Characterization of the human thyroid epigenome

by

Celia Siu

B.Sc., The University of British Columbia, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

The University of British Columbia
(Vancouver)

March 2017

Abstract

The thyroid gland, necessary for normal human growth and development, is essential for the regulation of metabolism. Its function – to produce and secrete appropriate levels of thyroid hormone – is simple; however accurate assessment of thyroid abnormality is challenging and a fundamental understanding of the normal thyroid is therefore needed. One way to characterize the normal functioning of the thyroid gland is to study the epigenome and resulting transcriptome within its constituent cells. In this study, we compare the consistency of chromatin state annotations across the epigenomes from the grossly uninvolved tumour-adjacent thyroid tissue of four human individuals using ChIP-seq and RNA-seq. We profile four activating (H3K4me1, H3K4me3, H3K27ac, H3K36me3) and two repressing (H3K9me3, H3K27me3) histone modifications, identify chromatin states using a hidden Markov model, produce a novel metric for model selection, and establish epigenomic maps of 19 chromatin states. We found that epigenetic features characterizing promoters and transcription elongation tend to more consistent across epigenomes and that epigenetically active genes consistent across all epigenomes tend to have higher expression than those that are not marked as epigenetically active in all samples. We also identified a set of 18 genes epigenetically active and consistently expressed in the thyroid that are likely relevant to thyroid function. Altogether, we believe the epigenomes presented in this work represent a useful resource to gain a deeper understanding of the underlying molecular biology of thyroid function and provide contextual information of thyroid and human epigenomic data for comparison and integration into future studies.

Preface

The present work was done at the Michael Smith Genome Sciences Centre under the supervision of Dr. Steven JM Jones.

A version of this dissertation has been submitted to a journal as the “Characterization of the human thyroid epigenome”. Contributions were made by Sam Wiseman, Sitanshu Gakkhar, Alireza Heravi-Moussavi, Misha Bilenky, Annaick Carles, Thomas Sierocinski, Angela Tam, Eric Zhao, Katayoon Kasaian, Richard A Moore, Andrew J Mungall, Blair Walker, Thomas Thomson, Marco A Marra, Martin Hirst, and Steven JM Jones. Aligned RNA-sequencing and ChIP-sequencing bam files were provided through the author’s participation in the Canadian Epigenetics, Environment and Health Research Consortium. With regards to the characterization of the human thyroid epigenome, I was the main analyst of the project and the novel quantitative selection metric was developed together with Sitanshu Gakkhar. The manuscript and further edits were done by myself, with ideas, directions, and supervision from Dr. Steven JM Jones.

Table of Contents

Abstract ii

Preface iii

Table of Contents iv

List of Tables vi

List of Figures vii

List of Abbreviations xi

Acknowledgements viii

Dedication ix

Introduction 1

The human thyroid 2

Anatomy 2

Function 2

Thyroid hormone action 3

Thyroid hormone regulation 3

Abnormalities 4

Tests 4

Epigenetics 6

Chromatin structure 6

Epigenetic modifications 7

International efforts to map the human epigenome 7

Related works 9

Hypothesis 10

Research chapters 11

Data 11

Methods 12

ChIP-sequencing 12

RNA-sequencing 13

ChIP-seq enrichment analysis 13
Determination of chromatin states 14
Novel quantitative metric for model selection 15
Promoters 16
Estimating transcript abundance and gene expression 16
Estimating gene variance 17
Gene annotations 17
Motifs 17
Results 18
Reference epigenomes of thyroid tissue 18
Defining chromatin states 18
Chromatin states correlate with genomic features 21
Chromatin states stability 22
Epigenetically marked promoters and relation with gene expression 22
Enhancers 24
Transcript abundance 25
Epigenetically active and consistently expressed genes in the thyroid 26
Chromatin state defined by both H3K9me3 and H3K27me3 28
Reliability of a single reference 30
Conclusion 31
Tables 32
Figures 38
Bibliography 58

List of Tables

Table 1. Thyroid donor information. 32

Table 2. The 19-state model: state labels and average genomic coverage. Genomic coverage values were averaged across 4 normal thyroids samples. 32

Table 3. Motifs significantly enriched in genomic DNA epigenetically consistent at enhancers type chromatin states: 8 & 9 = genic enhancers, 10 = active enhancer, and 11 weak enhancers. Motif enrichment was performed using HOMER software and the top 3 motifs for each enhancer type chromatin state is given (Benjamini corrected p-values < 0.03). State 9 has enrichment in only 1 motif. 33

Table 4. The 25 most abundant transcripts for protein coding genes in each sample. In total, there are 42 unique genes. Text highlighted in grey represent genes (n=6) not determined to be epigenetically active (i.e. labelled as active TSS state 1 in the same bin) across 4 samples. Non-highlighted genes (n=36) are considered epigenetically active. 34

Table 5. GO Biological Process annotation of 18 acitvely transcribed and consistently expressed genes in the thyroid that do not have high expression in 52 non-thyroid GTEx tissues. GO annotations were obtained from Metascape. 35

List of Figures

Figure 1. Thyroid hormone regulation showing the direction of stimulation (normal arrow) and inhibition (blunt arrow). 38

Figure 2. Plots showing the homogeneity cost used for model selection. Formulation for the homogeneity cost is presented in the methods section. Scores were computed for 26 ChromHMM generated candidate models. The number of hidden states ranged from k = 11 – 23 states. Input was treated as a control (left) and as a mark (right). 19 states with input as a control and 20 states with input as a mark produced the lowest models with the lowest homogeneity cost. 19 states with input as control was chosen for the model to use for further analysis. 39

Figure 3. 19-state model with input as control. Chromatin states were defined using the ChromHMM software. The figure shows: (A) chromatin state definitions, histone mark probabilities, transition probabilities, (B) CEMT_44 genomic feature enrichments, and (C) CEMT_44 neighborhood enrichments around RefSeq TSSs and TESs. Average genomic coverages are given in Table 2. 40

Figure 4 Boxplot showing the methylation levels across chromatin states. Fractional methylation calls were computed based on the $\frac{\text{number of CpG reads}}{\text{total number of CpG reads}}$ for each genomic bin. Values were summarized for each normal sample to which bisulfite-seq data was available at the time of analysis. 41

Figure 5. Screenshot of the UCSC Genome Browser showing tracks for the 19-state model around the thyroglobulin gene. These tracks can be viewed on the UCSC Genome Browser through a link provided in http://www.bcgsc.ca/data/thyroid. (A) The consistency of chromatin states across 4 epigenomes. We show the tracks for states 1 (active TSS) and 10 (active enhancer). The tracks for the remaining 17 states are hidden from view. (B) The overview of ChromHMM state segmentations for each sample. (C) Predefined tracks for gene annotations from RefSeq, UCSC, and Ensembl; CpG islands; and repeat elements by RepeatMasker. 42

Figure 6. Overview of epigenetic consistency across 4 thyroid epigenomes. The genome was divided into 15,181,508 bins. Each bin is 200bp in length and is

marked by a chromatin state. For a particular bin across different individuals, the chromatin state may be the same or it may be different. If a bin was partitioned as state 1 consistently across 4 samples, then the bin count for state 1 at x = 4 is incremented. If the states for a bin across 4 samples were {1, 1, 2, 1}, then the bin counts for state 1 at x = 3 and state 2 at x = 1 is incremented. We define a bin as epigenetically consistent when the chromatin state is the same across all individuals. (A) Histogram showing the number of genomic bins sharing the same state across 4 epigenomes. (B) Values from (A) scaled to 0 and 1 showing that states 1, 5, and 7 tends to more epigenetically consistent than every other state excluding quiescent state 19. (C) Heat map showing the average probability of finding a bin partitioned to the same chromatin state in 0, 1, 2, or 3 other epigenomes.....43

Figure 7. Association of chromatin state 1 "Active TSS" with protein coding genes. (A) Histogram showing the number of genomic bins partitioned to state 1 in 1, 2, 3, or 4 epigenomes. Orange represents state 1 bins located within promoters (TSS +/- 1kbp) of known protein coding genes. (B) Histogram showing the number of protein coding genes partitioned as state 1 across the 4 epigenomes; values are 6979, 947, 754, 1014, 10460. (C) Plot showing the percentile of expression (log10-scaled, values from CEMT_44) in the set of genes epigenetically active in 0, 1, 2, 3, and 4 epigenomes. Genes with 0 expression were removed. (D) Expression (log10-scaled, values from CEMT_44) across genes that are epigenetically active in 0, 1, 2, 3, and 4 samples. Genes with 0 expression were removed. (E) Proportion of genes in different brackets of expression (values from CEMT_44). Total number of genes in each bracket is shown on top. Color represents the number of samples sharing the same genomic bin. 44

Figure 8. Genes epigenetically active in only 1 sample. (A) Gene counts. (B) Probability of metascape gene set enrichments. 45

Figure 9. Average proportion of transcripts in the top 10,000 most abundant protein coding genes. Genes were ranked according to transcript abundances. The gene at rank 1 is the most abundant gene in a given sample. The average transcript proportion by gene rank were computed across 4 thyroid samples and is

shown in the blue line. The grey ribbon is the mean proportion of transcripts +/- 2 standard deviations. 46

Figure 10 Screenshot of the UCSC Genome Browser showing *RPS24* as being inconsistently marked as active TSS across different bins within the gene promoter. In comparison, *POLR3A* is consistently marked as active TSS within the gene promoter. 47

Figure 11. 137 epigenetically active and consistently expressed genes in the thyroid. (A) Epigenetically active and consistently expressed genes were identified based on criteria as follows: is epigenetically marked as state 1 across all 4 epigenomes, have high expression, and have low variance. (B) Metascape gene set enrichment of the 137 genes. 48

Figure 12. Heat map highlighting 18 genes (green cluster) epigenetically active and consistently expressed in the thyroid with low expression in 52 non-thyroid tissues obtained from GTEx. Blue represent FPKM >= 10, white represents otherwise. Genes in dark blue cluster are present in all tissues, whereas genes in the light blue cluster are present in a subset of predominately non-brain related tissues. 49

Figure 13. Chromatin state overlap enrichment of repeat regions close to and far from protein coding genes. We consider a gene as close if it is within 10kbp. Coordinates were obtained from RepeatMasker downloaded from the UCSC Table Browser. Overlap enrichment was performed using ChromHMM software. The enrichment values displayed is the average of values from 4 normal thyroid epigenomes. 50

Figure 14. The 19-state model applied to 15 normal colon epigenomes. (A) Histogram showing the number of genomic bins sharing the same state across 4 epigenomes. See Figure 6 for details. (B) Values from (A) scaled to 0 and 1. 51

Figure 15. Pearson correlation of state emissions between ChromHMM models. The 19-state model presented in this work is on the y-axis. The 18-state model published in Roadmap (Roadmap Epigenomics Consortium, et al., 2015) is on the x-axis. 52

Figure 16. Plot showing the heterogeneity cost for model selection on models trained on 15 normal colon reference epigenomes. Training was specified for k = 14 – 20 states. Input was treated as a control and training was done on 15 normal colon epigenomes using ChromHMM software. 53

Figure 17 The 19 states generated from the epigenomes of thyroid (left) and colon (right) samples differ. The states and emission probabilities were produced using ChromHMM. 54

Figure 18. Overview of histone modification consistency across 4 thyroid epigenomes. We used FindER (A) and MACS2 regular and broad (B, C) peak callers to find enriched regions. The genome was divided in 15,181,508 bins. Each bin is 200bp in length and was discretized into two levels: 1 indicating enrichment, and 0 indicating no enrichment. For a particular bin across different individuals, the enrichment of a particular histone may be present in all (x = 4), some (x = {1, 2, 3}), or no (x = 0) individual. 55

Figure 19 Overview of state consistency across 98 epigenomes published in (Roadmap Epigenomics Consortium, et al., 2015). The segmentations of the 18 Roadmap states for each epigenome were obtained from http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html 57

List of Abbreviations

AMED-CREST
: Japan Agency for Medical Research and Development Core Research for Evolutional Science and Technology.

ATC
: Anaplastic Thyroid Carcinoma.

bp
: Base pair.

CEEHRC
: Canadian Epigenetics, Environment and Health Research Consortium.

CEMT
: Centre for Epigenome Mapping Technologies.

CREST
: Core Research for Evolutional Science and Technology.

EEF1A1
: Eukaryotic Translation Elongation Factor 1 Alpha 1.

EGA
: European Genome-phenome Archive.

ENCODE
: Encyclopedia of DNA Elements.

eQTL
: Expression quantitative trait loci.

ESCRT-II
: Endosomal sorting complex required for transport II.

ETFB
: Electron transfer flavoprotein beta subunit.

FMR1
: Fragile X Mental Retardation 1.

FNA
: Fine needle aspiration.

FTC
: Follicular Thyroid Carcinoma.

FXR1
: FMR1 autosomal homolog 1.

H2AFY
: H2A histone family member Y.

HBA2
: Hemoglobin Subunit Alpha 2.

HBB
: Hemoglobin Subunit Beta.

HEROIC
: High-throughput Epigenetic Regulatory Organisation In Chromatin.

HMM
: Hidden Markov model.

HN12
: Humanin-Like 12.

HOX
: Homeobox.

IHEC
: International Human Epigenome Consortium.

KNIH
: Korea National Institute of Health.

MRT
: Malignant rhabdoid tumours, 6

MT1G
: Metallothionein 1G.

MTC
: Medullary Thyroid Carcinoma.

MTRNR2L12
: MT-RNR2-Like 12.

mU/L
: Milliunits per liter.

N4BP2L2
: NEDD4 binding protein 2 like 2.

NEDD4
: Neural Precursor Cell Expressed Developmentally Down-Regulated Protein 4.

NIH
: National Institutes of Health.

NSMCE1
: Non-SMC Element 1 Homolog, SMC5-SMC6 complex component.

NT5C2
: 5'-nucleotidase, cytosolic II.

PMF1
: Polyamine modulated factor 1.

pmol/L
: Picomoles per liter, 5

PTC
: Papillary Thyroid Carcinoma.

RXR
: Retinoid X receptor.

SCAF11
: SR-related CTD associated factor 11.

SNF8
: SNF8, ESCRT-II complex subunit.

SNV
: single nucleotide variants.

SORD
: Sorbitol dehydrogenase.

SPG11
: Spastic paraplegia 11 (autosomal recessive).

T3
: Triiodothyronine.

T4
: Thyroxine.

TCTN1
: Tectonic family member 1.

TG
: Thyroglobulin.

TOR1AIP1
: Torsin 1A interacting protein 1.

TPD52
: Tumor protein D52.

TPM
: Transcripts per million.

TRH
: TSH Releasing Hormone.

TSH
: Thyroid stimulating hormone.

TSS
: Transcription start site.

VEZT
: Vezatin, adherens junctions transmembrane protein.

WBSCR22
: Williams-Beuren syndrome chromosome region 22.

Acknowledgements

Dr. Steven Jones is a smart and busy man. I would like to thank him for all the support, guidance, opportunities he has given me throughout the course of this project. I also am thankful to my lab mates for keeping it interesting, geeking it out, and bringing cake to trainee meetings.

Thank you to my committee members, Dr. Martin Hirst and Dr. Sohrab Shah, for all their guidance, advice, and kind support. Thank you to Sitanshu Gakkhar, Alireza Heravi-Moussavi, and Misha Bilenky, for their insights, directions, and discussions we have whenever I'm stumped.

I am also thankful to Dr. Sam Wiseman, Jordan Wong, and Kaidi Liu for the opportunity for being involved with doing the analysis for the “Management of PET Diagnosed Thyroid Incidentalomas in British Columbia Canada: Critical Importance of the PET Report” (in submission to the Journal of Surgery).

Dedication

To family

2017-03-10

Introduction

The main role of the thyroid is to produce and secrete hormones necessary for growth and development. In humans, thyroid dysfunction has been associated with infertility and poor pregnancy outcomes (Crawford & Steiner, 2016). Furthermore, thyroid cancer, the most common endocrine malignancy, was previously the 14th 20 years ago but is now the 5th most frequent cancer in women (Vigneri, Malandrinoa, & Vigneri, 2015).

Routine assessment of a healthy or diseased thyroid function state is currently based on blood serum concentrations of thyroid related hormones such as Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), or Thyroxine (T4) within a predefined “normal” range (Führer, Brix, & Biebermann, 2015). However, the definition of a “normal” TSH range and similarly “normal” T3 and T4 concentrations remains the subject of debate in different countries worldwide (Führer, Brix, & Biebermann, 2015). Furthermore, the variability in individual factors such as sex, body mass index, exclusion of incident thyroid disease, ethnicity, and iodine and selenium intake impacts a more comprehensive definition of healthy thyroid hormone ranges within a given population (Führer, Brix, & Biebermann, 2015).

In summary, the incidence of thyroid abnormalities is increasing and accurate assessment of abnormal thyroids states across different individuals is challenging. Overall, a fundamental understanding of the normal thyroid is needed.

The human thyroid

Anatomy

The human thyroid is a relatively homogeneous tissue composing of two types of cells: follicular cells and parafollicular cells. Follicular cells are thyroid epithelial cells and are responsible for the production, storage, and secretion of thyroid hormones. Spatially, the follicular cells are arranged into spherical units known as follicles and the lumen of each follicle is the stored secretion of the follicular cells known as the colloid. On the other hand, parafollicular cells – sometimes referred to as C cells – are responsible for the production of calcitonin, a hormone playing a minor role in calcium homeostasis. Spatially, parafollicular cells are scattered sparsely around follicles and account for only a relatively small percentage of the thyroid. According to a review by (Eladio & Gershon, 1978), despite inter- and intra- species differences, sampling variation, functional differences in parafollicular cell activity, and the relative non-specificity of some histochemical procedures used to estimate C cell number, there is good agreement of the relatively small percentage – 1% of thyroid cells in mice, 1% in pig, 2% in rabbit, 13% in guinea pig, and 1-6% in rat – of C cells compared to the follicular cells.

With regards to blood, the thyroid gland is richly vascularized with blood supplying from the superior and inferior thyroid arteries and draining into the upper, middle, and lower thyroid veins (Jancic & Stosic, 2014). Blood flow is profuse, with little resistance for the exchange of nutrients, gases, and hormones. With regards to lymphatics, the drainage of lymph (produced in consequence by the exchange from the capillary bed) is also well developed. With regards to nerve supply, the thyroid is connected to the vagus and sympathetic nerves. According to (Harris & Donovan, 1961), the thyroid nerves are probably vasomotor in function and indirectly influence the thyroid gland by altering blood supply.

Function

The thyroid gland is the largest endocrine gland in the human body, it produces and secretes T3 and T4 thyroid hormones directly into the blood stream. These hormones are then used for the regulation of metabolism in every cell of the human body.

Predominantly, the thyroid gland produces the less active T4 form and within cells this form is converted to the more active T3 form by deiodinases (Panicker, 2011). According to (Harris & Donovan, 1961), a component of functional thyroid hormones is iodine and the follicular cell is the only cell in the human body to uptake iodine. In addition to the production and secretion of thyroid hormones, the thyroid gland serves as a reservoir to store thyroglobulin (a precursor of thyroid hormones), T3, and T4 in the colloid of the thyroid follicles (Jancic & Stosic, 2014).

Thyroid hormone action

Once released from the thyroid gland into circulation, T3 and T4 thyroid hormones are taken into cells by thyroid hormone transporters; examples include the monocarboxylate 8 (MCT8) transporter and the organic-anion transporting polypeptide 1C1 (OATP1C1) (Panicker, 2011). Once within cells, the less active T4 form is converted into the active T3 form by iodothyronine deiodinases: D1, D2, and D3 (Panicker, 2011). To regulate the transcription of thyroid responsive genes, T3 then moves to the cell nucleus where it binds thyroid hormone receptors resulting in a change in the formation and binding of the receptor to DNA (Panicker, 2011). According to (Panicker, 2011), the binding of thyroid hormone receptors are often heterodimeric with retinoid X receptor and the binding action is also influenced by co-regulator proteins which can bind once T3 is bound to the receptor.

Thyroid hormone regulation

With regards to thyroid hormone regulation, T3 and T4 thyroid hormones are released into the blood by the thyroid gland under the stimulation of TSH from the anterior pituitary gland (Panicker, 2011). When the level of TSH is low, TSH Releasing Hormone (TRH) is released by the hypothalamus to stimulate the anterior pituitary gland to produce TSH (Figure 1). As the levels of T3 and T4 increase, T3 and T4 negatively feedback to the anterior pituitary gland and hypothalamus to inhibit the production and release of TSH and TRH (Figure 1). Overall, T3, T4, TSH, and TRH hormone levels remain stable.

