we examined if particular states in H1 have a tendency to correspond to ones in IMR90 based on the number of exon-intron junctions that are common to the states in the two cell-types. We computed the p-value of transitions between the 6 states in H1 to the 4 states in IMR90 using a hyper-geometric distribution(Methods) and significant transitions, based on a p-value $< 2.2X10^{-308}$, are enumerated in Table 2. It appears that the chromatin state transitions are in keeping with the overall ranking in terms of splice site usage. For instance, state 2 in H1 and state 4 in IMR90 show significant transitions even though their chromatin modification patterns do not appear to be the same. However, both these clusters are ranked immediately after the “promoter-like” and “enhancer-like” states in terms of their splice site usage. Such a trend is in keeping with the fact that the change in splice site usage across the two cell-types is relatively small. For instance, if we assume any exon junction with splice site usage <0.9 to be called alternative, then only 1.92% of the total exons undergo any change at all in their splice site usage between H1 and IMR90.

We observed previously that we could obtain a higher accuracy of classification of alternatively spliced exons in H1, if we considered a negative set that was composed of constitutive exons in both H1 and IMR90, rather than just H1(74% vs 71%). This suggests that certain constitutive exons in H1 maybe “pre-marked” for alternative splicing in IMR90.In order to validate this, we created two sets of junctions – one that is alternatively spliced at usage levels <-0.9 in H1 but not IMR90, and another that is spliced at usage levels <-0.9 in IMR90 but not H1(Fig.2.11D,E,blue vs red). Both the acetylation rich clusters 1 and 6 in H1(Fig.2.9A)  are significantly enriched for celltype-specific retained junctions whether it is in H1 or IMR90 (Fig.2.11D). On the other hand in IMR90, the corresponding acetylation-rich clusters 1 and 2 (Fig.2.11A) are not significantly enriched for H1-exclusive retention events (Fig.2.11E). Hence, it may be that the states in H1 are pre-marked for alternative splicing in IMR90 since they are undifferentiated cells that contain the tendency for alternative splicing in future differentiated cells as well. Since IMR90 is a fully differentiated cell-type, it does not show similar tendencies.

Overall, it appears that only a small proportion (<2%) of exons undergo alternative splicing changes between H1 and IMR90.The chromatin modification patterns at exon-intron boundaries changes across H1 and IMR90 in such a manner so as to correspond to the splice site usage corresponding to the cluster, rather than the actual enrichment of various modifications. Also, constitutive exon-intron boundaries in H1 maybe pre-marked by an alternative splice site signature for use in later differentiated cell-types such as IMR90.

## Discussion

Chromatin modifications distinguishing promoters and enhancers have previously been identified as H3K4me1 and H3K4me3[24]. Besides these two, we find that several modifications, including histone acetylations that can reliably distinguish these regulatory elements. In particular, H3K9ac, H3K23ac and H3K14ac are promoter-specific, while H2BK120ac and H2BK20ac are enhancer-specific. Overall, histone acetylation is not only distinctive between the two regulatory elements but also informative enough to predict promoters and enhancers genome-wide. These observations potentially lead to several hypotheses regarding differences in mechanisms of functioning of these two regulatory elements. H2BK120 has been shown to have a ubiquitination modification that is present at active promoters and exclusive of H2BK120ac[85]. This exclusivity may explain the presence of H2BK120ac at enhancers, and suggest the lack of H2BK120Ub at these elements. Understanding the dynamics of the H2BK120 acetylase, KAT3[85] and the H2BK120 ubiquitin ligase, RNF20[86,87] may lead to further understanding of differences between enhancers and promoters.

Beside enhancers and promoters, acetylations were found to be quite informative in delineating gene bodies. Previously, only H4K16ac was characterized as being enriched in gene bodies [12]. [71].We find extensive enrichment of H2AK5ac, H2BK120ac, H3K14ac and H3K23ac along gene bodies, and acetylations alone can achieve 80% accuracy in predicting gene bodies. Some studies have shown PCAF to be regulating H3K14ac[88], also known to be part of an elongation-competent form of RNA-polymerase II[89]. This factor maybe involved in the maintenance of gene body acetylations in IMR90. Tip60 and HDAC6 have also been characterized as being within gene bodies [68], the former of which is known to acetylate H2AK5 [90]. Hence, given the patterns of acetylations within gene bodies, and prediction of genes enriched in these, there is a potential to generate hypotheses regarding the combinatorial localization of HATs and HDACs within specific genes.

Acetylations within the gene body are especially enriched near exon-intron junctions of retained exons. Indeed, we found that histone lysine acetylations alone can predict cell-type specific usage of exon-intron boundaries with up to 74% accuracy, which is ~8% higher than H3K36me3, a modification that has previously been described as quite distinctive between constitutive and alternatively-spliced exons [78,79]. But this accuracy of 74% likely reflects an upper bound, as acetylation-rich states in an earlier developmental state could be pre-marking alternative exons in a later state. For instance, many acetylation-rich, constitutive exons in H1 are alternatively spliced in IMR90. Such a hypothesis may be further tested by including detailed splicing and chromatin formation across many human cell-lines, both from early and late lineages. Further, there maybe regulatory elements distal to the actual splice-site that maybe regulating it's usage, which may be discovered using a chromosomal conformation captures technique such as 4C[91].

Hence, we observed patterns of histone acetylations that are specific to promoters, enhancers and genic regions. Such observations can suggest testable hypotheses regarding the enrichment of potential chromatin modifiers at various genomic elements that may lead to a better understanding of the mechanism of functioning of these elements.

## Methods

### Datasets and Processing

All datasets used, including 24 modifications in H1 and IMR90, various sequence-specific transcription factors and DNase-I hypersensitivity sites, were as used in the development of the RFECS algorithm[23]. In addition, the histone modification datasets in H9 can be accessed using GSE16256. Data normalization for histone modifications, determination of binding sites of transcription factors, training and prediction using RFECS, correlation clustering, visualization of chromatin patterns are also as previously described[23].

### Z-score normalization for comparing enhancers and promoters

We created a pooled set of equal numbers of distal p300 binding sites and known UCSC TSS overlapping DNase-I hypersensitivite sites, representing active enhancers and promoters respectively. We computed average histone modification levels, measured as input-adjusted RPKM(reads per kilobase per million), between -1 to +1 kb around each of these elements. The Z-score normalized profile for each element was calculated against the mean and standard deviation of the histone modification levels of the entire set of pooled elements. Hence, deviations of the mean z-score profile for the TSS class would be positive for TSS-preferred modifications while it would be negative for p300-preferred modifications. This would be the exact mirror image of the values of the mean z-score values for the p300-class.

### Genome-wide Prediction of promoters

In order to perform supervised prediction of promoters, we created a training set comprising of a set of UCSC TSS overlapping DNase-I hypersensitive sites as representative of the active promoter class, and a second class comprising of TSS-distal p300 binding sites as well as randomly selected non-p300 regions as background. We used input-adjusted RPKM values of histone modifications[23] measured in 100bp bins between -1 to +1 kb around the training set elements, as the input features for training this classifier. The RFECS classifier was then used to assign every 100bp bin in the genome "promoter" or "non-promoter" class based on a 50% voting percentage, after which promoter peaks were called in a genome-wide fashion as described previously for enhancers[23]. We validated our genome-wide promoter predictions by defining gold standard true positive(TP) and true negative(TN) sets. The former comprised of UCSC and Gencode annotated TSS overlapping DNase-I hypersensitivity sites in the particular cell-type while the latter (TN) comprised of p300 binding sites, cell-type specific TFs or DNase-I sites lying within gene desert regions. The true negative set was selected so as to comprise the elements most likely to be mistaken for promoters, due to the enrichment of active modifications. Training and prediction was performed using the RFECS methodology previously applied to prediction of enhancers.

### Computation of Variable Importance

We used the out-of-bag measure for variable importance[60] to compute importance of either all modifications or just acetylations for various classification or prediction tasks. Since not all modifications had the same replicates, we permuted replicates of each histone modification to create several different combinations and assessed the variable importance for each of these.

## RNA-seq data processing

We first mapped the Illumina-generated mRNA fragments (paired end reads) to the exon trio database TXdb, which we have previously built[83] using Bowtie version 1[32] for hits with no more than 2 mismatches. Our sequence mapping is based upon the human genome (hg19 assembly – Genome Reference Consortium GRCh37). The fragments are mapped to TXdb to be able to handle transcriptomic variability that arises from alternative splicing. TXdb represents every known contiguous sequence of exons in the human transcriptome as exonic trios and duos, such that mapping to this database allows us to quantify the splicing pattern in terms of the relative abundance of fragments of the different isoforms in this region, locally.

We ran the splicing analysis tool SpliceTrap version 0.90.5, with default parameters, which uses a Bayesian model to estimate inclusion ratios. SpliceTrap uses an inclusion ratio distribution model (estimated from high-confidence data) in order to reduce noise in the RNA Seq data without unnecessarily throwing away evidence from real transcriptomic events. Ultimately, it produces inclusion ratio estimates for all splicing events and classifies all local splicing decisions as constitutively spliced exon(CS), alternative donor site (AD), alternative acceptor site (AA), intronic retention (IR) or alternatively spliced exon (CA).

We chose to use SpliceTrap instead of other RNA-Seq analysis tools due to the facts that the SpliceTrap model is exclusively focused on optimizing a local, exon-centric splicing model (which is also our main focus), and that in our experience, SpliceTrap produces one of the most robust and consistent estimates of inclusion ratios among the tools we compared[83].

### Splice Site Usage

We created a measure of splice site usage by using labels associated with each exon-intron boundary to the various categories of splice sites – constitutively spliced exon (CS), alternative donor site (AD), alternative acceptor site (AA), intronic retention (IR) or alternatively spliced exon (CA). Each assignment is accompanied by an inclusion value of the exon with respect to the transcript under consideration. We assigned negative weights to all the cases where inclusion values represent increased inclusion such as IR, AA (3' end), AD(5' end) , and positive weights to the inclusion values that represent decreased inclusion such as AA(5' end), AD(3' end), CA and CS. The splice site usage value was defined as a weighted mean of the inclusion values, with the weights being the activity of the transcript under consideration. That is, splice site usage for a particular exon-intron boundary is:

$$SS = -\sum_{j\in A}\sum_{i\in Tj} Incl_i * FPKM_i + \sum_{j\in B}\sum_{i\in Tj} Incl_i * FPKM_i$$

I is a particular assignment of an exon with respect to a transcript $T_j$

$Incl_i$ is the inclusion value of exon-intron boundary in instance i

$FPKM_i$ is the RNA-seq FPKM value of the transcript I belonging to set $T_j$

A =[IR, AA (3' end), AD(5' end)]

B=[AA(5' end), AD(3' end), CA ,CS]

If there was no assignment for any of the seven cases due to weak coverage in that region, that term was set to 0.

### Identification of chromatin modification patterns at exon-intron boundaries

Using splice-trap, we obtained annotations for 286368 exon-intron boundaries in H1 and 246657 such boundaries in IMR90, of which 232919 boundaries had annotations in both cell-types. In each cell-type, we randomly selected a subset of 50000 sites (~25%) for unsupervised classification as larger number of sites required many more rounds of selection of the number of clusters to filter out the outliers. We performed fast k-means++ algorithm[84] at the exon-intron boundaries using RPKM-normalized histone modification levels in 100 bp bins between -2 to +2kb around the boundary as features, and determined the accurate number of clusters using the minimum value of the Davies-Bouldin measure[92]. We tested different randomly selected subsets of the data to ensure the results were robust. Further confirmation of the distinctiveness of each of these states was obtained by constructing RFECS classifiers for each cluster against all exon-intron boundaries not assigned to that cluster. We were able to show a 100% out-of-bag classification accuracy in H1 and over 95% in IMR90, for each cluster as compared to all others. We used these classifiers to assign all the boundaries that had not been used in the clustering to assign them to the appropriate state.

### Significance calculations for transitions of chromatin state at exon-intron boundaries between H1 and IMR90

For computing the significance of the transition from cluster I in H1 to cluster j in IMR90, we use a hyper-geometric distribution. Thus we model the probability by using the following analogies to the standard hyper-geometric distribution framework:

total exon-intron boundaries, N = total population

exon-intron boundaries belonging to cluster I in H1, m= elements having desired characteristic

exon-intron boundaries belonging to cluster j in IMR90, n=elements drawn without replacement from the population

exon-intron boundaries common to cluster I in H1 and cluster j in IMR90, x= number of elements drawn from the total population with the desired characteristic.

In Matlab, the p-value of transition from cluster I in H1 to cluster j in IMR90 was calculated as: p-value = 1 – hygecdf(x,N,n,m).

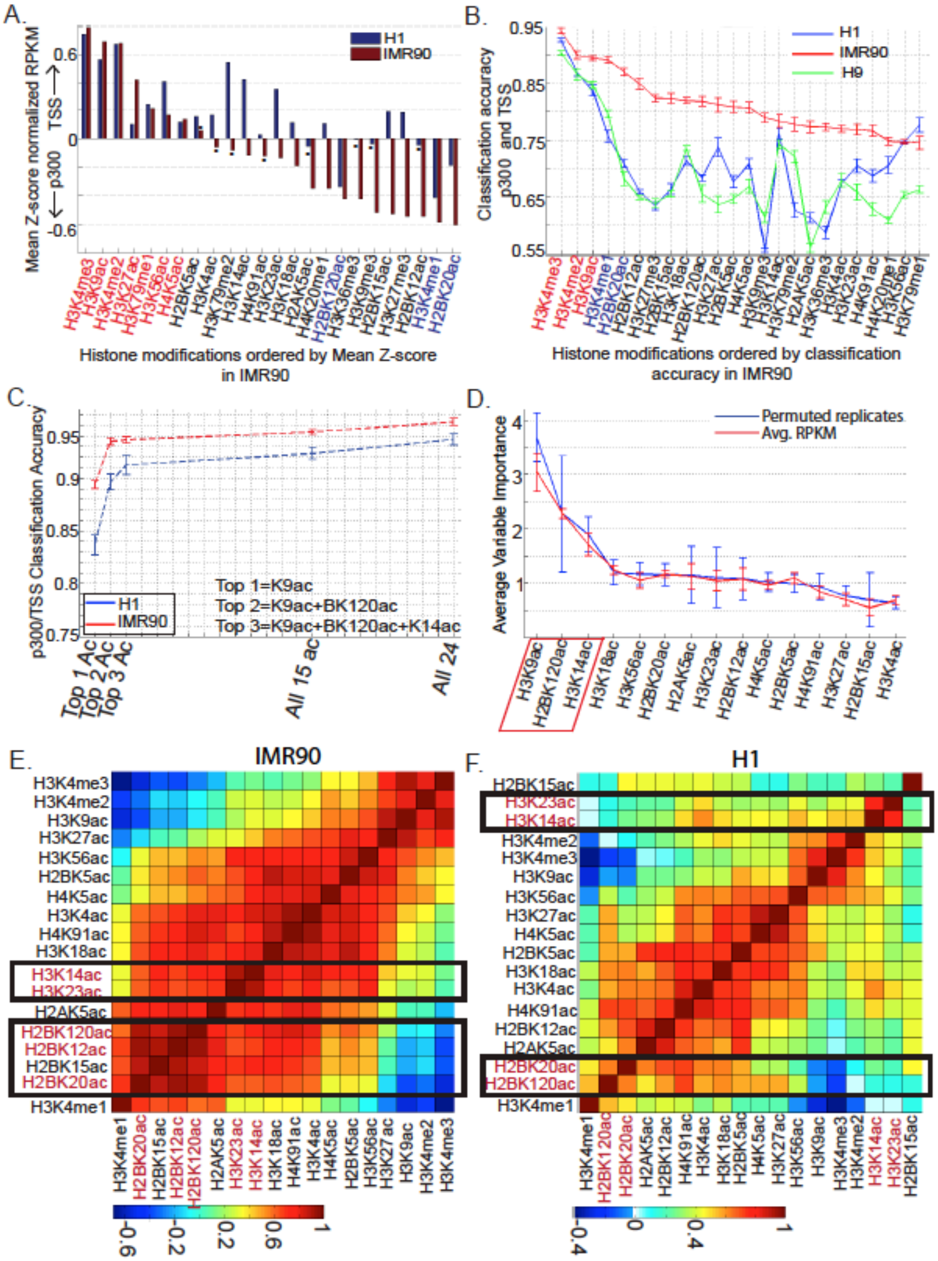

**Figure 3.1. Classification of distal enhancers and promoters**

A.) Preference of various histone modifications for either enhancer or promoter using a Z-score normalized score of histone modification levels measured as Input-subtracted RPKM(reads per kilobase per million). H1(blue) and IMR90(red) B.) Classification accuracy achieved using each of the 24 histone modifications individually to separate enhancers from promoters using RFECS in H1(blue), IMR90(red) and H9(green) cell-lines. C.) Comparison of classification accuracy of acetylations with that of all 24 modifications D.) Ordering of histone acetylations by their out-of-bag variable importance in classification of enhancers against promoters in H1. E,F.) Correlation clustering of histone acetylations at promoters and enhancers in E.) H1 and F.) IMR90.

