Alignment protocol

Individual libraries were combined across each lane of the flowcell for each read, and aligned using bwa[4] aln (0.7.4-r385) on the R1 and R2 separately. Following this, bwa sampe was used to combine the R1 and R2 sai files into paired end alignments. All ChIP sequencing data was aligned to the hg38 analysis set, downloaded from:

ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/hg38.analysisSet.chroms.tar.gz.

Duplicate sequences were flagged using the Picard MarkDuplicates tool (v1.123; http://broadinstitute.github.io/picard).

Peak calling protocol

Sequencing of input libraries was performed for each tissue sample as a means of correcting for background signal. Peaks were called using MACS2 (https://github.com/taoliu/MACS), using the following parameters: BAMPE, P = 0.05. We performed initial identification of enriched regions (peaks) using the MACS2[5] (version 2.1.1.20160309) callpeak function specifying a relaxed P-value threshold of 0.01, allowing one duplicate read per locus (keep-dup.2), pairing each ChIP-seq sample with its input control. We removed data aligning to genomic regions with aberrantly high signal due to copy number differences[6], and we defined low complexity regions using the "Duke Excluded Regions" and "DAC Blacklisted regions" tracks at the UCSC Genome Browser, which were lifted over to hg38. Statistics were gathered for each peak set, including peak counts at a wide range of MACS2-reported Q value and fold enrichment (FE) levels and total base coverage for selected levels. Because P and Q values are strongly affected by sequencing depth, FE provided the most consistent measure for thresholds across chromatin mark sets. From examination of peak counts and base coverage at various FE levels, we selected three significance levels: high (FE 4,4.5 or 5), target (FE 3,3.5 or 4), and low (FE 2.75 or 3). MACS2-generated narrowPeak files were converted to a BED6 format. All reported results are based on peaks at the target level. In preparation for model building, target level peaks were extracted and, for each mark, peaks from all tissues were merged using bedtools[7] (v2.26.0) in such a way as to preserve sample identity.

Modeling chromatin states

From the above merged consensus peaks for each experiment, we used ChromHMM[8] (v1.18) to build a 21-state model, using only LA tissue ChIP-seq data. To understand which genes are regulated by which states, we used the GENCODE annotation[9], version 29, and used bedtools closest to identify the distance between chromatin states and the closest protein-coding gene(s). For enhancers, we analyzed the closest gene even if it was >1 Mb in distance.