Abnormalities

Abnormal levels of thyroid hormone, thyroiditis (inflammation of the thyroid gland), degeneration, and neoplasms are some abnormalities associated with an unhealthy thyroid. Too much thyroid hormone is hyperthyroidism and too little is hypothyroidism. Hyperthyroidism may result from Grave's disease (also known as toxic diffuse goiter), an autoimmune disease characterized by the swelling of the neck and protrusion of the eyes resulting from an overactive thyroid gland (Harris & Donovan, 1961). On the other hand, hypothyroidism may result from Hashimoto's disease, an autoimmune disease characterized by chronic inflammation thyroiditis and an underactive and subsequent failure of the thyroid gland (Davies, Latif, & Yin, 2012). Two diseases characterized by the degeneration of the thyroid gland include Myxedema (or Gull's disease) occurring during adult age and cretinism occurring during childhood (Harris & Donovan, 1961). Neoplasms (or new and abnormal growth) of the thyroid may be benign such as goiter and follicular adenoma or malignant such as thyroid carcinomas. Goiter is the visible swelling of the neck due to enlargement of the thyroid gland (Harris & Donovan, 1961). Follicular adenomas are encapsulated benign tumors (Lai & Chen, 2015). The four main types of thyroid carcinoma include papillary thyroid carcinoma (PTC), follicular thyroid carcinoma (FTC), medullary thyroid carcinoma (MTC), and anaplastic thyroid carcinoma (ATC). The most frequent thyroid carcinoma (as well as most frequent endocrine carcinoma) is PTC and the one with the worst prognosis is ATC. Approximately 80-85% of all thyroid cancers are PTC and 90% of patients with ATC die within 6 months of diagnosis (Lin, 2011)

Tests

Blood tests, imaging tests, and fine-needle aspiration (FNA) biopsies are some ways in which the thyroid gland is assessed (Bomeli, LeBeau, & Ferris, 2010). Blood tests are used to assess thyroid function by measuring the levels of thyroid hormone circulating in the blood. A hormone is considered free if it is not bounded to proteins. According to the healthcare diagnostic provider LifeLabs (Canada) reference ranges (LifeLabs, 2016), for female and male adults over 20, the normal range is considered 0.32 – 5.04 milliunits per liter (mU/L) for TSH, 10.6 – 19.7 picomoles per liter (pmol/L) for free T4, and 3.00 –

5.90 pmol/L for free T3. These ranges can differ from lab to lab and geographically. Furthermore, a test result within laboratory reference limits is not necessarily normal for an individual (Andersen, Pedersen, Bruun, & Laurberg, 2002). With regards to imaging tests, ultrasounds, computed tomography (CT) scans, and positron emission tomography (PET) scans are used – sometimes in conjunction with the uptake of radioactive isotopes including 18-flurodeoxyglucose (F18-FDG) (Bertagna, et al., 2013) or radioactive iodine-131 (Harris & Donovan, 1961) – to look for abnormal growth. With regards to FNA biopsies, a hollow needle is inserted into the thyroid through the neck to collect a sample of cells for analysis. This technique is fairly inexpensive and excels at identifying PTC with diagnostic accuracy as high as 95% in skilled hands with experienced cytopathologic staff (Patel, et al., 2011). In comparison, distinguishing between a benign follicular adenoma, malignant FTC, and malignant follicular variant of PTC is problematic with FNA for these neoplasms cannot be differentiated cytopathologically (Patel, et al., 2011). Instead, surgery is performed to check for the presence of capsular and vascular invasion characteristic of malignant neoplasms (Patel, et al., 2011). According to (Patel, et al., 2011), inconclusive diagnosis from “suspicious” or “follicular-patterned lesion” occur roughly 10-30% of the time and 8-17% of these suspicious nodules are determined to be malignant after surgical removal. Nevertheless, FNA biopsy and cytologic analysis is an integral and invaluable tool in the comprehensive evaluation of the thyroid nodule (Cannon, 2011), but there is room for improvement with regards to thyroid related laboratory tests.

Epigenetics

Epigenetic processes play a role in the regulation of transcription and gene expression. The term epigenetics, originally introduced by Waddington in 1942, was used to describe the mechanism lying between and connecting the genotype and the phenotype (cited in (Jablonka & Lamm, 2012)). Today, epigenetics is a broad field of study referring to heritable changes in the regulation of gene activity and expression that are not dependent on the underlying DNA sequence. Previous studies have found epigenetic components in health and disease. For instance, in adult de novo acute myeloid leukemia, 44% of cases described DNA-methylation-related genes mutations (The Cancer Genome Atlas Research Network, 2013); in malignant rhabdoid tumours (MRT) associated with SMARCB1 loss, there was evidence for epigenetic reprogramming of homeobox (HOX) genes including the loss of H3K27me3 at HOX promoters and MRT-specific super enhancers at HOXA, HOXB, and HOXC clusters (Chun, et al., 2016); and in ependymomas, CpG hypermethylation at promoters containing CpG islands was found to be higher in ependymomas predominantly found in infants (which is associated with poor prognosis in spite of maximally aggressive therapy) than those found in older children and adults (Mack, et al., 2014).

Chromatin structure

Chromatin – a complex macromolecule made up of DNA, protein, and RNA – functions to package the DNA into a smaller volume within the nuclei of a cell. The basic repeating unit of chromatin, a nucleosome, consists of about 200 base pairs (bp) of DNA wrapped around a histone protein octamer. This octamer – made up of two copies of each core histone H2A, H2B, H3, and H4 – can be chemically modified to signal an activation or repression of transcription. Broadly, chromatin can be classified into two categories: euchromatin (a loose and transcription permissive structure) and heterochromatin (a dense and transcription repressed structure). Overall, DNA is packaged, reinforced to prevent DNA damage, and controlled by the interaction with proteins regarding replication and gene expression.

Epigenetic modifications

Epigenetic, referring to the reversible changes in chromatin and DNA that can regulate gene activity and expression, include the post-translational modifications of histone proteins at their N-terminal tails and DNA methylation. At the most basic level, when acetylated, positively charged histones tend to be less positively charged and will result in the loosening of negatively charged DNA from chromatin and allows access to transcription machinery for gene expression. Examples of histone modifications include H3K4m3, which have been associated with active promoters; H3K27ac to active promoters and enhancers; H3K4me1 to active enhancers; H3K36me3 to transcribed gene bodies; H3K9me3 to heterochromatin; and H3K27me3 to Polycomb repressed regions (Roadmap Epigenomics Consortium, et al., 2015). To map the putative locations of these marks onto the genome, chromatin immunoprecipitation sequencing (ChIP-seq) experiments – involving the crosslinking, shearing, extracting, un-crosslinking, amplifying, and sequencing of DNA – are used. Methylated DNA occurs at cytosine nucleotides and is often associated with gene silencing. In the human genome, there are roughly 28 million CpG dinucleotides (Ernst & Kellis, 2015) and over 28 thousand CpG islands (Karolchik D, 2004). To determine the pattern of DNA methylation mapping across the genome, bisulfite sequencing experiments – involving the treatment of DNA with bisulfite to convert un-methylated cytosines to thymines – are used.

International efforts to map the human epigenome

Genome wide epigenomic maps of functional elements encompassing promoters, enhancers, silencers, and transcription factor binding sites across an increasing number of different cell types and tissues have been generated (Roadmap Epigenomics Consortium, et al., 2015). Earlier projects focusing on understanding, cataloging, and identifying epigenetic processes include the Human Epigenome Project and the High-throughput Epigenetic Regulatory Organisation In Chromatin (HEROIC) consortium (2005-2010). Large international initiatives contributing to the mapping of the human epigenome include the International Human Epigenome Consortium (IHEC) (Stunnenberg, Consortium, & Hirst, 2016), which set standards for experimental setup,

meta data collection, data storage, and data analysis. IHEC also coordinates the contribution from seven international consortia – ENCODE (Dunham, et al., 2012), NIH Roadmap (Roadmap Epigenomics Consortium, et al., 2015), CEEHRC (CEEHRC, 2016), BLUEPRINT (Martens & Stunnenberg, 2013), DEEP, AMED-CREST, and KNIH – and aims to sequence and decipher 1000 human epigenomes of various cell types (Bujold, et al., 2016).

The Encyclopedia of DNA Elements (ENCODE) project – funded by US National Institutes of Health (NIH) – aims to identify all functional elements in the human genome (Dunham, et al., 2012). Building on the work of ENCODE, the NIH Roadmap Epigenomics Project – generating the largest collection so far of human epigenomes for primary cells and tissues (Roadmap Epigenomics Consortium, et al., 2015) – aims to analyze samples taken directly from human tissues and cells for the understanding of how epigenetic processes contribute to human biology and disease (Roadmap Epigenomics Consortium, et al., 2015). Similarly, BLUEPRINT, a European Union-funded project, focuses on the generation of at least 100 hematopoietic epigenomic maps of a wide variety of cell types from the blood of healthy and diseased individuals (Martens & Stunnenberg, 2013)

Other international efforts include the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network from Canada, the Japan Agency for Medical Research and Development Core Research for Evolutional Science and Technology (AMED-CREST) program from Japan, the German epigenome programme ‘DEEP’ from Germany, and the Korea National Institute of Health (KNIH) from South Korea. In summary, there is work focused towards the common goal of generating epigenomic references across the world.

Related works

Various works have been published to supplement and analyze epigenetic data. Tools such as ChromImpute (Ernst & Kellis, 2015) and Epigram (Whitaker, Chen, & Wang, 2015) have been developed to deal with missing epigenetic data by predicting, respectively, signal tracks for mark-sample combinations not experimentally mapped and the epigenome based on DNA motifs. Other tools such as ChromHMM (Ernst & Kellis, 2012) using a hidden Markov model or Segway (Hoffman, et al., 2012) using a Dynamic Bayesian Network have been developed to represent different combinations of epigenetic features by partitioning the epigenome into various defined chromatin states. Other methods of segmentation include the modelling of topological domains by wavelet transformations for predicting active and repressive states (Chen, Wang, Xuan, Chen, & Zhang, 2016) and the use of nucleotide-sequence-based Markov chains to refine the chromatin map produced by ENCODE (Lee & Park, 2016). With these chromatin states, tools such as ChromDiff (Yen & Kellis, 2015) have been developed to identify chromatin state differences across groups of epigenomes.

Groups generating epigenomes such as ENCODE (Dunham, et al., 2012) and Roadmap (Roadmap Epigenomics Consortium, et al., 2015) then use these tools to produce chromatin state reference annotations which are then further used in various studies. For instance, excess rare SNVs were observed to be significantly different in schizophrenia versus control cases at Polycomb prepressed states (González-Peñas J. , et al., 2016); rheumatoid arthritis associated SNPs were found in the enhancer chromatin state in memory but not naïve T cells (Orent, et al., 2016); and the enhancer states were compared between mouse and humans to reveal an immune basis of Alzheimer's disease (Gjoneska, et al., 2015). Furthermore, promoter related chromatin states have been used to profile core promoter elements (Lent, Lee, & Park, 2015) and the chromatin state of maternal and embryonic *Xenopus* were compared to highlight the extent maternal factors shape chromatin state in *Xenopus* embryos (Hontelez, et al., 2015).

There are also published studies comparing normal samples. For instance, (Roberto, et al., 2016) compared the microarray gene expression profile of normal cells surrounding tumors of thyroid cancer for neoplastic and non-neoplastic thyroid disease; (Wijetunga, et al., 2014) looked at the epigenetic variability in the same cell type between healthy individuals; and (Gascard, et al., 2015) studied the epigenetics and transcriptional determinants of the human breast between three normal individuals.

Hypothesis

Overall, a fundamental understanding of the normal thyroid is needed. One way to characterize the normal thyroid is to study its epigenome and matched transcriptome across different individuals. I hypothesized that the epigenetic features important in the function of the normal thyroid would be consistent between different individuals. We therefore want to understand and characterize regions of epigenetic regulation which are consistent and regions which are variable across the normal thyroids of different individuals. Overall, I characterized an available reference thyroid epigenome as a resource and reference of human epigenome data useful for comparison and integration of future studies.

Research chapters

Data

Four human adult thyroid specimens were provided from surgical resections conducted at St. Paul's Hospital, Vancouver, British Columbia. The pathologic findings in the glands included two follicular adenomas, one goiter, and one papillary carcinoma. The pathologic findings reflect the challenge of obtaining normal thyroid tissue from healthy individuals. In the case of the thyroid, a biopsy sample is obtained by a procedure called fine-needle aspiration (FNA), whereby a hollow needle is inserted into the thyroid – through the neck – to collect a sample of cells. Following the procedure, patients may feel sore and common complications include local pain, discomfort, and minor hematomas (abnormal collection outside the blood vessel). Furthermore, there have been reported cases of acute transient thyroid swelling (Norrenberg, et al., 2011) (Nakatake, Fukata, & Tajiri, 2012), cutaneous sinus formation (Akbaba, et al., 2014), and transient bradycardia (abnormally slow heart action) and faintness (Silverman, et al., 1986) following a FNA. Furthermore, the quantity of cells collected from a FNA may not be sufficient for genetic and epigenetic profiling and larger surgical operations may be required for sample collection. The specimens referred to as "normal" in this study are from microscopically uninvolved thyroid tissue in the resected thyroid glands.

The ChIP-seq and RNA-seq data of 4 adult human thyroid and 15 adult human colon samples were obtained from the Centre for Epigenome Mapping Technologies (CEMT) branch of the CEEHRC Network. Thyroid tissue sample donor information is presented in Table 1.

Methods

ChIP-sequencing

Human thyroid ChIP-seq data was as previously described (Pellacani, et al., 2016). In brief, this procedure involves the (1) cross linking of DNA to proteins using formaldehyde, (2) lysing of cells or tissues, (3) shearing of DNA into smaller fragments by sonification, (4) recovering of DNA-protein complexes by immunoprecipitation using specific antibodies, (5) reversing cross links, (6) ligating on sequence adaptors, (7) amplifying DNA using PCR, and (8) sequencing of DNA. The antibodies were obtained from Diagenode (x3), Abcam (x2), and Cell Signaling (x1); and the catalogue numbers are, respectively, C15410037/pAb-037-050 (H3K4me1), C15410056/pAb-056-050 (H3K9me3), C15410195/pAb-195-050 (H3K27me3), ab4729 (H3K27ac), ab9050 (H3K36me3), and 9751S (H3K4me3). DNA input fractions (obtained before immunoprecipitation) were also sequenced as control. For sample CEMT_86/87, one lane of sequencing was merged with native ChIP protocol. 75 base pair paired-end reads were sequenced on Illumina HiSeq 2500 (Illumina Inc., USA). Sequenced reads were then split by index, adaptors were trimmed, and the fastq files corresponding to the two mate pairs were generated for each index. The reads were aligned to GRCh37-lite reference using Burrows-Wheeler Aligner v0.5.7 (Li & Durbin, 2009), converted to bam format with SAMtools v0.1.13 (Li, et al., 2009), and annotated using CEEHRC in-house tools (including flagging of chastity failed reads) and Picard Tools' MarkDuplicates.jar v1.71 (Broad Institute). Processed datasets and all underlying raw DNA sequences have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/) under accession number EGAS00001000552. In this work, CEMT_40-45 and CEMT_86-87 were the thyroid samples and CEMT_33-34, CEMT_50-61, and CEMT_72 were the colon samples used for analysis. Detailed methodology for library construction, read alignment, 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.

RNA-sequencing

Human RNA-seq data was as previously described (Pellacani, et al., 2016). In brief, this procedure involves the (1) purification of RNA followed by poly-A RNA selection, (2) conversion of RNA to cDNA by random priming, (3) fragmentation of cDNA, (4) ligation of adaptors, (5) amplification of DNA by PCR, and (6) sequencing of DNA. 75 base pair paired-end reads were sequenced on Illumina HiSeq 2500 (Illumina Inc., USA). Adaptors were trimmed and the fastq files corresponding to the two mate pairs were generated. The reads were aligned to a genome + transcription reference using Burrows-Wheeler Aligner v0.5.7 (Li & Durbin, 2009) and converted to bam format with SAMtools v0.1.13 (Li, et al., 2009). The resulting bam files were repositioned to GRCh37-lite using JAGuaR v2.0.2 (Butterfield, et al., 2014) and annotated using CEEHRC in-house tools (including flagging of chastity failed reads) and Picard Tools' MarkDuplicates.jar v1.71 (Broad Institute). Processed datasets and all underlying raw DNA sequences have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/) under accession number EGAS00001000552. In this work, CEMT_40-45 and CEMT_86-87 were the thyroid samples used for analysis. Detailed methodology for library construction, read alignment, 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.

ChIP-seq enrichment analysis

We used FindER v1.0.0b (CEEHRC, 2016) and MACS2 v2.1.1.20160309 (Zhang Y, 2008) to find enriched regions. We called peaks (i.e. regions of enrichment) using FindER with default options. We called peaks using MACS2 “callpeak” with options as follows: “-B --nomodel --extsize 200 --SPMR -g hs”. For broad MACS2 peaks, we used the same options with an additional “--broad" argument.

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 DNA input control data for local adjustments to the binarization threshold. We have also used ChIP-seq DNA input control data as an additional 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 (as opposed to the default incremental expectation-maximization algorithm when "-p" is not specified). 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 neighbourhood 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.

Novel quantitative metric for model selection

To determine a model of chromatin states that most closely represented our thyroid data, we selected a model with the most discrete 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:

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

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

Promoters

In this study, promoters were defined to be regions around the annotated transcription start site (TSS) +/- 1kbp. We decided to use 1kbp from the TSS for this distance encapsulates the promoter signal as observed in the RefSeq TSS neighborhood enrichments generated by ChromHMM (Figure 3C). The coordinates for the TSS promoter regions were obtained from the Ensembl GRCh37 Release 75 Gene sets GTF file available at http://feb2014.archive.ensembl.org/info/data/ftp/index.html. Gene sets was filtered for "protein_coding" (source), "transcript" (feature), on chromosomes 1-22, X, and Y. In total, we obtain 81,732 transcripts deriving from 20,314 protein coding genes across the autosomes. Altogether, the 20,314 genes encompass 43% of the genome, with their exons and coding sequences representing 2.5% and 1.2% of the genome respectively.

Estimating transcript abundance and gene expression

We used Salmon v0.7.2 (Patro, Duggal, Love, Irizarry, & Kingsford, 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 75 nucleotides in length. The transcriptome used 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, tximport R package v1.2.0 (Soneson, Love, & Robinson, 2015) was used to sum up the Salmon estimated transcript abundances within genes. The tximport function of the tximport R package also computes gene level read counts by the same method.

Estimating gene variance

We used the regularized logarithm transformation (rlog) function of the DESeq2 R package v1.14.0 (Love, Huber, & Anders, 2014) 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). The rlog transformation behaves similarly to a log2 transformation for genes with high counts, while shrinking together the values for different samples and avoiding the problem of spreading apart of data for genes with low counts (Love, Huber, & Anders, 2014). Gene variance was calculated on the transformed read counts.

Gene annotations

We used Metascape v3.0 (Tripathi, et al., 2015) – available at http://metascape.org/ – to annotate gene lists.

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.

Motifs

We used HOMER v4.8 (Heinz, et al., 2010) to find enriched motifs in genomic regions using “findMotifsGenome.pl” with options as follows: “-size given”.

Results

Reference epigenomes of thyroid tissue

Reference epigenomes have been used to describe regions of functional interest such as promoter regions or transcription factor binding sites (Roadmap Epigenomics Consortium, et al., 2015) and they have been used to provide context to specific genomic locations such as single nucleotide variants (SNVs) or expression quantitative trait loci (eQTLs) (González-Peñas J. , et al., 2016). In this study, we have generated reference epigenomes from the thyroids of tumor and adjacent normal tissue of four human adult subjects. The pathology of the sample donors includes two follicular adenomas, one goiter, and one papillary carcinoma. In total, we have 56 histone modification ChIP-seq data sets covering six histone modifications and an input DNA control, 8 DNA methylation data sets, and 8 RNA-seq data sets. The six histone modifications consist of four activating (H3K4me1, H3K4me3, H3K27ac, and H3K36me3) and two repressing (H3K9me3 and H3K27me3) marks and they coincide with the core set of histone modifications analyzed from the 127 epigenomes of various tissues and cells in the NIH Roadmap project (Roadmap Epigenomics Consortium, et al., 2015).

Defining chromatin states

Histone modifications associate with different parts of the genome. In the 6 histone modifications used, H3K4me3 have been associated to promoters, H3K27ac to promoters and active enhancers, H3K4me1 to active enhancers, H3K36me3 to transcribed gene bodies, H3K9me3 to heterochromatin, and H3K27me3 to Polycomb repressed regions (Roadmap Epigenomics Consortium, et al., 2015). Undoubtedly, the distribution of different histone modifications reveal different epigenetic signals and tools such as ChromHMM (Ernst & Kellis, 2012) and Segway (Hoffman, et al., 2012) have been developed to represent combinations of epigenetic features by partitioning the epigenome into various chromatin states. In this study, we used ChromHMM (Ernst & Kellis, 2012) to partition the epigenomes into 19 chromatin states.

