Computational inference of H3K4me3 and H3K27ac domain length

Julian Zubek[1,2], Michael Stitzel[3], Duygu Ucar[3]*, Dariusz Plewczynski[1]*

[1] Centre of New Technologies, University of Warsaw, Poland

[2] Institute of Computer Science, Polish Academy of Sciences, Poland

[3] The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, Unites States of America

Corresponding authors*:

Dariusz Plewczynski
Centre of New Technologies, University of Warsaw, Poland
e-mail: d.plewczynski@cent.uw.edu.pl

Duygu Ucar,
The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, Unites States of America
e-mail: Duygu.ucar@jax.org

Abstract

Background. Recent epigenomic studies have shown that the length of a DNA region covered by an epigenetic mark is not just a byproduct of the assaying technologies and has functional implications for that locus. For example, expanded regions of DNA sequences that are marked by enhancer-specific histone modifications, such as acetylation of histone H3 lysine 27 (H3K27ac) domains coincide with cell-specific enhancers, known as super or stretch enhancers. Similarly, promoters of genes critical for cell-specific functions are marked by expanded H3K4me3 domains in the cognate cell type, and these can span DNA regions from 4-5kb up to 40-50kb in length. These expanded H3K4me3 domains are known as buffer domains or super promoters.

Methods. To ask what correlates with—and potentially regulates—the length of loci marked with these two important histone marks, H3K4me3 and H3K27ac, we built Random Forest regression models. With these models we computationally identified genomic and epigenomic patterns that are predictive for the length of these marks in seven ENCODE cell lines.

Results. We found that certain epigenetic marks and transcription factors explain the variability of the length of H3K4me3 and H3K27ac marks across different cell types, which implies that the lengths of these two epigenetic marks are tightly regulated in a given cell type. Our source code for the regression models and data can be found at our GitHub page: https://github.com/zubekj/broad_peaks

Discussion. Our Random Forest based regression models enabled us to estimate the individual contribution of different epigenetic marks and protein binding patterns to the length of H3K4me3 and H3K27ac deposition patterns; therefore potentially revealing genomic signatures at cell specific regulatory elements.

Introduction

Epigenomics refer to the heritable changes that are not stemming from the changes in the genomic DNA sequence. The major epigenetic mechanisms can be categorized into two distinct classes: 1) methylation of the cytosine residues of DNA (i.e., DNA methylation); and 2) the post-translational modifications of the histone proteins (i.e., histone modifications). Recent advancements in next generation sequencing (NGS) technologies have enabled the development of experimental methods to generate genome-wide profiles of epigenetic marks in many types of cells and organisms. These technologies enable profiling epigenetic states genomewide as collections of short sequence reads associated with a given epigenetic mark, which is then aligned to the reference genome. Each locus in the genome has therefore an associated read count, where regions with high abundance of reads (i.e., peaks) are captured to identify loci marked by the profiled epigenetic mark. It has been previously shown that the length of a particular locus covered by an epigenetic mark is related to the cell-specific functions of that region. For example, several research groups observed that long genomic segments marked by enhancer-specific histone modifications, such as acetylation of histone H3 lysine 27 (H3K27ac), are found at cell-specific enhancer loci [1, 2 , 3]. Similar observations have been made for DNA methylation profiles. Continuous regions of low DNA methylation (at least

