A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types

Maxwell W. Libbrecht*
Department of Computer Science and Engineering
University of Washington

Oscar Rodriguez*
Department of Genetics and Genomic Sciences
Icahn School of Medicine at Mount Sinai

Zhiping Weng
Program in Bioinformatics and Integrative Biology
University of Massachusetts Medical School

Jeffrey A. Bilmes
Department of Electrical Engineering
University of Washington

Michael M. Hoffman
Princess Margaret Cancer Centre
Department of Medical Biophysics
Department of Computer Science
University of Toronto

William S. Noble
Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington

November 5, 2016

Abstract

Semi-automated genome annotation methods such as Segway enable understanding of chromatin activity. Here we present chromatin state annotations of 164 human cell types using 1,615 genomics data sets. To produce these annotations, we developed a fully-automated annotation strategy in which we train separate unsupervised annotation models on each cell type and use a machine learning classifier to automate the state interpretation step. Using these annotations, we developed a measure of the functional importance of each genomic position called the "functionality score", which allows us to aggregate information across cell types into a multi-cell type view. This score provides a measure of importance directly attributable to a specific activity in a specific set of cell types. In contrast to evolutionary conservation, this measure is not biased to detect only elements shared with related species. Using the functionality score, we combined all our annotations into a single cell type-agnostic encyclopedia that catalogs all human functional regulatory elements, enabling easy and intuitive interpretation of the effect of genome variants on phenotype, such as in disease-associated, evolutionary conserved or positively selected loci. These resources, including cell type-specific annotations, enyclopedia and a visualization server, are publicly available online at http://noble.gs.washington.edu/proj/encyclopedia/.

*These authors contributed equally

1 Introduction

Sequencing-based genomics assays can measure many types of genomic biochemical activity, including transcription factor binding, chromatin accessibility, transcription, and histone modification. The scale and complexity of these data sets require integrative analysis with computational methods. A class of methods known as *semi-automated genome annotation* (SAGA) algorithms are widely used to perform such integrative modeling of diverse genomics data sets. These algorithms take as input a collection of genomics data sets and simultaneously partition the genome and label each segment with an integer state index such that positions with the same label have similar patterns of activity. These methods are "semi-automated" because a human performs a functional interpretation of the states after the annotation process. Examples of SAGA methods include HMMSeg (Day et al., 2007), ChromHMM (Ernst and Kellis, 2010), Segway (Hoffman et al., 2012) and others (Thurman et al., 2007; Lian et al., 2008; Filion et al., 2010; Lystig and Hughes, 2002; Schliep et al., 2003; Jiang et al., 2008; Mammana and Chung, 2015). These genome annotation algorithms have had great success in interpreting genomics data and have been shown to recapitulate known functional elements including genes, promoters and enhancers.

The wide availability of genomics data sets necessitates the development of SAGA strategies that scale to many cell types. The primary strategy previously used to annotate multiple cell types has been *concatenated* annotation, in which a shared model is trained across all cell types (Sheffield et al., 2013; Ho et al., 2014; Kundaje et al., 2015; Sohn et al., 2015; Zerbino et al., 2015). However, concatenated annotation has two limitations. First, it requires that all cell types have exactly the same data sets available. Second, it is very sensitive to artifactual differences between cell types, resulting in bias for or against each label. Later methods that share information across cell types—such as imputing results of unperformed assays (Ernst and Kellis, 2015), joint annotation (Biesinger et al., 2013; Zhang et al., 2016) and graph-based regularization (Libbrecht et al., 2015)—result in improved accuracy at identifying functional elements, but sharing information in this way complicates the interpretation of the resulting annotations.

To avoid these limitations, in this work we train one model separately for each cell type (Figure 1). This strategy allows us to use all available data in every cell type and removes the potential for issues resulting from experimental artifacts. Using this approach, we were able to annotate 164 human cell types using a total of 1,615 genomics data sets. In contrast, a concatenated approach applied to these data can use at most 570 data sets, which is achieved by annotating 114 cell types with a panel of five assay types each.

In order to fully automate the annotation process, we developed a method for automating the state interpretation step. Previously, performing separate annotations was impeded by the need to manually interpret the states learned by each trained model. We use a machine learning classifier that interprets each state using the information typically used in manual interpretation. We trained this classifier using preexisting human-interpreted annotations. This classifier allows the annotation process to proceed from raw data to final output in a fully automated way.

To enable easier understanding of these annotations, we propose a measure of the functional importance of each position, called the *functionality score*. The functionality score is defined based on the enrichment of each annotation state for evolutionary conservation, and therefore separates functional activity (such as promoters and enhancers) from non-functional activity (repressed regions). The functionality score provides a measure of importance that is directly attributable to a specific activity in a specific set of cell types. In addition, because the functionality score locally depends only on chromatin state, it is not biased to detect only elements shared with related species. Therefore, the functionality score is an important orthogonal measure of importance to conservation.

The functionality score also enables a new way of visualizing the activity of a locus across cell types that we call a *functionality score plot*. This plot emphasizes functional over non-functional activity, simultaneously displaying the putative importance of each genomic position as well as what type of activity is responsible for this importance. We have set up a publicly available server where a user can produce a functionality score visualization of any target genomic locus (http://noble.gs.washington.edu/proj/encyclopedia).

Finally, we combine our cell type-specific annotations to produce a single, cell type-agnostic encyclopedia using the functionality score. Past annotations simply characterize biochemical activity, producing a separate annotation for each cell type. However, when a researcher or clinician is interested in a given locus—for example, when studying a disease variant—they often do not know which cell type is most relevant. These users are often more interested in a cell type-agnostic view of genome function, such as the gene annotations

