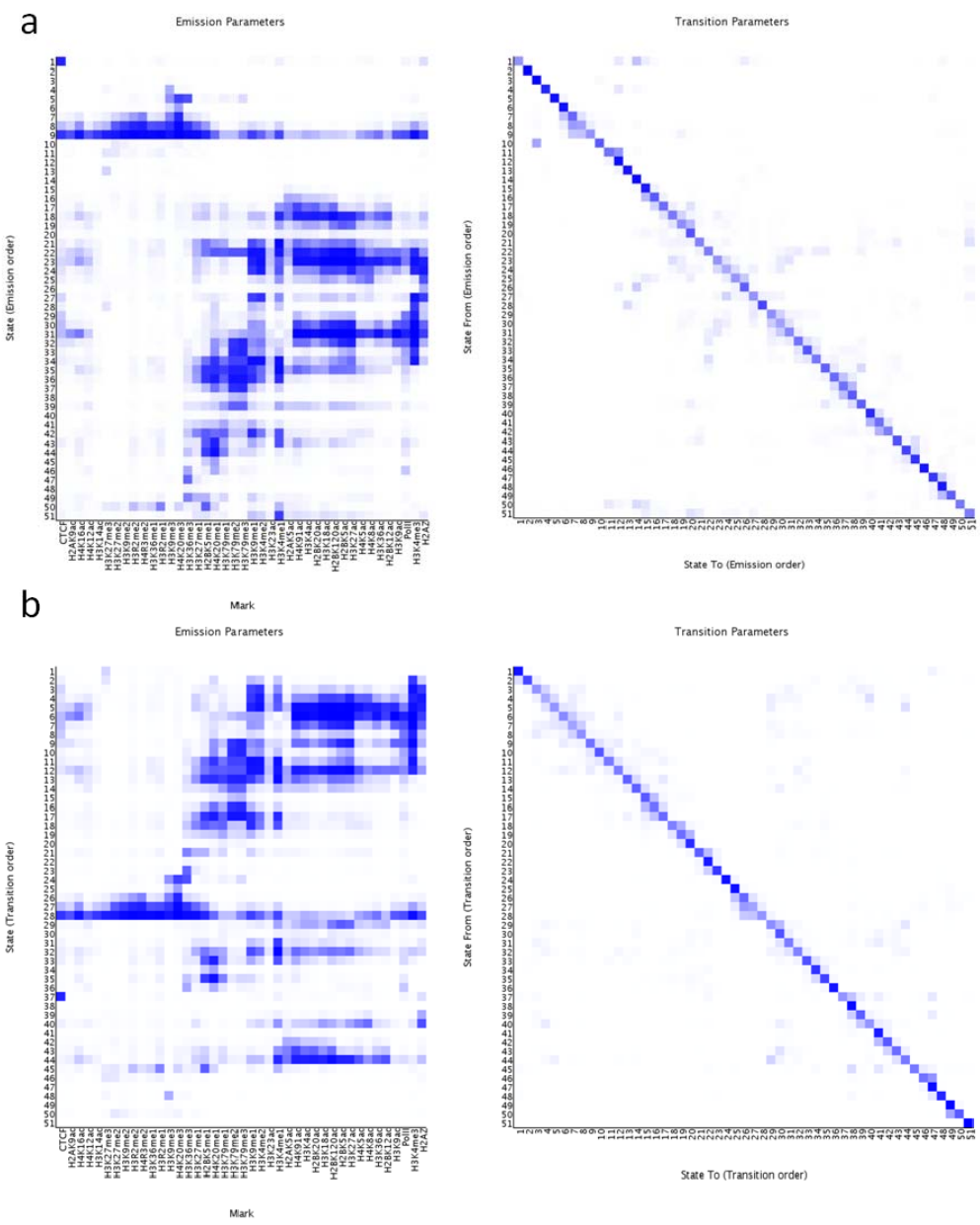
Nature Methods

ChromHMM: automating chromatin-state discovery and characterization

Jason Ernst & Manolis Kellis

Supplementary Figure 1	Automatic State Ordering by Emission and Transition Parameters.
Supplementary Figure 2	Comparing Emission and Transition based Ordering with Previous Manual State Ordering.
Supplementary Figure 3	Example Heat Map of Chromatin State Positional Enrichments.
Supplementary Figure 4	Example Model Emission Parameter Correlation Comparison Heat Map.
Supplementary Note	Use of Control Data in ChromHMM, Model Parameter Initialization, Automatic State Ordering, Data Sets in Enrichment Analysis
Supplementary Data	Example ChromHMM Report