ChromHMM (Ernst & Kellis, 2012), an implementation of a hidden Markov model (HMM), uses epigenetic features such as histone modifications to represent observed (or output) states and unobserved (or hidden) states to represent chromatin states. Generally, HMMs have 2 parameters: (1) emission probabilities representing the output (e.g. histone) probability of a hidden state, and (2) transition probabilities representing the probability of the next hidden state. Due to the nature of hidden states, the number of states (denoted by k) will need to be specified programmatically. In this study, we trained ChromHMM on k = 11 to 23 states. The number of hidden states used encompassed the number of states chosen by the NIH Roadmap Consortium for the anlsysis of epigenomic states across 111 cell types (Roadmap Epigenomics Consortium, et al., 2015): 15 states for 5 histone modifications, and 18 states for 6 histone modifications. Furthermore, there are 2 ways to treat the input DNA control using ChromHMM: (1) as an input feature directly in the model to help isolate regions of copy number variation and repeat associated artifacts or (2) as a control to locally adjust the input feature binarization threshold. Interestingly, when we trained 3 independent models using the same parameters for arbitrarily k = 15 states, the same model and same segmentation of chromatin states were produced. Inspecting the ChromHMM program further, we found that the randomization of the initial parameters has been seeded with a predefined integer, which will result in reproducible models. In total, we trained 26 different candidate models in order to select the final model for further analysis.

From the epigenomes, we produced a set of 26 candidate models. The task now is to select a model for further analysis. Two popular model selection methods include the Bayesian information criterion (BIC) and Akaike information criterion (AIC). These selection methods however tend to favour higher number of states which are biologically harder to distinctly interpret and does not capture sufficiently distinct interactions. According to (Hamada, Ono, Fujimaki, & Asai, 2015), using BIC for HMMs is not mathematically well founded because HMMs do not typically satisfy the regularity conditions of BIC. (Hamada, Ono, Fujimaki, & Asai, 2015) then proposed a factorized information criterion (FIC) for selecting the number of states produced by ChromHMM;

however, their result indicated more estimated chromatin states by FIC-HMM than what was selected for by the original ChromHMM analysis done in (Ernst, et al., 2011) and thus are again biologically harder to distinctly interpret. In comparison, the number of states chosen in (Roadmap Epigenomics Consortium, et al., 2015) was based on manual consideration on evaluation for the number of states which capture all key interactions between chromatin marks. Similarly, the number of chromatin states presented in (Hoffman, et al., 2013) was chosen by a manual compromise between capturing all of the potential complexity of chromatin mark combinations (which requires very large numbers of states) and generating models that are easily interpretable and maximally useful for interpreting genomic features (which requires maintaining a small number of states).

The number of states modelled in (Roadmap Epigenomics Consortium, et al., 2015), (Hoffman, et al., 2013), or (Ernst, et al., 2011) were selected by manual consideration. BIC, AIC, and FIC tend to favour higher number of states which are biologically harder to distinctly interpret. In this study, we devised a novel quantitative selection metric that will allow rapid assessment for the optimal number of states by choosing a model that is most well defined, maximizing the homogeneity of epigenetic features in chromatin states across samples (see methods). Using this novel quantitative selection metric to compute homogeneity cost, we found that the number of states chosen is similar to (Roadmap Epigenomics Consortium, et al., 2015), which has been trained on 111 primary human tissues and cell types. From the homogeneity cost (Figure 2), we found that 19 states with input as control and 20 states with input as a mark were the optimal number of states to use. For further analysis, we then proceeded with 19 states using the input as control based on (1) there were less states and (2) the Roadmap project (Roadmap Epigenomics Consortium, et al., 2015) treated input as control. Similar to the 18 state model published for 111 primary human tissues and cell types (Roadmap Epigenomics Consortium, et al., 2015), we found our model recapitulates many of the states with a few notable differences (Figure 3A): (1) we have 19 states while Roadmap has 18; (2) our model, in accordance to state enrichments described in the next section, has a repressed (state 15) and repeat (state 17) state not published in (Roadmap

Epigenomics Consortium, et al., 2015); and (3) we lack the bivalent TSS state published in (Roadmap Epigenomics Consortium, et al., 2015). Minor differences in state discrimination include having a second transcription state, but lacking a second active enhancer state; and having an extra flanking enhancer state, but lacking the weakly repressed Polycomb state.

Chromatin states correlate with genomic features

The chromatin states from the 19 epigenomic partitioning correlate with various known genomic features (Figure 3). For instance, states 1 – 4 are enriched in transcription initiation neighborhoods (Figure 3C), which indicates that states 1 – 4 are correlated with promoters. H3K36me3 associated emissions correlate with genes, introns, and exons in states 5 – 9, indicating these states being related to transcribed gene bodies. In comparison, states 9 - 12 have emission of H3K4me1 characteristic of enhancers. In state 16, the H3K4me1 and H3K27me3 emissions are indicate of a bivalent enhancer state. According to the overlap enrichment of genomic features (Figure 3B), there is a lack of gene enrichment in states 14 – 15, and 17 – 19. In state 17, there is emission for all histone marks, indicating that this state may be associated with repetitive regions such as in (Ernst, et al., 2011). In contrast, state 19 is likely an epigenetically unmarked state based on the rationale that state 19 has no emission in any of the histone marks while covering the greatest percentage of the genome (Table 2) and such a quiescent state of this magnitude is to be expected (Roadmap Epigenomics Consortium, et al., 2015). Based on a combination of histone mark emissions probabilities (Figure 3A), enrichment in genomic features (Figure 3B, Figure 3C), and comparison with published chromatin states (Roadmap Epigenomics Consortium, et al., 2015), we have labelled the states with biologically meaningful labels (Table 2). Furthermore, in an orthogonal experiment where the levels of methylation is measured, we found that the active TSS state (state 1) had, as expected, the lowest level of methylation across chromatin states (Figure 4). The chromatin state segmentations can be viewed on the UCSC Genome and Wash U Epigeniome Browsers through http://www.epigenomes.ca/data-release/ and a link provided in http://www.bcgsc.ca/data/thyroid (Figure 5).

Chromatin states stability

We do not know how much epigenetic variation exists in the population and thus sought to annotate stable and unstable states. In this study, we were interested in characterizing regions that were epigenetically consistent. The genome was divided into 15,181,508 genomic bins. Each bin is 200bp in length and represents a chromatin state. For a particular bin across different individuals, the chromatin state may be the same or it may be different. We define a bin as epigenetically consistent when the chromatin state is the same across all individuals. We find that only the promoter (state 1), transcribed (state 5), and weakly transcribed (state 7) states show consistency across the epigenomes of the normal thyroid tissue from the four individuals (Figure 6A, Figure 6B). Furthermore, state 19, the epigenetically unmarked or quiescent state covering the greatest percentage of the genome (Table 2), remained largely unchanged (Figure 6A). We also found the epigenetic consistency is reduced in the other states. The states lacking the most agreement across samples are states 4 and 17 (Figure 6C), which we labelled as, respectively, regions flanking downstream of TSS and repeats associated with artifacts.

Epigenetically marked promoters and relation with gene expression

The promoter state labelled as active TSS (state 1) was found to be the most epigenetically consistent state (Figure 6). 101,278 out of 15,181,508 genomic bins were partitioned to this state in at least one epigenome and 36.5% of the 101,278 bins were found to be epigenetically consistent across all four epigenomes. For any given epigenome, a bin partitioned as state 1 had an average of 57% probability of also being partitioned as state 1 in three, 19% in two, 13% in one, and 11% in no of the other epigenomes (Figure 6C).

We next associated bins partitioned as state 1 to genes if the bin is within a gene's promoter (TSS +/- 1kbp). A majority of state 1 bins (77.4%) were found within protein coding gene promoters (Figure 7A). This value increased to 91.2% when we consider only bins consistently partitioned as state 1 across all four epigenomes. 13,175 out of 20,154 known protein coding genes were associated to bins partitioned as state 1 in at

least one epigenome and 10,460 to bins partitioned as state 1 across all four epigenomes (Figure 7B). Collectively, state 1 bins capture the promoters of 65.4% known protein coding genes in the thyroid and this value drops to 51.9% when we consider only state 1 bins that are epigenetically consistent across the four epigenomes. It is striking that in a relatively simple tissue such as the thyroid, whose main role is to predominately produce thyroid hormone, approximately half of the known protein coding genes have epigenetically active promotes in all four samples. In comparison, we find roughly 100 to 300 genes epigenetically active in only one epigenome and when we annotate these genes using Metascape, we find functions related to matrix organization and immune response (Figure 8). One possible explanation of these functions may be due to the nature of the samples. From Figure 7, we find CEMT_40 and CEMT_42 associated with matrix organization. Looking at the meta data (Table 1), both these samples come from donors with follicular adenoma – a condition whereby a benign tumor is encapsulated by a thin fibrous capsule (McHenry & Phitayakorn, 2011) – which may explain the effect on matrix organization in these samples. With regards to the enrichment of immune response relate functions in CEMT_44 and CEMT_86, there might have been more B cells in circulation during sample collection.

We next grouped the genes by the epigenetic consistency of state 1 in gene promoters and compared the level of gene expression. A gene is epigenetically active if the promoter region is characterized by state 1 in at least one epigenome. We hypothesized that genes that are epigenetically active across all four samples will have higher expression than genes that are not epigenetically active in any samples. When we grouped expression by the number of epigenetically active promoters shared across samples, we found that the expression tends to be higher in genes partitioned as epigenetically active in more samples (Figure 7C, Figure 7D). Specifically, expression is on average 9.7-fold higher in genes characterized as epigenetically active than genes not characterized as epigenetically active in any samples. Furthermore, expression is on average 4.4-fold higher in genes that are epigenetically active across all four samples than genes that are epigenetically active in only one sample. Similarly, when we grouped genes into different brackets of expression, we found that genes with high

expression tend to be epigenetically active in all samples (Figure 7E). In one sample, 90.9% of genes with expression between 100 – 1,000 TPM is epigenetically active in all samples and this proportion drops to 44.3% for genes with expression between 1 – 10 TPM and 7.9% for genes with expression between 0.1 – 1 TPM (Figure 7E). Interestingly, we also find some genes, such as *MTRNR2L12,* have high expression despite not being determined as epigenetically active in any sample. In the findings, *MTRNR2L12* is within the top 12 most highly expressed genes across the four samples (Table 4); epigenetically marked as heterochromatin, repressed, and quiescent; and the closest genomic bin marked as active TSS is in one sample located more than 20kbp away. In the literature, MTRNR2L12 has been suggested to be a candidate blood marker of early Alzheimer's disease-like dementia in adults with Down syndrome (Bik-Multanowski, Pietrzyk, & Midro, 2015), but there were no mentions of thyroid or epigenetic regulation of MTRNR2L12. One possible explanation of the high expression despite not being determined as epigenetically active across the four samples is that such genes are constitutively expressed genes.

Enhancers

Chromatin states characterized as enhancers (states 8 – 11) were less consistent than states characterized as promoters (Figure 6). Nevertheless, enhancer type chromatin states include genic enhancers (state 8 and 9), active enhancers (state 10), and weak enhancers (state 11). Sequence analysis of these genomic DNA epigenetically consistent at enhancer type chromatin states indicate that the NF1 response element (CYTGGCABNSTGCCAR) was the most overrepresented sequence motif in enhancer states 8, 10, and 11. Other transcription factors response elements common across enhancer states 8, 10, and 11 are TLX (CTGGCAGSCTGCCA), PAX8 (GTCATGCHTGRCTGS), and PAX5 (GCAGCCAAGCRTGACH). In the literature, PAX8 was found to be involved with thyroid organogenesis and the maintenance of the thyroid differentiated state (Trueba, et al., 2005). PAX8 may also have diagnostic utility in thyroid epithelial neoplasms for it was strongly expressed in papillary carcinomas, follicular adenomas, follicular carcinomas, and 79% of anaplastic carcinomas (Nonaka,

Tang, Chiriboga, Rivera, & Ghossein, 2008). The top 3 motifs of each enhancer chromatin state are shown in (
Table **3**).

Transcript abundance

With regards to estimating transcript abundances, we found that the most highly expressed transcripts representing 95% of the RNA-seq reads are made of at least 7,194 genes and 10,000 genes account for an average of 98% of transcript reads detected (Figure 9). Across the 4 samples, the top 25 genes – accounting for 19% of transcripts – contains 42 unique genes and 10 of these genes are consistent across the 4 samples (Table 4).

In 3 out of 4 cases, the top gene – accounting for about 2.4% of transcripts – is the gene encoding the thyroid hormone precursor protein, thyroglobulin (TG). In CEMT_40, where *TG* is not the most abundant gene, we find genes encoding ribosomal proteins, Eukaryotic Translation Elongation Factor 1 Alpha 1 (EEF1A1), and Metallothionein 1G (MT1G) as being more abundant (Table 4). Across the 4 samples, *EEF1A1* is ranked within the top 4 and *MT1G* within the top 42 most abundant protein coding gene across the 4 samples. *EEF1A1,* encoding a protein which plays a key role in protein translation by interacting with aminoacyl-tRNA to bring it to the acceptor site of the ribosome in the first step of the elongation cycle, was found to have increased expression in *Solea senegalensi* upon T4 hormone treatment (Infante, Asensio, Cañavate, & Manchado, 2008), which may suggest *EEF1A1* is a thyroid sensitive gene and may be highly relevant to thyroid function. In contrast, functions for MT1G are related to metal-binding property, including detoxification of heavy metals, donation of zinc/copper to certain enzymes and transcription factors, and protection against oxidative stress (Fu, et al., 2013). Furthermore, it was suggested that MT1G acts as a tumor suppressor of thyroid carcinogenesis (Fu, et al., 2013). Overall, these genes appear to be related to metabolism – which is not unexpected in the thyroid – by either being directly involved in the synthesis of proteins or as protection against oxidative stress as a result of metabolic process.

In the top 25 most abundant genes across the 4 samples, most were partitioned as active TSS state 1 within the gene promoter across the same bin in the 4 samples (Table 4). Exceptions include *CD74*, *CLU*, *HBA2*, *HBB*, *MTRNR2L12*, and *RPS24* (highlighted with grey in Table 4). Although *CD74*, *CLU*, and *RPS24* were not partitioned as active TSS (state 1) in the same bin across the 4 epigenomes, we find these genes were nonetheless inconsistently marked as active TSS across different bins within the gene promoters for at least 3 epigenomes (Figure 10). With regards to the other 3 genes, *HBA2* was partitioned with repressed Polycomb (state 18) and bivalent enhancer (state 16) states; and *HBB* and *MTRNR2L12* were partitioned with heterochromatin (state 14) and quiescent (state 19) states. Hemoglobin Subunit Alpha 2 (HBA2) and Hemoglobin Subunit Beta (HBB) are proteins associated with blood. One rationale for their high expression, despite not being epigenetically marked with active chromatin states, is that the expression may be a consequence of impurities from blood during thyroid tissue collection. In contrast, not a lot is known about MT-RNR2-like 12 (MTRNR2L12) other than its alias as Humanin-Like 12 (HN12) and the one study suggesting potential value for *MTRNR2L12* to be used as a blood biomarker of early dementia in individuals with Down syndrome (Bik-Multanowski, Pietrzyk, & Midro, 2015).

Epigenetically active and consistently expressed genes in the thyroid

To further define the thyroid, we next identified a set of genes that were likely highly relevant to thyroid function. These genes are ideally epigenetically active and consistently expressed, as epigenetically active genes are presumed poised for transcription and consistently expressed genes with low expression variance across samples are considered to be under stringent transcriptional control. We consider a gene as epigenetically active if a bin within the gene promoter (TSS +/- 1kbp) is partitioned as state 1. Previously, we found 13,175 genes to be epigenetically active in at least one sample and 10,460 genes were found to be epigenetically active across all four samples (Figure 7B). We considered a gene as consistently expressed if (1) it is within the intersection of the top 2,000 most highly expressed gene in each sample and

(2) it is in the set of 2,000 genes with the lowest variance across the normal samples. Overall, the 2,000 most highly expressed genes have a minimum expression of 29 TPM and accounted for an average of 76% of protein coding RNA-seq transcripts. Within the top 2,000 genes across the four samples, there was a total of 3,024 genes and the intersection defined 1,183 genes across the four samples. Intersecting the set of 10,460 genes that are epigenetically active across all four samples, 1,183 genes that have high expression, and 2,000 genes with low variance, we arrived at a set of 137 genes (Figure 11A). Examining this set of genes using Metascape (Tripathi, et al., 2015), we predominantly find general processes such as functions related to metabolic processes, protein folding, transport, and secretion (Figure 11B). The top 3 Gene Ontology (GO) terms are RNA localization (GO:0006403), protein folding (GO:0006457), and negative regulation of cell death (GO:0060548).

To further prioritize the list of 137 genes, we filtered out genes expressed (FPKM >= 10) in 52 non-thyroid tissues using expression values from the GTEx project (Figure 12). Overall, we are left with 18 genes (Table 5). When we perform a gene set enrichment (Tripathi, et al., 2015) analysis on this set of 18 genes, no terms were found enriched. In Table 5, we present the GO annotation of individual genes. *ETFB*, *NT5C2*, *SNF8*, *SORD,* and *TOR1AIP1* appear to be terms related to metabolism, *N4BP2L2* to blood, and *TPD52* to the immune system. In the literature, spatacsin*,* encoded by *SPG11*, was identified to play critical roles in autophagic lysosome reformation, a pathway that generates new lysosomes (Chang, Lee, & Blackstone, 2014) and TPD52 has been predicted to regulate endolysosomal trafficking in secretory cell types (Byrne, Frost, Chen, & Bright, 2014). In the thyroid gland, thyroid hormone is produced (from the breakdown of biomolecules involving lysosomes) and secreted (playing important roles in secretory processes). Thus it is not unexpected for *SPG11* and *TPD52* to be of importance to normal thyroid function. With regards to DEPTOR, a mTOR inhibitor, it was suggested as having activity in controlling several molecular pathways, such as apoptosis, cell survival, autophagy, and endoplasmic reticulum homeostasis, and it was suggested to play a role as a transcriptional activator (Catena & Fanciulli, 2017). DEPTOR may also play a role in the transcriptional activation of thyroid responsive

genes. According to a review by (Claudel, Zollner, Wagner, & Trauner, 2011), FXR1 belongs to the nuclear receptor superfamily of transcription factors and can bind DNA as a heterodimer with retinoid X receptor (RXR) alpha. Similarly, thyroid hormone receptors binding with T3 can also often heterodimerize with RXR (Panicker, 2011). Taken together, we suggest the binding of FXR1 with RXR could influence transcription of thyroid responsive genes. With regards to *PMF1* (involved in polyamine homeostasis (Alvarez-Mugica, et al., 2013)), *H2AFY* (encoding a histone H2A variant (Jufvas, Stralfors, & Vener, 2011)), *NSMCE1, SCAF11* (involved in RNAPII elongation (Rebehmed, Revy, Faure, De Villartay, & Callebaut, 2014)), *TCTN1* (involved in embryonic development and growth (Wang, et al., 2015)), *TPGS2*, *VEZT* (encoding an adherens junction transmembrane protein (Sousa, et al., 2004)), and *WBSCR22* (involved in ribosome small subunit biosynthesis (Ounap, Kasper, Kurg, & Kurg, 2013)), we did not find any published studies linking these genes with the thyroid, which may suggest potential significance of these genes in the thyroid.

Chromatin state defined by both H3K9me3 and H3K27me3

In state 15 (labelled as “repressed”), we find emission of H3K9me3 and H3K27me3 (Figure 3A). In the literature, there is limited knowledge of regions containing both H3K9me3 and H3K27me3. Studies have suggested there may be a functional role of H3K9 and H3K27 methylation in coordinating and ensuring progressive lineage restriction during the enactment of the oligodendrocyte progenitor differentiation program (Liu, et al., 2015) or in a cooperative mechanism in maintaining silencing whereby H3K27me3-bound PRC2 stabilizes H3K9me3-anchored HP1A (Boros, Arnoult, Stroobant, Collet, & Decottignies, 2014). In another study, it was suggested the antibody used to enrich H3K27me3 has off target enrichment for H3K9me3 (Peach, Rudomin, Udeshi, Carr, & Jaffe, 2012), which would make H3K9me3 and H3K27me3 an artificial state. Overall, there remains a question whether state 15 is a functional chromatin state. According to our results (Figure 6A), the stability of this chromatin state across 4 epigenomes is low and that out of all bins partitioned as state 15, only 3.0% are consistent across 4 epigenomes. Similarly, a bin partitioned as state 15 has a 9% probability of finding the same state in the same bin across 3 other epigenomes (Figure

6C). The lack of conservation of state 15 between samples further questions whether this has any real biological function or whether it arises as a random chromatin state. In terms of transition probabilities, there exists probability for transitions to occur from state 14 (“heterochromatin”) to state 15 and from state 15 to itself and to states 14 and 19 (“quiescent”) (Figure 3A). Taken together, this suggests regions containing both H3K9me3 and H3K27me3 may be an intermediate state from heterochromatin to quiescent states.

