GMR
Genetics and Molecular Research
ISSN 16765680

Building the SeqChromMM Markov property atlas of the human genome by analyzing the 200-bp units of the 15 different chromatin regions of ENCODE

K.-E. Lee[1] and H.-S. Park[1,2]

[1]Bioinformatics Laboratory, Engineering School, Ewha Womans University, Seoul, Korea
[2]Center for Convergence Research of Advanced Technologies, Ewha Womans University, Seoul, Korea

Corresponding author: H.-S. Park
E-mail: neo@ewha.ac.kr

Genet. Mol. Res. 15 (3): gmr.15038992
Received July 20, 2016
Accepted August 1, 2016
Published August 30, 2016
DOI http://dx.doi.org/10.4238/gmr.15038992

ABSTRACT. We analyzed the publicly available ChromHMM BED files of the ENCODE project and tested the Markov properties of the different chromatin states in the human genome. Nucleotide frequency profiles of regional chromatin segmentations were analyzed, and Markov chains were built to detect Markov properties in the chromatin states of different ChromHMM regions. By estimating the transition probabilities of 200-base pair nucleotide sequences of the human genome, we constructed a nucleotide-sequence-based Markovian chromatin map called SeqChromMM.

Key words: Chromatin maps; Nucleotide frequency patterns; Markov chain; Noncoding DNA; Computational epigenetics

INTRODUCTION

Recently, Ernst et al. (2011) applied multivariate hidden Markov models to the biological assay dataset from the Broad Histone track of ENCODE, building annotation maps of 15 chromatin states, known as ChromHMM (Ernst et al., 2011). The 9 browser extensible data (BED) files (with the human genome GRCh35/hg19) are published on the ENCODE Analysis Data Hub website for public download (http://genome.ucsc.edu/ENCODE/downloads.html) (ENCODE Project Consortium, 2012). Thus, recent advances in computational epigenetics such as ChromHMM provide new insights into n-gram probabilistic language models for parsing non-coding DNA regions (Smith et al., 1983; Borodovskii et al., 1986).

In our recent study (Lee and Park, 2015), we performed preliminary experiments to test whether each of the 15 chromatin states contained in the commonly annotated chromatin regions of the ENCODE Tier 1 cell types possessed Markov properties by applying third to sixth order Markov chains. We presented a pioneering study to provide the Markovian order statistics of nucleotide sequences of the whole human genome (Lee and Park, 2015).

In this study, a follow-up to our previous study, extended this analysis and continued our ongoing efforts to build a Markov property map of chromatin blocks of the human genome by publishing a sequence-based Markovian chromatin map, called SeqChromMM. In building the SeqChromMM, we modified our initial Markov models by dissecting the ChromHMM blocks into 200-base pair (bp) units, analyzing the blocks in relation to different cell lines, and iteratively rebuilding the Markov chains. Our map is an important resource for statistical models necessary to develop algorithms to predict chromatin states or genes in relation to the vast amount of biological assays of large-scale epigenetic projects.

MATERIAL AND METHODS

Building preliminary Markov chains based on a BED file single ChromHMM

In our recent study (Lee and Park, 2015), we used the BED files of ChromHMM (Ernst et al., 2011; Lee and Park, 2014) to analyze sequence-based profiles to identify nucleotide sequences in the 15 chromatin states with Markov properties. We downloaded the ChromHMM BED files of the Broad Histone track of the ENCODE consortium and parsed the BED files to build various transition tables based on nucleotide frequency profiles for each of the 15 different ChromHMM regions. The 15 chromatin states were Active, Repressed, and Poised Promoters (states 1, 2, 3), Strong and Weak Enhancers (states 4, 5, 6, 7), Putative Insulators (state 8), transcribed regions (states 9, 10, 11), and large-scale repressed and inactive domains (states 12, 13, 14, 15).

Figure 1 displays 15 transition table Markov chains built from the commonly annotated ChromHMM regions of the erythrocytic leukemia (K562) and B-lymphoblastoid (GM12878) cell lines.