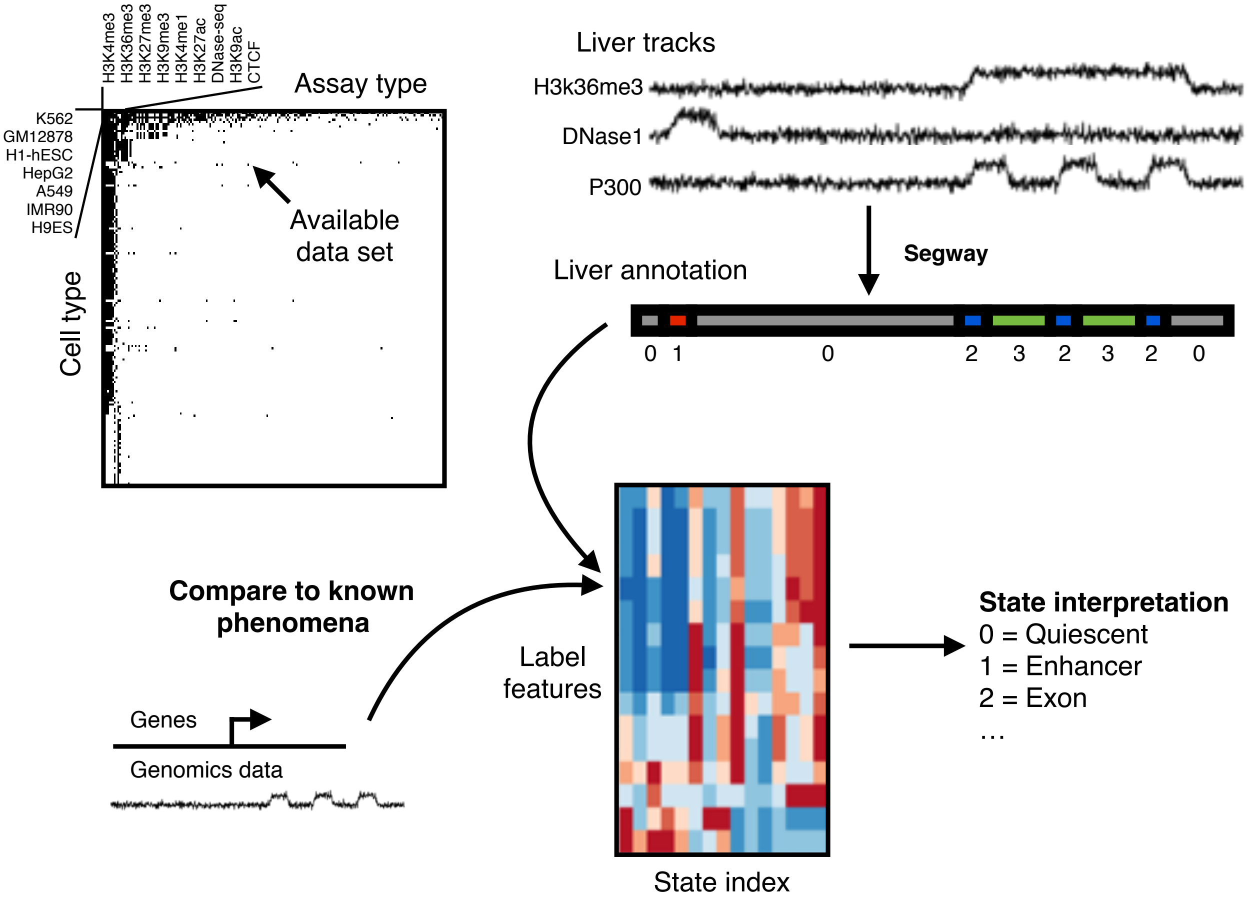


Figure 1: Schematic of annotation pipeline (Methods).

produced by Ensembl (Yates et al., 2016) or GENCODE (Harrow et al., 2006). In this view, there is a common, cell type-agnostic set of elements (genes or regulatory elements), and each element is annotated with its pattern of activity across cell types. We call this type of annotation an “encyclopedia” to distinguish it from a cell type-specific annotation and because it represents the goal of ENCODE (Encyclopedia of DNA Elements). We produce an encyclopedia of regulatory elements by collecting all high-functionality score segments in the genome and labeling each with its pattern of activity across cell types. This encyclopedia catalogs all measured human regulatory elements, enabling easy and intuitive interpretation of the effect of genome variants on phenotype, such as loci that are disease-associated, evolutionarily conserved or under positive selection.

2 Results

2.1 Annotation of 164 human cell types

We obtained all available ChIP-seq, DNase-seq and Repli-seq data from the ENCODE and Roadmap Epigenomics consortia (Figure 1, Supplementary File 1). Because we are interested in transcriptional regulation, we excluded measurements of post-transcriptional activity, such as RNA-seq, CAGE and RNA-binding protein assays. We also excluded measures of methylation because they are defined on only a subset of the genome (i.e. CpG loci) and 3C-based assays of chromatin conformation because they cannot be directly represented as a genomic signal track. We chose to annotate every cell type with sufficient data; specifically, we annotated any cell type that has either (1) five histone modification data sets or (2) at least one data set each from two of the following categories: histone ChIP-seq, transcription factor ChIP-seq or

Label	Description	Characteristic activity
Quiescent	Inactive region	None
ConstitutiveHet	Constitutive heterochromatin	H3K9me3, HP1
FacultativeHet	Facultative heterochromatin	H3K27me3, PRC
Transcribed	Transcribed region	H3K36me3
Promoter	Promoter	H3K4me3, H3K27ac
Enhancer	Enhancer	H3K4me1, H3K27ac, P300
RegPermissive	Region with weak marks of regulation	H3K4me1
Bivalent	Regulatory element with marks of both activation and repression	H3K27ac, H3K27me3
LowConfidence	Placeholder assigned to labels that were not classified with high confidence	–
Unclassified	Placeholder assigned to reference labels that do not fit the above vocabulary	–

Table 1: Description of each label type.

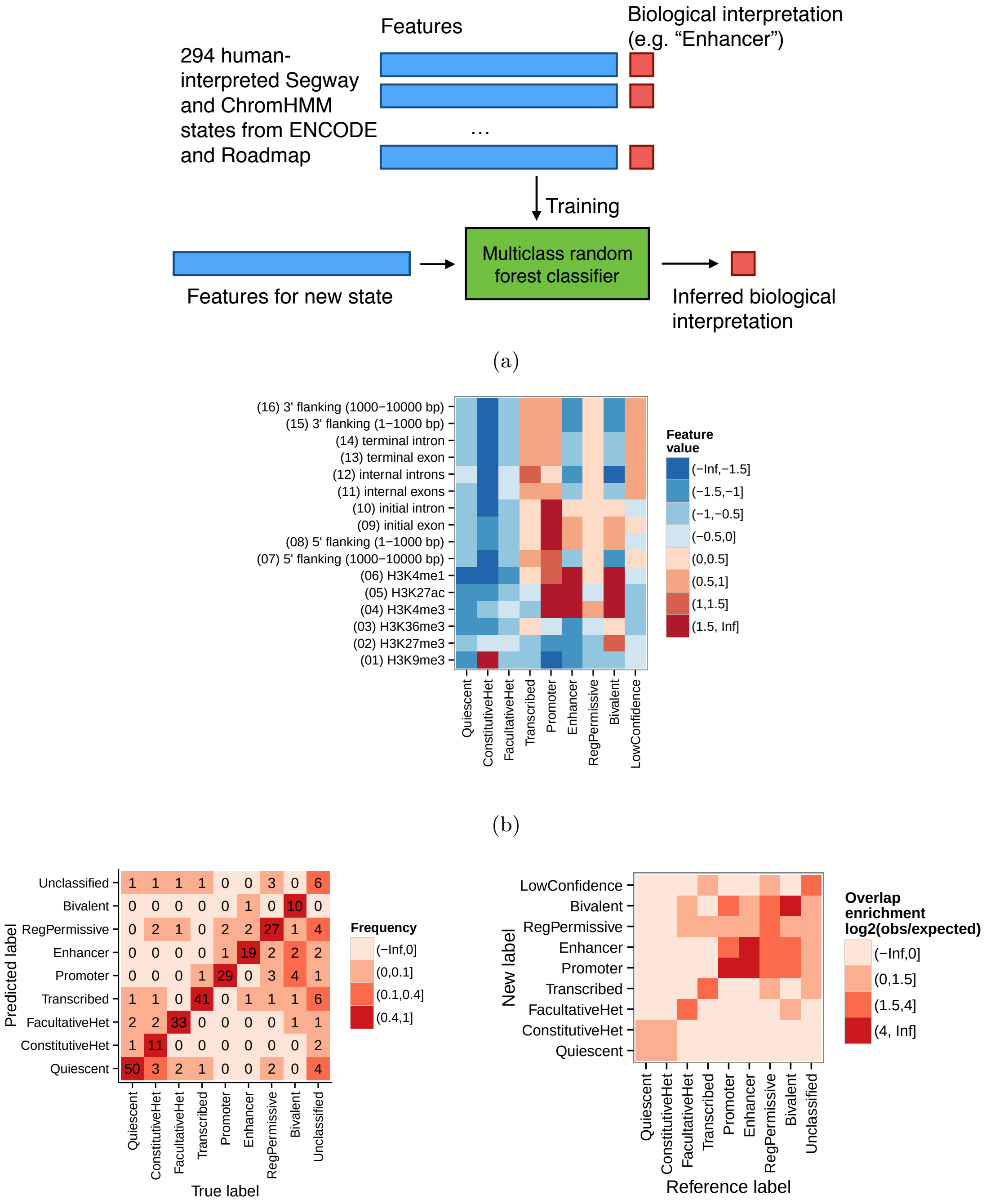
DNA accessibility.

