UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Methods for Comparative Genomic and Epigenomic Annotations across Multiple Species

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Adriana Cristina Arneson

2020

ABSTRACT OF THE DISSERTATION

Computational Methods for Comparative Genomic
and Epigenomic Annotations across
Multiple Species

by

Adriana Cristina Arneson
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2019
Professor Jason Ernst, Chair

In recent years Genome Wide Association Studies (GWAS) and large-scale whole genome sequencing case-control studies have led to the identification of a wealth of phenotype-associated and rare genetic variants. Interpreting the biological significance of these variants has been a significant challenge, especially since a large majority of their genomic locations fall within non-protein coding genomic regions. Here we present a computational method, ConsHMM, for annotating the genome at single-nucleotide resolution into a set of conservation states learned from the combinatorial and spatial patterns of species aligning and matching a reference genome in a multiple-sequence alignment. Conservation states have specific enrichments for orthogonal biological annotations and can be used for interpreting genetic variants. We provide here a comprehensive resource of conservation state annotations, the ConsHMM atlas, comprised of models and annotations for eight different organisms based on

several multiple-sequence alignments. At the epigenomic level, modifications such as DNA methylation have emerged as useful biomarkers for several phenotypes, but a large majority of these phenotypes have been studied predominantly in human samples. Leveraging sequence conservation among genomes, we have designed a methylation array that can query DNA methylation of many different mammals, and therefore facilitate cross species epigenetic studies. The array has been produced and used to profile 8730 samples from 145 different mammals. In summary, this work takes a comparative genomics based approach to expanding the available genomic and epigenomic annotations of multiple species.

The dissertation of Adriana Cristina Arneson is approved.

Kirk Lohmueller

Bogdan Pasaniuc

Xia Yang

Jason Ernst, Committee Chair

University of California, Los Angeles

2020

DEDICATION

This dissertation is dedicated to my husband Doug, my parents Cristina and Traian and my late grandparents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd like to first acknowledge my mentor and advisor Jason Ernst. Jason is one of the most patient, hard-working and intelligent people I have ever met and he has helped me transform from an overwhelmed first year student into a confident researcher. He has taught me how to ask the right questions and how to conduct research with exemplary integrity, transparency and commitment to reproducibility. I will forever cherish the lessons learned from Jason and will continue to strive to follow the example that he sets as a scientist. Words cannot express how grateful I am for his patience and dedication in advising me.

The other members of my committee Xia Yang, Bogdan Pasaniuc and Kirk Lohmueller have been instrumental throughout my graduate development. They are wonderful mentors who have always lent their time and expertise and whom I greatly admire both for their accomplishments as scientists and for their unwavering commitment to the students they advise.

I also have to thank my husband (and fellow graduate student) Doug for his constant love and support, without which I would be lost. Doug believed in me in times when I did not and has always been a beacon of light when I felt like I was stumbling through darkness. He is my soulmate, my teammate, my inspiration, and I am grateful for every moment we get to be together. Doug, I love you more than anything.

Friends are the family you choose, and I have to thank all my friends for providing me with the best support system anyone could ask for: Beth, Cynthia, Andres, Marvin and Sam, whom I have now known for a decade, and who have kept a bond alive across an entire continent. All the friends I made in graduate school have left their mark on me as a person and a researcher and I would not have completed this dissertation without them.

Si nu in ultimul rand, vreau sa le multumesc parintilor si bunicilor care m-au crescut si sustinut din prima zi. Nu as fi realizat nimic fara sprijinul lor si fara nenumaratele sacrificii pe care le-au facut pentru mine. Parintii mei mi-au dat intotdeauna toata dragostea lor, si pentru asta voi fi vesnic recunoscatoare. Mami, ai fost si vei fi intotdeauna un exemplu pentru mine. De la tine am invatat cum sa fiu puternica, independenta si muncitoare, dar si ce inseamna sa iti pui copiii mai presus de orice. Sper ca atunci cand va veni si randul meu sa fiu parinte, voi putea face macar jumatate din ceea ce ai facut tu pentru mine. Tati, mi-ai inspirat pasiunea pentru informatica, citit, benzi desenate, jocuri, limbi straine, si cate altele. Iti multumesc ca m-ai invatat sa rad din orice, sa nu iau totul in serios, sa raman mereu copil la suflet si pentru ca nu mi-ai spus niciodata ca fetele nu fac informatica. Va iubesc pe amandoi din toata inima.

Chapter 2 is a version of Arneson A, Ernst J (2019). Systematic discovery of conservation states for single nucleotide annotation of the human genome. *Communications Biology*, 2(1):248. AA is supported by US NIH grant T32CA201160. JE is supported by US NIH grants DP1DA044371, R01ES024995, U01HG007912, U01MH105578, US National Science Foundation CAREER Award #1254200, Kure It cancer research Kure-IT award and Alfred P. Sloan Fellowship.

Chapter 3 is a version of Arneson A, Felsheim B, Chien J, Ernst J (2020), ConsHMM Atlas: conservation state annotations for major genomes and human genetic variation. *Manuscript in preparation. AA is* supported by US NIH grant T32CA201160. JE is supported by US NIH grants DP1DA044371, R01ES024995, U01HG007912, U01MH105578, US National Science Foundation CAREER Award #1254200, Kure It cancer research Kure-IT award and Alfred P. Sloan Fellowship.

The work in **Chapter 4** was a collaboration with Dr. Steve Horvath and members of his group Mike Thompson, Joshua Zhang, Caesar Li, Ake Lu, Joseph Zoller and with Illumina Inc.,

represented by Bret Barnes. The study was supported by funding from the Paul G. Allen Frontiers Group, US NIH grant T32CA201160 to AA and US NIH grants DP1DA044371, R01ES024995, U01HG007912, U01MH105578, US National Science Foundation CAREER Award #1254200, Kure It cancer research Kure-IT award and Alfred P. Sloan Fellowship to JE.

VITA

EDUCATION

2019	PhD Candidate, Bioinformatics University of California, Los Angeles
2014	Bachelor of Arts, Computer Science Colgate University, Hamilton, NY

RESEARCH EXPERIENCE

2014 – Present	PhD trainee Department of Biological Chemistry University of California, Los Angeles
2017	Computational Biology Summer Student Worker Celgene Corporation La Jolla, California

PEER-REVIEWED PUBLICATIONS

Arneson A, Ernst J (2020). Systematic discovery of conservation states for single-nucleotide annotation of the human genome. ***Communications Biology***, 2(1):248.

Ay A, Holland J, **Sperlea A**, Devakanmalai GS, Knierer S, Sangervasi S, Stevenson A, Ozbudak EM (2014), Spatial gradients of protein-level time delays set the pace of the traveling segmentation***. Development****,* 141(21):4158-67

Ay A, Knierer S, Sperlea A, Holland J, Ozbudak EM (2013), Short-lived Her proteins drive robust synchronized oscillations in the zebrafish segmentation clock. ***Development*****,** 140(15):3244-53

PRE-PRINTS

Grujic O, Phung TN, Kwon SB, **Arneson A**, Lee Y, Lohmueller K, Ernst J (2019), Identification and characterization of constrained non-exonic bases lacking predictive epigenomic and transcription factor binding annotations. ***bioRxiv*****,** doi:10.1101/722876

MANUSCRIPTS IN PREPARATION

Arneson A, Felsheim B, Chien J, Ernst J (2020), ConsHMM Atlas: conservation state annotatios for major genomes and human genetic variation.

Arneson A, Li Z, Thompson M, Barnes B, Lu A, Zoller J, Zhang J, Ernst J, Horvath S, Development of a mammalian methylation array.

Chapter 1. Introduction

Comparative genomic approaches have long been at the core of efforts for annotating genomes and interpreting the biological function of DNA sequences[1]. These efforts have become increasingly important in recent years, as the flourishing of GWAS and large-scale whole genome sequencing case-control studies has led to an accumulation of documented phenotype-associated and rare variants, whose mechanistic role is poorly understood[2]. In particular, in non-protein coding regions of the genome, where the amino acid code is not applicable, variant interpretation becomes challenging. Methods based on comparative genomics use comparisons either between genomes of different species to identify regions where mutations are less likely to happen than would be expected based on models of neutral evolution [3–7]. Such methods can be applied genome-wide and produce a single nucleotide annotation of the importance of each base in a genome. Epigenomic annotations have been another source of genome interpretation, particularly in recent years, when consortiums such as Roadmap Epigenomics and ENCODE have accumulated a wealth of epigenomic datasets assays across cells and tissue types[8,9]. In particular the epigenetic modifications of DNA methylation has emerged as an important biomarker for several phenotypes, such as cancer status and biological age[10,11]. However, these biomarkers have been primarily studies in human samples, and a robust platform for studying epigenetic modifications across species is still lacking. Here, I present a set of computational comparative genomics methods that leverage sequence conservation to create biologically meaningful annotations of genomes and to create a methylation array for cross-species epigenomic studies.

In **Chapter 2** we present a comparative genomics whole genome annotation approach, ConsHMM, which applies a multivariate hidden Markov model to learn de novo 'conservation states' based on the combinatorial and spatial patterns of which species align to and match a

reference genome in a multiple species DNA sequence alignment. Unlike existing methods, ConsHMM takes a data-driven, unbiased approach that does not make any assumption about a neutral rate of evolution or an underlying phylogeny. We applied ConsHMM to a 100-way vertebrate sequence alignment to annotate the human genome at single nucleotide resolution into 100 conservation states. These states have distinct enrichments for other genomic information including gene annotations, chromatin states, repeat families, and bases prioritized by various variant prioritization scores. Constrained elements have distinct heritability partitioning enrichments depending on their conservation state assignment. ConsHMM conservation states are a resource for analyzing genomes and genetic variants.

In **Chapter 3** we apply ConsHMM to produce 21 additional genome annotations covering human and seven other organisms for a variety of multi-species alignments. Additionally, we have extended ConsHMM to generate allele specific annotations, which we used to produce conservation state annotations for every possible single nucleotide mutation in the human genome. Finally, we provide a web interface to interactively visualize parameters and annotation enrichments for ConsHMM models. These annotations and visualizations comprise the ConsHMM Atlas, which we expect will be a valuable resource for analyzing a variety of major genomes and genetic variation.

In **Chapter 4** we present the development of a mammalian methylation array that can be used to profile DNA methylation in several mammals. The mammalian array can facilitate comparative epigenomic analyses, as it is based on a set of human CpGs falling in conserved regions, for which we engineered probes that can accurately profile the same CpG in other species. We applied the array to a set of 8730 samples from 145 different species. Here we present an analysis of the array data from 10 mammals with more than 200 samples for each species. We find that the methylation signature of each sample encodes both the species and

tissue of origin information. We anticipate the array will be used for a multitude of comparative epigenomic studies.

Chapter 2. Systematic Discovery of Conservation States for Single-Nucleotide Annotation of the Human Genome

2.1 Introduction

The large majority of phenotype-associated variants implicated by genome-wide association studies (GWAS) are non-coding[12]. Identifying and interpreting causal non-coding variants is an important challenge[13]. Mapping of epigenomic data across different cell and tissue types has been one approach for annotating and interpreting the non-coding regions of genomes[8,9,14]. Using comparative genomics data to identify regions of evolutionary constraint has been a complementary approach for these purposes[3–5,15].

In addition to providing evolutionary information, comparative genomics data has the advantage of providing information at single-nucleotide resolution. Furthermore, it is cell type agnostic and thus informative even when the relevant cell or tissue type has not been experimentally profiled[2,16]. The most commonly used representations of this information are univariate scores and binary elements of evolutionary constraint, which are called based on a multiple species DNA sequence alignment and assumed models of evolution and selection[3,4,6,7,17]. Supporting the importance of these annotations, heritability analyses have recently implicated evolutionary constrained elements as one of the annotations most enriched for phenotype associated variants[18]. These scores and elements have also been highly informative features to integrative methods for prioritizing pathogenic variants[19–22]. Further improvements to pathogenic coding variant prioritization scores have been made by also using features defined directly from a multiple sequence alignment[23].

While useful, the representation of comparative genomics information into univariate scores or binary elements is limited in the amount of information it can convey about the underlying multiple sequence alignment at a specific base. This limitation has become more pronounced given the large number of species now available in multi-species alignments such as a 100-way alignment to the human genome[24]. Approaches have been developed to associate constrained elements, regions, or individual bases with specific branches in a phylogenetic tree[25–31]. While also useful, such directed approaches are biased to only representing certain types of patterns present in an alignment. An alternative approach learned patterns of different classes of mutations between human and only one non-human genome[32], and was only applicable at a broad region level.

Analogous to the many sequenced genomes available for comparative analysis, many different epigenomic datasets are available for annotating genomes. Approaches that define 'chromatin states' based on combinatorial and spatial patterns in these datasets have effectively summarized the information in them to provide *de novo* genome annotations[14,33–35]. Inspired by the success of these approaches, here we develop a method, ConsHMM, that extends the ChromHMM[34] method to systematically annotate genomes into 'conservation states' at single nucleotide resolution given a multiple species DNA sequence alignment. ConsHMM takes a relatively unbiased and flexible modeling approach that does not explicitly assume a specific phylogenetic relationship between species.

We applied ConsHMM to assign a conservation state to each nucleotide of the human genome. The states capture distinct enrichments for other genomic annotations such as gene annotations, CpG islands, repeat families, chromatin states, genetic variation, and bases prioritized by variant prioritization scores. The ConsHMM conservation state annotations are a resource for interpreting genomes and potential disease-associated variation, which complement both existing conservation and epigenomic-based annotations.

2.2 Results

Annotating the human genome into conservation states

We developed an approach, ConsHMM, to annotate a genome into conservation states at single nucleotide resolution based on a multiple species DNA sequence alignment (**Figure 2.1a, Methods**). At each position in a reference genome, ConsHMM encodes one of three observations for each non-reference species in the alignment: aligns with a nucleotide present that is the same as the reference genome, different than the reference genome, or does not have a nucleotide present at that position. ConsHMM then probabilistically models the combinatorial and spatial patterns in these observations using a multivariate hidden Markov model (HMM). In each state of the HMM, ConsHMM assumes that the probability of observing a specific combination of observations is determined by a product of independent multinomial random variables. The parameter values will generally differ between states, and ConsHMM learns them from the input. After the model is learned, ConsHMM assigns each nucleotide in the reference genome to the state that had the maximum posterior probability of generating the observations.

We applied ConsHMM to a 100-way Multiz vertebrate alignment with the human genome as the reference genome[24,36]. We focused our analysis here on a model learned using 100 states to balance recovery of additional biological features and model tractability (**Figures 2.2-2.10, Methods**). We verified that ConsHMM's transition parameters have a smoothing effect, which is consistent with applications of HMMs for constrained element detection[3,17], as the number of segments increased from 889 million to 1.06 billion when using an equivalent model without transition information, though most state assignments to individual bases were the same (**Figure 2.11, Methods**). We illustrate ConsHMM conservation state annotations at two loci, which shows that bases with similar existing constraint annotations can have different

conservation state assignments corresponding to very different underlying alignment patterns (**Figures 2.2b, 2.12**).

Major groups of conservation states

Hierarchically clustering the conservation states revealed eight notable subsets of states (**Figures 2.2a, 2.6, Supplementary Data 1, Methods**). The first subset was a single state (state 1) that showed high align and match probabilities through essentially all the vertebrates. The second subset showed relatively high align and match probabilities for all mammals and some non-mammalian vertebrates (states 2-4). The third subset showed relatively high align and match probabilities for most if not all mammals, but not non-mammalian vertebrates (states 5-22). The fourth subset showed high align probabilities for many mammalian species, but had low align probabilities for notable mammals such as mouse and rat for many of the states in the group (states 23-46). The lower mouse and rat probabilities relative to mammals that diverged earlier is consistent with increased substitution rates for mouse and rat[5]. The fifth subset showed high align probabilities for many mammalian species, but did not show high match probabilities (states 47-63). The sixth subset showed high align probabilities for most primates, but not for other species (states 64-89). The seventh subset showed high align probabilities for at most a subset of primates (states 90-99). The final subset was a single state (state 100) that showed high align and match probabilities for most primates and non-mammalian vertebrates, but low probabilities for non-primate mammals, consistent with a previous observation about the association of non-mammalian vertebrates with likely alignment artifacts[37].

Conservation states positional enrichments

Conservation states showed strong and distinct positional enrichments relative to annotated gene features including transcription start sites (TSS), transcription end sites (TES),

and exon start and end sites, for both protein coding genes and pseudogenes. Within 20 base pairs (bp) of exon starts of protein coding genes, seven states (states 1-4, 7, 28, and 54) had at least 13-fold enrichment for some position, which also held for exons in specific coding phases (**Figures 2.13a, 2.14a-c**). These states were the only states that had a majority of positions aligning for at least some non-mammalian vertebrates, while still having a majority of positions aligning for all primates (**Figure 2.2a, Supplementary Data 1**). Within exons, states 1 showed the strongest enrichment, consistent with its high matching probabilities through all vertebrates (**Figures 2.2b, 2.13a,b, 2.14a-e**). State 1 also had >40-fold enrichment at each of the three nucleotides immediately upstream of exon starts and six nucleotides downstream of exon ends (**Figures 2.13b, 2.14c**), corresponding to positions of the canonical 3' and 5' splice site sequences respectively, and consistent with their high conservation throughout vertebrates[38]. Downstream of the start of protein-coding exons, the enrichment profile for state 1 showed a 3-bp oscillation period, with a dip of enrichment at codon wobble positions. States 3 and 54 showed an inverse oscillation pattern, consistent with the states' high align probabilities through many vertebrates and lower match probabilities (**Figures 2.13a, 2.14a-c**).

Around the TSS of protein coding genes, state 28, which had moderate align and match probabilities for most vertebrates, had the maximum enrichment (>30-fold) (**Fig. 2.13c**). Consistent with this enrichment, state 28 also had a 32-fold enrichment for CpG islands. However, state 28 was also 20-fold enriched for CpG islands >2kb away from any TSS of protein coding genes and 10-fold enriched for TSS of protein coding genes >2kb away from a CpG island. This suggests that both of these features are contributing to the association or the presence of unannotated TSS overlapping CpG islands[39]. Relative to TES of protein coding genes, enrichment of state 2, which had high align and match probabilities for almost all vertebrates except for fish, peaked at almost 12-fold (**Figure 2.14f**).

Relative to pseudogene exon starts and ends, states 100 and 82, both associated with alignability to distal vertebrates without many mammals closer to human (**Figure 2.2b, Supplementary Data 1**), had enrichments peaking at greater than 100 and 38-fold respectively (**Figure 2.14g,h**). States 100 and 82 also showed the greatest enrichment relative to TSS of pseudogenes peaking at 184 and 68-fold respectively (**Fig. 2.13d**) and for TES of pseudogenes peaking at 199 and 61-fold respectively (**Figure 2.14i**).

Conservation states also had different positional enrichments relative to instances of regulatory motifs, with the enrichment varying at single nucleotide resolution (**Figure 2.13e,f, Methods**)[40]. For example, states 2 and 5 reached 1.8-fold enrichments at some nucleotides in the *POU5F1* and *STAT* motifs respectively, but had lower enrichments (1.4-1.5) at other nucleotides with lower information content. States 55-57, which had high align probabilities for most mammals and low match probabilities even for most primates, peaked in enrichment at the CG dinucleotide in the center of the *STAT* motif, consistent with their genome-wide CG dinucleotides enrichments (**Figures 2.13e, 2.15**).

Conservation state enrichments for different gene classes

We next investigated conservation states enrichments for different gene classes. For each state, we determined the top 5% of gene promoter regions overlapping the state, which controls for different state preferences in general for promoters. For those corresponding genes, we evaluated Gene Ontology (GO) enrichments, which revealed distinct enrichment patterns (**Figures 2.16b, 2.17, Methods**). For example, states 1-3, which all had high alignability through at least birds, had substantial differences in their gene preferences. Out of these states, state 1 and state 3, which had high matching through all vertebrates and mainly mammals respectively, were the only ones enriched for nucleosomes ($p<10^{-41}$; 10.5-fold) and sensory perception of smell genes ($p<10^{-300}$; 15.5-fold) respectively. State 2, which had high match probabilities

through all vertebrates except fish, was the state most enriched for cellular developmental processes ($p<10^{-30}$; 1.8-fold), which were not enriched in state 3. States with overall lower align or match probabilities also had notable enrichments. For example, state 89, which had moderate alignability for most non-primate mammals, but low matching even for primates, was the state most enriched for antigen binding ($p<10^{-14}$; 6.7-fold) consistent with antigen binding being associated with many species, but fast evolving[41].

Conservation state enrichments for repeat elements

The conservation state enrichments for bases in repeat elements ranged widely from 2-fold enrichment to 133-fold depletion (**Figures 2.2b, 2.10**)[24,42]. Of the 25 states in which only primate species had a majority of positions aligning, all but states 89 and 96 had an enrichment of 1.55 or greater for repeat elements, while the other 75 states all had a lower enrichment or were depleted (**Supplementary Data 1**). Neither state 89 nor 96 enriched for repeat elements. As noted above, state 89 is associated with fast evolving bases shared with some non-primate mammals, while state 96 is associated with assembly gaps (**Figure 2.10**).

Individual conservation states had distinct enrichments for different repeat classes (**Figure 2.18**). For instance, different states had maximal enrichments for the DNA, LINE, LTR, and SINE repeat classes (**Figure 2.16d**). State 74, which had high align and match probabilities for all primates, had the maximal enrichment of 5.6-fold for DNA repeats, while the enrichment for the other three classes were between 1.0 and 1.8-fold. State 86, which lacked alignability of a subset of primates, had the maximal of 3.0-fold enrichment for LINE repeats, while the enrichment for the other classes were between 0.6 and 1.6-fold. States 76 and 77 had maximal enrichments of 3.3 and 4.5-fold for LTR and SINE respectively compared to 1.1 and 2.1-fold for SINE and LTR respectively. States 76 and 77 both had high align probabilities through primates up to and including squirrel monkey, with the exception that state 77 lacked alignability to gorilla.

Despite these subtle differences in alignment probabilities, these states had substantial differences in their repeat enrichments.

Relationship of conservation states to chromatin states

To understand the relationship of conservation states to chromatin states we determined the median enrichment of each conservation state for 25-chromatin states defined across 127 samples using imputed data[9,43] (**Figures 2.16a**, **2.19**). Eleven conservation states were maximally enriched for at least one of the chromatin states. Conservation state 28 had the greatest enrichment for any chromatin state, with a 35-fold enrichment for an active promoter chromatin state, and was maximally enriched for four other promoter associated chromatin states. Conservation state 1 was maximally enriched (3.8 to 8.7-fold) for five chromatin states associated with transcribed and exonic regions[43], consistent with its maximal enrichment for annotated exons. Conservation state 2 was maximally enriched (3.1 to 4.7-fold) for five enhancer associated chromatin states, while conservation state 5 had high enrichments for these states and was maximally enriched (2.5-fold) for a chromatin state primarily associated with just signal of DNase I hypersensitive sites (DHS). These chromatin state enrichments highlight the multi-dimensional information that conservation states capture.

Conservation states and cell type specific DHS

We next investigated whether different conservation states capture distinct enrichment patterns for DHS across cell and tissue types. We analyzed DHS from the 53 samples considered above for which maps of experimentally observed DHS were available[9]. We hierarchically clustered the row normalized enrichment patterns of the 21 conservation states that exhibited at least 2-fold enrichment in one or more samples, revealing two major clusters of states (**Figure 2.16c**). One major cluster contained 14 states, with ten of the states having

maximum enrichment for a fetal sample and the remaining four states having maximum enrichment for the cell type Human Umbilical Vein Endothelial Cells (HUVEC). The second major cluster consisted of seven states, all of which were enriched for CpG islands (**Figures 2.2b**, **2.10**). The samples for which DHS had the greatest enrichments for states in this cluster also had the greatest enrichment for CpG islands (**Figure 2.16c, Methods**), but were biologically diverse in the type of cell or tissue and could potentially reflect technical differences.

Conservation states' relationship to constraint annotations

We next investigated the relationship of the conservation state annotations with constrained element sets from four methods (GERP++, SiPhy-omega, SiPhy-pi, and PhastCons) and univariate scores of evolutionary constraint from three methods (GERP++, PhastCons, and PhyloP). The PhastCons and PhyloP constraint annotations were defined on the same alignment as the conservation states. The available GERP++, SiPhy-omega, and SiPhy-pi constraint annotations were defined from different versions of Multiz alignments and only considered mammals.

States 1-5 all had >9.0-fold enrichment for each constrained element set and high mean constraint scores consistent with their high matching probabilities across all mammals (**Figures 2.2b, 2.20**). States 54 and 100 also had >6.0-fold enrichment for at least one constrained element set. State 100, which had high aligning and matching primarily in non-mammalian vertebrates, had 15-fold enrichment for PhastCons elements and high mean PhastCons and PhyloP scores, consistent with these scores being defined using non-mammalian vertebrates. State 54, which had high alignability through most vertebrates and low matching outside primates, enriched 4 to 7-fold for the constrained element sets, but did not show high mean base-wise scores particularly for the GERP++ and PhyloP scores, consistent with its enrichments for codon wobble positions. More generally, constrained element sets, except for

PhastCons, did not show biologically relevant variation at single nucleotide resolution in their enrichments around regulatory motifs and exon start and ends as the conservation state annotations did (**Figures 2.13a,e,f, 2.21**).

We compared biologically relevant information in conservation state and constraint annotations using established genome annotations. We evaluated their ability to recover annotated TSS, TES, and exon starts and ends separately for protein coding and pseudogenes (**Figure 2.22a-c, 2.23**). In almost all cases the conservation states provided greater information for recovering annotated gene features. The only exceptions were that PhyloP scores had higher precision at low recall levels for protein coding exon starts and ends, and that SiPhy-pi elements had slightly higher precision for TSS of protein coding genes at their one recall point.

We also evaluated recovering bases covered by DHS (**Figures 2.24-25, Methods**). When comparing DHS recovery from 53 samples in aggregate, the conservation states had greater precision at the same recall level than all the constraint scores and PhastCons elements, both genome-wide and for non-exonic bases. The precision for GERP++, SiPhy-pi and SiPhy-omega elements was higher at their single recall point (**Figure 2.24**). Similar results were seen for regions distal to TSS, except for some scores at low recall levels in the non-exonic comparison. The higher precision for GERP++, SiPhy-pi and SiPhy-omega elements in the aggregate evaluation over constraint scores, PhastCons elements, and conservation states might be related to the coarser resolution at which they were defined and also did not hold for all cell types (**Figures 2.21, 2.25**).

Conservation states also had complementary information about DHS to constrained elements, as constrained element enrichments for DHS varied substantially depending on their conservation state (**Figures 2.22d, 2.26-27**). For example, PhastCons elements bases in 35 states were depleted for Fetal Brain DHS in non-exonic regions, covering 10% of PhastCons bases, while PhastCons elements bases in 12 states bases were enriched over 5-fold, covering

37% of PhastCons bases. Additionally, bases not in a constrained element in some states had greater enrichments for DHS than bases in a constrained element in other states. Constrained elements also offered additional information, as in most cases bases that were in a constrained element in a given conservation state had greater enrichment for DHS than those that were not.

We also analyzed conservation state enrichments for previously defined subsets of PhastCons constrained non-exonic elements (CNEEs) based on a directed phylogenetic approach that assigned each element to a phylogenetic branch point of origin[25] (**Figure 2.28a**). Bases in elements assigned to the Tetrapod clade branch point of origin had a 37-fold enrichment for state 2, which had high aligning and matching through all vertebrates except fish, but also 51-fold enrichment for state 100, associated with likely alignment artifacts, demonstrating the heterogeneous nature of assignments from directed phylogenetic partitioning. We also evaluated the subsets of CNEEs enrichment for CpG islands within non-exonic regions (**Figure 2.28c**). The most enriched subset of CNEEs was 6.7-fold enriched covering 1.9% of non-exonic CpG islands. In comparison, conservation state 28 had a 37.6-fold enrichment, while covering 12.8% of such bases. A similar pattern of enrichments was observed when only considering CNEEs overlapping a PhastCons element called on the same alignment as the conservation states (**Figures 2.28b,d**). These results highlight that the conservation states capture additional biological information compared to directed phylogenetic based approaches.

Conservation states enrichments for prioritized variants

Various scores have been proposed to prioritize variants, including based on inter- or intra-species constraint or integration of diverse genomic annotations. However, a systematic understanding of different types of bases these scores prioritize is generally lacking. To address this, we analyzed conservation states' genome-wide enrichments of top 1%, 5%, and 10% prioritized bases by 12-different scores (CADD (v1.4), CDTS, DANN, Eigen, Eigen-PC,

FATHMM-XF, FIRE, fitCons, GERP++, PhastCons, PhyloP, and REMM). We also analyzed the enrichment specifically in non-coding regions for those scores and two non-coding only scores, LINSIGHT and FunSeq2 (**Figures 2.29a,b**, **2.30-32**)[3,4,6,19,21,22,44–50].

Bases prioritized by most scores had strong enrichments for specific conservation states. For example, state 1, which had high align and match probabilities across all vertebrates, had a 77.2-fold enrichment for CADD top 1% prioritized bases genome-wide, covering 46% of such bases. Despite the CADD score being based in part on many non-conservation annotations, this enrichment was greater than that observed for any inter-species constraint score. There was a general consistency in states with higher enrichment across the various measures. For example, in top 1% bases for the genome-wide analysis, only 13 states were among the top five most enriched by at least one of the 12 scores. Nine of these 13 states (states 1-5, 7, 28, 54, 100) were in the top five for at least three scores. However, there were also important enrichment differences between scores for these states, and in several cases a single score prioritized other states.

There was substantial disagreement among the scores of the relative importance of states 2 and 28, the most enhancer and promoter enriched states respectively, particularly in non-coding regions. For example, state 2 was the second or third most enriched state (24.9 to 47.2-fold) for CADD, Eigen, FATHMM-XF, GERP++, LINSIGHT, PhastCons, PhyloP, and REMM top 1% prioritized bases in non-coding regions. On the other hand, state 28 had lower enrichments (0.3 to 6.2-fold) and was not one of the top five most enriched states for any of those scores. In contrast, for CDTS, DANN, and Eigen-PC, state 28 was the first or second most enriched state (7.6 to 18.6-fold), while state 2 had lower enrichments (0.8 to 2.1-fold) and was not among the top five most enriched states.

There was a large disagreement in the state enrichments between variants prioritized by DANN and CADD for both the current and original versions of CADD (**Figures 2.31-33**). This

was despite DANN using the same framework as CADD except using a deep neural network[46]. Surprisingly, for top 1% non-coding variants, DANN showed a depletion for state 2, which had high matching probabilities through all vertebrates except fish, while having over four-fold enrichment for multiple states that showed high alignment or matching probabilities for only subsets of primates.

There were also notable enrichment differences for other states for which the biological importance was less apparent. For example, state 100, associated with likely alignment artifacts, in the top 1% non-coding region analysis had enrichments in the range 14.7 to 34.5-fold for FATHMM-XF, fitCons, PhastCons and PhyloP prioritized bases, while the enrichment for all other scores was at most 2.0-fold. Another example was state 54, which associated with wobble position within codons, and had a 21.1 fold enrichment in the top 1% genome-wide analysis for fitCons prioritized bases and was also the third most enriched state for CDTS, Eigen-PC, and FIRE, while depleting for GERP++ and REMM prioritized bases. These results highlight how the conservation states enable recognizing and characterizing distinct subsets of nucleotides that are selectively captured by different variant prioritization scores.

Conservation states and human genetic variation

Previous analyses have found a depletion of human genetic variation in evolutionarily constrained elements[5]. Consistent with that, the greatest depletion (3.3-fold) of common single nucleotide polymorphisms (SNPs) is in state 1, the state most enriched for constrained elements, while states 55-57 and 87-89 had the greatest enrichments for common SNPs (5 to 8-fold). These six states all had high align, but low match probabilities for most primates and had the greatest enrichment of CG dinucleotides (**Figure 2.15**). We observed similar patterns of enrichments and depletions for variants identified from whole genome sequencing[50], with their magnitude increasing with minor allele frequency (**Figure 2.34**).