Figure 1A explains the detailed approach doe building Markov chains. Initially, the frequency counts were used to build 15 preliminary transition tables for the 5th order Markov models based on the common regions of the two ChromHMM BED files. These transition tables were used as the basis of a global Markov chain classifier for exploring and ranking sub-optimal predictions.

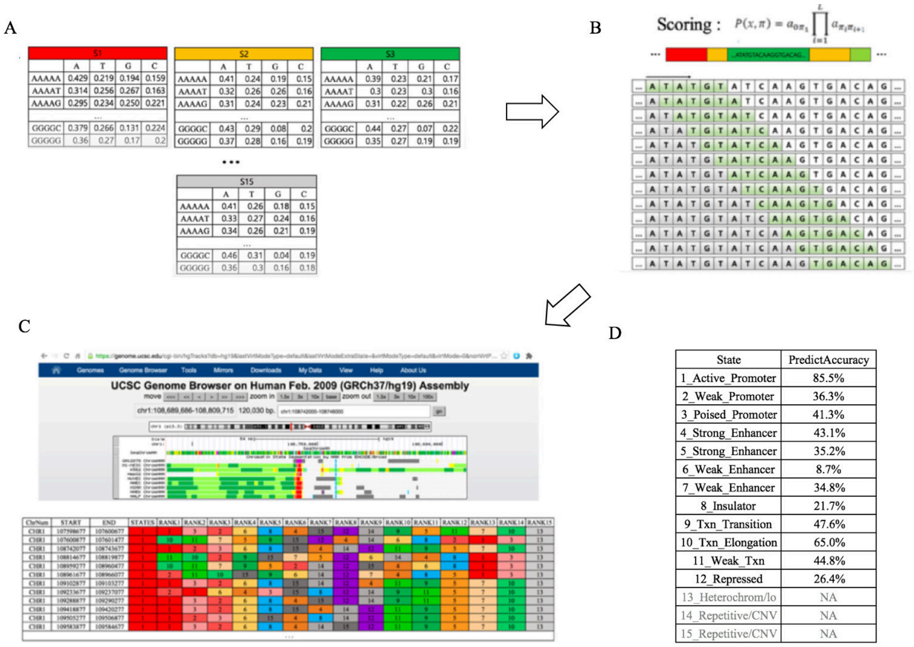
Figure 1B shows how a chromatin state was predicted based on the nucleotide frequency profiles. Given a random sequence in the state of a cell line, we calculated the sequence of chromatin states that maximized the following probability of the 15 Markov chain models, where is a transition probability:

$$P(x,\pi) = \alpha_{0\pi 1} \prod_{i=1}^{L} \alpha_{\pi i \pi+1}$$

As these Markov chains can be used as a Naive Bayes classifier, we calculated the sequence of each ChromHMM block that maximized our Markov models.

Figure 1C shows an example prediction result from chr1:108,686,689 to chr1: 108,809,715. For each ChromHMM block, the transition probabilities of the 15 Markov chain models were sorted in descending order according to their probability scores. Thus, the first column shows the assigned or predicted chromatin state that maximized P[x, p|M]. In this case, most ChromHMM blocks were predicted as the Active Promoter, or state 1.

Figure 1D shows the prediction result, a preliminary characterization of the Markov property of the 15 chromatin states, as the Markov property comes into the model as an assumption. Validity was typically verified empirically by statistical analysis. We measured the prediction accuracy for each chromatin state by adding the number of all correctly predicted blocks in the same chromatin broad group and dividing by the total number of testing blocks. The prediction accuracy of each state differed greatly, and the results clearly showed that some regions had a stronger Markov property than others. More detailed descriptions can be found in our previous study (Lee and Park, 2015).


