Characterization of the human thyroid epigenome

Celia Siu^{1,2}, Sam Wiseman³, Sitanshu Gakkhar¹, Alireza Heravi-Moussavi¹, Misha Bilenky¹, Annaick Carles⁴, Thomas Sierocinski⁴, Angela Tam¹, Eric Zhao¹, Katayoon Kasaian¹, Richard A Moore¹, Andrew J Mungall¹, Blair Walker⁵, Thomas Thomson⁶, Marco A Marra^{1,7}, Martin Hirst^{1,4} and Steven J M Jones^{1,7,8}

¹Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, Canada

²Department of Sciences, University of British Columbia, Vancouver, Canada

³Department of Surgery, St. Paul's Hospital & University of British Columbia, Vancouver, Canada

⁴Department of Microbiology & Immunology, Michael Smith Laboratories, University of British Columbia, Vancouver, Canada

⁵Department of Pathology and Laboratory Medicine, St. Paul's Hospital & University of British Columbia, Vancouver, Canada

⁶Department of Pathology and Laboratory Medicine, BC Cancer Agency & University of British Columbia, Vancouver, Canada

⁷Department of Medical Genetics, University of British Columbia, Vancouver, Canada

⁸Department of Molecular Biology & Biochemistry, Simon Fraser University, Burnaby, Canada

Correspondence should be addressed to S J M Jones
Email
sjones@bcgsc.ca

Abstract

The thyroid gland, necessary for normal human growth and development, functions as an essential regulator of metabolism by the production and secretion of appropriate levels of thyroid hormone. However, assessment of abnormal thyroid function may be challenging suggesting a more fundamental understanding of normal function is needed. One way to characterize normal gland function is to study the epigenome and resulting transcriptome within its constituent cells. This study generates the first published reference epigenomes for human thyroid from four individuals using ChIP-seq and RNA-seq. We profiled six histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3), identified chromatin states using a hidden Markov model, produced a novel quantitative metric for model selection and established epigenomic maps of 19 chromatin states. We found that epigenetic features characterizing promoters and transcription elongation tend to be more consistent than regions characterizing enhancers or Polycomb-repressed regions and that epigenetically active genes consistent across all epigenomes tend to have higher expression than those not marked as epigenetically active in all epigenomes. We also identified a set of 18 genes epigenetically active and consistently expressed in the thyroid that are likely highly relevant to thyroid function. Altogether, these epigenomes represent a powerful resource to develop 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.

Key Words

- thyroid
- epigenetics
- gene expression
- gene regulation
- ChIP-seq

Journal of Endocrinology
(2017) **235**, 153–165

Introduction

The normal human thyroid is a homogeneous tissue mainly composed of two cell types: follicular cells and parafollicular cells. Thyroid follicular cells are epithelial cells responsible for the production, storage and secretion of thyroid hormone. Parafollicular cells (also known as C cells) account for only a relatively small proportion of the thyroid cells (Eladio & Gershon 1978) and produce calcitonin.

The thyroid gland produces and secretes hormones necessary for growth and development and is involved with the regulation of metabolism. Assessment of thyroid function is based on blood serum concentrations of thyroid-related hormones (thyroid-stimulating hormone (TSH), triiodothyronine (T3) or thyroxine (T4)) in predefined normal ranges (Führer *et al.* 2015). However, the definition of a 'normal' TSH, T3 and T4 concentration range is controversial (Führer *et al.* 2015) when variability in individual factors such as sex, body mass index, exclusion of incident thyroid disease, ethnicity and iodine and selenium intake are considered. Furthermore, thyroid nodules are common, diagnosed in 5% of the general population by palpation and in 50% by ultrasound (Gharib & Papini 2007) suggesting frequent local heterogeneity within thyroid glands. The result is that accurate assessment of abnormal thyroid states across individuals is challenging.

One way to study thyroid function is to examine the epigenetics involved in the regulation of thyroid gene expression and transcription. Epigenetics, referring to the reversible changes in chromatin and DNA that can regulate gene activity and expression, include the post-translational modifications of histone proteins and DNA methylation. In cells, DNA is packaged into chromatin, a complex of DNA, proteins and RNA. The basic repeating unit of chromatin, a nucleosome, consists of about 200 base pairs (bp) of DNA wrapped around both a histone protein octamer and a linker protein. This histone octamer can be chemically modified to signal an activation or repression of transcription; such modifications include H3K4me3 associated with active promoters, H3K27ac with active enhancers and promoters, H3K4me1 with active enhancers, H3K36me3 with transcribed gene bodies, H3K9me3 with heterochromatin and H3K27me3 with Polycomb-repressed regions (Roadmap Epigenomics Consortium 2015). Undoubtedly, the distribution of different histone modifications reveals different epigenetic signals. 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 defined chromatin states.

