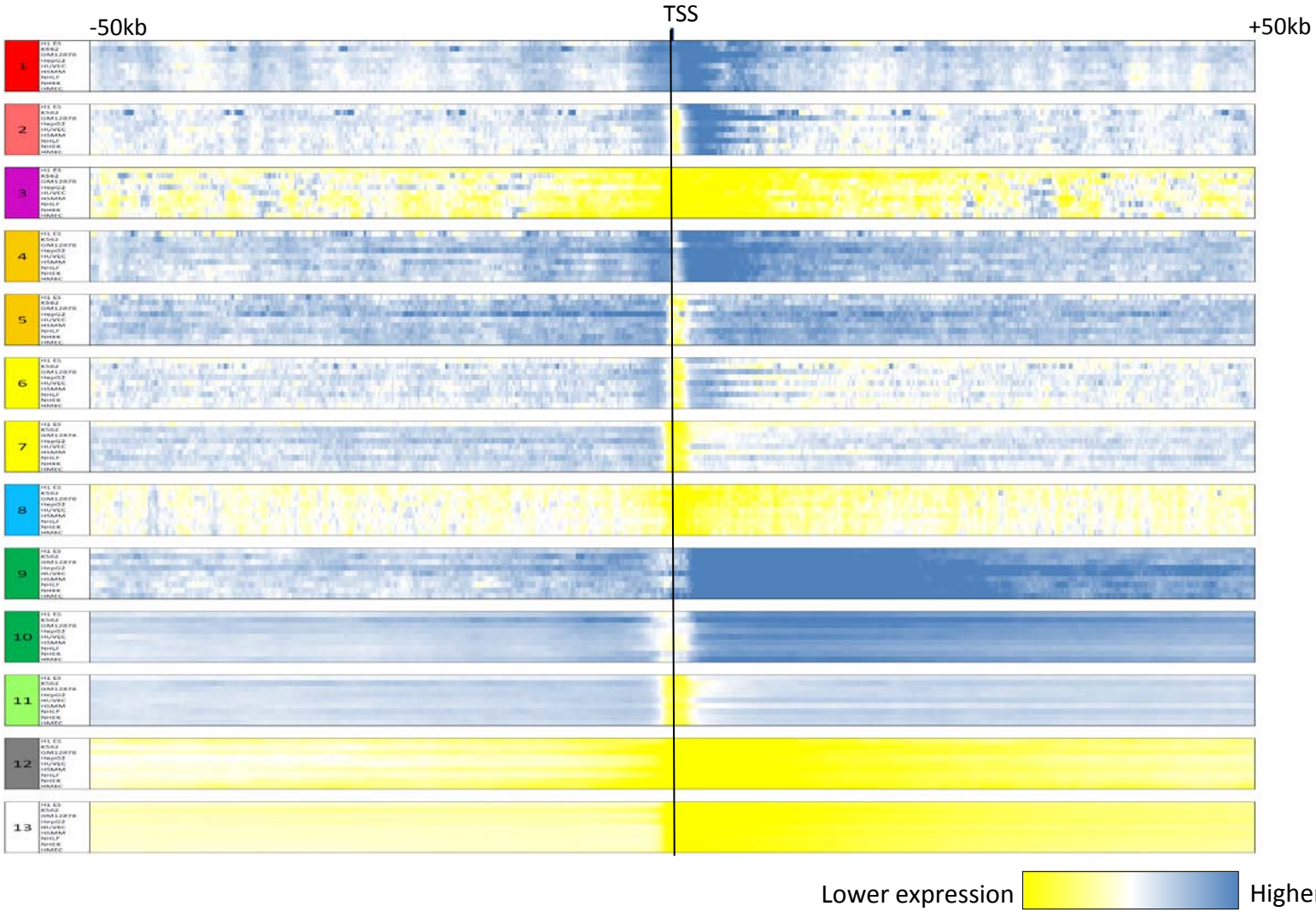
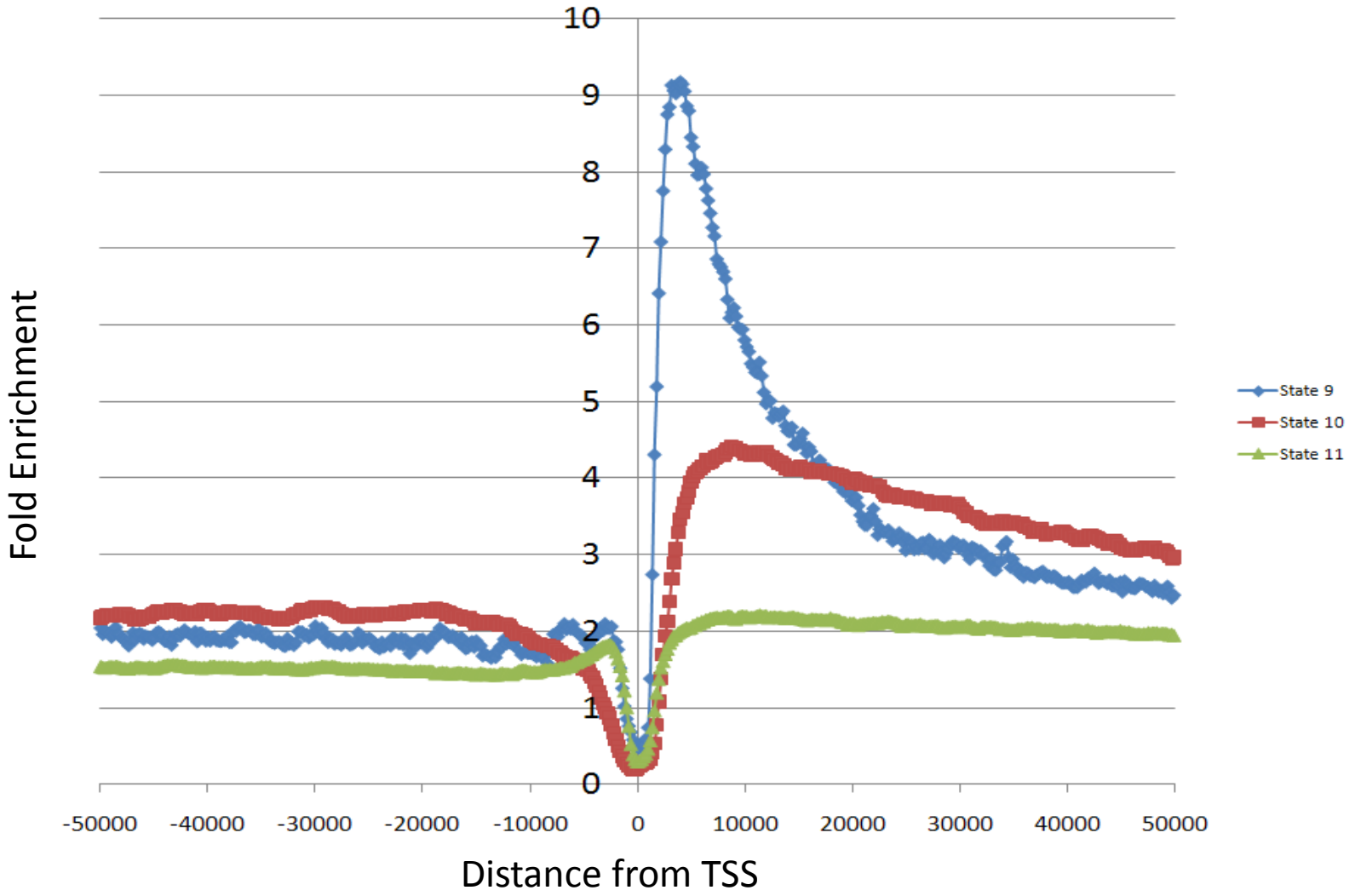
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	83	11	0	4	0	2	0	0	0	0	0	0	0	0	0
2	11	62	1	2	0	17	0	1	2	0	2	0	2	0	0
3	0	4	78	0	0	1	0	1	0	0	0	17	0	0	0
4	4	2	0	75	15	3	1	0	1	0	0	0	0	0	0
5	0	0	0	9	70	1	16	0	2	1	1	0	0	0	0
6	1	10	0	2	1	52	16	1	0	0	6	1	10	0	0
7	0	0	0	0	7	7	62	1	0	0	16	0	7	0	0
8	0	1	0	0	0	1	2	50	0	0	12	4	30	0	0
9	0	1	0	1	2	0	0	0	87	6	3	0	0	0	0
10	0	0	0	0	0	0	0	0	1	96	3	0	0	0	0
11	0	0	0	0	0	0	3	1	0	1	95	0	0	0	0
12	0	0	1	0	0	0	0	1	0	0	0	95	3	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0
14	0	0	0	0	0	0	0	0	0	0	5	2	6	80	7
15	0	0	0	0	0	0	0	0	0	1	4	1	15	12	65

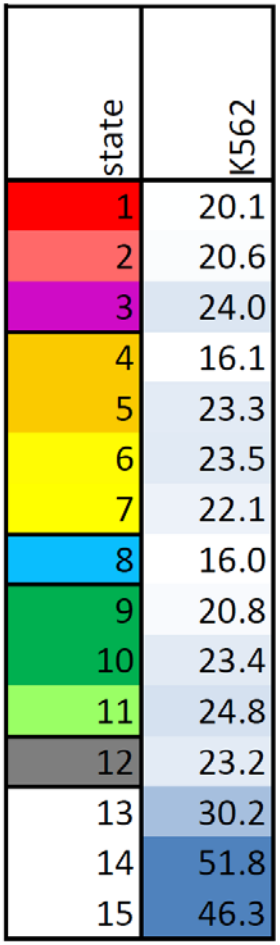
Supplementary Figure 1: HMM Transition Matrix. The above table indicates the transition probabilities of the HMM model. Through the transition probabilities the HMM is able to model spatial information about the data. Each row of the table corresponds to one state. The values in the row indicate the probability (multiplied by 100) that at the next 200 bp interval along the chromosome there will be a transition to the state of the column. Boxes are shown around the subgroups of states. From this matrix one can observe that states 10-13 have low probability of transitioning to a different state, which enables the model to capture the broader domains that these states represent.