We used the SAGA method Segway to annotate each cell type (Methods, Hoffman et al. (2012, 2013)). Segway is based on a dynamic Bayesian network (DBN) model. In order for the number of states to scale with the amount of data, we used the formula $(10 + 2 \cdot \sqrt{\text{number of tracks}})$ to determine the number of labels for a given cell type, which is roughly in line with previous SAGA annotations. We implemented two improvements to Segway for this work. First, we used a mixture of three Gaussians as the emission model for each track, as opposed to the single Gaussian distribution used in previous Segway analysis. A mixture of three Gaussians provides a more flexible model and therefore is more capable of capturing complex phenomena. Second, we applied a *mini-batch* training strategy. Previously, Segway training has been performed on a fraction of the genome (usually the ENCODE pilot regions, which cover 1% of the human genome). This procedure speeds up training, but it means that the training procedure has access to only a fraction of the available data. In this work, we instead applied training to a different random 1% of the genome (a "mini-batch") at each iteration. This strategy maintains fast training, while allowing the algorithm to access all the available data. Mini-batch training is frequently used in optimization (Murphy, 2012).

2.2 A machine learning approach recapitulates manual interpretation of annotation labels

Previously, SAGA annotations have generally been interpreted manually, but doing so individually for all 164 annotations would be impractical. To solve this problem, we developed a machine learning framework that automates state interpretation (Methods, Figure 2a). We did this by training a random forest classifier to recapitulate human interpretation, using existing interpreted SAGA annotations as training data. For each state, we derived a set of 17 features that encompass the information that has typically been used to interpret these states in the past, and used these features as the input to the classifier. We collected 10 existing Segway or ChromHMM annotations and manually interpreted four more from this work, for a total of 294 manually interpretated annotation states. We curated the biological interpretations of these states into a unified vocabulary of eight categories (listed in Table 1, and described in detail in the next section). We assigned the placeholder label "Unclassified" to 26/294 reference labels that do not fit into one of the eight categories. We trained the random forest classifier on the 294 reference states, then applied the classifier to each state from our 164 new annotations to obtain a biological interpretation of each new state. Note that a single training example in this framework corresponds to thousands of segments. We assigned the label "LowConfidence" to new states that the classifier predicted as Unclassified and to states not assigned with more than 25% confidence to any of the other categories.

This classifier recapitulated human interpretation very accurately (Figure 2c). Using a leave-one-out cross-validation strategy, the classifier achieved an accuracy of 226/294 (77%; 19% expected by chance). Moreover, most errors either involved the "Unclassified" placeholder meaning (33/70 errors) or involved mistaking similar types of activity for one another. The classifier based its assignments on the expected



(a)

(b)

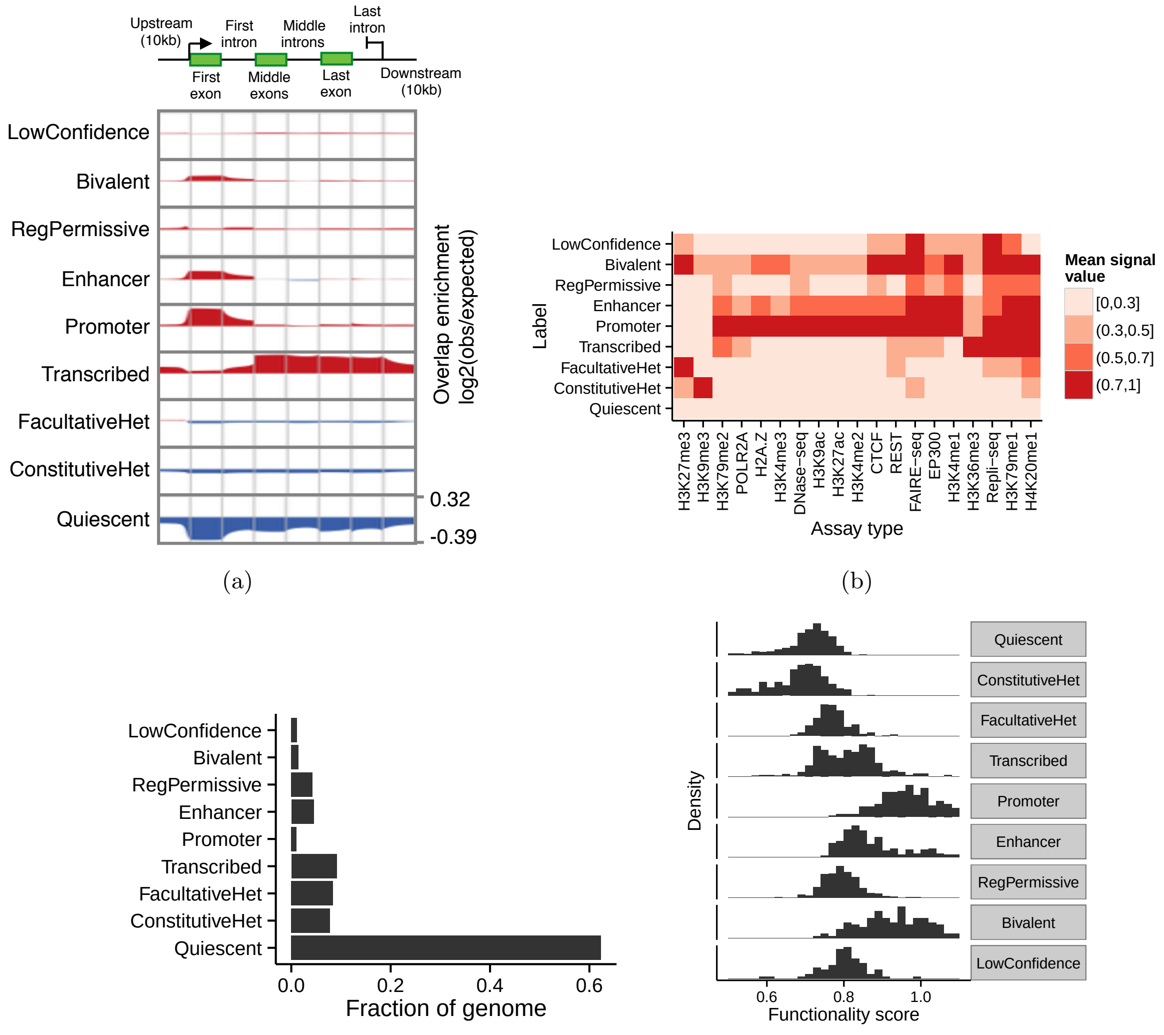
(c)