In our other analyses, when we ran a sequence analysis of genomic DNA epigenetically consistent at state 15, we found that ZNF692 response element (GTGGGCCCCA) was the most overrepresented sequence motif. When we computed the overlap enrichment of repeat regions, we found enrichment in LTR, LINEs, and SINEs far (>10kbp) from protein coding genes (Figure 13). When we applied the 19 state model to 15 colon epigenomes, we find that the conservation in state 15 drops to 0.2% across the 15 epigenomes, while active TSS state 1 remains at 12.4% (Figure 14). When we correlated the emission profiles of the 19 state model with Roadmap’s 18 state model, we find that state 15 correlates with Roadmap’s Quiescent, Heterochromatin, and ZNF states (Figure 15). If state 15 was an intermediate between heterochromatin state 14 and quiescent state 19, then it’s feasible for there to be correlation of state 15 with these Roadmap states. In other words, if state 15 was an intermediate between the heterochromatin and quiescent states, we expect state 15 to have characteristics of both the heterochromatin and quiescent states, resulting to the higher correlation of state 15 with Roadmap’s Quiescent and Heterochromatin states.

Reliability of a single reference

References should be representative and we should therefore expect the epigenetic features from the reference epigenomes to be consistent across different individuals. In this study, we compared the consistency of chromatin state annotations across the epigenomes from the thyroid tissue of different individuals. We characterized normal thyroid epigenomes into 19 chromatin states and found that some states – such as the promoter and transcription states – tend to be more epigenetically consistent and stable than others (Figure 6). Similar to the high consistencies of our active TSS and transcription states, (Lee & Park, 2016) predicted chromatin states from nucleotide frequency profiles of K562 or GM12878 and found that their Active Promoter and Transcribed chromatin states highly coincided with the annotations of other cell lines. Furthermore, the quiescent state remained largely unchanged across epigenomes, while every other state showed inconsistencies across the 4 epigenomes. In comparison with other chromatin state models, the same states were not always reproduced. For instance, state 15, labelled as repressed, was not observed in the Roadmap project (Roadmap Epigenomics Consortium, et al., 2015) but was observed in (Pellacani, et al., 2016). When we trained a new ChromHMM model on 15 colon samples, we found that (1) the optimal number of states differs between the thyroid (k=19) and the colon (k=16) (Figure 16) and (2) the 19 states produced from the colon samples differ from the 19 states produced from the thyroid samples (**Figure 17**). To account that the inconsistencies were not due to the partitioning of chromatin state by ChromHMM, we show similar findings as Figure 6 when peaks were called directly using FindER and MACS2 (Figure 18). Nonetheless, we found regions of consistency across the 4 epigenomes, but to isolate individual differences (Figure 8), we suggest the more biological replicates you have, the more relevant the consensus reference will be.

Conclusion

We characterized the normal thyroid epigenome into 19 chromatin states and compared the epigenetic features across four different individuals. We found that epigenetic features characterizing promoters and transcription elongation tends to be more consistent whereas every other feature tends to be more variable across the four individuals. Consistent with expectations, we also found that genes epigenetically active across all epigenomes tend to have higher expression than those that are not consistently epigenetically active. Furthermore, we identified a set of 18 genes epigenetically active and consistently expressed genes in the thyroid. Overall, we conclude the epigenomes presented in the paper and available at http://www.bcgsc.ca/data/thyroid represent a valuable resource to gain a deeper understanding of the molecular biology of thyroid function and provide contextual epigenetic information and integration within future studies.

Tables

Table 1. Thyroid donor information.

Sample ID	Donor age	Donor sex	Donor health status
CEMT_40	67	Female	Follicular Adenoma
CEMT_42	46	Female	Follicular Adenoma
CEMT_44	55	Male	Goiter
CEMT_86	44	Female	Papillary Carcinoma

Table 2. The 19-state model: state labels and average genomic coverage. Genomic coverage values were averaged across 4 normal thyroids samples.

State	Color (rgb)	Label	Average genomic coverage (%)
1	255,0,0	Active TSS	0.43
2	255,69,0	TSS	0.22
3	255,69,0	TSS Flanking	0.32
4	255,69,0	TSS Flanking downstream	0.14
5	0,128,0	Transcription 1	2.82
6	0,128,0	Transcription 2	2.40
7	0,100,0	Weak transcription	9.42
8	194,225,5	Genic enhancer 1	0.75
9	194,225,5	Genic enhancer 2	0.65
10	255,195,77	Active enhancer	1.91
11	255,255,0	Weak enhancer	1.80
12	255,255,0	Flanking enhancer	7.68
13	102,205,170	ZNF	0.31
14	138,145,208	Heterochromatin	5.32
15	138,145,208	Repressed	2.06
16	189,183,107	Bivalent enhancer	0.28
17	192,192,192	Repeats associated with artifacts	2.58
18	128,128,128	Repressed PolyComb	6.33
19	255,255,255	Quiescent	54.58

Table 3. Motifs significantly enriched in genomic DNA epigenetically consistent at enhancers type chromatin states: 8 & 9 = genic enhancers, 10 = active enhancer, and 11 weak enhancers. Motif enrichment was performed using HOMER software and the top 3 motifs for each enhancer type chromatin state is given (Benjamini corrected p-values < 0.03). State 9 has enrichment in only 1 motif.

State	TF	DNA binding domain	Consensus	log(p-value)
8	NF1	CTF	CYTGGCABNSTGCCAR	-29.3
8	Tlx?	NR	CTGGCAGSCTGCCA	-16.4
8	Pax8	Paired,Homeobox	GTCATGCHTGRCTGS	-11.1
9	Mef2c	MADS	DCYAAAAATAGM	-9.6
10	NF1	CTF	CYTGGCABNSTGCCAR	-114.5
10	Fosl2	bZIP	NATGASTCABNN	-71.5
10	Tlx?	NR	CTGGCAGSCTGCCA	-60.3
11	NF1	CTF	CYTGGCABNSTGCCAR	-293.6
11	Tlx?	NR	CTGGCAGSCTGCCA	-114.9
11	PAX6	Paired,Homeobox	NGTGTTCAVTSAAGCGKAAA	-84.0

Table 4. The 25 most abundant transcripts for protein coding genes in each sample. In total, there are 42 unique genes. Text highlighted in grey represent genes (n=6) not determined to be epigenetically active (i.e. labelled as active TSS state 1 in the same bin) across 4 samples. Non-highlighted genes (n=36) are considered epigenetically active.

Rank	mean	st. dev.	Gene			
	(%)	(%)	CEMT_40	CEMT_42	CEMT_44	CEMT_86
1	2.4	0.4	RPS29	TG	TG	TG
2	4.0	0.5	RPL39	MTRNR2L12	EEF1A1	EEF1A1
3	5.2	0.8	RPS27	EEF1A1	RPS27	B2M
4	6.4	1.1	EEF1A1	RPS27	MT1G	MTRNR2L12
5	7.4	1.6	MT1G	RPS29	B2M	RPS27
6	8.4	2.1	RPL41	TPT1	GPX3	RPL41
7	9.3	2.5	RPS3A	RPL41	TPT1	GPX3
8	10.1	2.9	TG	RPS3A	MTRNR2L12	CLU
9	10.9	3.3	TPT1	B2M	RPS29	ACTB
10	11.6	3.7	RPS18	RPL39	RPL41	TPT1
11	12.3	4.0	RPS21	RPL26	TPO	RPL10
12	12.9	4.4	MTRNR2L12	GPX3	RPS3A	RPS3A
13	13.5	4.8	RPL34	RPL27A	RPS24	HBA2
14	14.1	5.0	RPL27A	RPS21	RPL39	UBC
15	14.6	5.3	RPL26	RPL34	RPL37A	EMP1
16	15.1	5.5	B2M	TPO	RPL10	ACTG1
17	15.7	5.8	RPS15A	RPS24	RPL27A	CD74
18	16.1	5.9	RPL24	RPL37A	RPL26	TPO
19	16.6	6.1	RPS6	RPL17	ACTB	RPS18
20	17.1	6.3	RPS24	RPS15A	GNAS	RPS29
21	17.5	6.4	HBB	RPS18	CLU	RPL9
22	17.9	6.5	RPS27A	RPL10	RPL34	FOSB
23	18.4	6.7	RPL27	RPL18A	RPL13	RPL37A
24	18.8	6.8	RPS12	ACTB	RPL17	RPL34
25	19.2	7.0	RPL17	RPL24	ACTG1	FTL

Table 5. GO Biological Process annotation of 18 acitvely transcribed and consistently expressed genes in the thyroid that do not have high expression in 52 non-thyroid GTEx tissues. GO annotations were obtained from Metascape.

GENE SYMBOL	DESCRIPTION	GO BIOLOGICAL PROCESS
DEPTOR	DEP domain containing MTOR-interacting protein	GO:0045792 negative regulation of cell size; GO:0032007 negative regulation of TOR signaling; GO:0006469 negative regulation of protein kinase activity
ETFB	electron transfer flavoprotein beta subunit	GO:0033539 fatty acid beta-oxidation using acyl-CoA dehydrogenase; GO:0006635 fatty acid beta-oxidation; GO:0009062 fatty acid catabolic process
FXR1	FMR1 autosomal homolog 1	GO:2000637 positive regulation of gene silencing by miRNA; GO:0060148 positive regulation of posttranscriptional gene silencing; GO:0060964 regulation of gene silencing by miRNA
H2AFY	H2A histone family member Y	GO:0034184 positive regulation of maintenance of mitotic sister chromatid cohesion; GO:0061086 negative regulation of histone H3-K27 methylation; GO:0051572 negative regulation of histone H3-K4 methylation
N4BP2L2	NEDD4 binding protein 2 like 2	GO:1902037 negative regulation of hematopoietic stem cell differentiation; GO:1902035 positive regulation of hematopoietic stem cell proliferation; GO:1901533 negative regulation of hematopoietic progenitor cell differentiation

NSMCE1	NSE1 homolog, SMC5-SMC6 complex component	GO:2001022 positive regulation of response to DNA damage stimulus; GO:0006301 postreplication repair; GO:0016925 protein sumoylation
NT5C2	5'-nucleotidase, cytosolic II	GO:0046085 adenosine metabolic process; GO:0006195 purine nucleotide catabolic process; GO:0046040 IMP metabolic process
PMF1	polyamine modulated factor 1	GO:0007062 sister chromatid cohesion; GO:0000819 sister chromatid segregation; GO:0098813 nuclear chromosome segregation
SCAF11	SR-related CTD associated factor 11	GO:0000245 spliceosomal complex assembly; GO:0000398 mRNA splicing, via spliceosome; GO:0000377 RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
SNF8	SNF8, ESCRT-II complex subunit	GO:1903772 regulation of viral budding via host ESCRT complex; GO:0010797 regulation of multivesicular body size involved in endosome transport; GO:0043328 protein targeting to vacuole involved in ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway
SORD	sorbitol dehydrogenase	GO:0006062 sorbitol catabolic process; GO:0051160 L-xylitol catabolic process; GO:0019640 glucuronate catabolic process to xylulose 5-phosphate

SPG11	spastic paraplegia 11 (autosomal recessive)	GO:0048675 axon extension; GO:0008088 axo-dendritic transport; GO:1990138 neuron projection extension
TCTN1	tectonic family member 1	GO:0021956 central nervous system interneuron axonogenesis; GO:0021523 somatic motor neuron differentiation; GO:0021955 central nervous system neuron axonogenesis
TOR1AIP1	torsin 1A interacting protein 1	GO:0071763 nuclear membrane organization; GO:0032781 positive regulation of ATPase activity; GO:0043462 regulation of ATPase activity
TPD52	tumor protein D52	GO:0030183 B cell differentiation; GO:0030098 lymphocyte differentiation; GO:0042113 B cell activation
TPGS2	tubulin polyglutamylase complex subunit 2	
VEZT	vezatin, adherens junctions transmembrane protein	GO:0016337 single organismal cell-cell adhesion; GO:0098602 single organism cell adhesion; GO:0098609 cell-cell adhesion
WBSCR22	Williams-Beuren syndrome chromosome region 22	GO:0031167 rRNA methylation; GO:0000154 rRNA modification; GO:0001510 RNA methylation

Figures

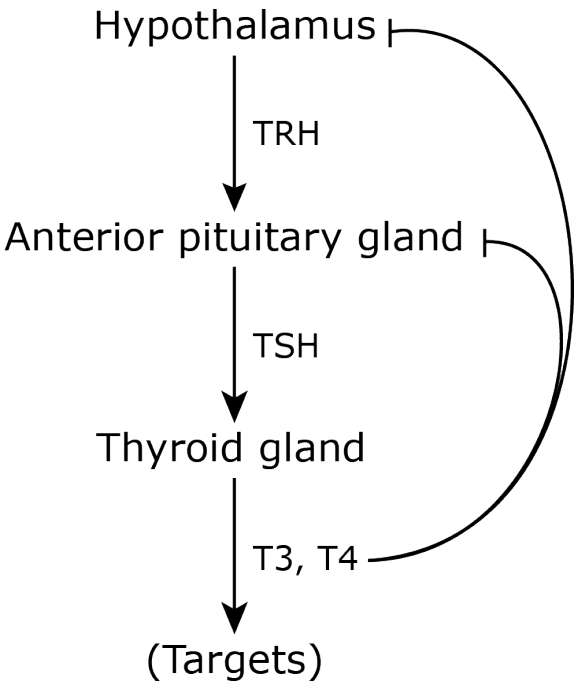


Figure 1. Thyroid hormone regulation showing the direction of stimulation (normal arrow) and inhibition (blunt arrow).

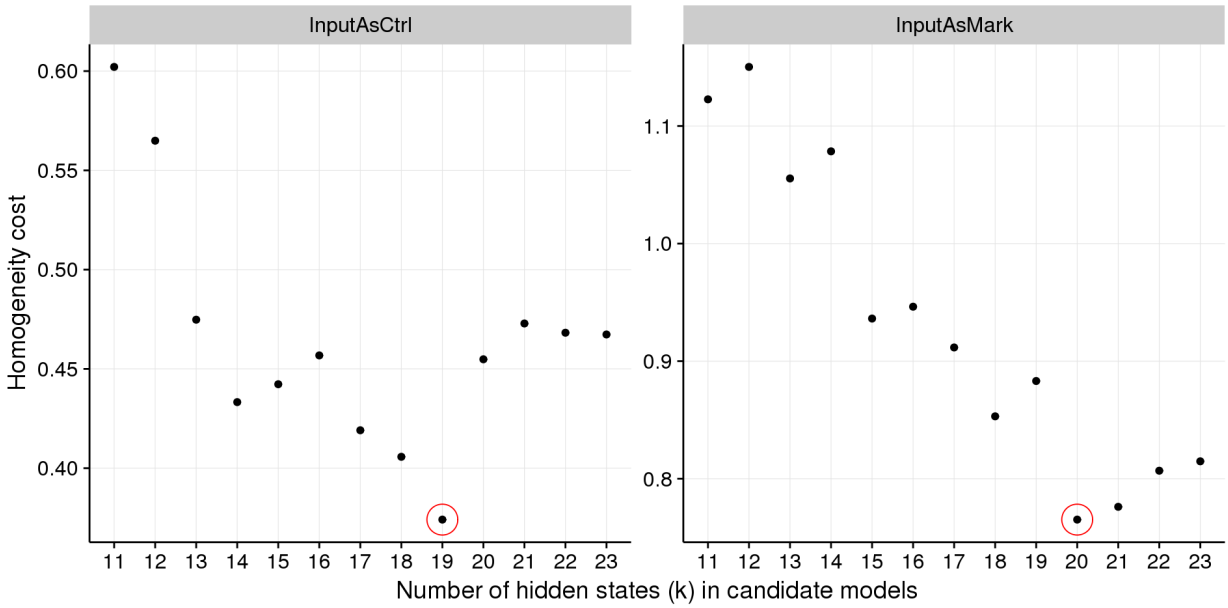


Figure 2. Plots showing the homogeneity cost used for model selection. Formulation for the homogeneity cost is presented in the methods section. Scores were computed for 26 ChromHMM generated candidate models. The number of hidden states ranged from k = 11 – 23 states. Input was treated as a control (left) and as a mark (right). 19 states with input as a control and 20 states with input as a mark produced the lowest models with the lowest homogeneity cost. 19 states with input as control was chosen for the model to use for further analysis.

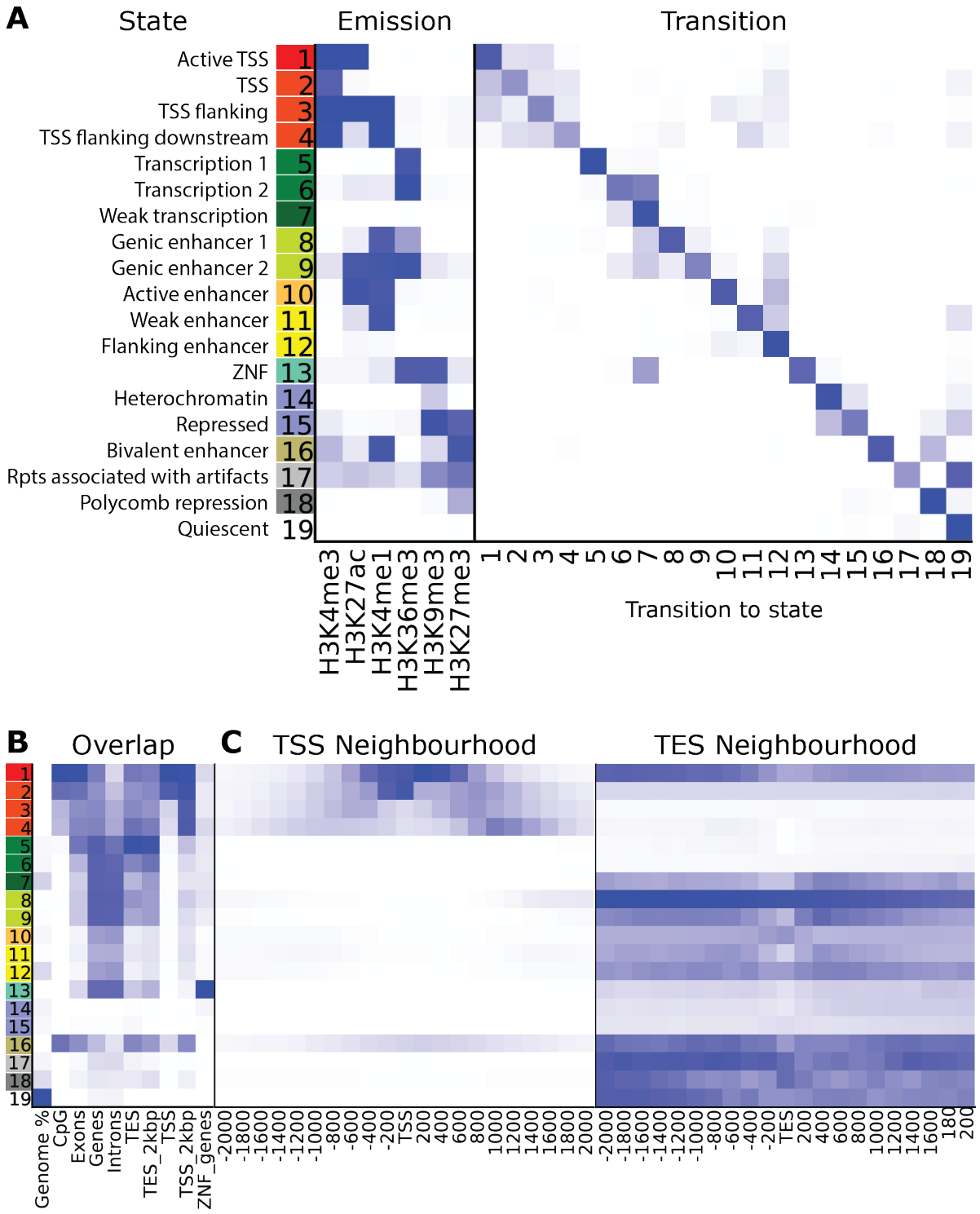


Figure 3. 19-state model with input as control. Chromatin states were defined using the ChromHMM software. The figure shows: (A) chromatin state definitions, histone mark probabilities, transition probabilities, (B) CEMT_44 genomic feature enrichments, and (C) CEMT_44 neighborhood enrichments around RefSeq TSSs and TESs. Average genomic coverages are given in Table 2.

