SUPPLEMENTAL MATERIAL

1. Spark resources

Software

The Spark Java application (currently v1.2.0) is available for download from:
http://www.bcgsc.ca/downloads/spark/v1.2.0/

Spark is also accessible via the Genboree Workbench. To access Spark, login into Genboree at www.genboree.org, click on "Workbench" and from the "Epigenome" Menu, select "Cluster by Spark" under the "Analyze Signals" option.

User Guide

A detailed user guide is available within the Spark graphical user interface under the Help menu and is also available from:
http://www.bcgsc.ca/downloads/spark/v1.2.0/supportingMaterial/manual.html

Tutorial Videos

There are two tutorial videos available:

1. An overview of Spark's core functionality:
   http://www.bcgsc.ca/downloads/spark/v1.2.0/supportingMaterial/Spark_Tutorial.mp4
2. An overview of the interface to Spark within the Genboree Workbench:
   http://www.bcgsc.ca/downloads/spark/v1.2.0/supportingMaterial/Genboree_Spark_Tutorial.mp4

2. Clustering analyses for download

All clustering analyses used in this manuscript can be downloaded from:
http://www.bcgsc.ca/downloads/spark/v1.2.0/supportingMaterial/H1_hESC_analysis.zip

The analyses are labeled as follows:

(a) TSSs_k2_split-Fig2	Clustering across TSS using *k*=2 followed by manual cluster splits. Used for Figure 2 in the main manuscript.

(b) TSSs_k3-FigS1	Alternate clustering across TSSs using *k*=3. Used for Figure S1 (below).
(c) YY1peaks_k2_split-Fig3	Clustering across YY1 peaks using *k*=2 followed by manual cluster splits. Used for Figure 3 in the main manuscript.

3. Choosing an initial *k* value

We generated Figure 2 in the main manuscript by first clustering using *k*=2 and subsequently interactively splitting cluster 1 (c1) into two sub-clusters (c1-1 and c1-2). Given that we expect three clusters in this data set (active, poised, and inactive), it is natural to instead use a single clustering step with *k*=3. In fact, such a clustering produces similar results (Supplemental Figure S1). However, it is worth pointing out that the separation between the groups from a biological standpoint is better with the two successive *k*=2 cluster steps. This is qualitatively apparent from visual inspection of the cluster members in the region browser (cluster analyses can be downloaded from the link above).

One underlying assumption of *k*-means clustering is that the clusters have similar spatial extent. This is not the case in this example as the bivalent (poised) regions represent a much smaller group (roughly < 1% of TSSs in our H1 analyses). As a result, the poised cluster produced in the *k*=3 case is larger and less consistent than the one generated by successive clustering with *k*=2, where the transcriptionally active and poised groups are initially clustered together and are more easily separated in a subsequent step. This example demonstrates the value of allowing the user to guide the clustering, and in general, we recommend starting with a small initial *k* value and refining the clusters with interactive splits.

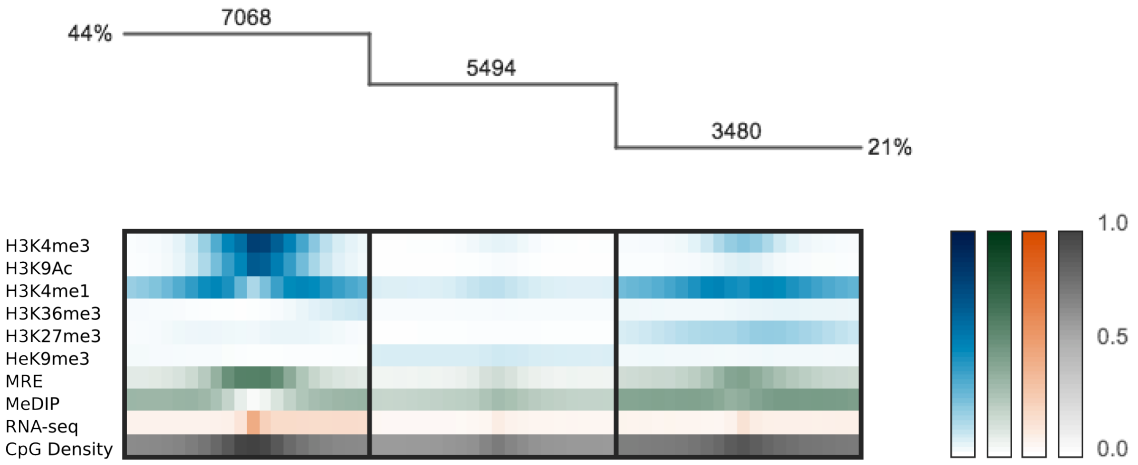

Figure S1. Clustering analysis centered on annotated TSSs using *k*=3. Figure features are as described for Figure 2 in the main text.