A
B
Scoring : $P(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} a_{\pi_i \pi_{i+1}}$
C
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
D

S1	A	T	G	C
AAAAA	0.429	0.219	0.194	0.159
AAAAT	0.314	0.256	0.267	0.163
AAAAG	0.295	0.234	0.250	0.221
...				
GGGGC	0.379	0.266	0.131	0.224
GGGGG	0.36	0.27	0.17	0.2

S2	A	T	G	C
AAAAA	0.41	0.24	0.19	0.15
AAAAT	0.32	0.26	0.26	0.16
AAAAG	0.31	0.24	0.23	0.21
...				
GGGGC	0.43	0.29	0.08	0.2
GGGGG	0.37	0.28	0.16	0.19

S3	A	T	G	C
AAAAA	0.39	0.23	0.21	0.17
AAAAT	0.3	0.23	0.3	0.16
AAAAG	0.31	0.22	0.26	0.21
...				
GGGGC	0.44	0.27	0.07	0.22
GGGGG	0.35	0.27	0.19	0.19

...

S15	A	T	G	C
AAAAA	0.41	0.26	0.18	0.15
AAAAT	0.33	0.27	0.24	0.16
AAAAG	0.34	0.26	0.21	0.19
...				
GGGGC	0.46	0.31	0.04	0.19
GGGGG	0.36	0.3	0.16	0.18

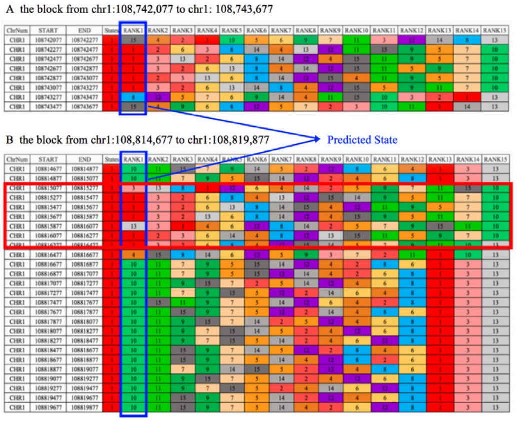
State	PredictAccuracy
1_Active_Promoter	85.5%
2_Weak_Promoter	36.3%
3_Poised_Promoter	41.3%
4_Strong_Enhancer	43.1%
5_Strong_Enhancer	35.2%
6_Weak_Enhancer	8.7%
7_Weak_Enhancer	34.8%
8_Insulator	21.7%
9_Txn_Transition	47.6%
10_Txn_Elongation	65.0%
11_Weak_Txn	44.8%
12_Repressed	26.4%
13_Heterochrom/lo	NA
14_Repetitive/CNV	NA
15_Repetitive/CNV	NA

Figure 1. Initial Markov chains built from the 15 chromatin states of ChromHMM (heterochromatin and repetitive/CNV regions were not population homogeneous, and later, excluded from the test set because they resulted in low entropy according to the Kullback-Leibler distance test (Kullback, 1987).

Analyzing the prediction results by dissecting the ChromHMM blocks into 200-bp units

To extend our previous study (Lee and Park, 2015), we evaluated the ChromHMM blocks for cases in which our prediction result and ChromHMM annotation in the K562 or GM12878 BED file did not match. To evaluate these blocks, the ChromHMM blocks were uniformly dissected into 200 bp, and a chromatin state for each of the individual regions was analyzed and assigned a predicted chromatin state at a nucleosome resolution of 200 bp. Based on the prediction results, we created the initial version of a sequence-based Markovian chromatin map, called SeqChromMM, where the number of entries was 13,223,025.

Figure 2A and B explain the exemplary prediction results of the two ChromHMM blocks dissected into units of 200 base pairs. The original annotations of both blocks in the K562 and GM12878 BED files were chromatin state 1. However, Figure 2A and B are a correctly and incorrectly predicted example, respectively.