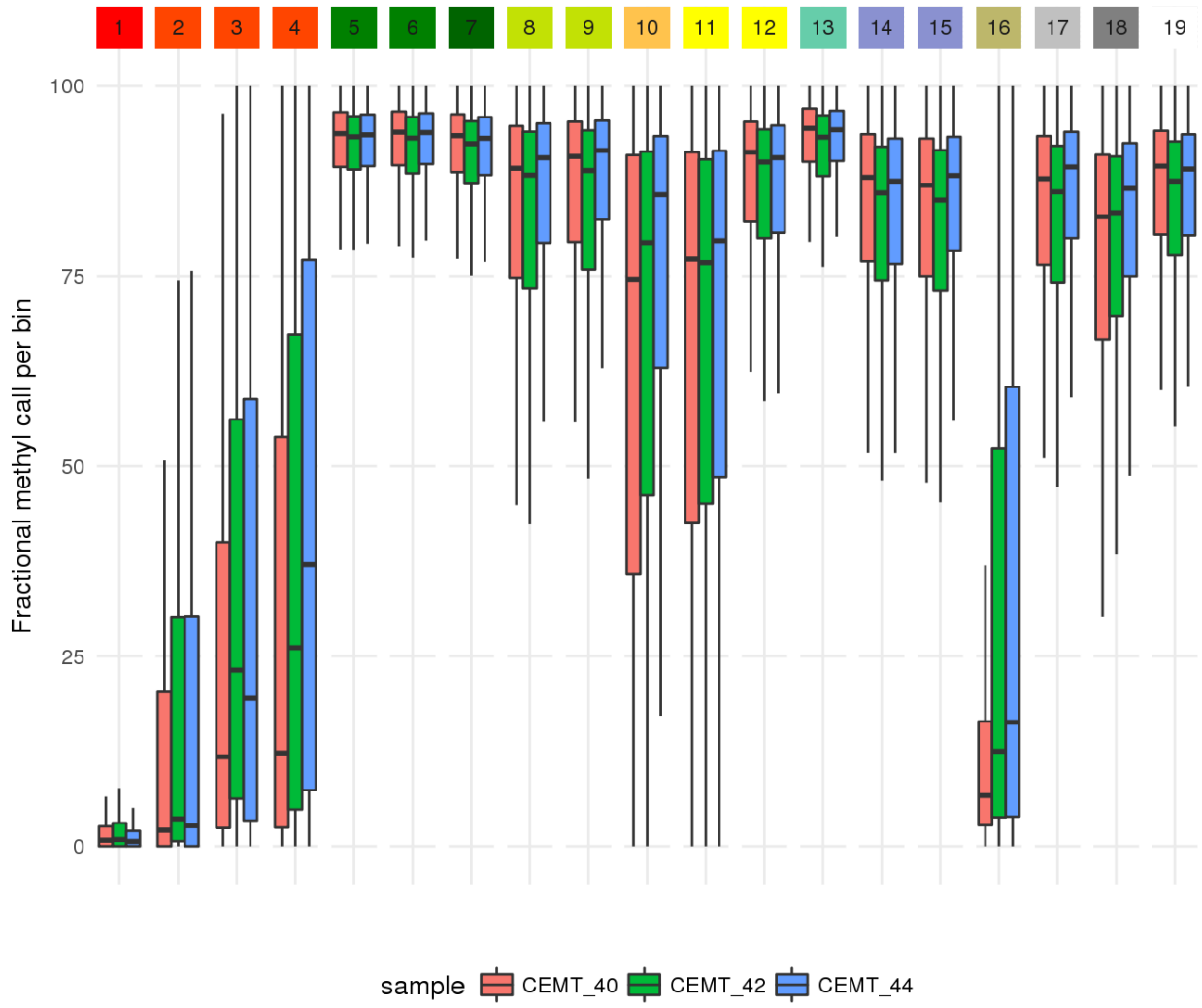


Figure 4 Boxplot showing the methylation levels across chromatin states. Fractional methylation calls were computed based on the $\frac{number\ of\ CpG\ reads}{total\ number\ of\ CpG\ reads}$ for each genomic bin. Values were summarized for each normal sample to which bisulfite-seq data was available at the time of analysis.

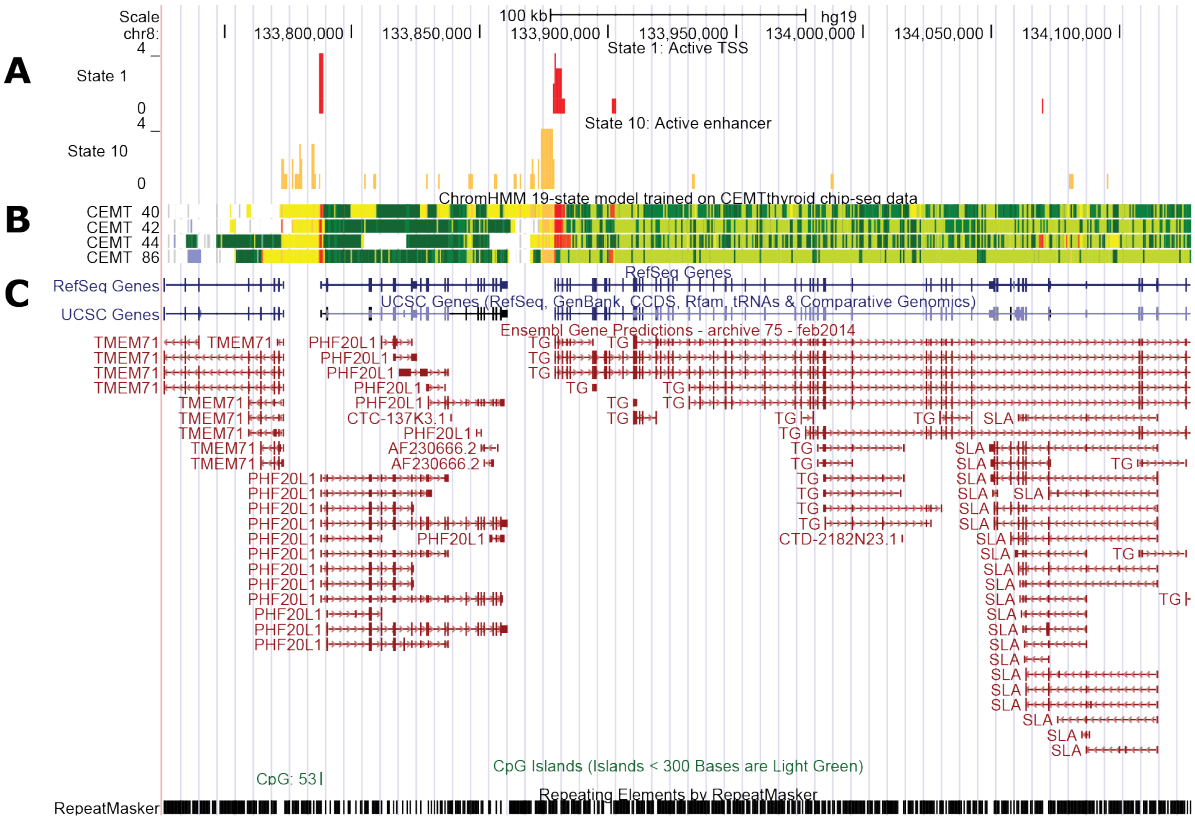


Figure 5. Screenshot of the UCSC Genome Browser showing tracks for the 19-state model around the thyroglobulin gene. These tracks can be viewed on the UCSC Genome Browser through a link provided in http://www.bcgsc.ca/data/thyroid. (A) The consistency of chromatin states across 4 epigenomes. We show the tracks for states 1 (active TSS) and 10 (active enhancer). The tracks for the remaining 17 states are hidden from view. (B) The overview of ChromHMM state segmentations for each sample. (C) Predefined tracks for gene annotations from RefSeq, UCSC, and Ensembl; CpG islands; and repeat elements by RepeatMasker.

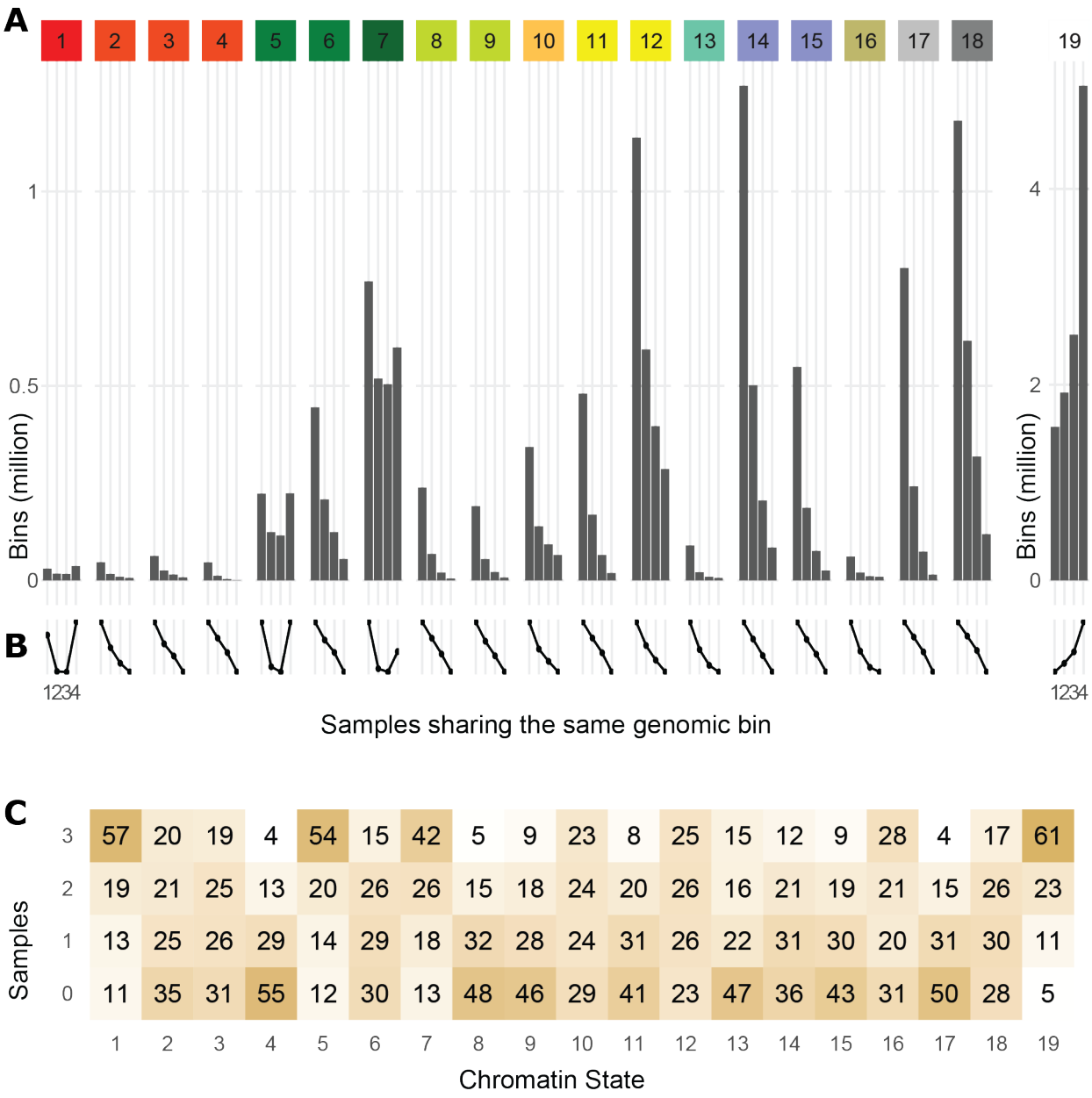


Figure 6. Overview of epigenetic consistency across 4 thyroid epigenomes. The genome was divided into 15,181,508 bins. Each bin is 200bp in length and is marked by a chromatin state. For a particular bin across different individuals, the chromatin state may be the same or it may be different. If a bin was partitioned as state 1 consistently across 4 samples, then the bin count for state 1 at x = 4 is incremented. If the states for a bin across 4 samples were {1, 1, 2, 1}, then the bin counts for state 1 at x = 3 and state 2 at x = 1 is incremented. We define a bin as epigenetically consistent when the chromatin state is the same across all individuals. (A) Histogram showing the number of genomic bins sharing the same state across 4 epigenomes. (B) Values from (A) scaled to 0 and 1 showing that states 1, 5, and 7 tends to more epigenetically consistent than every other state excluding quiescent state 19. (C) Heat map showing the average probability of finding a bin partitioned to the same chromatin state in 0, 1, 2, or 3 other epigenomes

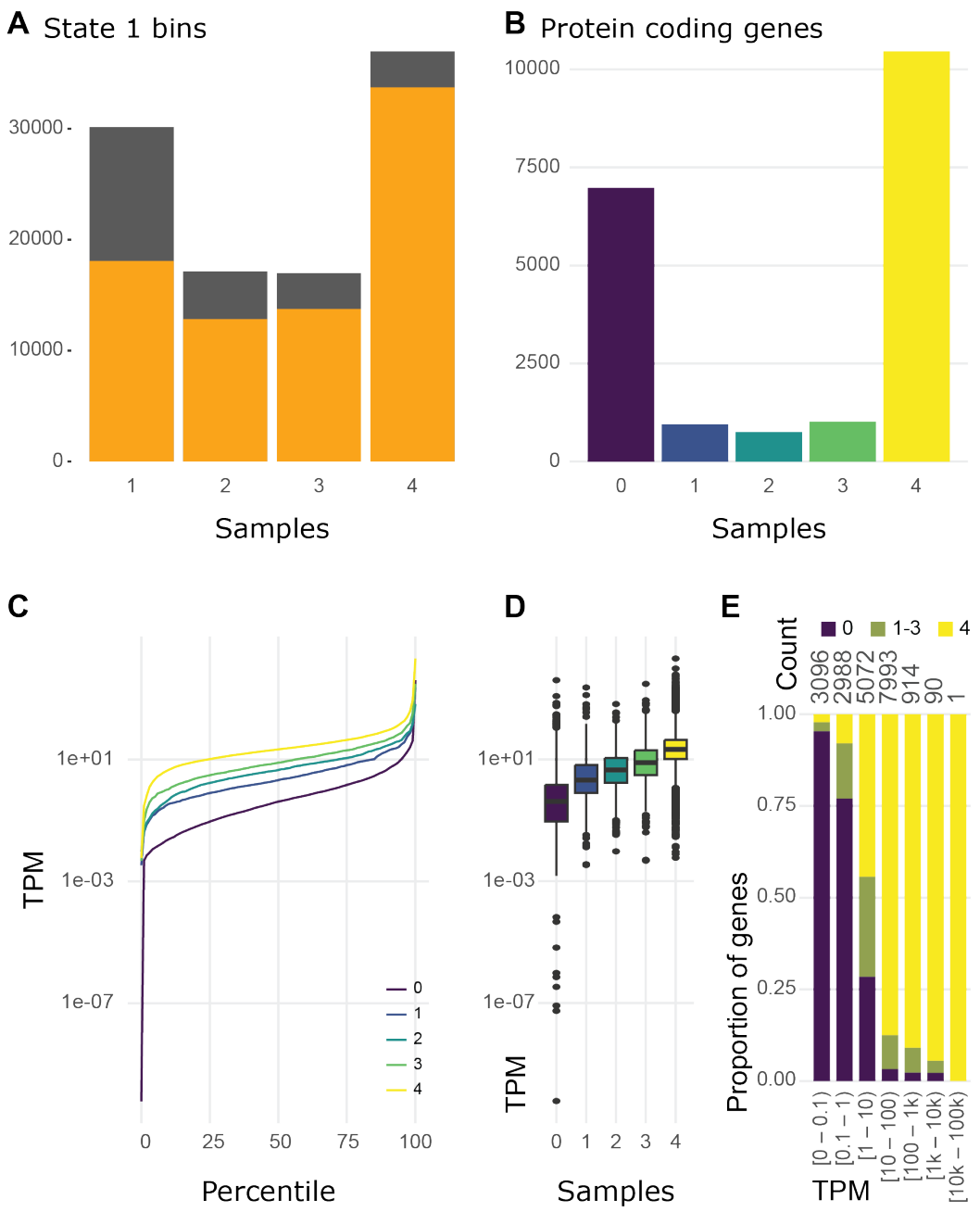


Figure 7. Association of chromatin state 1 "Active TSS" with protein coding genes. (A) Histogram showing the number of genomic bins partitioned to state 1 in 1, 2, 3, or 4 epigenomes. Orange represents state 1 bins located within promoters (TSS +/- 1kbp) of known protein coding genes. (B) Histogram showing the number of protein coding genes partitioned as state 1 across the 4 epigenomes; values are 6979, 947, 754, 1014, 10460. (C) Plot showing the percentile of expression (log10-scaled, values from CEMT_44) in the set of genes epigenetically active in 0, 1, 2, 3, and 4 epigenomes. Genes with 0 expression were removed. (D) Expression (log10-scaled, values from CEMT_44) across genes that are epigenetically active in 0, 1, 2, 3, and 4 samples. Genes with 0 expression were removed. (E) Proportion of genes in different brackets of expression (values from CEMT_44). Total number of genes in each bracket is shown on top. Color represents the number of samples sharing the same genomic bin.

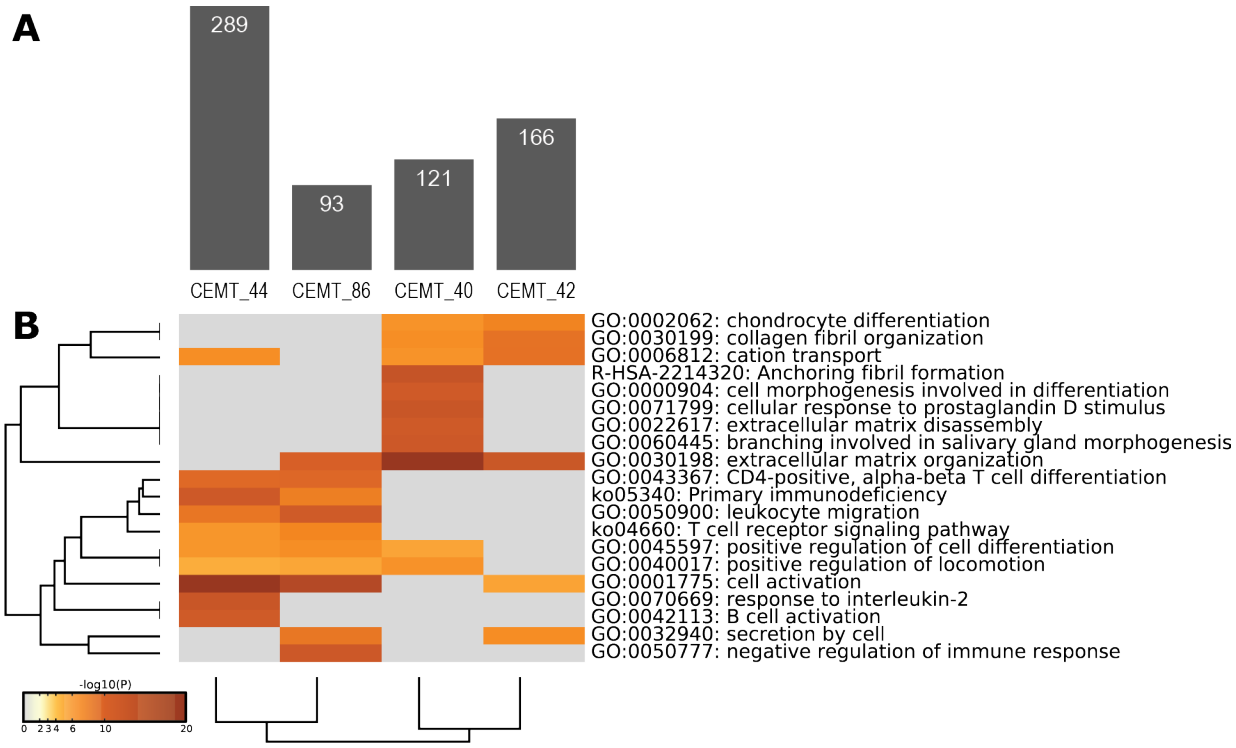


Figure 8. Genes epigenetically active in only 1 sample. (A) Gene counts. (B) Probability of metascape gene set enrichments.