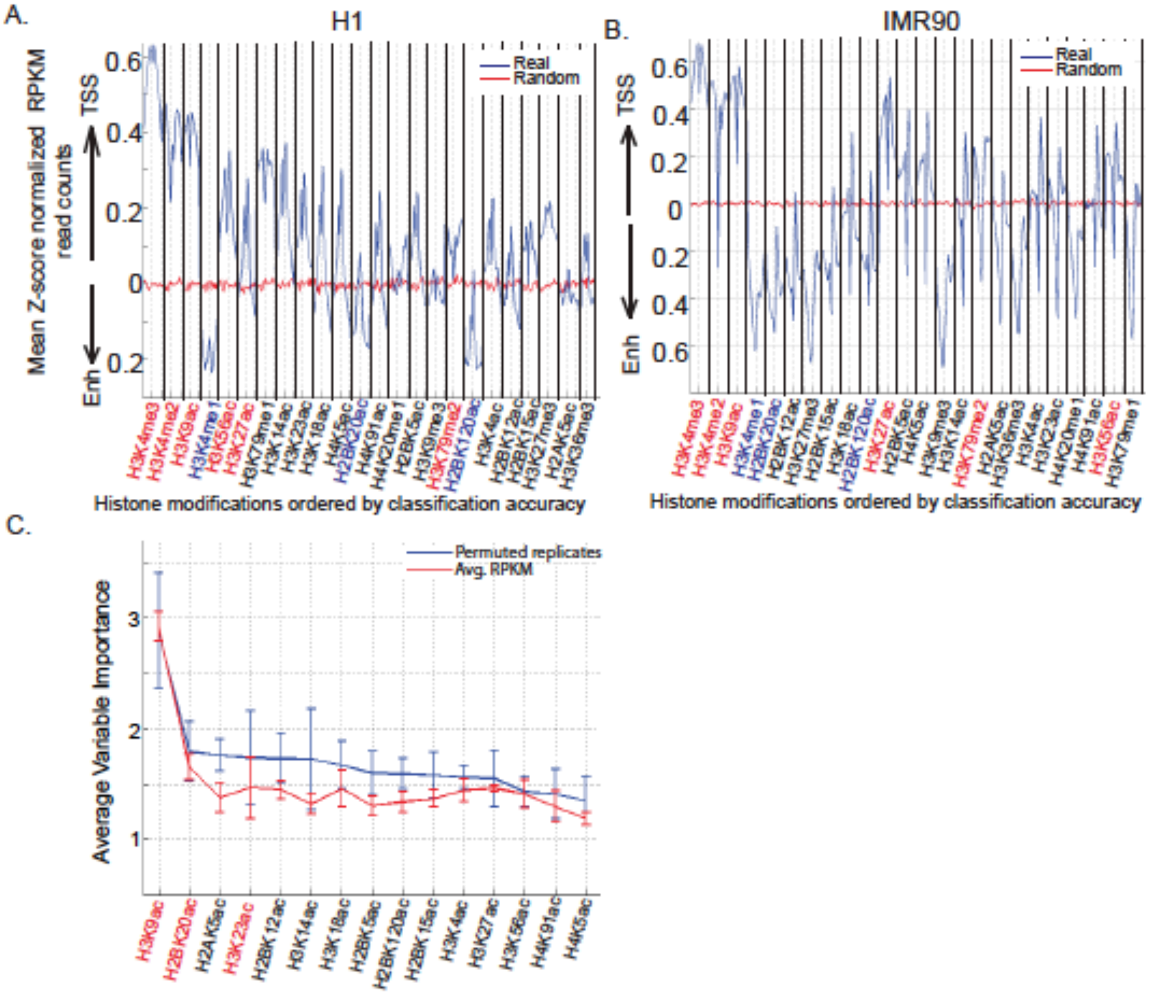

**Figure 3.2. Differential histone modifications between enhancers and promoters**
A,B.) Preference of various histone modifications from -1kb to +1kb around an enhancer or promoter using a Z-score normalized score (blue) as compared to the randomly shuffled class labels (red) in A.) H1 B.) IMR90 C.) Ordering of histone acetylations by their out-of-bag variable importance in classification of enhancers against promoters in IMR90.

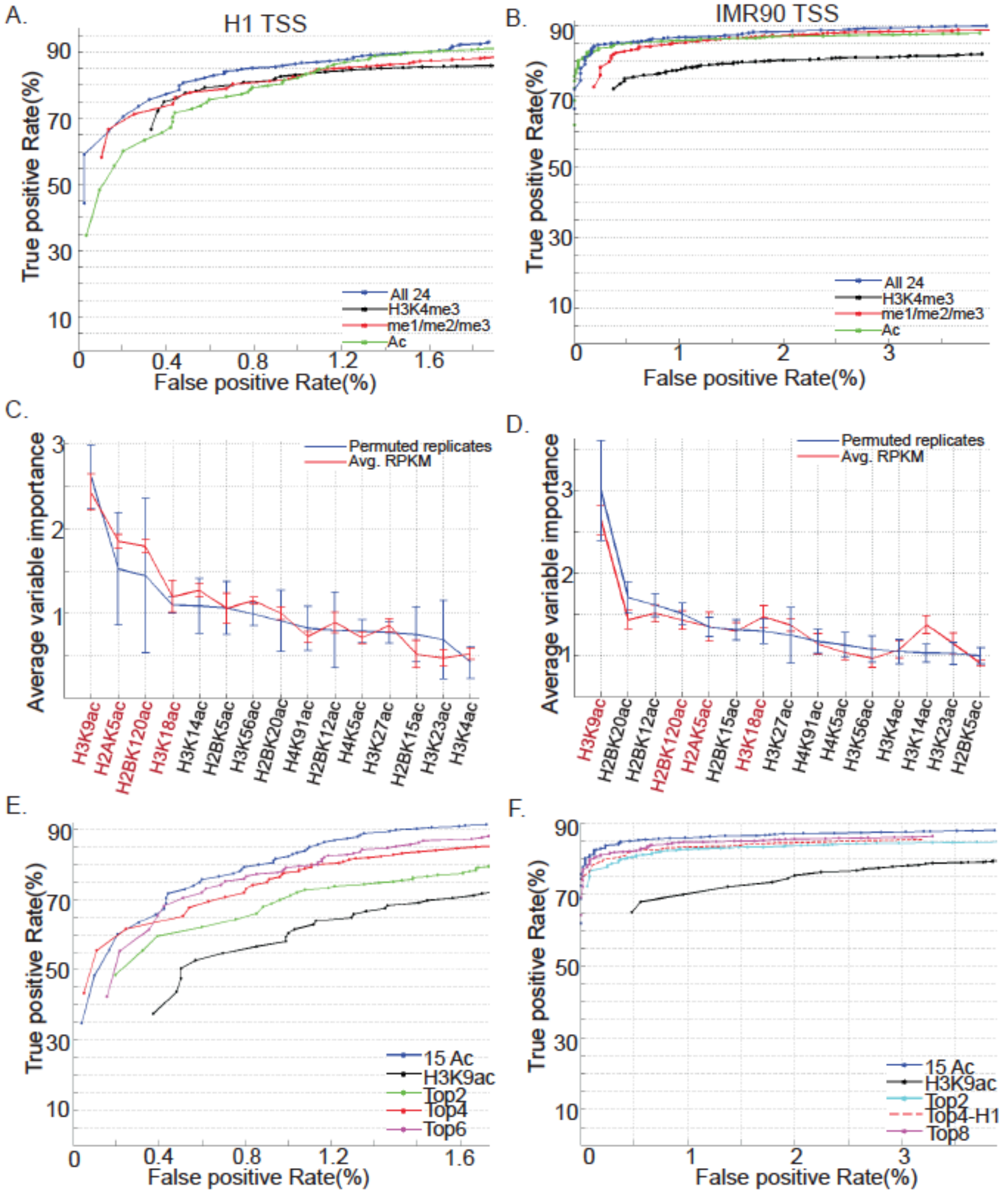

**Figure 3 3. Genome-wide prediction of promoters**

A,B.) Receiver operating characteristic(ROC) curves for prediction of promoters in A.) H1 and B.) IMR90 using all 24 modifications(blue),H3K4me3(black), H3K4me1/2/3 (red) or all 15 acetylations(green). C,D.) Out-of-bag variable importance for acetylations in making genome-wide prediction of promoters in C.) H1 and D.) IMR90. E,F.) ROC curves for prediction of promoters using minimal combinations of acetylations in E.) H1 and F.) IMR90.

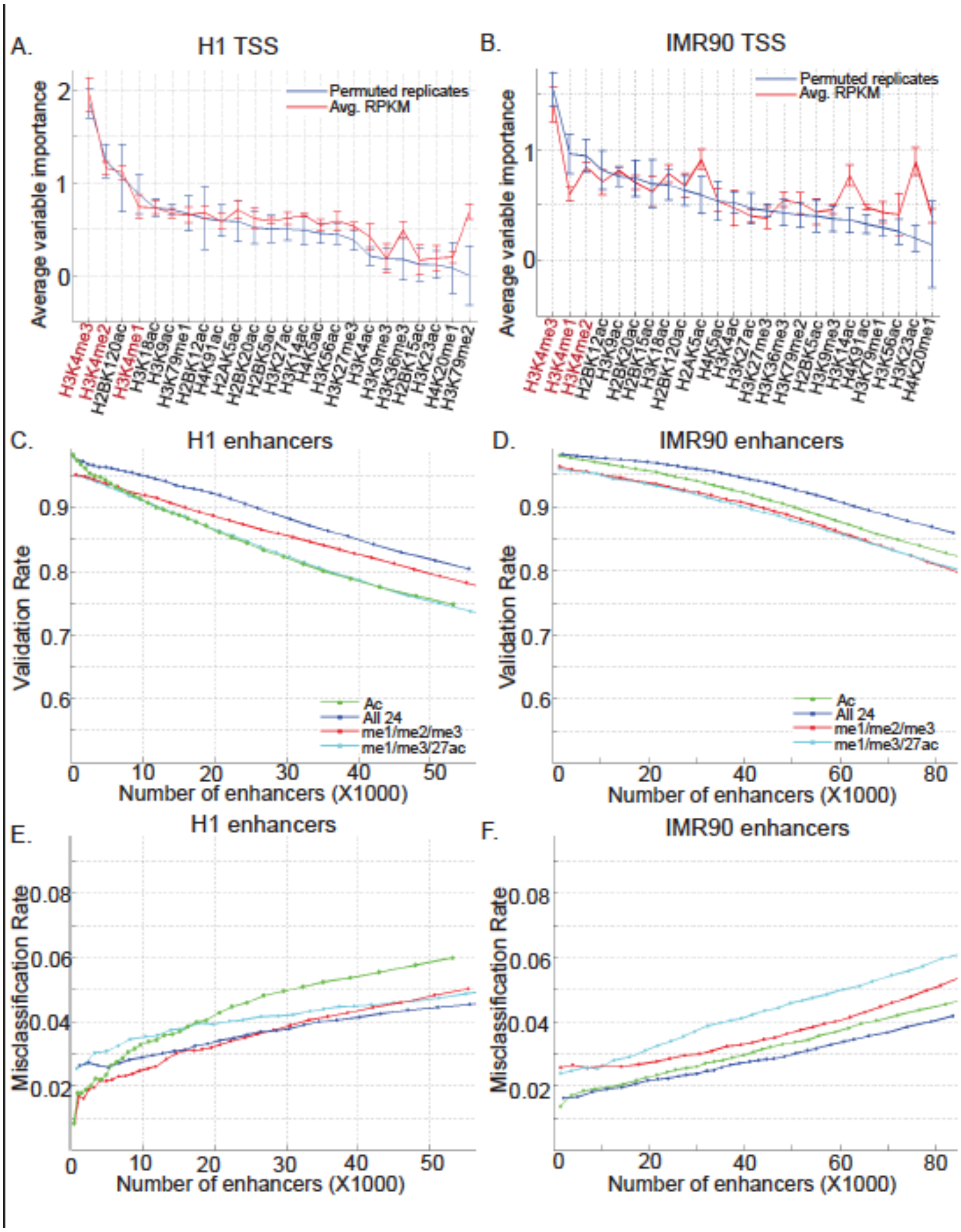

**Figure 3.4. Genome-wide prediction of promoters and enhancers**
A,B) Ordering of all 24 histone modifications by their out-of-bag variable importance for the prediction of promoters in A.) H1 and B.) IMR90. C,D) Validation rates and E,F) misclassification rates for enhancer predictions at various voting percentage cutoffs in C,E.) H1 and D,F.) IMR90 using all 24 modifications(blue), H3K4me1/2/3 (red), H3K4me1,H3K4me3 and H3K27ac(cyan) and 15 acetylations(green).

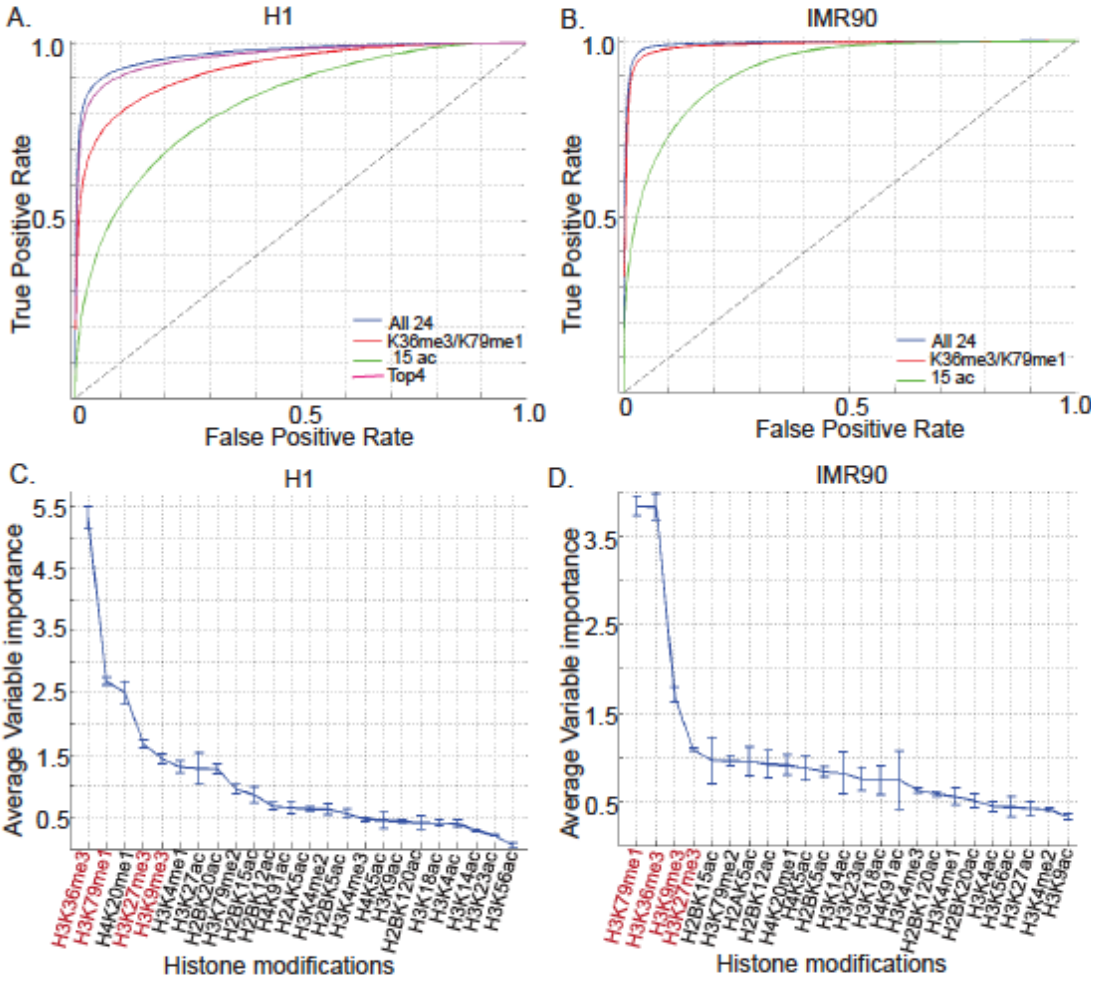

**Figure 3.5. Classification of genic from intergenic regions**
A,B.) ROC curves for classification of genic regions in A.) H1, B.) IMR90 using various combinations of modifications. C,D) Out-of-bag variable importance of all modifications in separating genic from intergenic regions in C.) H1 and D.) IMR90.

