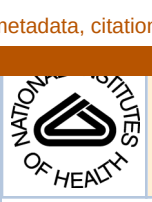

Published in final edited form as:
Nat Methods. ; 9(3): 215–216. doi:10.1038/nmeth.1906.

ChromHMM: automating chromatin state discovery and characterization

Jason Ernst[1,2,3] and **Manolis Kellis**[1,2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

Chromatin state annotation using combinations of chromatin modification patterns has emerged as a powerful approach for discovering regulatory regions and their cell type specific activity patterns, and for interpreting disease-association studies[1-5]. However, the computational challenge of learning chromatin state models from large numbers of chromatin modification datasets in multiple cell types still requires extensive bioinformatics expertise making it inaccessible to the wider scientific community. To address this challenge, we have developed ChromHMM, an automated computational system for learning chromatin states, characterizing their biological functions and correlations with large-scale functional datasets, and visualizing the resulting genome-wide maps of chromatin state annotations.

ChromHMM is based on a multivariate Hidden Markov Model that models the observed combination of chromatin marks using a product of independent Bernoulli random variables[2], which enables robust learning of complex patterns of many chromatin modifications. As input, it receives a list of aligned reads for each chromatin mark, which are automatically converted into presence or absence calls for each mark across the genome, based on a Poisson background distribution. An optional additional input of aligned reads for a control dataset can be used to either adjust the presence or absence threshold, or as an independent input feature **(**Supplementary Note**)**. Alternatively, the user can input files that contain calls from an independent peak caller. By default, chromatin states are analyzed at 200-base pair intervals that roughly approximate nucleosome sizes, but smaller or larger windows can be specified. We have also developed a new parameter initialization procedure that enables relatively efficient inference of comparable models across different numbers of states (Supplementary Note).

ChromHMM then outputs both the learned chromatin state model parameters and the state assignments for each genomic position. The learned emission and transition parameters are returned in both text and image format **(**Fig. 1**)**, automatically grouping states with similar emission parameters or proximal genomic locations, although a user-specified reordering can also be used **(**Supplementary Fig. 1-2, Supplementary Note**)**. ChromHMM enables the study of the likely biological roles of each chromatin state based on enrichment in diverse external annotations and experimental data, shown as heat maps and tables **(**Fig. 1**)**, both for direct genomic overlap and at various distances from a state **(**Supplementary Fig. 3**)**. ChromHMM also generates custom UCSC genome browser tracks[6] showing the resulting

Corresponding Author: Manolis Kellis (manoli@mit.edu).
[3]Current address: Department of Biological Chemistry, University of California Los Angeles, Los Angeles, California, USA

The software is written in Java enabling it to be run on virtually any computer, and is freely available with further documentation at http://compbio.mit.edu/ChromHMM.

chromatin state segmentation in dense view (single color-coded track), or expanded view (each state shown separately) (Fig. 1). All the files ChromHMM produces by default are summarized on a webpage that it also creates (Supplementary Data).

ChromHMM also enables the analysis of chromatin states across multiple cell types. When the chromatin marks are common across the cell types, a common model can be learned by a virtual 'concatenation' of the chromosomes of all cell types. Alternatively a model can be learned by a virtual 'stacking' of all marks across cell types, or independent models can be learned in each cell type. Lastly, ChromHMM supports the comparison of models with different number of states based on correlations in their emission parameters **(**Supplementary Fig. 4**)**.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Massachusetts Institute of Technology Computational Biology group and B. Bernstein for useful discussions related to this work. The work was supported by a NSF Postdoctoral Fellowship 0905968 to JE and grants from the National Institute of Health (NIH 1-RC1-HG005334 and NIH 1 U54 HG004570).

References

1. Day N, et al. Bioinformatics. 2007; 23:1424–1426. [PubMed: 17384021]
2. Ernst J, Kellis M. Nat Biotechnol. 2010; 28:817–825. [PubMed: 20657582]
3. Ernst J, et al. Nature. 2011; 473:43–49. [PubMed: 21441907]
4. Filion GJ, et al. Cell. 2010; 143:212–224. [PubMed: 20888037]
5. Roy S, et al. Science. 2010; 330:1787–1797. [PubMed: 21177974]
6. Kent WJ, et al. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]