66 five kilobase (kb) in length), namely DNA methylation valleys or canyons, have been associated
67 with transcription factors and with genes known to regulate development in embryonic stem
68 cells and genes with potential involvement in the regulation of hematopoiesis in hematopoietic
69 stem cells (HSCs) [4]. In addition, in a recent publication, we showed that longer domains (from
70 5 to 50kb) of histone H3 lysine 4 trimethylation (H3K4me3) preferentially mark genes
71 associated with cell identity and function in diverse cell types and organisms [5]. These
72 expanded H3K4me3 domains, termed broad domains, have also been shown to mark tumor
73 suppressors in normal cells when compared to tumor pairs [6].
74
75 These recent studies show that the long continuous segments of DNA covered by H3K4me3 and
76 H3K27ac can be used as an epigenetic signature for identifying cell-type-specific promoters and
77 enhancers and these epigenetic patterns might be established and maintained by the cell to
78 robustly regulate cell-type specific gene expression patterns and functions. However, what
79 potentially regulates the deposition or maintenance of such cell-specific epigenetic signatures
80 still remains an open question. Computationally identifying genomic and epigenomic
81 characteristics associated with the length of these histone marks might provide some hints
82 about potential mechanisms. For this purpose, to systematically identify genomic and
83 epigenomic characteristics associated with cell-specific epigenetic signatures, we build Random
84 Forest regression models to explain the length of H3K4me3 and H3K27ac domains using both
85 genomic and epigenomic information. Our regression models are different than our previous
86 method that is mainly comparing very long H3K4me3 domains to shorter domains [5]. Here, we
87 study the length of a domain as a spectrum without arbitrarily introducing a length cut-off.
88 Moreover, we also study H3K27ac patterns in addition to H3K4me3, which is the mark that has
89 been used for identifying cell-type-specific enhancers. To the best of our knowledge, this is the
90 first attempt to study the length of an epigenetic mark under the light of other genomic and
91 epigenomic characteristics of that site. A recent algorithm showed that in the existence of many
92 epigenetic marks, missing epigenetic marks could be imputed at 25-bp resolution using
93 regression tress [7]. Although this method and study was informative in showing that
94 epigenetic marks are not independently deposited from each other, it did not reveal whether
95 any of these correlations are important for the length of H3K4me3 and H3K27ac domains.
96 Moreover, this method was restricted to histone modification profiles and didn't study the
97 relationship between protein binding patterns and the length of these two important
98 epigenetic marks.
99
100 Using our regression models in seven ENCODE cell lines (human embryonic stem cell lines H1
101 and H9 and cancer cell lines K562, HCT116, GM12878, A549 and HeLa), we asked whether the
102 length of loci marked with H3K4me3 or H3K27ac domains can be predicted using other genomic
103 features of that loci including transcription factor (TFs) binding patterns and other epigenetic
104 marks using chromatin immunoprecipitation followed by sequencing (ChIP-seq) datasets. This
105 enabled us to systematically study the relationship between H3K4me3 and H3K27ac length
106 found at a locus and other genomic features of that locus. Moreover we also quantified the
107 importance of different genomic features in describing the epigenetic domain length. Our
108 analyses revealed that with integrative computational models we could predict the length of
109 H3K4me3 and H3K27ac domains with high efficacy (>0.9 correlation on the average between

110 observed and predicted). Moreover, with our models we can infer the most important genomic
111 and epigenomic features that are predictive for the length of these two epigenetic marks.
112

Materials & Methods

114

115 **ChIP-seq data pre-processing:** Publicly available datasets were obtained from ENCODE
116 consortium [8], in multiple ENCODE cell lines as well as in pancreatic islets (accession
117 numbers are listed in **Table S1**). To preprocess ChIP-seq datasets, we mapped reads to
118 reference human genome (hg19) using bowtie0.12.7 [9]. ChIP-seq peaks were called using
119 MACS2.08 [4] with the "—broad" option for histone marks. Peaks were assigned to the
120 gene with the closest transcription start site (TSS).

121 **Domain length prediction using Random Forest regression algorithm**
122 **Feature extraction:** For each histone modification domain and data feature being studied,
123 we used the fraction of overlapping basepairs (bps) in our models to eliminate the bias that
124 could be introduced by differences in domain lengths. For example, if an H3K4me3 domain
125 has length 100 base pairs and 15 of these bases also have CHD1 binding, fraction of overlap
126 for that domain for the CHD1 feature is (15/100=0.15). Note that a single H3k4me3 domain
127 can simultaneously overlap with multiple instances of the same feature (e.g., multiple CHD1
128 binding sites within a domain). For these instances, we used the sum of all of such overlaps.
129 For example, as depicted in the **Fig. 1C** domain B overlaps with two instances of domains A,
130 therefore the final score for this domain for feature A is = sum(bp_A)/length(B).
131

132

133

134 **Model building:** Random Forest regression with 100 trees was performed on the
135 constructed data sets. We employed these models to unfold the relationship between
136 domain length and the constructed features. We used an implementation available in scikit-
137 learn library [10]. Gini impurity was used as a quality measure for choosing best splits. At
138 each level $\sqrt{m}$ randomly chosen attributes were considered as candidates for split (where
139 m is the number of all attributes).
140

141 **Feature ranking:** Feature rankings enable identifying and ranking important features for
142 the regression problem at hand. When building a decision tree, the feature that leads to the
143 greatest decrease in Gini impurity score is chosen at each split. Importance score for each
144 feature is the mean decrease of impurity for all tree nodes, which is averaged over all trees
145 in the ensemble. Calculating importance scores enable comparison of attributes, but does
146 not state the significance of importance scores. To assess significance we employed a
147 Monte-Carlo technique based on contrast attributes, which are random permutations of the
148 original attributes. To quantify the significance of a feature importance score, we followed
149 the below procedure:
150   1. For each of the original attributes its values were permuted at random and added as
151   a new contrast attribute to the original dataset.
152   2. Random Forest model was trained on the dataset consisting of both original and
153   contrast attributes.