Genomewide 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 2015). This and other large-scale international projects have aimed to sequence and decipher the human epigenomes of various cell types in order to understand how epigenetic processes contribute to human biology and disease (Stunnenberg *et al.* 2016).

The goals of this work are to provide a resource of human thyroid epigenomic data and to introduce a novel quantitative metric for model selection. In this study, reference epigenomes were generated from the thyroid tissues of four individuals. Each specimen has a complete set of six histone marks (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3) profiled with ChIP-seq, a methylome, a transcriptome and a normal and disease-matched genomic sequence. We partitioned the epigenomes into various chromatin states and developed a novel quantitative metric for model selection. We selected a model for further analysis and compared chromatin state consistencies across four epigenomes. We found that the epigenetic features characterizing promoters and transcription elongation tend to be more consistent across samples. We also found that genes that are consistently epigenetically active across all individuals tend to have higher expression than genes not marked as epigenetically active or only active in a subset of epigenomes. The findings provide four reference thyroid epigenomes as a valuable resource for future study of the function and regulation of the human thyroid gland.

Materials and methods

Samples

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 (Supplementary Table 1, see section on supplementary data given at the end of this article). The pathologic findings reflect the challenge of obtaining normal thyroid tissue from healthy individuals.

The specimens referred to as 'normal' in this study are from microscopically uninvolved thyroid tissue in the resected thyroid glands.

ChIP sequencing and RNA sequencing

Human thyroid chromatin immunoprecipitation sequencing (ChIP-seq) and RNA sequencing (RNA-seq) data were collected as previously described (Pellacani *et al.* 2016). The antibodies for ChIP-seq were obtained from Diagenode (Denville, NJ, USA), Abcam and Cell Signaling Technologies. The catalog numbers for each company, respectively, are C15410037/pAb-037-050 (H3K4me1), C15410056/pAb-056-050 (H3K9me3), C15410195/pAb-195-050 (H3K27me3); ab4729 (H3K27ac), ab9050 (H3K36me3) and 9751S (H3K4me3). For sample CEMT_86/87, one lane of sequencing was merged with native ChIP protocol. With regards to RNA-seq, purification of RNA was followed by poly-A RNA selection. Conversion of RNA to cDNA was done by random priming. 75 base pair paired-end reads were sequenced on an Illumina HiSeq 2500 (Illumina Inc., San Diego, CA, USA). Alignment was to the GRCh37-lite reference and processed datasets and all underlying raw DNA sequences have been deposited at the European Genome-phenome Archive (EGA, www.ebi.ac.uk/ega/) under accession number EGAS00001000552. In this work, CEMT_40–45 and CEMT_86–87 were the normal and diseased thyroid samples utilized for analysis. Detailed methodology for ChIP-seq and RNA-seq library construction, read alignment and data processing is available in the Supplemental Experimental Procedures of Pellacani *et al.* (2016) at www.epigenomes.ca/protocols-and-standards or upon request.

Promoters

In this study, promoters were defined to be regions around the annotated transcription start site (TSS) $\pm$ 1 (kilobase pair) kbp. 1 kbp from the TSS was used given that this distance encapsulates the promoter signal as observed in the RefSeq TSS neighborhood enrichments generated by ChromHMM (Fig. 3C). The coordinates for the TSS promoter regions were obtained from the Ensembl GRCh37 Release 75 Gene sets GTF file that is available at <http://feb2014.archive.ensembl.org/info/data/ftp/index.html>. The gene set was filtered for 'protein_coding' (source) 'transcript' (feature) on chromosomes 1–22, X and Y. In total, we obtain 81,732 transcripts derived from 20,314 protein-coding genes across the standard

chromosomes. 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, gene expression and gene variance

Detailed methodology for estimating transcript abundance, gene expression and gene variance is in the [Supplementary Materials and methods](#). In brief, we used Salmon, v0.7.2 (Patro *et al.* 2017) to estimate transcript abundance from 75nt in length RNA-seq reads. The reference transcriptome was downloaded from the UCSC Table Browser. 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 sum up the Salmon estimated transcript abundances (and read counts) within genes for gene-level abundances, the tximport::tximport R function, v1.2.0 (Soneson *et al.* 2015) was used. The regularized logarithm transformation (rlog) function of the DESeq2 R package, v1.14.0 (Love *et al.* 2014) was then used to transform tximport generated read count data to render them homoskedastic. Gene variance was calculated on the rlog transformed read counts.