Predicted State

A the block from chr1:108,742,077 to chr1: 108,743,677

ChrNum	START	END	States	RANK1	RANK2	RANK3	RANK4	RANK5	RANK6	RANK7	RANK8	RANK9	RANK10	RANK11	RANK12	RANK13	RANK14	RANK15
CHR1	108742077	108742277	1	15	4	2	1	10	5	6	9	7	11	3	12	14	8	13
CHR1	108742277	108742477	1	1	2	6	3	8	14	4	13	12	11	15	9	5	7	10
CHR1	108742477	108742677	1	1	3	13	2	6	8	14	12	4	15	11	9	5	7	10
CHR1	108742677	108742877	1	1	3	2	6	13	8	4	14	12	15	11	5	9	7	10
CHR1	108742877	108743077	1	1	2	3	13	6	8	14	12	15	4	11	9	5	7	10
CHR1	108743077	108743277	1	1	3	2	6	13	14	8	4	12	15	5	9	11	7	10
CHR1	108743277	108743477	1	8	12	5	7	6	9	14	4	11	15	10	3	2	1	13
CHR1	108743477	108743677	1	15	4	9	6	8	12	5	7	2	3	10	11	1	14	13

B the block from chr1:108,814,677 to chr1:108,819,877

ChrNum	START	END	States	RANK1	RANK2	RANK3	RANK4	RANK5	RANK6	RANK7	RANK8	RANK9	RANK10	RANK11	RANK12	RANK13	RANK14	RANK15
CHR1	108814677	108814877	1	10	11	15	7	9	14	5	2	12	8	6	4	1	3	13
CHR1	108814877	108815077	1	10	11	7	9	5	15	14	12	2	4	6	8	1	3	13
CHR1	108815077	108815277	1	3	13	8	1	12	6	4	14	2	5	9	7	11	15	10
CHR1	108815277	108815477	1	1	3	2	6	8	4	13	12	15	14	9	5	11	7	10
CHR1	108815477	108815677	1	1	3	2	6	13	8	14	15	4	12	11	9	5	7	10
CHR1	108815677	108815877	1	1	3	2	13	6	8	14	12	4	15	9	5	11	7	10
CHR1	108815877	108816077	1	13	3	1	4	6	12	8	14	2	5	7	9	15	11	10
CHR1	108816077	108816277	1	1	2	3	6	14	4	8	12	13	15	11	5	9	7	10
CHR1	108816277	108816477	1	1	3	2	6	8	4	12	15	14	5	7	9	11	10	13
CHR1	108816477	108816677	1	4	15	8	14	6	12	5	9	3	7	2	11	1	10	13
CHR1	108816677	108816877	1	10	11	7	9	5	15	14	12	4	2	8	6	1	3	13
CHR1	108816877	108817077	1	10	11	7	9	5	15	12	14	4	2	6	8	1	3	13
CHR1	108817077	108817277	1	10	11	9	15	7	2	5	14	12	4	6	8	1	3	13
CHR1	108817277	108817477	1	10	11	7	9	15	5	12	2	14	6	4	8	1	3	13
CHR1	108817477	108817677	1	10	15	11	9	7	5	14	12	6	4	2	8	1	3	13
CHR1	108817677	108817877	1	10	11	15	9	7	5	12	2	14	4	8	6	1	3	13
CHR1	108817877	108818077	1	10	11	9	15	7	14	2	5	12	4	6	8	1	3	13
CHR1	108818077	108818277	1	10	11	9	7	15	14	5	2	4	12	6	8	1	3	13
CHR1	108818277	108818477	1	10	11	9	7	15	5	14	2	4	6	12	8	1	3	13
CHR1	108818477	108818677	1	10	11	15	9	7	14	5	12	2	8	4	6	1	3	13
CHR1	108818677	108818877	1	10	11	15	9	7	5	14	4	12	8	2	6	1	3	13
CHR1	108818877	108819077	1	10	11	7	9	14	15	5	2	12	4	8	6	1	3	13
CHR1	108819077	108819277	1	10	11	9	15	7	5	14	2	4	12	6	8	1	3	13
CHR1	108819277	108819477	1	10	11	9	7	15	5	2	4	14	6	12	8	1	3	13
CHR1	108819477	108819677	1	10	11	15	9	7	5	14	2	12	4	6	8	1	3	13
CHR1	108819677	108819877	1	10	11	15	9	7	5	14	4	2	6	12	8	1	3	13

Figure 2. Two example predicted blocks (compared to the K562 cell line), dissected into 200-bp regions.