Supplementary Figure 2: Chromatin states and gene expression. The above figure illustrates the relationship between the chromatin-states and the expression level of a gene, as a function of the position of the chromatin state relative to RefSeq TSS up to 50kb for the set of genes on the microarray. Each larger row corresponds to a chromatin state (the two repeat associated states are omitted, due to the small number of proximal genes) . Within each larger row are smaller rows corresponding to each of the nine cell types. A smaller rows shows for the presence of the chromatin state in that cell type, for each position relative to the TSS, the average expression level of those corresponding genes. Blue indicates higher expression, white intermediate expression values, and yellow low or no expression based on the raw log expression signal. For instance one can observe from this figure that the presence of active promoter state (state 1) over the TSS corresponds to a highly expressed gene, while in contrast the presence of the inactive/poised promoter state (state 3) over the TSS corresponds to a low or non expressed genes. The presence of the strong transcribed states (states 9-10) particularly downstream of the TSS corresponds to higher expression levels even for locations far from the TSS. In contrast the presence of states 12-13 downstream of the TSS strongly correlates with lower expression levels. The presence of enhancer states (States 4-7) are generally associated with higher expression of genes except when found in promoter regions. One can also observe that the strong enhancer states (States 4-5) have even greater correlation with expression than the weaker enhancer states (States 6 and 7).

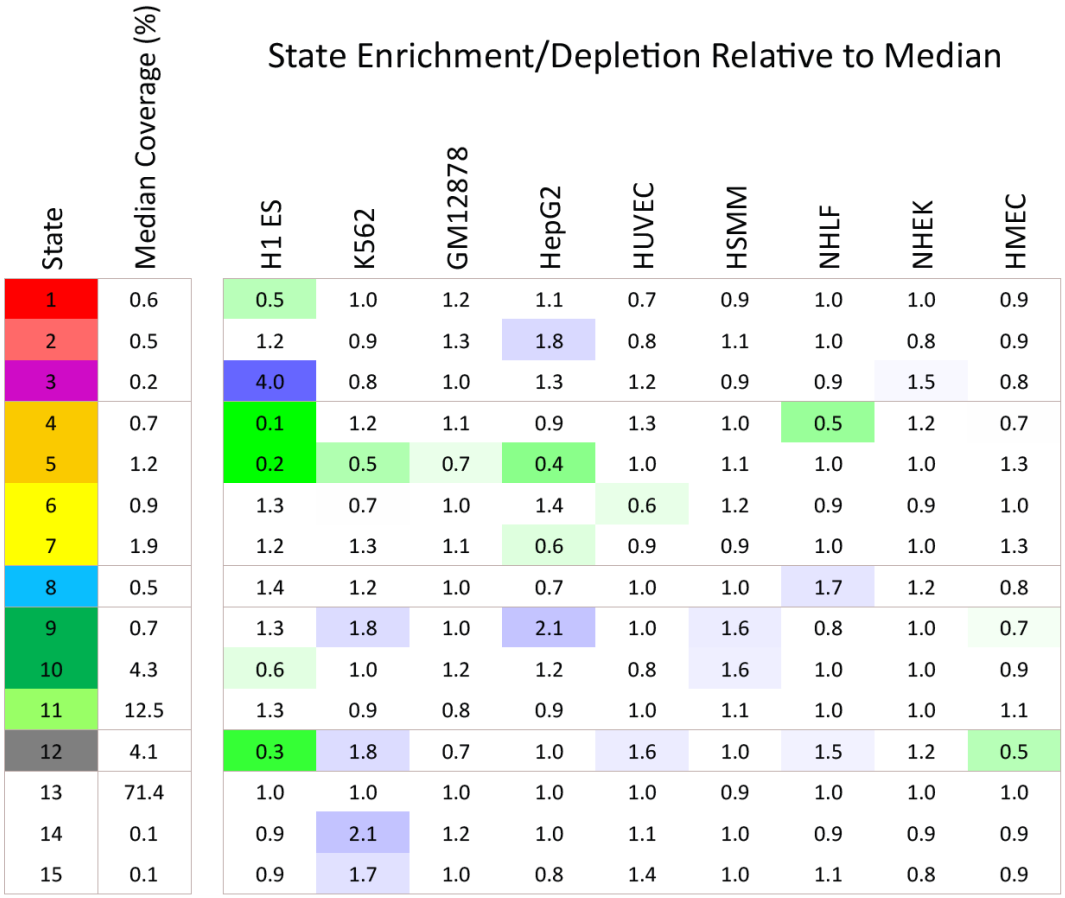


Supplementary Figure 3: Positional Enrichment of Transcription Associated States. Comparison of the fold enrichment for the three transcription associated states relative to TSS. As can be seen State 9 has greater enrichment closer to the 5' end of genes while state 10 has a higher enrichment in positions more distal to the 5' end of the genes, consistent with the interpretation that state 9 is more associated with transcriptional transition and state 10 with elongation. Both states have stronger enrichments than the weak transcribed state (State 11). The enrichment is based on the same set of genes as used in **Supplementary Fig. 2**.



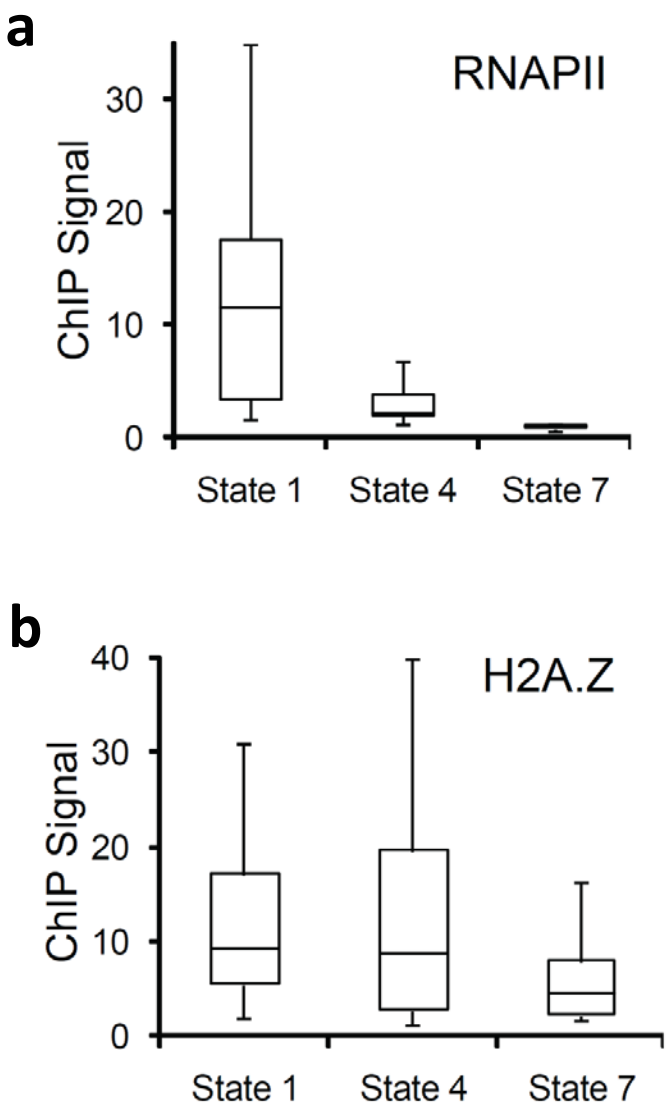
state	K562
1	20.1
2	20.6
3	24.0
4	16.1
5	23.3
6	23.5
7	22.1
8	16.0
9	20.8
10	23.4
11	24.8
12	23.2
13	30.2
14	51.8
15	46.3

Supplementary Figure 4. H3K9me3 enriched locations relative to whole cell extract in K562 cells. For each state, this table reports the percentage of 200 bp intervals that reside within regions of significant H3K9me3 enrichment relative to whole cell extract (p-value < 10^{-4}). Enriched regions were called using a sliding 5 kb window, and significance was computed based on a Poisson distribution where the mean parameter of the distribution was based on the genome wide average of H3K9me3 signal adjusted proportional to the local input enrichment. The results suggest H3K9me3 tends to enrich within states 13-15 .