154   3. Standard deviation (SD) of importance scores for contrast attributes was calculated.
155   The value equal to 2xSD was used as a cutoff for minimal significant difference
156   between importance scores.
157   4. Feature importance scores were sorted in decreasing order. Differences between
158   subsequent scores were calculated. We looked for the last pair of subsequent scores
159   for which the difference was larger than the cutoff value. The larger of the scores
160   from that pair constituted a threshold for dividing features into significant and not
161   significant.
162   This procedure enabled us to identify a small set of significant features for our models.
163
164

165

166
167

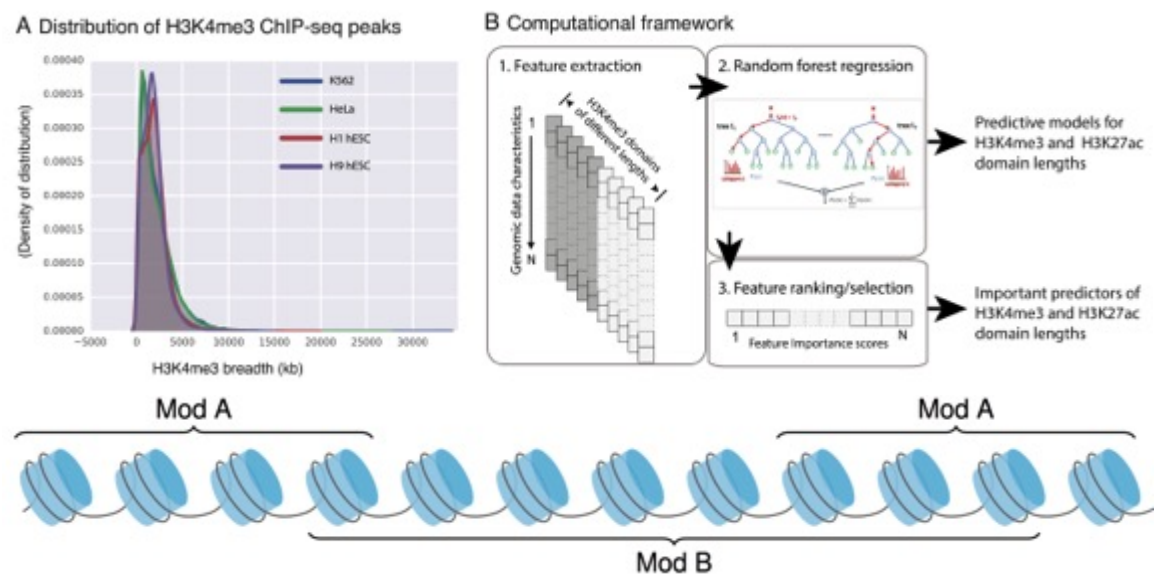



Figure 1: H3K4me3 domain lengths vary. A. H3K4me3 domain length distribution in four ENCODE cell lines. H3K4me3 domain lengths span a wide range and longer domains mark cell-type-specific promoters. **B.** Our three-step framework to model the length of H3K4me3 and H3K27ac domains. **C.** An example of domain type B, which overlaps with two instances of domains of type A.

168
169
170
171
172

Results

H3K4me3- and H3K27ac-domain lengths are informative for cell-type specificity

It has been observed that epigenetic marks can decorate loci with domains of varying length. For example, when we study the domain lengths for H3K4me3, we observe that the length distribution is not uniform and the length of these peaks vary between a few hundred and 20kb as shown in the length distribution of H3K4me3 domains in H1 human embryonic stem cells (hESCs) (**Fig. 1A**). The span of epigenetic mark deposition has gained attention in recent years with several studies, including ours, showing that particularly long stretches of DNA marked with H3K4me3 or H3K27ac specifically coincide with cell-type specific promoters or enhancers, respectively. [1, 2 , 3, 5, 11].