Figure 2A shows that the original block from chr1: 108,742,077 to chr1: 108,743,677 (a block size of 1600 bp) was dissected into 8 units of 200 bp, where each 200-bp unit was re-evaluated individually by our classifier. Most of the 8 200-bp units were constantly predicted as chromatin state 1, resulting in the correct prediction of the original block. However, as shown in Figure 2B, when the block from chr1: 108,814,677 to chr1: 108,819,877 (a block size of 5200 bp) was dissected into 26 units of 200 bp, most of the units were incorrectly predicted as chromatin state 11, resulting in the incorrect prediction of the original block. However, some parts of the 200-bp units (e.g., from chr1: 108,815,277 to chr1: 108,816,477) were correctly predicted as chromatin state 1.

Based on the observation that correctly predicted 200-bp units can be found even within an incorrectly predicted ChromHMM block (similar to the case shown in Figure 2B), we gathered more detailed distribution statistics of the percentage of the correctly predicted 200-bp units within the incorrectly predicted ChromHMM Blocks, as shown in Figure 3. For example, among the ChromHMM blocks of state 1 (Active Promoter), 81.1% were correctly predicted, while 78.9% were incorrectly predicted. Among the incorrectly predicted blocks, 16.7% contained at least one correctly predicted 200-bp unit.

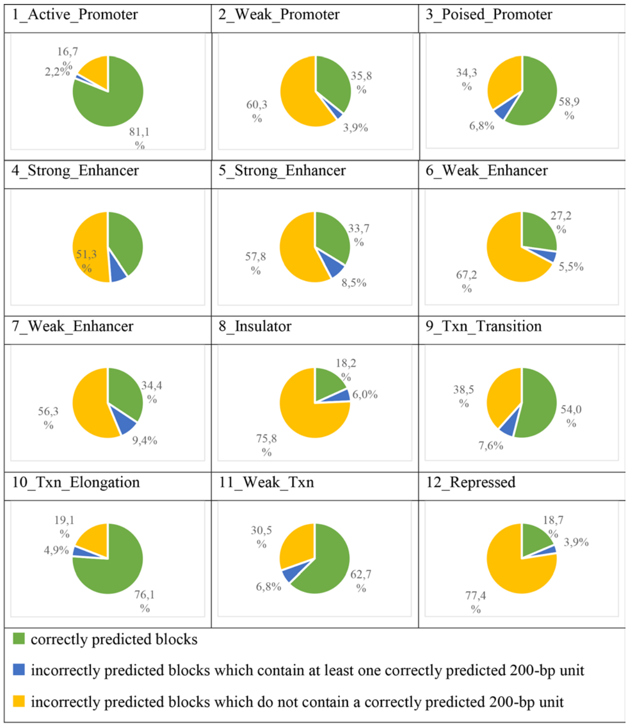


Figure 3. Distribution of percentage of the correctly predicted 200-bp units within the incorrectly predicted ChromHMM blocks of chromatin state 1 (excluding states 13, 14, and 15).

RESULTS

Publishing the SeqChromMM map

Finally, we iteratively excluded the incorrectly predicted 200-bp regions and rebuilt new Markov chains using the correctly predicted regions to produce 15 transition tables; we defined a correctly predicted block as one whose predicted result matches the majority annotation (more than 50%) of the 9 cell lines.

Figure 4 displays the distance matrix corresponding to newly built transition tables of the 15 states of SeqChromMM. According to the Chi-square distance plot, state 3 (Poised Promoter) was quite distant from states 1 and 2 (Active Promoter, Weak Promoter), although states 1, 2, and 3 all belonged to the same broad group of Promoters according to the ChromHMM document (Ernst et al., 2011). State 6 (Weak Enhancer) also showed different behavior compared to other enhancer states 4, 5, and 7. State 8 (Insulator) also showed a very different Markov property.