(d)

Figure 2: Results of label interpretation classifier. (a) Schematic of machine learning-based automatic classification strategy. (b) Association of label categories and classifier features. Color indicates mean feature value (standard deviation units). (c) Label classification confusion matrix. Numbers and colors indicate the number of reference labels with a particular category assigned to a predicted category by the classifier under leave-one-out cross-validation. Classifications off of the diagonal indicate mis-classifications. (d) Overlap enrichment of reference annotations with our annotations, in the cell types that have a reference annotation. Numbers and colors indicate the enrichment, calculated as the log2 of the number of bases that overlap between a given reference and new label category, divided by the number expected if the labels were distributed independently. Note the difference between (c) and (d): (c) measures whether the interpretation classifier assigns the same category as the reference annotation for a fixed label, whereas (d) measures the genomic similarity of two entirely separate genome annotations.

feature patterns (see next section; Figure 2b). Where one of our cell types was previously annotated by another SAGA effort, the annotations largely match (Figure 2d).

2.3 Annotations accurately recover known genome biology



(a)

(b)

(c)

(d)

Figure 3: Relationship of annotations to known genomic elements. (a) Enrichment of each label over an idealized gene. We calculated enrichment as the base-2 logarithm of the observed frequency of a label at a particular position along an annotation divided by the expected frequency of the label from its prevalence in the genome overall. Enriched positions are shown in red, and depleted positions are shown in blue. (b) Relationship of labels to selected input data sets. Color corresponds to the mean signal value of a given assay type at positions annotated with a given label, aggregated over cell types where the given assay type is available. Values are normalized such that the maximum and minimum in each column are 1 and 0, respectively. (c) Fraction of the genome covered by each label. (d) Distribution of functionality scores for states of each type.

Our classification strategy assigns each state to one of nine categories (Table 1). Our categories are largely consistent with previous annotation efforts (with a few differences described as follows) and capture most known types of genomic activity.

- Quiescent regions are characterized by a lack of any marks (Figures 2b, 3b) and cover about 60% of our annotations (Figure 3c). The high prevalence of Quiescent regions in our annotations is partly due to the fact that many of the cell types we annotated have only 5–10 available data sets. Quiescent regions are highly depleted around genes (Figure 3a).

- Constitutive heterochromatin (*ConstitutiveHet*) is characterized by the histone modification H3K9me3, is regulated by the HP1 complex and is thought to repress permanently silent regions such as centromeres and telomeres (Lachner et al., 2003). Our annotations of constitutive heterochromatin cover about 10% of the bases we annotated (Figure 3c) and are depleted around genes (Figure 3a).

- Facultative heterochromatin (*FacultativeHet*; also known as Polycomb-repressed chromatin) is characterized by the histone modification H3K27me3, is regulated by the Polycomb complex, and is thought to carry out cell type-specific repression (Morey and Helin, 2010; Pauler et al., 2009). Facultative heterochromatin covers about 15% of bases we annotated (Figure 3c). This type of element is only slightly depleted for annotated genes and evolutionary conservation (Figures 3a,3d), indicating that many regions repressed by facultative heterochromatin in a given cell type are active in a different cell type.

- We annotate active genes with the *Transcribed* label. Transcribed regions are characterized by the transcription-associated marks H3K36me3 and H3K79me2 (Figures 2b, 3b) and are highly enriched in annotated gene bodies (Figure 3a). The first exon of some genes is annotated as Promoter rather than Transcribed (Figure 3a); this is likely due either to promoter-associated marks extending into gene bodies or to imprecise transcription start site annotations. The input data sets do not distinguish exons from introns, so Transcribed labels include both types. For this reason, even though Transcribed regions contain highly-conserved coding exons, they exhibit only a moderate level of conservation as a whole (Figure 3d).

- *Promoter* regions are characterized by the promoter-associated marks H3K4me3 and H3K27ac, the binding of many transcription factors, and the binding of the RNA polymerase POL2RA (Figure 3b). They are highly enriched at the transcription start sites of annotated genes (Figure 3a), and are highly conserved (Figure 3d).

- *Enhancer* regions are characterized by the enhancer-associated marks H3K27ac and H3K4me1 and the binding of many transcription factors, including EP300 and CTCF (Figure 3b). Enhancer regions are enriched at the transcription start sites of annotated genes; this may be due either to promoters acting as enhancers in cell types where their proximal gene is inactive or to mis-annotation of some promoters as enhancers (Figure 3a).

- *Bivalent* regions are regulatory elements with both activating and repressive marks and are believed to be "poised" for activation in response to a developmental signal Bernstein et al. (2006). While both promoters and enhancers can be bivalent, we found that the two types of bivalent regions were difficult to distinguish, so we use a single category for both types. These regions are characterized by both the activating marks H3K27ac and TAF1 as well as the repressive marks H3K27me3 and EZH2 (Figures 2b, 3b).

- Previous annotation efforts have reported regulatory elements with marginal strength characterized by H3K4me1 without H3K27ac, which they have typically described as "weak enhancers" Ernst and Kellis (2010); Hoffman et al. (2013). This terminology has caused confusion (Kwasnieski et al., 2014) because it suggests that these elements either are called with low confidence, or promote expression to a lesser degree than "strong enhancers". In fact, it has not been verified that this pattern of activity (+H3K4me1, –H3K27ac) corresponds to either of these hypotheses. To avoid this confusion, we apply the term "permissive regulatory region" (*RegPermissive*) to these regions instead. Our RegPermissive annotations are mildly enriched in the vicinity of genes and are mildly enriched for conservation.

- As described above, we assigned the category "LowConfidence" to states that do not fit easily into one of the other categories. These tend not to simply be inactive regions, as those types would be able to be confidently categorized as Quiescent. As such, in aggregate, LowConfidence regions are neither enriched nor depleted relative to genes, and their level of conservation ranges from extremely un-conserved to a level comparable to promoters (Figures 3a, 3d). These regions may correspond to new element types or subtypes, and more work will be necessary to ascertain the function of each such state.

The *functionality score* measures the importance of a given type of activity to the organism's phenotype

To understand the functional importance of a given locus, it is important to distinguish functional activity that is relevant to phenotype from non-functional biochemical activity. For example, most promoter, enhancer and transcription activity is likely functional, whereas quiescent and heterochromatic regions are usually not functional in the repressed cell type. We use evolutionary conservation as a proxy for function: if positions with a given type of activity are usually conserved, this indicates that this type of activity is probably important for the organism's phenotype. To measure the putative functional importance of a given annotation label, we define the functionality score of a given annotation label as the 75th percentile of conservation of the positions annotated with the label. We chose the 75th percentile because, intuitively, we expect that functional regions will be enriched for conservation but that not every base will be conserved, so we would expect an upper quantile to be a more precise measure of functionality than the mean or median. We define this functionality score on the original integer annotation labels, thereby isolating this analysis from any imprecision in the label meaning interpretation step.