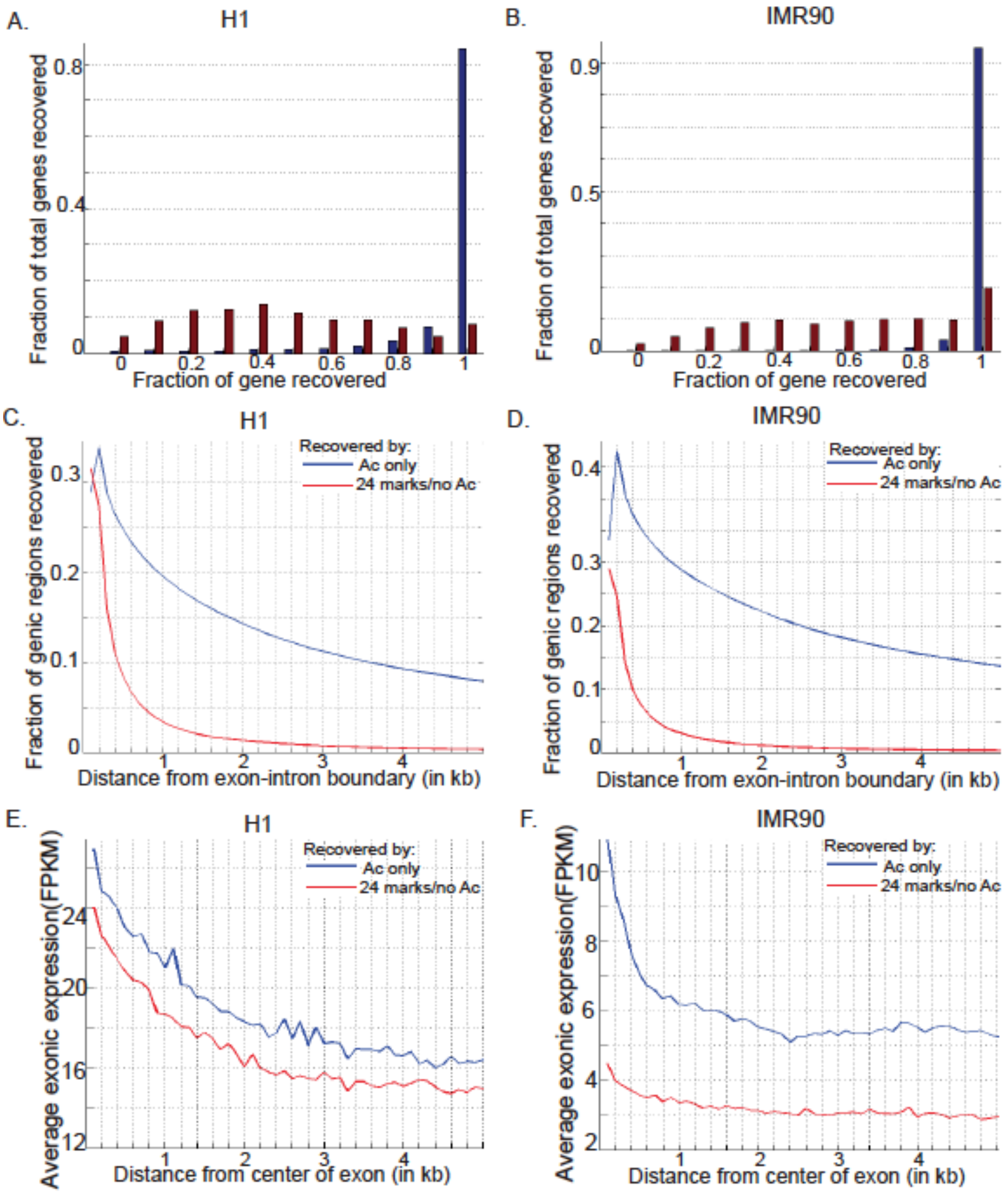

**Figure 3.6. Recovery of genic regions using acetylations**

A,B.) Fraction of gene body predicted in A.) H1 and B.) IMR90 using all 24 modifications(blue) or just acetylations(red). C,D.) Fraction of predicted 100bp bins lying at various distances from exon-intron boundaries using all 24 modifications(blue) or just acetylations(red) in C.) H1 and D.) IMR90. E,F.) Distance of predicted 100bp bins from exon-intron boundaries versus activity of exons in E.) H1 and F.) IMR90.

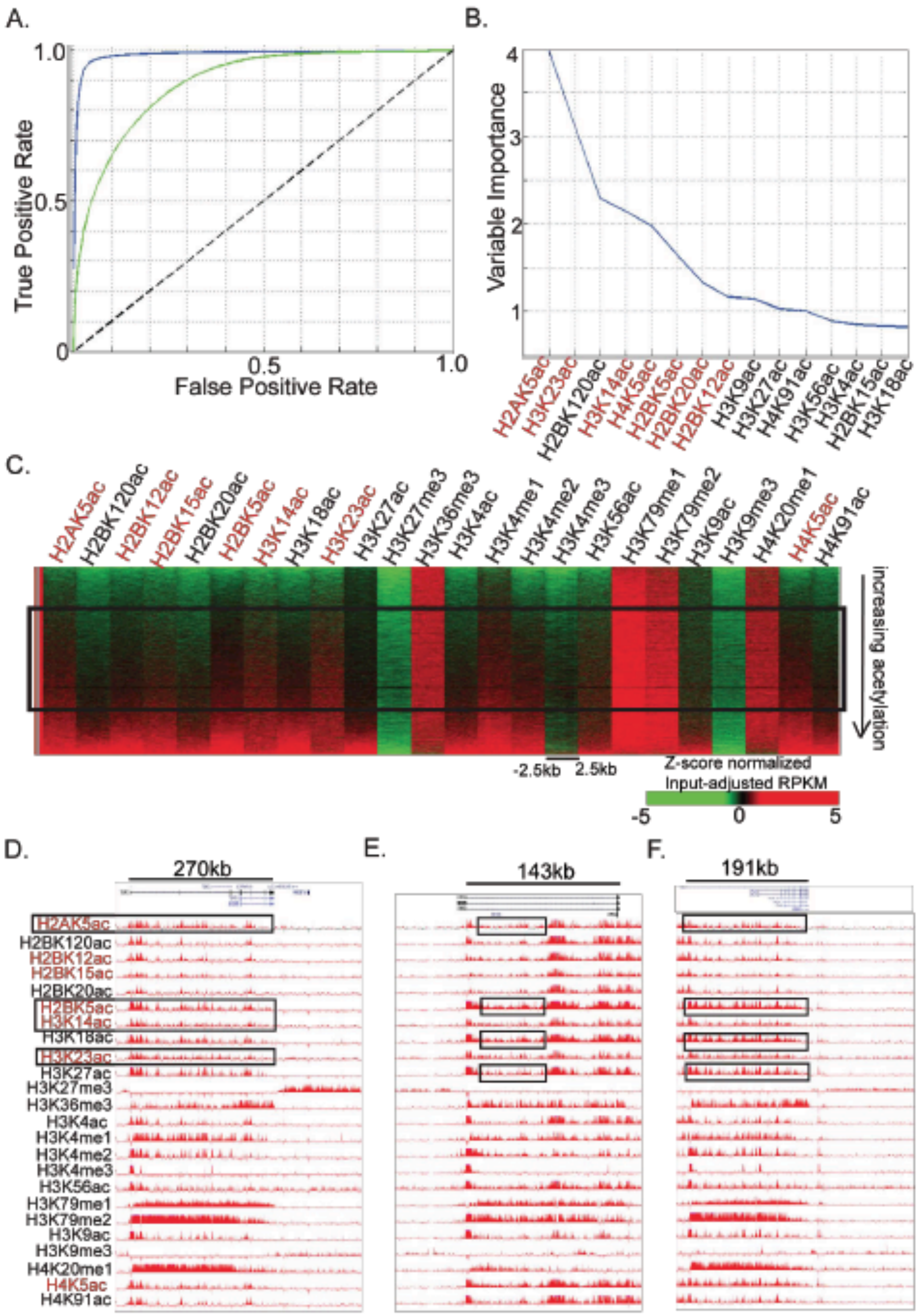

**Figure 3.7. Acetylations within the gene body distal to exon-intron boundaries and DNAse-I hypersensitive sites in IMR90**

A.) ROC curves showing classification of distal genic regions using all 24 modifications(blue) or only 15 acetylations(green). B) Variable Importance of acetylations in classification of distal genic regions C.) Heatmap showing enrichment of acetylations in genic regions as compared to intergenic ones using a Z-score normalized measure. D,E,F.) UCSC genome browser snapshot of genes D.) TEAD1, E.) CHRM2, and F.) CALD1,showing enrichment of acetylations as compared to neighboring intergenic regions.

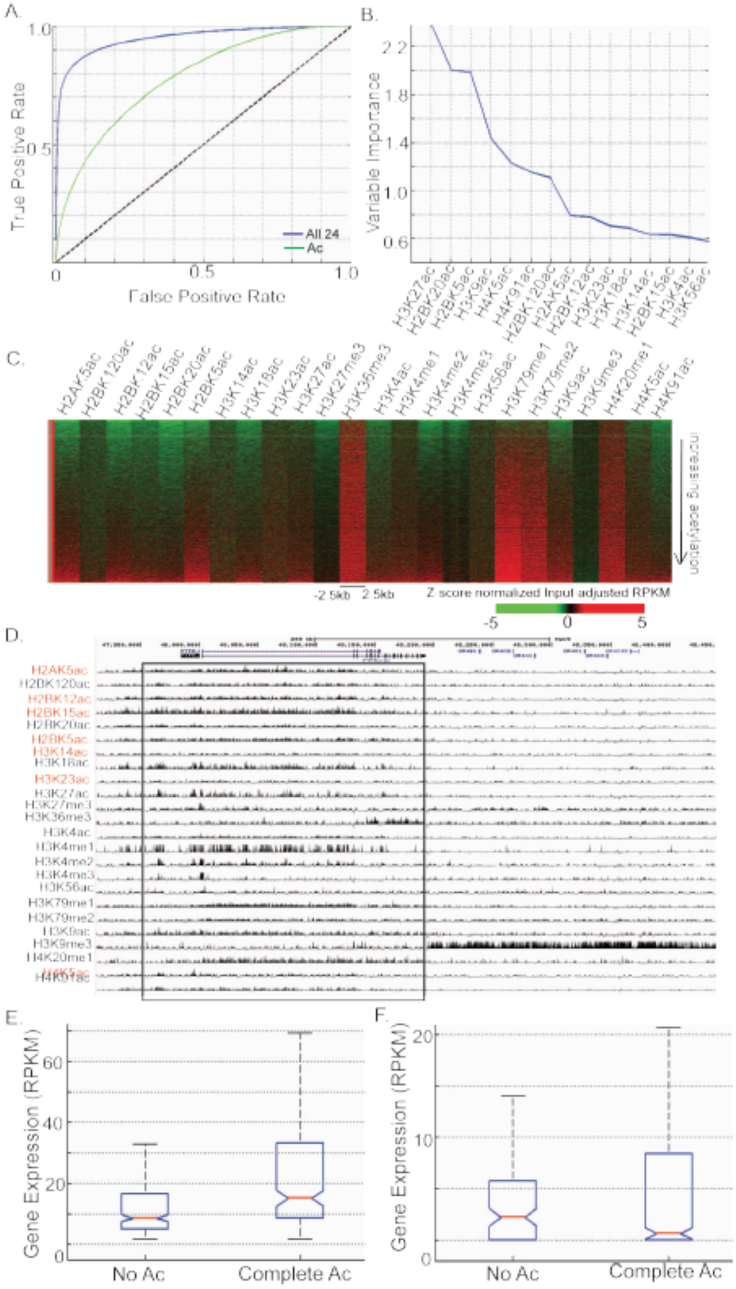

**Figure 3.8. Acetylations within the gene body distal to exon-intron boundaries and DNAse-I hypersensitive sites in H1**

A.) ROC curves showing classification of such distal genic regions using all 24 modifications(blue) or only 15 acteylations(green).B.) Variable Importance of acetylations in classification of distal genic regions C.) Heatmap showing enrichment of acetylations in genic regions as compared to intergenic ones using a Z-score normalized measure. D.) UCSC genome browser snapshot of gene PTPRJ showing enrichment of acetylations as compared to neighbouring intergenic region. E,F.) Gene expression levels versus fractional enrichment of acetylations within the gene body in E.) H1 and F.) IMR90.

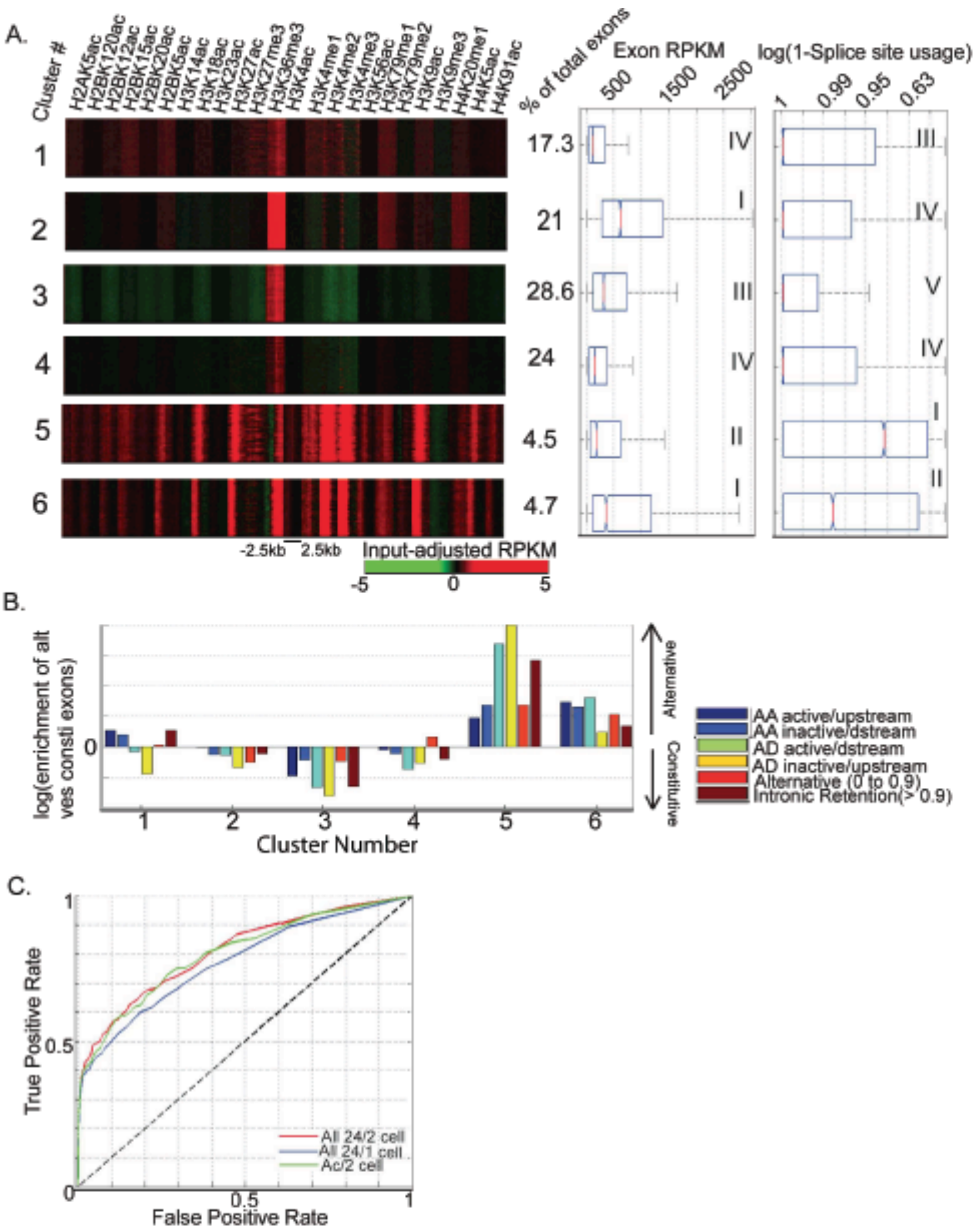

**Figure 3.9. Associations of chromatin modification patterns with splicing in H1**
A.) 6 distinct chromatin modification patterns at exon-intron junctions with corresponding levels of exonic activity(panel2) and splice site retention(panel3). B.) Association of various types of splice variants with each chromatin state C.) Maximum classification accuracy of each state against genic background using all 24 modifications(blue) or 15 acetylations (red).