The results of the SeqChromMM map have been published on the GitHub repository (https://github.com/KyungEunLee/SeqChromMM.git). We also uploaded the SeqChromMM data for display in the UCSC browser (chr1:108,689,686-108,809,715). Figure 5 shows a snapshot, although our SeqChromMM custom annotation tracks are currently viewable only on the machine from which they were uploaded because of the publication policies of the UCSC genome browser (Rosenbloom et al., 2015).

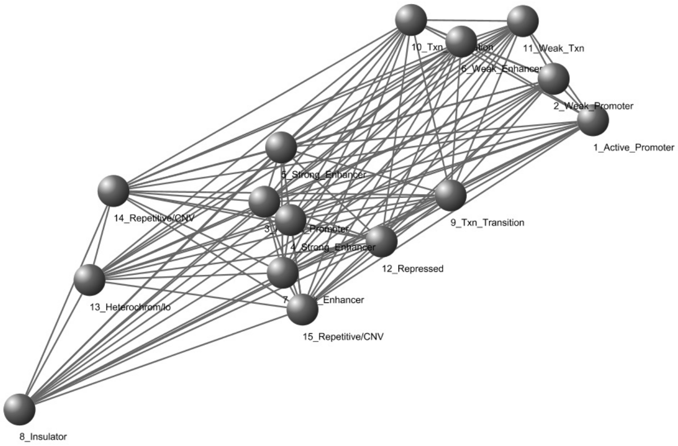
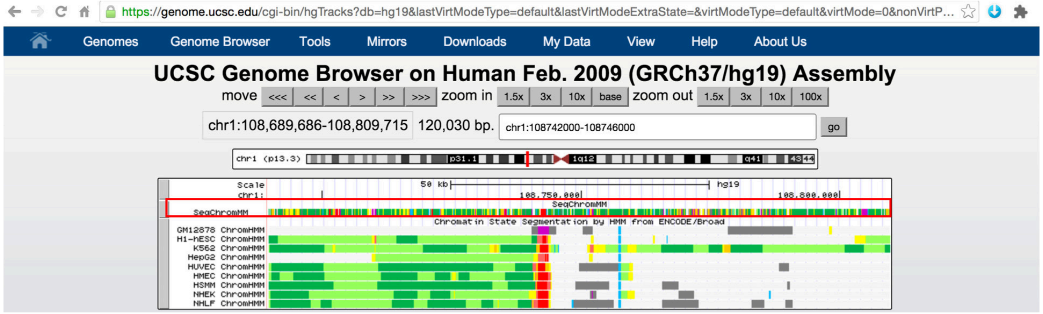


Figure 4. Distance matrix of the newly built transition tables of 15 Markov chains: the graph was drawn using Cytoscape (Shannon et al., 2003), an open source software platform for visualizing complex networks.


https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtP...

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:108,689,686-108,809,715 120,030 bp. chr1:108742000-108746000

Figure 5. Snapshot of the SeqChromMM Track visualized in the UCSC browser. There are 10 tracks, where the first track indicates the predicted annotations of SeqChromMM and the other tracks represent the ChromHMM annotations for 9 different cell types. We followed the same color code of ChromHMM. For example, state 1 segment is bright red (Active Promoter State), state 2 segment is light red (Weak Promoter State), state 3 segment is purple (Inactive/poised Promoter State), state 4 segment is orange (Strong enhancer State), and so on. The original 15 states and their associated segment colors can be found on the Broad ChromHMM epigenome project page (http://moma.ki.au.dk/genome-mirror/cgi-bin/hgTrackUi?db=hg18&g=wgEncodeBroadHmm).

DISCUSSION

SeqChromMM is an important resource because it can construct probabilistic models necessary to develop computational epigenetic algorithms to predict chromatin states, in relation to the vast amount of biological assay information generated through large-scale epigenetic projects.

By dissecting the original ChromHMM blocks into 200-bp units and publishing SeqChromMM, we can study the characteristics of chromatin barriers by examining the 200-bp units where the predicted states of the Markov chains change (Labrador and Corces, 2002; Lunyak et al., 2007; Cuddapah et al., 2009; Wang et al., 2012). However, more detailed evaluation of the issues associated with the boundary elements of chromatin states is necessary, and the 200-bp units of SeqChromMM can serve as a resource for investigating the characteristics of the chromatin boundaries in the whole human genome.