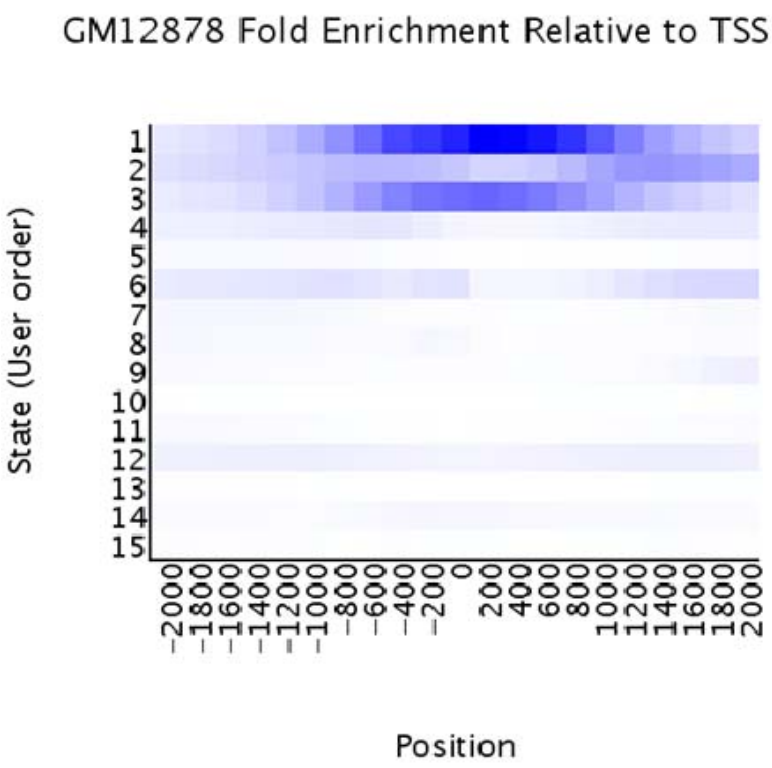


Supplementary Figure 1: Automatic State Ordering by Emission and Transition Parameters. **(a)** Visualization of the emission parameters (left) and transition parameters (right) of a previously published model[2] when the states of the model have been automatically ordered based on the emission parameters. **(b)** The same as in (a) except the states are ordered based on the transition parameters. See **Supplementary Note** for more details about the state ordering. In both cases the columns of the emission matrix have been ordered based on the same procedure to order the states in (a).

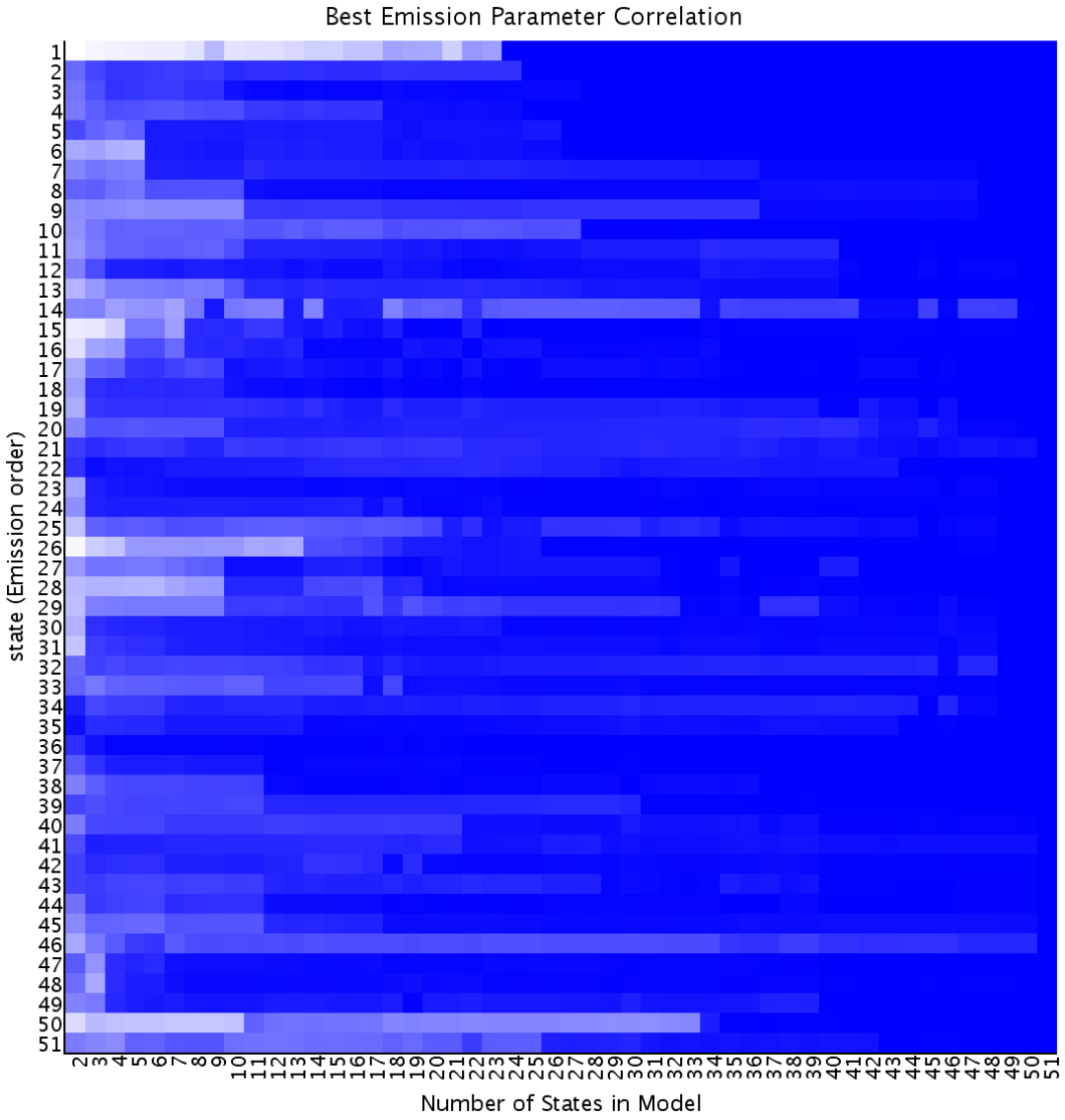
State Emission Order	state (Ernst and Kellis, 2010) Order	State Description
1	39	CTCF island; candidate insulator
2	40	Unmappable
3	41	Heterochr; nuclear lamina; most AT rich
4	47	L1/LTR repeats
5	28	ZNF genes; KAP-1 repressed state
6	48	Satellite repeat
7	49	Satellite repeat; moderate mapping bias
8	50	Satellite repeat; high mapping bias
9	51	Satellite repeat/rRNA; extreme mapping bias
10	42	Heterochr; nuclear lamina; ERVL repeats
11	44	Heterochr; ERVL repeats: lower gene/exon depletion
12	43	Heterochr; lower gene depletion
13	45	Specific repression
14	37	Non-repressive intergenic domains; Alu repeats
15	36	Active intergenic further from enhancers; Alu repeats
16	35	Active intergenic regions not enhancer specific
17	32	Candidate weak distal enhancer
18	29	Cand strong distal enh; higher open chr; higher target expr
19	30	Cand strong distal enh; high open chr; higher target expr
20	34	Proximal to active enhancers; Alu repeats
21	20	Candidate strong enhancer in transcribed regions
22	10	Transcribed promoter; high expr, near TSS
23	1	Promoter upstream high expr; potential enh looping
24	2	Promoter upstream med expr; potential enh looping
25	31	Intergenic H2AZ with open chr/TF binding. Cand. distal enh
26	38	H2AZ specific state
27	3	Promoter upstream low expr; potential enh looping
28	4	Repressed promoter
29	5	TSS low-med expr; most GC rich
30	6	TSS med expr
31	7	TSS high expr
32	8	Transcribed promoter; highest expr, TSS for active genes
33	9	Transcribed promoter; highest expr, downstream
34	11	Transcribed promoter; high expr, downstream
35	12	Transcribed 5' proximal, higher expr, open chr, TF binding
36	13	Transcribed 5' proximal, higher expr, open chr
37	15	Transcribed 5' proximal, high expr
38	16	Transcribed 5' proximal, med expr; Alu repeats
39	14	Transcribed 5' proximal, high expr, open chr
40	19	Transcribed less 5' proximal, lower expr; Alu repeats
41	18	Transcribed less 5' proximal, med expr
42	17	Transcribed less 5' proximal, med expr, open chr
43	21	Spliced exons/GC rich; open chr, TF binding
44	22	Spliced exons/GC rich
45	23	Spliced exons/GC rich; Alu repeats
46	27	End of transcription; exons; high expr
47	25	Transcribed further 5' distal; exons
48	26	Transcribed 5' distal; Alu repeats
49	24	Transcribed 5' distal; exons
50	46	Simple repeats (CA)n, (TG)n; L1/LTR repeats
51	33	Candidate distal enhancer