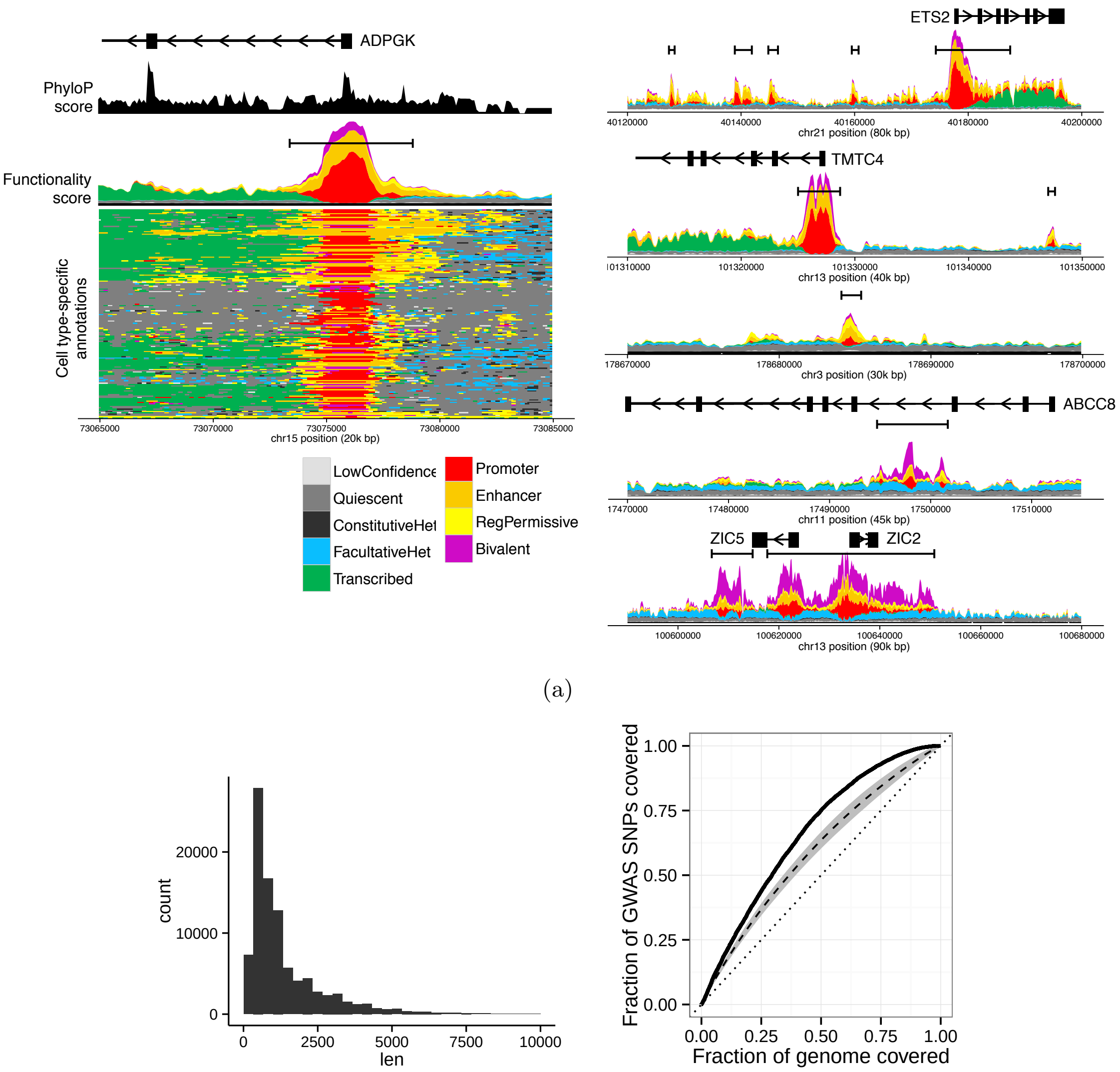
As expected, the functionality score differs greatly between label types. Regulatory elements are directly involved in gene regulation, and they typically have a high functionality score. Repressed regions generally have a low functionality score because these regions are inactive. Even though coding regions are generally the most conserved positions in the genome, transcribed regions have only an intermediate functionality score because these regions include introns as well as exons. Even though bivalent regions are likely repressed in the cell types they are active in, the functionality score accurately reflects the fact that this regulation is important to phenotype, and assigns a high score to these labels.

The functionality score is an alternative measure of functional importance to evolutionary conservation, but the functionality score has two advantages over conservation. First, the functionality score is directly attributable to a specific activity in a specific set of cell types; in contrast, conservation indicates only that a position is important, with no way to determine how it acts. Second, the functionality score can detect elements that became functional only in recent human evolution and therefore do not show conservation relative to other mammals. This means that it can detect recently-developed functional elements that are not conserved compared to other mammals. Applying it for this purpose requires developing a statistical model that can account for biases such as mappability and biased gene conversion, so we leave this for future work.

We propose the *functional activity plot* as a way to succinctly view the activity of a locus across all cell types (Figure 4a). Like a traditional annotation plot (Figure 4a), a functional activity plot displays the pattern of annotation labels across a given genomic locus, but a functionality plot additionally scales each label by its functionality score. This scaling makes it easy to see which positions show functionally important activity. For example, at a transcription start site, a functional activity plot clearly shows the promoter region, and the downstream transcription and upstream enhancers, while shrinking the upstream inactive region because it has shows no functional activity (Figure 4a). Other visualization tools that emphasize important activity have been proposed, but we believe that the importance of a position to phenotype, as represented by the functionality score, is the feature of a locus that a viewer is most frequently interested in (Discussion).

The Segway encyclopedia is an easy-to-use catalogue of all functional regulatory elements

We leveraged the functionality score to produce a cell type-agnostic encyclopedia of regulatory elements. This type of encyclopedia is inspired by a gene annotation in that it contains a single set of functional elements, where each element is marked with its pattern of activity across cell types. The encyclopedia differs from cell type-specific annotations, which annotate all bases of a given cell type and annotate both functional and non-functional activity. We defined contiguous segments with high functionality score as encyclopedia



(a)

(b)

(c)

Figure 4: Encyclopedia and functionality score. (a) Functionality score plots. Color indicates annotation label at a given position. In the cell type-specific annotation plot, cell types are clustered on the vertical axis. In functionality score plots, labels are vertically scaled by their functionality score, so that the overall height corresponds to a position's total functionality score. Black boundary bars indicate encyclopedia segments. Black phyloP area indicates the 75th percentile of phyloP scores within 100 bp of a given genomic position. Box-and-arrow pictograms indicate genes, where boxes indicate exons and arrows indicate the direction of transcription. We show functionality score plots for several example loci (left to right, top to bottom): promoter and start of neighboring gene; gene with upstream enhancers; gene with distant enhancer; distal enhancer; regulatory element that is repressed in many cell types; large regulatory domain. The vertical axis indicates the functionality score at a given position, colored proportionally to the fraction of the score that derives from each label type. Genome coordinates are relative to genome assembly GRCh37. (b) Distribution of lengths of encyclopedia segments. (c) Accuracy with which the functionality score predicts GWAS SNPs from the GWAS Catalog. We ordered genomic positions by their functionality score. The solid line indicates the number of GWAS SNPs that fall in the top X% of this list, for a given fraction X. The dashed line indicates the average performance when this ordering is performed using a single annotation (standard deviation over annotations indicated by grey area). The dotted line indicates random performance.

segments (Methods). This encyclopedia covers about 5% of the genome, and its segments range in size mostly between 300-10,000 bp.