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 or transcription factor-binding sites (Roadmap Epigenomics Consortium 2015). Reference epigenomes also have been used to provide context to genomic locations such as single nucleotide variants (SNVs) or expression quantitative trait loci (eQTLs) (González-Peñas *et al.* 2016). In this study, reference epigenomes from tumor and adjacent normal thyroid tissue of four human adult subjects. In total, we generated 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. H3K4me1, H3K4me3, H3K27ac, H3K36me3,

H3K9me3 and H3K27me3 are the six histone modifications of this study, and they coincide with the core set of histone modifications profiled as part of the International Human Epigenome Consortium (Stunnenberg *et al.* 2016). These data can be viewed on the UCSC Genome and Wash U Epigenome Browsers through www.epigenomes.ca/data-release/ and through a link provided in www.bcgsc.ca/data/thyroid (Fig. 1).

Defining chromatin states

ChromHMM (Ernst & Kellis 2012), an implementation of a hidden Markov model (HMM), uses epigenetic features such as histone modifications to represent observed states and unobserved, or hidden, states to represent chromatin states. Generally, HMMs have 2 parameters: (1) emission probabilities representing the observed (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) needs to be specified programmatically. In this study, we divided the genome into 15,181,508 genomic bins and trained ChromHMM on $k=11-23$ states (Supplementary Materials and methods). The number of hidden states used encompassed the number of states selected by the NIH Roadmap Consortium for the analysis of epigenomic states across 111 cell types (Roadmap Epigenomics Consortium 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. In total, we trained 26 candidate models in order to select the final model for further analysis. In this

Figure 1

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 www.bcgsc.ca/data/thyroid. (A) The overlap of ChIP-seq from six histone modifications belonging to sample CEMT_42 and CEMT_44. (B) The overlap of, respectively, H3K4me3 and H3K27ac ChIP-seq across four normal samples. (C) 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. (D) The overview of ChromHMM state segmentations for each epigenome. Definition of track colors are listed in Fig. 3A.

Figure 2

Plots showing the homogeneity cost used for model selection. Formulation for the homogeneity cost is presented in the Supplementary Methods. 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 homogeneity cost. 19 states with input as control were chosen for the model to use for further analysis.

study, we also introduce a novel quantitative selection metric (Supplementary Methods) that maximizes the homogeneity of epigenetic features in chromatin states across samples, selecting 19 states with input treated as control and 20 states with input treated as a mark as the optimal number of states to be utilized (Fig. 2).

Between the 19 states with input treated as control and 20 states with input treated as a mark, as selected by the novel quantitative selection metric, we proceeded with 19 states using input as control based on (1) there were less states and (2) the Roadmap project (Roadmap Epigenomics Consortium 2015) treated input as control. Similar to the 18-state model published for 98 primary human tissues and cell types (Roadmap Epigenomics Consortium 2015), we found our model recapitulates many of the states with a few notable differences (Fig. 3A): (1) we have 19 states while Roadmap has 18; (2) our model, in accordance to state enrichments described

in the next section, has repressed (state 15) and repeat (state 17) states not published in (Roadmap Epigenomics Consortium 2015) and (3) we lack the bivalent TSS state published in Roadmap Epigenomics Consortium (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 correlated with various known genomic features (Fig. 3). States 1–4 are enriched in regions of transcription initiation and promoters (Fig. 3C). H3K36me3-associated emissions correlate with genes, introns and exons in states 5–9, suggesting these states are related to transcribed gene bodies. In comparison, states 9–12 have emissions associated with H3K4me1,

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) average genomic coverage values, CEMT_44 genomic feature enrichments, and (C) CEMT_44 neighborhood enrichments around RefSeq TSSs and TESs.

which is generally considered to be associated with gene enhancers (Roadmap Epigenomics Consortium 2015), but may also be associated with other functionality (Cui *et al.* 2009, Cheng *et al.* 2014). In state 16, the H3K4me1 and H3K27me3 emissions are indicative of a bivalent enhancer state. According to the overlap enrichment of genomic features (Fig. 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, suggesting 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 upon the rationale that state 19 has no emission in any of the histone marks, while covering the greatest percentage of the genome. Based on a combination of histone mark emissions probabilities (Fig. 3A), enrichment in genomic features (Fig. 3B and C) and comparison with published chromatin states (Roadmap Epigenomics Consortium 2015), we have labeled the states with biologically meaningful labels (Fig. 3A). Furthermore, when the levels of DNA methylation were measured, we found that the active TSS state (state 1) had, as expected, the lowest level of methylation across chromatin states, which was consistent across all samples measured (Supplementary Fig. 1). The chromatin state segmentations can be viewed

on the UCSC Genome Browser through a link provided in www.bcgsc.ca/data/thyroid (Fig. 1).