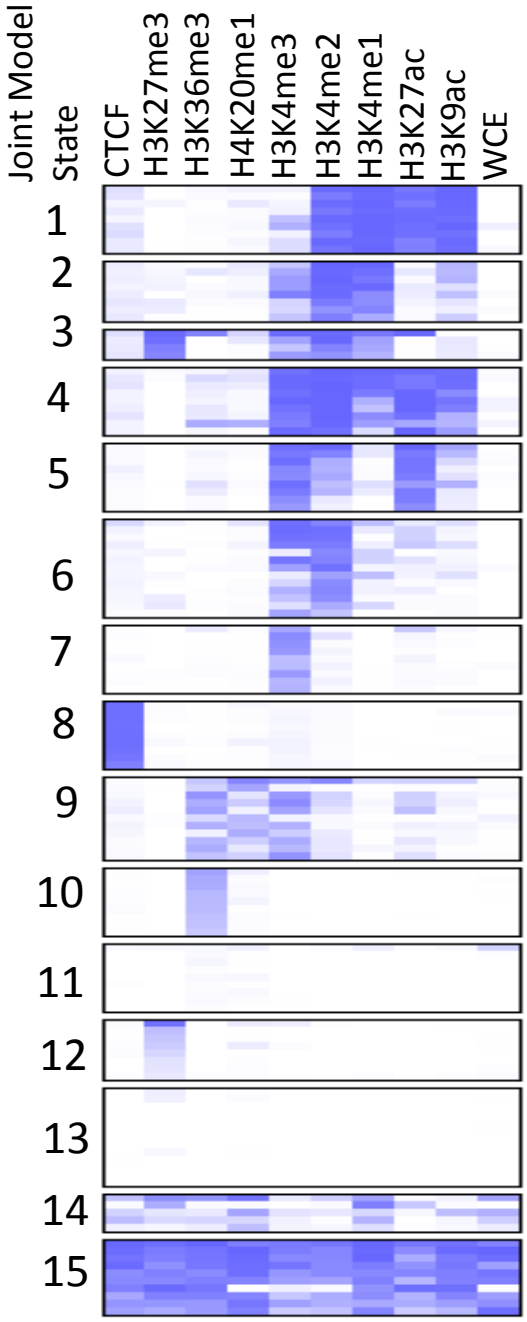


State	Median Coverage (%)	State Enrichment/Depletion Relative to Median								
		H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
1	0.6	0.5	1.0	1.2	1.1	0.7	0.9	1.0	1.0	0.9
2	0.5	1.2	0.9	1.3	1.8	0.8	1.1	1.0	0.8	0.9
3	0.2	4.0	0.8	1.0	1.3	1.2	0.9	0.9	1.5	0.8
4	0.7	0.1	1.2	1.1	0.9	1.3	1.0	0.5	1.2	0.7
5	1.2	0.2	0.5	0.7	0.4	1.0	1.1	1.0	1.0	1.3
6	0.9	1.3	0.7	1.0	1.4	0.6	1.2	0.9	0.9	1.0
7	1.9	1.2	1.3	1.1	0.6	0.9	0.9	1.0	1.0	1.3
8	0.5	1.4	1.2	1.0	0.7	1.0	1.0	1.7	1.2	0.8
9	0.7	1.3	1.8	1.0	2.1	1.0	1.6	0.8	1.0	0.7
10	4.3	0.6	1.0	1.2	1.2	0.8	1.6	1.0	1.0	0.9
11	12.5	1.3	0.9	0.8	0.9	1.0	1.1	1.0	1.0	1.1
12	4.1	0.3	1.8	0.7	1.0	1.6	1.0	1.5	1.2	0.5
13	71.4	1.0	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
14	0.1	0.9	2.1	1.2	1.0	1.1	1.0	0.9	0.9	0.9
15	0.1	0.9	1.7	1.0	0.8	1.4	1.0	1.1	0.8	0.9

Supplementary Figure 5. Prevalence of each chromatin state compared across the 9 cell types. Blue shading indicates states that are enriched in a given cell type; green shading indicates states that are depleted relevant to the median for the corresponding column. The states are relatively stable across the different cell types, with exception that ES cells are enriched for 'inactive/poised promoter' (State 3) and depleted for 'strong enhancer' and 'Polycomb-repressed' states (States 4, 5 and 12).



Supplementary Figure 6: Biochemical distinctions between states validated by RNAPII and H2A.Z occupancy levels. Predicted regulatory elements were subject to biochemical analysis to validate chromatin states. We tested elements annotated as state 1 (promoter-like, n=13), state 4 (strong enhancer, n=13) or state 7 (weak/poised enhancer, n=11) in K562 cells. We selected the elements randomly except that we excluded elements of state 4 that were within 2kb of a state 1 call, and elements of state 7 calls which were within 2kb of a state 1 or 4 call. ChIP-PCR was used to measure (a) RNAPII and (b) H2A.Z occupancy at these locations in K562 cells. The resulting data are presented in box plot format with Y-axis indicating ChIP-PCR enrichment ratio relative to input, and a negative control region. Boxes indicate 25th, 50th and 75th percentile value. Whiskers indicate 5th and 95th percentiles. Elements annotated as promoter-like are strongly enriched for RNAPII binding. Strong enhancers show a lesser degree of RNAPII enrichment. H2A.Z has been associated with histone exchange and nucleosome-free regions. H2A.Z enrichment is seen at promoter-like elements and strong enhancers, both of which exhibit notable nucleosome-free regions. A lower degree of enrichment is seen at weak enhancers, consistent with the observation that elements annotated as state 7 tend to lack a nucleosome-free region. ChIP-PCR data were collected in biological replicates, each evaluated in technical replicate.

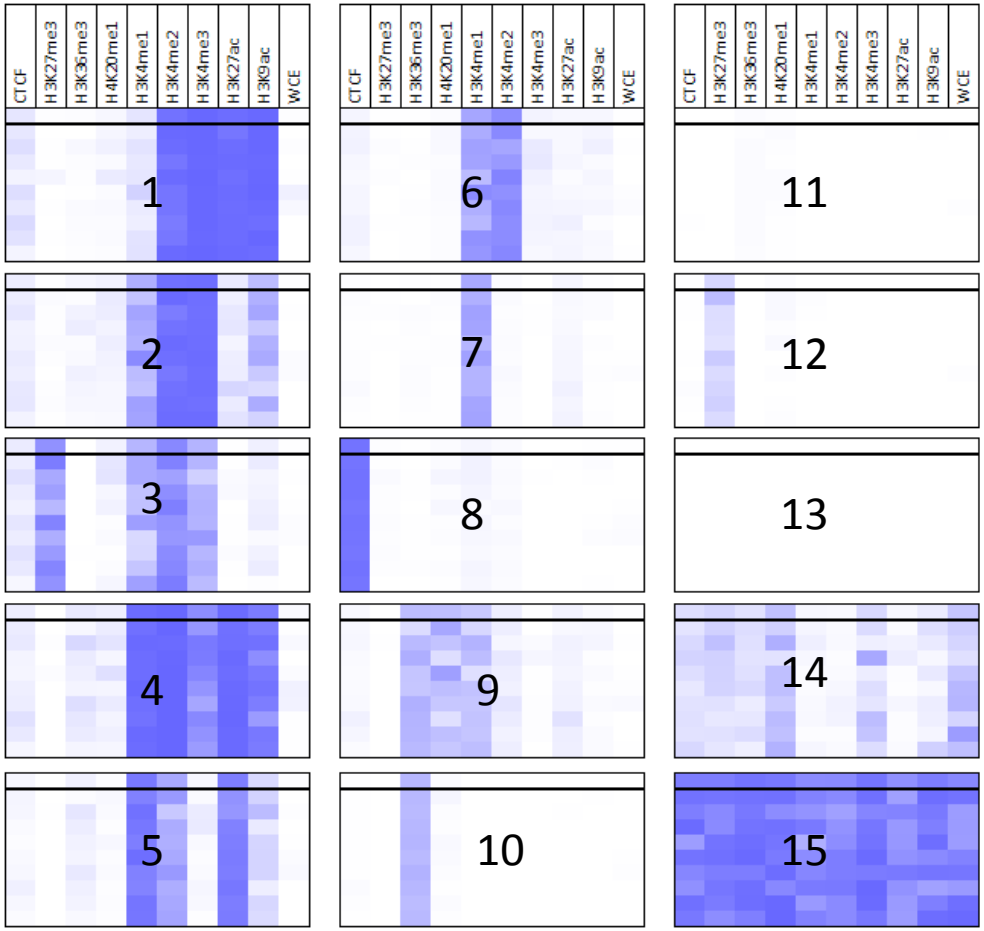


Best Correlation with Joint State from independently learned models

	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
1	0.99	1.00	1.00	0.99	0.97	1.00	1.00	0.99	0.98
2	0.98	0.94	1.00	0.98	0.95	0.94	0.94	0.96	0.97
3	1.00	0.65	0.85	0.73	0.98	0.71	0.74	0.98	0.72
4	0.83	0.98	0.99	0.98	0.98	0.97	0.97	0.99	0.91
5	0.84	0.92	0.99	0.97	1.00	0.99	0.99	0.94	0.95
6	0.95	0.93	0.96	0.93	0.84	0.99	0.99	0.97	0.96
7	0.99	1.00	1.00	0.87	0.98	0.99	0.99	0.99	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	0.82	0.97	0.85	0.82	0.84	0.81	0.68	0.85	0.93
10	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
11	0.94	0.90	0.85	0.98	0.92	0.90	0.83	0.90	0.94
12	0.99	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
13	0.94	0.92	0.92	0.93	0.97	0.93	0.92	0.94	0.95
14	0.74	0.81	0.68	0.59	0.85	0.77	0.74	0.92	0.76
15	0.82	0.47	0.68	0.82	0.68	0.46	0.77	0.56	0.67

Supplementary Figure 7: Independent Cell Type Learning. 15 state models were learned independently on the nine cell types. (**Left)** The heatmap on left displays the emission parameters for all 135 states from all 9 independently learned models. These 135 states are grouped and displayed based on the state from the jointly learned 15 state model for which they are closest to based on Euclidean distance of the emission parameters. The heatmap indicates that few states emerge from any of the independent model which are not well represented by a state from the joint model. Those states which are not well represented in the joint model were not reproducible in other cell types, and can be attributed to experimental artifacts such as cell type specific PCR bias. **(Right)** This table indicates for each cell type independently learned model (columns) the greatest correlation of any of its states for each state of the joint model (rows). The table shows that most states of the joint model were also recovered with high correlation from each of the independently learned models.