To demonstrate the utility of the Segway encyclopedia, we used it to interpret the mechanism of action of known disease variants. We used known disease variants from the GWAS Catalog (Welter et al., 2014). Each variant is a single nucleotide polymorphism (SNP) significantly associated with a given disease according to a genome-wide association study (GWAS). These SNPs are known to be genetically linked to a causative variant, but the causative variant is generally not immediately clear because GWAS cannot disentangle genetic linkage and generally do not genotype all variation. Moreover, even when the causal variant is known, it is not easy to tell what activity and tissue this variant acts through. The median functionality score of GWAS SNPs is higher than the functionality score of 74% of the genome as a whole (Figure 4c). Moreover, the functionality score derived from all cell types is much more sensitive than using the annotation of any single cell type (Figure 4c). Note that GWAS tag SNPs are usually not the causative variant themselves, but because our annotations are performed at 100 bp resolution, the functionality score is similar within the linkage disequilibrium region around each SNP. Results were similar when using a fixed-width window around each GWAS SNP (not shown). This analysis demonstrates that the encyclopedia is a powerful and easy-to-use tool for interpreting the function of a genomic position.

3 Discussion

In this work we applied the unsupervised genome annotation method Segway to annotate 164 human cell types. We proposed a new machine learning classifier that automatically assigns biological semantics to each label, converting annotation from a semi-automated to fully-automated process. The resulting annotations represent the largest consistent annotation to date, encompassing 1,616 data sets. We were able to use this large number of input data sets because—unlike previous "concatenated" annotations—our strategy of training separate models in each cell type does not require that each cell type have the same set of available data.

We defined a functionality score that measures the putative importance of each type activity to an organism's phenotype. This score can be used to facilitate visualization of the genome through a functional activity plot, and enabled the construction of the Segway encyclopedia. This encyclopedia forms a cell type-agnostic catalogue of all regulatory elements, analogous to the widely-used cell type-agnostic catalogues of genes such as GENCODE. The Segway encyclopedia will therefore be a useful resource for interpreting genome activity.

A downside of the supervised classification approach to label interpretation is that it, by definition, cannot be used to discover new types of biochemical activity. However, there are several reasons to believe that we have not missed many novel activity types this way. First, many diverse cell types have now been annotated using SAGA methods, and the discovered labels have almost entirely fit into our set of eight categories (with the exception of a small number of labels that we assigned as "Unclassified"). Moreover, previous annotations were generally performed on the most well-studied cell types and took up to one hundred data sets as input; in contrast, the majority of our cell types have less than ten input data sets. Second, our annotations use just 13-32 annotation labels. To the extent that novel types of genomic activity exist, they are most likely subtypes of known types, and therefore will become apparent when annotating the genome with more label types. Third, our classifier was able to confidently assign all but a small fraction of annotation labels to one of these categories; only those that we assigned as LowConfidence could not be assigned this way.

Our functionality score is in some ways analogous to existing variant effect predictors that predict the impact of a mutation at a given position, such as GERP (Cooper et al., 2005), CADD (Kircher et al., 2014), and others (Gagliano et al., 2014; Ionita-Laza et al., 2016). These methods are almost certainly more sensitive than the functionality score, but they are more difficult to interpret because most such scores are based on complex machine learning classifiers that weigh many factors. In contrast, the functionality score can be directly traced to a specific genomic element with a known pattern of activity across cell types. Therefore, a variant effect predictor is most effective when trying to determine whether a given variant is deleterious, whereas the functionality score (and the Segway encyclopedia) is the most effective tool for understanding what the function of a known important variant (such as a disease allele).

There has been much debate over whether regions that show activity in genomics data are truly "functional"

(Kellis et al., 2014; Brunet and Doolittle, 2014; Ward and Kellis, 2012; Green and Ewing, 2013; Ward and Kellis, 2013; Eddy, 2012, 2013; Mattick and Dinger, 2013; Niu and Jiang, 2013; Germain et al., 2014). In this manuscript we use the word “function” in concordance with Godfrey-Smith (1994). That is, a genomic element has a given function if and only if that function caused this type of element to be selected for in the recent past. High functionality score regions are true putative functional elements according to this definition because they mark regions with patterns of activity that are predictive of evolutionary conservation.

The proposed functional activity plot has some similarities to an *epilogos* visualization (http://compbio.mit.edu/epilogos/). In particular, both types of plots show the annotation labels over a given genomic position and scale the label axis by a measure of each position’s importance. An epilogos plot differs from a functional activity plot in that it scales the label axis by the “surprisal score”, a measure of the rarity of a distribution of labels, rather than the functionality score. Both scores have the effect of magnifying promoter and enhancer regions while shrinking quiescent regions. However, most users of genome annotations—such as those in medicine or population genetics—are generally interested in the functional importance of a given position. The surprisal score does not directly measure this functional importance because not all rare labels are important, and not all common labels are unimportant. For example, in our annotations, ConstitutiveHet and Transcribed regions have similar prevalence and therefore an epilogos plot would display them with similar importance. In contrast, the functionality score accurately reflects the fact that transcribed regions are usually functional, whereas constitutive heterochromatin is virtually always non-functional. Moreover, the surprisal score actually gives a higher score to a position that is quiescent in all cell types than one that is labeled as promoter or enhancers in a few cell types, because the latter distribution most closely matches the genomic average. The functionality score is a more accurate measure of importance because it is based on conservation, which is the most direct measure of functionality that we have access to.

4 Methods

4.1 Functional genomics data

We obtained genomics signal data sets from the ENCODE and Roadmap Epigenomics consortia (`http://encodeproject.org/`). These data sets were processed by the two consortia into real-valued data tracks, as described previously (Hoffman et al., 2013; Kundaje et al., 2015; Zhang et al., 2008). Briefly, the sequencing reads were mapped to human reference genome GRCh37/hg19 (Yates et al., 2016), reads were extended according to inferred fragment lengths, the number of reads overlapping each genomic position was computed, and assay type-specific normalizations were performed, such as computing fold enrichment over an input control for ChIP-seq. We manually curated these assays to unify assay type and cell type terminology and, when multiple assays were available, we arbitrarily chose a representative assay for each (cell type, assay type) pair. We applied the inverse hyperbolic sine transform $\operatorname{asinh}(x) = \ln(x + \sqrt{x^2 + 1})$ to all signal data (Johnson, 1949; Hoffman et al., 2012). This transform is similar to the log transform in that it decreases the magnitude of extremely large values, but unlike a log transform, asinh is defined at zero and amplifies the magnitude of small values less severely than the log transform does. Finally, we applied a Z-score normalization to each data set by subtracting the genome-wide mean and dividing by the standard deviation.