Stability of chromatin states

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. We found that promoter (state 1), transcribed (states 5 and 7), and quiescent (state 19) states were consistently marked across the normal thyroid epigenomes of four individuals (Fig. 4A and B). Strikingly other chromatin states were highly specific for an individual (Fig. 4). Furthermore, we found the epigenetic consistency is reduced in the other states, and the states lacking the most agreement across specimens are regions flanking downstream of TSS (states 4) and repeats associated with artifacts (state 17) (Fig. 4C).

Epigenetically marked promoters and relation with gene expression

The promoter state labeled as active TSS (state 1) was found to be the most epigenetically consistent state

Figure 4

Overview of epigenetic consistency across 4 thyroid epigenomes. The genome was divided into 15,181,508 bins. Each bin is 200 bp 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 four epigenomes, then the bin count for state 1 at $x=4$ is incremented. If the states for a bin across four epigenomes 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 four 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.

(Fig. 4C). 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 probability of 57, 19, 13 and 11% of also being partitioned as state 1 in three, two, one and zero other epigenomes, respectively (Fig. 4C).

We next associated bins partitioned as state 1 to genes if the bin is within a gene's promoter (defined as TSS $\pm$ 1kbp). A majority of state 1 bins (77.4%) were found within protein-coding gene promoters (Fig. 5A). 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 with bins partitioned as state 1 in at least one epigenome and 10,460 known protein coding genes to bins partitioned as state 1 across all four epigenomes (Fig. 5B).

We next grouped the genes by the epigenetic consistency of state 1 in gene promoters and compared their levels 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 epigenomes will have higher expression than genes that are not epigenetically active in any epigenome. When we grouped expression by the number of epigenetically active promoters shared across epigenomes, we found that indeed the expression tends to be higher in genes

partitioned as epigenetically active in more epigenomes (Fig. 5C and D; this behavior is also the case in the other samples). Specifically, expression is on average 9.7-fold higher in genes characterized as epigenetically active than genes not characterized as epigenetically active in any epigenome. Furthermore, expression is on average 4.4-fold higher in genes that are epigenetically active across all four epigenomes than genes that are epigenetically active in only one epigenome. Similarly, when we grouped genes into different brackets of expression, we found that genes with high expression tend to be epigenetically active in all epigenomes (Fig. 5E). We also find that 90.9% of genes with expression between 100 and 1000 TPM is epigenetically active in all epigenomes and this proportion drops to 44.3% for genes with expression between 1 and 10 TPM and 7.9% for genes with expression between 0.1 and 1 TPM (Fig. 5E).

Enhancers

Chromatin states characterized as enhancers (states 8–11) were less consistent than states characterized as promoters (Fig. 4). Nevertheless, we find regions epigenetically consistent across all thyroid specimens for genic (state 8 and 9), active (state 10) and weak (state 11) enhancer type chromatin states. Sequence analysis of the genomic DNA in regions shared by the four samples (2527 regions for state 8; 4663 for state 9; 22,604 for state 10; and 9463 for state 11) indicate that

Figure 5

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 $\pm$ 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 no expression were removed. (D) Expression (log10-scaled, values from CEMT_44) across genes that are epigenetically active in 0, 1, 2, 3, and 4 epigenomes. Genes with no 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 epigenomes sharing the same genomic bin.

the NF1 response element (CYTGGCABNSTGCCAR) was the most overrepresented sequence motif in enhancer states 8, 10 and 11. Other transcription factor response elements common across enhancer states 8, 10 and 11 are TLX (CTGGCAGSCTGCCA), PAX8 (GTCATGCHTGRCTGS) and PAX5 (GCAGCCAAGCRTGACH). In the literature, PAX8 has been 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 given its high expression in papillary carcinomas, follicular adenomas, follicular carcinomas and 79% of anaplastic carcinomas (Nonaka *et al.* 2008). The top 3 motifs of each enhancer chromatin state are shown in (Table 1).

Thyroid transcript abundance

With regards to estimating transcript abundances, we found that the most highly expressed transcripts, representing 95% of the protein coding RNA-seq reads, are made up of on average 7194 top genes and the top 10,000 genes account for an average of 98% of detected transcript reads (Fig. 6). Across the four specimens, the top 25 most highly expressed genes (accounting for an average of 19% of transcripts) collectively consists of 42 unique protein coding genes and 10 of these genes are consistent across the four specimens (Table 2). Furthermore, motif analysis of active enhancers around these genes is described in the [Supplementary Materials](#).