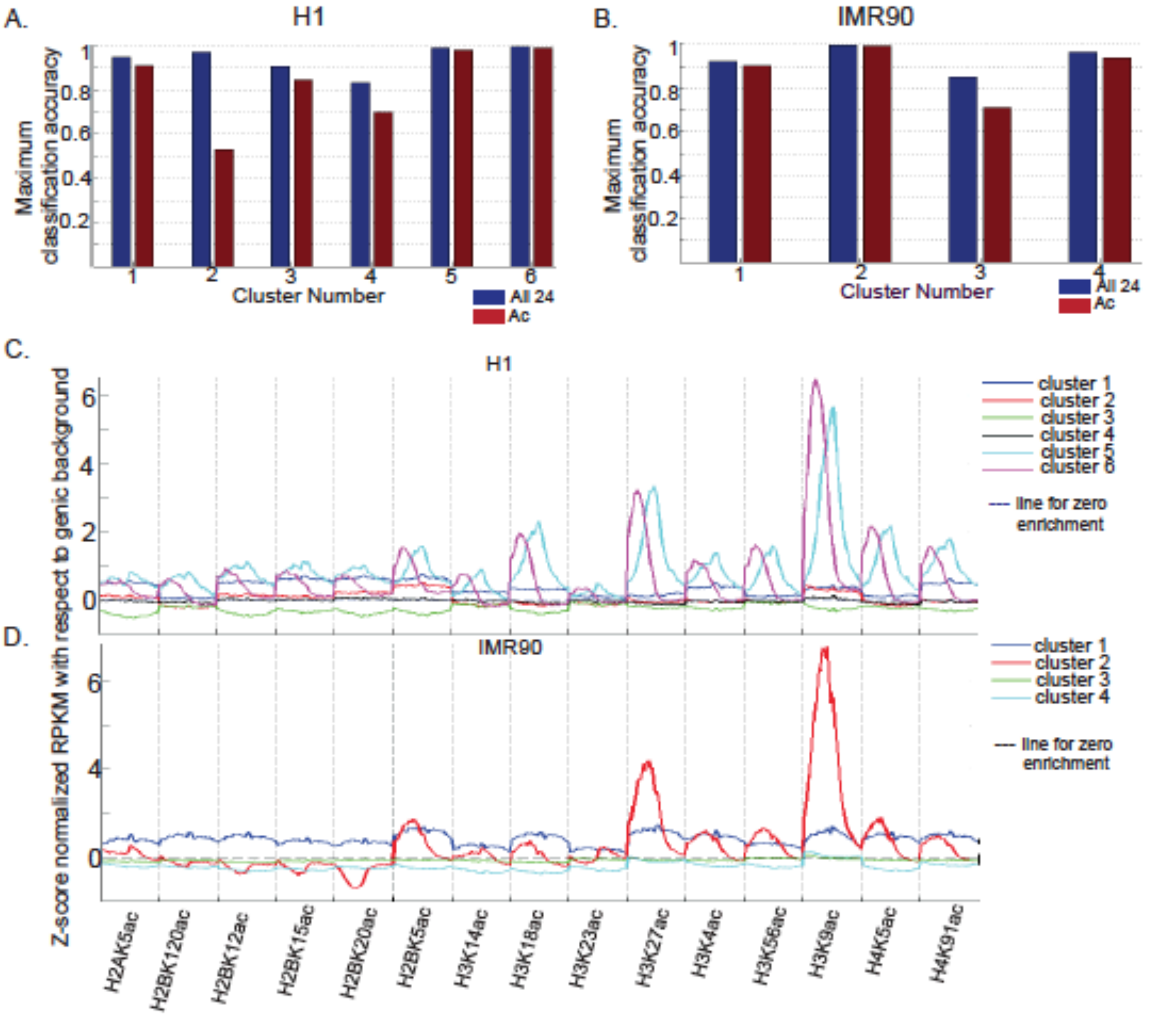

**Figure 3.10. Enrichment of acetylations at exon-intron boundaries for each chromatin state (Fig.5A,6A) with respect to genic background**

A,B.) Maximum classification accuracy of each state against genic background using all 24 modifications(blue) or 15 acetylations (red) in A.) H1 and B.) IMR90. C,D.) Histone acetylation RPKM levels at exon-intron boundaries Z-score normalized against levels at distal genic regions in C.) H1 and .D) IMR90.

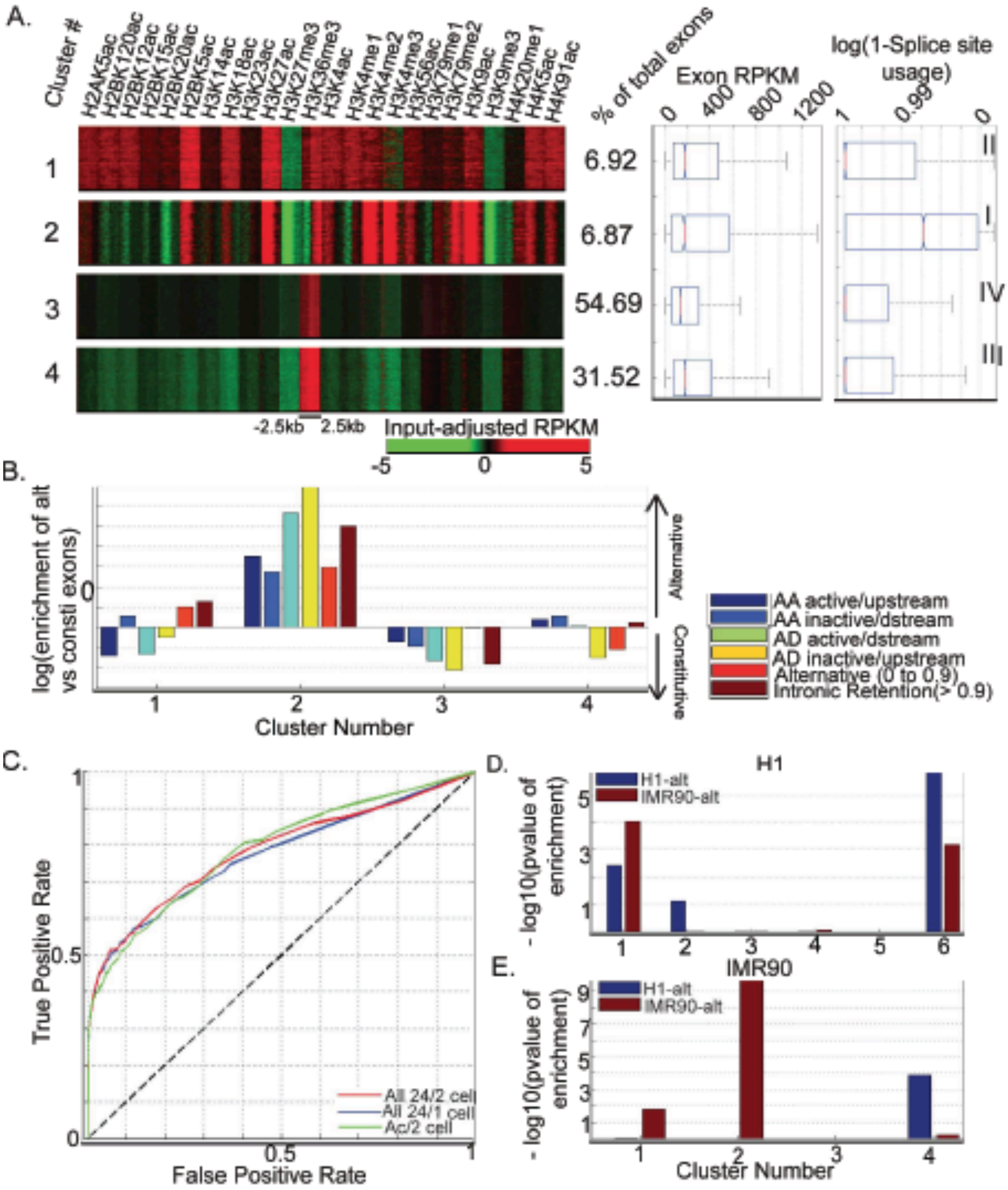

**Figure 3.11. Associations of chromatin modification patterns with splicing in IMR90**
A.) 4 distinct chromatin modification patterns at exon-intron junctions with corresponding levels of exonic activity and splice site retention. B.) Association of various types of splice variants with each chromatin state C.) Maximum classification accuracy of each state against genic background using all 24 modifications(blue) or 15 acetylations (red) D,E) Negative logarithm of the p-value of enrichment of alternatively spliced exons exclusive to H1(blue) or IMR90(red) in D.) H1 and E.) IMR90.

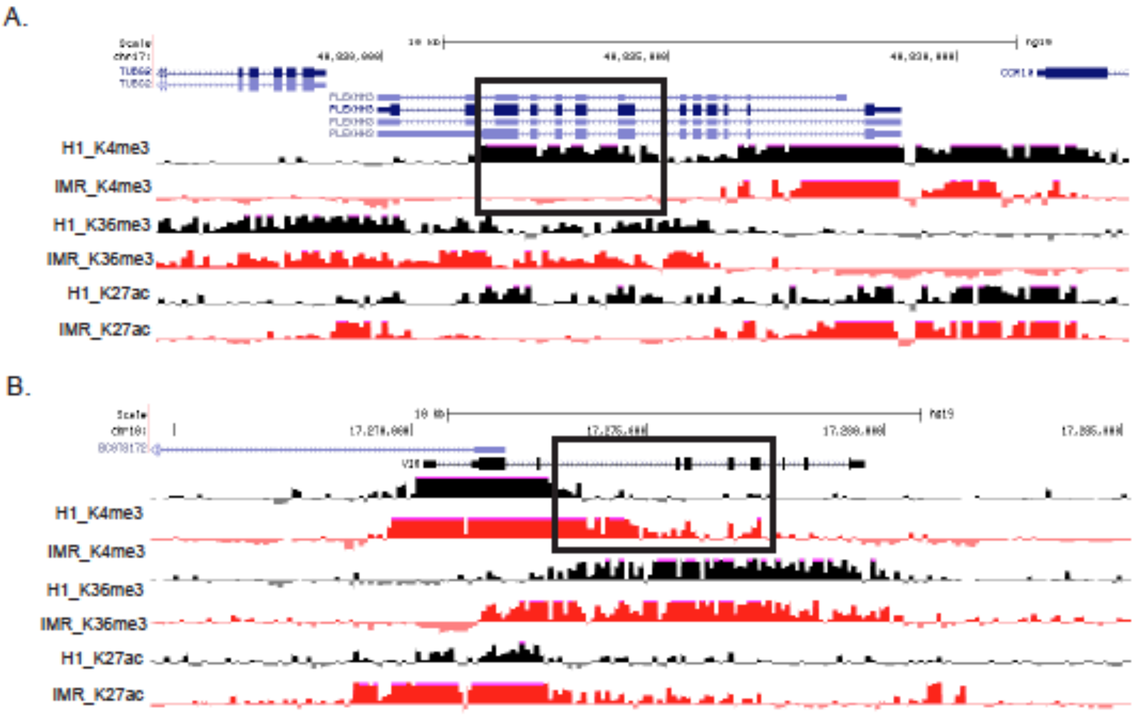

**Figure 3.12. “Promoter-like” chromatin states are associated with various splice variants** Snapshots from UCSC genome browser of A.) PLEKH3 showing alternative splicing of several exons in H1 as compared to IMR90 B.) VIM showing 87lternative splicing of several exons in IMR90 as compared to H1.

**Table 3.1. GO terms for Acetylation-rich genes in H1 and IMR90**

| GO term | description | H1 p-value | IMR90 p-value |
|---|---|---|---|
| GO:0006397 | mRNA processing | 5.90E-09 | 7.19E-04 |
| GO:0010467 | gene expression | 4.79E-05 | 4.79E-05 |
| GO:0003723 | RNA binding | 3.21E-04 | 1.03E-05 |

**Table 3.2. Significant chromatin state transitions at exon-intron junctions between H1 and IMR90**

| | IMR90 clust 1 | IMR90 clust 2 | IMR90 clust 3 | IMR90 clust 4 |
|---|---|---|---|---|
| H1 clust 1 | Yes | No | No | Yes |
| H1 clust 2 | No | No | No | Yes |
| H1 clust 3 | No | No | Yes | No |
| H1 clust 4 | Yes | No | Yes | No |
| H1 clust 5 | Yes | Yes | No | No |
| H1 clust 6 | No | Yes | No | No |

## ACKNOWLEDGEMENTS

Chapter 3, has been submitted for publication. Nisha Rajagopal, Jason Ernst, Pradipta Ray, Jie Wu, Michael Zhang, Manolis Kellis, Bing Ren. Distinct and Predictive histone lysine acetylation patterns at promoters, enhancers and gene bodies. The dissertation author was the primary investigator and author of this material.

# Chapter 4.Dynamic epigenetic signatures at regulatory elements in human tissues

## Abstract

Covalent histone modifications play important roles in embryonic lineage specification and development. Recent studies have shown diversity of such marks in various primary cell types. Here, we expanded the scope of available data by generating genome-wide maps of chromatin marks in 17 somatic tissues isolated from 4 human subjects. Combined with previously published datasets, we conducted comprehensive analyses to elucidate epigenetic differences and their potential function across an array of diverse cell types and tissues. We employed a unique random-forest based algorithm developed in-house to identify transcriptional promoter and enhancer elements in over 45 human cell types and tissues based on their chromatin states. We uncovered several novel features regarding the dynamics of regulatory sequences across tissues and along development lineages. Intriguingly, we discovered dual property *cis*-regulatory elements, which harbor the capacity to function both as promoters and enhancers in different cell types. Over 60,000 such elements exist among all analyzed cell types and tissues, some of which could give rise to cell type-specific isoforms. Furthermore, we observed significant differences between tissue-specific regulatory sequences as opposed to ubiquitous elements in the context of evolutionary conservation and association with disease variants.

Taken together, our results reveal variations of histone modification patterns between somatic tissues, namely at tissue-specific *cis*-regulatory elements. The dynamics of these epigenetic marks could substantially influence gene expression, which could potentially explain the distinct phenotypes of genotypically identical tissues.