Given the association of large domains with functionally important DNA elements, we built a computational framework that can i) assess whether—and to what extent—the length of an H3K4me3 (or H3K27ac) domain at a locus can be predicted from other genomic and epigenomic characteristics of that locus and ii) quantify the ability of each genomic and epigenomic characteristic to predict the length of domains of these two marks. Our framework consisted of three stages (summarized in **Fig. 1B**): 1) extracting genomic and epigenomic characteristics of each domain; 2) building a regression model (based on Random Forest regression) for the length of the domains as a function of other data characteristics; and 3) delineating the predictive genomic signature for domain length by prioritizing and selecting important predictors. The signatures we obtained will help us to identify candidate molecular mechanisms for setting or maintaining domain length of this epigenetic mark.

H3K4me3 and H3K27ac deposition lengths can be predicted with high accuracy by integrating other genomic datasets

For each H3K4me3 or H3K27ac domain, we extracted other genomic and epigenomic characteristics from ENCODE datasets. The complete set of datasets used in each cell type is listed in **Fig. 2**. We used the fraction of overlap (as explained in methods) to normalize feature overlaps based on domain length, so longer domains would not have an inflation of overlapping features. Next, in seven ENCODE cell lines (H1 and H9 hESCs, K562, HCT116, GM12878, A549, HeLa) as well as in human pancreatic islets, we built Random Forest Regression models to predict the length of H3K4me3 and H3K27ac domains. Our analyses showed that these models can predict the length of epigenetic marks with high accuracy (>0.9 correlation between observed and predicted lengths**; Fig. 2)**.

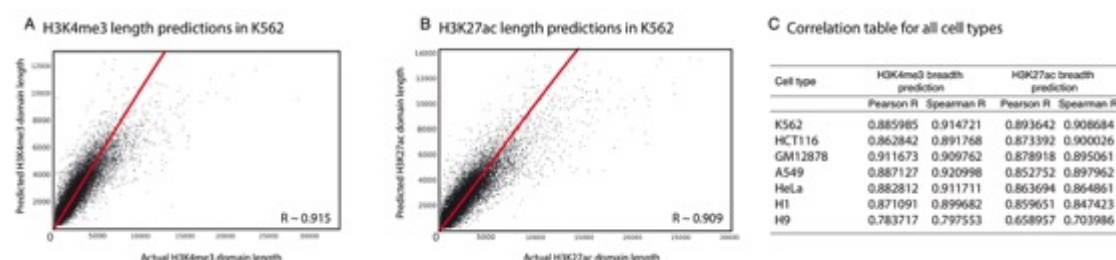



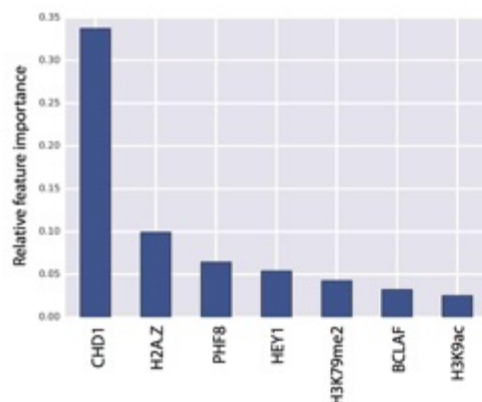
Figure 2: Our computational models can predict the length of H3K4me3 and H3K27ac domains with high precision. A. Predicted vs. observed H3K4me3 domains using our model in K562. **B.** Predicted vs. observed H3K27ac domains using our model in K562. **C.** Correlation between predicted and observed domain lengths in all studied cell types.

210

211

212 **Conserved genomic features are predictive for H3K4me3 and H3K27ac**
213 **deposition lengths**
214 Random Forest quantifies the importance of each feature by calculating its contribution to the
215 overall classification accuracy (see methods for details). We calculated feature importance
216 scores for each genomic and epigenomic data type we used in our models. Next, we assessed
217 the significance of these scores by comparing them with scores produced by random
218 permutations of the original features (i.e., contrast features). With this methodology, we
219 identified the significant predictors of length for H3K4me3 and H3K27ac in each studied cell
220 type. For example, we showed that in K562 CHD1 is one of the most important predictors for
221 both H3K4me3 and H3K27ac domains (**Figures 3 and 4**).

222

223 Our random forest regression models revealed that there are frequently repeating features that
224 are predictive for both H3K4me3 and H3K27ac epigenetic marks, which point out to putative
225 upstream or downstream associations between the length of epigenetic marks and these
226 predictive features. Certain genomic and epigenomic datasets frequently rank as important for
227 predicting the length of these two marks across cell types (**Fig. 3C-D**). For H3K4me3, CHD1- and
228 Pol2-binding and the H3K79me2 histone modification are the most important predictors as
229 assessed by feature importance scores. For H3K27ac, TEAD1- and CHD1-binding and the
230 presence of other acetylation marks (H3K4ac and H2AK5ac) are important genomic
231 characteristics.