State Transition Order	state (Ernst and Kellis, 2010) Order	State Description
1	45	Specific repression
2	4	Repressed promoter
3	3	Promoter upstream low expr; potential enh looping
4	2	Promoter upstream med expr; potential enh looping
5	1	Promoter upstream high expr; potential enh looping
6	7	TSS high expr
7	6	TSS med expr
8	5	TSS low-med expr; most GC rich
9	8	Transcribed promoter; highest expr, TSS for active genes
10	9	Transcribed promoter; highest expr, downstream
11	11	Transcribed promoter; high expr, downstream
12	10	Transcribed promoter; high expr, near TSS
13	12	Transcribed 5' proximal, higher expr, open chr, TF binding
14	14	Transcribed 5' proximal, high expr, open chr
15	16	Transcribed 5' proximal, med expr; Alu repeats
16	15	Transcribed 5' proximal, high expr
17	13	Transcribed 5' proximal, higher expr, open chr
18	17	Transcribed less 5' proximal, med expr, open chr
19	18	Transcribed less 5' proximal, med expr
20	19	Transcribed less 5' proximal, lower expr; Alu repeats
21	24	Transcribed 5' distal; exons
22	26	Transcribed 5' distal; Alu repeats
23	25	Transcribed further 5' distal; exons
24	28	ZNF genes; KAP-1 repressed state
25	48	Satellite repeat
26	49	Satellite repeat; moderate mapping bias
27	50	Satellite repeat; high mapping bias
28	51	Satellite repeat/rRNA; extreme mapping bias
29	30	Cand strong distal enh; high open chr; higher target expr
30	34	Proximal to active enhancers; Alu repeats
31	33	Candidate distal enhancer
32	20	Candidate strong enhancer in transcribed regions
33	21	Spliced exons/GC rich; open chr, TF binding
34	23	Spliced exons/GC rich; Alu repeats
35	22	Spliced exons/GC rich
36	27	End of transcription; exons; high expr
37	39	CTCF island; candidate insulator
38	37	Non-repressive intergenic domains; Alu repeats
39	38	H2AZ specific state
40	31	Intergenic H2AZ with open chr/TF binding. Cand. distal enh
41	36	Active intergenic further from enhancers; Alu repeats
42	35	Active intergenic regions not enhancer specific
43	32	Candidate weak distal enhancer
44	29	Cand strong distal enh; higher open chr; higher target expr
45	46	Simple repeats (CA)n, (TG)nL1/LTR repeats
46	44	Heterochr; ERVL repeats: lower gene/exon depletion
47	43	Heterochr; lower gene depletion
48	47	L1/LTR repeats
49	41	Heterochr; nuclear lamina; most AT rich
50	42	Heterochr; nuclear lamina; ERVL repeats
51	40	Unmappable

Supplementary Figure 2: Comparing Emission and Transition based Ordering with Previous Manual State Ordering. The figure shows the state ordering based on the emission parameters (left) and transition parameters (right) next to the manual state ordering of a previously presented model[2]. State colors and descriptions are as previously presented[2]. Many of the states near each other under the manual ordering are also near each other under the automatic orderings. Differences are also biologically reasonable for example the ZNF gene associated state (state 28) appearing with other repetitive states in the emission based ordering, and the specific repression state (state 45) appearing next to the repressed promoter (state 4) under the transition ordering.



Supplementary Figure 3: Example Heat Map of Chromatin State Positional Enrichments. An example heat map produced by ChromHMM showing the relative genome-wide enrichment of each state at positions within 2000 base pairs (bp) of RefSeq transcription start sites at a 200 bp resolution in the cell type GM12878 for a previously presented model[3].