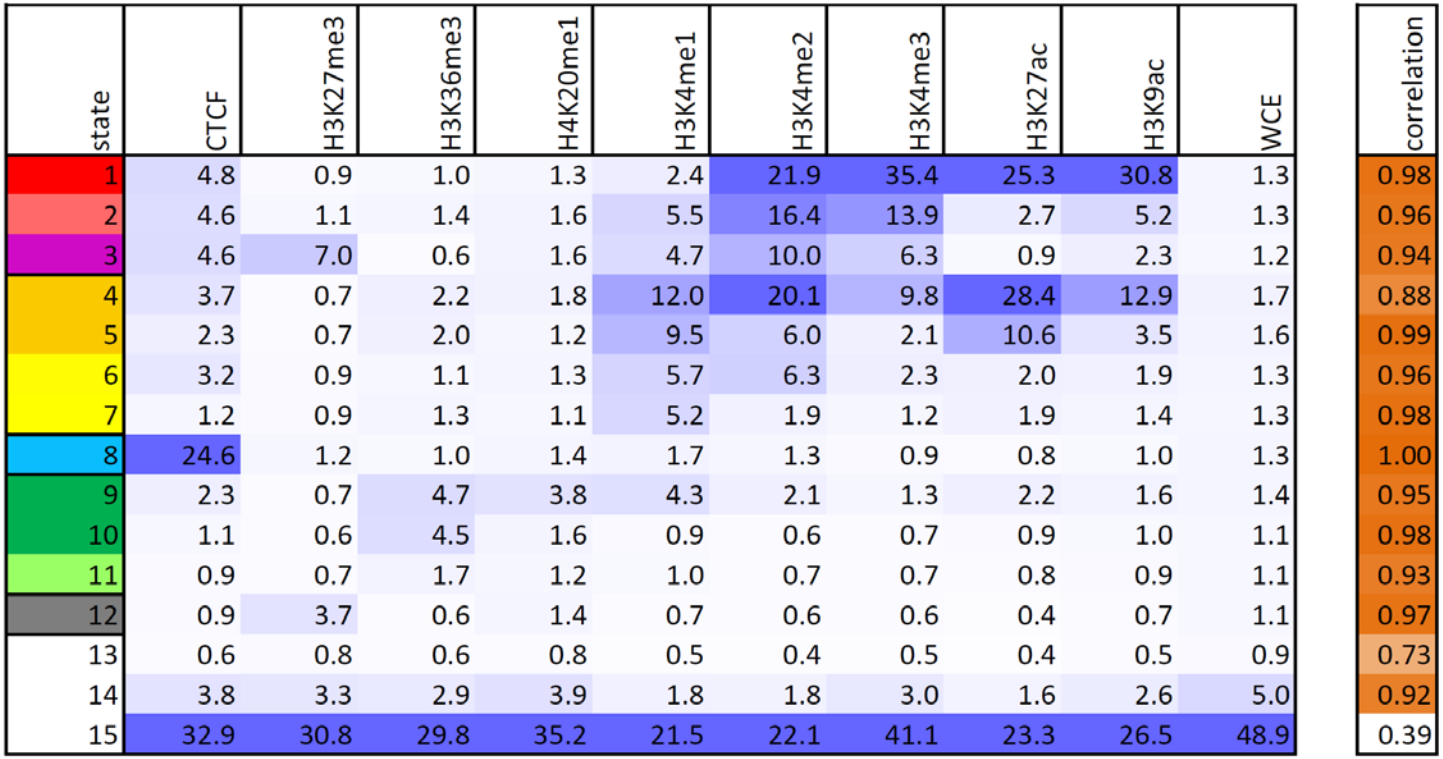
Observed Mark Frequency in States



Correlation with Joint Model Parameters

	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
1	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
2	0.99	0.99	0.99	1.00	0.98	0.99	0.97	0.99	0.99
3	0.99	0.95	1.00	0.96	0.98	0.93	0.97	0.99	0.98
4	1.00	0.99	0.99	1.00	1.00	0.99	0.99	1.00	1.00
5	0.98	0.95	0.99	1.00	1.00	0.99	0.99	1.00	1.00
6	1.00	0.98	0.99	0.98	0.99	1.00	0.99	0.99	1.00
7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	0.88	0.97	0.89	0.91	0.99	0.98	0.88	0.99	0.99
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11	0.82	0.99	0.95	0.96	0.95	0.89	0.90	0.96	0.96
12	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00
13	0.07	0.94	0.91	0.95	0.96	0.61	0.75	0.97	0.96
14	0.93	0.83	0.53	0.94	0.95	0.90	0.84	0.90	0.86
15	0.80	0.55	0.72	0.76	0.82	0.18	0.39	0.82	0.49

Supplementary Figure 8: Observed Frequency of Marks in a State are Consistent with Joint Model Emission Parameters (Left) Each box corresponds to the indicated state of the 15 state jointly learned model. The top row in the box indicates the joint model emission parameters for the state. The successive rows below the black horizontal line correspond to the observed frequency of each mark in the state across the 9 cell types in the order: H1 ES,K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK, HMEC. **(Right)** Each row corresponds to a state of the joint model and indicates for each cell type the correlation of the emission parameters with the observed frequency of the marks in the state for each cell type. For states 1-12 all correlations are >= 0.8 and almost all are >=0.95, a few lower correlation occur in state 13 which has low frequency for all marks, and repetitive states 14-15. Overall this indicates that the joint model emission parameters are reflective of the frequency mark observations in the individual cell types validating the modeling approach of considering a virtual concatenation of genomes.



state	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	correlation
1	4.8	0.9	1.0	1.3	2.4	21.9	35.4	25.3	30.8	1.3	0.98
2	4.6	1.1	1.4	1.6	5.5	16.4	13.9	2.7	5.2	1.3	0.96
3	4.6	7.0	0.6	1.6	4.7	10.0	6.3	0.9	2.3	1.2	0.94
4	3.7	0.7	2.2	1.8	12.0	20.1	9.8	28.4	12.9	1.7	0.88
5	2.3	0.7	2.0	1.2	9.5	6.0	2.1	10.6	3.5	1.6	0.99
6	3.2	0.9	1.1	1.3	5.7	6.3	2.3	2.0	1.9	1.3	0.96
7	1.2	0.9	1.3	1.1	5.2	1.9	1.2	1.9	1.4	1.3	0.98
8	24.6	1.2	1.0	1.4	1.7	1.3	0.9	0.8	1.0	1.3	1.00
9	2.3	0.7	4.7	3.8	4.3	2.1	1.3	2.2	1.6	1.4	0.95
10	1.1	0.6	4.5	1.6	0.9	0.6	0.7	0.9	1.0	1.1	0.98
11	0.9	0.7	1.7	1.2	1.0	0.7	0.7	0.8	0.9	1.1	0.93
12	0.9	3.7	0.6	1.4	0.7	0.6	0.6	0.4	0.7	1.1	0.97
13	0.6	0.8	0.6	0.8	0.5	0.4	0.5	0.4	0.5	0.9	0.73
14	3.8	3.3	2.9	3.9	1.8	1.8	3.0	1.6	2.6	5.0	0.92
15	32.9	30.8	29.8	35.2	21.5	22.1	41.1	23.3	26.5	48.9	0.39

Supplementary Figure 9: Mark Signal Enrichment. This figure shows for each state the median fold enrichment for each mark across the nine-cell types. The last column shows the correlation between the emission parameters of the model and the signal enrichments. While the general pattern of modifications is the same as with the emission parameters in Figure 1, states which have similar emission parameter values can show differences in enrichment signal; e.g., H3K4me3 in state 1 and 2. Thus, by considering marks jointly and in their spatial context the model captures signal information beyond the initial binarizations.

