Supplementary Information

Methods

Quantitative metric of model selection

To determine a model of chromatin states that most closely represented our thyroid derived data, we selected a model with the most discrete and inter-sample consistent output state emissions. In other words, the model that is the most well defined, maximizing the homogeneity of epigenetic features in chromatin states across samples. Concretely, choosing a model based on this selection metric will make it so that the set of epigenetic features associated with a region partitioned as state 2 in one sample will tend to be similar (or homogeneous) to the set of epigenetic features associated with a region partitioned as state 2 in another sample. We provide an R package (hmmpickr available at https://github.com/csiu/hmmpickr) to help users select such a model (doi:10.5281/zenodo.398681). Overall, we choose the model that has the lowest homogeneity cost, which we compute as follows:

Let H represent the total number of histone marks and h represent a particular histone mark. Here $h = \{1,2, \ldots H\}$. We represent a probability close to 0 or 1, representing respectively absent and present histone marks across regions of the same state, by taking the minimum of the emission probability of a histone mark for a state (E_{hk}) and 1 minus that probability. To increase the penalty on states that are not as well defined, state costs are squared. To account for the difference in the number of states across models, we normalize by the number of states (K) in each model. Overall, we represent

the homogeneity cost of the state (d_k) and the homogeneity cost of a model (D) as follows:

$$d_k = \sum_{h=1}^{H} \min\{1 - E_{hk},\ E_{hk}\}$$

$$D = \frac{\sum_{k=1}^{K} {d_k}^2}{K}$$

Determination of Chromatin states

We used ChromHMM v1.12 (Ernst & Kellis 2012), an implementation of a hidden Markov model, to learn combinatorial chromatin states jointly across 8 thyroid epigenomes (a normal and diseased thyroid sample from each of the 4 thyroid sample donors). ChromHMM was trained using 6 histone marks (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, and H3K27me3). For each ChIP-seq data set, read counts were computed in non-overlapping 200bp bins across the entire genome. In total there were 15,181,508 bins. Each bin was discretized using ChromHMM's BinarizeBam into two levels: 1 indicating enrichment, and 0 indicating no enrichment. The binarization was performed by comparing ChIP-seq read counts to ChIP-seq input control data for local adjustments to the binarization threshold. We have also used ChIP-seq input (obtained before immunoprecipitation) control data as an additional input feature directly in the model. Reads mapping to chromosome Y were discarded to ensure reads that were mismapped were not carried forward in the computation. Command "LearnModel" with options "-p 11" was specified to use 11 processors in parallel to train a model using a standard Baum-Welch training algorithm. We trained a total of 26 models with the number of states ranging from 11 to 23 states. The trained model was then used to

compute the posterior probability of each state for each genomic bin in each sample. The regions were labelled using the state with the maximum posterior probability. To assign biologically meaningful labels to the states, we used ChromHMM package to compute the overlap and neighborhood enrichments of each state relative to coordinates of known functional annotation obtained from the Epigenome Roadmap Project (Roadmap Epigenomics Consortium *et al.* 2015). The chromatin state models and browser tracks can be downloaded from http://www.bcgsc.ca/data/thyroid.

Estimating transcript abundance, gene expression, and gene variance

We used Salmon v0.7.2 (Patro *et al.* 2017) to estimate transcript abundance from RNA-seq reads. As input, Salmon takes a reference transcriptome and a set of raw sequence reads. Each read is 75nt in length. The transcriptome was downloaded from the UCSC Table Browser with options as follows: group “Genes and Gene Predictions”, track “GENCODE Genes V19”, table “Basic (wgEncodeGencodeBasicV19)”, and output format “sequence”. The function “salmon index” was used to index the reference transcriptome, while “salmon quant” was used to estimate transcript abundance measured in transcripts per million (TPM). To integrate the transcript-level abundance estimates into gene-level abundance estimates, the tximport function of the tximport R package v1.2.0 (Soneson *et al.* 2015) was used to sum up the Salmon estimated transcript abundances within genes. This was also repeated for read counts. The regularized logarithm transformation (rlog) function of the DESeq2 R package v1.14.0 (Love *et al.* 2014) was then used to transform tximport generated read count data to

render them homoskedastic (i.e. such that the variance of the errors over the samples are similar). Gene variance was calculated on the rlog transformed read counts.

Selecting genes that have low expression in non-thyroid tissue types

Gene expression of various tissue types were obtained from the Genotype-Tissue Expression (GTEx) project. Data was downloaded for “Query: Genes matching: ‘’, specifically expressed in any Organism part above the expression level cutoff: 0 in experiment E-MTAB-2919” at https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2919 on November 21, 2016. Expression is measured in Fragments Per Kilobase of transcript per Million mapped reads (FPKM). We consider a gene as lowly expressed if the FPKM is less than or equal to 10. Expression values were then binarized to “low” and “high” expression. Genes for 52 non-thyroid samples were then clustered and visualized on a heat map. The cluster of genes that had low expression across all non-thyroid samples were then considered the set of genes that had low expression in 52 non-thyroid tissue types.

Motif analysis of Active Enhancers nearby genes

Sequence analysis of the genomic DNA in regions marked as active enhancer (state 10) by all four samples nearby (within 2 Mbp) the highly expressed 42 unique protein coding genes indicated 9 transcription factor motifs (p-value < 0.05). These transcription factor motifs (and consensus sequence) are Fosl2 (NATGASTCABNN), Jun-AP1 (GATGASTCATCN), Fra1 (NNATGASTCATH), MafK (GCTGASTCAGCA), BATF (DATGASTCAT), Atf3 (DATGASTCATHN), RORgt (AAYTAGGTCA), ZSCAN22 (SMCAGTCWGAKGGAGGAGGC), and Bach2 (TGCTGAGTCA). After Benjamini multiple test correction, the first 3 motifs (i.e. Fosl2, Jun-AP1, and Fra1) were found to be significantly enriched.

A similar sequence analysis for the 10 highly expressed genes consistent across the four specimens indicated 0 transcription factor motifs (p-value < 0.05).

Furthermore, a similar sequence analysis for the 18 genes that we consider epigenetically active and consistently expressed in the thyroid that are likely highly relevant to thyroid function indicated 2 transcription factor motifs (p-value < 0.05). These transcription factor motifs (and consensus sequence) are LXRE (RGGTTACTANAGGTCA) and ZNF675 (ARGAGGMCAAAATGW). After Benjamini multiple test correction, no motifs were found to be significantly enriched.

The list of all motifs and their significance are found in the Supplementary Excel file.

Distinct patterns of repressive marks

Detailed methodology for bisulfite-seq and data processing is available in the Supplemental Experimental Procedures of (Pellacani et al. 2016), at http://www.epigenomes.ca/protocols-and-standards, or upon request. We used FindER v1.0.0b (available at http://www.epigenomes.ca/tools-and-software/finder/index.html) with default options to find enriched ChIP-seq regions and bedtools v2.24.0 (Quinlan *et al.* 2010) to integrate DNA methylation profiles with repressive chromatin marks. Integration was done by (1) finding a list of H3K4me3 or H3K27me3 repressed regions unique to each individual, (2) mapping the fractional methylation call per 200 bp genomic bin onto the list of repressed regions, and (3) comparing the DNA methylation profiles across individuals for each of the repressed regions.

Quinlan AR & Hall IM 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26** 841–842