States had opposite enrichment patterns for GWAS catalog variants[51] relative to the background of common SNPs (**Figure 2.29c,d**). Using this background, state 1 was most enriched for GWAS catalog variants, consistent with constrained elements enriching for GWAS variants[5]. States 55-57 and 87-89 showed the greatest depletion, suggesting that a variant in one of these states is less likely to be phenotypically associated.

We also applied the INSIGHT[52] model to obtain its estimates of the density of positive selection events and percentage of bases under selection within human populations in each conservation state (**Figure 2.35**). States 54-57 and 87-89 all had substantial density for positive selection event estimates. INSIGHT also estimated that 77% of states had more than 75% bases under-selection, while 13% had less than 50% bases under selection. Similar estimates held when restricting to bases in PhastCons elements and not in PhastCons elements (**Figure 2.35**). However, instead of a majority of states actually having a high percentage of bases under selection, this likely reflects that there is a relatively direct relationship between human variation information contained by the conservation states and INSIGHT's use of such information to quantify selection.

Conservation states and heritability partitioning

Previous analyses have suggested strong enrichments of constrained elements and DHS for phenotype heritability[18,53]. Given the differences in DHS enrichments of constrained elements across conservation states, we investigated whether constrained elements in conservation states most enriched for DHS had different phenotype heritability than those in other states. Specifically, we ranked the conservation states in descending order of their median enrichment within non-exonic bases for DHS from 123 experiments (**Figure 2.2b, Methods**)[8]. We then partitioned bases in PhastCons elements into two almost equal size sets based on whether they overlapped a top seven-ranked conservation state (states 1-5, 8, 28). We

computed the heritability for the two sets for eight phenotypes in the context of baseline annotations that include DHS[18]. For seven of the phenotypes, bases in constrained elements overlapping the top seven states had greater enrichment than those in the other states, often substantially so (**Fig. 2.29e**). These results suggest possible additional value of conservation states for isolating disease-associated variants.

2.3 Discussion

We introduced the ConsHMM method for genome annotation and used it to annotate the human genome at single nucleotide resolution into one of 100 conservation states. ConsHMM learns conservation states *de novo* using a multivariate HMM based on the combinatorial and spatial patterns of which species align and match a reference genome in a multi-species DNA sequence alignment. Conservation states had substantial enrichments for a wide range of other genomic annotations, functional genomics data, and human variation data.

ConsHMM differs from other commonly used comparative genomics based annotation approaches in several respects. One difference is that it takes an unsupervised approach that does not explicitly use a phylogenetic tree in its modeling. This leads to relatively unbiased, flexible and interpretable models. Despite not explicitly using a phylogenetic tree, many state patterns discovered are consistent with commonly assumed phylogenetic relationships of the species. While states' parameters often decreased with divergence time from human, there were some exceptions. Some of these exceptions corresponded to missing specific sub-clades of species, particularly those with long branch lengths. For example, in some states mouse and rat were absent, while more distally diverged mammals were present. Other states isolated likely artifacts in alignments that heavily enriched for pseudogenes. A second difference is that ConsHMM explicitly differentiates non-aligning bases from aligning non-matching bases, which allowed it, for example, to identify states such as those associated with third codon positions. A

third difference between the ConsHMM annotations and standard constraint measures is that the ConsHMM annotations are defined directly relative to the variant present in the genome being annotated. When applying ConsHMM to annotate the human genome, a mutation unique to human would be expected to have a much larger effect on the ConsHMM annotations than a mutation unique to a single other species. This would not in general be expected for constraint measures that treat the target genome for annotation in the same way as other genomes in an alignment. An interesting future direction would be to produce and analyze individual specific ConsHMM annotations.

ConsHMM annotations are complementary to existing binary elements and scores of evolutionary constraint based on phylogenetic modeling. Both bases within and outside of constrained elements are heterogeneous in their assigned conservation states. ConsHMM annotations provide additional information about the conservation patterns at each base. In many cases, the conservation states had greater information than constraint scores or elements for predicting external annotations. Notably, ConsHMM identified a conservation state strongly enriched for TSS and CpG islands that was not well captured by phylogenetic modeling approaches. For other annotations, such as DHS, the relative information depended on the constrained element set or score being compared. Importantly, the DHS information provided by the states was complementary to information in the constrained elements. Furthermore, we observed that bases in constrained elements showed substantially different enrichments for phenotype-associated heritability, depending on their conservation state. The conservation state annotations also provide a useful framework for understanding the types of bases prioritized by constraint scores or other types of variant prioritization scores, since the corresponding conservation patterns are defined systematically in an unbiased way, at single nucleotide resolution and capture a diverse set of biological features.

ConsHMM is both inspired by, and provides complementary information to, ChromHMM. While the annotations produced by two methods have fundamental differences, they also exhibited substantial cross-enrichments. In general, conservation states have the advantages of providing information at single nucleotide resolution and about bases active in cell types that have not been experimentally profiled, while chromatin states have the advantage of directly providing cell type specific information.

We expect many applications for the ConsHMM method and annotations. The ConsHMM method can be readily applied to alignments to other reference species or alignments by other methods[29]. The ConsHMM annotations are a resource to interpret other genomic datasets or variant prioritization scores. A possible avenue for future work would be to integrate the conservation states with other genomic annotations to produce a variant prioritization score. An effective strategy for that would need to be powered to retain the rich information in the conservation state annotations, and would also need to be based on a principle sufficiently independent from how the conservation states are defined to enable a meaningful integration and prioritization. This work represents a step towards improving whole genome annotations, including of non-coding regions and variants, which will be of continued importance towards understanding disease.

2.4 Methods

Modeling conservation states with ConsHMM

ConsHMM takes as input an *N-way* multi-species sequence alignment to a designated reference genome. For each base in the reference genome, i, ConsHMM encodes information from the multiple species alignment into a vector, v_i, of length N-1. An element of the vector, $v_{i,j}$, corresponds to one of three possible observation for a non-reference species j at position i. The three possible observations are: (1) the non-reference species aligns with a non-indel nucleotide

symbol present matching the reference nucleotide, (2) the non-reference species aligns with a non-indel nucleotide symbol present, but does not match the reference nucleotide, or (3) the non-reference species does not align with a non-indel nucleotide symbol present.

ConsHMM assumes that these observations are generated from a multivariate HMM where the emission parameters are assumed to be generated by a product of independent multinomial random variables, corresponding to each non-reference species in the alignment. Formally, the model is defined based on a fixed number of states K, and number of species in the multiple sequence alignment N. For each state k (k = 1,…,K), non-reference species j (j = 1,...,N-1) and possible observation m (m = 1, 2, or 3 as described above), there is an emission parameter: $p_{k,j,m}$ corresponding to the probability in state k for species j of having observation m. For each possible observation m, let $I_m(v_{i,j})$ = 1 if $v_{i,j} = m$, and 0 otherwise. Let $b_{t,u}$ be a parameter for the probability of transitioning from state t to state u. Let $c \in C$ denote a chromosome, where C is the set of all chromosomes in the reference genome of the multiple species alignment, and let L_c be the number of bases on chromosome c. Let a_k (k = 1,…,K) be a parameter for the probability of the first base on a chromosome being in state k. Let $s_c \in S_c$ be a hidden state sequence on chromosome c and S_c be the set of all such possible state sequences. Let c_h denote position h on chromosome c. Let s_{c_h} denote the hidden state at position c_h for state sequence s_c.

We learn a setting of the model parameters that aims to optimize

$$P(v|a,b,p) = \prod_{c \in C} \sum_{s_c \in S_c} a_{s_{c_1}} \left(\prod_{i=2}^{L_c} b_{s_{c_{i-1}}, s_{c_i}} \right) \prod_{h=1}^{L_c} \prod_{j=1}^{N-1} \prod_{m=1}^{3} p_{s_{c_h},j,m}^{I_m(v_{c_h,j})}$$

Once a model is learned, each nucleotide is assigned to the state with maximum posterior probability. To conduct the model learning and state assignments, ConsHMM calls an

extended version of the ChromHMM[34] software, originally designed to solve an analogous problem of annotating a genome into chromatin states based on combinatorial and spatial patterns of the presence of different chromatin marks. The modeling in ConsHMM differs from the typical use of ChromHMM in three main respects: (1) the observation for each feature comes from a three-way multinomial distribution as opposed to a Bernoulli distribution, (2) it is applied at single nucleotide resolution as opposed to 200-bp resolution, (3) it is applied with more features than ChromHMM models have used in the past. (2) and (3) raise scalability issues in terms of time and memory, which we addressed in an updated version of ChromHMM (see below).

To apply ChromHMM in the context of three-way multinomial distributions, ConsHMM represents the three possible observations at position *i* for a species *j* with two binary variables, y_{ij} and z_{ij}, corresponding to aligning and matching the reference genome respectively. y_{ij} has the value of 1 if the other species aligns to the reference with a non-indel nucleotide and 0 otherwise. z_{ij} has the value of 1 if the other species has the same nucleotide as the reference sequence and has a value of 0 if the other species has a different nucleotide present than the reference. In the case in which y_{ij}=0, there is no nucleotide to compare to the reference and that value of the z_{ij} variable is considered missing (encoded with a '2' for ChromHMM). If the value of an observed variable is missing, ChromHMM excludes the Bernoulli random variable corresponding to the observation from the emission distribution calculation at that position. For each state *k* and species *j*, ChromHMM thus learns two parameters, $f_{k,j}$ and $g_{k,j}$. $f_{k,j}$ corresponds to the probability that at a given position in state *k*, species *j* aligns to the reference genome with a non-indel nucleotide, that is $P(y_{i,j}=1|\ s_i=k)$. $g_{k,j}$ corresponds to the probability that at a given position in state *k*, species *j* matches the reference genome conditioned on species *j* aligning with a non-indel nucleotide, that is $P(z_{i,j} = 1|\ y_{i,j}=1$ and $s_i=k)$. This representation is equivalent to the three-way multinomial distribution, $(p_{k,j,1},\ p_{k,j,2},\ p_{k,j,3})$ described above where $p_{k,j,1} = P(y_{i,j}=1,$

$z_{i,j}=1 \mid s_i = k)$, $p_{k,j,2} = P(y_{i,j}=1, z_{i,j}=0 \mid s_i = k)$, and $p_{k,j,3} = P(y_{ij}=0 \mid s_i = k)$, since $p_{k,j,1} = f_{k,j} \times g_{k,j}$, $p_{k,j,2} = f_{k,j} \times (1-g_{k,j})$, and $p_{k,j,3} = 1 - f_{k,j}$.

Multiple species sequence alignment choice

ConsHMM can be applied to any multiple species sequence alignment which is available in multiple alignment format (MAF) or which can be converted into this format. For the results presented here we applied it to the 100-way Multiz vertebrate alignment with human (hg19) as the reference genome[24,36].

Scaling-up ConsHMM to single base resolution

Since for our application ConsHMM needs to run ChromHMM at single base resolution ('-b 1' flag) with 198 features after our binary encoding (2 for each non-human species in the 100-way alignment), we had to address scalability issues in terms of both memory and time. To address the memory issue we modified ChromHMM to support only loading in main memory input for chromosomes it is actively processing, as previously ChromHMM would only support loading all data into main memory upfront. This option can now be accessed in ChromHMM through the '-lowmem' flag. To reduce the time required we used 12-parallel processors ('-p 12' flag) and we trained on a different random subset of the human genome on each iteration of the Baum-Welch algorithm. We divided each chromosome into 200kb segments (with the exception of the last segment of each chromosome which was less than this) in order to form random subsets of the human genome. We modified ChromHMM to allow training for each iteration on a randomly selected subset of 150 of these segments ('-n 150' flag), corresponding to 30MB per iteration. We ran this for 200 iterations by adding the '-d -1' flag, which removed one of ChromHMM's default stopping criterion based on computed likelihood change on the sampled data, since the likelihood is now expected to both increase and decrease between iterations as

different segments are sampled. These new options were included in version 1.13 of ChromHMM. The unique code to ConsHMM v1.0 is written in Python. The code of ConsHMM shared with ChromHMM is written in Java and included with ConsHMM.

Generating genome-wide annotations

After ConsHMM learned a state model, we used it to segment and annotate the human genome at base-pair resolution into conservation states. Each base in the human genome is classified into the state with the highest posterior probability. ConsHMM does this by running the MakeSegmentation command of ChromHMM. Due to computational constraints, the segmentation could not be generated for entire chromosomes at once. Instead, we ran MakeSegmentation on the same 200kb partitioning made for learning the model. We then merged the resulting files together using ConsHMM's mergeSegmentation.py command with slice size parameter set to 200,000 ('-s 200000' flag) and the number of states parameter set to 100 ('-n 100 flag').

Computing enrichments for external annotations

All overlap enrichments for external annotations were computed using the ChromHMM OverlapEnrichment command. OverlapEnrichment computes enrichments for an external annotation in each state assuming a uniform background distribution. Specifically the fold enrichment of a state for an external annotation is

$$\frac{\%\ of\ external\ annotation\ bases\ falling\ in\ that\ state}{\%\ of\ genome\ falling\ in\ that\ state}$$

Positional enrichments of states relative to an anchor point from an external annotation were computed using the ChromHMM NeighborhoodEnrichment command at single base

resolution ('-b 1' flag), single base spacing from the anchor point ('-s 1') and using the '-l' and '-r' flags to specify the size of the region of interest around the anchor point. The '-lowmem' flag was also used for computing the enrichments for OverlapEnrichment and NeighborhoodEnrichment.

External data sources for enrichment analyses

The external annotations of repeat elements were obtained from the UCSC genome browser RepeatMasker track[24,42]. We generated an annotation for whether a base overlapped any repeat element, as well as separate annotations for bases falling in each class and family of repeat elements. The gene annotations were obtained from GENCODE v19 for hg19[54]. CpG island annotations were obtained from the UCSC genome browser. Annotations of SNPs with >=1% minor allele frequency were obtained from the commonSNP147 track from the UCSC genome browser, which is based on dbSNP build 147. GWAS catalog variants were obtained from the NHGRI-EBI Catalog, accessed on Dec 5, 2016[51]. For annotations of DNase I Hypersensitive Sites (DHS) processed by the Roadmap Epigenomics Consortium, we used Macs2 narrowPeak calls[9]. The Fetal Brain and HepG2 DHS used were of epigenome samples E082 and E118 respectively. For the median non-exonic DHS enrichments and ranking of states in the heritability partitioning analysis we used narrowPeak calls from the ENCODE consortium[8]. In the cases where ENCODE provided more than one replicate for a cell or tissue type, we used the first replicate.

PhyloP and PhastCons scores and constrained element calls were obtained from the UCSC genome browser. Assembly gap annotations were obtained from the Gap track from the UCSC genome browser. The context-dependent tolerance score (CDTS) used was that based on a cohort of 7784 unrelated individuals, following the analyses in Ref. 47, which focused on

this version of the score. The CDTS and variants from this cohort were both lifted from hg38 to hg19 using the liftOver tool from the UCSC genome browser[24].

Choice of number of states

We learned models with each number of states between 2 and 100 states. We set 100 as the maximum number of states we would consider for computational tractability and maintaining a manageable number of states for analysis. The choice of a maximum of 100 also corresponds to the number of species used and allows for the possibility of each state to cover 1% of the genome. We analyzed the Bayesian Information Criterion (BIC) for models with each number of states between 2 and 100, and found that the BIC generally decreases as the number of states increases in the range considered (**Figure 2.3**). The BIC was calculated using the BIC_HMM function from the HMMpa R package[55]. Analyzing the 100-state model's internal confidence estimate of its state assignments also supported a larger number of states. Specifically, for each state in the 100-state model we computed the average posterior probability of that state at each base in the genome assigned to it, and confirmed consistently high average posterior probability values in the range [0.92,1.00] with a median of 0.97 (**Figure 2.3**). The posterior probabilities were computed by running the MakeSegmentation command in ChromHMM with the '-printposterior' flag. We also investigated if additional states in models with larger number of states were biologically relevant. Specifically, we computed enrichments for various external annotations for models with each number of states between 2 and 100 to determine if biologically relevant enrichments were only robustly observed in models with more than a certain number of states. In the case of CpG islands, we observed that only models with at least 87 states consistently obtained >15 fold enrichment and only models with at least 95 states consistently obtained >30 fold enrichment (**Figure 2.4**). We saw a similar pattern of increasing enrichments for annotated TSS for models with large number of states. We therefore

decided to analyze the largest model, 100 states, that we were considering. We note that annotations based on chromatin states used fewer number of states, but were also defined on fewer features at a coarser resolution and had a less uniform genome coverage[14,33,43].

State clustering

We clustered the states based on the correlation of vectors containing the values $f_{k,j}$ and $f_{k,j} \times g_{k,j}$ for each species j defined above. State clustering was performed using the hclust hierarchical clustering function from the cba R package[56]. The leaves of the resulting hierarchical tree were ordered according to the optimal leaf ordering algorithm[57] implemented in the order.optimal R function from the cba package. We then cut the tree such that the 8 major groups of states were designated. The full tree is provided in **Figure 2.6**.

Genome segmentation using uniform transition probabilities

For analyzing the effect of the transition probabilities on the genome segmentation, we created a separate model, which was the same model we used in the main analyses, except we set all transition probabilities to 0.01, corresponding to each state having an equal probability of transitioning to any state including itself. We then created a new genome segmentation by running the MakeSegmentation command in ChromHMM with this new model. For each state, we counted how many of the bases assigned to it in the original annotation were also assigned to it in the annotation created with the uniform transitions, and divided this number by the number of bases in the state in the original annotation. This calculation provided a fraction from 0 to 1. We also reported the number of segments produced by each model, where a segment is defined to be one or more consecutive bases all assigned to the same state, such that any immediately adjacent bases are assigned to a different state or states.

GO enrichments

For each state and each protein-coding gene based on GENCODE, we computed the number of bases in that state that are within +/- 2kb of the gene's TSS. In the case of genes with multiple annotated TSS, we used the outermost TSS. We then created a ranking of genes for every state by sorting the genes in descending order of this number of bases. For each state, we then created a set of 969 genes that represent the top 5% of genes in the state among the 19,397 genes we considered. We performed a GO enrichment analysis (ontology and annotations files from Nov. 24th, 2016) for the top 5% genes in each state using the STEM v1.3.10 software in batch mode with default options and the set of all genes considered as background[58]. STEM computed an uncorrected p-value based on the hypergeometric distribution for each term displayed in the figures summarizing the analysis. STEM also reported corrected p-values for testing multiple GO terms for a single state based on randomization to three significant digits, which was less than 0.001 for all p-values mentioned in the main text.

Transcription factor binding site motif enrichments

We computed the fold enrichment of the conservation states within 15 bases upstream and downstream of the center point of the *POU5F1* and *STAT* known transcription factor-binding site motifs[40]. The enrichment was computed relative to the background regions of the genome that were used to identify the motifs, which excluded repeat elements, coding sequence, and 3' untranslated regions (UTRs). We used the *known1* version of the motifs for both *POU5F1* and *STAT*.

Clustering of cell-type specific DHS enrichments

For the clustering of DHS analysis, we first computed the fold enrichments of all conservation states for DHS for 53 samples processed by the Roadmap Epigenomics

consortium[9], of which 16 were originally generated by the ENCODE project consortium[8]. We then selected the subset of states that had a fold enrichment of at least 2 in at least one sample, leading to a subset of 21 conservation states. To more directly focus on each state's relative enrichments across samples, we $\log_2$ transformed each enrichment value, and then normalized the enrichments for each state by subtracting the mean enrichment across samples and dividing by the standard deviation. We then hierarchically clustered the states based on the correlation of their enrichments across samples and hierarchically clustered the samples based on their correlations across states using the pheatmap R package[59]. We also computed for each sample the fold enrichment of DHS bases for bases in CpG islands, as the ratio between the percent of DHS bases in CpG islands and the percent of the genome falling in CpG islands.

Precision recall analysis for recovery of gene annotations

We randomly split the 200kb genome segments used for training the model and segmentation into two halves corresponding to training and testing data. For each target set in the precision-recall analyses, we ordered the ConsHMM states in decreasing order of their enrichment for the target among the training set bases. We then used that ordering to iteratively add the testing set bases in each state to form cumulative sets of bases predicted to be of the target set, and computed the precision and recall for them. For each constraint score, we computed the precision-recall curve for predicting the target set in the test data using two methods. For the first method, we directly ordered bases in descending order of their assigned score. For the second method, we split the sorted scores into 400 bins such that each bin contains on average 0.25% of the genome, which was the size of the smallest state of the ConsHMM model (0.25% of the genome in state 100). Specifically, we assigned all bases in the genome where the score was not defined to one bin and then divided the remaining bases uniformly among the 399 other bins based on their score. In some cases, score increments

were at the boundary between two bins at their provided floating-point precision, or overlapped multiple bins. In these cases, we uniformly split the target bases assigned to that score increment into multiple bins proportionally to the overall percentage of the score increment falling in each bin. We then treated the 400 bins as 400 states and followed the same procedure described for the ConsHMM states. We also computed the precision and recall of bases in each constrained element set for predicting the target set on the testing data.

Precision recall analysis for recovery of DHS

For the precision recall analysis for recovery of DHS analysis for a single cell type, we followed the same procedure described above. We also separately evaluated recovery of DHS bases when restricting the analysis to non-exonic regions. Additionally, both genome-wide and within non-exonic regions, we evaluated the recovery of DHS bases when restricting the analysis to bases distal to a TSS, defined as more than 2kb from a TSS. For the analysis of the recovery of DHS aggregated across cell and tissue types we concatenated DHS from 53 cell or tissue types processed by the Roadmap Epigenomics Consortium into one annotation in which each combination of chromosome and cell or tissue type effectively becomes a new chromosome. We then split the concatenated data into training and testing sets as described above. We computed the enrichments of the ConsHMM states and scores split into bins as detailed above, but multiplying the size of each state and bin by the number of DNase I hypersensitivity data sets. The precision and recall values for the ConsHMM states, constraint scores considered directly, constraint scores split into bins, and constrained element sets were then computed on the testing data.

Enrichment analysis for phylogenetically partitioned CNEEs

We lifted over the CNEEs from Ref. 22 from hg18 coordinates to hg19, using the liftOver tool from the UCSC genome browser with default settings[24]. These elements were previously partitioned into subsets based on the inferred branch point of origin in a phylogenetic tree[25]. We computed the enrichments of the conservation states for all the CNEEs and for each subset of the CNEEs separately, using the OverlapEnrichment command from ChromHMM at single nucleotide resolution ('-b 1' flag) and using the low memory option ('-lowmem'). We also computed analogous enrichments for CNEEs overlapping PhastCons elements called on the same 100-way alignment that the conservation states were annotated based on. To compute the enrichments of CNEEs for bases in CpG islands we created an annotation consisting of a state for each CNEE subset and one additional state for bases not assigned to any CNEE. We then ran the same OverlapEnrichment command as above to compute enrichments of CNEE bases for non-exonic CpG islands, and non-exonic bases in general. The reported enrichment of CpG islands is the ratio of these two enrichments, effectively computing an enrichment relative to the non-exonic background. The set of non-exonic bases for the enrichment analysis was generated by excluding all bases annotated as an exon in GENCODE v19.

Heritability partitioning analysis

The heritability partitioning was performed using the LD-score regression ldsc software[18]. We partitioned the PhastCons constrained elements into two halves based on a ranking of the conservation states. We focused on the PhastCons constrained elements for this analysis, since it was the only element set defined on the same alignments as the conservation states. We focused on halves since the LD-score regression estimates can be unstable for annotations covering too small of a percentage of the genome[18]. To determine the two halves we ranked the conservation states in descending order of median fold-enrichment of non-exonic bases for DHS from 123 experiments from the University of Washington ENCODE group[8]. We

then divided bases in PhastCons elements between the top 7 ranked states (1-5, 8 and 28), which contain 51.9% of bases in PhastCons elements, and the bottom 93 states, which contain the other 48.1% of bases in PhastCons elements. We applied ldsc to these two sets for 8 traits (age at menarche, body mass index (BMI), coronary artery disease, educational attainment, height, low-density lipoprotein (LDL) levels, schizophrenia and smoking behavior), all of which were previously considered in heritability partitioning analysis[18]. We followed the procedure for partitioning heritability as done in Ref. 15, including using the baseline annotation set and 500 base-pair windows around annotations to dampen the artificial inflation of heritability in neighboring regions caused by linkage disequilibrium. The baseline annotation set contains a range of annotations including DHS. For our analysis, we first removed the constrained element set already included in the baseline annotation set, then added our two halves of PhastCons elements and finally ran the ldsc software on the full set of annotations.

Enrichment analysis for variant prioritization scores

For each variant prioritization score included in the conservation state enrichment analysis of prioritized bases, we extracted the top 1%, 5% and 10% of all the bases ranked by each score, both genome-wide and just in non-coding regions. The non-coding regions were defined as the intersection of where the LINSIGHT and FunSeq2 scores provided a value, as these two scores were only defined on non-coding regions. This intersection results in a set of bases covering 90% of the genome that excludes coding regions in addition to other regions filtered for technical reasons by either of the two methods[22,44]. For each score we chose the score threshold that gave us a size for the top set that was as close as possible to the target percentage, which did not always exactly match the target percentage due to the precision of the scores. If a score did not provide a value for a particular base being considered, then that base was assigned to the lowest value of that score, but would still be counted when

establishing the percentage thresholds. For the scores that provided separate score values for alternate alleles at a certain position, we used the maximum of the values for all alleles. The state enrichments were then computed using the OverlapEnrichment command from ChromHMM at single base resolution ('-b 1' flag) and with the low memory option ('-lowmem' flag). For the analysis restricted to non-coding regions, we also computed the enrichment of the states for this background region using the same command. The enrichment for each score in a state was then divided by the enrichment of the background region for the state. For the Eigen and Eigen-PC scores we used version 1.1, for FunSeq2 we used version 2.1.6, and for CADD we used both v1.0 and v1.4.

INSIGHT analysis

The INSIGHT[52] package was used with parameters of 15% allele frequency threshold, 100 minimum neutral flanking sites and the optimizer method BFGS_DIRECT for the OPT_METHOD flag.

Data availability

The ConsHMM conservation state annotations of hg19 are available at https://doi.org/10.6084/m9.figshare.8162036.v1 and https://github.com/ernstlab/ConsHMM. The input multiple species alignment for producing the conservation state annotations is available at http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/. The following URLs contain data sets that were used in the downstream analyses: 25-state chromatin state annotations: http://compbio.mit.edu/roadmap; CADD score v1.0: http://krishna.gs.washington.edu/download/CADD/v1.0/whole_genome_SNVs.tsv.gz; CADD score v1.4: http://krishna.gs.washington.edu/download/CADD/v1.4/GRCh37/whole_genome_SNVs.tsv.gz;

CDTS score: http://www.hli-opendata.com/noncoding/coord_CDTS_percentile_N7794unrelated.txt.gz, http://www.hli-opendata.com/noncoding/SNVusedForCDTScomputation_N7794unrelated_allelicFrequency0.001truncated.txt.gz; CNEEs from Ref. 22: http://www.stanford.edu/~lowec/data/threePeriods/hg19cnee.bed.gz; DANN score: https://cbcl.ics.uci.edu/public_data/DANN/data/; EIGEN and Eigen-PC score: https://xioniti01.u.hpc.mssm.edu/v1.1/; ENCODE DHS: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/; FATHMM-XF score: http://fathmm.biocompute.org.uk/fathmm-xf/; FIRE score: https://sites.google.com/site/fireregulatoryvariation/; fitCons score: http://compgen.cshl.edu/fitCons/0downloads/tracks/i6/scores/; FunSeq2 score: http://org.gersteinlab.funseq.s3-website-us-east-1.amazonaws.com/funseq2.1.2/hg19_NCscore_funseq216.tsv.bgz; GENCODE v19: https://www.gencodegenes.org/releases/19.html; GERP++ scores and constrained element calls: http://mendel.stanford.edu/SidowLab/downloads/gerp/; GWAS catalog variants: https://www.ebi.ac.uk/gwas/; LINSIGHT score: http://compgen.cshl.edu/~yihuang/tracks/LINSIGHT.bw; Motif instances and background: http://compbio.mit.edu/encode-motifs/; REMM score: https://zenodo.org/record/1197579/files/ReMM.v0.3.1.tsv.gz; Roadmap Epigenomics DHS: http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/; SiPhy-omega and SiPhy-pi constrained element calls (hg19 liftOver): https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info

Code availability

The ConsHMM software is available through https://github.com/ernstlab/ConsHMM. The ChromHMM software used for enrichment analyses and on top of which ConsHMM is built is available at http://www.biolchem.ucla.edu/labs/ernst/ChromHMM/. The STEM software used for GO enrichment analysis is available at http://sb.cs.cmu.edu/stem/. The ldsc software used for the heritability partitioning analysis is available at https://github.com/bulik/ldsc. The INSIGHT software used for selection analyses is available at http://compgen.cshl.edu/INSIGHT/downloads/INSIGHTpackage/.

2.5 Figures

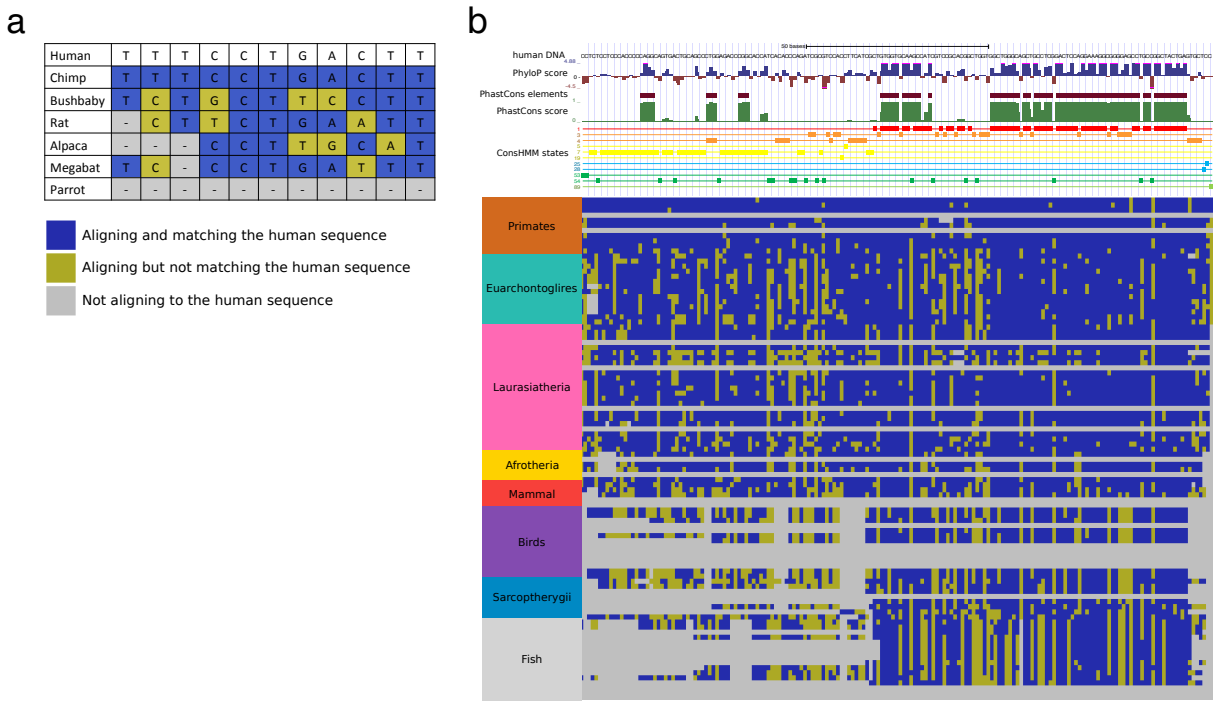


Figure 2.1 Illustration of ConsHMM modeling approach**.**

(a) The input to ConsHMM is a multiple species alignment, which is illustrated for a toy example of 6 species aligned to the human sequence. At each position and for each species ConsHMM represents the information as one of three observations: (1) aligns with a non-indel nucleotide matching the human sequence shown in blue, (2) aligns with a non-indel nucleotide not matching the human sequence shown in yellow, or (3) does not align with a non-indel nucleotide shown in gray. **(b)** Illustration of conservation state assignments at the locus chr22:25,024,640-25,024,812 in hg19. Only states assigned to at least one nucleotide in the locus are shown. Below the conservation state assignments is a color encoding of the input multiple species alignment according to panel (a). The major clade of species as annotated on the UCSC genome browser[21] are labeled and ordered based on divergence from human. Above the conservation state assignments are PhastCons constrained elements and scores and PhlyoP constraint scores. This figure and **Figure 2.12** together illustrate that positions of nucleotides that have the same status in terms of being in a constrained element or not or have similar constraint scores can be assigned to different conservation states depending on the patterns in the underlying multiple species alignment.