## Introduction

The human body is comprised of more than 200 cell types. Although possessing essentially the same genome, these cells have vastly diverse transcriptional profiles, which allow them to have distinct properties relevant to their functions. Transcriptional regulation of particular sets of genes play critical roles in driving differentiation and providing cell type specification[93]. Until recently, few studies have done systematic profiling of the transcriptome and the factors, which shape cell identity. The work done by the ENCODE consortium recently elucidated much of this aspect in an array of human cell types[16]. Specifically, when analyzing all transcripts by RNA-seq, Djebali et al. discovered that although three quarters of the genome is transcribed, a large proportion of transcripts are restricted to particular cell types[94]. Much of the previous studies regarding tissue or cell type specific transcriptional regulation have focused on networks of transcriptional factors (TFs), which activate target genes by binding to regulatory sequences[95]. However, in recent years mounting evidence has shown in addition to TF binding, epigenetic factors such as histone modifications and DNA methylation also play critical roles in the process of cell-type specification[34,96,97,98]. DNA accessibility at regulatory sequences, which has strong correlations with chromatin state[99], was found to predict cell-type specific expression quite accurately [100].

It has been shown that activity of cis-regulatory elements can be delineated by analyzing the enrichment of covalent histone modifications. For instance, histone H3 lysine 4 trimethylation (H3K4me3) is associated with active transcriptional promoters whereas trimethylation of histone H3 lysine 27 (H3K27me3) and lysine 9 (H3K9me3) are associated with repression[101]. H3K4 monomethylation (H3K4me1) was discovered to indicate enhancer elements[24]. In addition, active enhancers are also marked by H3K27 acetylation (H3K27ac)[102,103]. Recently, several studies have employed genome-wide techniques to investigate the role of epigenetic modifications in gene regulation in the context of cellular differentiation.

As a part of the Roadmap Epigenome project, our group studied the epigenomic changes that are associated with early embryonic differentiation by analyzing the chromatin states, DNA methylomes and transcriptomes of H1 human embryonic stem cells (ESC) and derived mesendoderm (ME), neural progenitors (NPC), trophectoderm and mesenchymal stem cells (MSC)[34]. As in our study, Gifford et al. also independently found that differentiation of human embryonic stem cells results in lineage-specific epigenetic remodeling[96].

Furthering these findings to in vivo tissues, Zhu et al. produced chromatin state maps for multiple histone modifications in many diverse cell types and tissues, which demonstrated global chromatin state changes which occur as a response to developmental or environmental cues[98]. Consistent with the diverse profiles of stem cells and various differentiated cell types, the chromatin states are highly dynamic between tissue types as well. Although histone modifications have been analyzed in certain somatic tissues and early embryonic cell types, the epigenetic differences at cis-regulatory elements across various somatic tissues have not been thoroughly studied.

In this study, we focused on the differential histone modifications of cis-regulatory elements across distinct cell and somatic tissue types. We generate extensive datasets profiling 6 core histone modifications across 17 human tissues isolated from 4 individual human donors. Combining previously generated datasets of the Roadmap Epigenome consortium[98], we conducted comprehensive analyses to elucidate epigenetic differences and their potential function across an array of 32 cell types, including 6 cell-lines and 26 primary tissues. These cell-types included early as well as late lineages, and have representation in all 3 germ layers (Fig.3.1,Table 3.1). Using the RFECS algorithm[23], we predicted transcriptional enhancers and promoters from histone modification profiles in all 32 cell-types.

One of the major distinctions between promoters and enhancers was considered to be the ability of promoters to serve as the point of assembly of the transcriptional machinery and initiation of transcription[104]. Recent years have shown that enhancers (whose chromatin is marked by high levels of H3K4me1 and low levels of H3K4me3) may also be transcribed to produce short bidirectional transcripts, called eRNAs[39] which may or may not be polyadenylated[105]. Interestingly, a study in mouse showed the production of elongated polyA+ transcripts from intragenic enhancers giving rise to additional isoforms for the gene within which the enhancer lay[106]. Such findings of transcription at enhancers blurs the lines between the strict definitions of enhancers and promoters, and provokes the question if the same regulatory sequence cannot function as an enhancer and promoter under different circumstances. Upon comparing enhancer and promoter predictions among the 32 cell-types mentioned above, we discovered over 60000 dual property cis-regulatory elements, which harbor the capacity to function as a promoter in one cell-type and an enhancer in another.

Evolution of regulatory DNA sequences may underlie the morphological diversity of animal species[107,108]. Hence, understanding the association of evolutionary conservation of regulatory sequences with tissue function could be particularly beneficial in understanding what distinguishes humans from other species. Based on the cis-regulatory map of 32 cell-types, we report novel associations of evolutionary conservation with tissue-specificity and disease-causing mutations.

Overall, our results reveal variations of histone modification patterns between somatic tissues, at tissue-specific cis-regulatory elements, which could potentially explain the distinct phenotypes of genotypically identical tissues.

## Results

### Genome-wide prediction of enhancers and promoters

RFECS[23] was used to predict enhancers and promoters in the 6 cell-lines and 26 human tissues(Fig.3.1) under consideration using the 6 profiled core histone modifications. Further, we filtered any enhancer prediction that had a promoter prediction lying with 2.5 kb of it within the same cell-type. This gave us a total of 311032 enhancers and 174805 promoters across 32 cell-types (Table 3.2).

We identified highly active enhancers using the enrichment of H3K27ac modification. We used fast k-means++ clustering[84] of all predicted enhancers using the H3K27ac profile as the input feature to assign “present” and “absent” values of H3K27ac to each enhancer in each cell-type. Enhancers have been known drive cell-type specific gene expression[30] hence similarities among enhancers in various lineages can reflect the similarities of the lineages themselves. With this in mind, we performed hierarchical clustering and optimal ordering of the 32 cell-types using the overlap between the highly active enhancers(as defined by H3K27ac) in each cell-type as a similarity measure(Fig.3.2A,Table 3.1).At the top-most level in Fig.3.2A, 3 broad clusters are observed- an early lineage cluster with H1, Mesendoderm ,NPC and Trophoblast (in blue); a cluster with all 5 brain tissues (in red); and the largest cluster with all mesoendodermal late lineages (in green). At the next level, the mesendodermal cluster clearly separates into the one cluster with cell-lines MSC(mesenchymal stem cells) and IMR90; and another with all the primary tissues. Going further down the tree, various compartments of the heart such as Right auricle(RA),Right ventricle(RV) and Left ventricle(LV) can be seen to cluster together, while other sets of endodermal or mesodermal lineages can also be either clustered or ordered together(Fig.3.2A). Hence, clustering of enhancers reflects many known similarities among the cell-types and maybe further considered to reflect hitherto unknown similarities among lineages.

Tissue-specific enhancers were defined as those enhancers where H3K27ac was “present” in atmost 2 cell-types. We obtained 126950 number of tissue-specific enhancers and 13218 tissue-specific promoters, defined in the same way as enhancers (Table 3.2). It is worth noting that while nearly 41% of enhancers are tissue-specific, only about 8% of promoters are tissue-specific, which is in keeping with prior beliefs regarding the highly tissue-specific nature of enhancers as compared to promoters. A heatmap showing enrichment of H3K27ac, H3K4me1 and H3K4me3 at all of these tissue-specific enhancers confirms that we did indeed detect enhancers with H3K27ac and H3K4me1 signal that is highly specific to the cell-type under consideration. (Fig.3.2B). We searched for transcription factor binding motifs enriched at the tissue-specific enhancers using HOMER[109]. At a p-value cutoff of 10^-10, we discovered significant motifs at tissue-specific enhancers, several of which associated with terms that would be expected to belong to the tissue type based on previous literature. For instance, Nr5a2 was among the top motifs in pancreas[110] and ovary[111], HNF4A in liver[112],TBX20 in heart right ventricle[113] and so on. The remaining list of enriched motifs in each cell-type maybe downloaded from enhancer.ucsd.edu/nisha/human_tissue_motifs .

Overall we predicted 311032 genome-wide enhancers and 174805 promoters in 32 cell-types using chromatin modification profiles . We clustered cell-types based on the enhancer-predictions and were able to observe several expected groupings such as those of the brain tissues, early lineages or heart tissues. Further, we also found lists of tissue-specific enhancers and promoters that can be further examined for tissue-specific properties such as enrichment of cell-type specific transcription factor binding sites.

### Identification of dual Property cis-regulatory elements

A previous study in mouse had shown that intragenic enhancers could act as alternative promoters[106]. Upon comparison of the genome-wide predictions of enhancers and promoters, we found that nearly 20% of enhancers active in one cell-type or tissue, were predicted as promoters in another tissue, and about 36% of predicted promoters were enhancers in another cell-type(Fig.3.3A). Based on the total number of switches between enhancers and promoters, we tried to identify cell-types where switches were significant using a hyper-geometric distribution(p-value<10^-3). We found that enhancers in skeletal muscle, spleen and lung tissues showed significant switches to promoters in over 10 cell-types. Enhancers undergoing switches in the skeletal muscle, spleen and lung were 28%,27% and 19% of the total enhancers in the cell-type, indicating this might be a dominant mechanism for enhancer creation within these 3 cell-type. This is in keeping with the heatmap showing enrichment of modifications at tissue-specific enhancers(Fig.3.3B), where substantial subsets of tissue-specific enhancers in lung(LG), spleen(SX) and skeletal muscle(Sk.Mu) appear to have enrichment of H3K4me3 in other cell-types. However, in order to obtain a more accurate view of this phenomenon, we consider as examples all predicted enhancers in lung and skeletal muscle that undergo switches to predicted promoters in other cell-types. The patterns of enrichment of H3K4me1 and H3K4me3 in 32-celltypes for these enhancers in skeletal muscle(Fig.3.3B,panel1) and in lung(Fig.3.3B,panel 2) show an enrichment of H3K4me1 and depletion of H3K4me3 in skeletal muscle and lung, respectively as would be expected for enhancers. However, H3K4me3, which is a promoter-preferred mark, is strongly enriched for a majority of these sites in nearly all other cell-types or tissues(Fig.3.3B). This provides visual confirmation for our predictions of dual property elements based on chromatin modification patterns.

If indeed enhancers are being used as promoters, they should be accompanied by changes in transcript levels or creation of new isoforms based on the creation of a new promoter. We used cufflinks[29] to identify changes at the isoform level for H1 and IMR90 cell-types. We found that among 112 enhancers in H1 that underwent a switch to promoters in IMR90 at a resolution of 500bp, 36% were accompanied by the creation of novel isoforms, and another 34% showed atleast 2-fold increase in the isoform level upon switching to promoters. On the other hand, among the 110 enhancers in IMR90 that were previously promoters in H1, 51% showed loss of the isoform, and 14% showed a 2-fold loss in the isoform level upon switch to enhancer in IMR90. As an example, we show the creation of a novel isoform of ALDH3B1 in IMR90, upon switching of an enhancer in H1 to promoter in IMR90, as indicated by the gain of H3K4me3 in IMR90(Fig.3.3C,black box). The reverse in seen in SLCA12A8, where a transcript in H1 is lost in IMR90, upon conversion of the promoter in H1 to an enhancer  in IMR90 (Fig.3.3D,black box).

In summary we observe a novel class of dual-property elements that can act as enhancers in one cell-type and promoters in another. These elements are often involved in the creation of novel isoforms in the cell-type where the element acts as a promoter. Further, certain cell-types such as lung and skeletal muscle show a significant proportion of enhancers that are derived from promoters in other cell-types.

**Factor involved in the switching between enhancers and promoters**

The observation that the same regulatory sequence can have different functions in different cell-types leads to the question of how this is achieved. In order to understand the mechanism involved in this, we need to identify potential transcriptional regulators that maybe involved in the process. First, we looked for motifs that maybe enriched at each of these elements. For each cell-type, we looked for motifs that were enriched at the enhancers switching to promoters as compared to other enhancers that were active within the cell-type. In Fig.3.4A, the motifs that showed up as significant at enhancers in over 15 cell-types are shown. Similarly, among the set of promoters active in each cell-type, we considered the differential enrichment of motifs at promoters that behave as enhancers in other cell-types and also found a strong enrichment for motifs at these elements. This discovery is noteworthy as promoters are usually quite motif-poor. In Fig3.4B, the motifs that showed up as significant at promoters in over 15 cell-types are shown. Since the motifs in Fig.3.4A,B are present in a majority of cell-types, the associated transcription factors may have a fundamental role to play in the mechanism of switching.

The CTCF motif is found to be enriched in all 32 cell-types for promoters that switch to enhancers as compared to those that do not(Fig.3.4B).It is also significantly enriched at enhancers switching to promoters as compared to those that don’t in over 25 cell-types. This indicates that CTCF maybe playing an important role in the in the specification of dual-property elements. We performed ChIP-seq to identify genome-wide binding sites of CTCF in H1 and IMR90.It can be seen that over 25% dual-property enhancers in both H1 and IMR90 are bound to CTCF, as compared to less than 15% of the remaining enhancers(Fig.3.4C). This difference was found to be highly significant using a hypergeometric distribution(p-value<2.2X10^-308). On the other hand at promoters, there was no significant difference in CTCF binding between promoters with dual property versus those without at a p-value cutoff of 10^-5(Fig.3.4C). This indicates that CTCF binding maybe playing a role in marking enhancers that have the potential to be promoters but not the other way round.

CTCF has been known to bind upstream of CPG island promoters which have a distinct transcription-associated chromatin organization as compared to non-CPG promoters[114].In order to see if our dual property-elements were also enriched within CPG islands, we obtained predictions from the UCSC genome browser[69] made using the sensitive criteria[115]. We found that 19% of enhancers with the potential to act as promoters were enriched within CPG islands as compared to 2% of enhancers that did not. (fig.3.4D). On the other hand, 23% of dual-property promoters overlapped CPG islands as compared to 12% of the remaining promoters(Fig.3.4D).This increased association of dual property-elements with CPG islands was found to be highly significant using a hyper-geometric distribution(p-value<10^-5), but this overlap was not significant for enhancers or promoters that did not have the dual-property(p-value=1).

Another observation of interest was that 11 motifs belonging to the ETS family of transcription factors[116] were enriched at enhancers with the potential to be promoters in other cell-types. While 3 of these motifs, Elk1, Elk4 and Elf1 were observed at dual-property enhancers in over 15 cell-types, the remaining such as ETV1,EWS,ERG,GABPA,FLI-1 etc were enriched at dual-property enhancers in atleast 1 cell-type

A recent study investigated the binding of KDM5C, a histone lysine demethylase that binds to enhancers as well as promoters, allowing for the maintenance of enhancer status, while preventing over-activation of promoters. We compared our motifs at dual-property elements with motifs found at KDM5C sites in the paper (Fig.3.4E) and found that 5 out of the 8 motifs identified at KDM5C binding sites were also enriched at our dual-property elements in atleast 1 cell-type. These motifs included NRF1,SP1 and ELK1 present enriched at over 15 dual-property elements as compared to other enhancers(Fig.3.4A vs

Fig.3.4E), and cMyC enriched at dual-property elements as compared to promoters in over 15 cell-types as well(Fig.3.4B vs Fig.3.4E). In addition GABPA was found to be enriched in 3 cell-types at dual property elements with respect to remaining enhancers. In order to verify that KDM5C was indeed playing a role at dual-property elements, we performed ChIP-seq for KDM5C in H1 and computed the average normalized profile surrounding dual-property elements versus the remaining enhancers and promoters(Fig.3.4F).H1 enhancers that switch to promoters in other cell-types have a significant enrichment for KDM5C within -1 to +1 kb around the enhancer, as compared to the enhancers that do not(Fig.3.4F,panel1,p-value=4.2X10^-75,Wilcoxon test).This may indicate a mechanism to ensure that sequences with the potential to act as promoters are prevented in doing so by the presence of a histone lysine demethylase. On the other hand, dual-property promoters in H1 are not differentially bound by KDM5C(Fig.3.4,panel2,red vs blue).

In summary, dual-property elements have a significant overlap with CPG islands and show binding of CTCF. CTCF binding appears to differentially mark enhancers with the potential to act as promoters. Motifs belonging to the ETS family of transcription factors were strongly enriched at dual-property enhancers. Several motifs enriched at the dual-property elements were seen to be associated with the recruitment of a lysine demethylase, KDM5C. Indeed, significant enrichment of KDM5C binding was observed around enhancers that have the potential to act as promoters as compared to the enhancers that do not show such potential.