Supplementary Figure 4: Example Model Emission Parameter Correlation Comparison Heat Map. An example heat map produced by ChromHMM comparing a set of models with different numbers of states. Each row corresponds to a state from a 51 state model[2], and each column a model with a different number of states. The intensity of a cell indicates the maximum emission parameter correlation of any state in the model of the column with the state of the row from the 51 state model. The states are numbered to match the emission parameter ordering in **Supplementary Figure 1a.** From this heat map one can quickly observe for instance that state 1 which is highly specific to CTCF is not well captured in models in the set with fewer than 24 states.

Supplementary Note

Use of Control Data in ChromHMM

Control data can be used by ChromHMM as an input feature directly in the model which can help isolate regions of copy number variation and repeat associated artifacts[3] or it can be used to locally adjust the binarization threshold. The binarization threshold is determined based on an expected number of reads per bin, a significance threshold based on a Poisson background distribution, and also optionally requiring a minimum fold enrichment over the expected number of reads. Without control data the expected number of reads per bin is the genome-wide average number of reads per bin. When using an input control this expectation is multiplied by the local enrichment for control reads. When computing the local enrichment for control reads, a pseudocount, default of 1, is first added to the count for each bin to smooth the data from 0. The local enrichment for control data is then computed as the total control counts within a fixed number of bins in both directions, default is five 200-base pair bins, divided by the expected number control counts in the same number of bins if control reads were uniformly distributed.

Model Parameter Initialization

When learning the parameters of an HMM through likelihood maximization an initial set of parameters needs to be specified. The local maximum and learned model obtained can depend on the choice of initial parameters. We have previously shown that for a fixed model size, that while many of the chromatin states recovered are largely invariant to the initial set of parameters not all are[2]. The fact that different independent initializations can lead to different states being recovered can confound comparing models of different sizes. To address this issue we previously proposed a two-pass nested initialization strategy that pruned states from a high scoring model based on states obtained from models learned from random initializations in the first pass[2]. The approach however had the drawbacks of requiring a maximum number of states to consider to be specified a priori and being more computationally expensive than a single pass approach.

ChromHMM supports a new initialization strategy which can be used to learn models from a nested set of emission parameters in parallel directly in a single pass. The idea behind the approach is to iteratively partition the bins of the genome in an informative way as defined by an entropy measure. Each iteration refines a previous partition based on the presence of a selected mark. At each iteration, an emission vector corresponding to the mark frequencies in a newly created subset of bins is added to the nested set of emission parameters. Opposed to directly working off of the binary present or absent vector associated with each bin, spatial

information is incorporated into the initialization procedure by only counting a mark as being present at a bin if it is also present at the next bin. In more detail and formally:

Inputs to the initialization procedure:
Let M be the number of marks in the model
Let K be the number of states in the model
Let C denote the set of chromosomes
Let T_c denote the number of bins in chromosome c in C
Let c_t denote the bin t on chromosome c
Let v_{c_t} denote the observation vector at position t on the chromosome c
Let $v_{c_t,m}$ denote the binary 0/1 observation for the m^{th} mark
Let α be a smoothing constant with a default value of 0.02

Outputs of the parameter initialization procedure:
Let $p_{k,m}$ denote the emission probability in state k for mark m.
Let $b_{i,j}$ denote the transition probability from state i to state j.
Let a_i denote the probability of starting in state i.

Parameter initialization procedure:

1. Create a smoothed emission vector corresponding to all marks being absent:
   $p_{1,m} = \frac{\alpha}{2}$ for $m = 1, \ldots, M$
2. Create a transformed set of observation vectors that only retains '1' calls observed at consecutive bins:
   Define M element vectors d_{c_t} for each c in C and $t = 1, \ldots, T_c - 1$
   where $d_{c_t,m} = v_{c_t,m} \times v_{c_{t+1},m}$ for $m = 1, \ldots, M$
3. Initialize all non-last bins to initially be in group 1:
   Set $s_{c_t} = 1$ for each c in C and $t = 1, \ldots, T_c - 1$
4. For $i = 2$ to K do
   a. Find the best group and mark to split on to increase total entropy
      that is select $u < i$ and m that maximizes:
      $$(h_{u,m,0} + h_{u,m,1}) log_2(h_{u,m,0} + h_{u,m,1}) - (h_{u,m,0}) log_2(h_{u,m,0}) - (h_{u,m,1}) log_2(h_{u,m,1})$$
      where:
      $$h_{u,m,v} = \frac{|\{c_t | (s_{c_t} = u) \wedge (d_{c_t,m} = v)\}|}{\sum_{c \in C}(T_c - 1)}$$
   b. Create a new group i containing the elements in group u which contain mark m that is:

For all c_t such that if $(s_{c_t} = u) \wedge (d_{c_t,m} = 1)$ let $s_{c_t} = i$

c. Create an emission vector based on a smoothed frequency of the mark being present in the newly created group:

$$p_{i,m} = (1 - \alpha) \times \frac{|\{c_t | (s_{c_t} = i) \wedge (d_{c_t,m} = 1)\}|}{|\{c_t | s_{c_t} = i\}|} + \frac{\alpha}{2}$$

end for

5. Initialize transition parameters based on the final group assignments:

Let $b_{i,j} = (1 - \alpha) \times \frac{|\{c_t | s_{c_t} = i \wedge s_{c_{t+1}} = j\}|}{|\{c_t | s_{c_t} = i\}|} + \frac{\alpha}{K}$ for $i = 1, \ldots, K$ and $j = 1, \ldots, K$

6. Initialize the initial probability parameters based on the final group assignments:

Let $a_i = (1 - \alpha) \times \frac{|\{c | s_{c_1} = i\}|}{|C|} + \frac{\alpha}{K}$ for $i = 1, \ldots, K$

The above procedure is limited in the values of K it can handle by the number of unique d_{c_t} vectors, but in practice the desired values of K will generally be less than this constraint.

ChromHMM also supports initializing the parameters randomly from uniform distributions, and the nested state pruning initialization method previously described[2] generalized to allow correlation or euclidean based distance and smoothing of the parameters away from 0.