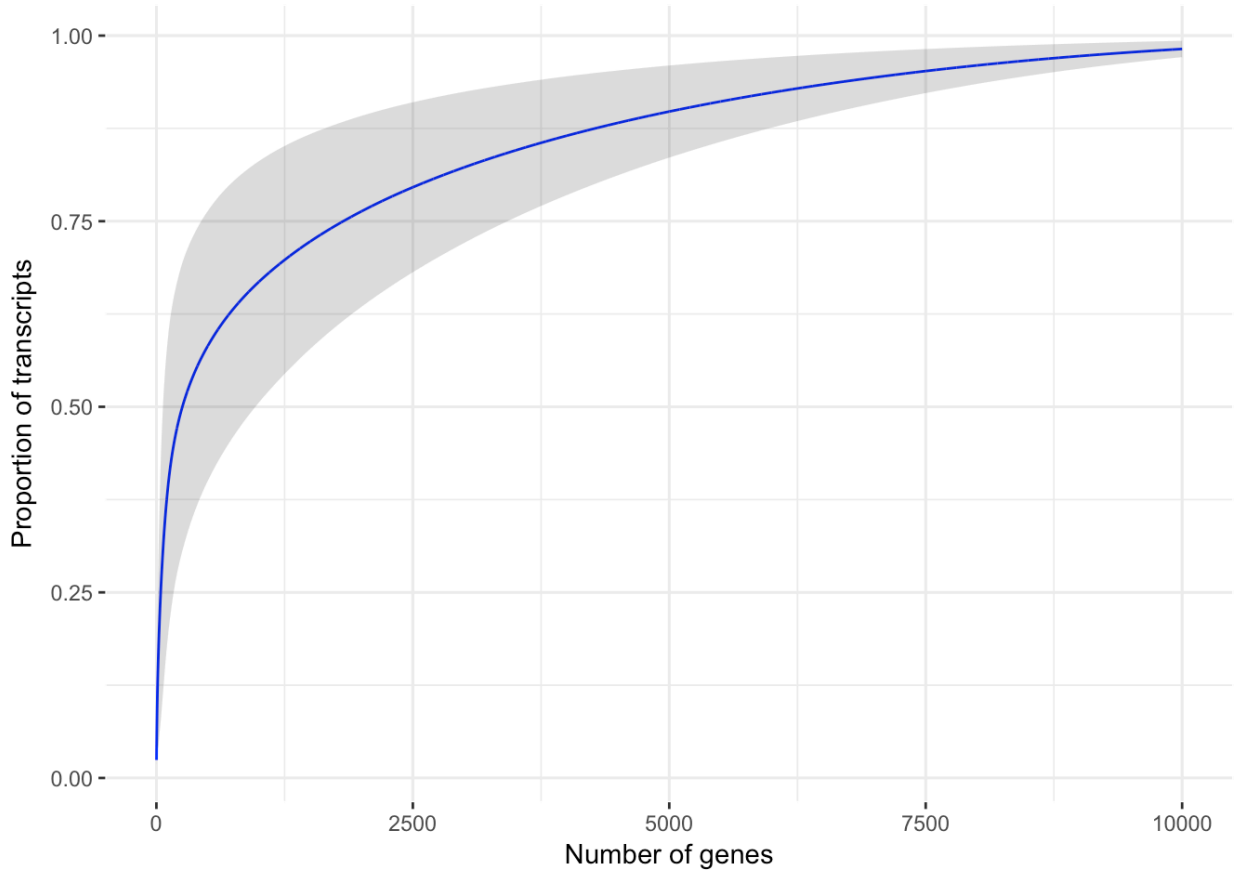


Figure 9. Average proportion of transcripts in the top 10,000 most abundant protein coding genes. Genes were ranked according to transcript abundances. The gene at rank 1 is the most abundant gene in a given sample. The average transcript proportion by gene rank were computed across 4 thyroid samples and is shown in the blue line. The grey ribbon is the mean proportion of transcripts +/- 2 standard deviations.

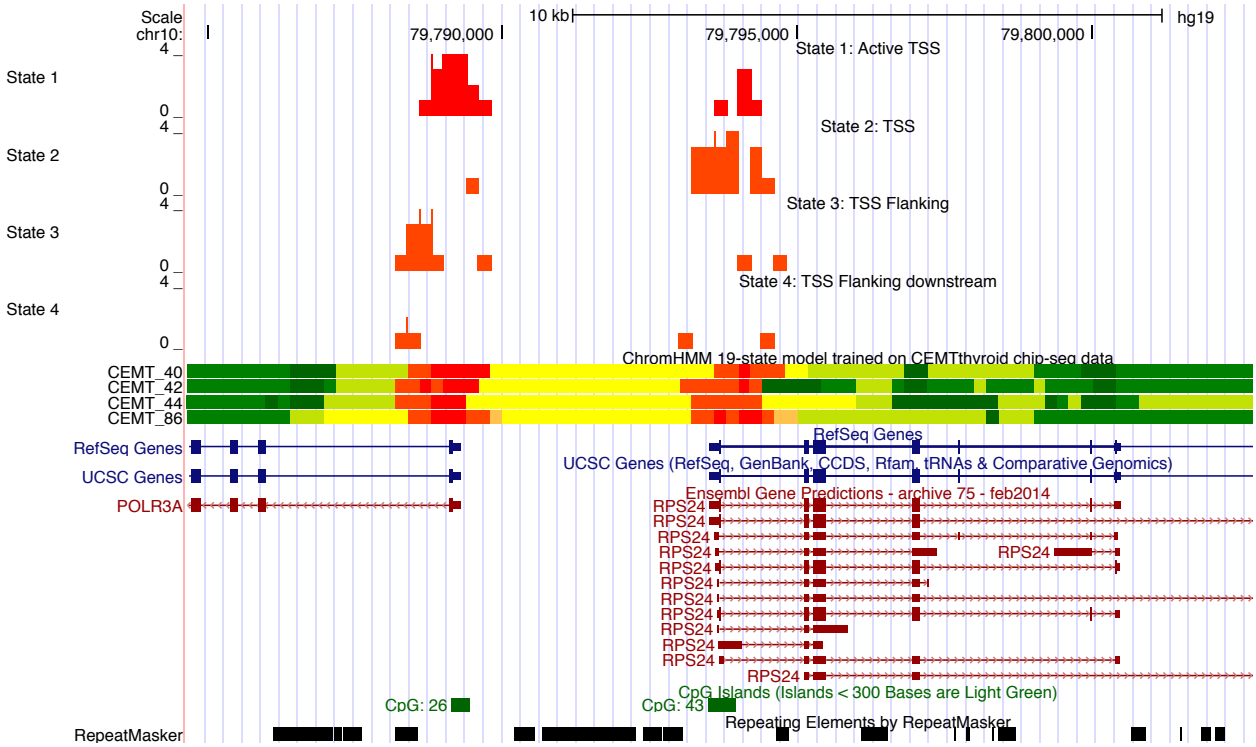


Figure 10 Screenshot of the UCSC Genome Browser showing *RPS24* as being inconsistently marked as active TSS across different bins within the gene promoter. In comparison, *POLR3A* is consistently marked as active TSS within the gene promoter.

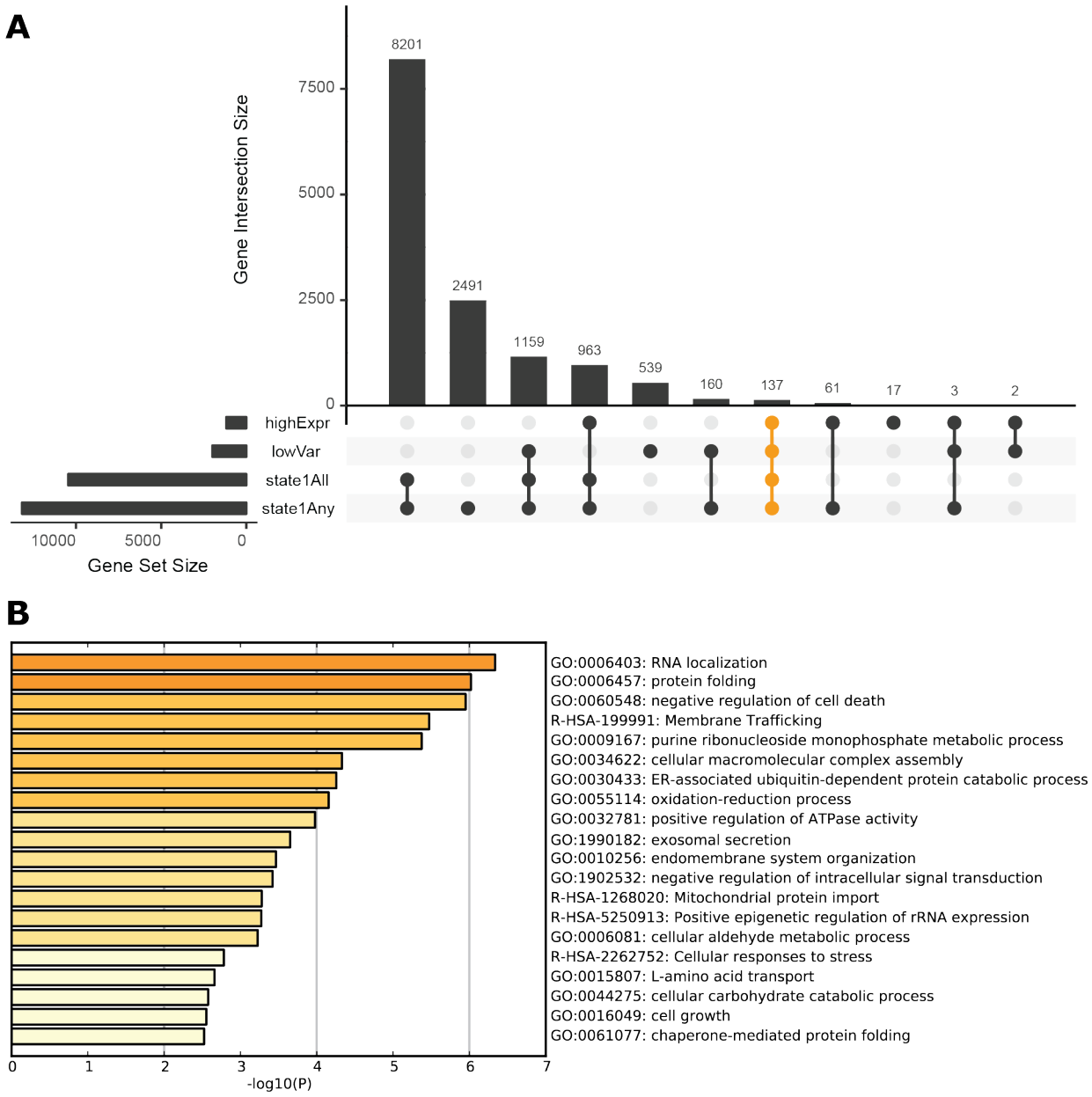


Figure 11. 137 epigenetically active and consistently expressed genes in the thyroid. (A) Epigenetically active and consistently expressed genes were identified based on criteria as follows: is epigenetically marked as state 1 across all 4 epigenomes, have high expression, and have low variance. (B) Metascape gene set enrichment of the 137 genes.

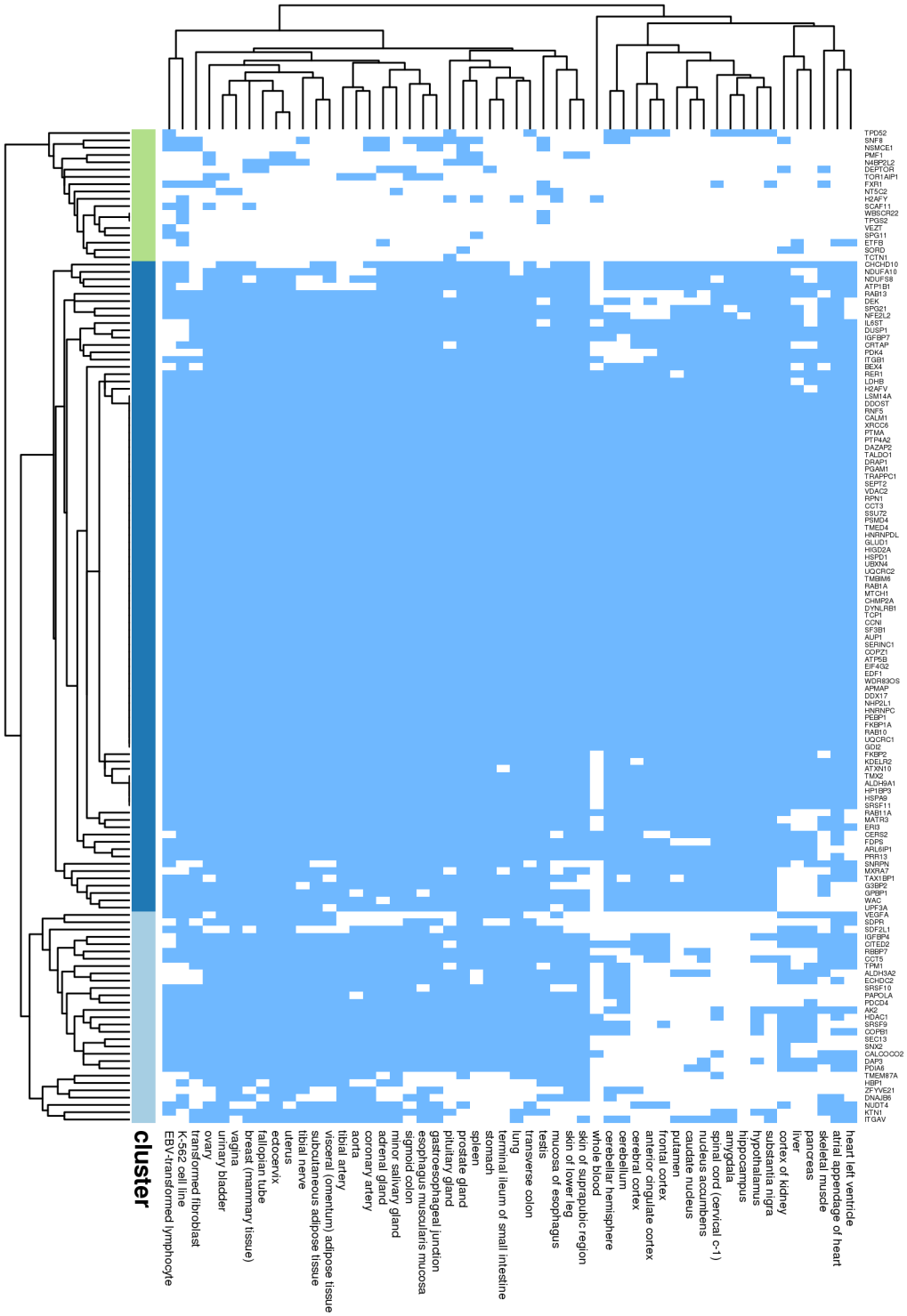


Figure 12. Heat map highlighting 18 genes (green cluster) epigenetically active and consistently expressed in the thyroid with low expression in 52 non-thyroid tissues obtained from GTEx. Blue represent FPKM >= 10, white represents otherwise. Genes in dark blue cluster are present in all tissues, whereas genes in the light blue cluster are present in a subset of predominately non-brain related tissues.

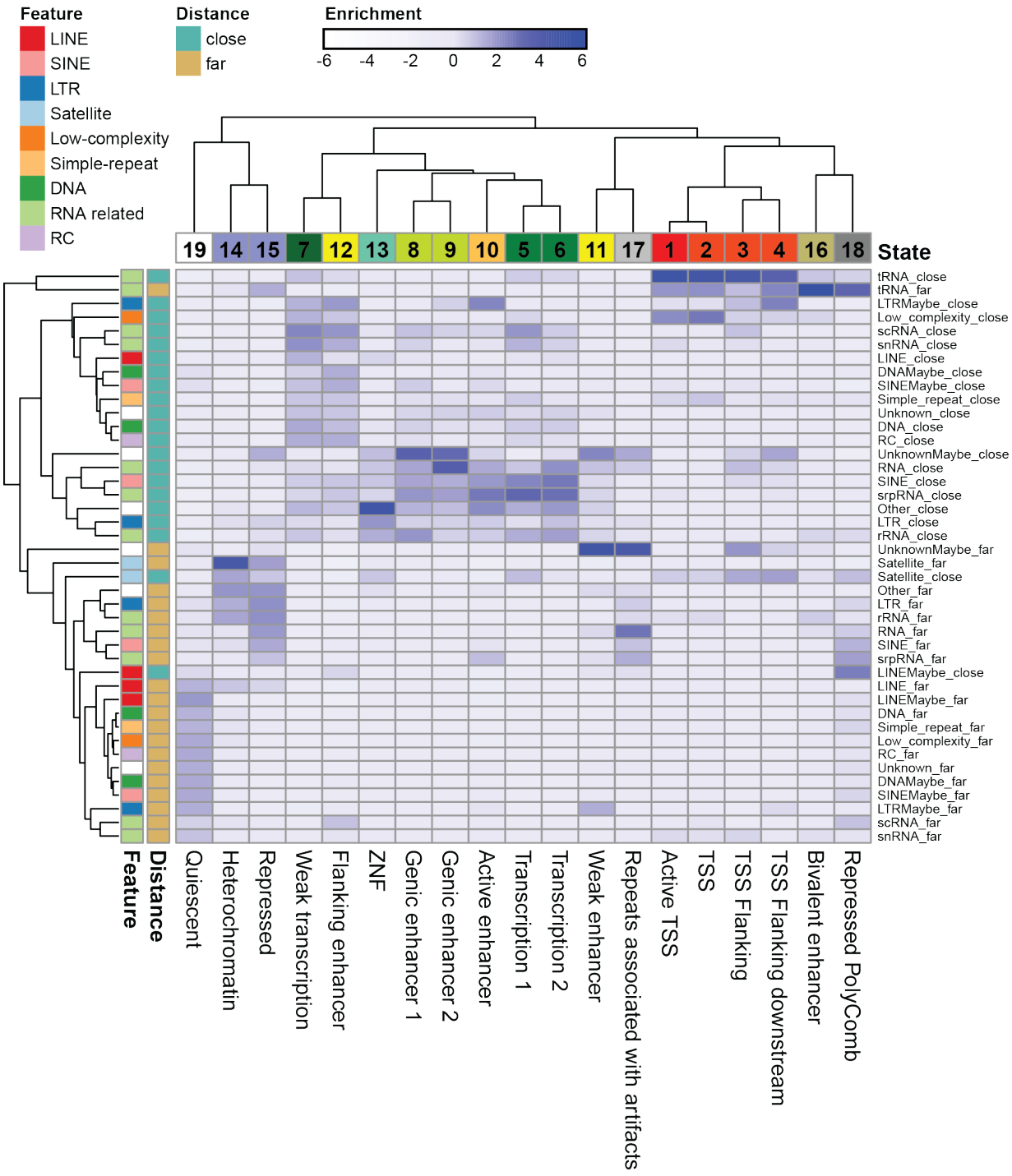


Figure 13. Chromatin state overlap enrichment of repeat regions close to and far from protein coding genes. We consider a gene as close if it is within 10kbp. Coordinates were obtained from RepeatMasker downloaded from the UCSC Table Browser. Overlap enrichment was performed using ChromHMM software. The enrichment values displayed is the average of values from 4 normal thyroid epigenomes.