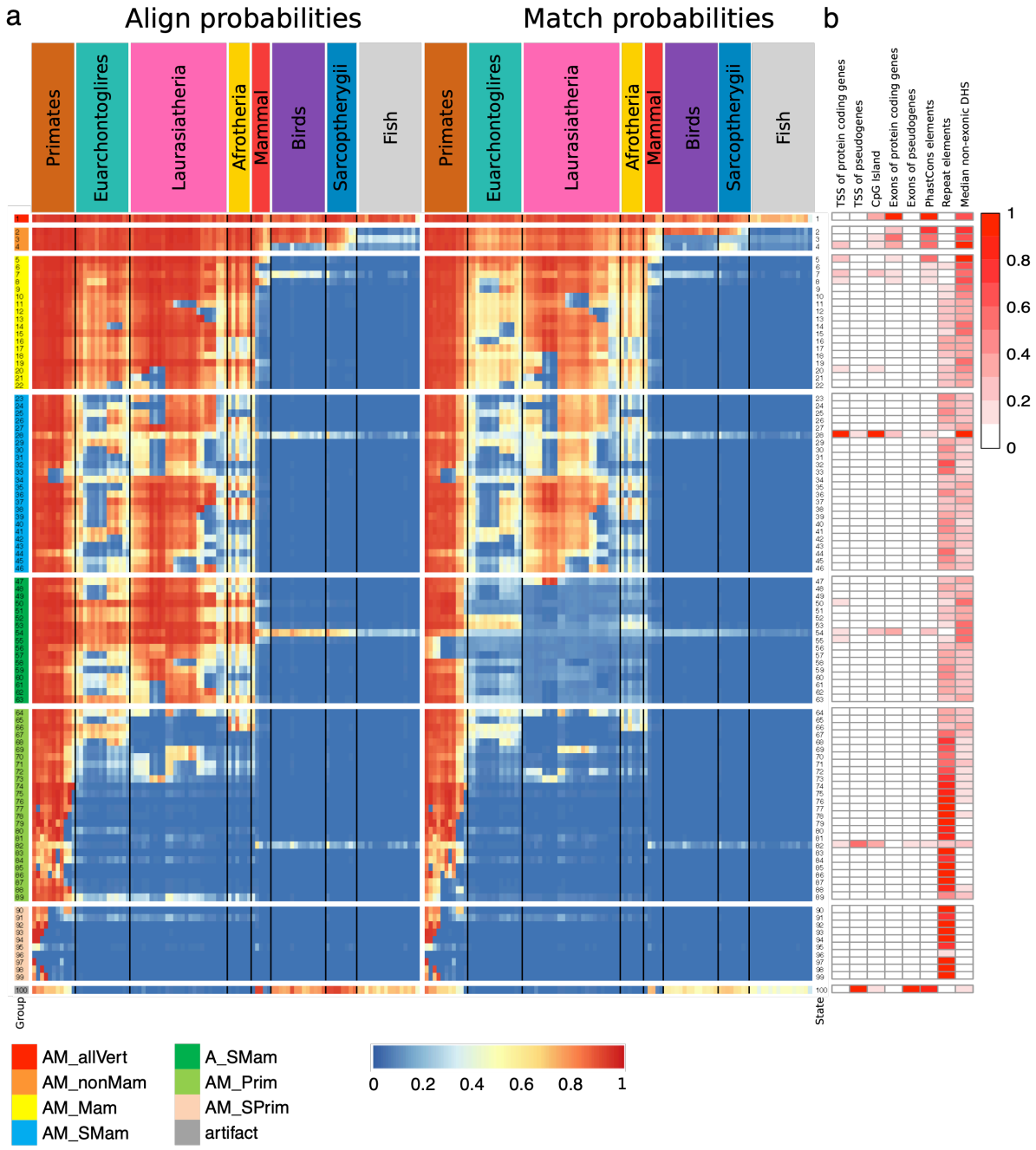


Figure 2.2 Conservation state emission parameters learned by ConsHMM and enrichments for other genomic annotations**.**

(a) Each row in the heatmap corresponds to a conservation state. For each state and species, the left half of the heatmap gives the probability of aligning to the human sequence, which is one minus the probability of the not aligning emission. Analogously, the right half of the heatmap gives the probability of the matching emission. Each individual column corresponds to one species with the individual names displayed in **Figure 2.7**. For both halves, species are grouped by the major clades and ordered based on the hg19.100way.nh phylogenetic tree from the UCSC genome browser, with species that diverged more recently shown closer to the left[21].The

conservation states are ordered based on the results of applying hierarchical clustering and optimal leaf ordering[54]. The states are divided into eight major groups based on cutting the dendrogram of the clustering. The full dendrogram and an explanation of the group mnemonics is available in **Figure 2.6**. The groups are indicated by color bars on the left hand side and a white row between them. Transition parameters between states of the model can be found in **Figure 2.8**. **(b)** The columns of the heatmap indicate the relative enrichments of conservation states for external genomic annotations (**Methods**). For each column, the enrichments were normalized to a [0,1] range by subtracting the minimum value of the column and dividing by the range and colored based on the indicated scale on the right. Values for these enrichments and additional enrichments can be found in **Figure 2.10** and enrichments for individual repeat classes and families can be found in **Figure 2.18**.

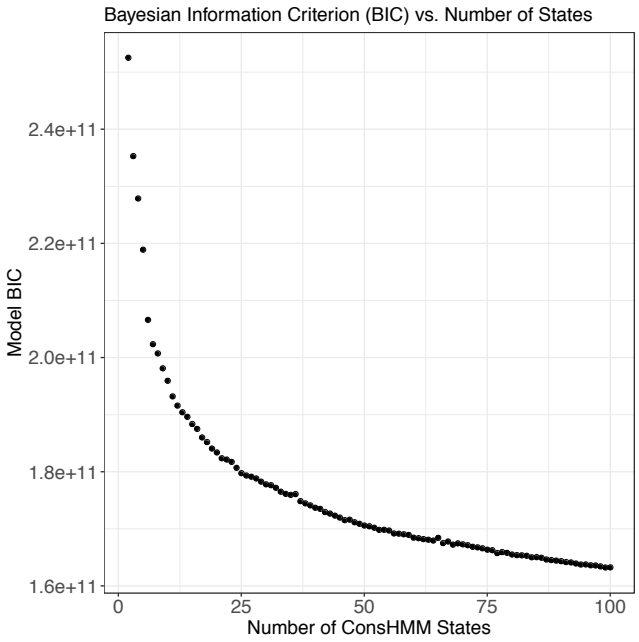


Figure 2.3 BIC as a function of number of states in the model.

The BIC criterion computed for models with each number of states from 2 to 100. For this criterion lower values correspond to preferred models.

State	Average posterior probability of state at positions assigned to state
1	1.00
2	0.99
3	0.99
4	0.98
5	0.96
6	0.96
7	0.97
8	0.92
9	0.93
10	0.94
11	0.98
12	0.97
13	0.97
14	0.95
15	0.92
16	0.92
17	0.95
18	0.94
19	0.97
20	0.98
21	0.98
22	0.96
23	0.97
24	0.98
25	0.97
26	0.97
27	0.96
28	0.98
29	0.97
30	0.98
31	0.97
32	0.97
33	0.98
34	0.98
35	0.97
36	0.96
37	0.97
38	0.97
39	0.97
40	0.97
41	0.96
42	0.95
43	0.97
44	0.96
45	0.97
46	0.97
47	0.93
48	0.95
49	0.96
50	0.96
51	0.95
52	0.93
53	0.95
54	0.98
55	0.97
56	0.96
57	0.96
58	0.96
59	0.95
60	0.95
61	0.96
62	0.96
63	0.96
64	0.97
65	0.98
66	0.99
67	0.98
68	0.98
69	0.97
70	0.95
71	0.94
72	0.96
73	0.98
74	0.99
75	0.99
76	0.99
77	1.00
78	1.00
79	0.99
80	0.98
81	0.98
82	0.99
83	0.99
84	0.99
85	0.99
86	1.00
87	0.97
88	0.98
89	0.95
90	0.99
91	0.98
92	1.00
93	1.00
94	1.00
95	0.99
96	1.00
97	1.00
98	1.00
99	1.00
100	1.00

Figure 2.4 Average posterior probability of ConsHMM state assignments

The value listed for each state is the average posterior probability of that state for all bases in the genome assigned to that state.

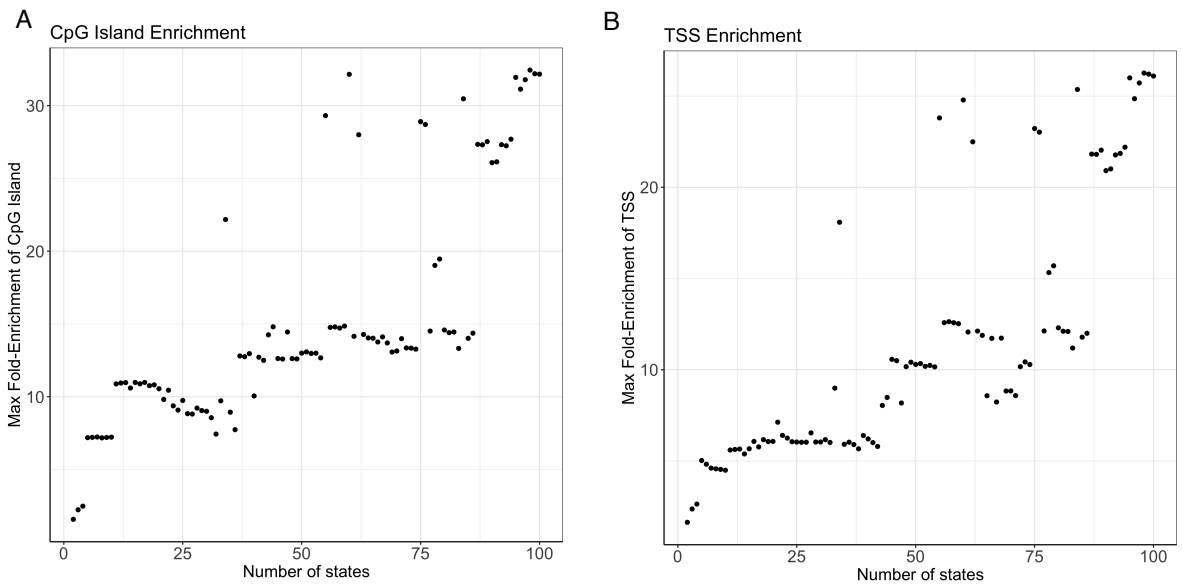


Figure 2.5 Maximum CpG and TSS state enrichments as a function of number of states in the model.

The figures show the maximum fold enrichment for **(a)** CpG islands and **(b)** TSS of any state in a model as a function of the number of states in the model. The figure shows that states with substantially higher enrichment for these annotations are only found consistently in models with a large number of states. There were isolated cases of models with a moderate number of states also exhibiting high enrichment. However, since similar enrichment levels were not captured in models with similar numbers of states this suggest the possibility that other biologically relevant states might be missing from these models.

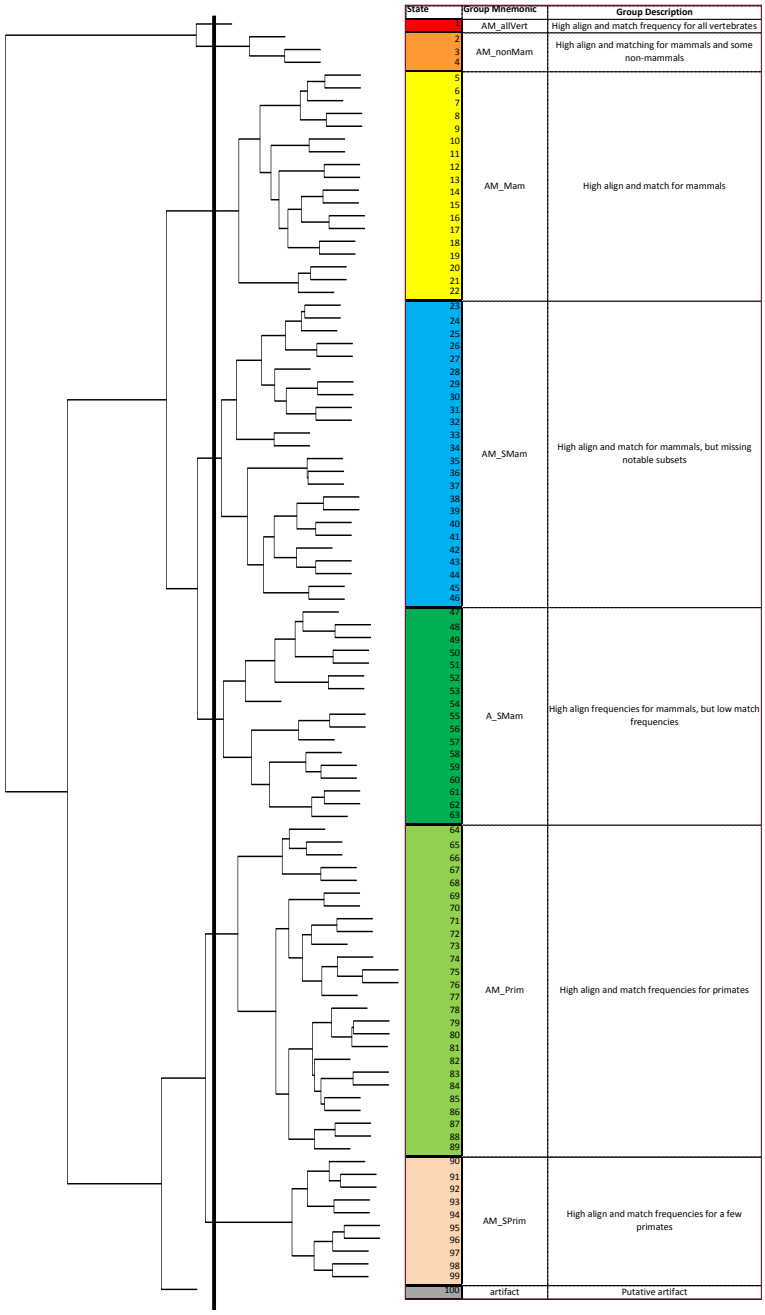
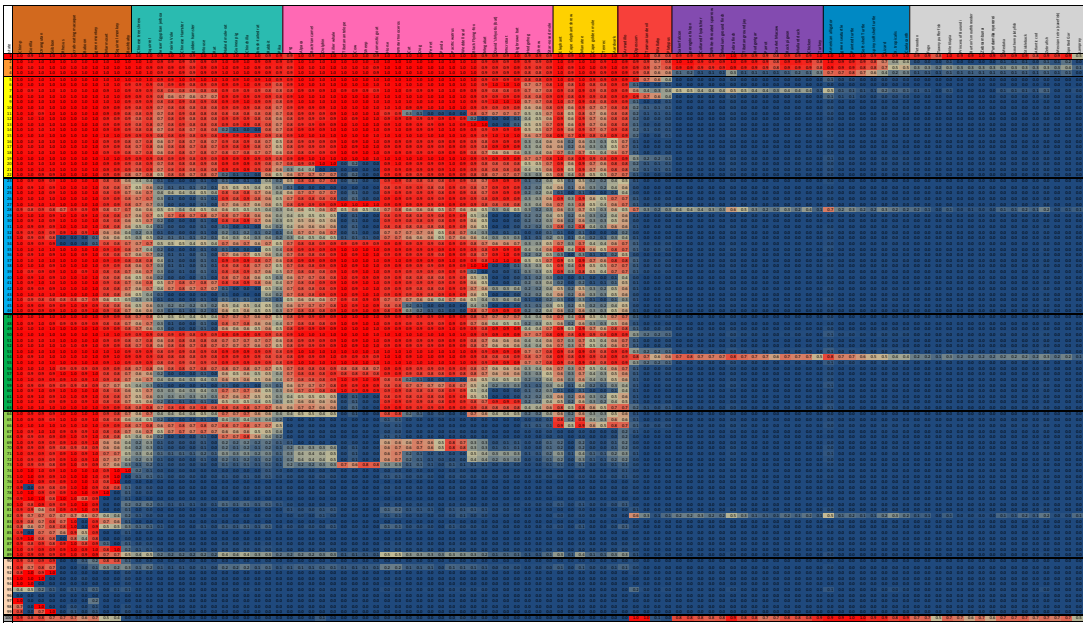


Figure 2.6 Hierarchical clustering and grouping of conservation states.

The dendrogram on left displays a hierarchical clustering of the states based on the values in **Figure 2.2a**, with the leaves ordered based on optimal leaf ordering[57]. The thick black line indicates where the dendrogram was cut to form the eight major groups of conservation states, each receiving a different color shown on right. To the right of the state numbers are state group abbreviations from **Figure 2.2a**. To the right of the state labels is a high level description of the general patterns of the parameters of the state groups. Notable enrichments associated with specific states are summarized in Supplementary Data 1.

a

Align Probabilities



b

Match Probabilities

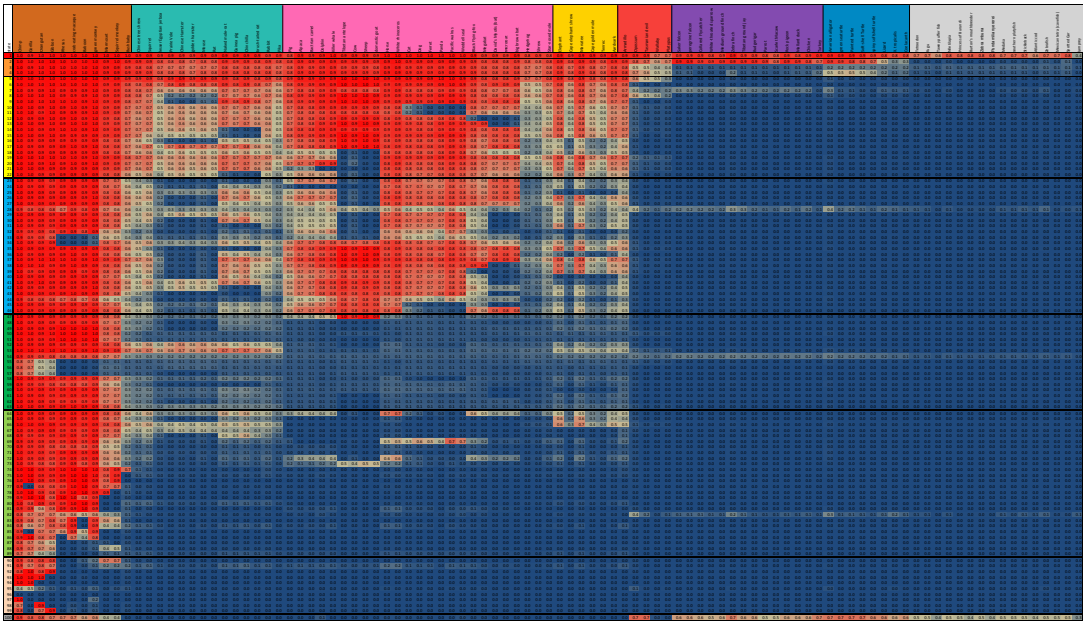


Figure 2.7 Representation of the emission parameters

This is a more detailed view of the representation of the emission parameters shown in the heatmap in **Figure. 2.2a**. In this figure the actual probabilities values and the individual species names are also displayed.

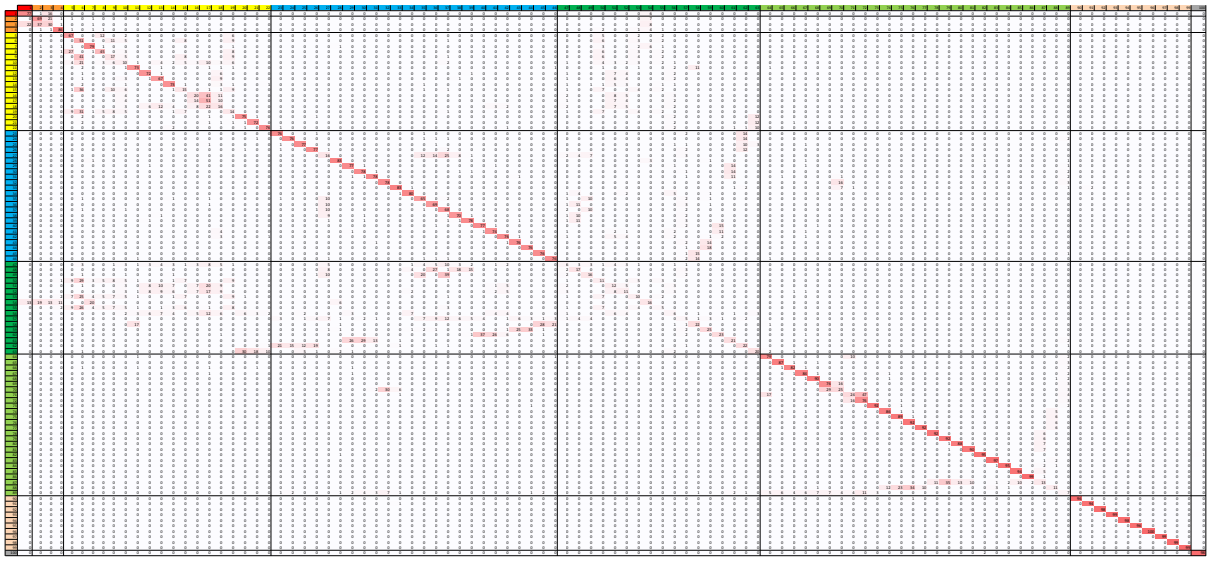


Figure 2.8 Conservation state model transition probabilities.

The figure displays the conservation state model transition probabilities. The values correspond to the probability when in the state of the row to transition to the state of the column at the next base. Probabilities are displayed multiplied by 100, rounded to the nearest integer and shaded based on their value, with darker red corresponding to greater transition probabilities. Transition probabilities along the diagonal show the probability of remaining in the state at a neighboring position, which were often the highest values for some states. For states associated with low matching probabilities relative to the alignment probabilities such as the A_SMam subgroup (states 47-63) the probability of remaining in the same state was low. Transition probabilities to stay in the same state were highest in some states only showing substantial alignability at most within primates, thus the model can use spatial information through these transition probabilities to better differentiate instances of states with relatively similar emission probabilities.

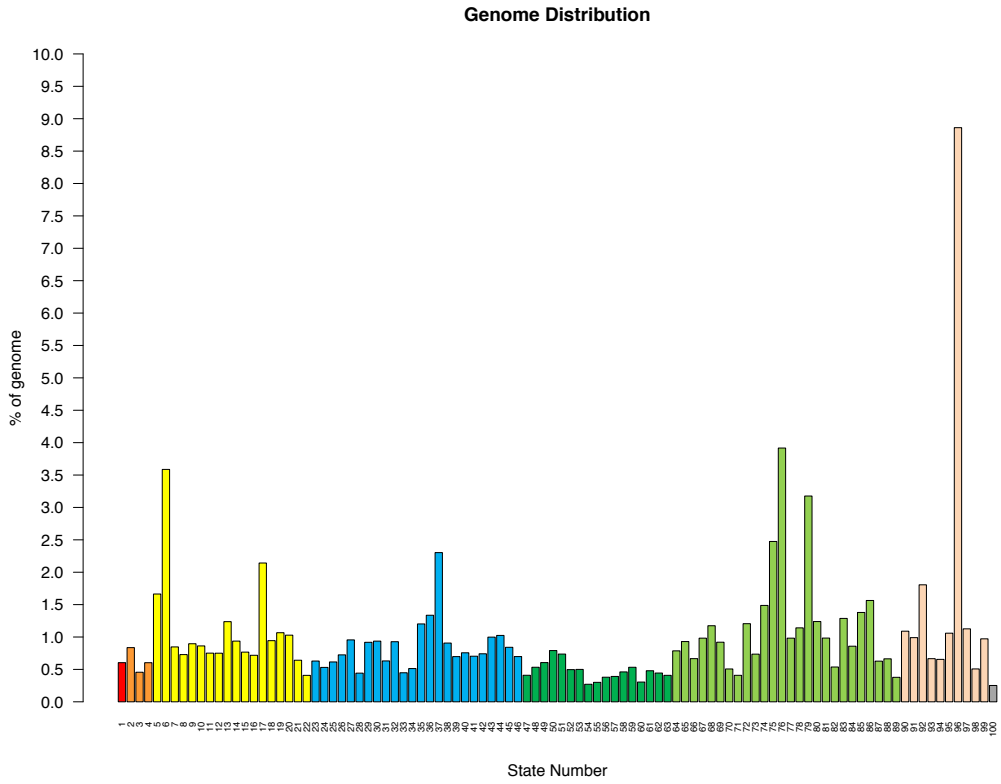
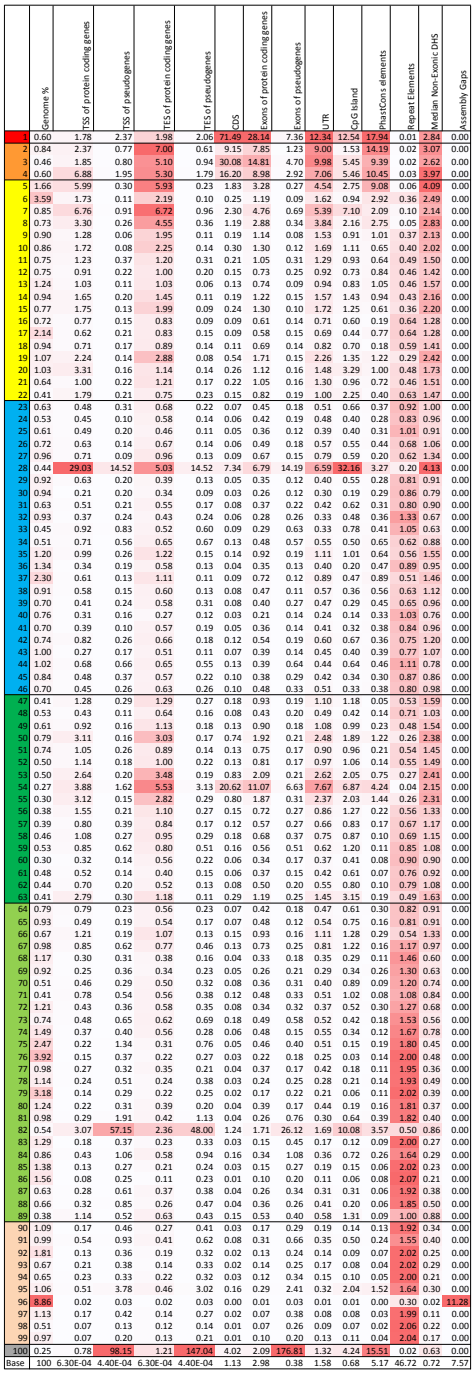


Figure 2.9 Distribution of the genome in each state.

The graph displays the percent of the genome assigned to each conservation state. The median state coverage was 0.76% of the genome. All states except state 96 were in the range of 0.25% to 3.92% of the genome. State 96 was the largest state covering 8.86% of the genome and was associated with assembly gaps (**Figure 2.10**).



