ChromHMM: automating chromatin-state discovery and characterization

To the Editor: 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. To address this challenge, we 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. One can use an optional additional input of aligned reads for a control dataset to either adjust the threshold for present or absent calls, or as an additional input mark. 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 also developed an improved parameter-initialization procedure that enables relatively efficient inference of comparable models across different numbers of states (**Supplementary Note**).

ChromHMM outputs both the learned chromatin-state model parameters and the chromatin-state assignments for each genomic position. The learned emission and transition parameters are returned in both text and image format (**Fig. 1**), automatically grouping chromatin states with similar emission parameters or proximal genomic locations, although a user-specified reordering can also be used (**Supplementary Figs. 1–2** and **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 chromatin state (**Supplementary Fig. 3**). ChromHMM also generates custom genome browser tracks[6] that show the resulting 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 (**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 chromatin states based on correlations in their emission parameters (**Supplementary Fig. 4**).

We wrote the software in Java, which allows it to be run on virtually any computer. ChromHMM and additional documentation is freely available at http://compbio.mit.edu/ChromHMM/.

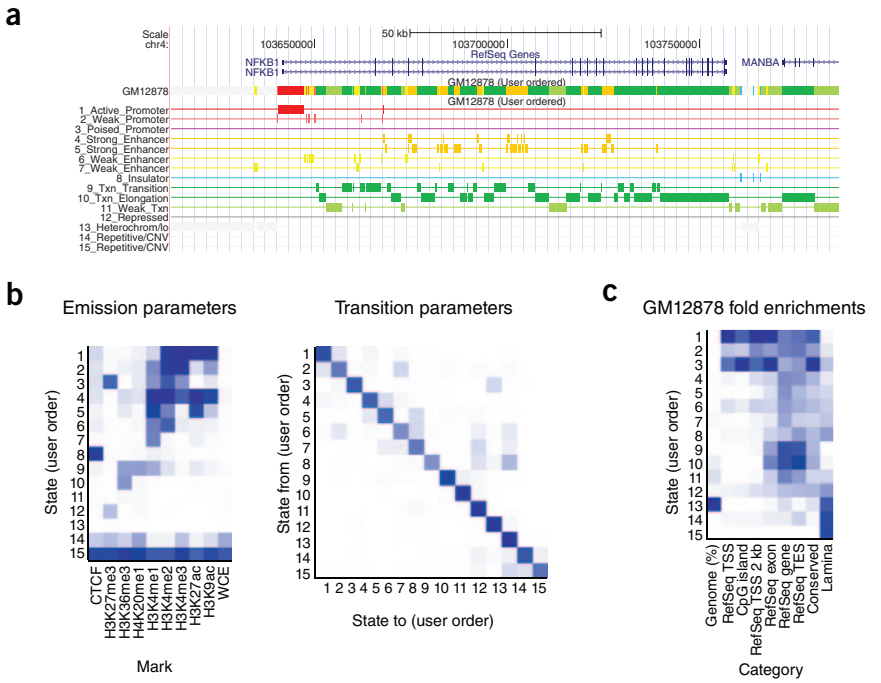

Figure 1 | Sample outputs of ChromHMM. (**a**) Example of chromatin-state annotation tracks produced from ChromHMM and visualized in the UCSC genome browser[6], including dense view (top; single track), expanded view (bottom; separate tracks). (**b**,**c**) Heat maps for model parameters (**b**) and for chromatin-state functional enrichments (**c**). The columns indicate the relative percentage of the genome represented by each chromatin state and relative fold enrichment for several types of annotation. CTCF, CTC-binding factor; WCE, whole-cell extract; TSS, transcription start site; TES, transcript end site; and GM12878 is a lymphoblastoid cell line. Data in this example correspond to a previous model learned across nine cell types[3].

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank members of the Massachusetts Institute of Technology Computational Biology group and B.E. Bernstein for useful discussions related to this work. The work was supported by a US National Science Foundation postdoctoral fellowship 0905968 to J.E. and grants from the US National Institutes of Health (1-RC1-HG005334 to M.K. and 1 U54 HG004570 to B.E. Bernstein).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jason Ernst[1–3] & Manolis Kellis[1,2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. [3]Present address: Department of Biological Chemistry, University of California Los Angeles, Los Angeles, California, USA.
e-mail: manoli@mit.edu

1. Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. *Bioinformatics* **23**, 1424–1426 (2007).
2. Ernst, J. & Kellis, M. *Nat. Biotechnol.* **28**, 817–825 (2010).
3. Ernst, J. *et al. Nature* **473**, 43–49 (2011).
4. Filion, G.J. *et al. Cell* **143**, 212–224 (2010).
5. Roy, S. *et al. Science* **330**, 1787–1797 (2010).
6. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).