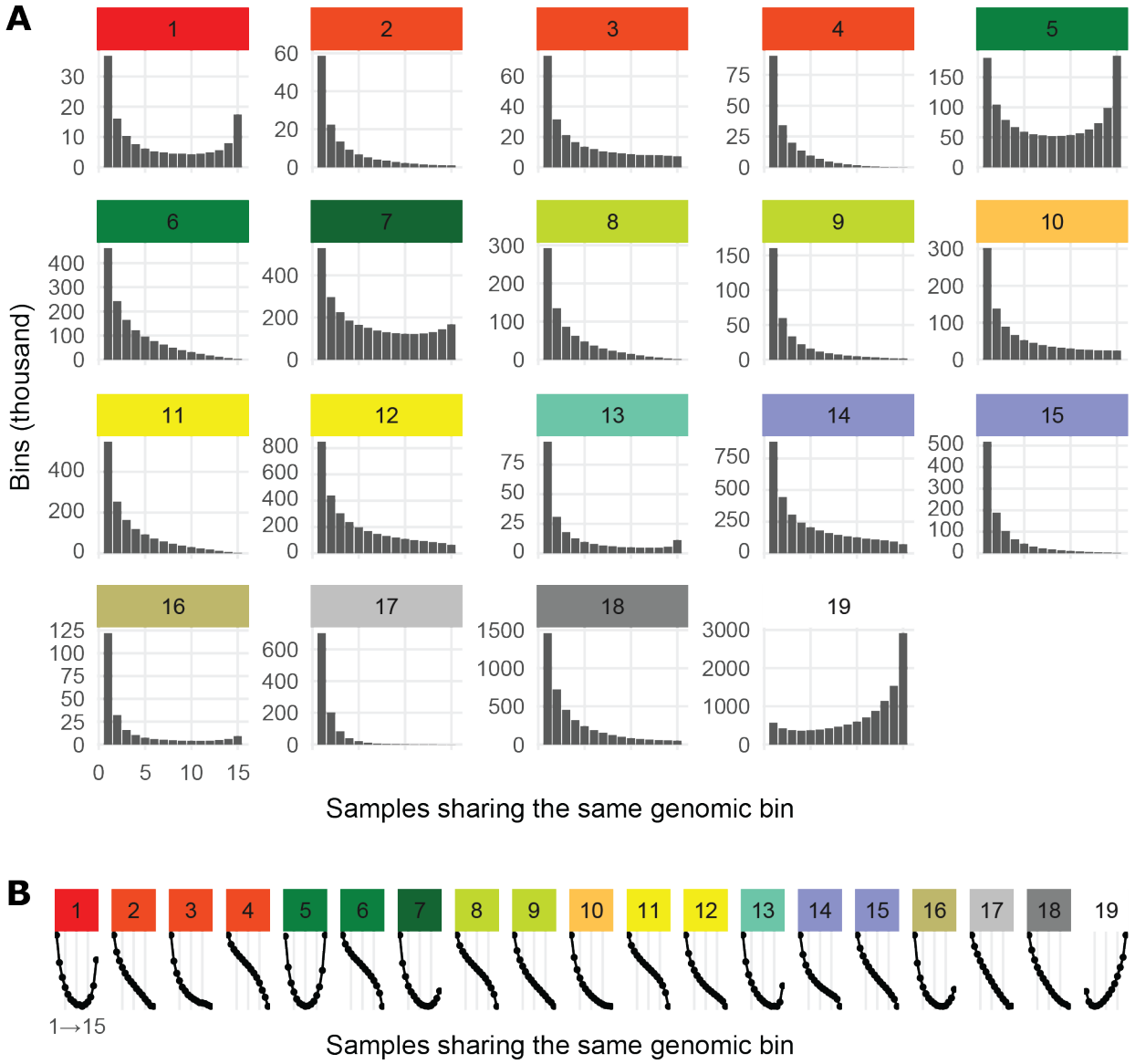


Figure 14. The 19-state model applied to 15 normal colon epigenomes. (A) Histogram showing the number of genomic bins sharing the same state across 4 epigenomes. See Figure 6 for details. (B) Values from (A) scaled to 0 and 1.

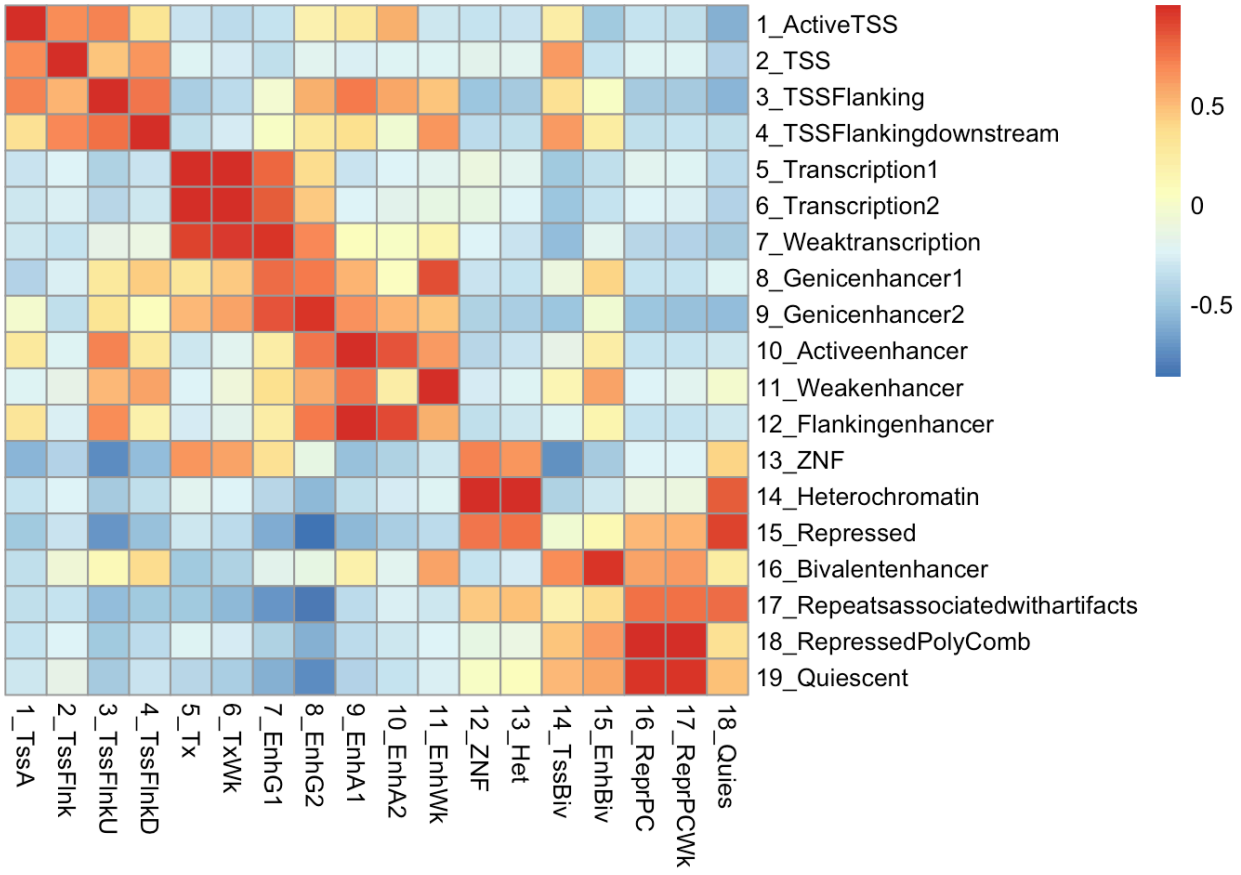


Figure 15. Pearson correlation of state emissions between ChromHMM models. The 19-state model presented in this work is on the y-axis. The 18-state model published in Roadmap (Roadmap Epigenomics Consortium, et al., 2015) is on the x-axis.

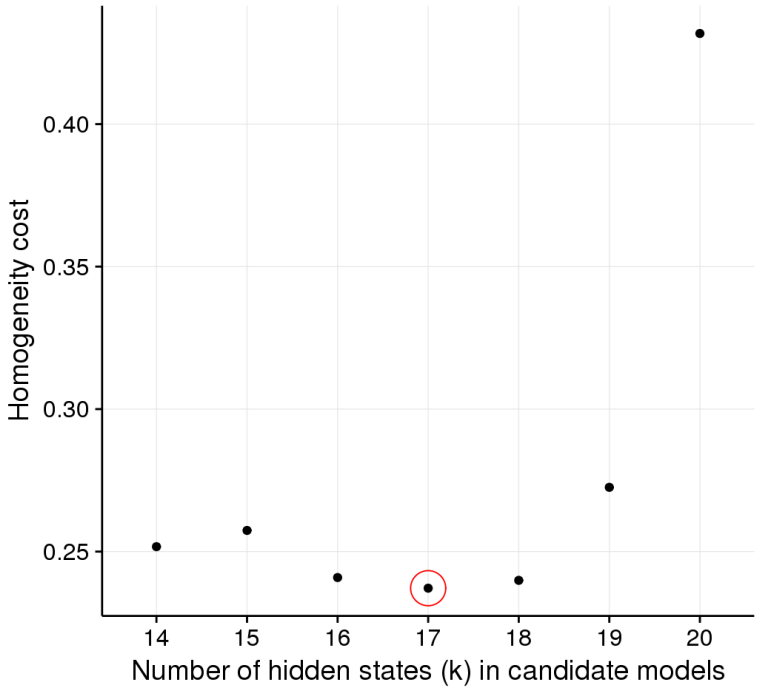


Figure 16. Plot showing the heterogeneity cost for model selection on models trained on 15 normal colon reference epigenomes. Training was specified for k = 14 – 20 states. Input was treated as a control and training was done on 15 normal colon epigenomes using ChromHMM software.

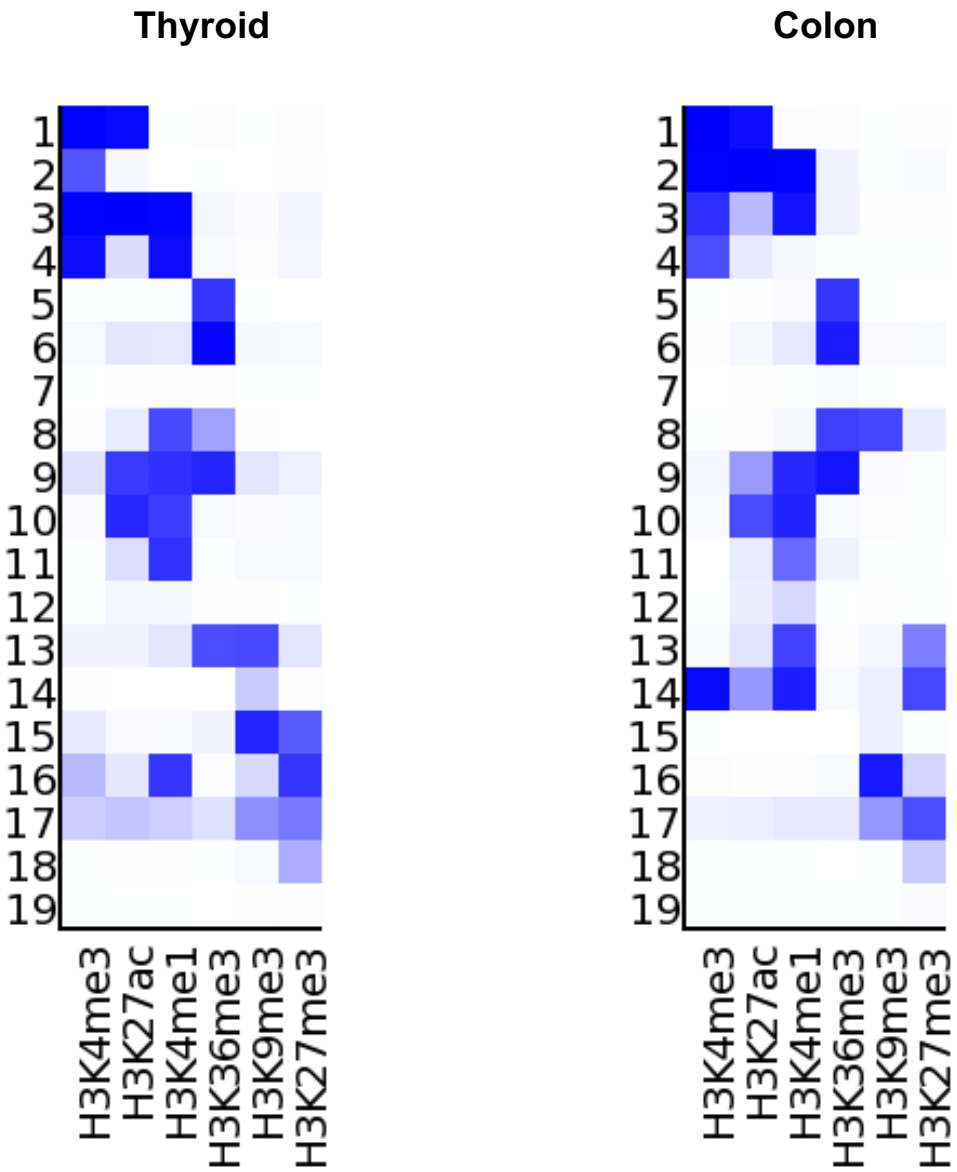


Figure 17 The 19 states generated from the epigenomes of thyroid (left) and colon (right) samples differ. The states and emission probabilities were produced using ChromHMM.

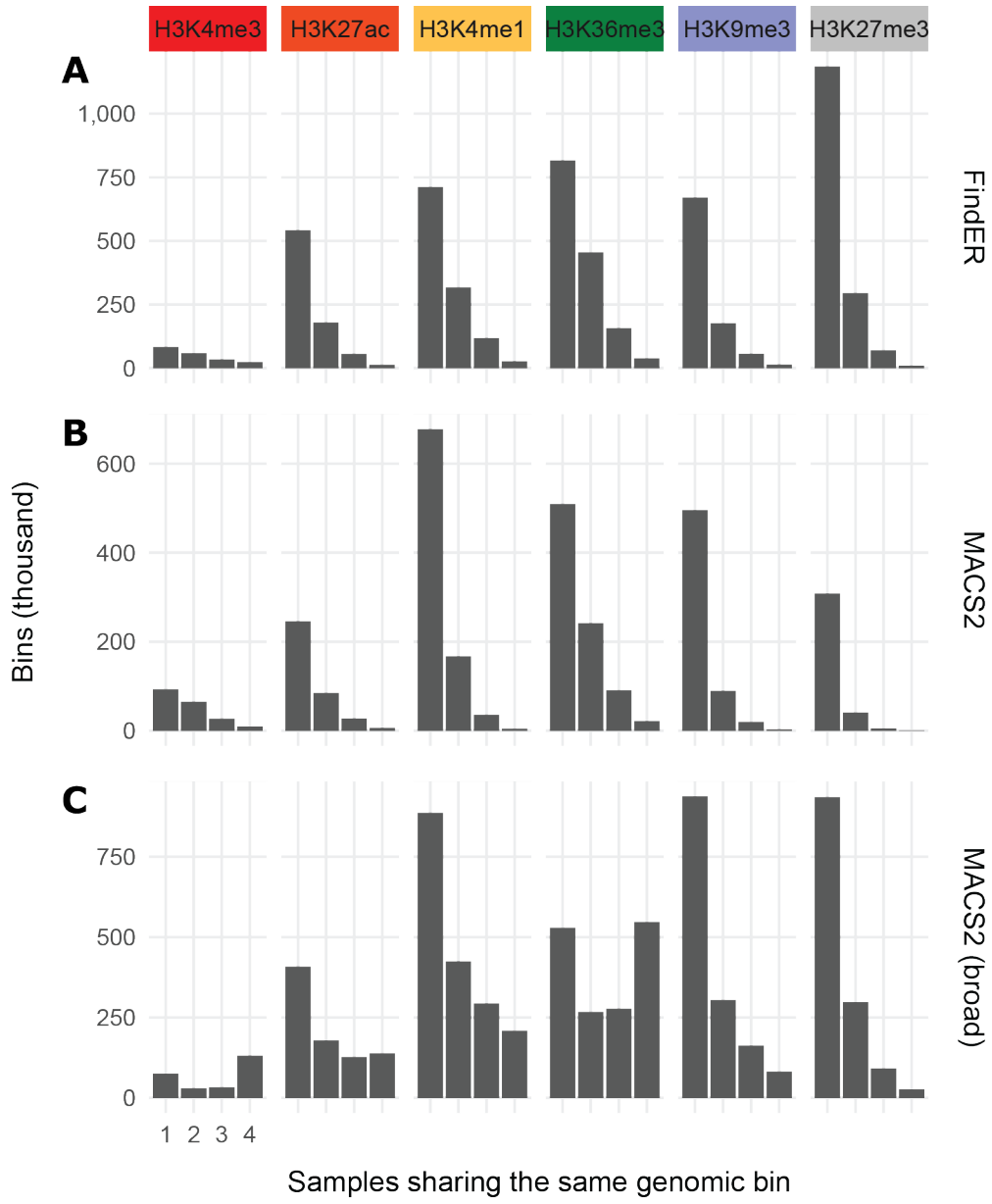
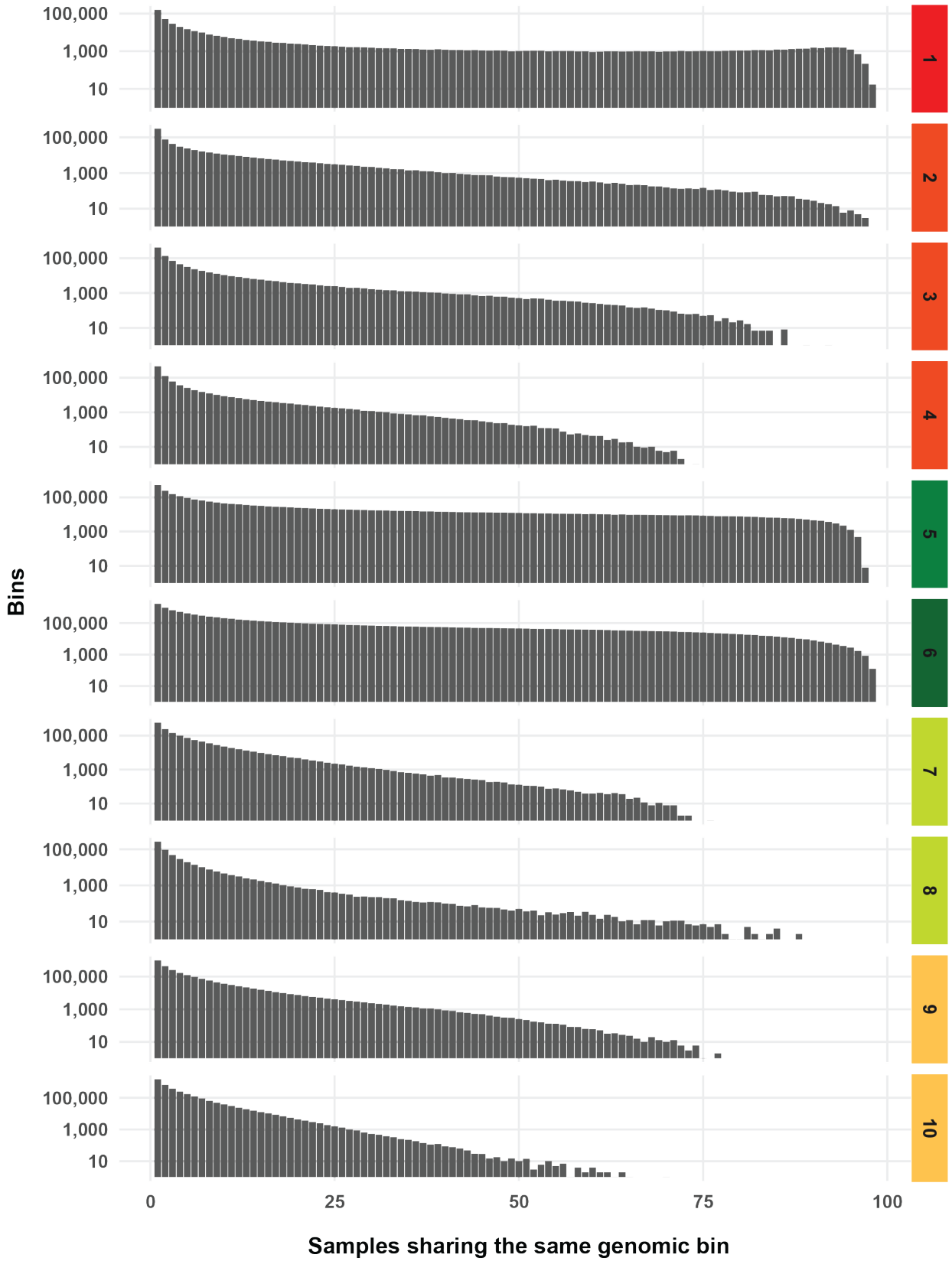


Figure 18. Overview of histone modification consistency across 4 thyroid epigenomes. We used FindER (A) and MACS2 regular and broad (B, C) peak callers to find enriched regions. The genome was divided in 15,181,508 bins. Each bin is 200bp in length and was discretized into two levels: 1 indicating enrichment, and 0 indicating no enrichment. For a particular bin across different individuals, the enrichment of a particular histone may be present in all (x = 4), some (x = {1, 2, 3}), or no (x = 0) individual.