Epigenetically active and consistently expressed genes in the thyroid

To further characterize the thyroid, we identified a set of genes that were likely highly relevant for 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 specimens are considered to be under stringent transcriptional control. We consider a gene as epigenetically active if a bin within the gene promoter is partitioned as state 1. Previously, we found 13,175 genes to be epigenetically active in at least one epigenome and 10,460 genes to be epigenetically active across all four epigenomes (Fig. 5B). We considered a gene as consistently expressed if (1) it was within the intersection of the top 2000 most highly expressed gene in each specimen and (2) it is in the set of 2000 genes with the lowest variance across the normal specimens. Overall, the 2000 most highly expressed genes have a minimum expression of 29 TPM and accounted for an average of 76% of the protein-coding RNA-seq transcripts. Within the top 2000 genes from each of the four specimens, there was a total of 3024 genes and the intersection defined 1183 genes across the four specimens. Intersecting the set of 10,460 genes that are epigenetically active across all four epigenomes, 1183 genes that have high expression, and 2000 genes with low variance, we arrived at a set of 137 genes (Fig. 7A). Examining this set of genes using Metascape (Tripathi *et al.* 2015), we find predominantly general processes such as metabolic processes, protein folding, transport and secretion (Fig. 7B). The top 3 Gene Ontology (GO)

Table 1 Top 3 motifs enriched in genomic DNA epigenetically consistent at enhancers type chromatin states.

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	NGTGTCAVTSAGCGKAAA	−84.0

States 8 & 9=genic enhancers, 10=active enhancer, and 11 weak enhancers. TF stands for transcription factor. Motif enrichment was performed using HOMER software, state 9 has enrichment in only 1 motif, and Benjamini corrected *P*-values<0.03. The list of all motifs (and corrected *P*-values) are available in the supplements.

Figure 6

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 specimen. The average transcript proportion by gene rank were computed across 4 thyroid specimens and is shown by the curved line. The gray ribbon is the mean proportion of transcripts $\pm$ 2 standard deviations.

terms are RNA localization (GO:0006403), protein folding (GO:0006457) and negative regulation of cell death (GO:0060548). To prioritize the list of 137 genes, we used the Genome-Tissue Expression (GTEx) project to filter out genes expressed (FPKM $\geq$ 10) in 52 non-thyroid tissues (Supplementary Fig. 2). This left 18 genes that we consider epigenetically active and consistently expressed in the thyroid that are likely highly relevant to thyroid function (Table 3).

Discussion

This study has generated the first high quality, published and deeply sequenced reference epigenomes for human thyroid tissue. Each reference epigenome has a complete set of six histone marks (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3) profiled with ChIP-seq, a complete bisulfite converted methylome, a transcriptome and a matched genomic sequence. From these reference epigenomes, we characterized the normal epigenomes into 19 chromatin states and compared the consistency of chromatin state annotations of different

Table 2 The 25 most abundant protein coding gene transcripts in each human specimen.

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

The mean and standard deviation (s.d.) are summary statistics of the proportion of transcripts across the four specimens. In total, there are 42 unique genes. Genes with an asterisk (*) represent genes not epigenetically active (i.e. labeled as active TSS state 1 in the same bin) across 4 specimens ($n=6$), while those without an asterisk represent genes that are epigenetically active across 4 specimens ($n=36$).

Figure 7

137 epigenetically active and consistently expressed genes in the thyroid. (A) Epigenetically active and consistently expressed genes were identified based on criteria as follows: TSS 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.

individuals. We found that some states, such as active TSS (state 1) and transcription (state 5), were more epigenetically consistent and stable than others. 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 cell lines and found that their active promoter and transcribed chromatin states highly coincided with the active promoter and transcribed chromatin state annotations of other cell lines. Furthermore, the quiescent state (state 19) remained largely unchanged across epigenomes, while every other state tended to be variable between epigenomes. Although the epigenetic state, consistent with active promoters showed high levels of consistency between samples, this was not observed for the majority of the remaining states. We also examined whether the differential modification of repressive histone marks (H3K9me3 and H3K27me3) between samples was also correlated with differential DNA methylation at these sites. As indicated by Supplementary Fig. 3, we observed that differential histone modification at these sites did not obviously correlate with differential DNA methylation. This lack of consistency in what should be identical tissues has not been previously characterized. It is not clear whether this is a unique feature of the thyroid or terminally differentiated tissues. It is possible that the states are much more consistent in developing and pluripotent cells where gene regulation may need to be under more stringent control.