## Evolutionary conservation, tissue-specificity and disease

Evolution of regulatory DNA sequences may underlie the morphological diversity of animal species[107,108]. Hence, understanding the association of evolutionary conservation of regulatory sequences with tissue function could be particularly beneficial in understanding what distinguishes humans from other species.

At first, we compared the evolutionary conservation of tissue-specific and ubiquitous regulatory elements to find which type of function arose earlier in evolution. Based on the concept of founder gene formation, a technique called phylostratigraphy has been developed that can assign human genes to their evolutionary strata[117,118] .We used this assignment to determine if tissue-specific promoters were assigned to a particular strata as compared to ubuqitous ones. In Fig.3.5A, 19 phylostrata have been defined based on the species assigned to each of them[118]. Tissue-specificity at genes assigned to each of these strata was measured using enrichment of either H3K4me3 (panel1, Fig3.5A) or H3K27ac(panel 2,Fig3.5A) a the 26 human tissues.The strata showing most significant enrichment of ubiquitous promoters (promoters with highest fraction of tissues in which it  shows enrichmhent of an active mark) as compared to all other strata was strata 2 or the eukaryota stage(Fig.3.5A), whether the active mark considered was H3K4me3(Wilcoxon test p-value<10^-169) or H3K27ac(Wilcoxon test p-value<10^-182). The strata showing most significant enrichment of tissue-specific promoters (promoters with lowest fraction of tissues in which it shows enrichment of an active mark) as compared to all other strata was strata 19 or the primates stage.(H3K27ac:p-value<10^-175;H3K4me3:p-value<10^-186). A general trend of progressive increase of the median tissue-specificity of promoters is seen between eukaryota and mammalia stages, when tissue-specificity is measured as the inverse of the fraction of tissues in which a promoter shows enrichment of H3K27ac(panel 2,Fig3.5A).

Next, we measured evolutionary conservation at predicted enhancers and promoters using a phastCons measure based on sequence comparisons of 44 vertebrates[119,120]. Tissue-specific enhancers or promoters were defined as those that showed enrichment of H3K27ac in atmost 2 among the 32 cell-types, Ubiquitous promoters were defined as those that showed enrichment of H3K27ac in all cell-types while ubiquitous enhancers were those that showed enrichment in over 15 cell-types. We computed the average phastCons score from -0.5 to +0.5kb surrounding either promoter or enhancer and examined the distribution of the scores for tissue-specific(red) and ubiquitous categories(Fig.3.5B,C).It is quite clear in promoters, that the distribution of tissue-specific TSS are biased towards the lower end of the phastCons scores, while the ubiquitous promoters are biased towards the higher end of the phastCons scores(Fig.3.5B,red vs blue). In enhancers on the other hand, the same trends are seen for phastCons scores less than 400 with the tissue-specific enhancers being biased towards the lower conservation end. However, there is a distinct enrichment of tissue-specific enhancers at high phastCons scores between 400 and 600 that is absent in ubiquitous enhancers. We performed Fast k-means++ clustering[84] and observed separation of the high conservation category described above from the rest of the enhancers at the number of clusters k equal to 2. We examined enrichment of various tissue-specific enhancers within that category, and found that enhancers specific to the early lineages H1, mesendoderm and neuroprogenitor cells as well as to the brain were enriched in the highly conserved category at p-value <10^-3 using a hypergeometric distribution. Skeletal muscle and stomach muscle were also significantly enriched in this highly conserved category. Enhancers specific to multiple tissues were observed to have significantly lower conservation levels than average. We found left and right ventricles of the heart to be among these tissues with significantly lower conservation levels(Wilcoxon test p-value<10^-3), which is in keeping with previous observations of heart enhancers being weakly conserved[121].

Diseases are often studied using animal models. Hence, the extent to which the causative mutations related to a genetic disease maybe recapitulated in an animal model is a question of great interest. We begin to probe this question by considering the distribution of SNPs(single nucletide polymorphisms) associated with disease based on genome-wide association studies(GWAS) [122] within the tissue-specific enhancers. In Fig.3.5D, it is seen that tissue-specific enhancers containing GWAS SNPs are biased towards lower levels of phastCons scores as compared to the tissue-specific enhancers without these SNPs. Enhancers with the GWAS SNPs also lack the distinctive enrichment between phastCons scores 400 to 600 that is observed with other tissue-specific enhancers(Fig.3.5D). On the whole, it appears that disease-associated SNPs localize within lowly conserved enhancers and may have a tendency to be more human-specific.

Overall, we found that tissue-specific enhancers and promoters are less conserved than ubiquitously active ones. However, a distinct category of highly conserved tissue-specific enhancers exists that is enriched for enhancers specific to early lineages, brain and muscle tissues. Phylostratigraphic analysis revealed that ubiquitous promoters probably evolved during the eukaryota stage while the tissue-specific ones arose in the later stages with the maximum concentration in the primate stage. Tissue-specific enhancers associated with disease mutations appear to be lowly conserved, possibly arising in human-specific enhancers.

## Discussion

In this study, we predicted enhancers and promoters in 26 human primary tissues and 6 cell-lines, including 5 early developmental lineages. Further we identified enhancers that were highly specific to each of these cell-types and provided evidence for the accuracy of our predictions by examining enrichment of motifs at tissue-specific enhancers, which associated with terms that would be expected to belong to the tissue type based on previous literature.

We discovered a novel class of dual property cis-regulatory elements that could function as enhancers in one cell-type and promoters in another. We verified the observation by noting that several novel transcripts and isoforms arose from the cell-type in which the element was a promoter, but did not exist or were present at low levels in the cell-type that was an enhancer. Further validations that the dual property elements can function as enhancers will be carried out using reporter assays[24]. We discovered the association of such elements with CTCF and CPG islands as well as histone demethylase, KDM5C. Many binding motifs for the ETS family of transcription factors were found to be enriched in the dual-property elements. The ETS family has a lot of redundancy in it’s binding sites and mainly achieves specificity through it’s binding partner[123]. This leads to a multitude of co-regulators that can drive gene-specific responses in many different cell-types[116]. As a next step, we should validate the localization of these ETS transcription factors at dual-property elements using ChIP-seq, as well as find their binding partners in the enhancer as well as promoter state. A comprehensive examination of other components involved in this mechanism maybe carried out by performing ChIP-seq for the various motifs discovered. Mass spectrometry and co-immunoprecipitation experiments maybe performed to find the components of the various complexes that are bound to the dual-property elements both in their enhancer as well as promoter states.

Phylostratigraphic analysis[117] revealed that ubiquitous promoters probably evolved during the eukaryota stage while the tissue-specific ones arose in the later stages with the maximum concentration in the primate stage. Overall, we found that tissue-specific enhancers and promoters are less conserved than ubiquitously active ones. However, a distinct category of highly conserved tissue-specific enhancers exists that is enriched for enhancers specific to early lineages, brain and muscle tissues. Tissue-specific enhancers associated with disease mutations appear to be lowly conserved, possibly arising in human-specific enhancers. Further analysis will be carried out to determine which diseases have SNPs localized within lowly conserved enhancers and which of them have SNPS within the highly conserved enhancers. This would be coupled with the knowledge of the cell-type in which these enhancers are active and potentially any known transcription factor binding motif that the disease-causing mutation maybe disrupting.

## Methods

Data normalization of histone modifications was as described in chapter 1.

Enhancer and promoter predictions were carried out as described in chapter 1 and 2 respectively using the RFECS algorithm trained on p300 binding sites in H1 with 6 histone modifications as features – H3K27ac, H3K4me1, H3K4me3, H3K27me3, H3K36me3 and H3K9me3.

Presence or absence of a modification at a regulatory element (enhancer or promoter) in a particular cell-type was computed by clustering all the enhancers or promoters predicted in 32 cell-types. The input feature was the normalized RPKM(reads per kilobase per million) level in 100 bp bins of the modification from -2 to +2 kb around the regulatory element We performed fast k-means++ clustering[84] on the list of all predicted regulatory elements and discovered the optimal number of clusters using the Davies-Bouldin measure[92]. All elements within the cluster that had the maximum positive enrichment of the modification were assigned a “present” value. In order to measure the tissue-specificity of a particular element, H3K27ac was used as the active modification. Here, each enhancer or promoter was assigned a “present” or “absent” value based on the method above. If an enhancer or promoter had H3K27ac present in atmost 2 cell-types of the 32 investigated, we considered this as tissue-specific to the cell-types in which it had H3K27ac present.

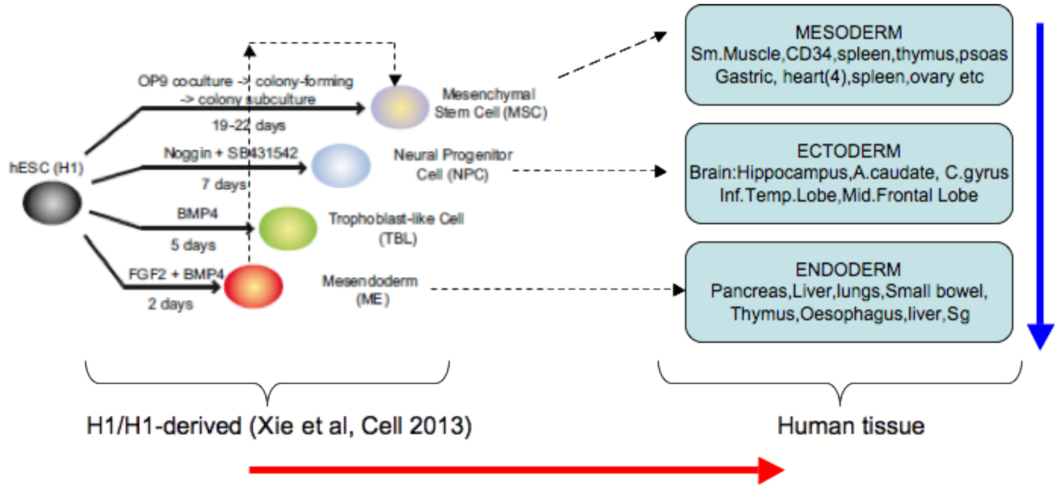
OP9 coculture -> colony-forming
-> colony subculture
19-22 days
Mesenchymal
Stem Cell (MSC)
hESC (H1)
Noggin + SB431542
7 days
Neural Progenitor
Cell (NPC)
BMP4
5 days
Trophoblast-like Cell
(TBL)
FGF2 + BMP4
2 days
Mesendoderm
(ME)
MESODERM
Sm.Muscle,CD34,spleen,thymus,psoas
Gastric, heart(4),spleen,ovary etc
ECTODERM
Brain:Hippocampus,A.caudate, C.gyrus
Inf.Temp.Lobe,Mid.Frontal Lobe
ENDODERM
Pancreas,Liver,lungs,Small bowel,
Thymus,Oesophagus,liver,Sg
H1/H1-derived (Xie et al, Cell 2013)
Human tissue

**Figure 4.1. Cell-types and tissues showing stage of development and germ-layer**

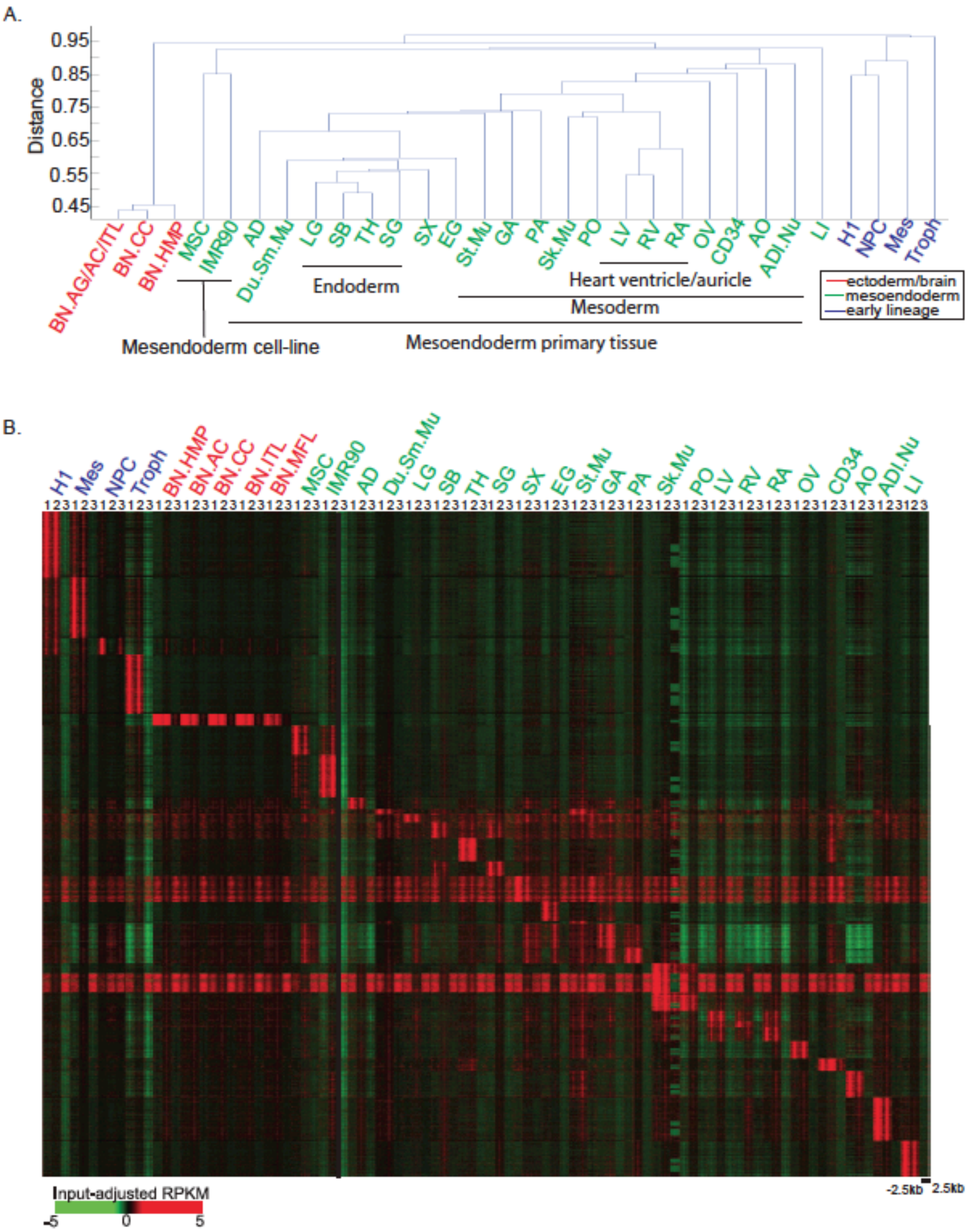

**Figure 4.2. Tissue-specific enhancers and similarity of cell-types**
A.) Hierarchical clustering of cell-lines and tissues based on overlap of strong enhancers predicted in each cell-type. B.) Heatmap showing enrichment of H3K27ac, H3K4me1 and H3K4me3 at tissue-specific enhancers in each cell-type.