% overlap +/- 2kb TSS

state	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC	median
1	83	75	76	79	87	83	83	83	85	83
2	67	42	41	41	62	65	58	55	64	58
3	44	20	55	30	48	49	47	46	56	47
4	23	26	11	15	24	12	24	19	29	23
5	4	7	3	3	3	2	4	3	3	3
6	18	18	16	15	23	15	17	17	18	17
7	4	6	4	5	5	3	3	4	4	4
8	3	4	4	3	4	3	4	3	3	3
9	4	4	3	2	3	4	4	4	3	4
10	1	1	1	1	1	1	1	1	1	1
11	2	3	2	2	2	2	2	2	2	2
12	14	4	10	6	6	7	6	7	9	7
13	1	1	1	1	1	1	1	1	1	1
14	3	3	3	3	3	6	3	3	2	3
15	2	6	1	3	1	1	1	1	1	1

Fold enrichment for CpG island

State	H1	K562	GM12878	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Median
1	73.4	51.8	46.5	47.3	71.3	60.1	61.3	58.4	54.1	58.4
2	33.3	19.5	14.7	20.2	28.5	34.1	22.0	26.9	40.0	26.9
3	44.6	19.3	60.8	27.1	58.7	65.7	56.2	58.6	82.7	58.6
4	6.6	3.5	1.5	2.1	7.0	2.1	5.8	2.6	8.7	3.5
5	0.6	0.5	0.2	0.2	0.5	0.2	0.5	0.2	0.2	0.2
6	3.6	7.6	4.7	2.3	4.8	1.7	3.9	3.8	5.5	3.9
7	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.2	0.4	0.3
8	1.0	2.8	2.1	1.2	2.2	1.4	1.5	1.2	0.9	1.4
9	1.6	0.5	0.5	0.5	0.4	0.6	0.8	0.9	1.7	0.6
10	0.9	0.4	0.5	0.6	0.8	0.4	1.4	0.8	0.5	0.6
11	0.4	0.9	0.6	0.8	0.4	0.6	0.3	0.4	0.5	0.5
12	1.2	1.7	4.6	2.0	2.1	2.5	2.2	2.8	3.6	2.2
13	0.1	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.2
14	3.0	1.7	2.7	2.8	4.1	9.0	3.7	2.9	2.3	2.9
15	4.6	2.2	0.8	0.7	1.1	0.7	1.0	1.1	1.0	1.0

Fold Enrichment for Gene Body

state	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC	Median
1	0.1	0.2	0.4	0.3	0.1	0.2	0.2	0.2	0.2	0.2
2	0.4	0.7	1.2	0.6	0.6	0.5	0.7	0.6	0.5	0.6
3	0.6	0.7	0.4	0.5	0.4	0.4	0.4	0.5	0.4	0.4
4	1.1	1.1	1.6	1.4	1.3	1.5	1.4	1.3	1.2	1.3
5	1.4	1.3	1.7	1.6	1.5	1.4	1.6	1.4	1.3	1.4
6	1.1	0.8	1.3	0.8	1.2	1.1	1.3	1.0	1.1	1.1
7	1.4	1.2	1.4	1.2	1.5	1.2	1.5	1.2	1.2	1.2
8	1.1	1.1	1.0	1.0	1.1	1.0	1.1	1.1	1.0	1.1
9	2.5	2.4	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
10	2.6	2.6	2.6	2.6	2.6	2.5	2.6	2.6	2.6	2.6
11	1.9	1.7	1.9	1.9	2.1	1.8	2.0	1.9	1.8	1.9
12	0.8	1.1	0.9	0.8	0.7	0.8	0.7	0.9	0.9	0.8
13	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7	0.8	0.7
14	0.5	0.8	0.5	0.4	0.5	0.4	0.3	0.3	0.3	0.4
15	0.1	0.8	0.4	0.4	0.4	0.3	0.4	0.1	0.2	0.4

Supplementary Figure 10: Annotation enrichment consistency across cell types. **(Left)** The table shows for each state (rows) and cell type (columns) the % of the state that overlaps a promoter within 2kb of a RefSeq TSS. The far right column is the median across the nine cell types. **(Middle)** The fold enrichment for each state in each cell type for overlapping an annotated CpG island. The far right column is the median enrichment across cell types. **(Right)** The fold enrichment for each state in each cell type for overlapping a RefSeq annotated gene that is not within +/-2kb of a TSS. The far right column is the median enrichment across cell types. With few exceptions these values are highly consistent across cell types.

a

State in second cell type

State in first cell type

state	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	110.9	28.0	10.9	10.5	0.4	3.0	0.3	0.3	0.3	0.0	0.1	0.4	0.1	0.4	1.2
2	28.0	53.0	23.5	11.6	1.4	11.7	1.2	2.6	3.3	0.2	0.5	1.1	0.2	1.3	2.0
3	10.9	23.5	111.6	4.3	0.8	5.4	1.0	3.4	0.4	0.1	0.3	5.6	0.2	4.0	2.5
4	10.5	11.6	4.3	26.7	9.2	6.4	3.6	1.6	5.2	0.8	1.0	1.0	0.3	0.8	1.6
5	0.4	1.4	0.8	9.2	12.5	4.0	5.2	1.7	4.2	1.2	1.5	0.9	0.5	0.5	0.7
6	3.0	11.7	5.4	6.4	4.0	11.9	3.4	3.9	2.6	0.3	1.1	1.5	0.5	1.1	2.9
7	0.3	1.2	1.0	3.6	5.2	3.4	4.8	1.2	2.4	0.6	1.8	1.4	0.6	0.5	0.3
8	0.3	2.6	3.4	1.6	1.7	3.9	1.2	82.1	1.8	0.5	0.5	0.9	0.4	2.9	2.3
9	0.3	3.3	0.4	5.2	4.2	2.6	2.4	1.8	27.4	4.6	2.1	0.4	0.1	2.1	1.6
10	0.0	0.2	0.1	0.8	1.2	0.3	0.6	0.5	4.6	12.7	2.2	0.2	0.1	0.4	0.2
11	0.1	0.5	0.3	1.0	1.5	1.1	1.8	0.5	2.1	2.2	3.3	0.7	0.5	1.0	0.2
12	0.4	1.1	5.6	1.0	0.9	1.5	1.4	0.9	0.4	0.2	0.7	6.0	0.8	1.5	0.2
13	0.1	0.2	0.2	0.3	0.5	0.5	0.6	0.4	0.1	0.1	0.5	0.8	1.2	0.2	0.0
14	0.4	1.3	4.0	0.8	0.5	1.1	0.5	2.9	2.1	0.4	1.0	1.5	0.2	293.0	77.0
15	1.2	2.0	2.5	1.6	0.7	2.9	0.3	2.3	1.6	0.2	0.2	0.2	0.0	77.0	740.7

Pairwise fold enrichments

b

State in second cell type

state	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	63.8	15.3	2.3	6.6	0.4	2.7	0.5	0.2	0.2	0.1	1.6	1.7	4.2	0.1	0.1
2	16.1	28.9	5.0	7.3	1.4	10.5	2.3	1.5	2.8	0.9	6.3	4.6	12.0	0.2	0.2
3	6.3	12.8	23.7	2.7	0.8	4.9	1.9	2.0	0.3	0.3	4.1	24.6	14.7	0.7	0.2
4	6.1	6.3	0.9	16.8	8.9	5.8	7.0	0.9	4.3	3.3	13.2	4.6	21.5	0.1	0.2
5	0.2	0.8	0.2	5.8	12.2	3.6	9.9	1.0	3.5	5.2	18.9	3.9	34.7	0.1	0.1
6	1.7	6.4	1.2	4.0	3.9	10.7	6.6	2.3	2.2	1.4	14.4	6.6	38.1	0.2	0.3
7	0.2	0.7	0.2	2.3	5.0	3.1	9.2	0.7	2.0	2.7	22.2	6.3	45.3	0.1	0.0
8	0.2	1.4	0.7	1.0	1.7	3.5	2.4	48.8	1.5	2.1	6.8	3.8	25.5	0.5	0.2
9	0.2	1.8	0.1	3.3	4.1	2.4	4.6	1.1	22.7	20.3	27.1	1.8	10.2	0.3	0.2
10	0.0	0.1	0.0	0.5	1.2	0.3	1.2	0.3	3.8	56.0	27.7	0.7	8.2	0.1	0.0
11	0.1	0.3	0.1	0.7	1.5	1.0	3.4	0.3	1.8	9.7	42.0	3.2	35.9	0.2	0.0
12	0.2	0.6	1.2	0.7	0.9	1.4	2.8	0.5	0.3	0.7	9.3	26.2	55.0	0.2	0.0
13	0.0	0.1	0.0	0.2	0.5	0.5	1.2	0.2	0.1	0.5	6.4	3.4	86.8	0.0	0.0
14	0.2	0.7	0.9	0.5	0.5	1.0	1.0	1.7	1.7	1.9	12.5	6.7	15.7	47.6	7.3
15	0.7	1.1	0.5	1.0	0.7	2.6	0.5	1.4	1.3	0.9	2.0	0.7	3.5	12.5	70.5
Expected	0.6	0.5	0.2	0.6	1.0	0.9	1.9	0.6	0.8	4.4	12.6	4.4	71.2	0.2	0.1

Pairwise transition frequencies

Supplementary Figure 11: Pairwise state assignment transition enrichments and frequencies. a. The likelihood that a genomic position assigned to a given chromatin state in one cell type will assume the indicated chromatin state in a second cell type. Values are relative enrichments calculated in aggregate over all 72 pairs of cell types. Blue shading indicates pairs of states that are more frequently assigned to the same position in two different cell types. For example, positions assigned to an active promoter state (State 1) tend to remain in a promoter state (States 1-3) in other cell types, and rarely change to transcribed, repressed, or inactive states. **b.** This figure shows the frequency of pairwise states (%) of transitioning from the state in the row of the matrix to the state in the column of the matrix between two cell types computed in aggregate based on all 72 pairs of states. The bottom 'Expected' row indicates the expected frequency if state assignments were independent between cell types. Dividing a value in the cell of the matrix by a value in the expected row gives the pariwise enrichments values found in Figure 2a.