Using HMMs to partition the epigenome into chromatin states is reliant on the number of hidden states available for partitioning, and different number of hidden states produces different models. There are various methods for model selection. Two popular model selection methods are the Bayesian information criterion (BIC) and the Akaike information criterion (AIC), but a

weakness of these selection methods is that they tend to favor higher number of states that are biologically more difficult to distinctly interpret and does not capture sufficiently distinct interactions. Another metric for selecting the number of states produced by ChromHMM is the Factorized Information Criterion (FIC) proposed by Hamada *et al.* (2015). However, FIC-HMM indicated more estimated chromatin states than what was selected for by the original ChromHMM analysis done by Ernst *et al.* (2011) and thus are again biologically more difficult to distinctly interpret. In comparison, the number of states chosen in Roadmap Epigenomics Consortium (2015) was based upon the manual consideration on evaluation for the number of states, which capture all key interactions between chromatin marks. As a result, Roadmap used 15 states for 5 marks and 18 states for 6 marks. Similarly, the 25 states presented in Hoffman *et al.* (2013) was selected 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. In our study, we devised a novel quantitative selection metric that will allow rapid assessment for the optimal number of states (Supplementary Materials and methods). Overall, we partitioned the thyroid epigenome into 19 states.

In state 15 (labeled as 'repressed'), we find emission of H3K9me3 and H3K27me3 (Fig. 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

Table 3 GO Biological Process annotation of 18 actively transcribed and consistently expressed genes in the thyroid that do not have high expression in 52 non-thyroid GTEx tissues.

Gene	Description	GO biological process (from Metascape)
<i>DEPTOR</i>	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
<i>ETFB</i>	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
<i>FXR1</i>	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
<i>H2AFY</i>	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
<i>N4BP2L2</i>	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
<i>NSMCE1</i>	NSE1 homolog, SMC5-SMC6 complex component	GO:2001022 positive regulation of response to DNA damage stimulus; GO:0006301 postreplication repair; GO:0016925 protein sumoylation
<i>NT5C2</i>	5'-nucleotidase, cytosolic II	GO:0046085 adenosine metabolic process; GO:0006195 purine nucleotide catabolic process; GO:0046040 IMP metabolic process
<i>PMF1</i>	Polyamine modulated factor 1	GO:0007062 sister chromatid cohesion; GO:0000819 sister chromatid segregation; GO:0098813 nuclear chromosome segregation
<i>SCAF11</i>	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
<i>SNF8</i>	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
<i>SORD</i>	Sorbitol dehydrogenase	GO:0006062 sorbitol catabolic process; GO:0051160 L-xylitol catabolic process; GO:0019640 glucuronate catabolic process to xylulose 5-phosphate
<i>SPG11</i>	Spastic paraplegia 11 (autosomal recessive)	GO:0048675 axon extension; GO:0008088 axo-dendritic transport; GO:1990138 neuron projection extension
<i>TCTN1</i>	Tectonic family member 1	GO:0021956 central nervous system interneuron axonogenesis; GO:0021523 somatic motor neuron differentiation; GO:0021955 central nervous system neuron axonogenesis
<i>TOR1AIP1</i>	Torsin 1A interacting protein 1	GO:0071763 nuclear membrane organization; GO:0032781 positive regulation of ATPase activity; GO:0043462 regulation of ATPase activity
<i>TPD52</i>	Tumor protein D52	GO:0030183 B cell differentiation; GO:0030098 lymphocyte differentiation; GO:0042113 B cell activation
<i>TPGS2</i>	Tubulin polyglutamylase complex subunit 2	
<i>VEZT</i>	Vezatin, adherens junctions transmembrane protein	GO:0016337 single organismal cell-cell adhesion; GO:0098602 single organism cell adhesion; GO:0098609 cell-cell adhesion
<i>WBSCR22</i>	Williams-Beuren syndrome chromosome region 22	GO:0031167 rRNA methylation; GO:0000154 rRNA modification; GO:0001510 RNA methylation

progenitor differentiation program (Liu *et al.* 2015) and in a cooperative mechanism in maintaining silencing whereby H3K27me3-bound PRC2 stabilizes H3K9me3-anchored HP1A (Boros *et al.* 2014). In another study, it was suggested that the antibody used to enrich H3K27me3 has off-target enrichment for H3K9me3 (Peach *et al.* 2012). According to our observations (Fig. 4A and B), the stability of this chromatin state across four epigenomes is low and that out of all bins partitioned as state 15, only 3.0% are shared across four epigenomes. Similarly, a bin partitioned as state 15 has a 9% probability of finding the same state in the same bin across three

other epigenomes (Fig. 4C). The lack of conservation of state 15 between epigenomes leads to the question as to whether it 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 heterochromatin state 14 to state 15 and from state 15 to itself, state 14 and quiescent state 19 (Fig. 3A). These observations suggest that regions containing both H3K9me3 and H3K27me3 may be an intermediate state from heterochromatin to quiescent states.

Overall, 10,460 genes were found to have epigenetically active promoters across all four epigenomes,

1014 genes across three epigenomes, 754 genes across two epigenomes, 947 genes across one epigenome and 6979 genes across no epigenomes (Fig. 5B). It is striking that in a relatively homogeneous tissue such as the thyroid gland, whose main function is to produce thyroid hormone, approximately half of the known protein coding genes have epigenetically active promoters in all four specimens.