We used all available data sets that measure features of chromatin state: histone modification ChIP-seq, transcription factor ChIP-seq, measures of DNA accessibility (FAIRE-seq, DNase-seq), and replication time (Repli-seq, Repli-chip). We chose not to include measures of the cell’s RNA state such as RNA-seq and measures of RNA-binding proteins because doing so would convolve regulatory and transcription state into a single annotation and would therefore make interpretation more difficult. In order to remove cell types that had only one or two available assays, we chose to annotate a given cell type only if it satisfied the criterion that either (1) there were at least five histone modification ChIP-seq assay available, or (2) there was at least one assay each in at least two of the categories transcription factor ChIP-seq, histone ChIP-seq, and DNase-seq. After performing these filtering steps, we had 1538 total tracks composed of 196 assay types and 164 cell types (median 7 tracks per cell type).

Reference	Method	Multi-cell type annotation type	Cell types	Labels	Interpreted labels
Hoffman et al. (2013)	Segway	Independent	GM12878, H1-hESC, HepG2, HUVEC, K562	25	125
Hoffman et al. (2013)	ChromHMM	Concatenated	GM12878, H1-hESC, HeLa-S3, HepG2, HUVEC, K562	25	25
Ernst et al. (2011)	ChromHMM	Concatenated	GM12878, H1-hESC, HepG2, HMEC, HSMM, HUVEC, K562, NHEK, NHLF	15	15
Kundaje et al. (2015)	ChromHMM	Concatenated	GM12878, H1-hESC, HMEC, HUVEC, K562, NHLF	15,18,25	58
This work	Segway	Independent	H1-hESC, AG04449, HSMM, RIGHT_ATRIUM	13-27	71

Table 2: Reference SAGA annotations used to train interpretation classifier. The annotations from Kundaje et al. (2015) were performed on 111 cell types, so we chose a representative six from which to compute features.

4.2 Segway model

We used Segway, a semi-automated genome annotation method, to produce annotations of the genome with integer labels (Hoffman et al., 2012). Segway takes as input a set of functional genomics data sets represented as real-valued tracks defined over the genome. The software simultaneously partitions the genome into segments and assigns an integer label to each segment such that genomic positions with the same label exhibit similar patterns in the genomics tracks. The Segway method is described in detail in previous work (Hoffman et al., 2012, 2013).

In this work, we extended Segway to use a mixture of C Gaussian distributions at each position. Previously, Segway associated each data track and each label with a single-component Gaussian distribution. In particular, let $X_{i,j} \in \mathbb{R}$ be the observed signal value for track j and position i, and let $Y_i \in \{1, \dots, K\}$ be the latent label at position i, where K is the user-specified number of labels. Previously, Segway assumed that $X_{i,j}|Y_i \sim N(\mu_{j,Y_i}, \sigma_j)$, where μ_{j,Y_i} is the learned Gaussian mean for track j and label Y_i, and σ_j is the learned track-specific variance for track j. (Segway uses variance parameters that are shared for each track—rather than K label-specific variance parameters—to avoid overfitting and local optima caused by variance parameters approaching zero.) To extend this specification to Gaussian mixtures, we define C mean parameters μ_{c,j,Y_i} for each label-track pair, and we define C variance parameters $\sigma_{c,j}$ for each track. In addition, we define a latent random variable $Z_{i,j} \in \{1, \dots, C\}$ that indicates which Gaussian component the observation associated with track j at position i was drawn from. We assume that $Z_{i,j} \sim \text{Multinomial}(m_{1,j,k}, \dots, m_{C,j,k})$ and $X_{i,j} \sim N(\mu_{Z_{i,j},j,Y_i}, \sigma_{Z_{i,j},j})$, and we learn the parameters $m_{c,j,k}$ using expectation-maximization training. To avoid local optima caused by any $\sigma_{c,j}$ approaching zero, we prevented the training process from assigning values < 0.001 to the parameter $\sigma_{c,j}$. We used three-component mixtures for all of our annotations in this work ($C = 3$) because it appeared to represent a good trade-off between flexibility and complexity.

We ran this modified version of Segway on each of the 164 cell types to produce annotations. For a cell type with M available data sets, we asked Segway to assign $10 + 2\sqrt{M}$ different labels. We chose this number of labels to let the complexity of the model vary with the amount of input data, following previous work. We performed all annotations at 100 bp resolution, allowing Segway to train for a maximum of 25 iterations. At each iteration, we chose a random 1% of the genome (a “mini-batch”) to use to update parameters—this speeds up training while allowing the model to use all available data. We used ten random parameter initializations for each cell type and selected the model with the highest likelihood after training. We then used the trained models to annotate the whole genome of each cell type.

4.3 Biological meaning classifier

For use in training our biological meaning classifier, we curated a collection of published manually-interpreted SAGA annotations. We obtained five Segway annotations and five ChromHMM annotations, for a total of 223 annotation labels that have had their biological meaning interpreted (Table 2). We additionally manually annotated four of our annotations in order to provide more training data, resulting in 71 additional labels. We mapped the biological meanings from these annotations to a unified vocabulary of eight biological meanings (Table 1) by combining synonyms for the same activity (such as "Polycomb repressed region" and "Facultative heterochromatin") and combining labels of the same type of activity that were artifactually divided by the simple single-component Gaussian models used by previous SAGA methods (such as "Weak transcription" and "Transcription"). Notably, we designated the meaning "RegPermissive" for regions that exhibit some signs of regulatory activity but that do not have the characteristic marks of either promoters or enhancers. These regions were previously designated as "weak enhancers" or "promoter flanking". We avoided using vocabulary indicating strength (such as "weak enhancer") because these terms have been inconsistently used in the past to refer to either (1) weak enrichment for the associated data sets or (2) enrichment for some but not all of the characteristic tracks (such as "weak enhancers" enriched for H3K4me1 but not H3K27ac). These strength-associated terms can be misinterpreted as indicating a level of confidence of strength of biological activity of the associated element, neither of which do we sufficient evidence to claim (Kwasnieski et al., 2014). For the 26/294 labels that did not fit into our vocabulary (including labels of "Artifact", "Insulator", "Genic enhancer" and "FAIRE"), we assigned the classification "Unclassified". This process resulted in 294 training examples, each associated with one of eight biological meanings.

For each annotation label, we defined features that encompass the information typically used to interpret annotation labels. Specifically, we used the following 17 features: mean value of H3K27me3, H3K4me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3 (six features), and log enrichment of the label in the following positions relative to GENCODE genes: 1–10k bp 5' flanking, 1 bp–1k bp 5' flanking, initial exon, initial intron, internal exons, internal introns, terminal exon, 1 bp–1k bp 3' flanking, and 1–10k bp 3' flanking (`http://gencodegenes.org`, version 24, Harrow et al. (2006)). The enrichment of a given label l at a set of loci c is defined as

$$\log_2 \frac{n_{\text{overlap}} + 1}{(n_l n_c / n) + 1},$$

where n_l and n_c are the number of bases that l and c cover respectively, $n_{overlap}$ is the number of bases that they overlap, and n is the total number of bases in the genome.