We can also investigate the overall variability in chromatin states in each 200-bp unit across the 9 cell lines in relation to the predicted states of SeqChromMM. We found that many positions of frequently variable chromatin states were the main sources of prediction errors. We also observed that our predicted states coincided with the annotations of other cell lines in most cases, although our initial Markov models were trained solely based on the BED files of the K562 or GM12878 cell lines. In general, Active Promoter and Transcribed chromatin states were highly constitutive. We also observed that Weak chromatin states were typically adjacent to Strong chromatin states.

Thus, sequence-based analysis dedicated to the prediction of epigenetic information is useful. Computational epigenomic predictions can substitute for experimental data to a certain degree. Prediction algorithms build computational models of epigenetic information from experimental data and can therefore act as a preliminary step toward statistical modeling of an epigenetic mechanism.

CONCLUSION

The field of epigenetics encompasses many diverse opinions, even regarding its definition. The proportion of genetically and epigenetically determined traits is also widely debated.

In this study, we extended our previous study of a conditional characterization of the Markov property of the publicly available 15 chromatin states of the ENCODE project and published the SeqChromMM map. The Markovian chromatin map has been made publicly available through GitHub and the UCSC Browser. We are collecting evidence not only from biological assays but also from DNA sequences, which can play an important role in determining the chromatin states of the human genome.

However, it was not possible to evaluate combinations of different chromatin states in cell lines within these Markov chain models in this study. Thus, SeqChromMM will be continuously improved by further investigating these issues.

ACKNOWLEDGMENTS

Research partially supported by the University-Industry Collaboration Program of Ewha Womans University and Digitaloptics Corp. (#2-2016-0872-001-1), and the Interdisciplinary Graduate Research Program in Systems Genomics (#2012M3A9D1054744) of Korea.

REFERENCES

Borodovskii MY, Sprizhitskii YA, Golovanov EI and Aleksandrov AA (1986). Statistical patterns in primary structures of functional regions in the in *E. coli. Mol. Biol.* 4: 826-883.

Cuddapah S, Jothi R, Schones DE, Roh TY, et al. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* 19: 24-32. http://dx.doi.org/10.1101/gr.082800.108

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74. http://dx.doi.org/10.1038/nature11247

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43-49. http://dx.doi.org/10.1038/nature09906

Kullback S (1987). Letter to the Editor: The Kullback-Leibler distance. *American Statistician* 41: 340-341.

Labrador M and Corces VG (2002). Setting the boundaries of chromatin domains and nuclear organization. *Cell* 111: 151-154. http://dx.doi.org/10.1016/S0092-8674(02)01004-8

Lee KE and Park HS (2014). A review of three different studies on hidden markov models for epigenetic problems: a computational perspective. *Genomics Inform.* 12: 145-150. http://dx.doi.org/10.5808/GI.2014.12.4.145

Lee KE and Park HS (2015). Preliminary testing for the Markov property of the fifteen chromatin states of the Broad Histone Track. *Biomed. Mater. Eng.* 26 (Suppl 1): S1917-S1927. http://dx.doi.org/10.3233/BME-151494

Lunyak VV, Prefontaine GG, Núñez E, Cramer T, et al. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* 317: 248-251. http://dx.doi.org/10.1126/science.1140871

Rosenbloom KR, Armstrong J, Barber GP, Casper J, et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* 43: D670-D681. http://dx.doi.org/10.1093/nar/gku1177

Shannon P, Markiel A, Ozier O, Baliga NS, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13: 2498-2504. http://dx.doi.org/10.1101/gr.1239303

Smith TF, Waterman MS and Sadler JR (1983). Statistical characterization of nucleic acid sequence functional domains. *Nucleic Acids Res.* 11: 2205-2220. http://dx.doi.org/10.1093/nar/11.7.2205

Wang J, Lunyak VV and Jordan IK (2012). Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res.* 40: 511-529. http://dx.doi.org/10.1093/nar/gkr750