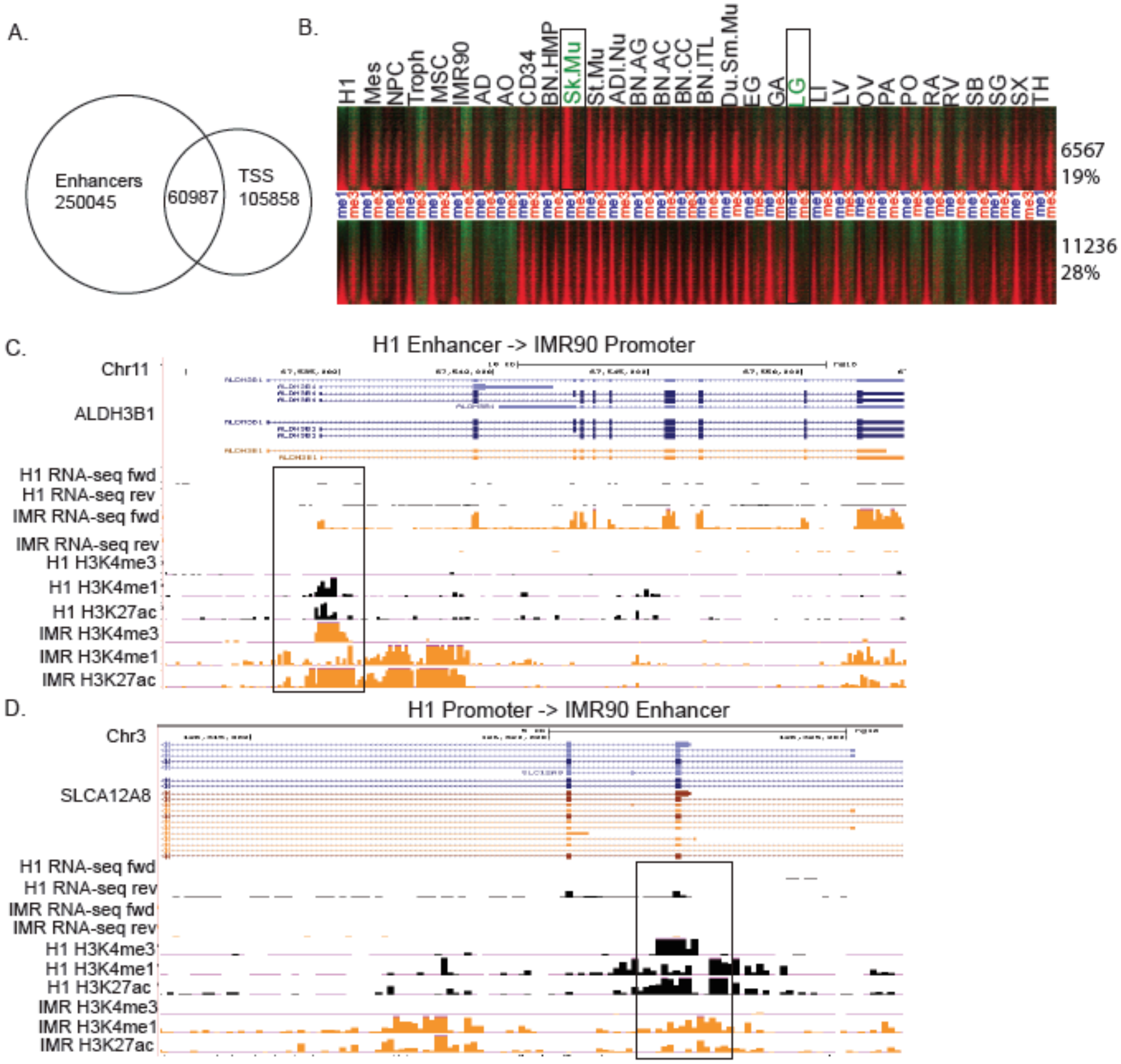

**Figure 4.3. Identification of dual property cis-regulatory elements**

A.) Overlap between predicted enhancers and promoters across 32 cell-types. B.) Heatmap showing enrichment of H3K4me1 and H3K4me3 across 32 cell-types at enhancers in Skeletal muscle(panel 1) and lung tissue(panel 2) that are predicted as promoters in other cell-types. C,D.) UCSC genome browser snapshot showing histone modification changes accompanying the switching of C.) an enhancer in H1 to promoter in IMR90 accompanied by creation of a novel isoform of gene ALDH3B1. D.) a promoter in H1 to enhancer in IMR90 accompanied by loss of expression of an isoform of gene SLCA12A8.

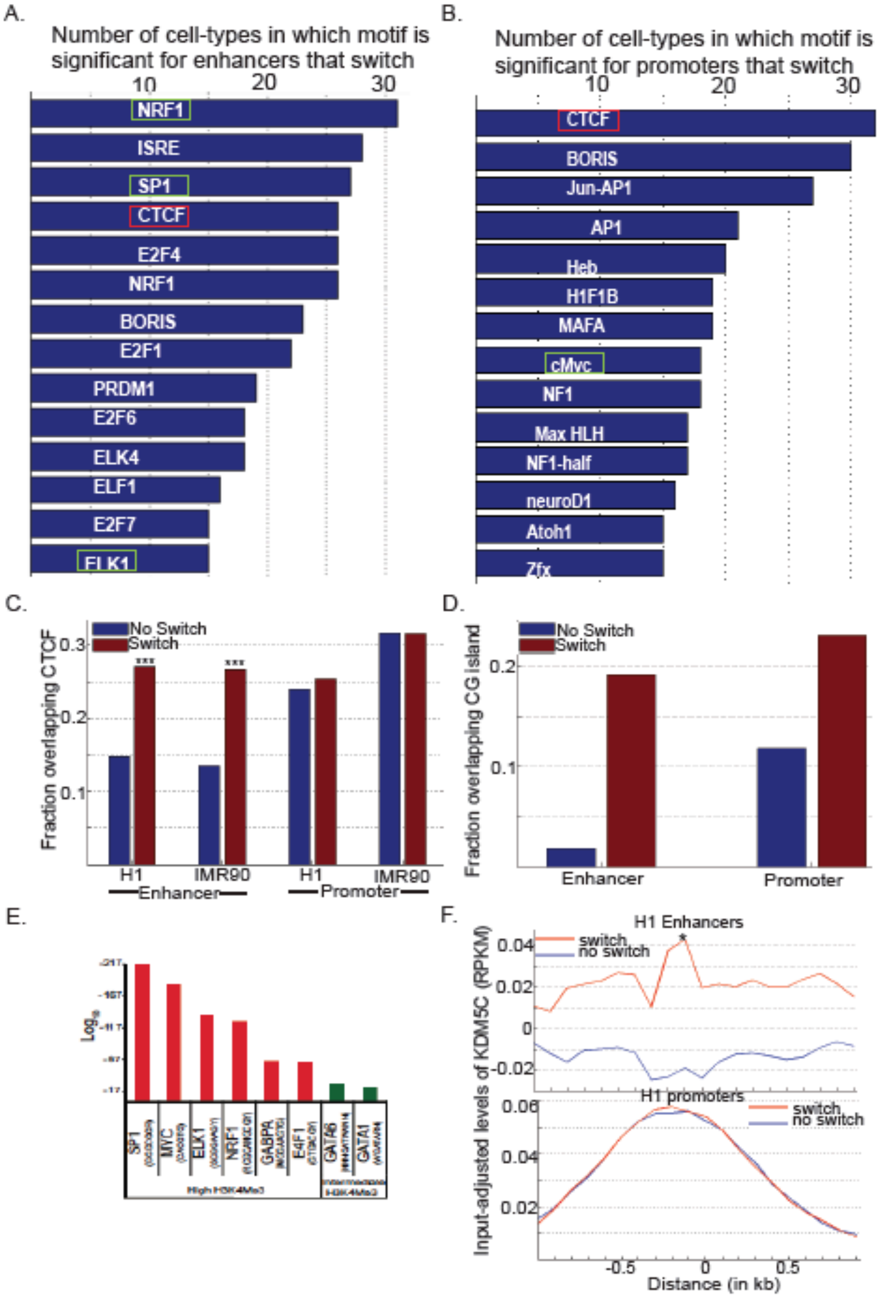

**Figure 4.4. Factors involved in the functioning of dual property cis-regulatory elements**
A,B.) Number of cell-types in which a motif is significant for A.) an enhancer that can behave as promoter as compared to a background of enhancers that cannot. B.) a promoter that can behave as enhancer as compared to a background of promoters that cannot. C.) Fraction of dual-property elements(red) in H1 and IMR90 that are bound by CTCF as compared to remaining enhancers or promoters(blue).D.) Fraction of dual-property elements(red) in all 32 cell-types that overlap CPG islands as compared to remaining enhancers or promoters(blue).E.) Motifs found to be enriched at KDM5C binding sites in mouse embryonic stem cells with varying levels of H3K4me3[124].F.) Input-adjusted levels of KDM5C binding (in reads per kilobase per million) at dual-property elements(red) in H1 as compared to remaining enhancers(panel1,blue) or promoters(panel2,blue).

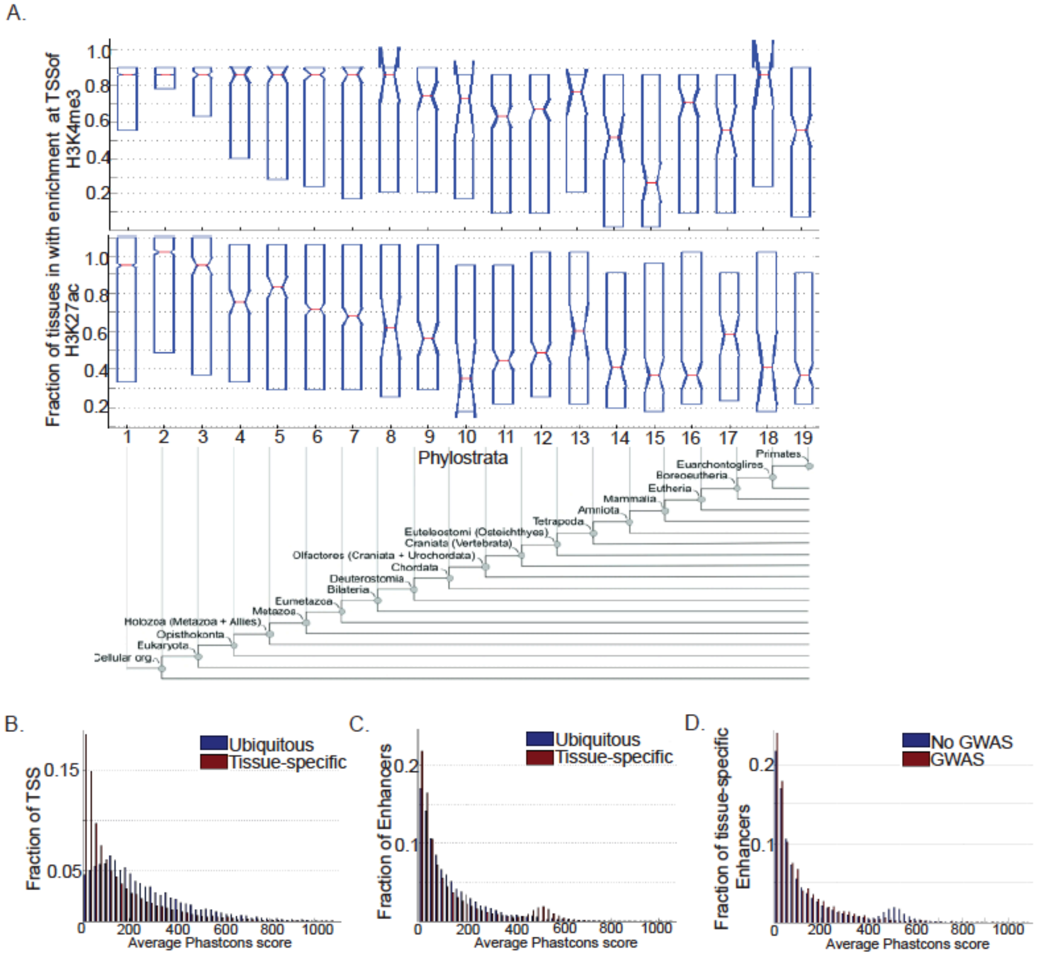

**Figure 4 5. Association of evolutionary conservation with tissue-specificity and disease**
A.) Tissue-specificity of TSS associated with each evolutionary strata assigned using the phylostratigraphy method(ref,panel 3). Tissue specificity maybe defined as the inverse of the fraction of tissues in which the TSS is active based on either enrichment of H3K4me3(panel 1) or H3K27ac(panel 2). B,C.) Differences in distribution of average vertebrate phastcons score between B.)enhancers and C.) TSS that are either ubiquitous and those that are tissue-specific as measured based on the enrichment of H3K27ac around the element. D.) Differences in distribution of average vertebrate phastcons score between tissue-specific enhancers containing GWAS SNPs and those lacking GWAS SNPs.

**Table 4.1. Abbreviations for the 32 cell-types**

| Abbrev | Full name |
|---|---|
| H1 | Embryonic stem cell |
| Mes | Mesendoderm |
| NPC | Neural progenitor cell |
| Tro | Trophoblast |
| MSC | Mesenchymal stem cell |
| IMR90 | Fetal lung fibroblast |
| AD | Adrenal Gland |
| AO | Aorta |
| CD34 | CD34+ blood cells |
| BN.HMP | Brain hippocampus |
| Sk.Mu | Skeletal muscle |
| St.Mu | Stomach muscle |
| ADI.Nu | Adipose nuclei |
| BN.AG | Brain An |
| BN.AC | Brain anterior caudate |
| BN.CC | Brain cerebral cortex |
| BN.ITL | Brain |
| Duo.Sm.Mu | Duodenum smooth muscle |
| EG | Oesophagus |
| GA | Gastric |
| LG | Lung |
| LI | Liver |
| LV | Left ventricle |
| OV | Ovary |
| PA | Pancreas |
| PO | Psoas |
| RA | Right auricle |
| RV | Right ventricle |
| SB | Small Bowel |
| SG | Sigmoid tissue |
| SX | Spleen |
| TH | Thymus |

**Table 4.2. Predicted enhancers and promoters in 32 cell-types**

| | Total | Tissue-specific |
|---|---|---|
| Enhancers | 311032 | 126950 |
| Promoters | 174805 | 13218 |

## ACKNOWLEDGEMENTS

Chapter 4 in part, is being prepared for submission for publication. Nisha Rajagopal, Danny Leung, Wei Xie, Pradipta Ray, Ah Young Lee, Chia-An Yen, Michael Q.Zhang, Bing Ren. Dynamic epigenetic signatures at regulatory elements in human tissues. The dissertation author was one of two primary investigators and author of this material.

# Chapter 5.Future Directions

In this thesis, we developed a random-forest based algorithm with two-fold advantage. First, we were able to accurately predict genome-wide enhancers and promoters from chromatin modifications. Second, we were able to identify the most informative set of modifications required for characterizing any genomic element. The latter enabled us not only to find the minimal set of modifications required to predict enhancers, promoters and gene bodies, but also helped elucidate the distinctive localization of histone acetylations at each of these regions.

For purposes of enhancer prediction, we used p300 binding sites as a training-set. Comparisons with using CBP, another well-known enhancer-binding coactivator[125], in CD4+ T-cells showed no significant difference in the prediction of genome-wide enhancers(data not shown). While these two co-activators are generally considered to be representative of genome-wide enhancers, there are also other transcriptional co-activators that are known to be part of enhancer binding complexes[126]. In addition to p300 and CBP, these could potentially be used to train the RFECS classifier. We may potentially discover other classes of enhancers with different patterns of modifications.

Recent studies have shown the existence of a class of enhancers termed “super-enhancers” that cover much larger genomic regions than regular enhancers and have a much higher density of transcription factor binding density.

They are characterized by binding of the mediator complex and seem to play key roles in the control of mammalian cell identity[127,128]. In our method of enhancer prediction, we defined a point location as the peak of a predicted enhancer, by assuming any region within a -2 to +2 kb window that was predicted as an enhancer, could potentially be part of the same enhancer[23]. Extending the algorithm to include the width of the enhancers in the output could be useful, as it may allow us to define super-enhancers that that are involved in the control of cellular state in a multitude of cell-types. This could potentially be achieved by measuring the density of transcription factor binding along the genome by integrating additional sources of data such as cell-type specific transcription-factor binding and DNase-I hypersensitivity. Further, the width of histone modification domains associated with enhancers such as H3K4me1 and H3K27ac[24,102] could also associated with the density of binding of known key regulators of the cellular state.

Recent years have shown that enhancers maybe transcribed to produce short bidirectional transcripts, called eRNAs[39] which may or may not be polyadenylated[105]. A study in mouse also showed the production of elongated polyA+ transcripts from intragenic enhancers[106]. It would be of interest to use the RFECS algorithm to find if there are distinctive chromatin modification profiles associated with transcription at enhancers.