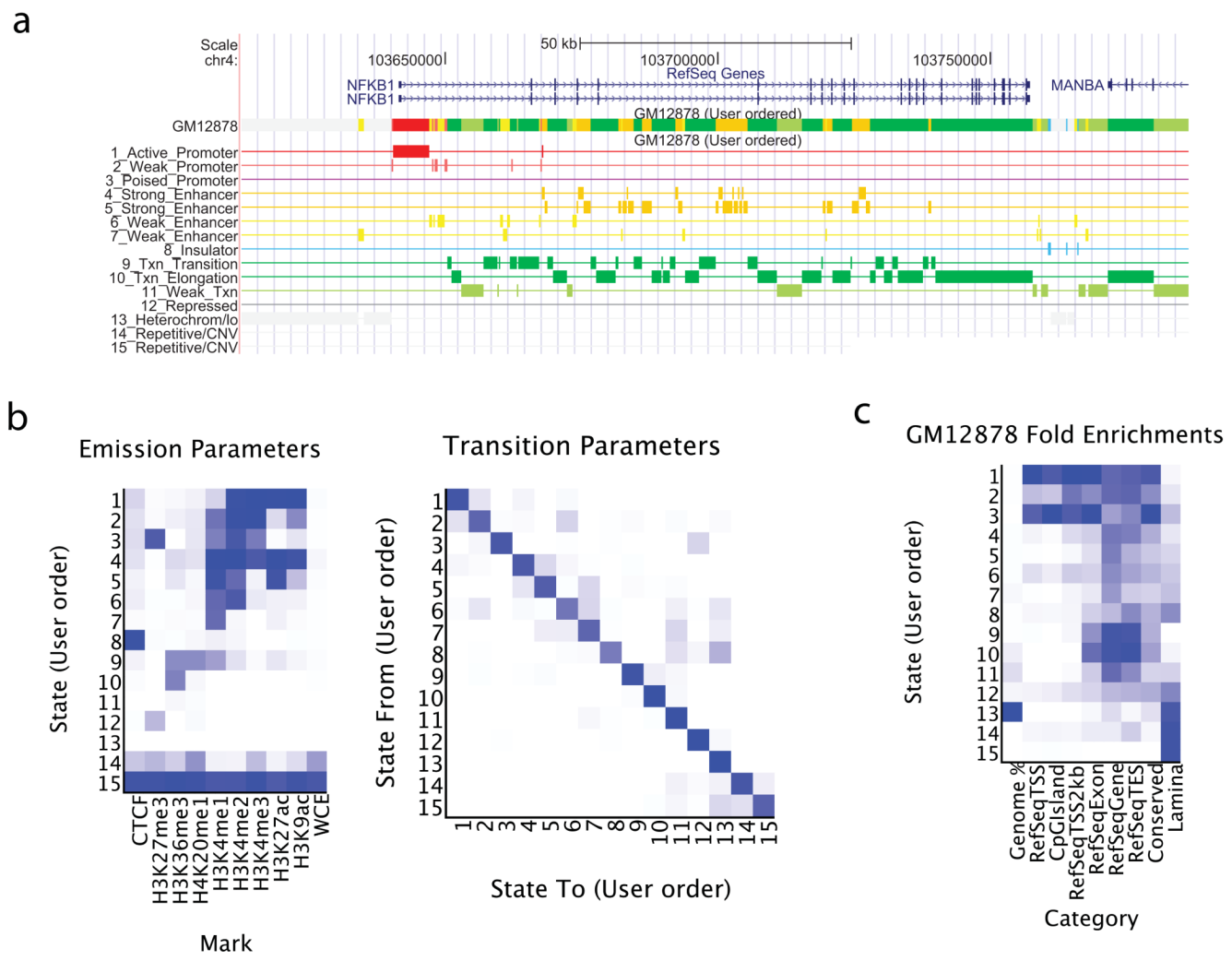

Figure 1. Sample Outputs of ChromHMM

(**a**) Example of state annotation tracks produced from ChromHMM and visualized in the UCSC genome browser[6], including a dense view of the segmentation as a single track (top), and an expanded view of the segmentation showing each state as a separate track (bottom). (**b**) Heat maps automatically produced by ChromHMM show emission (left) and transition (right) parameters. (**c**) Example heat map for state functional enrichments automatically generated by ChromHMM. The columns indicate the relative percentage of the genome represented by each state (first column) and relative fold enrichment for: RefSeq transcription start sites (TSS); CpG Islands; 2000 base pair intervals around the TSS; exons; genes; transcript end sites (TES); evolutionary conservation; and nuclear lamina associated regions **(Supplementary Note)**. **a-c**. Example shown corresponds to a previous model learned across nine cell types[3].