232

233

234

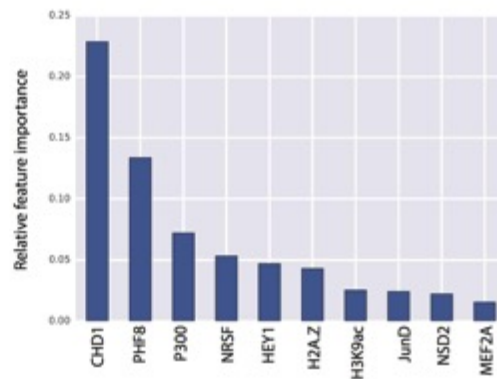
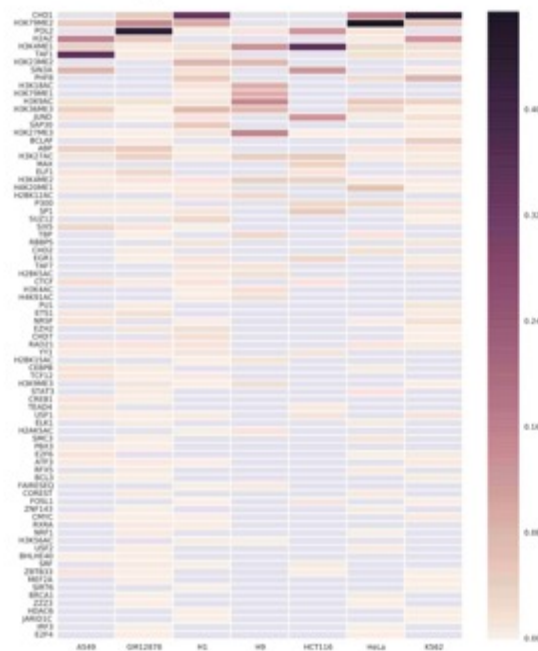
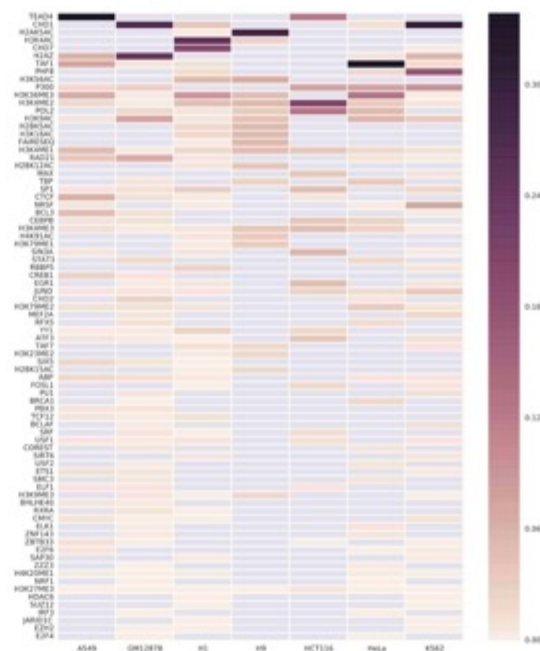
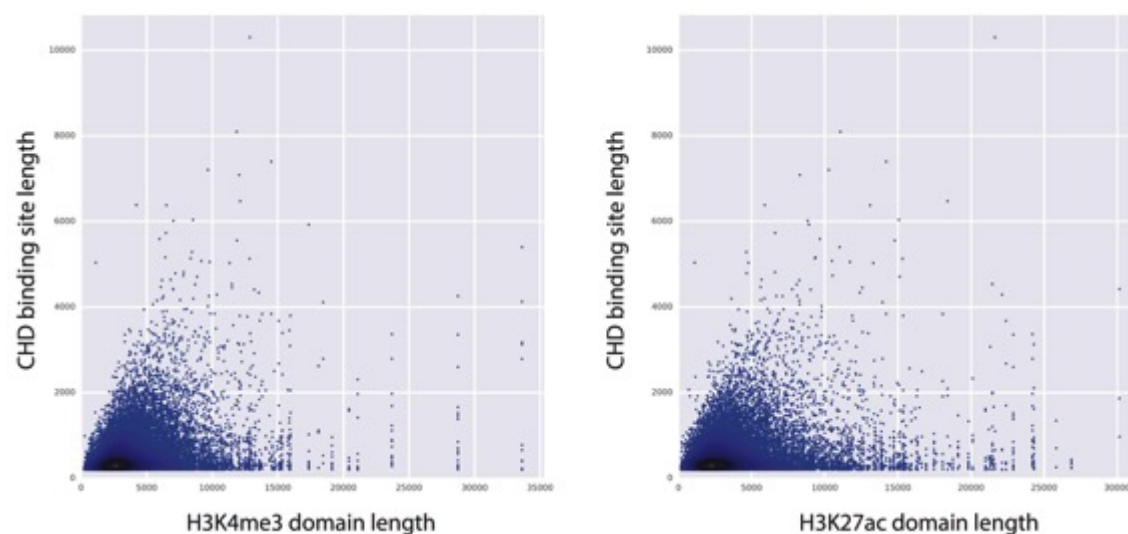