Automatic State Ordering

ChromHMM offers the ability to automatically order the states of a model based on either the emission or transition parameters. ChromHMM determines the state order based on an approximation algorithm to minimize the total distance between consecutive states in the ordering.

If ordering is based on the emission parameters, then the distance between two states is defined as

$$d(x,y) = \sqrt{\left(1 - \rho\left(p_x, p_y\right)\right)}$$

where ρ is the pearson correlation coefficient and p_x is the emission parameters in state x.

When ordering based on the transition parameters the distance between two states is defined as

$$d(x,y) = 2 - (b_{x,y} + b_{y,x})$$

where $b_{x,y}$ is defined as the probability of transitioning from state x to state y.

The approximation algorithm considers each state as the first state in the order and then greedily selects the nearest unselected state to it, and continues selecting the closest state to the last selected state among any unselected state. The ordering resulting from the choice of first state that has the smallest total distance is used.

Data Sets in Enrichment Analysis

In **Figure 1c** in the main text, the external data are RefSeq transcription start sites (TSS)[7], CpG Islands, regions within 2000 base pairs of a RefSeq TSS, RefSeq Exons, RefSeq genes, RefSeq transcript end sites, SiPhy omega conserved elements[8], and nuclear lamina associated domains[9] all obtained from the UCSC genome browser[6].

Supplementary References

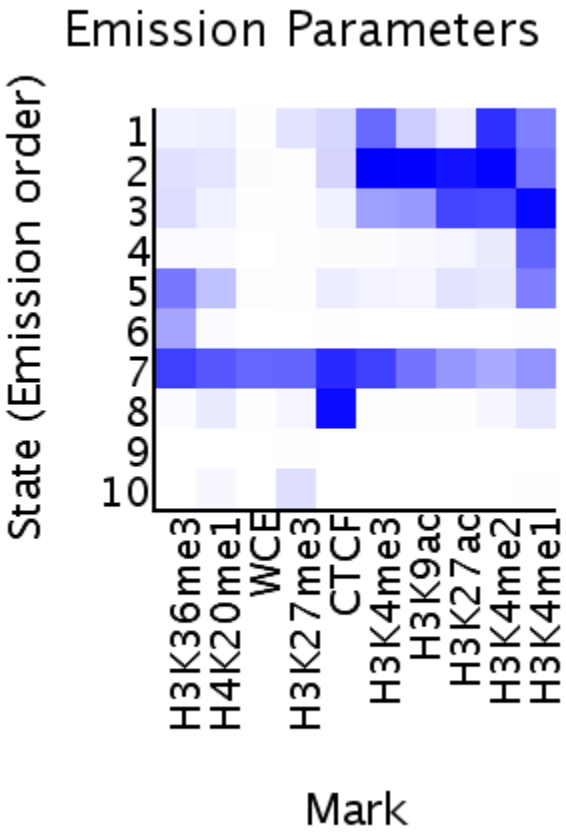
7. Pruitt, K.D., Tatusova, T., Klimke, W., & Maglott, D.R. *Nucleic Acids Res* **37(Database issue)**: D32-36 (2009).
8. Lindblad-Toh, K. *et al. Nature* **478**: 476-482 (2011).
9. Guelen, L. *et al*. *Nature* **453**: 948-951 (2008).

Supplementary Data

ChromHMM Report

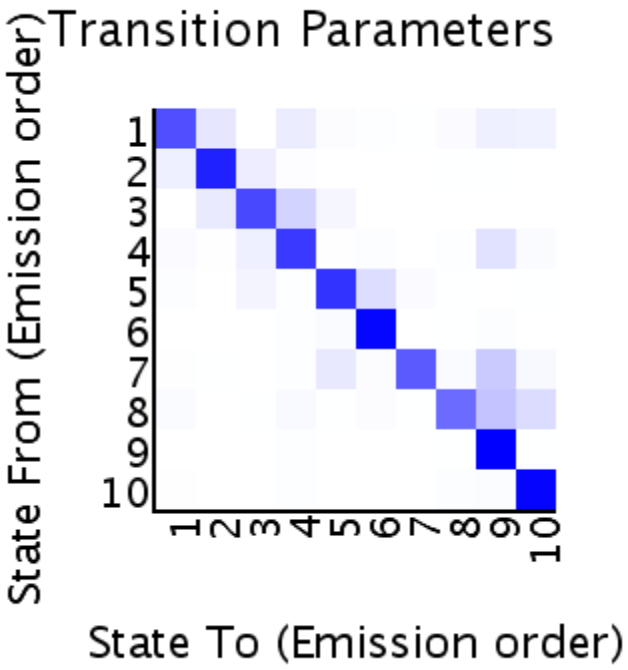
Input Directory: SAMPLEDATA_HG18
Output Directory: OUTPUTSAMPLE
Number of States: 10
Assembly: hg18
Full ChromHMM command: LearnModel SAMPLEDATA_HG18 OUTPUTSAMPLE 10 hg18