Pairwise enrichment between individual replicates of different cell types

state	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	116.5	35.2	11.4	14.7	1.1	5.3	0.5	0.4	0.5	0.0	0.2	0.6	0.1	0.8	0.9
2	35.2	70.2	30.4	14.7	1.8	17.9	1.5	2.4	4.4	0.2	0.4	1.1	0.1	2.6	2.7
3	11.4	30.4	231.6	3.9	0.8	9.9	1.9	4.6	0.3	0.1	0.2	7.6	0.2	6.4	2.2
4	14.7	14.7	3.9	40.9	13.7	8.4	4.4	1.7	5.8	0.7	1.0	1.1	0.3	0.9	1.2
5	1.1	1.8	0.8	13.7	16.6	5.3	6.7	1.8	5.4	1.1	1.4	0.9	0.5	0.6	0.7
6	5.3	17.9	9.9	8.4	5.3	17.1	4.5	3.5	4.3	0.4	1.2	1.4	0.5	1.8	2.6
7	0.5	1.5	1.9	4.4	6.7	4.5	6.1	1.4	3.9	0.8	1.8	1.3	0.6	0.7	0.6
8	0.4	2.4	4.6	1.7	1.8	3.5	1.4	79.4	2.2	0.7	0.7	1.0	0.4	3.3	3.4
9	0.5	4.4	0.3	5.8	5.4	4.3	3.9	2.2	41.3	6.4	2.5	0.5	0.2	2.9	1.8
10	0.0	0.2	0.1	0.7	1.1	0.4	0.8	0.7	6.4	14.8	3.3	0.2	0.1	0.6	0.2
11	0.2	0.4	0.2	1.0	1.4	1.2	1.8	0.7	2.5	3.3	3.5	0.6	0.5	1.2	0.2
12	0.6	1.1	7.6	1.1	0.9	1.4	1.3	1.0	0.5	0.2	0.6	7.3	0.8	1.8	0.2
13	0.1	0.1	0.2	0.3	0.5	0.5	0.6	0.4	0.2	0.1	0.5	0.8	1.2	0.2	0.0
14	0.8	2.6	6.4	0.9	0.6	1.8	0.7	3.3	2.9	0.6	1.2	1.8	0.2	307.8	123.6
15	0.9	2.7	2.2	1.2	0.7	2.6	0.6	3.4	1.8	0.2	0.2	0.2	0.0	123.6	1095.8

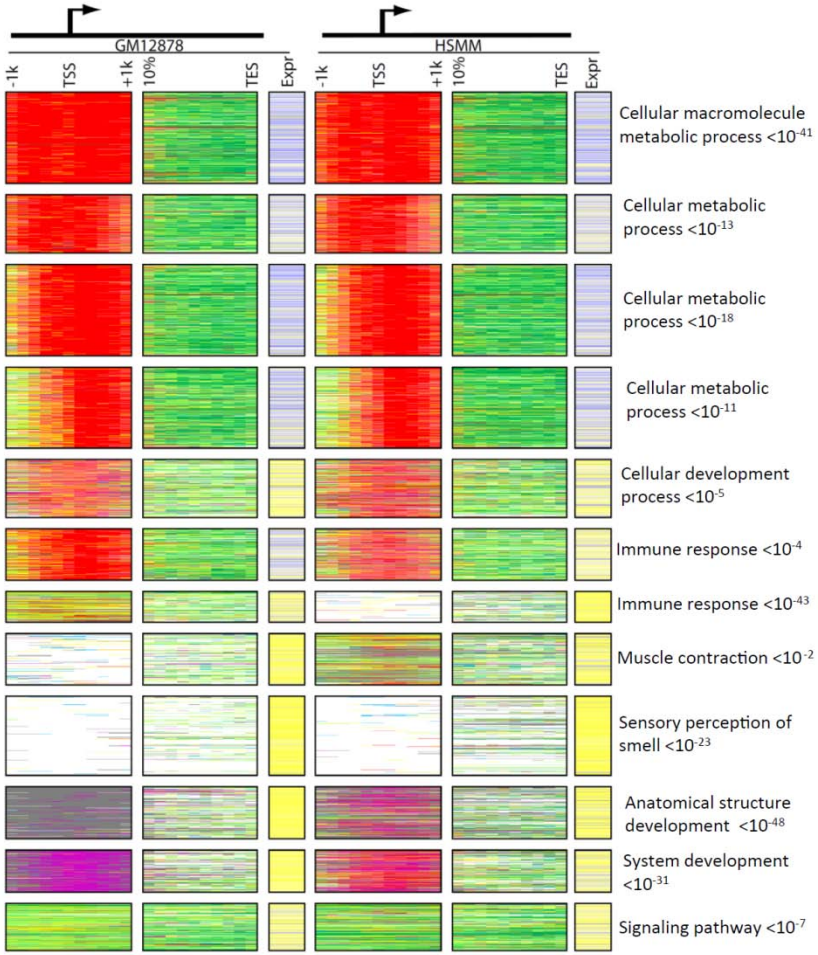
Pairwise enrichment between individual replicates of the same cell type

state	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	144.2	27.6	0.7	19.2	1.1	2.6	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.5	0.5
2	27.6	101.8	21.1	15.6	3.9	24.0	1.5	1.6	3.7	0.0	0.1	0.2	0.0	2.5	3.2
3	0.7	21.1	413.2	0.8	0.6	13.0	5.0	4.3	0.4	0.0	0.1	7.4	0.1	8.3	2.9
4	19.2	15.6	0.8	149.6	26.5	4.6	0.6	0.1	1.5	0.0	0.0	0.0	0.0	0.2	2.5
5	1.1	3.9	0.6	26.5	56.9	9.8	10.1	0.6	6.2	0.1	0.2	0.0	0.0	0.5	0.4
6	2.6	24.0	13.0	4.6	9.8	38.4	11.7	3.1	5.0	0.1	0.7	0.4	0.1	1.1	3.1
7	0.1	1.5	5.0	0.6	10.1	11.7	19.9	1.3	6.9	0.4	1.8	0.4	0.2	0.7	0.5
8	0.1	1.6	4.3	0.1	0.6	3.1	1.3	91.8	1.7	0.7	0.7	1.0	0.3	3.4	2.2
9	0.0	3.7	0.4	1.5	6.2	5.0	6.9	1.7	94.2	5.6	1.7	0.0	0.0	4.8	1.4
10	0.0	0.0	0.0	0.0	0.1	0.1	0.4	0.7	5.6	16.1	3.8	0.0	0.0	0.5	0.0
11	0.0	0.1	0.1	0.0	0.2	0.7	1.8	0.7	1.7	3.8	4.5	0.3	0.3	1.2	0.1
12	0.0	0.2	7.4	0.0	0.0	0.4	0.4	1.0	0.0	0.0	0.3	14.5	0.7	1.7	0.1
13	0.0	0.0	0.1	0.0	0.0	0.1	0.2	0.3	0.0	0.0	0.3	0.7	1.2	0.2	0.0
14	0.5	2.5	8.3	0.2	0.5	1.1	0.7	3.4	4.8	0.5	1.2	1.7	0.2	348.6	135.6
15	0.5	3.2	2.9	2.5	0.4	3.1	0.5	2.2	1.4	0.0	0.1	0.1	0.0	135.6	1103.9

Ratio of enrichments from different cell types to the same cell type

state	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.8	1.3	16.0	0.8	1.1	2.1	8.9	5.2	10.5	142.5	28.8	142.7	82.6	1.5	1.8
2	1.3	0.7	1.4	0.9	0.5	0.7	1.0	1.5	1.2	7.0	3.2	5.5	5.4	1.0	0.8
3	16.0	1.4	0.6	5.2	1.3	0.8	0.4	1.1	0.9	135.7	2.3	1.0	3.2	0.8	0.8
4	0.8	0.9	5.2	0.3	0.5	1.9	6.9	20.8	3.8	141.1	75.9	433.2	262.3	4.3	0.5
5	1.1	0.5	1.3	0.5	0.3	0.5	0.7	3.2	0.9	8.3	5.7	47.2	22.7	1.4	1.8
6	2.1	0.7	0.8	1.9	0.5	0.4	0.4	1.1	0.9	3.4	1.6	3.8	3.3	1.5	0.9
7	8.9	1.0	0.4	6.9	0.7	0.4	0.3	1.0	0.6	2.0	1.0	3.0	2.9	1.0	1.1
8	5.2	1.5	1.1	20.8	3.2	1.1	1.0	0.9	1.3	1.0	1.0	1.0	1.2	1.0	1.6
9	10.5	1.2	0.9	3.8	0.9	0.9	0.6	1.3	0.4	1.1	1.5	9.4	20.1	0.6	1.3
10	142.5	7.0	135.7	141.1	8.3	3.4	2.0	1.0	1.1	0.9	0.9	16.5	3.1	1.4	3.9
11	28.8	3.2	2.3	75.9	5.7	1.6	1.0	1.0	1.5	0.9	0.8	2.1	1.5	1.0	1.8
12	142.7	5.5	1.0	433.2	47.2	3.8	3.0	1.0	9.4	16.5	2.1	0.5	1.3	1.0	1.6
13	82.6	5.4	3.2	262.3	22.7	3.3	2.9	1.2	20.1	3.1	1.5	1.3	1.0	1.5	1.1
14	1.5	1.0	0.8	4.3	1.4	1.5	1.0	1.0	0.6	1.4	1.0	1.0	1.5	0.9	0.9
15	1.8	0.8	0.8	0.5	1.8	0.9	1.1	1.6	1.3	3.9	1.8	1.6	1.1	0.9	1.0