A major finding of our study was the association of histone acetylation patterns with various genomic elements. In particular, we found that retention of exon-intron junctions was significantly associated with the presence of histone acetylations. However, such a pattern of modifications could be pre-marking splice sites that are currently constitutive but maybe alternatively spliced in later developmental cell-types. It would be valuable to understand to what extent this pre-marking can occur. We can do this by comparing changes in splice site retention across multiple cell-types of early and late lineages with matched chromatin and RNA-seq datasets. Further, adding stimulus to a particular cell-type and measuring transcriptomic as well as epigenomic changes simultaeneously, could enable the association of changes in chromatin modification patterns with changes in splice-site usage, at different genomic locations.

Upon comparing enhancers and promoters across a large panel of 26 primary tissues and 6 cell-lines, we discovered a significant proportion of enhancers were promoters in other lineages. Further, we found certain characteristics associated with this class such as the presence of CPG islands, binding of CTCF and possible enrichment of KDM5C. We also obtained lists of other motifs of transcription factors that maybe potentially involved. The next stage in this study would be to understand the mechanism by which such a switch occurs. This might be learned by knocking out various factors suspected to be involved in facilitating the switch and determining which of them are necessary. Either by applying a stimulus to a particular cell-type, or by studying various stages of a developmental pathway, we may even be able to obtain a temporal dimension to the transition of enhancer to promoter state which could further enable understanding the order of recruitment of various factors involved in the process.

# References

1. Bird A (2007) Perceptions of epigenetics. Nature 447: 396-398.

2. Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G, Ecker JR (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res.

3. Burgess DJ (2013) Epigenetics: Mechanistic insight into epigenetic inheritance. Nat Rev Genet 14: 442.

4. Bonasio R, Tu S, Reinberg D (2010) Molecular signals of epigenetic states. Science 330: 612-616.

5. Saze H (2012) Transgenerational inheritance of induced changes in the epigenetic state of chromatin in plants. Genes Genet Syst 87: 145-152.

6. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251-260.

7. Grant PA (2001) A tale of histone modifications. Genome Biol 2: REVIEWS0003.

8. Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, Lu Z, Ye Z, Zhu Q, Wysocka J, Ye Y, Khochbin S, Ren B, Zhao Y (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 146: 1016-1028.

9. Mersfelder EL, Parthun MR (2006) The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. Nucleic Acids Res 34: 2653-2662.

10. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693-705.

11. Ruthenburg AJ, Li H, Patel DJ, Allis CD (2007) Multivalent engagement of chromatin modifications by linked binding modules. Nat Rev Mol Cell Biol 8: 983-994.

12. Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat Struct Mol Biol 14: 1025-1040.

13. Wysocka J, Swigut T, Xiao H, Milne TA, Kwon SY, Landry J, Kauer M, Tackett AJ, Chait BT, Badenhorst P, Wu C, Allis CD (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature 442: 86-90.

14. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

15. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22: 1813-1831.

16. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9: e1001046.

17. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28: 1045-1048.

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

19. Hon G, Ren B, Wang W (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol 4: e1000201.

20. Teng L, Tan K (2012) Finding combinatorial histone code by semi-supervised biclustering. BMC Genomics 13: 301.

21. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.

22. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223-227.

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

24. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

25. Won KJ, Chepelev I, Ren B, Wang W (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC Bioinformatics 9: 547.

26. Firpi HA, Ucar D, Tan K (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 26: 1579-1586.

27. Fernandez M, Miranda-Saavedra D (2012) Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res.

28. Kawaji H, Severin J, Lizio M, Forrest AR, van Nimwegen E, Rehli M, Schroder K, Irvine K, Suzuki H, Carninci P, Hayashizaki Y, Daub CO (2011) Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. Nucleic Acids Res 39: D856-860.

29. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7: 562-578.

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

31. Zentner GE, Henikoff S (2013) Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol 20: 259-266.

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

33. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190-1195.

34. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen C-a, Klugman S, Yu P,

Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung W-Y, Chi NC, Antosiewicz-Bourget JE, Slukvin I, Stewart R, Zhang MQ, Wang W, Thomson JA, Ecker JR, Ren B (2013) Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell.

35. Levine M (2010) Transcriptional enhancers in animal development and evolution. Curr Biol 20: R754-763.

36. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EE (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat Genet 44: 148-156.

37. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457: 854-858.

38. Jin F, Li Y, Ren B, Natarajan R (2011) PU.1 and C/EBP(alpha) synergistically program distinct response to NF-kappaB activation through establishing monocyte specific enhancers. Proc Natl Acad Sci U S A 108: 5290-5295.

39. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465: 182-187.

40. Won KJ, Agarwal S, Shen L, Shoemaker R, Ren B, Wang W (2009) An integrated approach to identifying cis-regulatory modules in the human genome. PLoS One 4: e5501.

41. Gonzalez S, Montserrat-Sentis B, Sanchez F, Puiggros M, Blanco E, Ramirez A, Torrents D (2012) ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. Bioinformatics 28: 763-770.

42. Busser BW, Taher L, Kim Y, Tansey T, Bloom MJ, Ovcharenko I, Michelson AM (2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. PLoS Genet 8: e1002531.

43. Girgis HZ, Ovcharenko I (2012) Predicting tissue specific cis-regulatory modules in the human genome using pairs of co-occurring motifs. BMC Bioinformatics 13: 25.

44. Meireles-Filho AC, Stark A (2009) Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. Curr Opin Genet Dev 19: 565-570.

45. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA (2008) Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40: 158-160.

46. Heintzman ND, Ren B (2009) Finding distal regulatory elements in the human genome. Curr Opin Genet Dev 19: 541-549.

47. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Afzal V, Simpson PC, Rubin EM, Black BL, Bristow J, Pennacchio LA, Visel A (2012) Large-scale discovery of enhancers from human heart tissue. Nat Genet 44: 89-93.

48. Janknecht R, Hunter T (1996) Versatile molecular glue. Transcriptional control. Curr Biol 6: 951-954.

49. Panne D (2008) The enhanceosome. Curr Opin Struct Biol 18: 236-242.

50. Korzus E, Torchia J, Rose DW, Xu L, Kurokawa R, McInerney EM, Mullen TM, Glass CK, Rosenfeld MG (1998) Transcription factor-specific requirements for coactivators and their acetyltransferase functions. Science 279: 703-707.

51. He A, Kong SW, Ma Q, Pu WT (2011) Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci U S A 108: 5632-5637.

52. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 21: 456-464.

53. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311-322.

54. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S,

Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

55. Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, James KD, Lefebvre GC, Bruce AW, Dovey OM, Ellis PD, Dhami P, Langford CF, Weng Z, Birney E, Carter NP, Vetrie D, Dunham I (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res 17: 691-707.

56. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817-825.

57. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B (2012) A map of the cis-regulatory sequences in the mouse genome. Nature.

58. Breiman L (2001) Random Forests. Machine Learning 45: 5-32.

59. Zhang C (2012) Ensemble machine learning : methods and applications. New York: Springer.

60. Bylander T (2002) Estimating generalization error on two-class datasets using out-of-bag estimates. Machine Learning 48: 287-297.

61. Lemmond TD, Chen BY, Hatch AO, Hanley WG (2010) An Extended Study of the Discriminant Random Forest

Data Mining. In: Stahlbock R, Crone SF, Lessmann S, editors: Springer US. pp. 123-146.

62. Do T-N, Lenca P, Lallich S, Pham N-K (2010) Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees

Advances in Knowledge Discovery and Management. In: Guillet F, Ritschard G, Zighed D, Briand H, editors: Springer Berlin / Heidelberg. pp. 39-55.

63. Spackman KA. Signal detection theory: Valuable tools for evaluating inductive learning; 1989; San Mateo, CA. Morgan Kauffman. pp. 160-163.

64. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.

65. Cotney J, Leng J, Oh S, Demare LE, Reilly SK, Gerstein MB, Noonan JP (2012) Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. Genome Res 22: 1069-1080.

66. Teng L, Firpi HA, Tan K (2011) Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. Nucleic Acids Res 39: 7371-7379.

67. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

68. Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W, Zhao K (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. Cell 138: 1019-1031.

69. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

70. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods 3: 511-518.

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

72. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S, Antosiewicz-Bourget J, Ye Z, Espinoza C, Agarwahl S, Shen L, Ruotti V, Wang W, Stewart R, Thomson JA, Ecker JR, Ren B (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 6: 479-491.

73. Saldanha AJ (2004) Java Treeview--extensible visualization of microarray data. Bioinformatics 20: 3246-3248.

74. Dinarello CA, Fossati G, Mascagni P (2011) Histone deacetylase inhibitors for treating a spectrum of diseases not related to cancer. Mol Med 17: 333-352.

75. Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, Durham T, Zhang X, Donaghey J, Epstein CB, Regev A, Bernstein BE (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell 147: 1628-1639.

76. Listerman I, Sapra AK, Neugebauer KM (2006) Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. Nat Struct Mol Biol 13: 815-822.

77. Lynch KW (2006) Cotranscriptional splicing regulation: it's not just about speed. Nat Struct Mol Biol 13: 952-953.

78. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet 41: 376-381.

79. Hon G, Wang W, Ren B (2009) Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol 5: e1000566.

80. Hnilicova J, Hozeifi S, Duskova E, Icha J, Tomankova T, Stanek D (2011) Histone deacetylase activity modulates alternative splicing. PLoS One 6: e16727.

81. Gunderson FQ, Merkhofer EC, Johnson TL (2011) Dynamic histone acetylation is critical for cotranscriptional spliceosome assembly and spliceosomal rearrangements. Proc Natl Acad Sci U S A 108: 2004-2009.

82. Zhou HL, Hinman MN, Barron VA, Geng C, Zhou G, Luo G, Siegel RE, Lou H (2011) Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. Proc Natl Acad Sci U S A 108: E627-635.

83. Wu J, Akerman M, Sun S, McCombie WR, Krainer AR, Zhang MQ (2011) SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics 27: 3010-3016.

84. Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. New Orleans, Louisiana: Society for Industrial and Applied Mathematics. pp. 1027-1035.

85. Gatta R, Dolfini D, Zambelli F, Imbriano C, Pavesi G, Mantovani R (2011) An acetylation-mono-ubiquitination switch on lysine 120 of H2B. Epigenetics 6: 630-637.

86. Hwang WW, Venkatasubrahmanyam S, Ianculescu AG, Tong A, Boone C, Madhani HD (2003) A conserved RING finger protein required for histone H2B monoubiquitination and cell size control. Mol Cell 11: 261-266.

87. Zhu B, Zheng Y, Pham AD, Mandal SS, Erdjument-Bromage H, Tempst P, Reinberg D (2005) Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation. Mol Cell 20: 601-611.

88. Lau OD, Courtney AD, Vassilev A, Marzilli LA, Cotter RJ, Nakatani Y, Cole PA (2000) p300/CBP-associated factor histone acetyltransferase processing of a peptide substrate. Kinetic analysis of the catalytic mechanism. J Biol Chem 275: 21953-21959.

89. Cho H, Orphanides G, Sun X, Yang XJ, Ogryzko V, Lees E, Nakatani Y, Reinberg D (1998) A human RNA polymerase II complex containing factors that modify chromatin structure. Mol Cell Biol 18: 5355-5363.

90. Jeong KW, Kim K, Situ AJ, Ulmer TS, An W, Stallcup MR (2011) Recognition of enhancer element-specific histone methylation by TIP60 in transcriptional activation. Nat Struct Mol Biol 18: 1358-1365.

91. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 38: 1341-1347.

92. Davies DLB, D.W. (1979) A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 224-227.

93. Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12: 283-293.

94. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y,

Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR (2012) Landscape of transcription in human cells. Nature 489: 101-108.

95. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T, Frampton GM, Drake AC, Leskov I, Nilsson B, Preffer F, Dombkowski D, Evans JW, Liefeld T, Smutko JS, Chen J, Friedman N, Young RA, Golub TR, Regev A, Ebert BL (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144: 296-309.

96. Gifford CA, Ziller MJ, Gu H, Trapnell C, Donaghey J, Tsankov A, Shalek AK, Kelley DR, Shishkin AA, Issner R, Zhang X, Coyne M, Fostel JL, Holmes L, Meldrim J, Guttman M, Epstein C, Park H, Kohlbacher O, Rinn J, Gnirke A, Lander ES, Bernstein BE, Meissner A (2013) Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. Cell 153: 1149-1163.

97. Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. Nat Rev Genet 14: 204-220.

98. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, Bennett DA, Houmard JA, Muoio DM, Onder TT, Camahort R, Cowan CA, Meissner A, Epstein CB, Shoresh N, Bernstein BE (2013) Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell 152: 642-654.

99. Shu W, Chen H, Bo X, Wang S (2011) Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. Nucleic Acids Res 39: 7428-7443.

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

101. Sims RJ, 3rd, Nishioka K, Reinberg D (2003) Histone lysine methylation: a signature for chromatin function. Trends Genet 19: 629-639.

102. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107: 21931-21936.

103. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470: 279-283.

104. Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. Annu Rev Biochem 72: 449-479.

105. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8: e1000384.

106. Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, McGowan SJ, De Gobbi M, Hosseini M, Vernimmen D, Brown JM, Gray NE, Collavin L, Gibbons RJ, Flint J, Taylor S, Buckle VJ, Milne TA, Wood WG, Higgs DR (2012) Intragenic enhancers act as alternative promoters. Mol Cell 45: 447-458.

107. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8: 206-216.

108. Liao BY, Weng MP (2012) Natural selection drives rapid evolution of mouse embryonic heart enhancers. BMC Syst Biol 6 Suppl 2: S1.

109. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576-589.

110. Benod C, Vinogradova MV, Jouravel N, Kim GE, Fletterick RJ, Sablin EP (2011) Nuclear receptor liver receptor homologue 1 (LRH-1) regulates pancreatic cancer cell growth and proliferation. Proc Natl Acad Sci U S A 108: 16927-16931.

111. Kawabe S, Yazawa T, Kanno M, Usami Y, Mizutani T, Imamichi Y, Ju Y, Matsumura T, Orisaka M, Miyamoto K (2013) A novel isoform of liver receptor homolog-1 is regulated by steroidogenic factor-1 and the specificity protein family in ovarian granulosa cells. Endocrinology 154: 1648-1660.

112. Liu T, Zhang S, Xiang D, Wang Y (2013) Induction of hepatocyte-like cells from mouse embryonic stem cells by lentivirus-mediated constitutive expression of Foxa2/Hnf4a. J Cell Biochem.

113. Jensen B, Wang T, Christoffels VM, Moorman AF (2013) Evolution and development of the building plan of the vertebrate heart. Biochim Biophys Acta 1833: 783-794.

114. Vavouri T, Lehner B (2012) Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biol 13: R110.

115. Bock C, Walter J, Paulsen M, Lengauer T (2007) CpG island mapping by epigenome prediction. PLoS Comput Biol 3: e110.

116. Sharrocks AD (2001) The ETS-domain transcription factor family. Nat Rev Mol Cell Biol 2: 827-837.

117. Domazet-Loso T, Brajkovic J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23: 533-539.

118. Domazet-Loso T, Tautz D (2008) An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol 25: 2699-2707.

119. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

120. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493-496.

121. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Bristow J, Ren B, Black BL, Rubin EM, Visel A, Pennacchio LA (2010) ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet 42: 806-810.

122. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

123. Verger A, Duterque-Coquillaud M (2002) When Ets transcription factors meet their partners. Bioessays 24: 362-370.

124. Outchkourov NS, Muino JM, Kaufmann K, van Ijcken WF, Groot Koerkamp MJ, van Leenen D, de Graaf P, Holstege FC, Grosveld FG, Timmers HT (2013) Balancing of histone H3K4 methylation states by the Kdm5c/SMCX histone demethylase modulates promoter and enhancer function. Cell Rep 3: 1071-1079.

125. Ogryzko VV, Schiltz RL, Russanova V, Howard BH, Nakatani Y (1996) The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell 87: 953-959.

126. Naar AM, Lemon BD, Tjian R (2001) Transcriptional coactivator complexes. Annu Rev Biochem 70: 475-501.

127. Loven J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 153: 320-334.

128. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153: 307-319.