Model Parameters



- Emission Parameter SVG File
- Emission Parameter Tab-Delimited Text File

Nature Methods: doi:10.1038/nmeth.1906



- Transition Parameter SVG File
- Transition Parameter Tab-Delimited Text File

- All Model Parameters Tab-Delimited Text File

Genome Segmentation Files

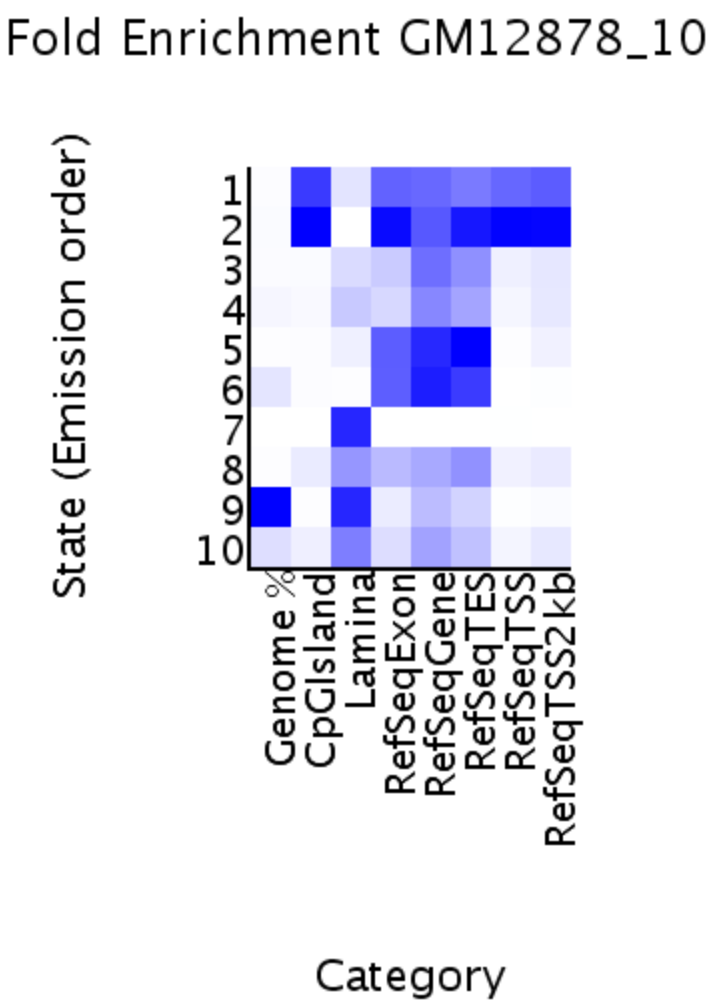
- GM12878_10 Segmentation File (Four Column Bed File)
- K562_10 Segmentation File (Four Column Bed File)

Custom Tracks for loading into the UCSC Genome Browser:
- GM12878_10 Browser Custom Track Dense File
- GM12878_10 Browser Custom Track Expanded File
- K562_10 Browser Custom Track Dense File
- K562_10 Browser Custom Track Expanded File