Figure 3: We identified genomic features that play an important role in H3K4me3 and H3K27ac domain lengths. Significant predictors of H3K4me3 domain length in K562 (**A**) and in all studied cell types (**C**). Significant predictors of H3K27ac domain length in K562 (**B**) and in all studied cell types (**D**). Missing datasets in the studied cell types are represented in white.

Elongated CHD1 binding might promote robust and increased expression patterns

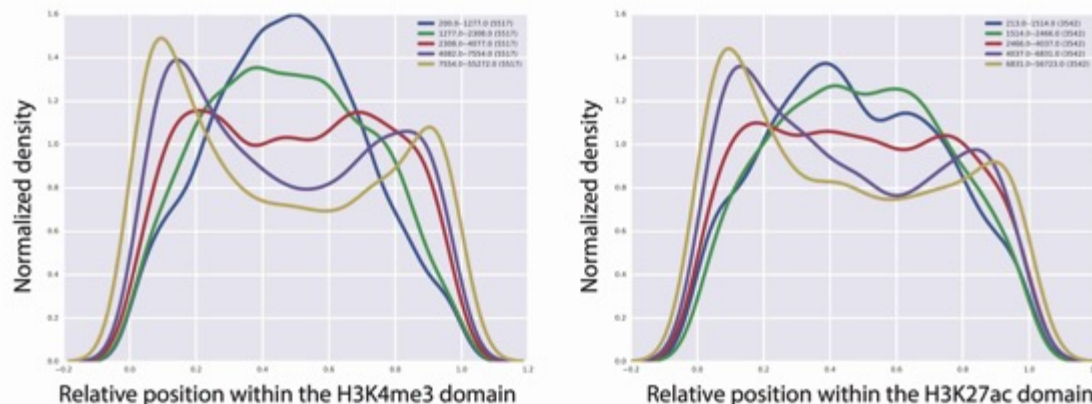
Among the top predictors, CHD1 is an H3K4me3-associated chromatin remodeler and it is shown to be essential for the expression of key developmental genes in mouse embryonic stem (ES) cells as well as for reprogramming of fibroblasts to the pluripotent stem cells [12]. Moreover, CHD1 binding is shown to be essential for the assembly of preinitiation complexes

245 (PIC), which is necessary to start transcription. PIC includes co-activators like Mediator complex
246 and the TFIID complex, along with Pol2 and the general transcription factors [13]. Given the
247 importance of CHD1 in regulating PIC assembly and ultimately gene expression, we have
248 conducted further analyses to study the relationship between CHD1 binding and the length of
249 H3K27ac and H3K4me3 domains. Our analyses showed that there is a positive correlation
250 between the length of CHD1 and the lengths of both H3K27ac and H3K4me3 domains (**Fig. 4**),
251 implying that elongated CHD1 binding takes place at regions where longer H3K4me3 and longer
252 H3K27ac deposition is observed. Broad H3K4me3 domains have been associated with increased
253 and robust expression of genes, similarly genes marked by longer H3K27ac domains (i.e.,
254 super/stretch enhancers) have been associated with increased expression. Our findings show
255 that longer H3K4me3 and H3K27ac domains have longer deposition of CHD1 binding; which
256 might imply a more robust assembly of mediator complex and PIC enabled by the elongated
257 binding of CHD1 protein. This robust assembly of PIC at these loci might explain their robust
258 and increased expression patterns.
259

260
261 **Figure 4: CHD1 binding length and H3K4me3 (and H3K27ac) domain lengths are correlated.** X axes
        represents the length of H3K4me3 domains (left panel) and the length of H3K27ac domains (right panel). Y axes
262   represents the length of CHD1 bound loci.