	Genome %	TSS of protein coding genes	TSS of pseudogenes	TES of protein coding genes	TES of pseudogenes	CDS	Exons of protein coding genes	Exons of pseudogenes	UTR	CpG Island	PhastCons elements	Repeat Elements	Median Non-Exonic DHS	Assembly Gaps
1	0.60	1.78	2.37	1.98	2.06	71.49	28.14	7.36	12.34	12.54	17.94	0.01	2.84	0.00
2	0.84	2.37	0.77	7.00	0.61	9.15	7.85	1.23	9.00	1.53	14.19	0.02	3.07	0.00
3	0.46	1.85	0.80	5.10	0.94	30.08	14.81	4.70	9.98	5.45	9.39	0.02	2.62	0.00
4	0.60	6.88	1.95	5.30	1.79	16.20	8.98	2.92	7.06	5.46	10.45	0.03	3.97	0.00
5	1.66	5.99	0.30	5.93	0.23	1.83	3.28	0.27	4.54	2.75	9.08	0.06	4.09	0.00
6	3.59	1.73	0.11	2.19	0.10	0.25	1.19	0.09	1.62	0.94	2.92	0.36	2.49	0.00
7	0.85	6.76	0.91	6.72	0.96	2.30	4.76	0.69	5.39	7.10	2.09	0.10	2.14	0.00
8	0.73	3.30	0.26	4.55	0.36	1.19	2.88	0.34	3.84	2.16	2.75	0.05	2.83	0.00
9	0.90	1.28	0.06	1.95	0.11	0.19	1.14	0.08	1.53	0.91	1.01	0.37	2.13	0.00
10	0.86	1.72	0.08	2.25	0.14	0.30	1.30	0.12	1.69	1.11	0.65	0.40	2.02	0.00
11	0.75	1.23	0.37	1.20	0.31	0.21	1.05	0.31	1.29	0.93	0.64	0.49	1.50	0.00
12	0.75	0.91	0.22	1.00	0.20	0.15	0.73	0.25	0.92	0.73	0.84	0.46	1.42	0.00
13	1.24	1.03	0.11	1.03	0.06	0.13	0.74	0.09	0.94	0.83	1.05	0.46	1.57	0.00
14	0.94	1.65	0.20	1.45	0.11	0.19	1.22	0.15	1.57	1.43	0.94	0.43	2.16	0.00
15	0.77	1.75	0.13	1.99	0.09	0.24	1.30	0.10	1.72	1.25	0.61	0.36	2.20	0.00
16	0.72	0.77	0.15	0.83	0.09	0.09	0.61	0.14	0.71	0.60	0.19	0.64	1.28	0.00
17	2.14	0.62	0.21	0.83	0.15	0.09	0.58	0.15	0.69	0.44	0.77	0.64	1.28	0.00
18	0.94	0.71	0.17	0.89	0.14	0.11	0.69	0.14	0.82	0.70	0.18	0.59	1.41	0.00
19	1.07	2.24	0.14	2.88	0.08	0.54	1.71	0.15	2.26	1.35	1.22	0.29	2.42	0.00
20	1.03	3.31	0.16	1.14	0.14	0.26	1.12	0.16	1.48	3.29	1.00	0.48	1.73	0.00
21	0.64	1.00	0.22	1.21	0.17	0.22	1.05	0.16	1.30	0.96	0.72	0.46	1.51	0.00
22	0.41	1.79	0.21	0.75	0.23	0.15	0.82	0.19	1.00	2.25	0.40	0.63	1.47	0.00
23	0.63	0.48	0.31	0.68	0.22	0.07	0.45	0.18	0.51	0.66	0.37	0.92	1.00	0.00
24	0.53	0.45	0.10	0.58	0.14	0.06	0.42	0.19	0.48	0.40	0.28	0.83	0.96	0.00
25	0.61	0.49	0.20	0.46	0.11	0.05	0.36	0.12	0.39	0.40	0.31	1.01	0.91	0.00
26	0.72	0.63	0.14	0.67	0.14	0.06	0.49	0.18	0.57	0.55	0.44	0.68	1.06	0.00
27	0.96	0.71	0.09	0.96	0.13	0.09	0.67	0.15	0.79	0.59	0.20	0.62	1.34	0.00
28	0.44	29.03	14.52	5.03	14.52	7.34	6.79	14.19	6.59	32.16	3.27	0.20	4.13	0.00
29	0.92	0.63	0.20	0.39	0.13	0.05	0.35	0.12	0.40	0.55	0.28	0.81	0.91	0.00
30	0.94	0.21	0.20	0.34	0.09	0.03	0.26	0.12	0.30	0.19	0.29	0.86	0.79	0.00
31	0.63	0.51	0.21	0.55	0.17	0.08	0.37	0.22	0.42	0.62	0.31	0.80	0.90	0.00
32	0.93	0.37	0.24	0.43	0.24	0.06	0.28	0.26	0.33	0.48	0.36	1.33	0.67	0.00
33	0.45	0.92	0.83	0.52	0.60	0.09	0.29	0.63	0.33	0.78	0.41	1.05	0.63	0.00
34	0.51	0.71	0.56	0.65	0.67	0.13	0.48	0.57	0.55	0.50	0.65	0.62	0.88	0.00
35	1.20	0.99	0.26	1.22	0.15	0.14	0.92	0.19	1.11	1.01	0.64	0.56	1.55	0.00
36	1.34	0.34	0.19	0.58	0.13	0.04	0.35	0.13	0.40	0.20	0.47	0.89	0.95	0.00
37	2.30	0.61	0.13	1.11	0.11	0.09	0.72	0.12	0.89	0.47	0.89	0.51	1.46	0.00
38	0.91	0.58	0.15	0.60	0.13	0.08	0.47	0.11	0.57	0.36	0.56	0.63	1.12	0.00
39	0.70	0.41	0.24	0.58	0.31	0.08	0.40	0.27	0.47	0.29	0.45	0.65	0.96	0.00
40	0.76	0.31	0.16	0.27	0.12	0.03	0.21	0.14	0.24	0.14	0.33	1.03	0.76	0.00
41	0.70	0.39	0.10	0.57	0.19	0.05	0.36	0.14	0.41	0.32	0.38	0.84	0.96	0.00
42	0.74	0.82	0.26	0.66	0.18	0.12	0.54	0.19	0.60	0.67	0.36	0.75	1.20	0.00
43	1.00	0.27	0.17	0.51	0.11	0.07	0.39	0.14	0.45	0.40	0.39	0.77	1.07	0.00
44	1.02	0.68	0.66	0.65	0.55	0.13	0.39	0.64	0.44	0.64	0.46	1.11	0.78	0.00
45	0.84	0.48	0.37	0.57	0.22	0.10	0.38	0.29	0.42	0.34	0.30	0.87	0.86	0.00
46	0.70	0.45	0.26	0.63	0.26	0.10	0.48	0.33	0.51	0.33	0.38	0.80	0.98	0.00
47	0.41	1.28	0.29	1.29	0.27	0.18	0.93	0.19	1.10	1.18	0.05	0.53	1.59	0.00
48	0.53	0.43	0.11	0.64	0.16	0.08	0.43	0.20	0.49	0.42	0.14	0.71	1.03	0.00
49	0.61	0.92	0.16	1.13	0.18	0.13	0.90	0.18	1.08	0.99	0.23	0.48	1.54	0.00
50	0.79	3.11	0.16	3.03	0.17	0.74	1.92	0.21	2.48	1.89	1.22	0.26	2.38	0.00
51	0.74	1.05	0.26	0.89	0.14	0.13	0.75	0.17	0.90	0.96	0.21	0.54	1.45	0.00
52	0.50	1.14	0.18	1.00	0.22	0.13	0.81	0.17	0.97	1.06	0.14	0.55	1.49	0.00
53	0.50	2.64	0.20	3.48	0.19	0.83	2.09	0.21	2.62	2.05	0.75	0.27	2.41	0.00
54	0.27	3.88	1.62	5.53	3.13	20.62	11.07	6.63	7.67	6.87	4.24	0.04	2.15	0.00
55	0.30	3.12	0.15	2.82	0.29	0.80	1.87	0.31	2.37	2.03	1.44	0.26	2.31	0.00
56	0.38	1.55	0.21	1.10	0.27	0.15	0.72	0.27	0.86	1.27	0.22	0.56	1.33	0.00
57	0.39	0.80	0.39	0.84	0.17	0.12	0.57	0.27	0.66	0.83	0.17	0.67	1.17	0.00
58	0.46	1.08	0.27	0.95	0.29	0.18	0.68	0.37	0.75	0.87	0.10	0.69	1.15	0.00
59	0.53	0.85	0.62	0.80	0.51	0.16	0.56	0.51	0.62	1.20	0.11	0.85	1.08	0.00
60	0.30	0.32	0.14	0.56	0.22	0.06	0.34	0.17	0.37	0.41	0.08	0.90	0.90	0.00
61	0.48	0.52	0.14	0.40	0.15	0.06	0.37	0.15	0.42	0.61	0.07	0.76	0.92	0.00
62	0.44	0.70	0.20	0.52	0.13	0.08	0.50	0.20	0.55	0.80	0.10	0.79	1.08	0.00
63	0.41	2.79	0.30	1.18	0.11	0.29	1.19	0.25	1.45	3.15	0.19	0.49	1.63	0.00
64	0.79	0.79	0.23	0.56	0.23	0.07	0.42	0.18	0.47	0.61	0.30	0.82	0.91	0.00
65	0.93	0.49	0.19	0.54	0.17	0.07	0.48	0.12	0.54	0.75	0.16	0.81	0.91	0.00
66	0.67	1.21	0.19	1.07	0.13	0.15	0.93	0.16	1.11	1.28	0.29	0.54	1.33	0.00
67	0.98	0.85	0.62	0.77	0.46	0.13	0.73	0.25	0.81	1.22	0.16	1.17	0.97	0.00
68	1.17	0.30	0.31	0.38	0.16	0.04	0.33	0.18	0.35	0.29	0.11	1.46	0.60	0.00
69	0.92	0.25	0.36	0.34	0.23	0.05	0.26	0.21	0.29	0.34	0.26	1.30	0.63	0.00
70	0.51	0.46	0.29	0.50	0.32	0.08	0.36	0.31	0.40	0.89	0.09	1.20	0.74	0.00
71	0.41	0.78	0.54	0.56	0.38	0.12	0.48	0.33	0.51	1.02	0.08	1.08	0.84	0.00
72	1.21	0.43	0.36	0.58	0.35	0.08	0.34	0.32	0.37	0.52	0.30	1.27	0.68	0.00
73	0.74	0.48	0.65	0.62	0.69	0.18	0.49	0.58	0.52	0.42	0.18	1.53	0.56	0.00
74	1.49	0.37	0.40	0.56	0.28	0.06	0.48	0.15	0.55	0.34	0.12	1.67	0.78	0.00
75	2.47	0.22	1.34	0.31	0.76	0.05	0.46	0.40	0.51	0.15	0.19	1.80	0.45	0.00
76	3.92	0.15	0.37	0.22	0.27	0.03	0.22	0.18	0.25	0.03	0.14	2.00	0.48	0.00
77	0.98	0.27	0.32	0.35	0.21	0.04	0.37	0.17	0.42	0.18	0.11	1.95	0.36	0.00
78	1.14	0.24	0.51	0.24	0.38	0.03	0.24	0.25	0.28	0.21	0.14	1.93	0.49	0.00
79	3.18	0.14	0.29	0.22	0.25	0.02	0.17	0.22	0.21	0.06	0.11	2.02	0.39	0.00
80	1.24	0.22	0.31	0.39	0.20	0.04	0.39	0.17	0.44	0.19	0.16	1.81	0.37	0.00
81	0.98	0.29	1.91	0.42	1.13	0.04	0.26	0.76	0.30	0.64	0.39	1.82	0.40	0.00
82	0.54	3.07	57.15	2.36	48.00	1.24	1.71	26.12	1.69	10.08	3.57	0.50	0.86	0.00
83	1.29	0.18	0.37	0.23	0.33	0.03	0.15	0.45	0.17	0.12	0.09	2.00	0.27	0.00
84	0.86	0.43	1.06	0.58	0.94	0.16	0.34	1.08	0.36	0.72	0.26	1.64	0.29	0.00
85	1.38	0.13	0.27	0.21	0.24	0.03	0.15	0.27	0.19	0.15	0.06	2.02	0.23	0.00
86	1.56	0.08	0.25	0.11	0.23	0.01	0.10	0.20	0.11	0.06	0.08	2.07	0.21	0.00
87	0.63	0.28	0.61	0.37	0.38	0.04	0.26	0.34	0.31	0.31	0.06	1.92	0.38	0.00
88	0.66	0.32	0.85	0.26	0.47	0.04	0.36	0.26	0.41	0.20	0.06	1.85	0.50	0.00
89	0.38	1.14	0.52	0.63	0.43	0.15	0.53	0.40	0.58	1.31	0.09	1.00	0.88	0.00
90	1.09	0.17	0.46	0.27	0.41	0.03	0.17	0.29	0.19	0.14	0.13	1.92	0.34	0.00
91	0.99	0.54	0.93	0.41	0.62	0.08	0.31	0.66	0.35	0.50	0.24	1.55	0.40	0.00
92	1.81	0.13	0.36	0.19	0.32	0.02	0.13	0.24	0.14	0.09	0.07	2.02	0.25	0.00
93	0.67	0.21	0.38	0.14	0.33	0.02	0.14	0.25	0.17	0.08	0.04	2.02	0.29	0.00
94	0.65	0.23	0.33	0.22	0.32	0.03	0.12	0.34	0.15	0.10	0.05	2.00	0.21	0.00
95	1.06	0.51	3.78	0.46	3.02	0.16	0.29	2.41	0.32	2.04	1.52	1.64	0.30	0.00
96	8.86	0.02	0.03	0.02	0.03	0.00	0.01	0.03	0.01	0.01	0.00	0.30	0.02	11.26
97	1.13	0.17	0.42	0.14	0.27	0.02	0.07	0.38	0.08	0.08	0.03	1.99	0.11	0.00
98	0.51	0.07	0.13	0.12	0.14	0.01	0.07	0.26	0.09	0.07	0.02	2.06	0.22	0.00
99	0.97	0.07	0.20	0.13	0.21	0.01	0.10	0.20	0.13	0.11	0.04	2.04	0.17	0.00
100	0.25	0.78	98.15	1.21	147.04	4.02	2.09	176.81	1.32	4.24	15.51	0.02	0.63	0.00
Base	100	6.30E-04	4.40E-04	6.30E-04	4.40E-04	1.13	2.98	0.38	1.58	0.68	5.17	46.72	0.72	7.57

Figure 2.10 Conservation state enrichment values.

The first column displays the state number followed by the percentage of genome in each state, and then the values of the fold enrichments represented in the heatmap in **Figure 2.2b** along with additional columns for: enrichments for TES of protein coding genes, TES of pseudogenes, coding sequence (CDS), untranslated regions (UTRs), and assembly gaps. The last row gives the % of the genome covered by the annotation in each column. In the case of median non-exonic DHS, the last row is the median genome coverage across the DNase I experiments.

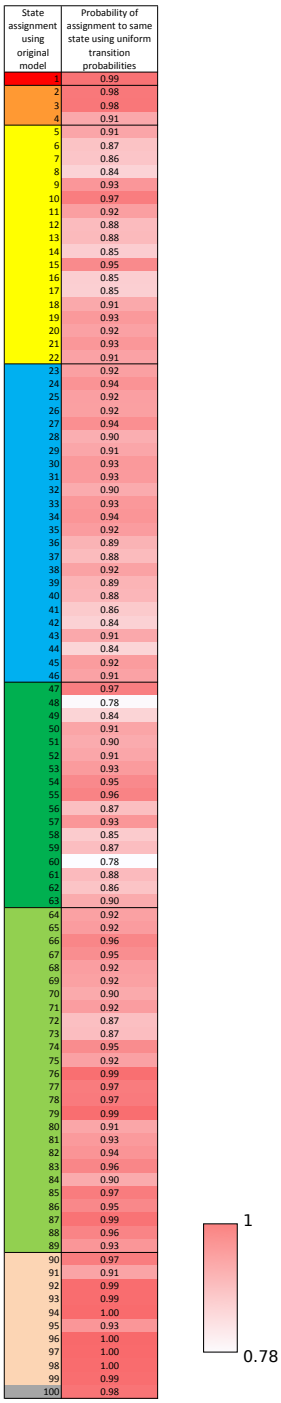


Figure 2.11 Comparison of state annotations from the learned model versus the learned model except with uniform probabilities.

The first column shows the state IDs. The second column displays for each state the fraction of bases annotated to the state using the learned model that were also assigned when using the learned model except with uniform transition probabilities. For all states the majority of bases were assigned to the same state when using the model with uniform transition probabilities with the fraction ranging between 0.78 and 1.00.

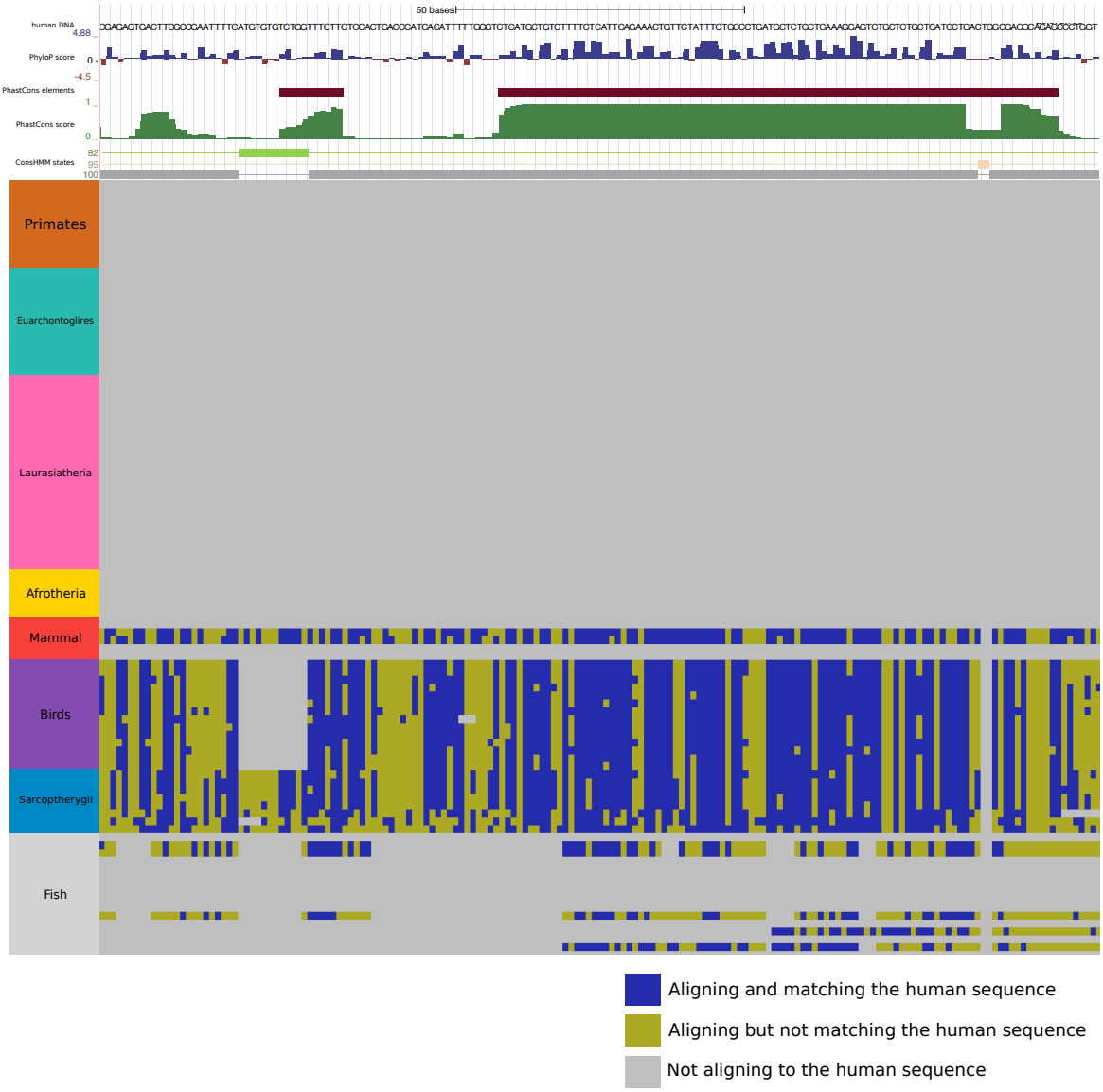


Figure 2.12 Illustration of conservation state assignments at an additional locus.

Similar to **Figure 2.2b**, but illustrating the conservation state assignment at a different locus: chr22:20,673,589-20,673,761. The top tracks are the DNA sequence, the PhyloP score, PhastCons elements, and then PhastCons score. Below this set of tracks are the conservation state assignments with only the states assigned to at least one nucleotide in the locus shown. Below the conservation state assignments is a color encoding of the input multiple sequence alignment. The major clade of species as annotated on the UCSC genome browser2 are labeled and ordered based on divergence from human. The figure is an example of positions with high constraint scores from PhyloP and PhastCons, while the multiple sequence alignment lacks alignment to most mammals, which is suggestive of alignment artifacts. ConsHMM states 82 and 100 capture the pattern of non-mammalian vertebrates aligning and/or matching the human genome, without most mammals. State 95 captures the pattern of all species having low alignment and matching probabilities and relatively proximal to states with higher probabilities of alignment and matching (**Figures 2.7-8**).

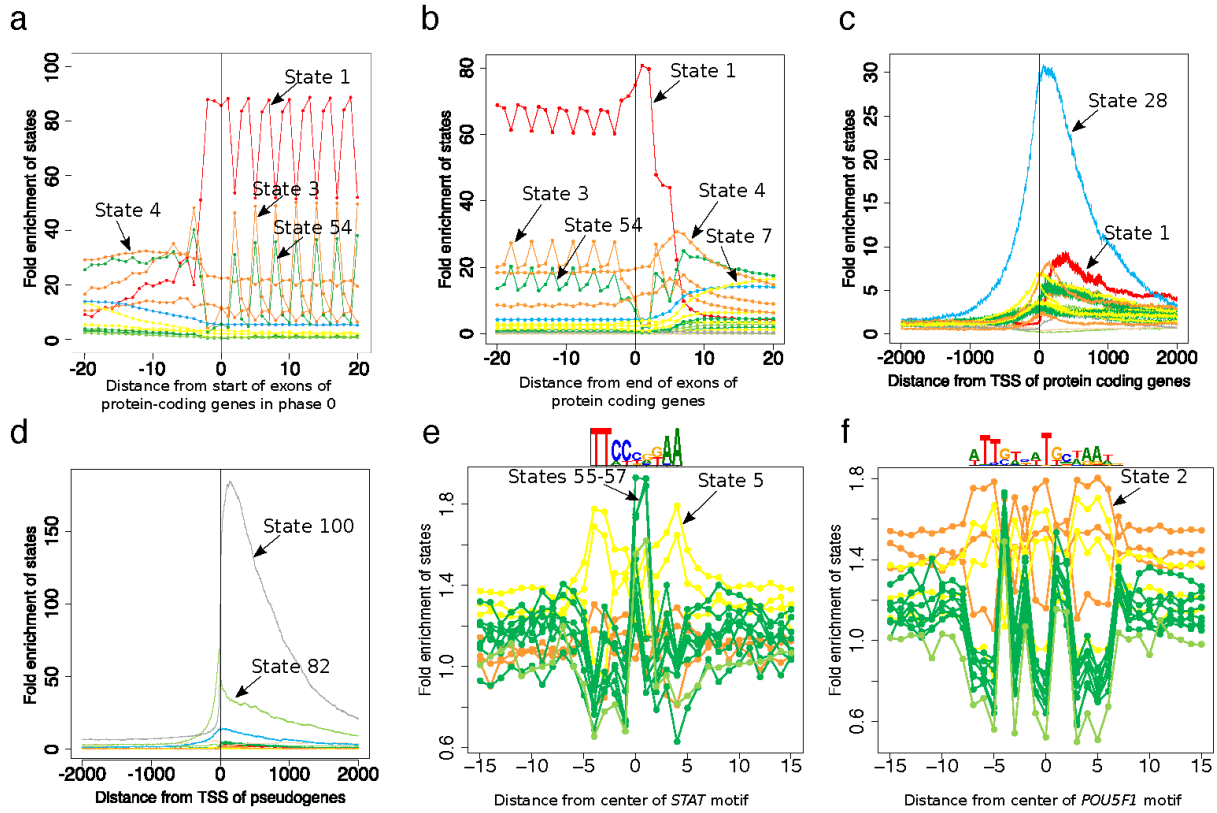


Figure 2.13 Conservation state emission parameters learned by ConsHMM and enrichments for other genomic annotations.

(a) Each row in the heatmap corresponds to a conservation state. For each state and species, the left half of the heatmap gives the probability of aligning to the human sequence, which is one minus the probability of the not aligning emission. Analogously, the right half of the heatmap gives the probability of the matching emission. Each individual column corresponds to one species with the individual names displayed in **Figure 2.7**. For both halves, species are grouped by the major clades and ordered based on the hg19.100way.nh phylogenetic tree from the UCSC genome browser, with species that diverged more recently shown closer to the left[21].The conservation states are ordered based on the results of applying hierarchical clustering and optimal leaf ordering[54]. The states are divided into eight major groups based on cutting the dendrogram of the clustering. The full dendrogram and an explanation of the group mnemonics is available in **Figure 2.6**. The groups are indicated by color bars on the left hand side and a white row between them. Transition parameters between states of the model can be found in **Figure 2.8**. **(b)** The columns of the heatmap indicate the relative enrichments of conservation states for external genomic annotations (**Methods**). For each column, the enrichments were normalized to a [0,1] range by subtracting the minimum value of the column and dividing by the range and colored based on the indicated scale on the right. Values for these enrichments and additional enrichments can be found in **Figure 2.10** and enrichments for individual repeat classes and families can be found in **Figure 2.18**.

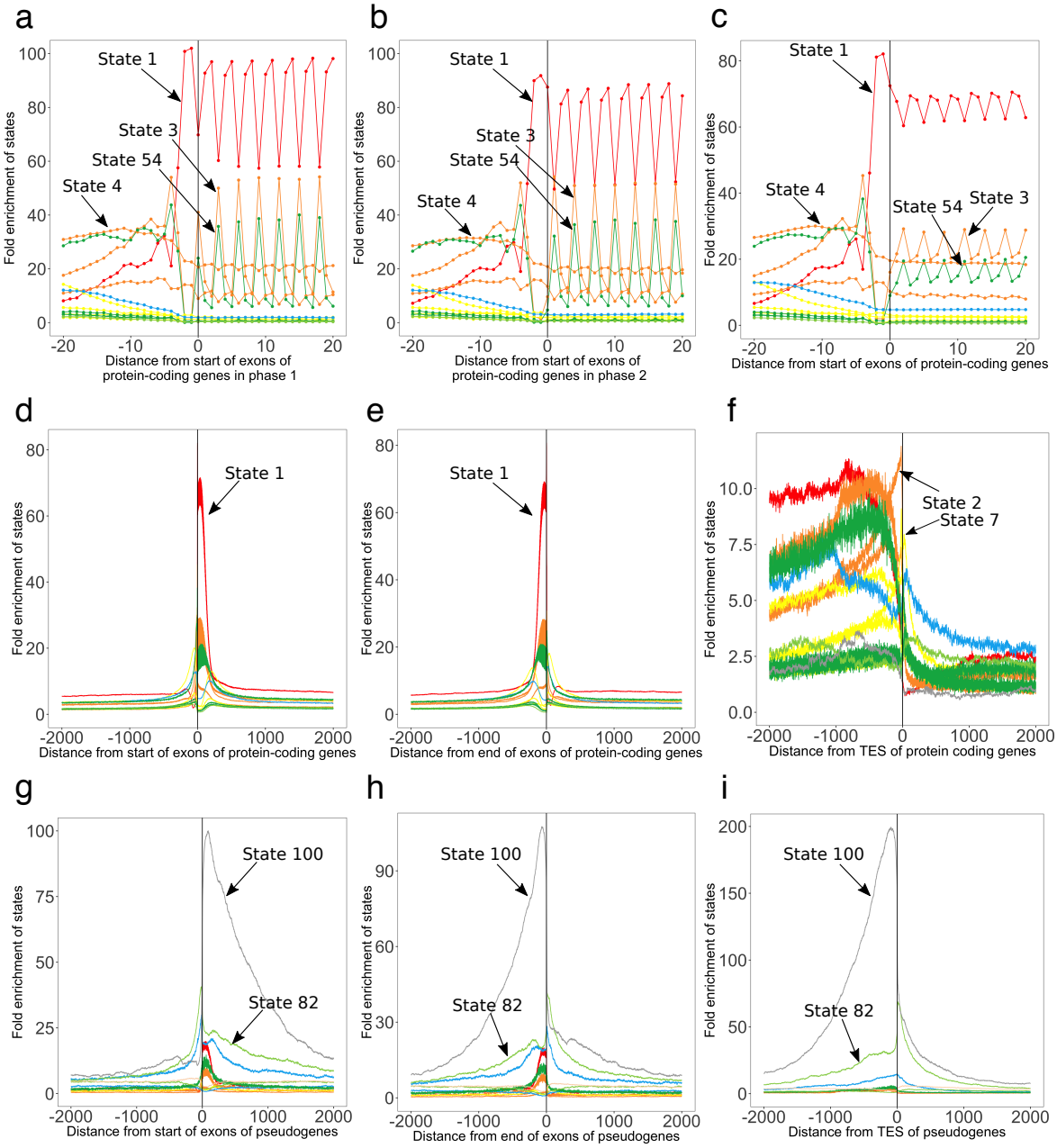


Figure 2.14 Additional conservation state positional enrichments.

The figure shows additional positional enrichment plots similar to what was shown in Fig. 3. These additional enrichment plots include enrichments relative to start of exons of protein coding genes for (a) phase 1 and (b) phase 2 exons, (c) all exons, and a zoomed out view of enrichments relative to (d) the start and (e) end of all exons of protein coding genes. Also shown are enrichment plots relative to (f) TES of protein coding genes, (g) start and (h) end of exons of pseudogenes as well as (i) TES of pseudogenes. Enrichments were computed relative to a genome-wide background. The subset of states included in the figure was composed of the states that had at least a 3 fold enrichment at some position within +/-2kb from the anchor point.

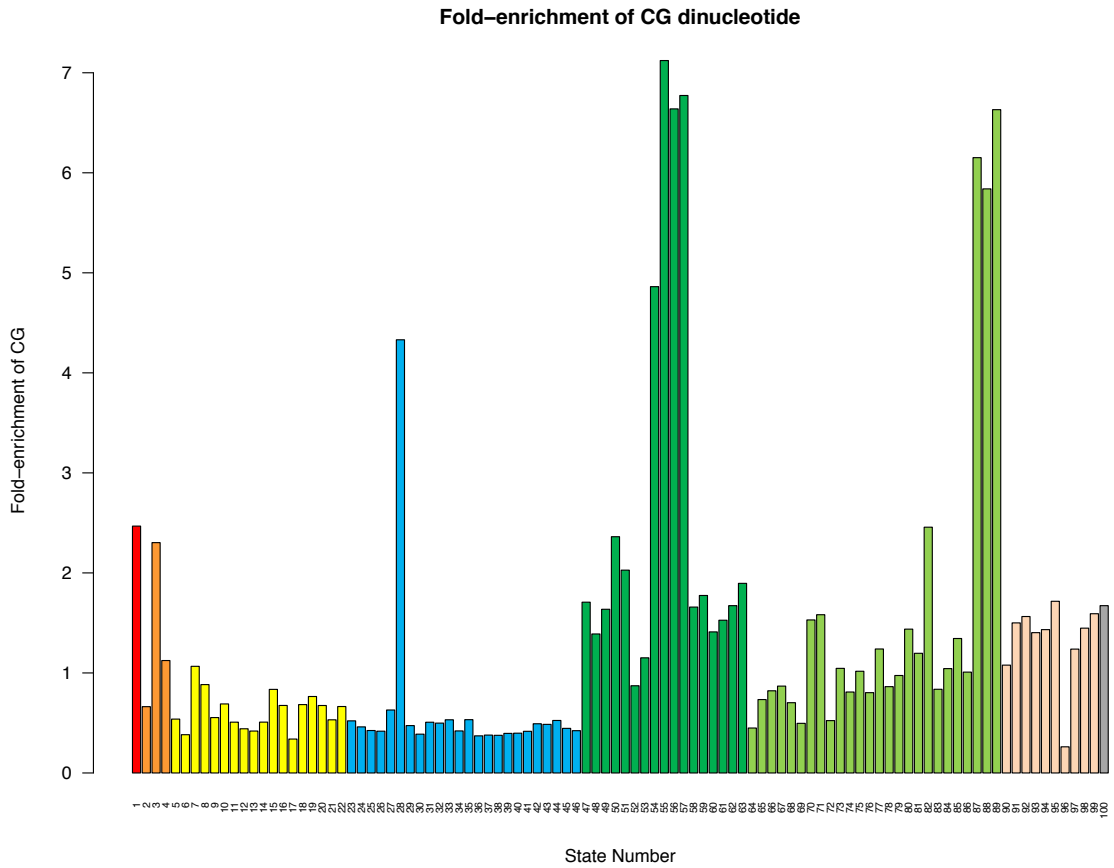


Figure 2.15 Enrichment of CG dinucleotides in the states.

The bar graph shows for each state its fold enrichment for CG dinucleotides. States 55-57 and 87-89 had the highest enrichments followed by states 28 and 54.

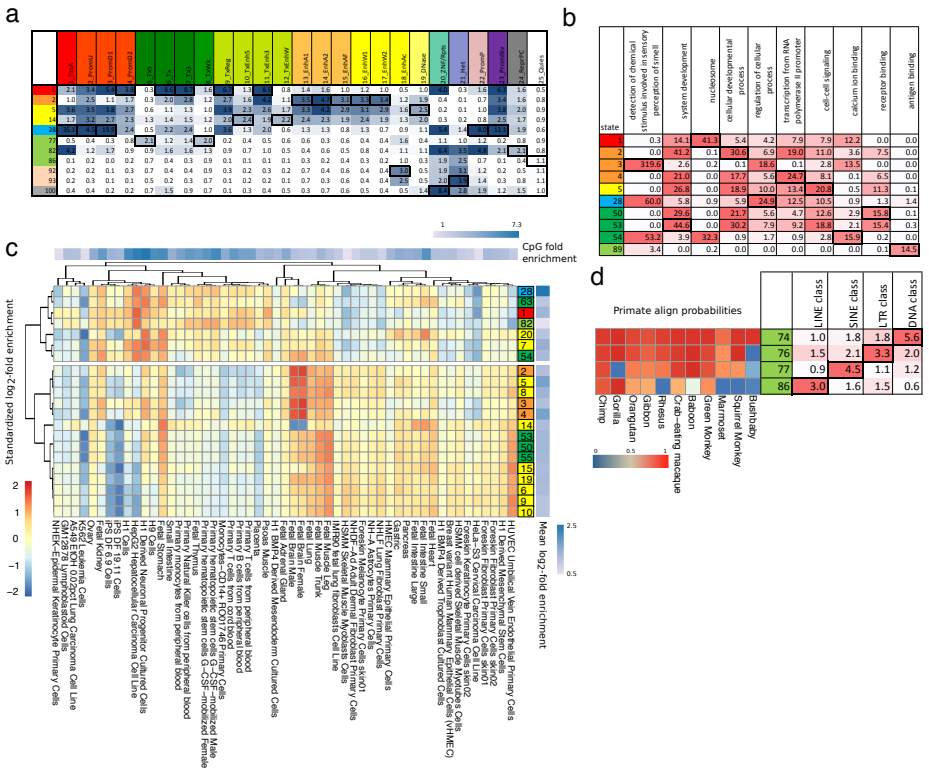


Figure 2.16 Conservation states enrichment for chromatin states, GO terms, DHS and repeat elements.