Bins
100,000
1,000
10
1
2
3
4
5
6
7
8
9
10
0
25
50
75
100
Samples sharing the same genomic bin

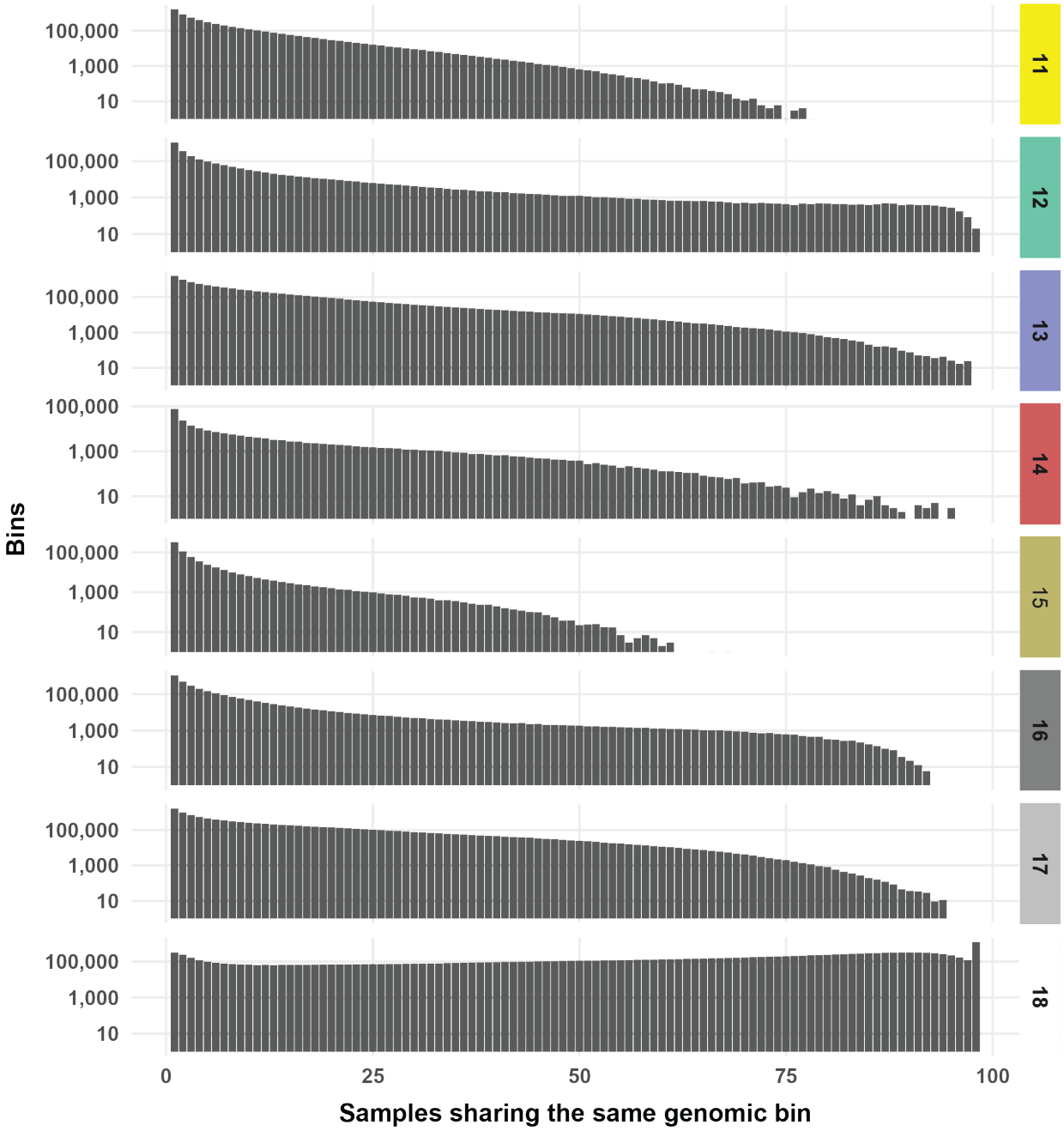


Figure 19 Overview of state consistency across 98 epigenomes published in (Roadmap Epigenomics Consortium, et al., 2015). The segmentations of the 18 Roadmap states for each epigenome were obtained from http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html

Bibliography

Akbaba, G., Omar, M., Polat, M., Özcan, Ö., Bellı, A. K., Şahan, M., & Çullu5, N. (2014). Cutaneous sinus formation is a rare complication of thyroid fine needle aspiration biopsy. *Case Reports in Endocrinology, 2014*, 923438.

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology, 33*(8), 831-838.

Allen, B. M. (1919). The Relation of the Pituitary and Thyroid Glands of Bufo and Rana to Iodine and Metamorphosis. *Biological Bulletin, 36*(6), 405-417.

Alvarez-Mugica, M., Fernandez-Gomez, J. M., Cebrian, V., Fresno, F., Escaf, S., & Sanchez-Carbayo, M. (2013). Polyamine-modulated Factor-1 Methylation Predicts Bacillus Calmette-Guerin Response in Patients with High-grade Non-muscle-invasive Blagged Carcinoma. *European Urology, 63*, 364-370.

Andersen, S., Pedersen, K. M., Bruun, N. H., & Laurberg, P. (2002). Narrow individual variations in serum T4 and T3 in normal subjects: A clue to the understanding of subclinical thyroid disease. *87*(3), 1068-1072.

Bertagna, F., Treglia, G., Piccardo, A., Giovannini, E., Bosio, G., Biasiotto, G., . . . Giubbini, R. (2013). F18-FDG-PET/CT thyroid incidentalomas: A wide retrospective analysis in three Italian centres on the significance of focal uptake and SUV value. *Endocrine, 43*, 678-685.

Bik-Multanowski, M., Pietrzyk, J. J., & Midro, A. (2015). MTRNR2L12: A candidate blood marker of early Alzheimer's disease-like dementia in adults with down syndrome. *Journal of Alzheimer's Disease, 46*(1), 145-150.

Bik-Multanowski, M., Pietrzyk, J. J., & Midro, A. (2015). MTRNR2L12: A candidate blood marker of early Alzheimer's disease-like dementia in adults with down syndrome. *Journal of Alzheimer's Disease, 46*(1), 145-150.

Bomeli, S. R., LeBeau, S. O., & Ferris, R. L. (2010). Evaluation of the thyroid nodule. *Otolaryngologic Clinics of North America, 43*(2), 229-238.

Boros, J., Arnoult, N., Stroobant, V., Collet, J.-F., & Decottignies, A. (2014). Polycomb repressive complex 2 and H3K27me3 cooperate with H3K9 methylation to

maintain heterochromatin protein 1α at chromatin. *Molecular and cellular biology, 34*(19), 3662-3674.

Boulard, M., Storck, S., Cong, R., Pinto, R., Delage, H., & Bouvet, P. (2010). Histone variant macroH2A1 deletion in mice causes female-specific steatosis. *Epigenetics & Chromatin, 3*, 8.

Broad Institute. (n.d.). Picard [Computer software]. Retrieved from https://broadinstitute.github.io/picard.

Brown, D. D., & Cai, L. (2007). Amphibian metamorphosis. *Developmental Biology, 306*(1), 20-33.

Bujold, D., Anderson de Lima Morais, D., Gauthier, C., Cote, C., Caron, M., Kwan, T., . . . Bourque, G. (2016). The International Human Epigenome Consortium Data Portal. *Cell Systems, 3*, 496–499.

Butterfield, Y., Kreitzman, M., Thiessen, N., Corbett, R., Li, Y., Pang, J., . . . Birol, I. (2014). JAGuaR: Junction Alignments to Genome for RNA-Seq Reads. *PLoS One, 9*(7), e102398.

Byrne, J. A., Frost, S., Chen, Y., & Bright, R. K. (2014). Tumor protein D52 (TPD52) and cancer -oncogene understudy or understudied oncogene? *Tumor Biology, 35*(8), 7369-7382.

Cannon, J. (2011). The Significance of Hurthle Cells in Thyroid Disease. *The Oncologist, 16*, 1380-1387.

Catena, V., & Fanciulli, M. (2017). Deptor: not only a mTOR inhibitor. *Journal of Experimental & Clinical Cancer Research, 36*(1), 12.

CEEHRC. (2016). Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network.

CEEHRC. (2016). FindER: A sensitive analytical tool to study epigenetic modifications and protein-DNA interactions from ChIP-Seq data. Retrieved from CEEHRC: http://www.epigenomes.ca/tools-and-software/finder

Chang, J., Lee, S., & Blackstone, C. (2014). Spastic paraplegia proteins spastizin and spatacsin mediate autophagic lysosome reformation. *The Journal of Clinical Investigation, 124*(12), 5249-5262.

Chen, Y., Wang, Y., Xuan, Z., Chen, M., & Zhang, M. Q. (2016). De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Research, 44*(11).

Chun, H. J., Lim, E. L., Heravi-Moussavi, A., Saberi, S., Mungall, K. L., Bilenky, M., . . . Marra, M. A. (2016). Genome-Wide Profiles of Extra-cranial Malignant Rhabdoid Tumors Reveal Heterogeneity and Dysregulated Developmental Pathways. *Cancer Cell, 29*(3), 394-406.

Claudel, T., Zollner, G., Wagner, M., & Trauner, M. (2011). Role of nuclear receptors for bile acid metabolism, bile secretion, cholestasis, and gallstone disease. *Biochimica et Biophysica Acta, 1812*(8), 867-878.

Consortium, R. E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*, 317–330.

Crawford, N. M., & Steiner, A. Z. (2016). Thyroid Autoimmunity and Reproductive Function. *Seminars in Reproductive Medicine*.

Davies, T. F., Latif, R., & Yin, X. (2012). New genetic insights from autoimmune thyroid disease. *Journal of Thyroid Research, 2012*, 6.

Dunham, I., Kundaje, A., Aldred, S., Collins, P., Davis, C., Doyle, F., . . . Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74.

Eladio, N. A., & Gershon, M. D. (1978). *International Review of Cytology* (Vol. 52). (G. H. Bourne, & J. F. Danielli, Eds.) New York: Academic Press, Inc.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods, 9*(3), 215-216.

Ernst, J., & Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology, 33*(2), 364-376.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., . . . Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43-49.

Führer, D., Brix, K., & Biebermann, H. (2015). Understanding the Healthy Thyroid State in 2015. *European Thyroid Journal, 4*(suppl 1), 1-8.

Fu, J., Lv, H., Guan, H., Ma, X., Ji, M., He, N., . . . Hou, P. (2013). Metallothionein 1G functions as a tumor suppressor in thyroid cancer through modulating the PI3K/Akt signaling pathway. *BMC Cancer, 13*, 462.

Gascard, P., Bilenky, M., Sigaroudinia, M., Zhao, J., Li, L., Carles, A., . . . Hirst, M. (2015). Epigenetic and transcriptional determinants of the human breast. *Nature Communications*(6), 6351.

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., & Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature, 518*, 365-369.

González-Peñas, J., Amigo, J., Santomé, L., Sobrino, B., Brenlla, J., Agra, S., . . . Costas, J. (2016). Targeted resequencing of regulatory regions at schizophrenia risk loci: Role of rare functional variants at chromatin repressive states. *Schizophrenia Research, 174*(1-3), 10-16.

González-Peñas, J., Amigo, J., Santomé, L., Sobrino, B., Brenlla, J., Agra, S., . . . Costas, J. (2016). Targeted resequencing of regulatory regions at schizophrenia risk loci: Role of rare functional variants at chromatin repressive states. *174*, 10-16.

Gudernatsch, J. F. (1912). Feeding Experiments on Tadpoles. *Archiv für Entwicklungsmechanik der Organismen, 35*(3), 457-483.

Hamada, M., Ono, Y., Fujimaki, R., & Asai, K. (2015). Learning chromatin states with factorized information criteria. *Bioinformatics, 31*(15), 2426-2433.

Harris, G., & Donovan, B. (1961). The Thyroid Gland. In C. H. Best, & N. B. Taylor, *The Physiological Basis of Medical Practice* (7th Edition ed., pp. 997-1021). Baltimore: The Williams & Wilkins Company.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., . . . Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell, 38*(4), 576-589.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods, 9*(5), 473-476.

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., . . . Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research, 41*(2), 827-841.

Hontelez, S., van Kruijsbergen, I., Georgiou, G., van Heeringen, S. J., Bogdanovic, O., Lister, R., & Veenstra, G. J. (2015). Embryonic transcription is controlled by maternally defined chromatin state. *Nature Communications, 6*, 10148.

IHEC. (n.d.). *Japan*. Retrieved 11 23, 2016, from IHEC: International Human Epigenome Consortium: http://ihec-epigenomes.org/about/ihec-countries/jp/

Infante, C., Asensio, E., Cañavate, J. P., & Manchado, M. (2008). Molecular characterization and expression analysis of five different elongation factor 1 alpha genes in the flatfish Senegalese sole (Solea senegalensis Kaup): differential gene expression and thyroid hormones dependence during metamorphosis. *9*, 19.

Jablonka, E., & Lamm, E. (2012). Commentary: The epigenotype - a dynamic network view of development. *International Journal of Epidemiology, 41*, 16-20.

Jancic, S. A., & Stosic, B. Z. (2014). Cadmium Effects on the Thyroid Gland. *Vitamins and Hormones, 94*, 391-425.

Jufvas, A., Stralfors, P., & Vener, A. V. (2011). Histone variants and their post-translational modifications in primary human fat cells. *PLoS ONE, 6*(1), e15960.

Karolchik D, H. A. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research, 32*(Database issue), D493-6.

Kasaian, K. (2015). *Genomic analysis of head and neck endocrine glands.* PhD Thesis, University of British Columbia.

Lai, X., & Chen, S. (2015). Identification of Novel Biomarker and Therapeutic Target Candidates for Diagnosis and Treatment of Follicular Adenoma. *Cancer Genomics & Proteomics, 12*, 271-281.

Lee, K., & Park, H. (2016). Building the SeqChromMM Markov property atlas of the human genome by analyzing the 200-bp units of the 15 different chromatin regions of ENCODE. *15*(3).

Lent, H., Lee, K.-E., & Park, H.-S. (2015). Building the Frequency Profile of the Core Promoter Element Patterns in the Three ChromHMM Promoter States at 200bp Intervals: A Statistical Perspective. *Genomics & informatics, 13*(4), 152-155.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Subgroup, 1. G. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

LifeLabs. (2016). *LifeLabs Clinical Laboratories Burnaby Reference Laboratory Reference Intervals.*

Lin, R.-Y. (2011). Thyroid cancer stem cells. *Nature Review Endocrinology, 7*, 609-616.

Liu, J., Magri, L., Zhang, F., Marsh, N. O., Albrecht, S., Huynh, J. L., . . . Casaccia, P. (2015). Chromatin Landscape Defined by Repressive Histone Methylation during Oligodendrocyte Differentiation. *The Journal of Neuroscience, 35*(1), 352-365.

Lorzadeh, A., Bilenky, M., Hammond, C., Knapp, D. J., Li, L., Miller, P. H., . . . Hirst, M. (2016). Nucleosome Density ChIP-Seq Identifies Distinct Chromatin Modification Signatures Associated with MNase Accessibility. *Cell Reports, 17*(8), 2112-2124.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*, 550.

Mack, S. C., Witt, H., Piro, R. M., Gu, L., Zuyderduyn, S., Stutz, A. M., . . . Taylor, M. D. (2014). Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *506*(7489), 445-450.

Martens, J. H., & Stunnenberg, H. G. (2013). BLUEPRINT: Mapping human blood cell epigenomes. *Haematologica, 98*(10), 1487-1489.

McHenry, C., & Phitayakorn, R. (2011). Follicular Adenoma and Carcinoma of the Thyroid Gland. *The Oncologist, 16*(5), 585-593.

Nakatake, N., Fukata, S., & Tajiri, J. (2012). Acute transient thyroid swelling after fine-needle aspiration biopsy: Three cases during only 6 weeks - A rare complication? *Clinical Endocrinology, 77*, 152-157.

Nonaka, D., Tang, Y., Chiriboga, L., Rivera, M., & Ghossein, R. (2008). Diagnostic utility of thyroid transcription factors Pax8 and TTF-2 (FoxE1) in thyroid epithelial neoplasms. *Modern Pathology, 21*(10), 192-200.

Norrenberg, S., Rorive, S., Laskar, P., Catteau, X., Delpierre, I., Avni, F. E., & Salmon, I. (2011). Acute transient thyroid swelling after fine-needle aspiration biopsy: rare complication of unknown origin. *Clinical Endocrinology, 75*, 568-570.

Orent, W., Mchenry, A. R., Rao, D. A., White, C., Klein, H.-U., Bassil, R., . . . Elyaman, W. (2016). Rheumatoid arthritis-associated RBPJ polymorphism alters memory CD4+ T cells. *Human Molecular Genetics, 25*(2), 404-417.

Ounap, K., Kasper, L., Kurg, A., & Kurg, R. (2013). The Human WBSCR22 Protein Is Involved in the Biogenesis of the 40S Ribosomal Subunits in Mammalian Cells. *PLoS ONE, 8*(9), e75686.

Panicker, V. (2011). Genetics of thyroid function and disease. *The Clinical Biochemist Reviews, 32*, 165-175.

Patel, M. R., Stadler, M. E., Deal, A. M., Kim, H. S., Shores, C. G., & Zanation, A. M. (2011). STT3A, C1orf24, TFF3: Putative markers for characterization of follicular thyroid neoplasms from fine-needle aspirates. *Laryngoscope, 121*(5), 983-989.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2016). Salmon provides accurate , fast , and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv*.

Peach, S. E., Rudomin, E. L., Udeshi, N. D., Carr, S. A., & Jaffe, J. D. (2012). Quantitative assessment of chromatin immunoprecipitation grade antibodies directed against histone modifications reveals patterns of co-occurring marks on histone protein molecules. *Molecular & cellular proteomics, 11*(5), 128-137.

Pei, L., Xie, P., Zhou, E., Yang, Q., Luo, Y., & Tang, Z. (2011). Overexpression of DEP domain containing mTOR-interacting protein correlates with poor prognosis in differentiated thyroid carcinoma. *Molecular Medicine Reports, 4*, 817-823.

Pellacani, D., Bilenky, M., Kannan, N., Heravi-Moussavi, A., Knapp, D. J., Gakkhar, S., . . . Eaves, C. J. (2016). Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *17*(8), 2060-2074.

Pellacani, D., Bilenky, M., Kannan, N., Heravi-Moussavi, A., Knapp, D. J., Gakkhar, S., . . . Eaves, C. J. (2016). Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Reports, 17*(8), 2060-2074.

Rebehmed, J., Revy, P., Faure, G., De Villartay, J. P., & Callebaut, I. (2014). Expanding the SRI domain family: A common scaffold for binding the phosphorylated C-terminal domain of RNA polymerase II. *FEBS Letters, 588*(23), 4431-4437.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *518*, 317-330.

Roberto, R., Vittorio, S., Assunta, M., Giovanna, D. M., Stefania, T., Carmela, M., . . . Mario, T. (2016). Gene expression profiling of normal thyroid tissue from patients with thyroid carcinoma. *Oncotarget*, 3-5.

Silverman, J., West, R., Larkin, E., Park, H., Finley, J., Swanson, M., & Fore, W. (1986). The Role of Fine-Needle Aspiration Biopsy in the Rapid Diagnosis and Management of Thyroid Neoplasm. *Cancer, 57*(6), 1164-1170.

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research, 4*, 1521.

Sousa, S., Cabanes, D., El-Amraoui, A., Petit, C., Lecuit, M., & Cossart, P. (2004). Unconventional myosin VIIa and vezatin, two proteins crucial for Listeria entry into epithelial cells. *Journal of Cell Science, 117*, 2121-2130.

Stunnenberg, H. G., Consortium, T. I., & Hirst, M. (2016). The International Human Epigenome Consortium (IHEC): A Blueprint for Scientific Collaboration and Discovery. *Cell, 167*(5), 1145-1149.

Tennstedt, P., Bölch, C., Strobel, G., Minner, S., Burkhardt, L., Grob, T., . . . Simon, R. (2014). Patterns of TPD52 overexpression in multiple human solid tumor types analyzed by quantitative PCR. *International Journal of Oncology, 44*, 609-615.

The Cancer Genome Atlas Research Network. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine, 368*(22), 2059-2074.

Tripathi, S., Pohl, M., Zhou, Y., Rodriguez, -F. A., Wang, G., Stein, D., . . . Chanda, S. (2015). Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host & Microbe, 18*, 723-735.

Trueba, S. S., Auge, J., Mattei, G., Etchevers, H., Martinovic, J., Czernichow, P., . . . Attie-Bitach, T. (2005). PAX8, TITF1, and FOXE1 gene expression patterns during human development: New insights into human thyroid development and thyroid dysgenesis-associated malformations. *Journal of Clinical Endocrinology and Metabolism, 90*(1), 455-462.

Vigneri, R., Malandrinoa, P., & Vigneri, P. (2015). The changing epidemiology of thyroid cancer: why is incidence increasing? *Current Opinion in Oncology, 1*, 1-7.

Wang, Z., Gao, Y., Liu, Y., Chen, J., Wang, J., Gan, S., . . . Cui, X. (2015). Tectonic-1 contributes to the growth and migration of prostate cancer cells in vitro. *International Journal of Molecular Medicine, 36*(4), 931-938.

Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., . . . Meaney, M. J. (2004). Epigenetic programming by maternal behavior. *Nature neuroscience, 7*(8), 847-854.

Whitaker, J. W., Chen, Z., & Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nature Methods, 12*(3), 365-372.

Wijetunga, N. A., Delahaye, F., Zhao, Y. M., Golden, A., Mar, J. C., Einstein, F. H., & Greally, J. M. (2014). The meta-epigenomic structure of purified human stem cell populations is defined at cis-regulatory sequences. *Nature communications, 5*, 5195.

Yen, A., & Kellis, M. (2015). Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *6*, 7973.

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., & Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics, 25*(15), 1952-1958.

Zhang Y, L. T. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology, 9*(9), R137.