State Enrichments

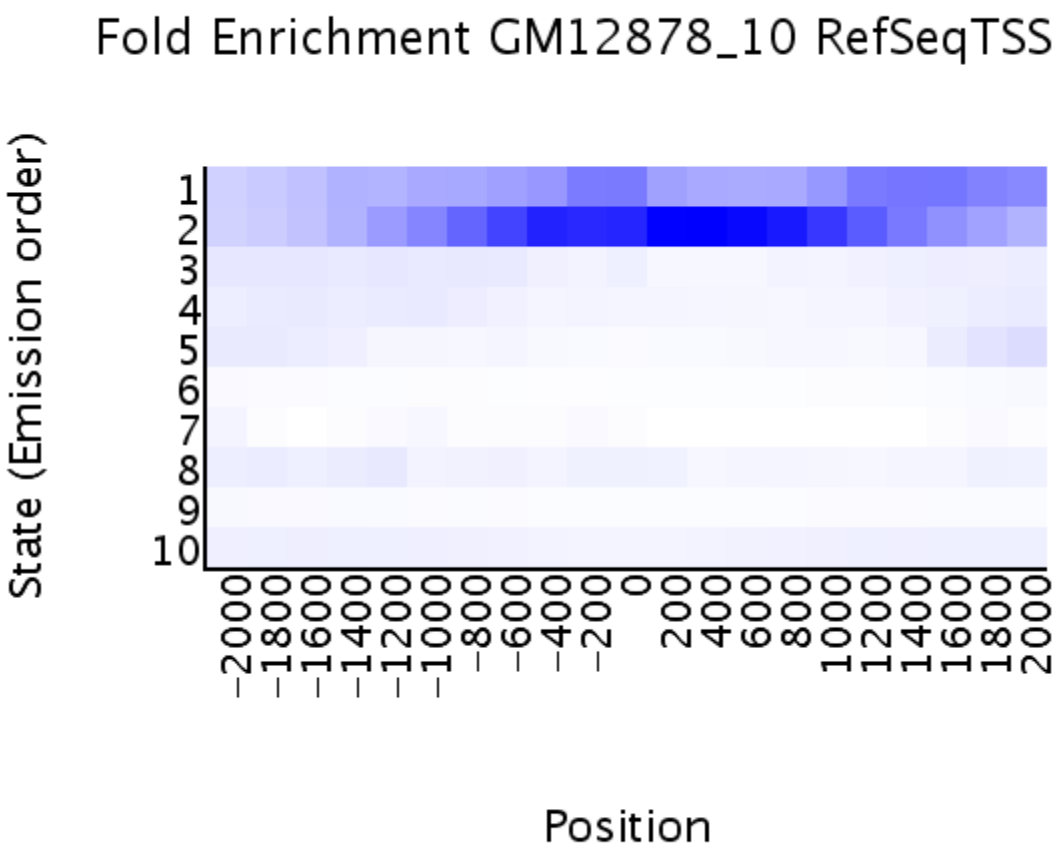
GM12878_10 Enrichments

Nature Methods: doi:10.1038/nmeth.1906



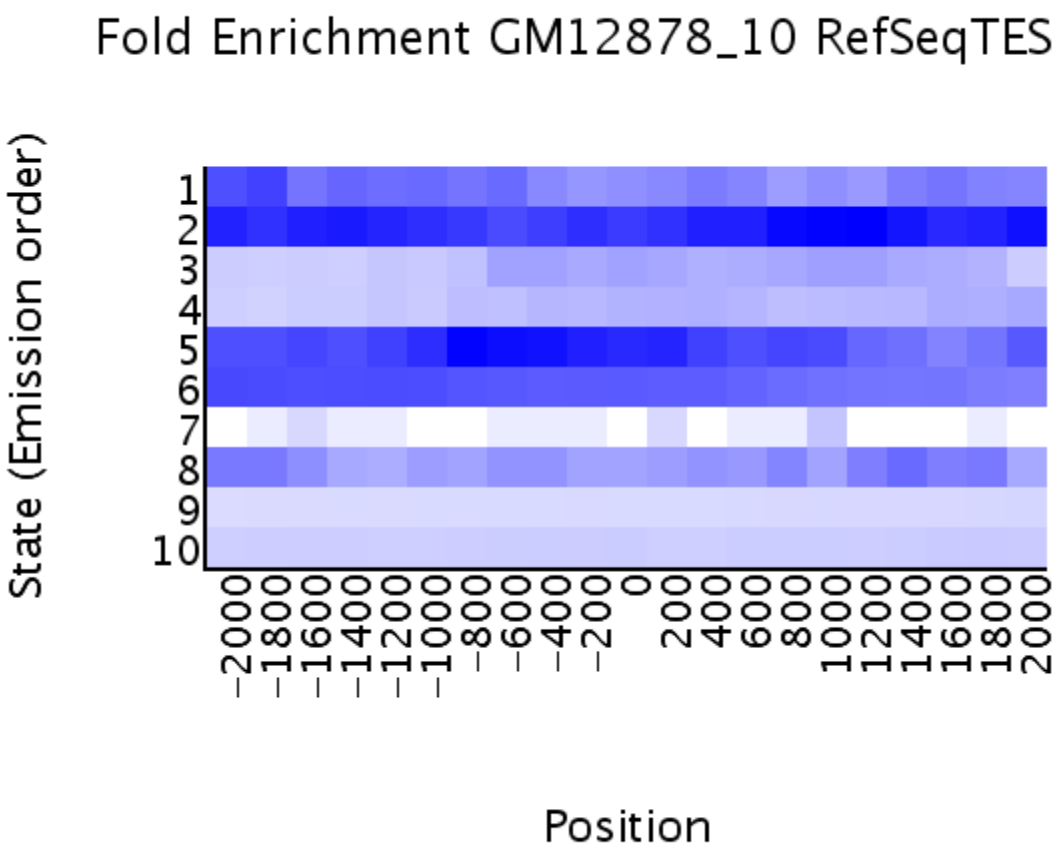
- GM12878_10 Overlap Enrichment SVG File
- GM12878_10 Overlap Enrichment Tab-Delimited Text File

Nature Methods: doi:10.1038/nmeth.1906



- GM12878_10_RefSeqTSS_neighborhood Enrichment SVG File
- GM12878_10_RefSeqTSS_neighborhood Enrichment Tab-Delimited Text File

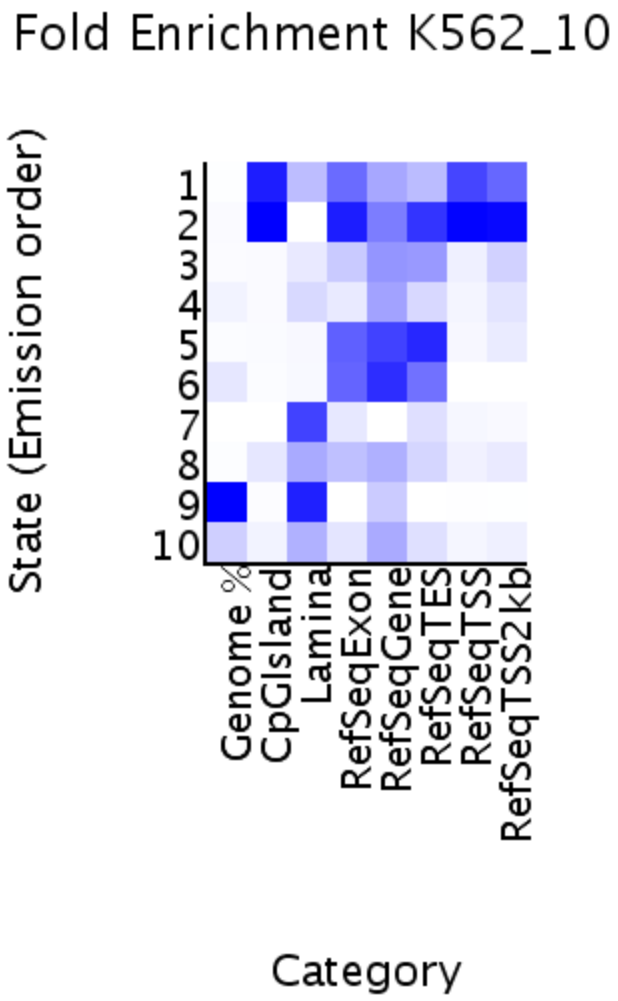
Nature Methods: doi:10.1038/nmeth.1906



- GM12878_10_RefSeqTES_neighborhood Enrichment SVG File
- GM12878_10_RefSeqTES_neighborhood Enrichment Tab-Delimited Text File