(a) Median fold enrichment of conservation states (rows) for one of 25 chromatin states from a previously defined chromatin state model defined across 127 samples of diverse cell and tissue types (columns)[40]. Only conservation states that had the maximum value for at least one chromatin state are shown, and those values are boxed. See **Figure 2.19** for the enrichments of all conservation states. (b) –log10 p-value (uncorrected) of the conservation states (rows) for the GO term (columns) where each conservation state is associated with its top 5% genes based on promoter regions (**Methods**). Only GO terms which were the most significantly enriched term for some conservation state among terms the state was maximally significant for are shown, restricted to the top 10 terms based on the significance of the enrichment. Only conservation states that had the most significant enrichment for one of the displayed GO terms are shown, with the maximal enrichments boxed. The full set of conservation states with additional GO terms are in **Figure 2.17**. **(c)** Relative enrichments of conservation states for DHS across cell and tissue types. Only conservation states with at least a 2 fold enrichment in one sample considered are shown. Enrichment values were log2 transformed and then row normalized by subtracting the mean (right heatmap) and dividing by the standard deviation. States and experiments were then hierarchically clustered and revealed two major clusters. In the top cluster conservation states showed the greatest enrichment for experiments in which the DHS also strongly enriched for CpG islands (top heatmap). In the bottom cluster conservation states had the strongest relative preference for fetal related samples or HUVEC. **(d)** Fold enrichment of conservation states with the maximal enrichment for LINE, SINE, LTR or DNA repeats next to the state align probabilities for primates. These states all had low align probabilities outside of primates, but their differences among primates corresponded to substantial differences in repeat enrichments.[28]

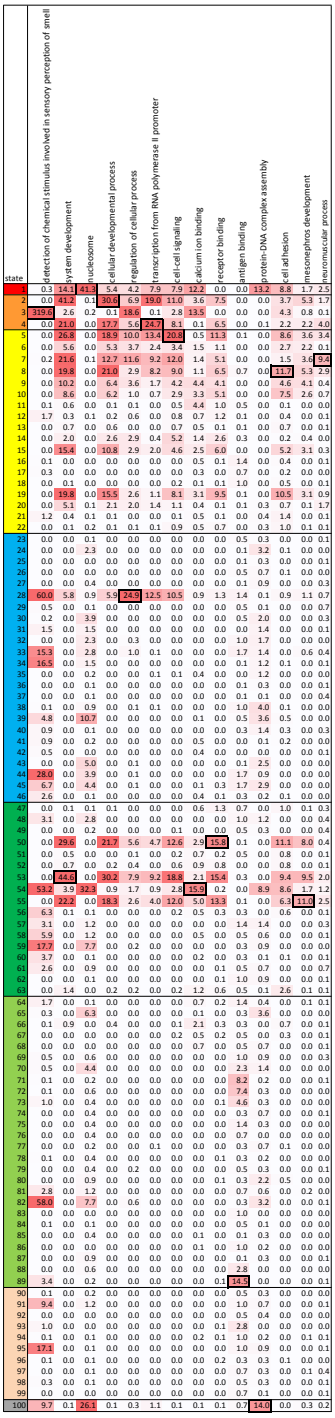


Figure 2.17 GO term enrichment p-values.

This figure is an extended version of **Figure 2.16b**, now showing –log10 p-values for GO term enrichments for all states. The criteria for including a GO term were the same as in **Figure 2.16b**, except the top 10 term criterion was relaxed to include four additional terms. Values that are boxed mark the top most significant enrichment for the term of the column.

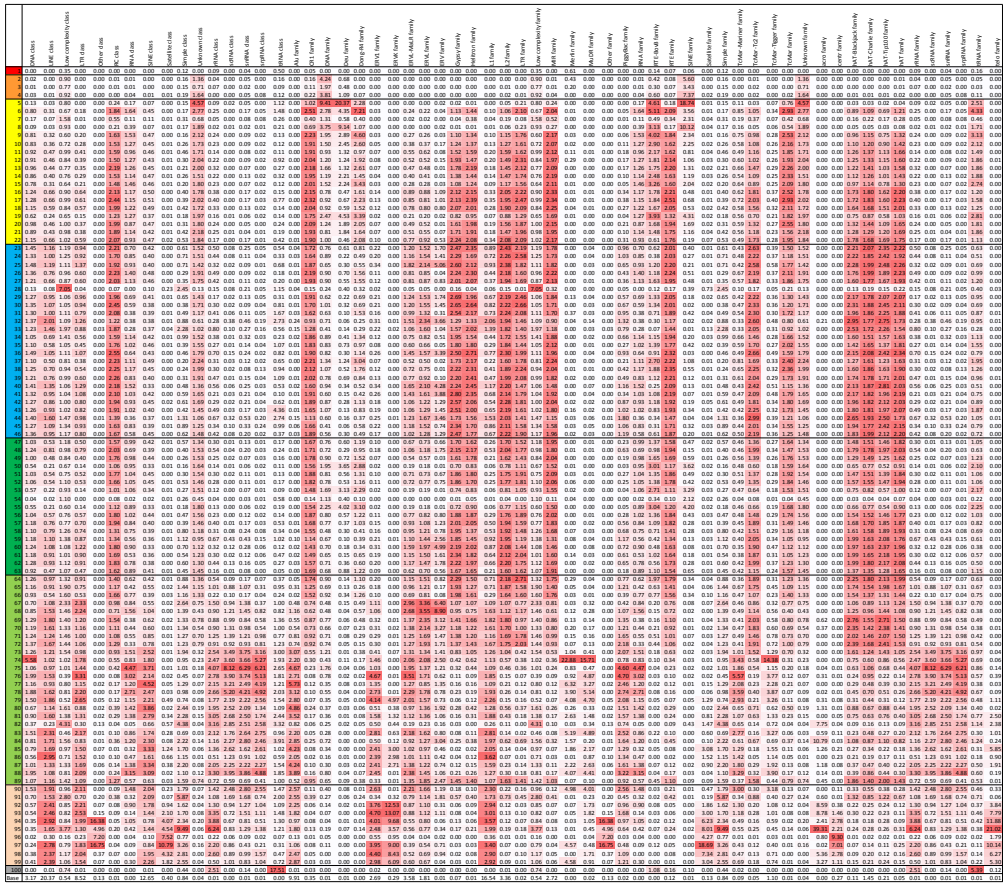


Figure 2.18 Conservation state enrichments for RepeatMasker classes and families of repeats.

The rows correspond to different conservation states and columns correspond to different repeat classes or families. The first 16 columns are repeat classes, and the remaining are repeat families. The values correspond to fold enrichment for the repeat class or family for the conservation state. Values are shaded in a column specific manner. The last row gives the % of the genome the repeat class or family covers.

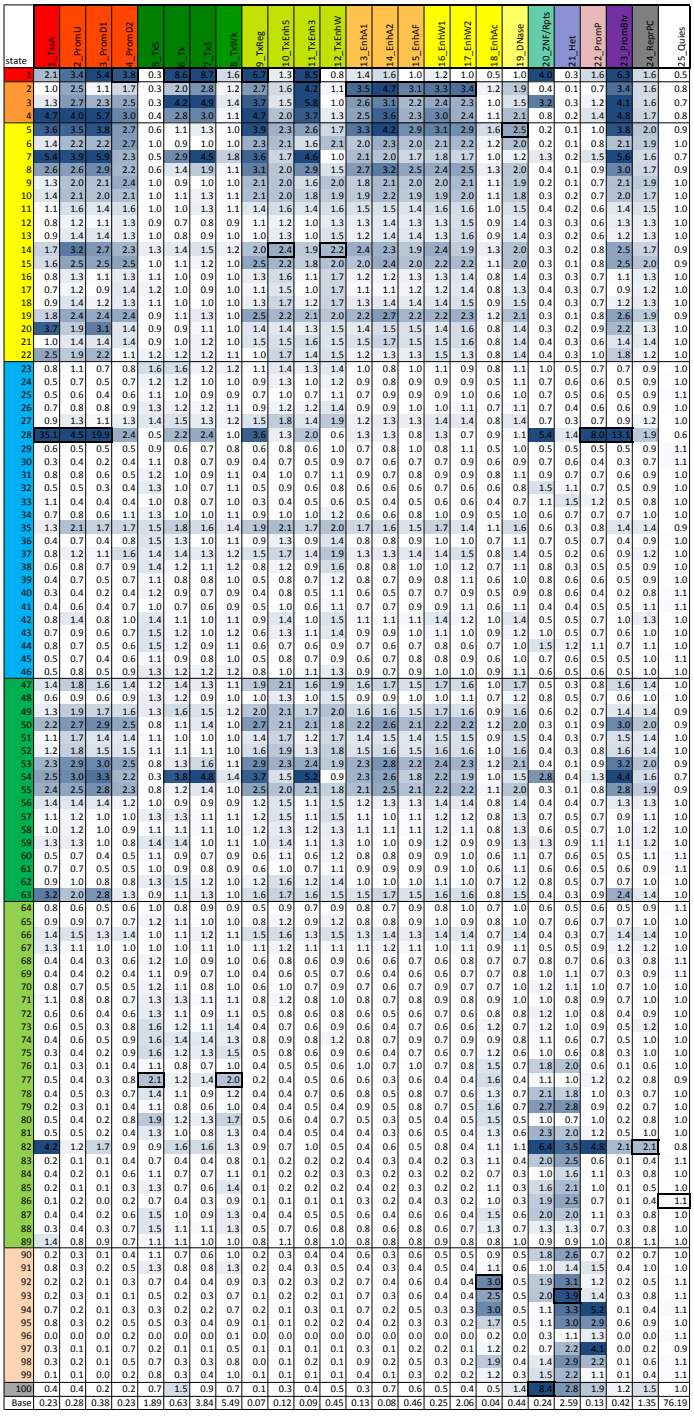


Figure 2.19 Conservation state enrichments for chromatin states.

This is an extended version of Fig. 4a. Each row corresponds to one of the conservation states. Each column corresponds to one of 25 chromatin states from the Roadmap Epigenomics imputation based chromatin state model defined across 127 samples of diverse cell and tissue types[9]. The values reported are the median fold enrichment across the 127 samples of the chromatin state for the conservation state. Values that are boxed were the maximum value of the column.

a

	GERP++ elements	PhastCons elements	SiPhy-Omega elements	SiPhy-Pi elements
1	14.32	17.94	21.98	16.49
2	12.62	14.19	16.40	13.18
3	11.98	9.39	14.71	12.60
4	10.99	10.45	13.54	11.55
5	9.84	9.08	11.27	9.77
6	3.37	2.92	3.22	3.48
7	3.77	2.09	3.10	3.56
8	5.77	2.75	4.02	4.48
9	2.09	1.01	1.44	1.99
10	1.65	0.65	0.99	1.62
11	0.85	0.64	0.54	0.84
12	1.09	0.84	0.70	1.19
13	1.35	1.05	0.99	1.29
14	1.17	0.94	0.88	1.21
15	1.61	0.61	0.96	1.64
16	0.34	0.19	0.16	0.35
17	0.65	0.77	0.49	0.64
18	0.41	0.18	0.18	0.47
19	2.86	1.22	1.94	2.70
20	1.34	1.00	1.00	1.29
21	1.02	0.72	0.65	0.88
22	0.33	0.40	0.24	0.35
23	0.09	0.37	0.09	0.07
24	0.09	0.28	0.08	0.07
25	0.11	0.31	0.11	0.11
26	0.31	0.44	0.22	0.26
27	0.34	0.20	0.15	0.30
28	3.00	3.27	2.27	1.91
29	0.16	0.28	0.12	0.20
30	0.12	0.29	0.09	0.11
31	0.09	0.31	0.07	0.10
32	0.03	0.36	0.03	0.05
33	0.07	0.41	0.07	0.06
34	0.47	0.65	0.37	0.44
35	0.51	0.64	0.35	0.39
36	0.17	0.47	0.16	0.11
37	0.91	0.89	0.65	0.77
38	0.45	0.56	0.30	0.32
39	0.33	0.45	0.18	0.34
40	0.09	0.33	0.08	0.10
41	0.24	0.38	0.17	0.27
42	0.17	0.36	0.15	0.23
43	0.12	0.39	0.10	0.14
44	0.05	0.46	0.05	0.05
45	0.08	0.30	0.08	0.15
46	0.13	0.38	0.13	0.12
47	0.31	0.05	0.08	0.51
48	0.16	0.14	0.06	0.15
49	0.48	0.23	0.18	0.39
50	2.85	1.22	1.86	2.58
51	0.45	0.21	0.19	0.47
52	0.35	0.14	0.13	0.42
53	2.38	0.75	1.36	2.25
54	7.67	4.24	7.31	7.77
55	3.09	1.44	2.17	2.87
56	0.45	0.22	0.20	0.46
57	0.23	0.17	0.09	0.20
58	0.14	0.10	0.05	0.19
59	0.05	0.11	0.02	0.07
60	0.06	0.08	0.02	0.12
61	0.07	0.07	0.02	0.11
62	0.08	0.10	0.03	0.08
63	0.56	0.19	0.23	0.54
64	0.16	0.30	0.12	0.14
65	0.06	0.16	0.03	0.09
66	0.36	0.29	0.21	0.33
67	0.16	0.16	0.08	0.17
68	0.02	0.11	0.02	0.06
69	0.02	0.26	0.02	0.05
70	0.01	0.09	0.00	0.04
71	0.03	0.08	0.01	0.05
72	0.03	0.30	0.02	0.05
73	0.01	0.18	0.01	0.07
74	0.07	0.12	0.03	0.08
75	0.02	0.19	0.01	0.05
76	0.01	0.14	0.01	0.02
77	0.01	0.11	0.00	0.03
78	0.00	0.14	0.00	0.03
79	0.01	0.11	0.00	0.02
80	0.01	0.16	0.01	0.06
81	0.01	0.39	0.01	0.03
82	0.33	3.57	0.16	0.08
83	0.00	0.09	0.00	0.01
84	0.00	0.26	0.00	0.03
85	0.00	0.06	0.00	0.01
86	0.00	0.08	0.00	0.01
87	0.01	0.06	0.00	0.03
88	0.02	0.06	0.01	0.04
89	0.07	0.09	0.02	0.09
90	0.01	0.13	0.01	0.02
91	0.02	0.24	0.01	0.07
92	0.01	0.07	0.01	0.02
93	0.01	0.04	0.00	0.01
94	0.01	0.05	0.01	0.01
95	0.04	1.52	0.02	0.03
96	0.00	0.00	0.00	0.00
97	0.00	0.00	0.00	0.00
98	0.00	0.02	0.00	0.00
99	0.00	0.04	0.00	0.01
100	1.96	15.51	0.99	0.31
Base	6.775	5.174	4.042	5.565

b

	mean PhyloP score	mean PhastCons core	mean GERP++ score
1	5.173	0.926	4.191
2	2.308	0.738	3.030
3	0.617	0.484	1.211
4	1.357	0.549	2.194
5	1.173	0.488	2.371
6	0.636	0.180	1.130
7	0.292	0.126	0.339
8	0.132	0.155	0.074
9	-0.029	0.065	-0.186
10	-0.243	0.042	-1.062
11	0.112	0.051	-0.076
12	0.197	0.062	0.155
13	0.272	0.076	0.350
14	0.221	0.070	0.235
15	-0.249	0.040	-0.921
16	-0.287	0.018	-0.829
17	0.401	0.066	0.541
18	-0.430	0.016	-1.440
19	-0.215	0.073	-1.025
20	0.166	0.070	0.052
21	0.111	0.055	-0.107
22	0.057	0.038	-0.175
23	0.130	0.045	-0.043
24	0.095	0.037	-0.116
25	0.085	0.035	-0.128
26	0.118	0.043	-0.054
27	-0.409	0.019	-1.299
28	0.148	0.180	-0.240
29	0.071	0.032	-0.147
30	0.111	0.036	-0.060
31	0.101	0.040	-0.075
32	0.175	0.053	0.056
33	0.104	0.051	-0.121
34	0.142	0.056	-0.073
35	0.233	0.059	0.135
36	0.228	0.051	0.159
37	0.276	0.071	0.317
38	0.214	0.053	0.181
39	0.122	0.044	-0.028
40	0.132	0.041	-0.011
41	0.129	0.039	-0.013
42	0.113	0.038	-0.021
43	0.145	0.046	-0.012
44	0.185	0.058	0.054
45	0.122	0.039	-0.049
46	0.135	0.043	-0.017
47	-1.126	0.004	-3.166
48	-0.509	0.015	-1.702
49	-0.475	0.020	-1.812
50	-0.310	0.071	-1.652
51	-0.469	0.018	-1.851
52	-0.530	0.012	-1.917
53	-0.484	0.045	-2.083
54	-0.338	0.219	-2.198
55	-0.376	0.082	-1.613
56	-0.593	0.018	-1.908
57	-0.619	0.017	-1.721
58	-0.641	0.011	-1.857
59	-0.653	0.015	-1.489
60	-0.674	0.010	-1.794
61	-0.631	0.010	-1.793
62	-0.600	0.012	-1.754
63	-0.618	0.015	-2.166
64	0.078	0.035	-0.144
65	-0.022	0.030	-0.347
66	-0.061	0.030	-0.548
67	-0.088	0.024	-0.531
68	-0.025	0.028	-0.273
69	0.171	0.048	0.033
70	-0.611	0.016	-1.079
71	-0.661	0.013	-1.275
72	0.183	0.050	0.046
73	-0.071	0.035	-0.184
74	0.040	0.042	-0.142
75	0.038	0.043	-0.043
76	0.124	0.106	0.102
77	0.068	0.068	0.035
78	0.093	0.089	0.057
79	0.117	0.137	0.098
80	0.030	0.044	-0.047
81	0.074	0.093	0.029
82	0.064	0.210	-0.180
83	0.064	0.108	0.057
84	0.019	0.056	-0.045
85	0.072	0.122	0.065
86	0.037	0.145	0.054
87	-1.178	0.034	-0.463
88	-1.067	0.024	-0.695
89	-0.883	0.012	-1.473
90	0.022	0.100	-0.003
91	-0.064	0.047	-0.168
92	0.004	0.146	-0.003
93	0.006	0.165	0.004
94	0.000	0.188	-0.007
95	0.030	0.182	-0.026
96	0.000	0.120	0.000
97	-0.006	0.196	-0.001
98	0.001	0.169	0.002
99	-0.003	0.140	-0.002
100	2.005	0.795	0.170

Figure 2.20 Conservation states enrichments for evolutionary constrained element calls and average constraint scores.

(a) Each row corresponds to a conservation state and each column corresponds to a different constrained element set. The values correspond to the fold enrichment for bases in a constrained element set for the conservation state. The constrained element sets are from left to right GERP++, PhastCons, SiPhy-omega, and SiPhy-pi. The bottom row gives the percentage of the genome of each constrained element set. **(b)** Each row corresponds to a conservation state and each column corresponds to a different score of constraint. The values correspond to the average constraint score in the conservation state. The constraint scores are from left to right: PhyloP8, PhastCons5, and GERP++.

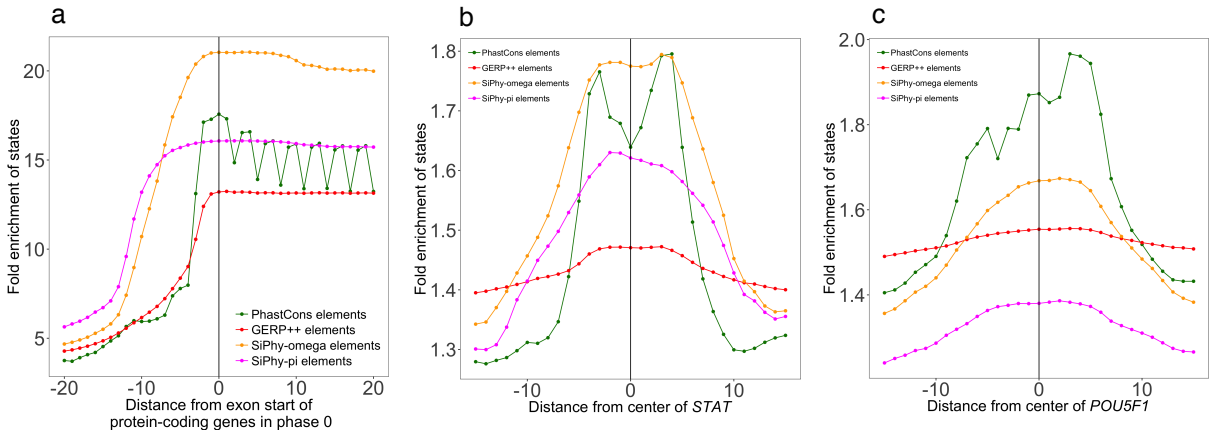


Figure 2.21 Positional enrichment of constrained element sets.

Analogous to what was shown for conservation states in **Figure 2.13a,e,f**, the graphs show the positional fold enrichments of GERP++, PhastCons, SiPhy-omega and SiPhy-pi constrained elements calls around **(a)** the start of exons of protein coding genes, **(b)** center of instances of a STAT motif, and **(c)** center of instances of a POU5F1 motif. Of these only PhastCons element calls are able to exhibit relevant single nucleotide enrichment variation, as was seen with the conservation states.

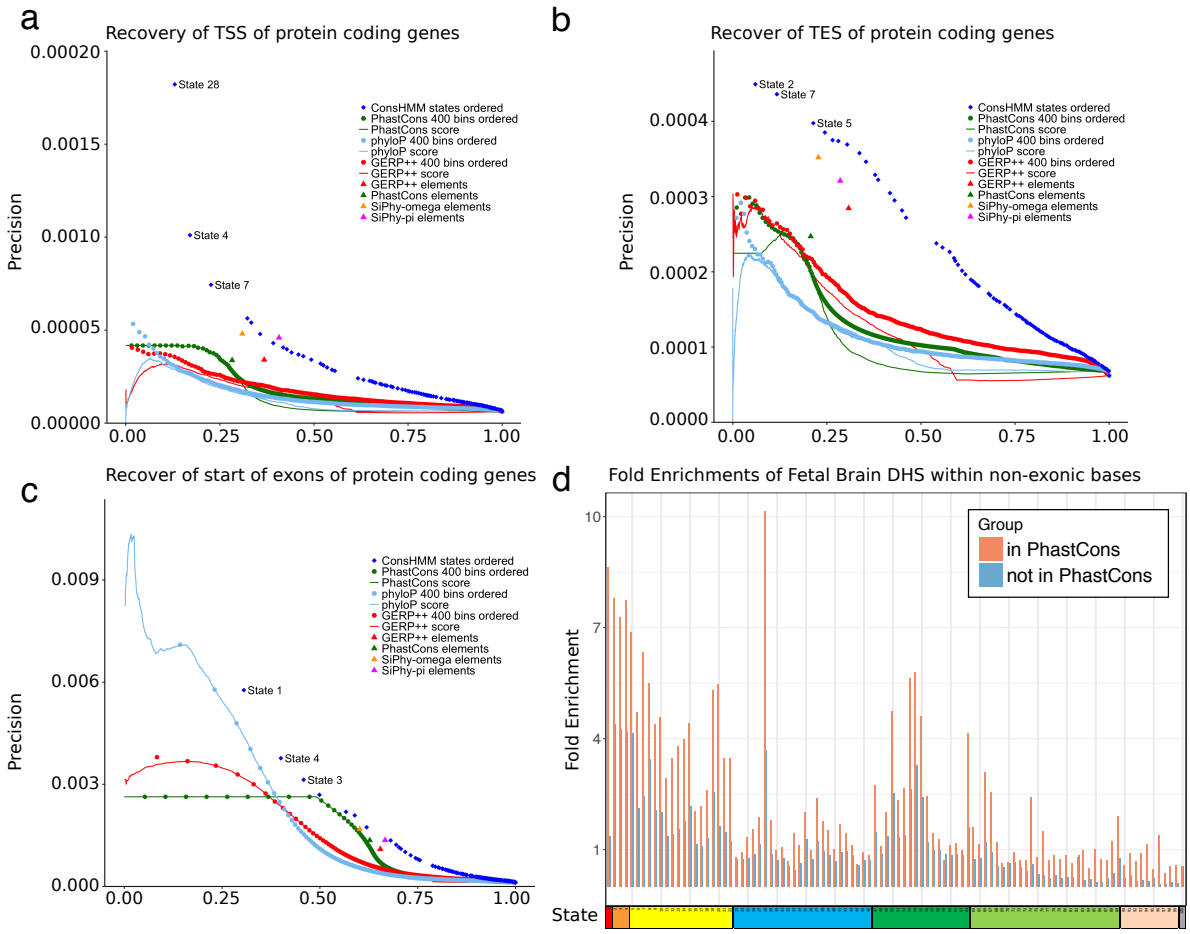


Figure 2.22 Relationship of conservation states with constrained elements and scores.

Precision-recall plots for recovery of **(a)** TSS of protein coding genes, **(b)** TES of protein coding genes, and **(c)** the start of exons of protein coding genes. Recovery based on ordering ConsHMM conservation states for their enrichment for the target set in the training data, then cumulatively adding the states in that ranked order and evaluating on the test data is shown with a series of blue dots (**Methods**). The first few conservation states added are labeled with their state number. Recovery based on ranking from highest to lowest value of constraint scores is shown with continuous lines. Recovery based on score partitioning into 400 bins and subsequent ordering based on enrichment for the target set in the training data, then cumulatively adding bins in that ranked order and evaluating on the test data is shown in a series of dots of the same color as the continuous line corresponding to the score. Recovery of target test bases by a constrained element set is shown with a single dot for each constrained element set. See Figures **2.23-25** for plots based on additional targets. **(d)** The graph shows the fold enrichment for Fetal Brain DHS[5] within the non-exonic portion of each conservation state, separately for those bases in a PhastCons constrained element (pink) and bases not in such an element (blue). Enrichments within constrained elements varied substantially depending on the conservation state. For a given conservation state, bases in a constrained element had greater enrichments than bases not in a constrained element, illustrating complementary information of conservation states and constrained elements. See **Figure 2.26** for graphs based on different element sets or DHS data and **Figure 2.27** for these enrichments plotted against the size of the set.

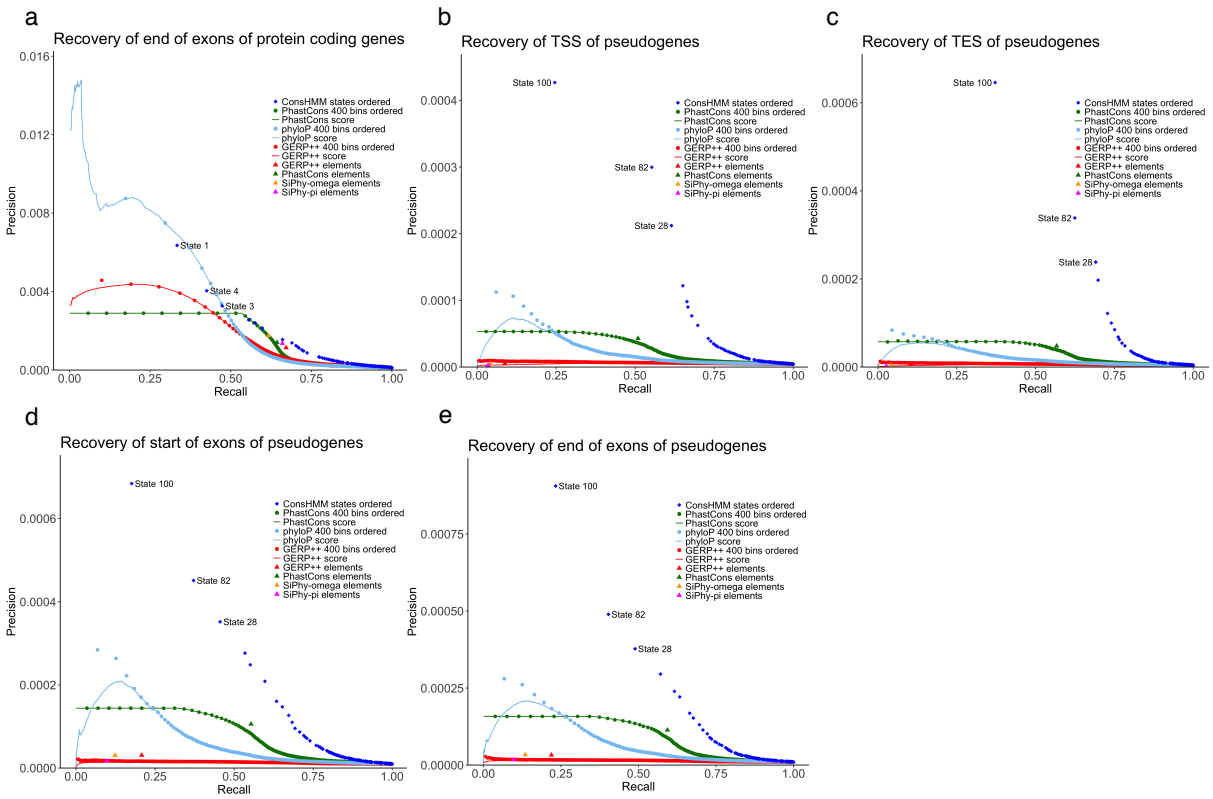


Figure 2.23 Precision-recall recovery of conservation states and constrained element and scores for additional gene annotations.

Analogous precision-recall plots to those shown in **Figure 2.22a-c**, shown here for (**a)** ends of exons of protein coding genes, **(b)** TSS of pseudogenes, **(c)** TES of pseudogenes, **(d)** start of exons of pseudogenes, and **(e)** end of exons of pseudogenes. Precision-recall values were computed using the same procedure as for **Figure 2.22** (**Methods**). The first few conservation states added are labeled.