Broader H3K4me3 and H3K27ac are bounded by CTCF mediated interactions

264 CTCF, is a TF that is important for the insulator activity. Insulators are boundary elements that
265 restrict the regulatory interactions between enhancers and promoters within a boundary
266 element.   An intriguing hypothesis based on the high predictive value of CTCF for the length of
267 H3K4me3 and H3K27ac domains is that the longer H3K27ac and H3K4me3 domains tend to be
268 enriched at the boundary of chromatin domains marked by the CTCF binding. To test this, we
269 analyzed CTCF mediated chromatin interactions obtained by the Chromatin Interaction Analysis
270 by Paired-End Tag Sequencing (ChIA-PET) technology in GM12878 cell line [14]. We split
271 H3K4me3 and H3K27ac domains into five bins based on their domain length and for each of
272 these bins, we studied the localization of CTCF mediated chromatin interactions with respect to
273 the center of the domain overlapping the CTCF fragment. Our analyses revealed an interesting
274 pattern, where we observed that with increasing length of H3K27ac and H3K4me3 domains,

275 CTCF interactions tend to cluster at the beginning or at the end of the domains (**Figure 5**), which
276 implies that longer H3K4me3 and H3K27ac domains tend to be at the boundary of CTCF defined
277 boundary domains.

278

279
280
281 **Figure 5: Localization of CTCF interactions over H3K4me3 (left panel) and H3K27ac (right panel) domains.** X
282 axis represents the relative position on the domains. Y axis represents the CTCF interaction coverage. Yellow
283 represents the longest domains, blue represents the shortest domains.

284

285 # Discussion

286 Recent epigenomic studies, including ours, have shown that the length of certain epigenetic
287 mark deposition domains — H3K4me3 for promoters and H3K27ac for enhancers — might have
288 functional implications and can be predictive in identifying promoters and enhancers that are
289 critical for the underlying cell's identity and function. However, we still do not know the
290 mechanisms that establish and maintain the boundaries of these domains, therefore it is still
291 intriguing to investigate whether these domain lengths correlated with other genomic and
292 epigenomic datasets using regression models, which might be important players in setting and
293 maintaining these domains. Our analyses showed that in fact these domain lengths could be
294 predicted with high accuracy using publically available Transcription Factor (TF) binding and
295 histone modification ChIP-seq datasets in seven ENCODE cell lines as well as in human
296 pancreatic islet samples. This high accuracy of our regression models show that the length of
297 these domains correlates well with the genomic and epigenomic states of these sites within a
298 cell and not a byproduct of assays or analyses. We conducted further analyses to understand in
299 which ways these important predictors might be related to broad H3K4me3 and H3K27ac
300 domains. Our analyses revealed that the nature of the relationship is based on the genomic
301 feature in consideration. Our analyses with CHD1 binding revealed that elongated CHD1 binding
302 is associated with longer H3K27ac and H3K4me3 domains, which might contribute to the robust
303 and increased expression of genes associated with these domains via a robust assembly of PIC.
304 On the other hand, for CTCF domains, we noticed that broader H3K4me3 and H3K27ac domains
305 tend to locate towards the edges of CTCF defined boundary elements. Although our
306 computational analyses cannot infer causality in the establishment of CTCF domains and
307 broader deposition of epigenetic marks, we still observe that the deposition of domain length is
308 not independent of the regulatory boundaries defined by CTCF binding.

309
310
311

Conclusions

313 Our analyses revealed that there are frequently occurring predictors for H3K4me3 and H3K27ac
314 domain lengths. Among the predictors we identified, CHD1-binding is important for the
315 prediction of both H3K27ac and H3K4me3 domain lengths. CHD1 is essential for the assembly of
316 preinitiation complexes (PIC) and ultimately for regulating gene expression levels. We have
317 conducted further analyses to show that broader H3K4me3 and H3K27ac domains have elongated
318 CHD1 binding, which might play a role in their robust and increased expression patterns In the
319 future, it would be interesting to study the effect of CHD1 knockdown or precise targeting on
320 the expression levels of genes marked by broad H3K4me3 and H3K27ac domains. We also
321 showed that broader H3K4me3 and H3K27ac domains localize towards the end of the
322 regulatory boundary elements defined by CTCF binding by studying chromatin interactions
323 datasets generated by ChIA-PET technology. In summary, our computational analyses revealed
324 interesting features of epigenetic mark length for two important histone modification marks
325 (H3K4me3 and H3K27ac) and showed in which ways these signals correlate with TF binding
326 patterns and chromatin interactions.