With regards to the set of 18 epigenetically active and consistently expressed genes of the thyroid (Table 3), when we perform a gene set enrichment analysis (Tripathi *et al.* 2015), no terms were found enriched. In Table 3, we present the GO annotation of the individual genes. Interestingly, thyroglobulin (TG), the thyroid prohormone, is excluded from this list of 18 genes that are highly relevant to thyroid function. TG was also not the highest expressed gene in all four specimens (Table 2) and showed high variability, ranging from 8000–20,000 TPM. From the 18 genes, *ETFB*, *NTSC2*, *SNF8*, *SORD* and *TOR1AIP1* appear to be related to metabolism, *N4BP2L2* to blood and *TPD52* to the immune system (Table 2). In the literature, spatacsin, encoded by *SPG11*, was identified to play critical roles in autophagic lysosome reformation, a pathway that generates new lysosomes (Chang *et al.* 2014) and *TPD52* has been predicted to regulate endolysosomal trafficking in secretory cell types (Byrne *et al.* 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 *et al.* (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). Thus, we suggest the binding of FXR1 with RXR could influence transcription of thyroid-responsive genes. With regards to *PMF1*, *H2AFY*, *NSMCE1*, *SCAF11*, *TCTN1*, *TPGS2*, *VEZT* and *WBSR22*, we did not find any reports linking these genes with the thyroid, which may suggest potential significance of these genes in the thyroid.

In conclusion, we characterized the normal thyroid epigenome into 19 chromatin states and compared the epigenetic features across four unique thyroid specimens. In general, normal thyroid tissue from non-pathologic human thyroid glands is challenging to obtain for study. However, in spite of the limitation of the specimens being microscopically normal thyroid tissue from thyroid glands with areas of pathology, we defined and found a set of epigenetic features conserved across different individual specimens. We found that epigenetic features characterizing promoters and transcription elongation tend to be more consistent. We also found that every other epigenetic feature tends to be more variable across four individuals, highlighting differences between individuals and the need for biological replicates when deriving reference epigenomes. Furthermore, we found that genes epigenetically active across all epigenomes tend to have higher expression than genes not consistently epigenetically active and we identified a set of 18 genes that are epigenetically active and consistently expressed by the thyroid. Overall, we developed a novel quantitative model selection metric and believe the epigenomes presented in this report represent a valuable resource that will allow for the development of a deeper understanding of the molecular biology that underlies thyroid function and provides important contextual epigenetic information for comparison and integration into future studies.

Supplementary data

This is linked to the online version of the paper at <http://dx.doi.org/10.1530/JOE-17-0145>.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This work was supported by Genome British Columbia and the Canadian Institutes of Health Research as part of the Canadian Epigenetics, Environment and Health Research Consortium Network (grant numbers EP1-120589 and EP2-120591); and the CIHR Foundation Scheme (grant number FDN-143288). C Siu was supported by the CIHR Bioinformatics Training Program for Health Research and the Canada Graduate Scholarships-Master's Program.

Acknowledgements

Aligned RNA-sequencing and ChIP-sequencing bam files were provided through the author's participation in the Canadian Epigenetics, Environment and Health Research Consortium.