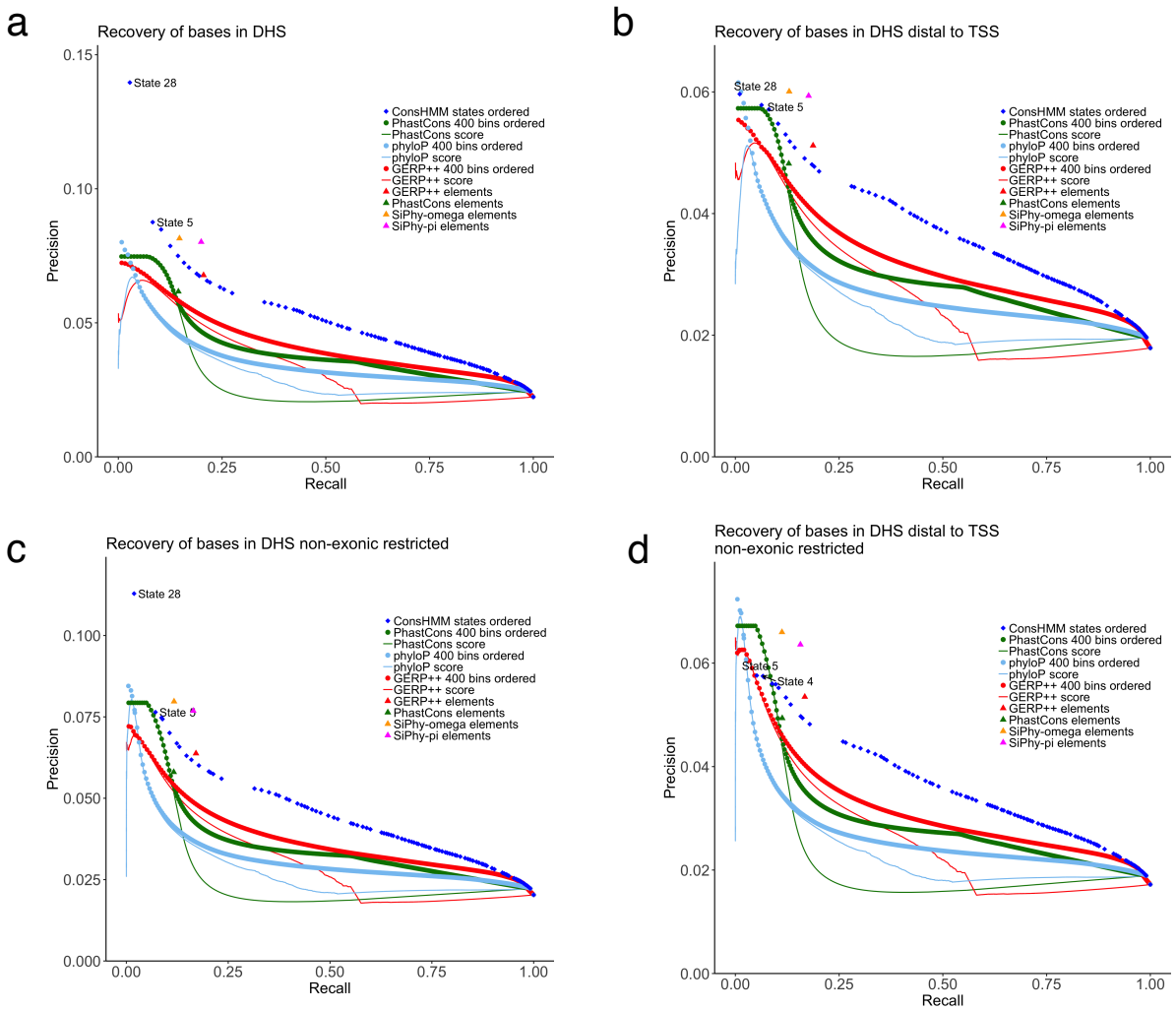


Figure 2.24 Precision-recall recovery of conservation states and constrained element sets and scores for a concatenation of DHS bases in 53 cell and tissue types.

Analogous precision-recall plots to those shown in **Figure 2.22a-c** and **Figure 2.23** shown here for DHS bases concatenated across experiments shown **(a)** without restriction and **(b-d)** with the following restrictions for the target and background: **(b)** bases more than 2kb away from a TSS, **(c)** non-exonic bases, **(d)** non-exonic bases more than 2kb away from a TSS.

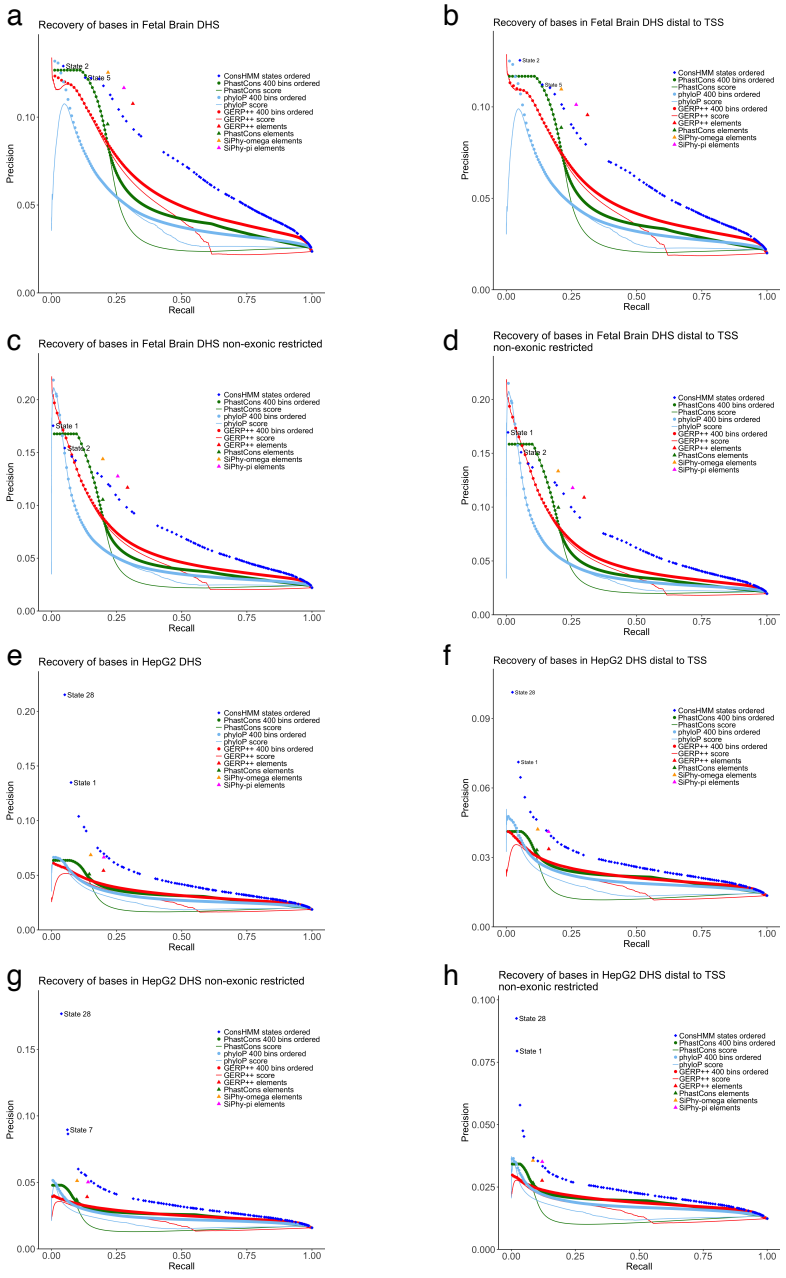


Figure 2.25 Precision-recall recovery of conservation states and constrained element and scores for DHS in two cell types.

Analogous precision-recall plots to those shown in Figure 2.22a-c, 2.23 and 19 shown here for **(a)** Fetal Brain DHS, **(b)** Fetal Brain DHS when restricting target and background to bases more than 2kb away from a TSS, **(c)** Fetal Brain DHS when restricting target and background to non-exonic bases of the genome, **(d)** Fetal Brain DHS when restricting target and background to non-exonic bases of the genome that are more than 2kb away from a TSS, **(e)** HepG2 DHS, **(f)** HepG2 DHS when restricting target and background to bases more than 2kb away from a TSS, **(g)** HepG2 when restricting target and background to non-exonic bases of the genome, **(h)** HepG2 DHS when restricting target and background to non-exonic bases of the genome that are more than 2kb away from a TSS.

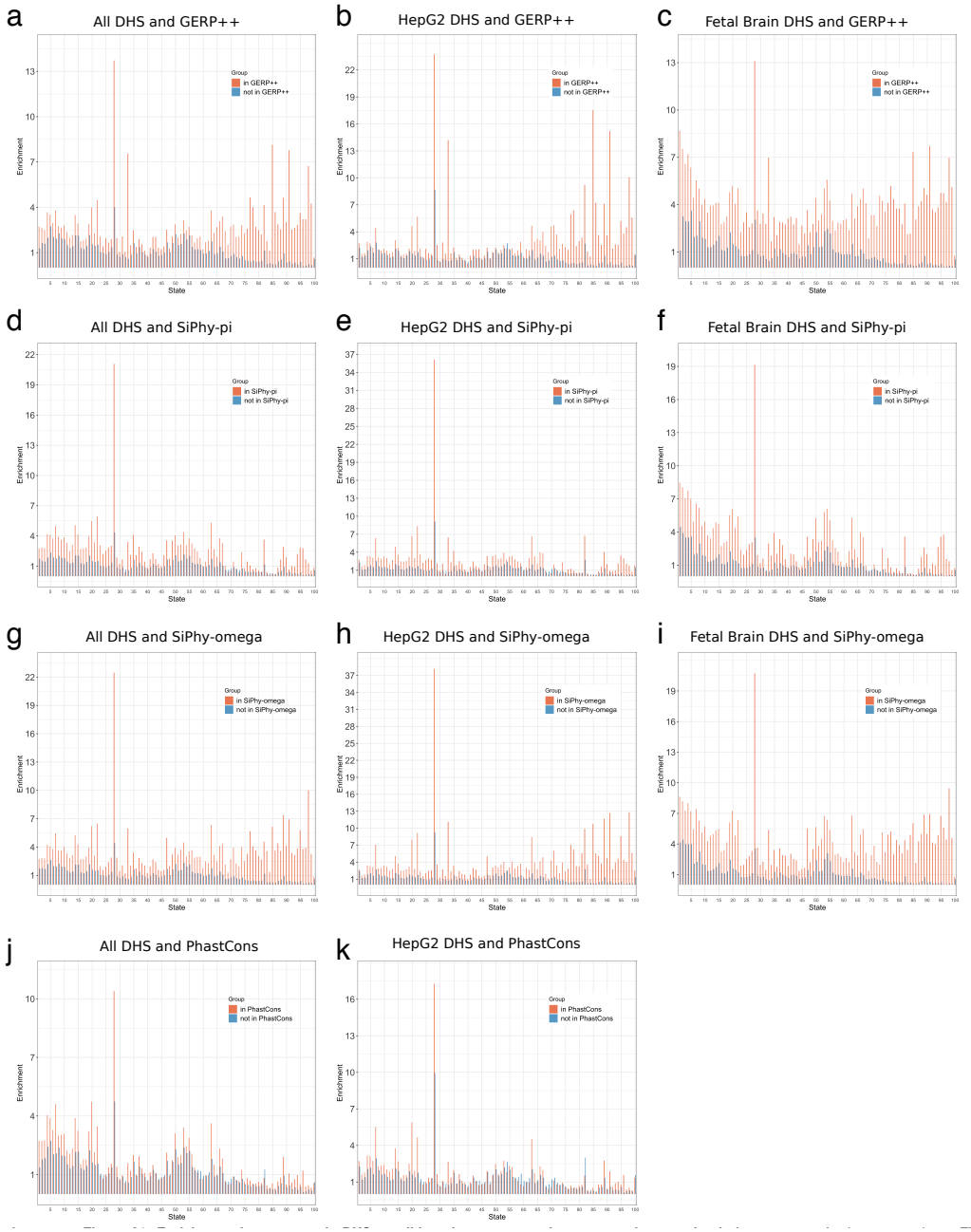


Figure 2.26 Enrichment for non-exonic DHS conditioned on conservation state and constrained element sets.

Analogous graphs to **Figure 2.22d,** showing enrichments for bases in DHS in the non-exonic portion of each conservation state conditioned on whether it is in a constrained element or not. Enrichments are shown here for **(a-c)** bases in and out of GERP++ elements for **(a)** geometric mean over 53 cell and tissue types, **(b)** HepG2 DHS, **(c)** Fetal Brain DHS; **(d-f)** bases in and out of SiPhy-pi elements for **(d)** geometric mean over 53 cell and tissue types **(e)** HepG2 DHS, **(f)** Fetal Brain DHS; **(g-i)** bases in and out of SiPhy-omega elements for **(g)** geometric mean over 53 cell and tissue types **(h)** HepG2 DHS, **(i)** Fetal Brain DHS; **(j-k)** bases in and out of PhastCons elements for **(j)** geometric mean over 53 cell and tissue types **(k)** HepG2 DHS.

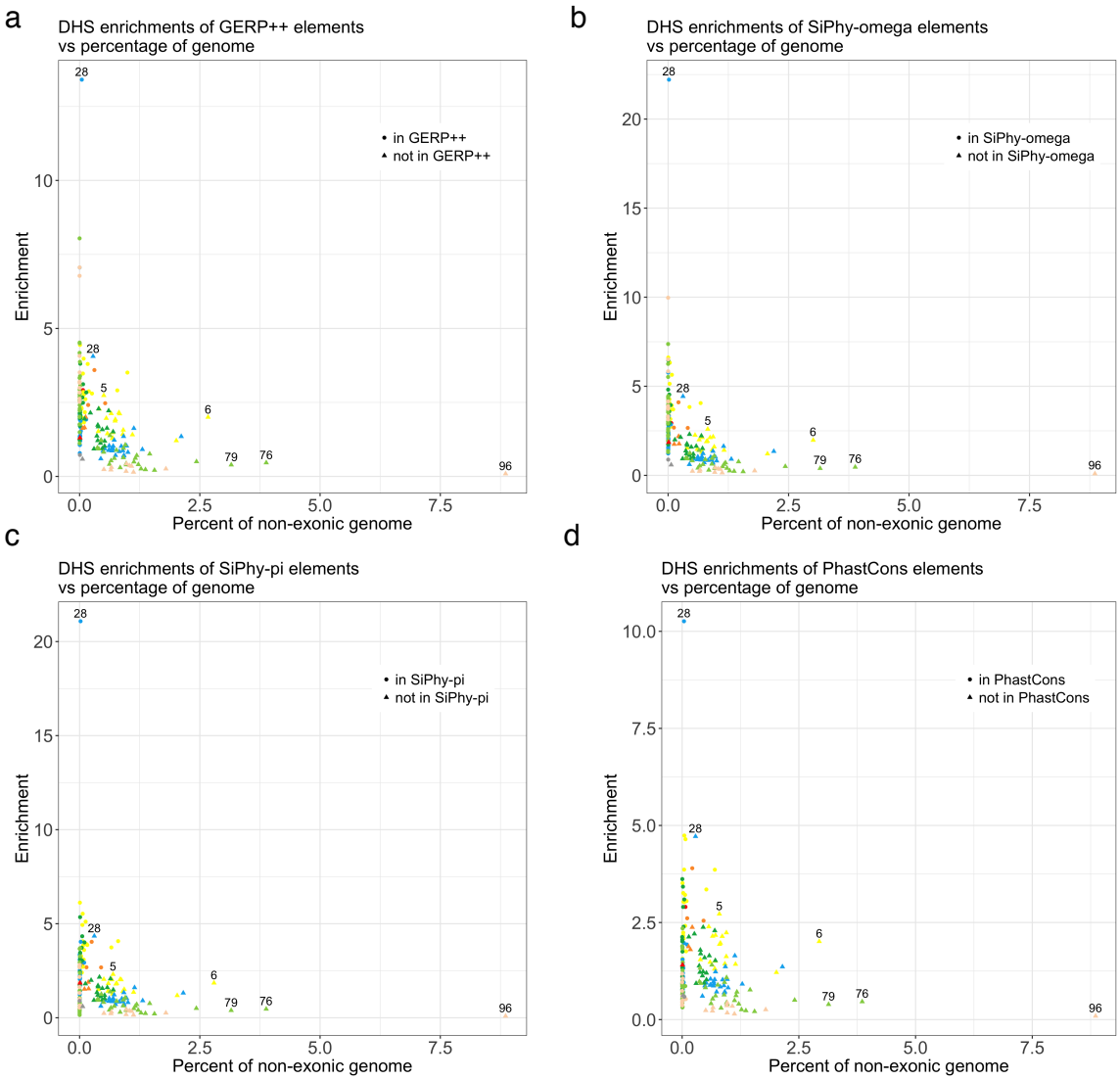
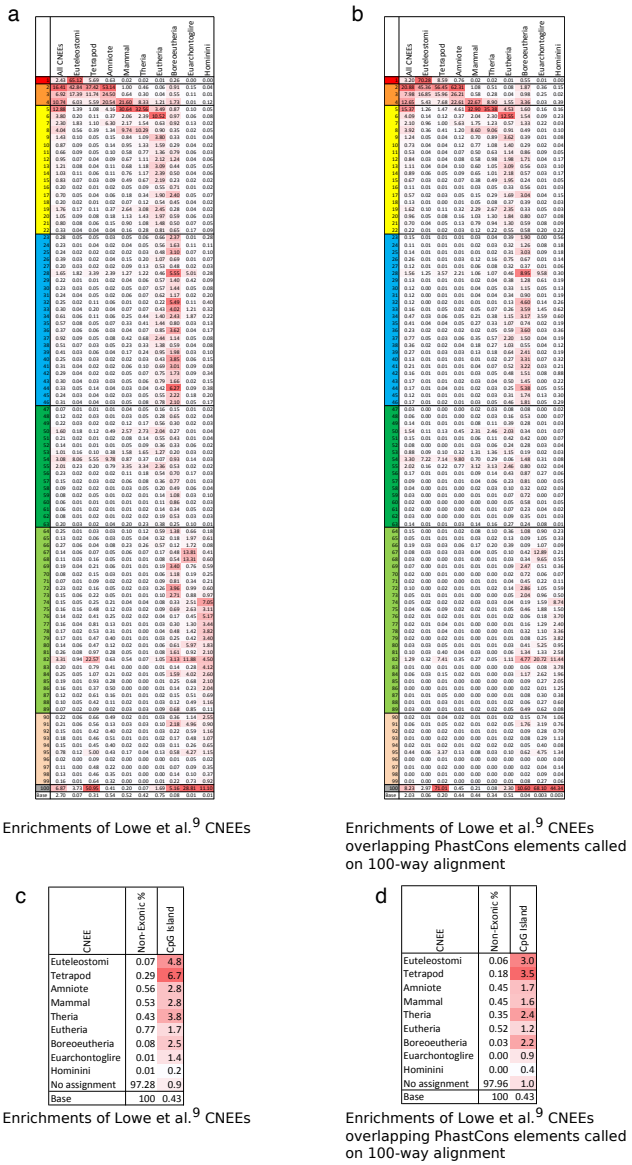


Figure 2.27 Enrichment for non-exonic DHS conditioned on conservation state and constrained element sets versus percent of non-exonic genome covered.

The enrichments in **Figure 2.26a,d,g,j** are shown here on the y-axis (geometric mean over 53 cell and tissue types), with the x-axis corresponding to the median percentage of the non-exonic genome covered by the bases falling in each category across 53 cell and tissue types. The coloring of a point corresponds to the coloring of states in **Figure 2.2**. The shape of a point corresponds to whether it is inside or outside a constrained element set according to the legend shown. Labeled points either have an enrichment >10 fold, cover >2.5% of the non-exonic genome, or are subsets of states not in a constrained element set with an enrichment >2 fold. The figure shows substantial variation of enrichments for bases both inside and outside of constrained elements depending on the conservation state, including for subsets covering non-negligible portions of the non-exonic genome. The constrained element sets used for each panel are **(a)** GERP++ elements, **(b)** SiPhy-pi elements, **(c)** SiPhy-omega elements, and **(d)** PhastCons elements.



Enrichments of Lowe et al.[9] CNEEs

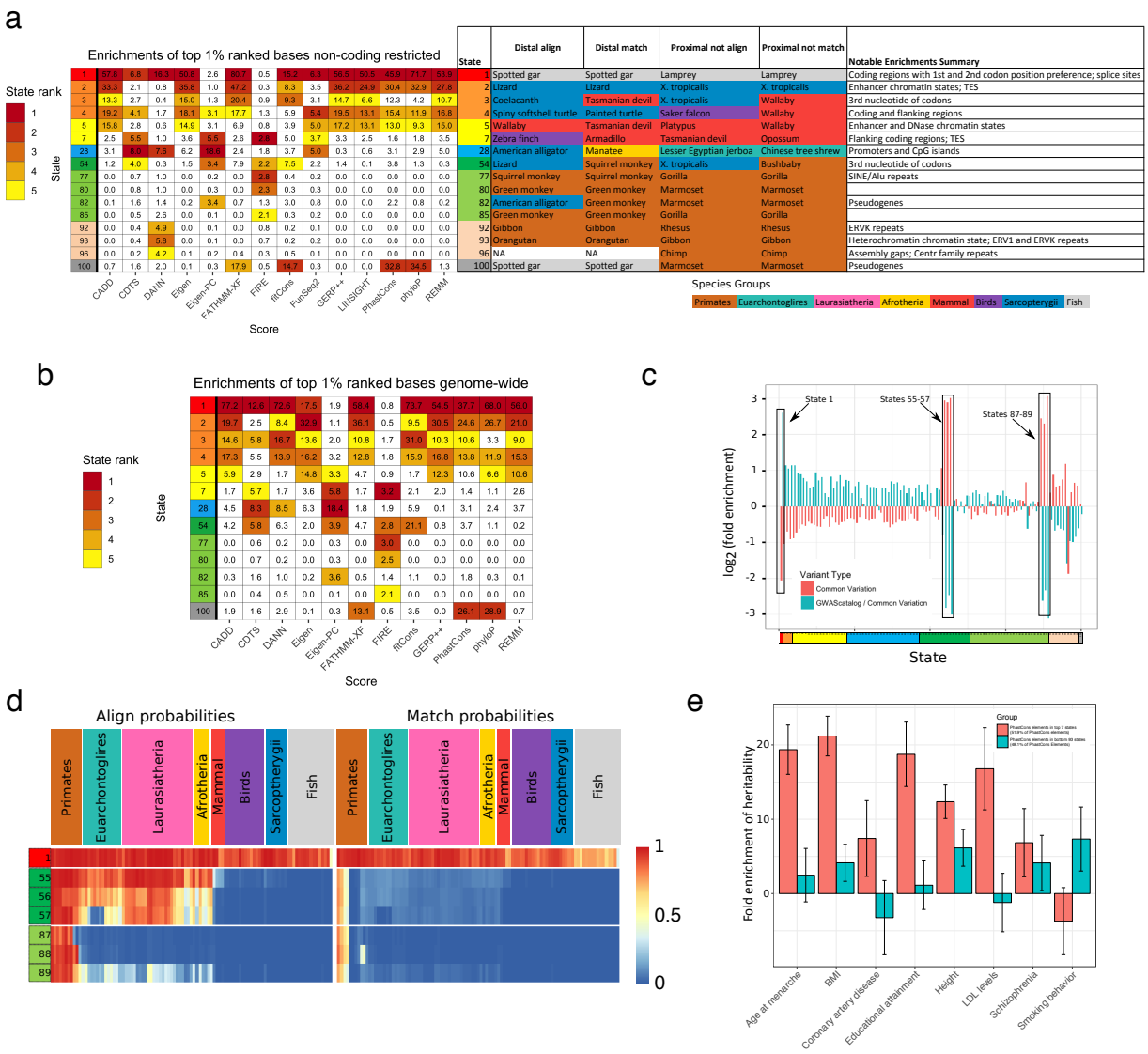
Enrichments of Lowe et al.[9] CNEEs overlapping PhastCons elements called on 100-way alignment

Enrichments of Lowe et al.[9] CNEEs

Enrichments of Lowe et al.[9] CNEEs overlapping PhastCons elements called on 100-way alignment

Figure 2.28 Relationship between conservation states and CNEEs from Ref. 9.

The conservation state enrichments for **(a)** the set of CNEEs defined in Ref. 9 and **(b)** the subset restricted to those that overlap PhastCons element bases called on the 100-way alignment in which the conservation states were defined. The first column of each panel gives the enrichment of all CNEEs included, followed by the enrichments of subset of CNEEs, based on the branch of origin of the CNEE. The second through last columns are sorted in order of distance of branch point to human. The last row gives the % of the genome covered by the annotation in the column. **(c,d)** The enrichments within non-exonic regions of bases in CpG islands of **(c)** the set of CNEEs defined in Ref. 9 and **(d)** the subset restricted to those that overlap PhastCons element bases called on the 100-way alignment in which the conservation states were defined. The first column of each panel gives the percentage of non-exonic bases found in each subset, and the second columns gives the enrichment for bases within non-exonic CpG islands. The last row gives the % of the non-exonic genome covered by the annotation in the column.



State	Distal align	Distal match	Proximal not align	Proximal not match	Notable Enrichments Summary
1	Spotted gar	Spotted gar	Lamprey	Lamprey	Coding regions with 1st and 2nd codon position preference; splice sites
2	Lizard	Lizard	X. tropicalis	X. tropicalis	Enhancer chromatin states; TES
3	Coelacanth	Tasmanian devil	X. tropicalis	Wallaby	3rd nucleotide of codons
4	Spiny softshell turtle	Painted turtle	Saker falcon	Wallaby	Coding and flanking regions
5	Wallaby	Tasmanian devil	Platypus	Wallaby	Enhancer and DNase chromatin states
7	Zebra finch	Armadillo	Tasmanian devil	Opossum	Flanking coding regions; TES
28	American alligator	Manatee	Lesser Egyptian jerboa	Chinese tree shrew	Promoters and CpG islands
54	Lizard	Squirrel monkey	X. tropicalis	Bushbaby	3rd nucleotide of codons
77	Squirrel monkey	Squirrel monkey	Gorilla	Gorilla	SINE/Alu repeats
80	Green monkey	Green monkey	Marmoset	Marmoset	
82	American alligator	Green monkey	Marmoset	Marmoset	Pseudogenes
85	Green monkey	Green monkey	Gorilla	Gorilla	
92	Gibbon	Gibbon	Rhesus	Rhesus	ERVK repeats
93	Orangutan	Orangutan	Gibbon	Gibbon	Heterochromatin chromatin state; ERV1 and ERVK repeats
96	NA	NA	Chimp	Chimp	Assembly gaps; Centr family repeats
100	Spotted gar	Spotted gar	Marmoset	Marmoset	Pseudogenes

Figure 2.29 Conservation states' association with human genetic variation.

(a) Fold enrichments of bases ranked in the top 1% of the non-coding genome by 14 variant prioritization scores. Only states among the top five most enriched states for at least one score are shown. The enrichment of the top five ranking states for each score is colored according to their ranking. The table provides a summary of the align and match probabilities and notable enrichments of each state. The 'Distal align' and 'Distal match' columns contain the species most distal to human that has an alignment and matching probability in the state greater than 0.5, respectively. The 'Proximal not align' and 'Proximal not match' columns contain the species closest to human that has an alignment and matching probability in the state lower than 0.5, respectively. The species are colored by the major clades indicated below. An expanded version including all states is available in **Supplementary Data 1**. **(b)** Enrichments of bases ranked in the top 1% genome-wide by 12 variant prioritization scores. The criteria for selecting states to display and coloring enrichments was the same as panel (a). Enrichments for prioritized bases at additional thresholds and for all states both genome-wide and for the non-coding genome are in **Figures 2.30-32**. **(c)** The log2 fold enrichment of each state for common

SNPs (pink) and GWAS catalog variants relative to common SNPs (blue). **(d)** The representation of state emission parameters from **Figure 2.2a** for the subset of states highlighted in panel (c). **(e)** Heritability partitioning enrichments from the method of Ref. 15 applied on two disjoint subsets of bases in PhastCons elements, with eight phenotypes previously analyzed with heritability partitioning in the context of a baseline annotation set (Methods). The two sets are PhastCons elements overlapping one of the seven conservation states showing the greatest enrichment for DHS in its non-exonic portion (states 1-5, 8, and 28) covering 51.9% of PhastCons bases (pink) and bases in PhastCons elements overlapping the remaining 93 states covering 48.1% of PhastCons bases (blue). Error bars represent standard errors around the enrichment estimate using jackknife resampling.

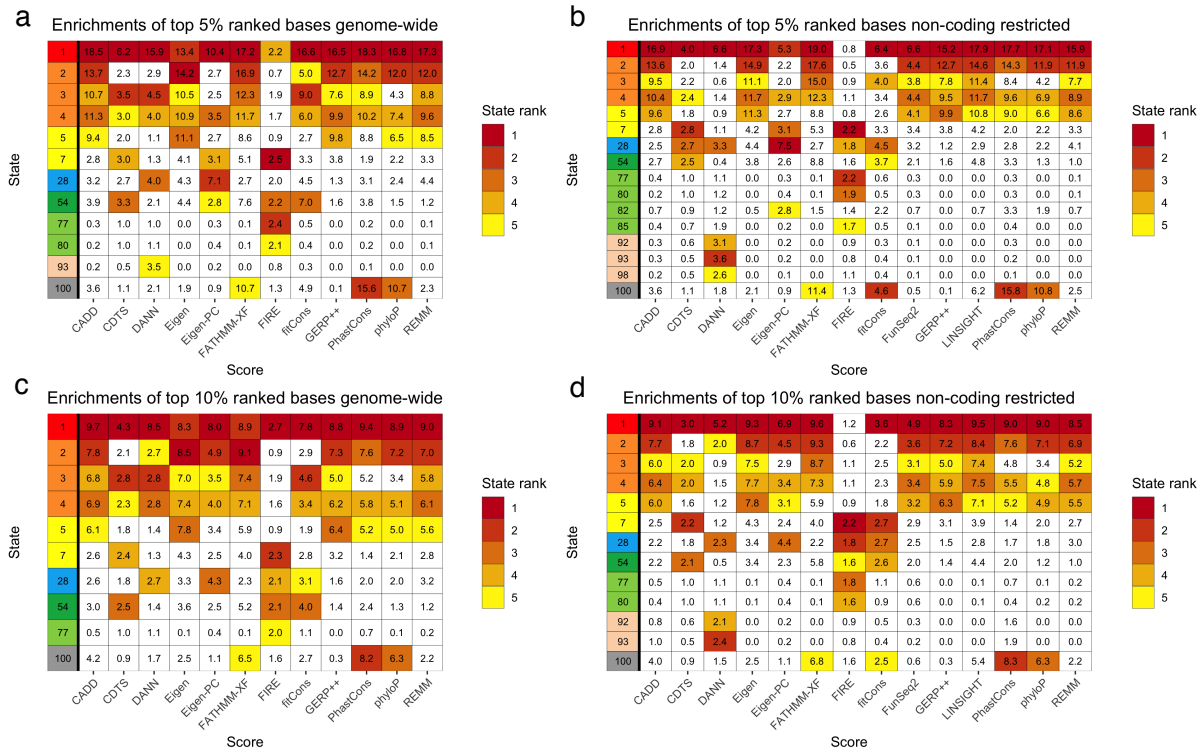


Figure 2.30 Enrichments of selected conservation states for bases prioritized by variant prioritization scores.

Analogous figures to those shown in **Figure 2.29a,b** of conservation state enrichments of top 1% bases of variant prioritization scores except shown here for: **(a)** bases ranked in top 5% genomewide, **(b)** bases ranked in the top 5% of the genome restricted to non-coding bases, **(c)** bases ranked in the top 10% genome-wide, and **(d)** bases ranked in the top 10% of the genome restricted to non-coding bases. Only states which were one of the five most enriched states by at least one variant prioritization score are shown. Coloring of enrichments is based on the rank of the state for the score as indicated in the color legend.