327
328
329
330
331
332

333

References

1. Hnisz, D., B.J. Abraham, T.I. Lee, A. Lau, V. Saint-Andre, A.A. Sigova, H.A. Hoke, and R.A. Young, *Super-enhancers in the control of cell identity and disease.* Cell, 2013. **155**(4): p. 934-47.PMCID: 3841062

2. Chapuy, B., M.R. McKeown, C.Y. Lin, S. Monti, M.G. Roemer, J. Qi, P.B. Rahl, H.H. Sun, K.T. Yeda, J.G. Doench, E. Reichert, A.L. Kung, S.J. Rodig, R.A. Young, M.A. Shipp, and J.E. Bradner, *Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma.* Cancer Cell, 2013. **24**(6): p. 777-90.PMCID: 4018722

3. Parker, S.C., M.L. Stitzel, D.L. Taylor, J.M. Orozco, M.R. Erdos, J.A. Akiyama, K.L. van Bueren, P.S. Chines, N. Narisu, N.C.S. Program, B.L. Black, A. Visel, L.A. Pennacchio, F.S. Collins, A. National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, and N.C.S.P. Authors, *Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17921-6.PMCID: 3816444

4. Jeong, M., D. Sun, M. Luo, Y. Huang, G.A. Challen, Rodriguez B., Zhang X., Chavez L., Wang H., Hannah R., Kim S.B., Yang L., Ko M., Chen R., Göttgens B., Lee J.S., Gunaratne P., Godley L.A., Darlington G.J., Rao A., Li W., Goodell M.A., *Large conserved domains of low DNA methylation maintained by Dnmt3a.* Nat Genet, 2014. **46**(1): p. 17-23.PMCID: 3920905

5. Benayoun, B.A., E.A. Pollina, D. Ucar, S. Mahmoudi, K. Karra, E.D. Wong, K. Devarajan, A.C. Daugherty, A.B. Kundaje, and E. Mancini, *H3K4me3 breadth is linked to cell identity and transcriptional consistency.* Cell, 2014. **158**(3): p. 673-688

6. Chen, K., Z. Chen, D. Wu, L. Zhang, X. Lin, J. Su, B. Rodriguez, Y. Xi, Z. Xia, and X. Chen, *Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes.* Nature genetics, 2015

7. Ernst, J. and M. Kellis, *Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues.* Nature biotechnology, 2015. **33**(4): p. 364-376

8. Hoffman, M.M., J. Ernst, S.P. Wilder, A. Kundaje, R.S. Harris, M. Libbrecht, B. Giardine, P.M. Ellenbogen, J.A. Bilmes, E. Birney, R.C. Hardison, I. Dunham, M. Kellis, and W.S. Noble, *Integrative annotation of chromatin elements from ENCODE data.* Nucleic Acids Res, 2013. **41**(2): p. 827-41.PMCID: 3553955

9. Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.PMCID: 2690996

10. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830

11. Bernstein, B.E., A. Meissner, and E.S. Lander, *The mammalian epigenome.* Cell, 2007. **128**(4): p. 669-81

12. Gaspar-Maia, A., A. Alajem, F. Polesso, R. Sridharan, M.J. Mason, A. Heidersbach, J. Ramalho-Santos, M.T. McManus, K. Plath, and E. Meshorer, *Chd1 regulates open*

378 *chromatin and pluripotency of embryonic stem cells.* Nature, 2009. **460**(7257): p.
379 863-868

380 13. Lin, J.J., L.W. Lehmann, G. Bonora, R. Sridharan, A.A. Vashisht, N. Tran, K. Plath, J.A.
381 Wohlschlegel, and M. Carey, *Mediator coordinates PIC assembly with recruitment of*
382 *CHD1.* Genes & development, 2011. **25**(20): p. 2198-2209

383 14. Tang, Z., Luo O.J., Li X., Zheng M., Zhu J.J., Szalaj P., Trzaskoma P., Magalska A.,
384 Wlodarczyk J., Ruszczycki B., Michalski P., Piecuch E., Wang P., Wang D., Tian S.Z.,
385 Penrad-Mobayed M., Sachs L.M., Ruan X., Wei C.L., Liu E.T., Wilczynski G.M.,
386 Plewczynski D., Li G., Ruan Y., *CTCF-Mediated Human 3D Genome Architecture*
387 *Reveals Chromatin Topology for Transcription.* Cell, 2015. **163**(7): p. 1611-27

388

389

390