For the 69 cell types missing one of these histone modifications, we substituted data from the most-similar cell type with that data set available. To find a substitute for a given assay type A and cell type C, we calculated the similarity between each cell type C' that had data for A. Specifically, we calculated the similarity between C and C' as the mean Pearson correlation between all assay types that are present in both cell types. We chose the instance of A from the most similar cell type as the substitute. Note that, while imputing missing data using methods like this is likely too noisy to produce good position-specific measures of chromatin state, doing so likely preserves the genome-wide patterns of these marks and therefore can be used for the interpretation of genome-wide labels derived from Segway.

We trained a multi-class decision tree classifier to predict the biological meaning of each label from its 17 features. We used the random forest implementation from scikit-learn (Pedregosa et al., 2011) using the "entropy" splitting criterion and regularized such there were at least 10 training example associated with each decision tree leaf (`min_samples_leaf=10`). We chose this regularization parameter using leave-one-out cross-validation. We applied this classifier to each of our new annotations, deriving features for these new annotations in the same way as for the reference annotations. For a given label, if the classifier assigned the meaning "Unclassified" or assigned less than 25% probability to any one meaning, we assigned the meaning "LowConfidence".

4.4 Functionality score and encyclopedia

We defined a functionality score for each annotation label that indicates the degree to which the label is likely to mark functionally important elements. For a given annotation label ℓ, we collected the 46-species placental mammal phyloP scores, a measure of evolutionary conservation, for all genomic positions annotated by ℓ (Siepel et al., 2006). We defined the functionality score of ℓ to be the 75th percentile of the absolute values of these phyloP scores. Functionality scores range from 0.365–1.215. We defined the functionality score of a genomic position p to be the sum of the functionality score for all 164 annotation labels that cover p.

We used the functionality score to define encyclopedia segments. We defined an encyclopedia segment to be any contiguous genomic segment with a high total functionality score. Specifically, we defined the score of a given position k as $s(k) = f_k - Z$ where f_k is the functionality score of k, and the total score of a segment $[i, j]$ as $s([i, j]) = \sum_{k=i}^{j} s(k)$. We defined an encyclopedia segment any segment $[i, j]$ such that $s([i, j]) > S$ and $s([i', j']) \leq s([i, j])$ for all $i' < i, j' > j$. We required that each segment have no subsegment $[i', j']$ $(i' > i; j' < j)$ such that $s([i', j']) \geq D$ to avoid merging neighboring segments. We further placed a minimum on the mean segment score $s([i, j])/(j - i) \geq M$ and a minimum on the segment length $j - i \geq L$. We chose $Z = 0.775$, $D = 1$, $S = 5$, $M = 0.02$, and $L = 500$ bp such that the resulting segments matched our intuition about the size and frequency of functional regulatory elements.

References

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al.. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315–326.

Biesinger J, Wang Y, and Xie X. 2013. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics* **14**: S4.

Brunet TD and Doolittle WF. 2014. Getting function right. *Proceedings of the National Academy of Sciences* **111**: E3365–E3365.

Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, and Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**.

Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, and Noble WS. 2007. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**: 1424–1426.

Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. *Current Biology* **22**: R898–R899.

Eddy SR. 2013. The ENCODE project: missteps overshadowing a success. *Current Biology* **23**: R259–R261.

Ernst J and Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology* **28**: 817–825.

Ernst J and Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology* **33**: 364–376.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh 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.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al.. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**: 212–224.

Gagliano SA, Barnes MR, Weale ME, and Knight J. 2014. A Bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. *PloS ONE* **9**: e98122.

Germain PL, Ratti E, and Boem F. 2014. Junk or functional DNA? ENCODE and the function controversy. *Biology & Philosophy* **29**: 807–831.

Godfrey-Smith P. 1994. A modern history theory of functions. *Noûs* **28**: 344–362.

Green P and Ewing B. 2013. Comment on “Evidence of abundant purifying selection in humans for recently acquired regulatory functions”. *Science* **340**: 682.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JGR, Storey R, Swarbreck D, et al.. 2006. GENCODE: Producing a reference annotation for ENCODE. *Genome Biology* **7**: S4.

Ho JWK, Liu T, Jung YL, Alver BH, Lee S, Ikegami K, Sohn K, Minoda A, Tolstorukov MY, Appert A, et al.. 2014. Comparative analysis of metazoan chromatin architecture. *Nature* **512**: 449–452.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, and Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**: 473–476.

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 Res* **41**: 827–41.

Ionita-Laza I, McCallum K, Xu B, and Buxbaum JD. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics* **48**: 214–220.

Jiang K, Thorsen O, Peters A, Smith B, and Sosa CP. 2008. An efficient parallel implementation of the hidden Markov methods for genomic sequence-search on a massively parallel system. *IEEE Transactions on Parallel and Distributed Systems* **19**: 15–23.

Johnson NL. 1949. Systems of frequency curves generated by methods of translation. *Biometrika* pp. 149–176.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al.. 2014. Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences* **111**: 6131–6138.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, and Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**: 310–315.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al.. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Kwasnieski JC, Fiore C, Chaudhari HG, and Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research* **24**: 1595–1602.

Lachner M, O'Sullivan RJ, and Jenuwein T. 2003. An epigenetic road map for histone lysine methylation. *Journal of Cell Science* **116**: 2117–2124.

Lian H, Thompson W, Thurman RE, Stamatoyannopoulos JA, Noble WS, and Lawrence C. 2008. Automated mapping of large-scale chromatin structure in ENCODE. *Bioinformatics* **24**: 1911–1916.

Libbrecht M, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, and Noble WS. 2015. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. *Genome Research* **25**: 544–557.

Lystig TC and Hughes JP. 2002. Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics* **11**: 678–689.

Mammana A and Chung HR. 2015. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome biology* **16**: 1.

Mattick JS and Dinger ME. 2013. The extent of functionality in the human genome. *The HUGO Journal* **7**: 1.

Morey L and Helin K. 2010. Polycomb group protein-mediated repression of transcription. *Trends in Biochemical Sciences* **35**: 323–332.

Murphy KP. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Niu DK and Jiang L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochemical and Biophysical Research Communications* **430**: 1340–1343.

Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, and Barlow DP. 2009. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Research* **19**: 221–233.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al.. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**: 2825–2830.

Schliep A, Schönhuth A, and Steinhoff C. 2003. Using hidden markov models to analyze gene expression time course data. *Bioinformatics* **19**: i255–i263.

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, and Furey TS. 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research* **23**: 777–788.

Siepel A, Pollard KS, and Haussler D. 2006. New methods for detecting lineage-specific selection. In *Annual International Conference on Research in Computational Molecular Biology*, pp. 190–205. Springer.

Sohn KA, Ho JW, Djordjevic D, Jeong Hh, Park PJ, and Kim JH. 2015. hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics* p. btv117.

Thurman RE, Day N, Noble WS, and Stamatoyannopoulos JA. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Research* **17**: 917–927.

Ward LD and Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**: 1675–1678.

Ward LD and Kellis M. 2013. Response to comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". *Science* **340**: 682.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al.. 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**: D1001–D1006.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al.. 2016. Ensembl 2016. *Nucleic Acids Research* **44**: D710–D716.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, and Flicek PR. 2015. The Ensembl regulatory build. *Genome Biology* **16**: 1.

Zhang Y, An L, Yue F, and Hardison RC. 2016. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Research* p. gkw278.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al.. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.