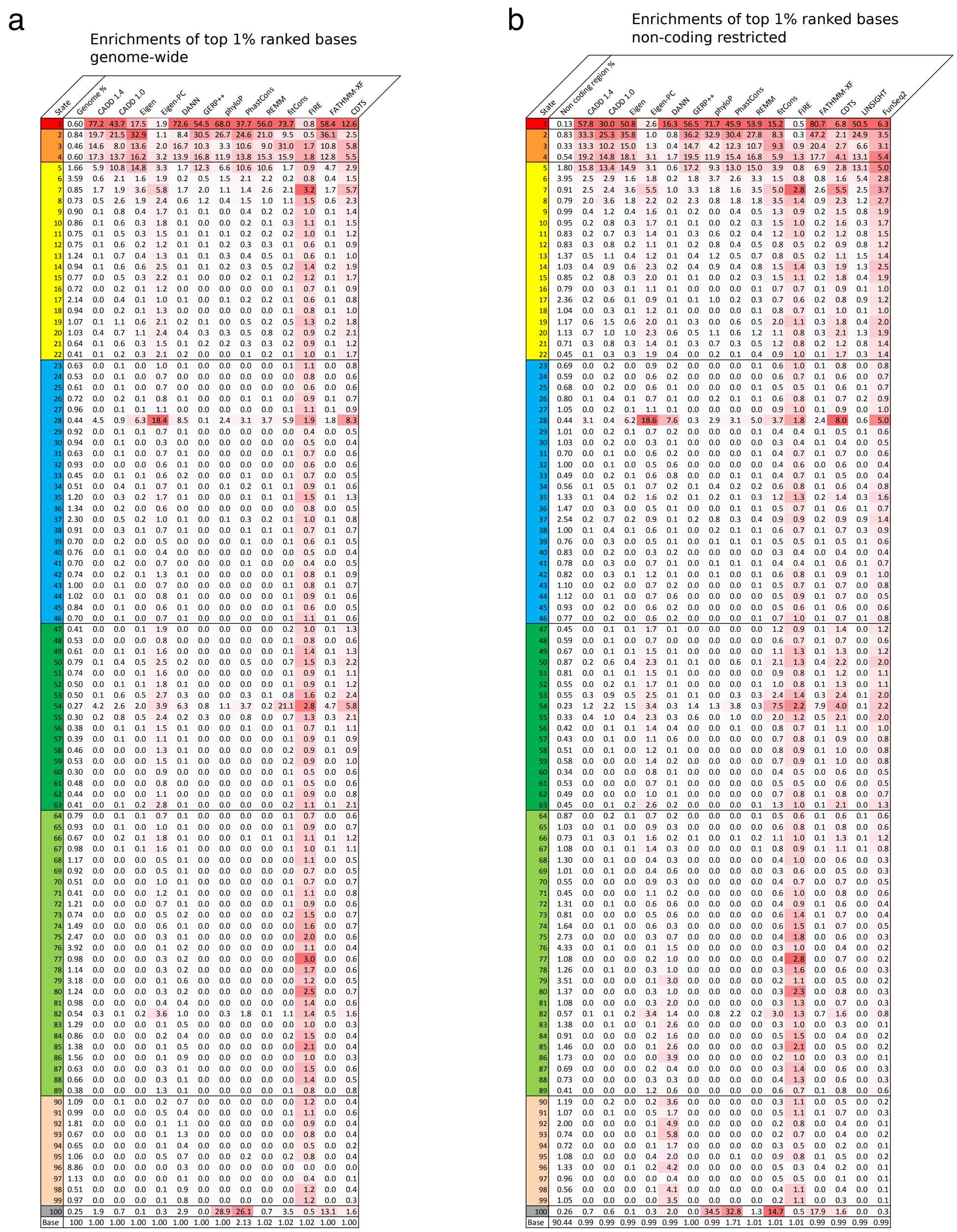


Figure 2.31 Enrichment of all conservation states for top 1% of bases prioritized by variant prioritization scores.

The figure displays for top 1% prioritized bases **(a)** genome-wide and **(b)** in non-coding regions by variant prioritization scores the same enrichments as shown in **Figure 2.29a,b** except here enrichments for all conservation states are shown and also included is the original version of the CADD score in addition to v1.4. Coloring of enrichments is based on their value in a column specific manner. The second columns gives the percentage of the background region used to compute the enrichments falling in each state, which is for **(a)** the whole genome and for **(b)** bases scored by both LINSIGHT and FunSeq2. The last line in both heatmaps gives the actual percentage of the background set covered by each set of prioritized bases, which can differ from 1.00% because of how ties were handled.

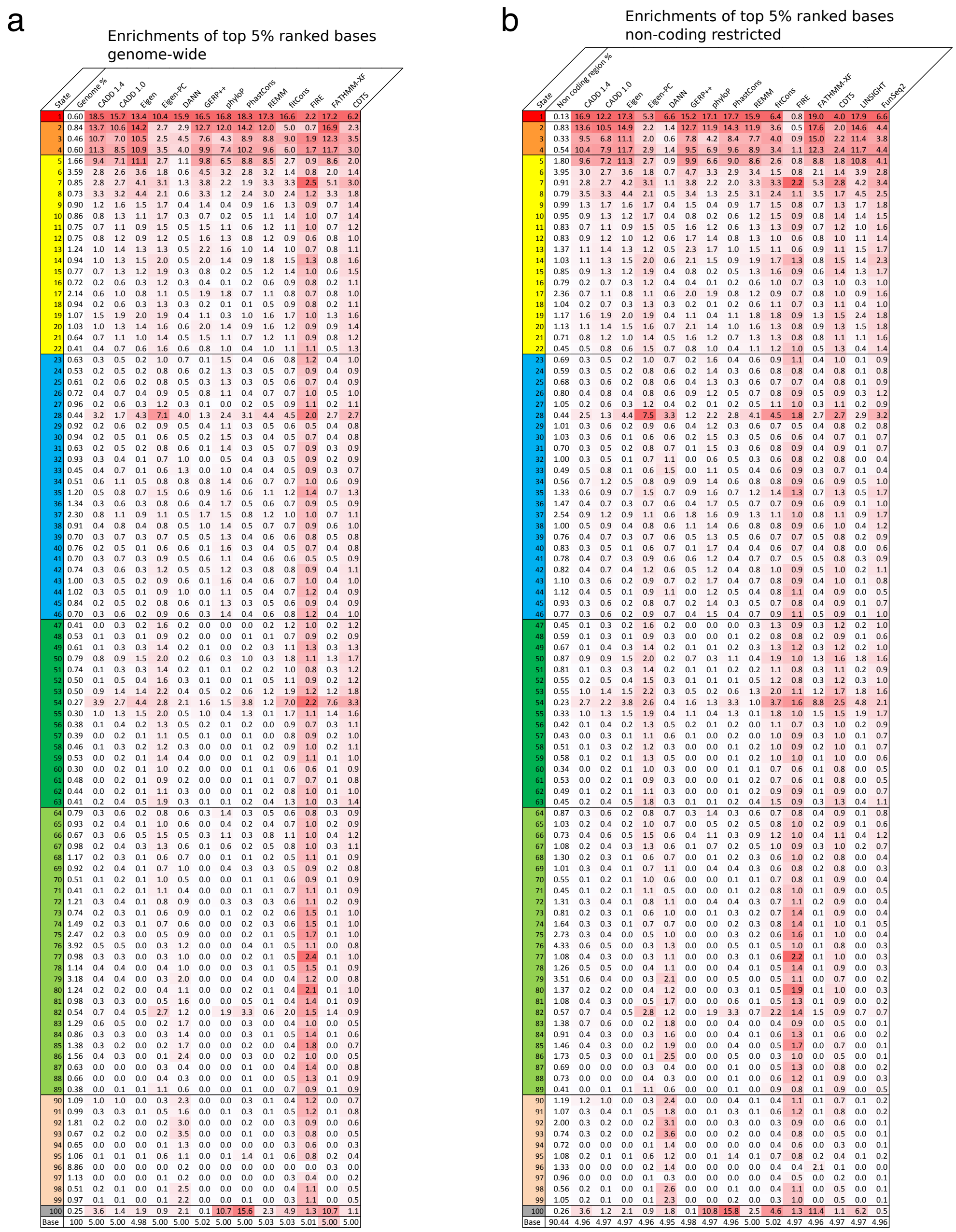


Figure 2.32 Enrichment of all conservation states for top 5% of bases prioritized by variant prioritization scores.

The figure displays for top 5% prioritized bases **(a)** genome-wide and **(b)** in non-coding regions by variant prioritization scores the same enrichments as shown in **Figure 2.30a,b** except here enrichments for all conservation states are shown and also included is the original version of the CADD score in addition to v1.4. Coloring of enrichments is based on their value in a column specific manner. The second columns gives the percentage of the background region used to compute the enrichments falling in each state, which is for **(a)** the whole genome and for **(b)** bases scored by both LINSIGHT and FunSeq2. The last line in both heatmaps gives the actual percentage of the background set covered by each set of prioritized bases, which can differ from 5.00% because of how ties were handled.

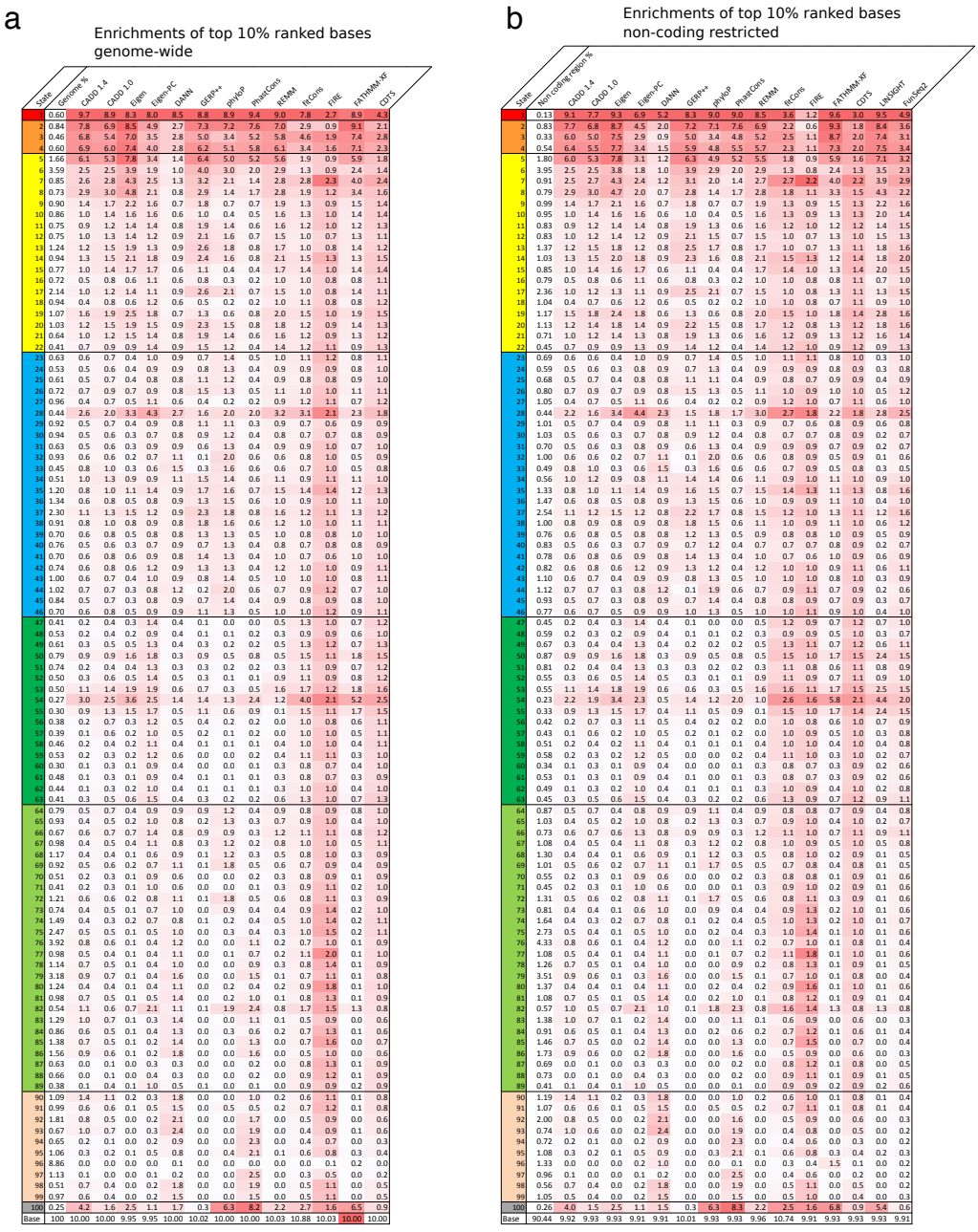
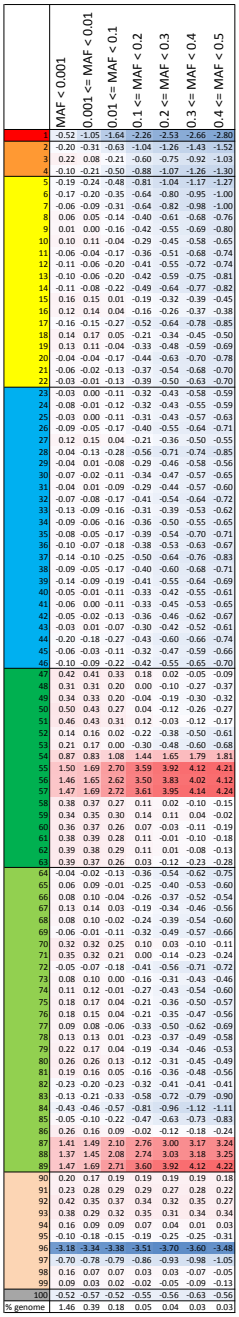


Figure 2.33 Enrichment of all conservation states for top 10% of bases prioritized by variant prioritization scores.

The figure displays for top 10% prioritized bases **(a)** genome-wide and **(b)** in non-coding regions by variant prioritization scores the same enrichments as shown in **2.30c,d** except here enrichments for all conservation states are shown and also included is the original version of the CADD score in addition to v1.4. Coloring of enrichments is based on their value in a column specific manner. The second columns gives the percentage of the background region used to compute the enrichments falling in each state, which is for **(a)** the whole genome and for **(b)** bases scored by both LINSIGHT and FunSeq2. The last line in both heatmaps gives the actual percentage of the background set covered by each set of prioritized bases, which can differ from 10.00% because of how ties were handled.



	MAF < 0.001	0.001 <= MAF < 0.01	0.01 <= MAF < 0.1	0.1 <= MAF < 0.2	0.2 <= MAF < 0.3	0.3 <= MAF < 0.4	0.4 <= MAF < 0.5
1	-0.52	-1.05	-1.64	-2.26	-2.53	-2.66	-2.80
2	-0.20	-0.31	-0.63	-1.04	-1.26	-1.43	-1.52
3	0.22	0.08	-0.21	-0.60	-0.75	-0.92	-1.03
4	-0.10	-0.21	-0.50	-0.88	-1.07	-1.26	-1.30
5	-0.19	-0.24	-0.48	-0.81	-1.04	-1.17	-1.27
6	-0.17	-0.20	-0.35	-0.64	-0.80	-0.95	-1.00
7	-0.06	-0.09	-0.31	-0.64	-0.82	-0.98	-1.00
8	0.06	0.05	-0.14	-0.40	-0.61	-0.68	-0.76
9	0.01	0.00	-0.16	-0.42	-0.55	-0.69	-0.80
10	0.10	0.11	-0.04	-0.29	-0.45	-0.58	-0.65
11	-0.06	-0.04	-0.17	-0.36	-0.51	-0.68	-0.74
12	-0.11	-0.06	-0.20	-0.41	-0.55	-0.72	-0.74
13	-0.10	-0.06	-0.20	-0.42	-0.59	-0.75	-0.81
14	-0.11	-0.08	-0.22	-0.49	-0.64	-0.77	-0.82
15	0.16	0.15	0.01	-0.19	-0.32	-0.39	-0.45
16	0.12	0.14	0.04	-0.16	-0.26	-0.37	-0.38
17	-0.16	-0.15	-0.27	-0.52	-0.64	-0.78	-0.85
18	0.14	0.17	0.05	-0.21	-0.34	-0.45	-0.50
19	0.13	0.11	-0.04	-0.33	-0.48	-0.59	-0.69
20	-0.04	-0.04	-0.17	-0.44	-0.63	-0.70	-0.78
21	-0.06	-0.02	-0.13	-0.37	-0.54	-0.68	-0.70
22	-0.03	-0.01	-0.13	-0.39	-0.50	-0.63	-0.70
23	-0.03	0.00	-0.11	-0.32	-0.43	-0.58	-0.59
24	-0.08	-0.01	-0.12	-0.32	-0.43	-0.55	-0.59
25	-0.03	0.00	-0.11	-0.31	-0.43	-0.57	-0.63
26	-0.09	-0.05	-0.17	-0.40	-0.55	-0.64	-0.71
27	0.12	0.15	0.04	-0.21	-0.36	-0.50	-0.55
28	-0.04	-0.13	-0.28	-0.56	-0.71	-0.74	-0.85
29	-0.04	0.01	-0.08	-0.29	-0.46	-0.58	-0.56
30	-0.07	-0.02	-0.11	-0.34	-0.47	-0.57	-0.65
31	-0.04	0.01	-0.09	-0.29	-0.44	-0.57	-0.60
32	-0.07	-0.08	-0.17	-0.41	-0.54	-0.64	-0.72
33	-0.13	-0.09	-0.16	-0.31	-0.39	-0.53	-0.62
34	-0.09	-0.06	-0.16	-0.36	-0.50	-0.55	-0.65
35	-0.08	-0.05	-0.17	-0.39	-0.54	-0.70	-0.71
36	-0.10	-0.07	-0.18	-0.38	-0.53	-0.63	-0.67
37	-0.14	-0.10	-0.25	-0.50	-0.64	-0.76	-0.83
38	-0.09	-0.05	-0.17	-0.40	-0.60	-0.68	-0.71
39	-0.14	-0.09	-0.19	-0.41	-0.55	-0.64	-0.69
40	-0.05	-0.01	-0.11	-0.33	-0.42	-0.55	-0.61
41	-0.06	0.00	-0.11	-0.33	-0.45	-0.53	-0.65
42	-0.05	-0.02	-0.13	-0.36	-0.46	-0.62	-0.67
43	-0.03	0.01	-0.07	-0.30	-0.42	-0.52	-0.61
44	-0.20	-0.18	-0.27	-0.43	-0.60	-0.66	-0.74
45	-0.06	-0.03	-0.11	-0.32	-0.47	-0.59	-0.66
46	-0.10	-0.09	-0.22	-0.42	-0.55	-0.65	-0.70
47	0.42	0.41	0.33	0.18	0.02	-0.05	-0.09
48	0.31	0.31	0.20	0.00	-0.10	-0.27	-0.37
49	0.34	0.33	0.20	-0.04	-0.19	-0.30	-0.32
50	0.50	0.43	0.27	0.04	-0.12	-0.26	-0.27
51	0.46	0.43	0.31	0.12	-0.03	-0.12	-0.17
52	0.14	0.16	0.02	-0.22	-0.38	-0.50	-0.61
53	0.21	0.17	0.00	-0.30	-0.48	-0.60	-0.68
54	0.87	0.83	1.08	1.44	1.65	1.79	1.81
55	1.50	1.69	2.70	3.59	3.92	4.12	4.21
56	1.46	1.65	2.62	3.50	3.83	4.02	4.12
57	1.47	1.69	2.72	3.61	3.95	4.14	4.24
58	0.38	0.37	0.27	0.11	0.02	-0.10	-0.15
59	0.34	0.35	0.30	0.14	0.11	0.04	-0.02
60	0.36	0.37	0.26	0.07	-0.03	-0.11	-0.19
61	0.38	0.39	0.28	0.11	-0.01	-0.10	-0.18
62	0.39	0.38	0.29	0.11	0.01	-0.08	-0.13
63	0.39	0.37	0.26	0.03	-0.12	-0.23	-0.28
64	-0.04	-0.02	-0.13	-0.36	-0.54	-0.62	-0.75
65	0.06	0.09	-0.01	-0.25	-0.40	-0.53	-0.60
66	0.08	0.10	-0.04	-0.26	-0.37	-0.52	-0.54
67	0.13	0.14	0.03	-0.19	-0.34	-0.46	-0.56
68	0.08	0.10	-0.02	-0.24	-0.39	-0.54	-0.60
69	-0.06	-0.01	-0.11	-0.32	-0.49	-0.57	-0.66
70	0.32	0.32	0.25	0.10	0.03	-0.10	-0.11
71	0.35	0.32	0.21	0.00	-0.14	-0.23	-0.24
72	-0.05	-0.07	-0.18	-0.41	-0.56	-0.71	-0.72
73	0.08	0.10	0.00	-0.16	-0.31	-0.43	-0.46
74	0.11	0.12	-0.01	-0.27	-0.43	-0.54	-0.60
75	0.18	0.17	0.04	-0.21	-0.36	-0.50	-0.57
76	0.18	0.15	0.04	-0.21	-0.35	-0.47	-0.56
77	0.09	0.08	-0.06	-0.33	-0.50	-0.62	-0.69
78	0.13	0.13	0.01	-0.23	-0.37	-0.49	-0.58
79	0.22	0.17	0.04	-0.19	-0.34	-0.46	-0.53
80	0.26	0.26	0.13	-0.12	-0.31	-0.45	-0.49
81	0.19	0.16	0.05	-0.16	-0.36	-0.48	-0.56
82	-0.23	-0.20	-0.23	-0.32	-0.41	-0.41	-0.41
83	-0.13	-0.21	-0.33	-0.58	-0.72	-0.79	-0.90
84	-0.43	-0.46	-0.57	-0.81	-0.96	-1.12	-1.11
85	-0.05	-0.10	-0.22	-0.47	-0.63	-0.73	-0.83
86	0.26	0.16	0.09	-0.02	-0.12	-0.18	-0.24
87	1.41	1.49	2.10	2.76	3.00	3.17	3.24
88	1.37	1.45	2.08	2.74	3.03	3.18	3.25
89	1.47	1.69	2.71	3.60	3.92	4.12	4.22
90	0.20	0.17	0.19	0.19	0.19	0.19	0.18
91	0.23	0.28	0.29	0.29	0.27	0.28	0.22
92	0.42	0.35	0.37	0.34	0.32	0.35	0.27
93	0.38	0.29	0.32	0.35	0.31	0.34	0.34
94	0.16	0.09	0.09	0.07	0.04	0.01	0.03
95	-0.10	-0.18	-0.15	-0.19	-0.25	-0.25	-0.31
96	-3.18	-3.34	-3.38	-3.51	-3.70	-3.60	-3.48
97	-0.70	-0.78	-0.79	-0.86	-0.93	-0.98	-1.05
98	0.16	0.07	0.07	0.03	0.03	-0.07	-0.05
99	0.09	0.03	0.02	-0.02	-0.05	-0.09	-0.13
100	-0.52	-0.57	-0.52	-0.55	-0.56	-0.63	-0.56
% genome	1.46	0.39	0.18	0.05	0.04	0.03	0.03

Figure 2.34 Conservation state enrichments for single nucleotide variants from Ref. 10.

Rows 1-100 corresponds to states, color coded based on their group, and the last line represents the percentage of the genome covered by each annotation in the columns. The table displays the log2 fold enrichments of all the single nucleotide variants from whole genome sequencing data of 7794 unrelated individuals that were used to generate the context dependent tolerance score. The variants are grouped into disjoint sets according to minor allele frequency (MAF). Depletions are shown in shades of blue and enrichments in shades of red. The large depletions in state 96 are due to the state capturing assembly gaps.

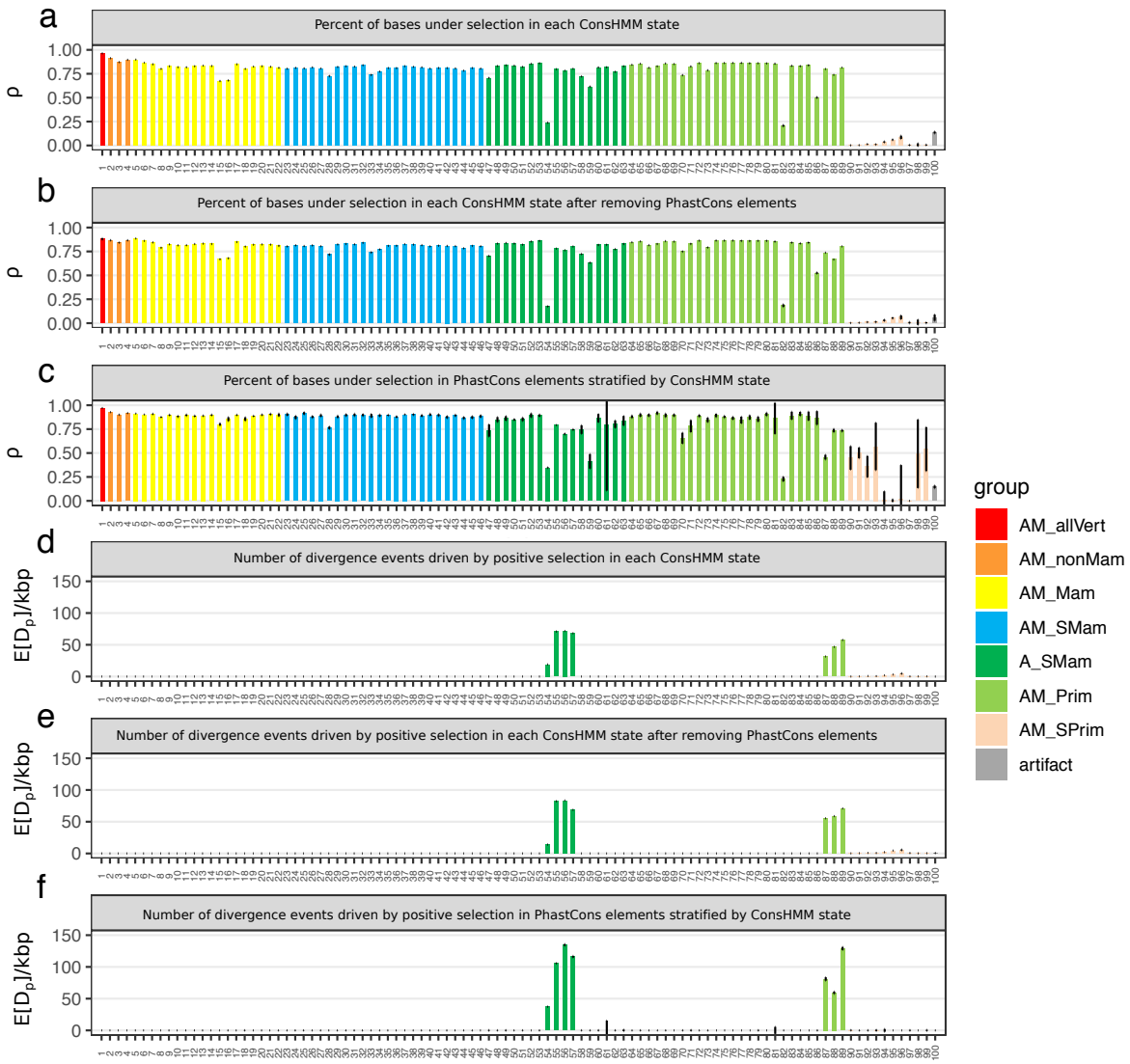


Figure 2.35 Results of running the INSIGHT model.

(a-c) The estimated fraction of bases under selection (ρ) as estimated by the INSIGHT method within **(a)** each conservation state, **(b)** each conservation state after removing bases in PhastCons elements and **(c)** each state restricted only to bases in PhastCons elements. **(d-f)** The estimated number of divergence events driven by positive selection per kilobase-pair (E[Dp]/kbp) as estimated by the INSIGHT method within **(d)** each conservation state, **(e)** each state after removing bases in PhastCons elements and **(f)** each state restricted only to bases in PhastCons elements. States are colored according to their group as indicated on the right. Error bars represent one standard error around each parameter estimate.

Chapter 3. ConsHMM Atlas: conservation state annotations for major genomes and human genetic variation

3.1 Introduction

We recently introduced the ConsHMM method[60] to annotate reference genomes at single-nucleotide resolution into a number of different 'conservation states' based on the combinatorial and spatial patterns of which species have a nucleotide aligning to and/or matching the reference genome in a multi-species DNA sequence alignment. ConsHMM does this using a multivariate hidden Markov model (HMM), building off the widely used ChromHMM approach for modeling epigenomic data[34], without making any explicit phylogenetic modeling assumptions. Each nucleotide in the reference genome receives an annotation corresponding to the state of the HMM with the maximum posterior probability.

ConsHMM annotations are complementary to previous whole genome comparative genomic annotations, which have primarily focused on univariate scores or binary element calls of constraint[3,7,61,62]. We previously applied ConsHMM to annotate one reference genome, human hg19, based on a 100-way vertebrate alignment[60]. The conservation states had diverse and biologically meaningful enrichments for other genomic annotations, and were also able to isolate putative artifacts in the underlying multiple sequence alignment, which can confound other constraint annotations.

Here we report applying ConsHMM to produce an additional 21 genome annotations for different reference genomes and based on different multi-species DNA sequence alignments. In addition to human, seven other organisms are represented in these additional genome annotations. Additionally, we have extended the ConsHMM software to also produce allele specific annotations opposed to only position specific annotations based on the reference allele. We have applied this to produce annotations for each possible single-nucleotide mutation for

every nucleotide in the human genome. To aid in the analysis of different ConsHMM models we have created a web-interface for interactive visualization of model parameters and annotation enrichments. These new annotations of the human genome and variation as well as model organism genomes and visualization tool comprise the ConsHMM Atlas, which we expect to be a valuable resource to community for analyzing various genomes and genetic variation.

3.2 New Approaches

ConsHMM annotations for additional organisms and multiple-sequence alignments

We generated an additional 21 ConsHMM genome annotations that in addition to human genome include annotations for the mouse, rat, dog, zebrafish, fruit fly, C. elegans and S. cervisiae genomes (**Table 3.1, Supplementary Data 2**). For some species we generated multiple different genome annotations that corresponded to different sets of species in the multi-species alignment, different alignment methods used to generate the alignment, or different assemblies of the reference genome. All alignments we used were obtained from the UCSC genome browser or Ensembl[63,64]. We applied ConsHMM as previously described[60], except setting the number of states for a model based on the number of species in the alignment (**Methods**).

We highlight as an illustrative example of one of the new ConsHMM models that we learned, the model based on the 60-way Multiz alignment of 59 vertebrates to the mouse mm10 genome (**Figure 3.1a**). In this model, which has 60 states, ConsHMM identified a number of noteworthy states showing enrichment for other external genomic annotations (**Figure 3.1b, Supplementary Data 2**). For example, a state that showed high aligning and matching probabilities in all the species in the alignment, state 60, was the most enriched state for exons (34.9 fold). A different state showed a pattern of moderate probabilities of aligning and matching for almost all species, and showed strong enrichment for CpG islands (50.4 fold) and TSS (34

fold). Another state, state 1, had high aligning probabilities only in distal species to mouse, which is likely capturing alignment artifacts, though still had a 12 fold enrichment for PhastCons constrained element calls. There were three other states (states 2, 3, and 13), which had similar though weaker versions of the state 1 alignment pattern and also enriched for PhastCons elements (3.8-6.4 fold). These different state patterns and corresponding enrichment were similar to those found for a previously analyzed human conservation state annotations[60].

Allele Specific ConsHMM annotations

Previously ConsHMM could only generate position specific conservation states based on the allele present in the reference genome. As ConsHMM models the observation of whether the nucleotide present in each other species matches the reference genome, an alternate allele at a position could potentially lead to a very different conservation state assignment. Allele specific annotations could thus be informative to studying genetic variation, but directly applying ConsHMM for every observed variant would not be computationally practical.

To address this challenge we extended ConsHMM to be able to compute conservation state assignments for any alternate allele with high accuracy under two assumptions. The first assumption is that it is sufficient to assume an alternate allele would not cause changes to the multi-species alignment except for the nucleotide present in the reference genome. The second assumption is that it is sufficient to consider a small local window around each variant to derive a state annotation opposed to segmenting 200kb at time as previously done[60]. We empirically verified this second assumption by considering a range of window sizes upstream and downstream of a variant and showing that a window of size 21 (10 bases upstream and 10 bases downstream) obtained 99.6% agreement in the conservation state assignments compared to applying ConsHMM as previously applied for a set of 40,000 common variants (**Methods**, **Figure 3.2**).

Using this extended version of ConsHMM we produce allele specific conservation state annotations for each possible single nucleotide alternate allele for both the hg19 and hg38 human reference genomes based on ConsHMM models trained on 100-way vertebrate alignments. To demonstrate the additional information in having allele specific conservation state assignments beyond just the reference genome we consider the set of positions that were assigned to a state in the human hg38 model associated with high probability of aligning through mammals, but a high probability of matching in a only a few primates, state 36. We then analyzed for different subsets of positions the frequency at which an alternate allele caused the conservation state assignment to a very different state that had high probability of both aligning and matching in many mammals, state 5 (**Figure 3.3a**). For only 0.8% of possible alternate alleles for reference allele in state 36 did we see the conservation state assignment change to state 5. Interestingly, we saw this percentage increase substantially for subsets of positions with unique annotations. Among positions in Fetal Brain DNase I hypersensitive sites[9] the percentage was 2% and for those in GERP++ constrained elements[6] it was 5% (**Figure 3.3b**). The percentage increased to 7% for those annotated as both. The percentage increased even further to 12% for previously annotated bases in human accelerated regions (HAR)[5]. Similar percentages were found when using other sets of conserved elements and another Fetal Brain DNase I hypersensitivity data set (**Figure 3.4**). These results highlight how allele specific conservation state assignments provide additional information beyond the conservation state assignment from the reference allele.