Supplementary Figure 12: Pairwise State Enrichments in Biological Replicates. This figure considers the pairwise chromatin state co-occurrence enrichment between cell types as compared to an estimate of what would be expected from a biological replicate within the same cell type. To estimate this we took the model learned jointly based on pooling replicates, and applied the model to individual replicates based on making presence/absence calls for the marks using just the data from the individual replicate to give replicate specific state assignments. **(Top)** The analogous enrichment grid to Figure 2a of the main text, but now based on pairs of state assignments from individual replicates from different cell types. **(Center)** The same figure as top figure, but based on pairs replicates from the same cell type. **(Bottom)** The ratio of the enrichment for a pair of states for different cell types as compared to within the same cell type. For instance from this one can observe the enrichment for observing the strong enhancer state 4 assignment and the repressed state 12 assignment is over 400 times more likely to occur when consider individual replicates from two different cell types than two different replicates from the same cell type.



Supplementary Figure 13. Pairwise chromatin state dynamics between cell types. Chromatin states were assigned to gene TSSs (200 bp windows within 1 kb) and primary transcripts (proportional intervals from 10% to the transcript end site, TES) based on the GM12878 and HSMM datasets (see **Methods**). The genes were then clustered on the TSS chromatin states. For each gene, mRNA expression values are indicated (blue for high expression; yellow for low expression). For each cluster, a top enriched GO functional category is indicated, along with a p-value.

GM12878	HSMM	CategoryName	#Genes	Corrected p-value	Fold
13	13	sensory perception of smell	347	3.7E-257	6.7
1	1	cellular macromolecule metabolic process	2687	6.8E-165	1.4
12	3	nervous system development	92	1.9E-26	3.7
1	13	immune response	37	4.3E-19	7.2
4	13	immune response	30	7.0E-16	7.7
12	1	anatomical structure morphogenesis	47	9.9E-16	4.4
3	1	anatomical structure development	84	1.2E-11	2.3
13	2	homophilic cell adhesion	18	1.7E-11	12.5
3	3	nervous system development	66	2.2E-11	2.8
12	12	system development	132	3.5E-11	1.9
3	2	anatomical structure morphogenesis	74	2.5E-10	2.5
12	2	anatomical structure development	88	9.2E-08	2.0
1	2	cellular metabolic process	686	1.7E-07	1.2
13	12	G-protein coupled receptor protein signaling pathway	55	9.5E-07	2.5
5	13	immune response	12	4.5E-06	9.2
2	1	anatomical structure morphogenesis	71	8.4E-04	1.9
7	12	defense response	11	5.0E-03	5.9
13	8	homophilic cell adhesion	5	6.0E-03	24.5
7	13	immune response	16	1.1E-02	3.9
4	2	lipid localization	6	2.7E-02	12.8
13	1	extracellular structure organization	7	7.3E-02	8.5
12	13	ectoderm development	11	7.5E-02	4.8
13	6	system development	35	7.7E-02	2.0
13	11	regulation of body fluid levels	13	7.8E-02	4.1
2	2	small molecule metabolic process	117	9.6E-02	1.4
13	7	muscle contraction	8	9.9E-02	6.8
1	3	lymphocyte differentiation	6	1.7E-01	9.4
11	7	secretion	7	1.8E-01	7.1

Supplementary Figure 14: Pairwise Promoter State Transitions Reveal Enrichment of Gene Ontology Categories. This table shows all pairwise promoter state transitions between GM12878 and HSMM cells that result in a gene set that corresponds significantly to a GO category. For each pair of states the biological GO category with the most significant correspondence by p-value is shown, along with the number of intersecting genes and fold enrichment.

a

Biological processes enriched for genes with the most enhancers in vincinity

GO Slim Category	Number of genes	Corrected p-value	Fold
anatomical structure morphogenesis	68	6E-11	2.3
multicellular organismal development	116	5E-10	1.8
cellular component organization	102	1E-07	1.7
regulation of biological process	212	2E-07	1.3
embryonic development	33	3E-07	2.7

b

Biological processes enriched for active promoters without enhancers

GO Slim Category	Number of genes	Corrected p-value	Fold
transcription	292	7E-11	1.4
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	396	7E-07	1.3
biosynthetic process	387	3E-04	1.2

c

Biological processes enriched for genes with TSS in States 3 or 12 in three or more cell types

GO Slim Category	Number of genes	Corrected p-value	Fold
multicellular organismal development	1008	7E-117	1.9
cell differentiation	630	6E-75	2.0
anatomical structure morphogenesis	488	3E-64	2.1
cell-cell signaling	317	1E-60	2.5
cell communication	555	4E-50	1.8
embryonic development	227	3E-35	2.3
ion transport	289	2E-31	2.0
behavior	195	8E-28	2.3

Supplementary Figure 15: Different gene classes appear to differentially utilize promoter and enhancer regulation. **(a)** Table lists GO categories enriched among genes with the highest number of enhancers in their vicinity. Specifically, we considered those genes with the highest proportion of sequence within 20 kb (not including +/- 5 kb) of the TSS assigned to state 4 or 5. Table indicates number of overlapping genes, Bonferroni corrected p-value, and fold enrichment. **(b)** Table lists GO categories enriched among genes with an active promoter, but no state 4 or 5 assignments within a 20 kb window (not including +/- 5 kb) of the TSS in any of the cell types. **(c)** Table lists GO categories enriched among genes whose TSS is assigned to state 3 (poised promoter) or 12 (Polycomb-repressed) in at least 3 cell types.

Linked Predictions

Cluster	Category	p-value	Fold
a	gene silencing	3.E-02	12.5
b	DNA conformation change	2.E-11	3.3
c	chromatin assembly	2.E-18	6.5
d	DNA conformation change	5.E-04	3.8
e			
f	immune response	3.E-24	2.0
g			
h	lipid metabolic process	1.E-14	2.0
i	lipid metabolic process	3.E-10	2.1
j			
k	angiogenesis	3.E-09	2.7
l	anatomical structure morphogenesis	3.E-12	2.3
m	cellular component movement	6.E-07	2.2
n	hemidesmosome assembly	7.E-03	7.7
o	anatomical structure development	6.E-12	1.5
p	anatomical structure development	6.E-11	1.7
q	ectoderm development	3.E-11	3.7
r	ectoderm development	9.E-15	3.4
s	ectoderm development	7.E-07	3.8
t	cellular macromolecular complex sub.	2.E-05	2.7

Proximal Gene Based

Cluster	Category	p-value	Fold
a			
b	chromatin organization	2.E-04	1.3
c	nucleosome assembly	7.E-07	3.0
d	nucleosome organization	4.E-02	2.0
e			
f	lymphocyte activation	1.E-08	1.6
g	programmed cell death	3.E-04	1.5
h	lipid metabolic process	3.E-06	1.3
i	lipid metabolic process	1.E-05	1.4
j	apoptosis	6.E-04	1.5
k	angiogenesis	2.E-04	1.5
l	blood vessel morphogenesis	9.E-08	2.4
m	blood vessel morphogenesis	8.E-05	1.9
n	vasculature development	6.E-08	2.2
o	anatomical structure development	7.E-12	1.3
p	anatomical structure morphogenesis	3.E-09	1.5
q	intracellular signaling pathway	9.E-08	1.5
r	ectoderm development	7.E-04	1.6
s	regulation of cellular process	7.E-03	1.1
t			

Supplementary Figure 16: GO Categories assignments for enhancer clusters. Tables list GO categories enriched for genes associated with enhancer clusters from Figure 2c, where the genes assignments were based either on predicted enhancer-gene linkages by correlation (**left**) or on distance (**right**; analogous to **Fig 2c**). Enhancer-gene linkages predicted by correlation were assigned at a threshold of 2.5, where the enhancer element must be between 5 kb and 125 kb of the TSS (see Methods). For the distance based linkages, enhancers were assigned to the nearest gene, excluding genes within 5 kb of the TSS. The table lists a top enriched GO category based on p-value for each cluster and reports the corrected p-value and fold.