References

- Boros J, Arnoult N, Stroobant V, Collet JF & Decottignies A 2014 Polycomb repressive complex 2 and H3K27me3 cooperate with H3K9 methylation to maintain heterochromatin protein 1 α at chromatin. *Molecular and Cellular Biology* **34** 3662–3674. (doi:10.1128/MCB.00205-14)
- Byrne JA, Frost S, Chen Y & Bright RK 2014 Tumor protein D52 (TPD52) and cancer -oncogene understudy or understudied oncogene? *Tumor Biology* **35** 7369–7382. (doi:10.1007/s13277-014-2006-x)
- Catena V & Fanciulli M 2017 Deptor: not only a mTOR inhibitor. *Journal of Experimental and Clinical Cancer Research* **36** 12. (doi:10.1186/s13046-016-0484-y)
- Chang J, Lee S & Blackstone C 2014 Spastic paraplegia proteins spatzin and spatacsin mediate autophagic lysosome reformation. *Journal of Clinical Investigation* **124** 5249–5262. (doi:10.1172/JCI77598)
- Cheng J, Blum R, Bowman C, Hu D, Shilatfard A, Shen S & Dynlacht BD 2014 A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Molecular Cell* **53** 979–992. (doi:10.1016/j.molcel.2014.02.032)
- 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** 867–878. (doi:10.1016/j.bbdis.2010.12.021)
- Cui K, Zang C, Roh TY, Schones DE, Childs RW, Peng W & Zhao K 2009 Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **5** 80–93. (doi:10.1016/j.stem.2008.11.011)
- Eladio NA & Gershon MD 1978 Histochemical studies of mammalian thyroid parafollicular cells. Distribution and number. In *International Review of Cytology*, vol 52, pp 12–14. Eds GH Bourne & JF Danielli. New York, NY, USA: Academic Press, Inc.
- Ernst J & Kellis M 2012 ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9** 215–216. (doi:10.1038/nmeth.1906)
- Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011 Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473** 43–49. (doi:10.1038/nature09906)
- Führer D, Brix K & Biebermann H 2015 Understanding the healthy thyroid state in 2015. *European Thyroid Journal* **4** 1–8. (doi:10.1159/000431318)
- Gharib H & Papini E 2007 Thyroid nodules: clinical importance, assessment, and treatment. *Endocrinology Metabolism Clinics of North America* **36** 707–735. (doi:10.1016/j.ecl.2007.04.009)
- González-Peñas J, Amigo J, Santomé L, Sobrino B, Brenlla J, Agra S, Paz E, Páramo M, Carracedo Á, Arrojo M, et al. 2016 Targeted resequencing of regulatory regions at schizophrenia risk loci: role of rare functional variants at chromatin repressive states. *Schizophrenia Research* **174** 10–16. (doi:10.1016/j.schres.2016.03.029)
- Hamada M, Ono Y, Fujimaki R & Asai K 2015 Learning chromatin states with factorized information criteria. *Bioinformatics* **31** 2426–2433. (doi:10.1093/bioinformatics/btv136)
- Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H & Glass CK 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38** 576–589. (doi:10.1016/j.molcel.2010.05.004)
- Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA & Noble WS 2012 Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9** 473–476. (doi:10.1038/nmeth.1937)
- Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. 2013 Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research* **41** 827–841. (doi:10.1093/nar/gks1284)
- 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. *Genetics and Molecular Research* **15**. (doi:10.4238/gmr.15038992)
- Liu J, Magri L, Zhang F, Marsh NO, Albrecht S, Huynh JL, Kaur J, Kuhlmann T, Zhang W, Slesinger PA, et al. 2015 Chromatin landscape defined by repressive histone methylation during oligodendrocyte differentiation. *Journal of Neuroscience* **35** 352–365. (doi:10.1523/JNEUROSCI.2606-14.2015)
- Love MI, Huber W & Anders S 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15** 550. (doi:10.1186/s13059-014-0550-8)
- 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** 192–200. (doi:10.1038/modpathol.3801002)
- Panicker V 2011 Genetics of thyroid function and disease. *Clinical Biochemist Reviews* **32** 165–175.
- Patro R, Duggal G, Love MI, Irizarry RA & Kingsford C 2017 Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14** 417–419. (doi:10.1038/nmeth.4197)
- Peach SE, Rudomin EL, Udesi ND, Carr SA & Jaffe JD 2012 Quantitative assessment of chromatin immunoprecipitation grade antibodies directed against histone modifications reveals patterns of co-occurring marks on histone protein molecules. *Molecular and Cellular Proteomics* **11** 128–137. (doi:10.1074/mcp.m111.015941)
- Pellacani D, Bilenky M, Kannan N, Heravi-Moussavi A, Knapp DJHF, Gakkhar S, Moksa M, Carles A, Moore R, Mungall AJ, et al. 2016 Analysis of normal human mammary epigenomes reveals cell-specific active enhancer states and associated transcription factor networks. *Cell Reports* **17** 2060–2074. (doi:10.1016/j.celrep.2016.10.058)
- Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015 Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330. (doi:10.1038/nature14248)
- Soneson C, Love MI & Robinson MD 2015 Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4** 1521. (doi:10.12688/f1000research.7563.1)
- Stunnenberg HG, The International Human Epigenome Consortium & Hirst M 2016 The International Human Epigenome Consortium (IHEC): a blueprint for scientific collaboration and discovery. *Cell* **167** 1145–1149. (doi:10.1016/j.cell.2016.11.007)
- Tripathi S, Pohl MO, Zhou Y, Rodriguez-Frandsen A, Wang G, Stein DA, Moulton HM, Dejesus P, Che J, Mulder LCF, et al. 2015 Meta- and orthogonal integration of influenza 'OMICs' data defines a role for UBR4 in virus budding. *Cell Host and Microbe* **18** 723–735. (doi:10.1016/j.chom.2015.11.002)
- Trueba S, Auge J, Mattei G, Etchevers H, Martinovic J, Czernichow P, Vekemans M, Polak M & 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** 455–462. (doi:10.1210/jc.2004-1358)

Received in final form 2 August 2017

Accepted 14 August 2017

Accepted Preprint published online 14 August 2017