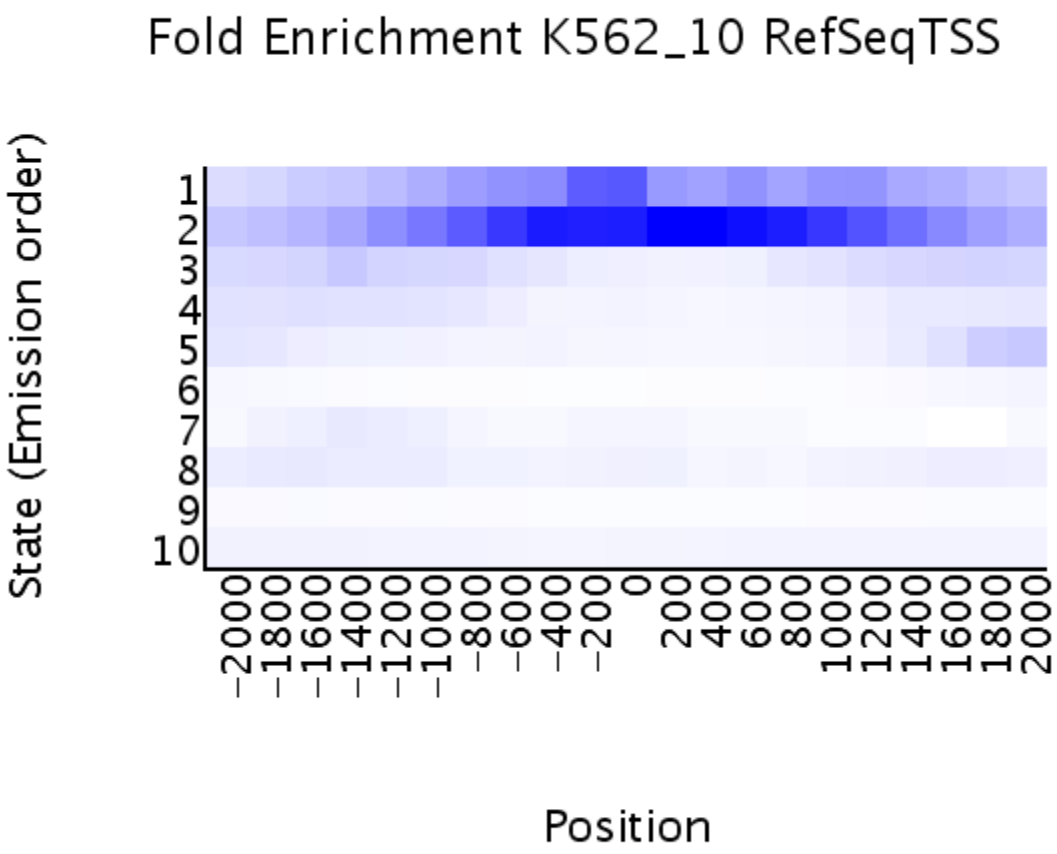
K562_10 Enrichments

Nature Methods: doi:10.1038/nmeth.1906



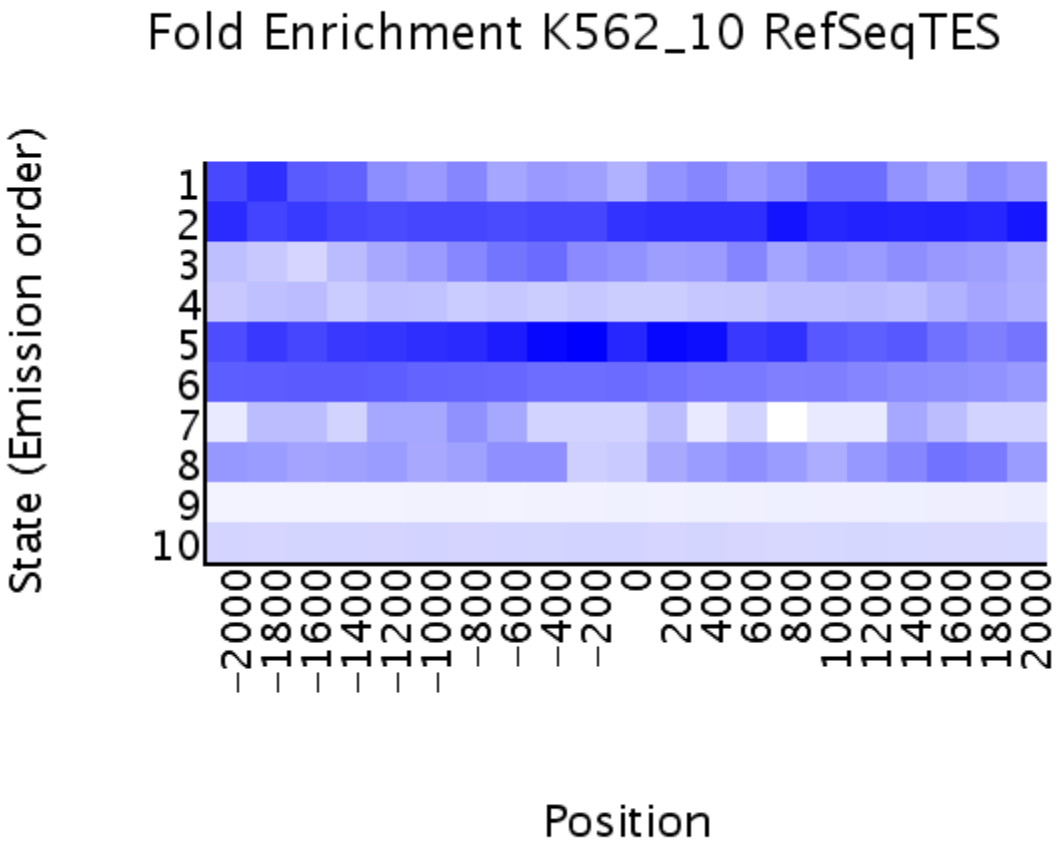
- K562_10 Overlap Enrichment SVG File
- K562_10 Overlap Enrichment Tab-Delimited Text File

Nature Methods: doi:10.1038/nmeth.1906



- K562_10_RefSeqTSS_neighborhood Enrichment SVG File
- K562_10_RefSeqTSS_neighborhood Enrichment Tab-Delimited Text File

Nature Methods: doi:10.1038/nmeth.1906



- K562_10_RefSeqTES_neighborhood Enrichment SVG File
- K562_10_RefSeqTES_neighborhood Enrichment Tab-Delimited Text File

Nature Methods: doi:10.1038/nmeth.1906