Schadt Liver - HepG2

Window Expansion	Fold (distance controlled)	p-val	#eQTL Pairs Linked	#Link Eligible	Expected (random)	Expected by Distance	**Observed Fraction Linked**
0	1.9	0.03	11	49	0.06	0.12	**0.22**
600	1.7	0.02	16	73	0.06	0.13	**0.22**
1000	1.7	0.01	21	98	0.06	0.13	**0.21**

Pickrell LCL - GM12878

Window Expansion	Fold (distance controlled)	p-val	#eQTL Pairs Linked	#Link Eligible	Expected (random)	Expected by Distance	**Observed Fraction Linked**
0	2.3	1.4E-02	9	32	0.07	0.12	**0.28**
600	2.3	4.0E-03	12	43	0.07	0.12	**0.28**
1000	1.9	2.0E-02	12	51	0.07	0.12	**0.24**

Montgomery LCL - GM12878

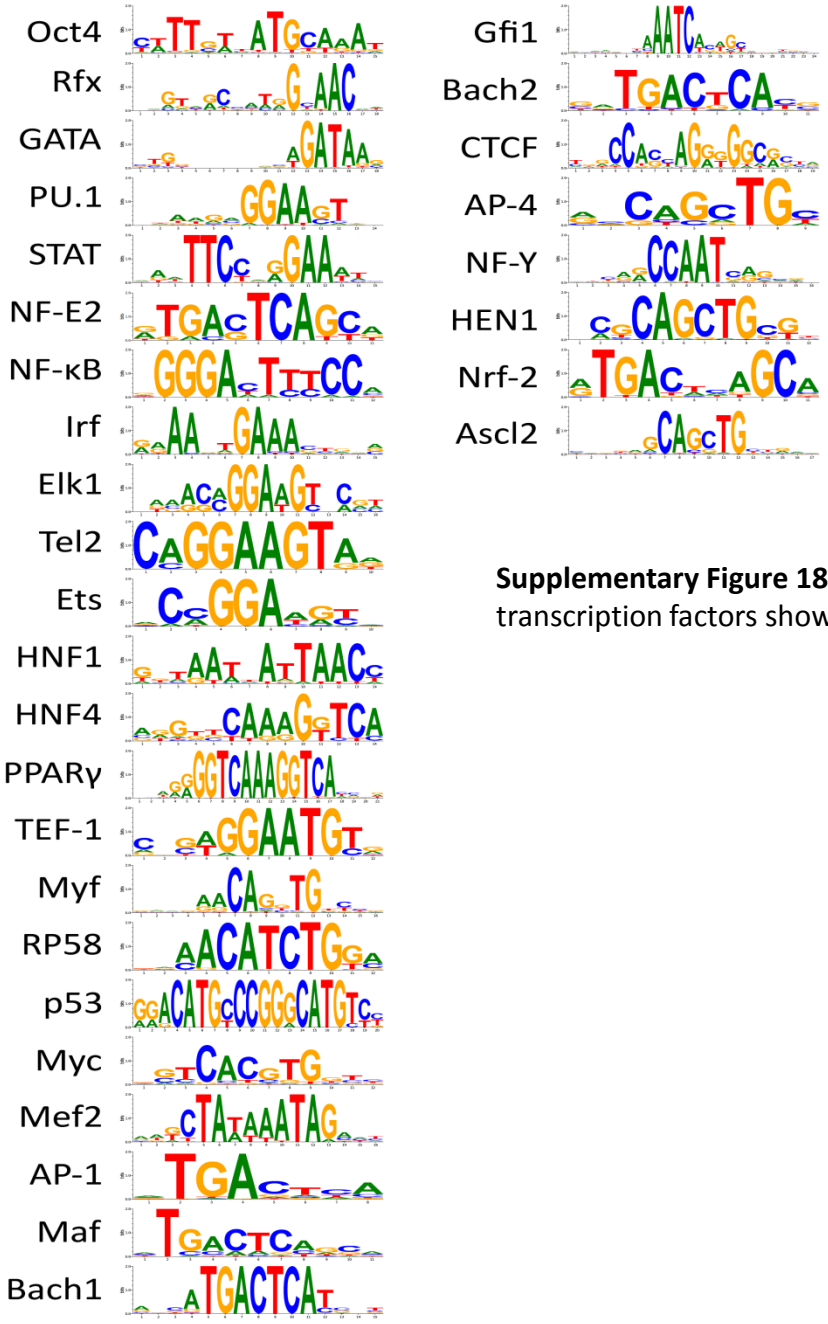
Window Expansion	Fold (distance controlled)	p-val	#eQTL Pairs Linked	#Link Eligible	Expected (random)	Expected by Distance	**Observed Fraction Linked**
0	4.6	<1.0E-10	51	120	0.07	0.09	**0.43**
600	4.7	<1.0E-10	78	197	0.07	0.08	**0.40**
1000	4.4	<1.0E-10	84	233	0.07	0.08	**0.36**

Veyrieras LCL - GM12878

Window Expansion	Fold (distance controlled)	p-val	#eQTL Pairs Linked	#Link Eligible	Expected (random)	Expected by Distance	**Observed Fraction Linked**
0	1.4	3.2E-03	61	387	0.07	0.11	**0.16**
600	1.4	1.2E-03	89	578	0.07	0.11	**0.15**
1000	1.3	2.5E-03	104	722	0.07	0.11	**0.14**

Supplementary Figure 17: Correspondence between enhancer-gene linkages from chromatin state mapping and cis-regulatory locus-gene linkages from QTL studies. We validated our predicted enhancer-gene linkages using QTLs data, based on liver[20] or lymphoblastoid cells[21-23]. We specifically evaluated the agreement between these studies and linkages drawn for strong enhancer states in HepG2 and GM12878 cells, respectively. We considered the fraction of all possible genetic linkages based on QTL mapping that corresponded to a chromatin-based linkage (link score at least 2.5; see Methods), treating each QTL study separately. The respective tables present (i) the fold enrichment of corresponding linkages relative to that expected for distance matched controls; (ii) p-value for significance of the correspondence (binomial distribution, based on distance matched controls); (iii) the total number of QTL SNP-gene linkages supported by the chromatin datasets; (iv) the total number of QTL SNP-gene linkages tested (ie, SNPs 5 to 125kb from a TSS with expression data); (v) the fraction of SNP-gene linkages that would be expected to be supported at random based on the overall frequency of enhancers and linked genes in the region from 5 to 125 kb of TSSs; (vi) the fraction of SNP-gene linkages that would be expected to be supported at random based on the actual distance distribution; and (vii) the actual fraction of linkages that were supported (bold). The first row of the table considers only SNP-gene pairs overlapping a state 4 or 5 assignment exactly. The second and third rows extend the set of SNPs to those pairs within 600bp or 1000 bp of a state 4 or 5 assignment. If multiple state 4 or 5 assignments were found in an expanded window of a SNP then we used the closest one.

Motif Logos



Supplementary Figure 18: Motif Logos. This figure shows the logos of the motifs used for the transcription factors shown in Figure 3.

state	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	Control	Genome %	% +/-TSS 2kb	DNase (CD4T; Fold)	Transcript (Fold)
1	16	2	2	6	17	93	99	96	98	2	0.30	81.0	37.6	0.2
2	12	2	6	9	53	94	95	14	44	1	0.47	63.2	30.8	0.6
3	13	72	0	9	48	78	49	1	10	1	0.06	53.9	17.2	0.3
4	11	1	15	11	96	99	75	97	86	4	0.19	37.3	33.1	1.2
5	5	0	10	3	88	57	5	84	25	1	0.42	8.7	21.4	1.5
6	7	1	1	3	58	75	8	6	5	1	0.38	24.2	14.2	1.4
7	2	1	2	1	56	3	0	6	2	1	1.12	9.5	7.1	1.3
8	92	2	1	3	6	3	0	0	1	1	0.21	4.2	18.8	1.0
9	5	0	43	43	37	11	2	9	4	1	3.03	7.2	2.9	2.4
10	1	0	47	3	0	0	0	0	0	1	3.20	0.6	0.5	2.6
11	0	0	3	2	0	0	0	0	0	0	13.39	2.8	0.7	2.0
12	1	27	0	2	0	0	0	0	0	0	1.74	10.0	0.8	0.8
13	0	0	0	0	0	0	0	0	0	0	75.33	1.2	0.2	0.7
14	22	28	19	41	6	5	26	5	13	37	0.16	47.0	15.9	0.5
15	85	85	91	88	76	77	91	73	85	78	0.01	9.7	24.9	0.3

Supplementary Figure 19: Application of the model to data in CD4T Cells: We applied the chromatin state model learned from the nine cell types to the common subset of previously published marks for CD4 T-cells[3,66] using Rabbit Ig-G[67] as input control. The columns on right indicate: (1) the % of the genome assigned to each state; (2) the % of each state within 2kb of a RefSeq TSS; (3) the fold enrichment for open chromatin as determined by DNase in CD4 T-cells[68], and (4) RefSeq transcripts excluding the promoter regions. Even though the model was trained using a distinct set of data for 9 cell types, we observe similar emission parameters for the states, similar genome coverages, and similar correspondences to gene annotations and chromatin accessibility, as seen for the original cell types.

Supplementary References

66. Wang, Z. et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics* **40**, 897-903 (2008).

67. Wang*, Z. et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1091-1031 (2009).

68. Boyle, A. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008).