Web-interface for visualization of parameters and annotation enrichments of ConsHMM models

We created a web interface built on an R shiny app in which one can browse a representation of emission parameters of ConsHMM models and annotation enrichments:

https://ernstlab.shinyapps.io/conshmm (**Figure 3.5**). Users can access the models trained on each of the reference genomes and multiple sequence alignments listed in **Table 3.1**. The app generates an interactive heatmap containing for each state and species the model probability for that species aligning the reference genome and also the probability of having a nucleotide matching the human reference. The interface allows a user to select a subset of states and/or species in the alignment to display, for ease of visualization. Lastly, the interface allows users to display precomputed enrichments of states for external annotations. These include enrichments for existing annotations of gene bodies, exons, transcription start and end sites, and the PhastCons elements called on the same alignment, when available.

3.3 Methods

Data and code availability

The Ensembl multiple sequence alignments were downloaded from ftp://ftp.ensembl.org/pub/release-97/maf/ensembl-compara/ and ftp://ftp.ensembl.org/pub/release-75/maf/ensembl-compara/[29]. The UCSC multiple sequence alignments listed in **Table 3.1** were downloaded from https://hgdownload.soe.ucsc.edu/downloads.html[64]. The Ensembl multiple sequence alignments listed in **Table 3.1** were downloaded from ftp://ftp.ensembl.org/pub/release-97/maf/ensembl-compara/ and ftp://ftp.ensembl.org/pub/release-75/emf/ensembl-compara/[29].

SiPhy-omega, SiPhy-pi constrained element calls, and HAR calls were downloaded from https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info[5,17].

Fetal Brain DNase I Hypersensitivity Sites were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/[9].

PhastCons constrained element calls, RefSeq and CpG Island annotations and dbSNP v150 variants were obtained from the UCSC genome browser. The ConsHMM model parameters and the corresponding genomic segmentations and annotations are available at http://www.biolchem.ucla.edu/labs/ernst/ConsHMMAtlas. The allele specific state annotations for the human genome can also be found at the same URL. The ConsHMM software is available at https://github.com/ernstlab/ConsHMM.

Learning ConsHMM annotations for reference genomes

We used ConsHMM v1.0 as described in Arneson and Ernst (2019) to learn the models parameters, to generate the segmentation and annotation of the reference genomes, and to compute the enrichments for external annotations. We used the same parameters except the number of states parameters. The number of states we used for each alignment depended on the number of species in the alignment. Specifically, if the alignment had more than 50 species, then the number of states was equivalent to the number of species in the alignment; if the alignment had between 25 and 49 species, then the number of states was set to 50; if the alignment had less than 25 species, then the number of states was set to 25. This set of rules allows for the number of states to be dependent on the number of species in the alignment, while also ensuring a sufficient, but not excessive, number of states for alignments with smaller number of species.

Creating allele-specific ConsHMM annotations

To generate allele specific ConsHMM annotations we used ConsHMM v1.1, containing the new updateInitialParams and ReassignVariantState commands and ChromHMM v1.20. The updateInitialParams takes as input the parameters of a ConsHMM model and a genome wide segmentation, and outputs an updated parameter set where the initial state parameters are

replaced by the genome wide frequency of each state in the segmentation, to better reflect the state assignment prior for a variant at any position in the genome. The ReassignVariantState command takes as input the ConsHMM model outputted by updateInitialParams, a file containing the multiple alignment on which the model is based, and a parameter *W*. The initial state parameters were obtained using the updateInitialparam command. The file containing the multiple alignment is in a format processed by the parseMAF command of ConsHMM. The parameter *W* indicates *W* flanking bases upstream and also *W* bases downstream of the allele.

We note that since ConsHMM uses an HMM the state assignment at a position of interest can depend on the observations at neighboring positions. We investigated the effect of different choices of *W* by first sampling a set of 40,000 common variants from dbSNP[65] that are further than 200kb apart, the segment size previously used with ConsHMM for genome segmentations, and applying ConsHMM with the alternate allele for those common variants[60]. We did this with the ConsHMM model for hg38 based on the 100-way vertebrate alignment. We then compared the agreement in the conservation state assignment when we apply ReassignVariantState with values of W between 1 and 10 and found that the agreement between the procedures plateaued at 99.6%. The final allele-specific annotations were generated using *W* = 10, for each possible nucleotide as the reference at the base in the center of the window. For variants in which the flanking region extends past the beginning or end of chromosomes, the missing bases upstream or downstream of the position of interest were marked as positions where the multiple sequence alignment is empty, which ConsHMM encodes as positions where no species align to the reference species.

3.4 Tables

Organism	Target assembly	Browser	Alignment Method	Alignment
C. elegans	ce11	UCSC	MultiZ	25 nematode genomes with C. elegans
D. melanogaster	dm6	UCSC	MultiZ	26 insects with D. melanogaster
Human	Hg38/GRCh38	UCSC	MultiZ	99 vertebrate genomes with human
Human	Hg38/GRCh38	UCSC	MultiZ	30 mammalian (27 primate) genomes with human
Human	Hg19/GRCh37	Ensembl Release 75	PECAN	21 amniota vertebrates
Human	Hg19/GRCh37	Ensembl Release 75	EPO_LOW_COVERAGE	37 eutherian mammals
Human	Hg38/GRCh38	Ensembl Release 97	PECAN	54 amniota vertebrates
Human	Hg38/GRCh37	Ensembl Release 97	EPO	38 mammals
Human	Hg38/GRCh37	Ensembl Release 97	EPO_LOW_COVERAGE	91 eutherian mammals
Mouse	mm10/GRCm38	UCSC	MultiZ	59 vertebrate genomes with mouse
Mouse	mm10/GRCm38	Ensembl Release 97	PECAN	54 amniota vertebrates
Mouse	mm10/GRCm38	Ensembl Release 97	EPO	38 mammals
Mouse	mm10/GRCm38	Ensembl Release 97	EPO_LOW_COVERAGE	91 eutherian mammals
Rat	rn6/Rnor_6.0	UCSC	MultiZ	19 vertebrate genomes with rat
Rat	rn6/Rnor_6.0	Ensembl Release 97	PECAN	54 amniota vertebrates
Rat	rn6/Rnor_6.0	Ensembl Release 97	EPO	38 mammals
Rat	rn6/Rnor_6.0	Ensembl Release 97	EPO_LOW_COVERAGE	91 eutherian mammals
Zebrafish	Zv9/DanRer7	UCSC	MultiZ	7 genomes with zebrafish
Zebrafish	Zv9/DanRer7	Ensembl Release 75	EPO_LOW_COVERAGE	10 teleost fish
Zebrafish	GRCz11/DanRer11	Ensembl Release 97	EPO	Ensembl 97 - 25 fish
S. cerevisiae	sacCer3	UCSC	MultiZ	6 yeast species to S. cerevisiae
Dog	canFam3/CanFam3.1	Ensembl Release 97	EPO	38 mammals

Table 3.1 List of organisms and respective multiple sequence alignments.
The ‘Organism’ column contains the species used as the reference in the multiple sequence alignment. The ‘Target assembly’ column contains the genome draft version used for each organism. The ‘Browser’ column contains the genome browser from where the multiple sequence alignment was downloaded. The ‘Alignment method’ column contains the name of the multiple sequence alignment method used to generate the alignment. The ‘Alignment’ column contains a summary of the species in the multiple sequence alignment.

3.5 Figures

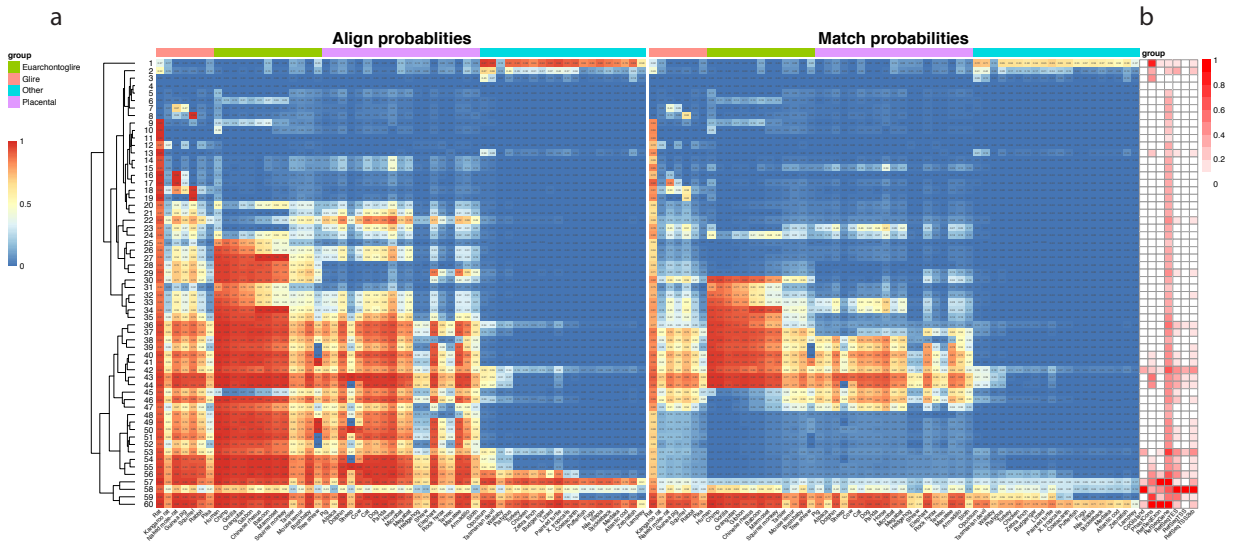


Figure 3.1 Conservation state emission parameters of a ConsHMM model based on a 60-way alignment of vertebrates to mouse and enrichments for other genomic annotations.

(a) The rows of the heatmap correspond to conservation states and the columns of the heatmap correspond to species. For each state and species, the left half of the heatmap contains the probability of that species aligning to the mouse sequence (one minus the probability of not aligning). The right half of the heatmap contains the probability of a species matching the mouse sequence. Species are ordered by phylogenetic distance to mouse and grouped by major clades. States are ordered by ConsHMM hierarchical clustering. **(b)** The columns of the heatmap indicate the relative enrichments of conservation states for CpG Islands, PhastCons elements, RefSeq exons, genes, transcription start and end sites. (**Methods**). The relative enrichments were calculated by subtracting the minimum value of the column from each enrichment and dividing by the range of the column. **Supplementary Data 2** the values of these enrichments.

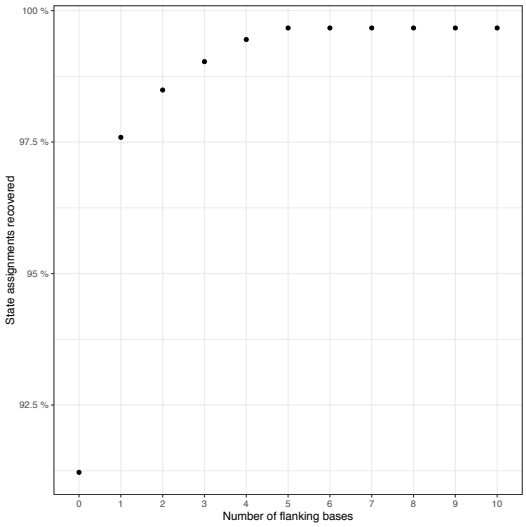
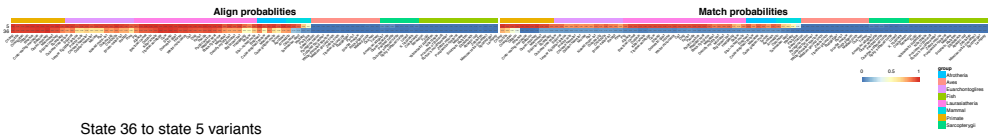


Figure 3.2 The agreement between state assignments using the segmentation of a local window centered around a variant and the segmentation of 200kb segments in the entire genome.

The x-axis represents the number of bases upstream and the number of bases downstream of the variant of interest used. The y-axis represents the percentage of the 40,000 variants tested for which the alternate allele gets assigned to the same state with the two approaches. This comparison was performed for the 100 state ConsHMM model trained on a 100-way multiple sequence alignment of 99 other species to the hg38 human genome.

a



b

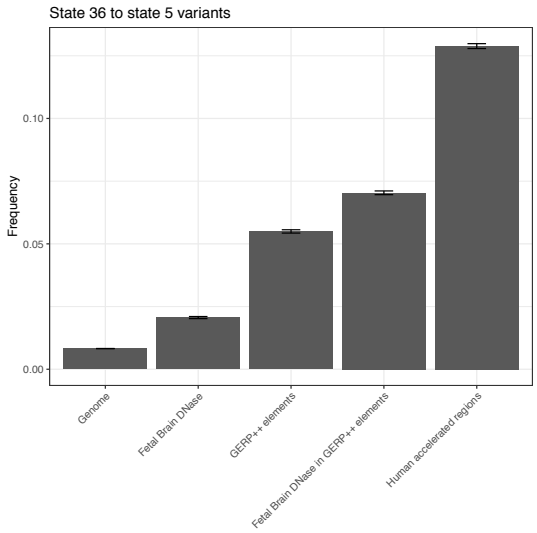


Figure 3.3 Characteristics of genetic variants leading to a change from state 36 to state 5 from the reference to alternate allele.

(a) Emission parameters of states 5 and 36 from a 100 state model based on a 100-way vertebrate alignment to the hg38 human genome. The heatmap is structured in the same way as the heatmap in **Figure 3.1**. **(b)** The 'Genome' category shows the frequency of observing the state assignment change to state 5 out of all possible alternate alleles for variants whose reference allele is in state 36. The rest of the columns shows the same frequency computed when restricting variants to those positioned in a Fetal Brain DNase peaks, GERP++ elements, the intersection of Fetal Brain DNase peaks and GERP++ elements and human accelerated regions. Error bars represent a 95% binomial confidence interval computed using a normal approximation of the error around the estimate.

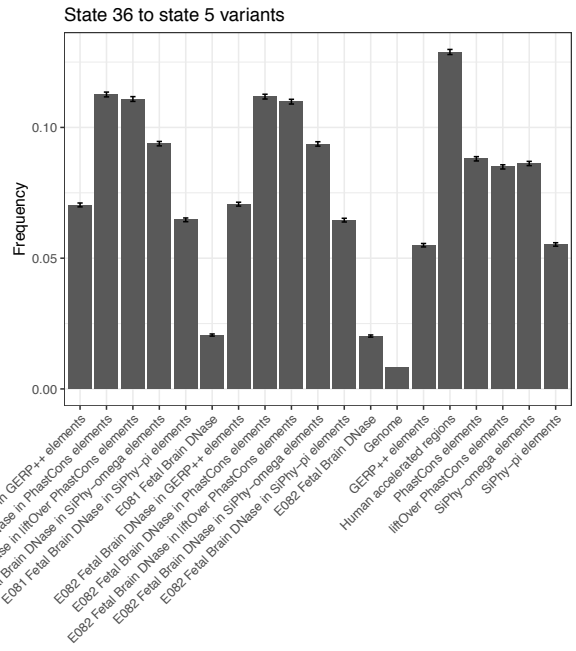


Figure 3.4 Extended characteristics of genetic variants leading to a change from state 36 to state 5 from the reference to alternate allele.

The frequencies in this plot were calculated analogously to the frequencies in **Figure 3.3**. Two Fetal Brain DNase I Hypersensitivity assays were used (Roadmap identifiers E081 and E082) and four different sets of conserved elements were used (GERP++, PhastCons, SiPhy-Pi, SiPhy-omega). Two sets of PhastCons elements were included: the elements called on the 100-way multiple sequence alignment of 99 vertebrates to the hg38 human genome, and the elements called on the 100-way multiple sequence alignment of 99 vertebrates to the hg19 human genome, which were lifted over to hg38.

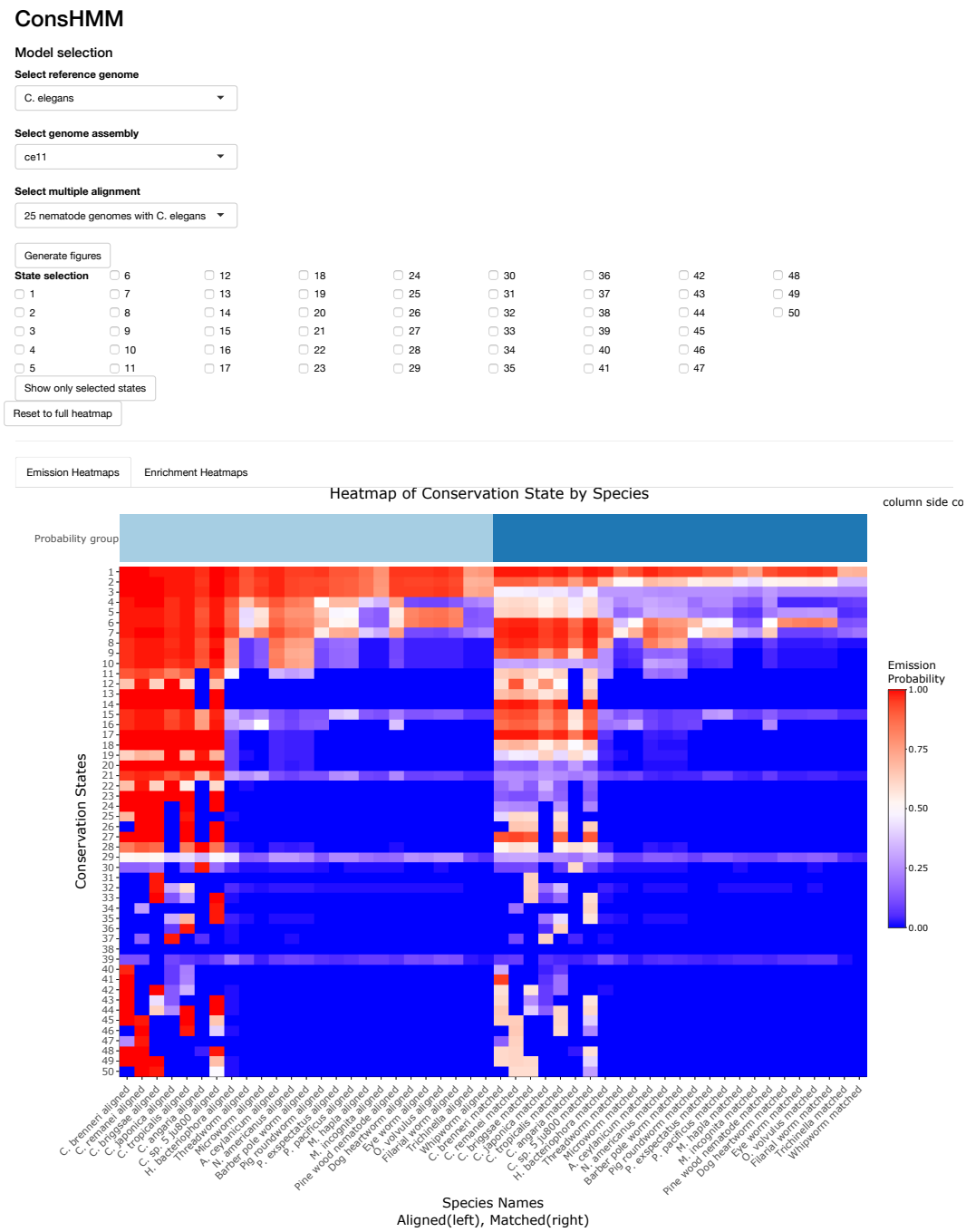


Figure 3.5 Screenshot of the ConsHMM R Shiny App.

The screenshot captures a representation of the emission probabilities of a 50 state model based on a 26-way alignment of nematodes with C. elegans. The dropdown menu at the top of the webpage allows users to select a different reference organism, genome and multiple

sequence alignment for which to generate the same heatmap. Each row in the heatmap corresponds to a state and each column corresponds to a species. The rows are sorted by ConsHMM hierarchical clustering, and the columns are sorted by phylogenetic distance to the reference genome in the alignment. The left half of the heatmap contains the probability of a species aligning the reference genome in the alignment. The right half of the heatmap contains the probability of a species matching the reference genome in the alignment. The checkboxes in the ‘state selection’ area of the app allow users to subset the heatmap to certain states of interest. The app also provides heatmaps summarizing important biological enrichments of each state.

Chapter 4. Design of a mammalian methylation array for cross-species epigenetic studies

4.1 Introduction

DNA methylation by the attachment of a methyl group to cytosines is one of the most widely studied epigenetic modifications in vertebrates, due to its implications in regulating gene expression across many biological processes[66,67].

The two most widely used technologies for obtaining DNA methylation levels are bisulfite sequencing[68] and microarray based methylation chips[69]. Whole genome bisulfite sequencing is an expensive assay, causing reduced representation bisulfite sequencing (RRBS) to become the prevalent sequencing approach. RRBS effectively queries only a small number of nucleotides on the genome but still provides a genome wide methylation profile. However, the sequencing depth required for robust RRBS measurements drives up costs. Due to this, for human samples, array chips containing an increasing number of probes have been the most reliable and widely used technology[70].

The first human methylation chip (Illumina Infinium 27K) was introduced over ten years ago[71] but no analogous chip has been presented for other species. This delay may reflect the fact that it was not economical to design a methylation chip for non-human species. Even if costs were no impediment, the development of species-specific arrays could hinder cross species comparisons as the measurement platforms would be different.

As mammalian genomes tend to have a lot of similarities, we sought to create an array that can measure DNA methylation in many mammalian species by leveraging cross-species sequence similarities. To accomplish this, we developed an algorithm, Conserved Methylation Array Probe Selector (CMAPS) that takes as input a set of probes that could target a human CpG site and a multiple sequence alignment of other species to the human genome, and

selects a set of probes sequences which can be used to query methylation in many mammals by the addition of degenerate bases. Using these sequences we have created a methylation array containing 38,000 probes, which has been used to profile DNA methylation in 8730 samples across 145 species. We show that the methylation signature of each sample contains information about both the species and tissue of origin and discuss a set of epigenetic questions which can be answered using CMAPS and the mammalian array.

4.2 Design

Mammalian array overview

Current methylation array technology produced by Illumina Inc. can contain two types of probes: Infinium I and Infinium II, with the latter being newer technology requiring only one silica bead to query the methylation of a CpG, while the former requires two beads. Two variations of each of the two types of probes can be designed for each CpG, depending on whether the probe is designed on the forward or reverse genomic strand, for a total of four total probe options for each CpG. The probes allow for up to three degenerate bases, which are positions that can be designed to tolerate variation in the sequence being interrogated. The number of degenerate bases tolerated is a function of a proprietary design score computed by Illumina, and the number of underlying CpGs in the case of Infinium 2 probes, since degenerate bases are used to account for possible differences in methylation status when multiple CpGs occur within an Infinium 2 probe.

Here we present the CMAPS algorithm, which uses the degenerate base technology to adapt probe sequences that can query human CpGs, so that the probes can now tolerate mutations and hybridize to DNA from other species as well. The CMAPS algorithm finds a specific set of degenerate bases for a human probe by analyzing a multiple sequence alignment of other genomes to the human genome. For the purposes of this project, CMAPS was applied

to the subset of 50 mammals within a 100-way alignment of 99 vertebrate genomes with human genome from the UCSC Genome Browser[64]. However, the algorithm can take as input any multiple sequence alignment with any reference genome and a parameter denoting how many degenerate bases can be introduced in each probe, and provide conserved probes and degenerate base selections.

For each CpG site, CMAPS selects those species in which the CpG is conserved and for which the difference between the sequence targeting the CpG in the human genome and the species' genome is within a number of mismatches that can be covered by degenerate bases. (**Figure 4.1a, Methods**). For each CpG site in the human genome we selected the Infinium 1 probe out of the two options (upstream or downstream of the CpG) that covered the most species based on the CMAPS algorithm, and analogously for Infinium 2. We first included all Infinium 2 probes that were targeting the mm10 mouse genome, such that the chip maximizes utility for one of the most widely used model organisms. We then sorted the CpG sites in descending order of the number of species covered with the Infinium 2 probe, and added an additional 17,000 probes that were not already selected due to targeting mm10, for a total of 53,000 probes. We then ranked the probes on the Illumina EPIC array in descending order of the number of species they can target using the degenerate bases picked by the CMAPS algorithm, and selected an additional 3,000 probes that had not already been picked based on the earlier criteria. Including probes on the EPIC array can allow testing of the array by comparing probe behavior on the mammalian versus EPIC array. Lastly, we sorted the CpG sites in descending order of number of species they can target and picked the top 4,000 Infinium 1 probes that targeted CpG sites that had not already been included. The Infinium 1 probes were selected to allow us to query CpG dense regions such as CpG islands, as the underlying CpG count of an Infinium 1 probe does not count against the number of SNVs permitted. This resulted in a set of 60,000 probes (**Figure 4.1b**).

A probe sequence targeting a certain CpG can map to multiple locations in a genome, which could result in a confounded signal coming from multiple CpG sites. This issue can be compounded by the fact that each of our probes can have up to 2^(# of degenerate bases) versions due to the degenerate base design. For 16 high quality genomes we computed for each probe how many of its versions map uniquely in that genome. We then filtered probes down by asking that all versions of a probe have to map uniquely in at least 80% of the species they were designed to target out of the 16 tested genomes, unless the probe targets at least 40 species, in which case the mapping criterion was discarded. This reduced the set of working probes to 35,988 probes. Two thousands additional probes were selected based on their utility for human biomarker studies. These CpGs, which were previously implemented in human Illumina Infinium arrays (EPIC, 450K, 27K), were selected due to their relevance for estimating age, blood cell counts, or the proportion of neurons in brain tissue.

4.3 Results

Properties of the custom chip

Not all probes on the array are expected to work for all species, but rather each probe is designed to cover a certain subset of species, such that overall all species have a high number of probes. Out of the 62 mammalian species considered by CMAPS from a multiple sequence alignment, 46 of them have CpSs that are targeted by more than 10,000 probes on the array, and 36 have CpGs that are targeted by more than 20,000 probes (**Table 4.1**).

Although the probes were selected based on sequence conservation criteria, we verified that the CpGs targeted by the final probe set are a representative set of the all CpGs in the human genome with respect to presence in CpG islands and average methylation level across tissues. We found that the distribution of CpG island density of islands containing a probe on the array contains less dense islands than the genome wide distribution, but that the shift from the

genome wide distribution is not large (**Figure 4.1c**). We also found that the probes on the array can target CpGs across a large range of fractional methylation levels., based on the analysis of fractional methylation called from whole genome bisulfite sequencing data across 37 tissues[9]. (**Figure 4.1d)**.

Calibration Studies

To confirm that the array is able to accurately profile methylation across different species, we created a set of DNA samples from human, mouse and rat which were engineered such that the fractional methylation at all CpG sites in their genomes is 0%, 25%, 50%, 75% and 100% (**Methods**). We profiled all these samples using the mammalian array. The distribution of the intensity of the probes in each human sample is centered around the known fractional methylation of the sample (**Figure 4.2a**). However, as expected, the distributions in the mouse and rat samples of all the probes show different patterns in these two species compared to the human samples, because many probes in the design of our array do not map to these genomes (**Figure 4.2b-c**). To confirm the accuracy of our design, for each species we removed the probes that do not map to that genome from the analysis, and normalized the array data using the SeSaMe package. After this procedure, the distribution of probe intensity in each sample becomes similar to those of the human samples, validating that the probes on the array behave as expected (**Figure 4.2d-f**).

DNA methylation signal encodes species and tissue type

DNA samples from 10 mammalian species (human, vervet monkey, olive baboon, mouse, horse, sheep, dog, pig, naked mole rat, killer whale), with more than 200 samples for each species, were processed using the mammalian array (**Table 4.2**). In this analysis, we focused our attention on the probes that can be mapped uniquely to the genomes of each of these

species (**Methods**). In a tSNE representation, we find that samples of the same species predominantly cluster together (**Figure 4.3a**), and that samples of the same tissue also tend to cluster together (**Figure 4.3b**). Interestingly, these two sources of variance seem to be orthogonal. For example, in one instance, mouse samples separate into several different clusters, each belonging to a different tissue. In another instance, blood samples separate into several different clusters, each belonging to a different species. We also find that a lot of the clusters exhibit some symmetry, which is explained by sex differences (**Figure 4.3c**).

4.4 Methods

Conserved Methylation Array Probe Selector (CMAPS)

The CMAPS algorithm was applied to the Multiz alignment of 99 vertebrates with the hg19 human genome downloaded from the UCSC Genome Browser.[24] For the purpose of this chip, only the 62 mammalian species in this alignment were considered. The design scores for each CpG in the human genome and each possible type of probe at each location were provided by Illumina and taken as input by CMAPS. For each CG site in the human genome, we computed the maximum number of species that could be targeted by each of the four different possible probe designs in human, considering each possible placing of the maximum number of tolerated mutations. For each probe option we tried all possibilities for placing the maximum number of potential variants, and greedily chose the allele that covers the most species at a particular position. More specifically, the algorithm for selecting the number of species covered by a probe is explain in pseudocode below:

The function get_max_species makes a greedy choice for the nucleotide at a certain SNV by picking whichever nucleotide is contained by the majority of non-human species in the alignment at that position.

```
function get_optimal_nucleotide(SNV_pos, multiple_sequence_alignment):
        max_species = 1
        for X in {A, C, T, G} \ {human nucleotide at SNV_pos}
        count_species = number of species with X at SNV_pos in the
multiple_sequence_alignment
                if count_species > max_species:
                        max_species = count_species
                        optimal_nucleotide = X
        return optimal_nucleotide
```

In the pseudocode below, SNV_set iterates over all possible positions of SNVs in a particular probe, given the design score and probe type constraints.

```
cur_max_species = 1
for SNV_set in all positions in probe:
        alt_nucleotide_list = []
        for SNV_pos in SNV_set:
                alt_nucleotide_list.append(get_max_species(SNV_pos,
multiple_sequence_alignment))
        num_species = number of species fully matching human given SNV_set and
alt_nucleotide_list
        if num_species > cur_max_species:
                cur_max_species = num_species
```

final_SNV_set = SNV_set

Since the get_max_species function makes greedy choices this may not be the true maximal subset of species for a probe, but this method is relatively computationally inexpensive and produced satisfactory species coverage for our purposes.

Mapping probes to genomic coordinates

We downloaded fasta sequence files for completely sequenced genomes from three public repositories (UCSC, Ensembl, and NCBI). We included only the most recent draft of each species' genome. In instances where the same genome draft for a species was available from multiple sources, we retrieved only one according via this preferential ordering: Ensembl over UCSC over NCBI. A complete list of all the genomes used, the public repository they were downloaded from, and the access date can be found in **Supplementary Data 3**.

After downloading these sequence files, where necessary, we concatenated multiple chromosome or sequence fragment files into a single fasta file for each species. We utilized the BSBolt software package from https://github.com/NuttyLogic/BSBolt to perform the alignments. For each species' genome sequence, BSBolt creates an 'in silico' bisulfite-treated version of the genome. As many of the currently available genomes are in a low quality assembly state (e.g. thousands of contigs or scaffolds), we used the utility "Threader" (which can be found in BSBolt's forebear BSseeker2[72] as a standalone executable) to reformat these fasta files into concatenated and padded pseudo-chromosomes.

The set of nucleotide sequences of the designed probes, which includes degenerate base positions, was explicitly expanded into a larger set of nucleotide sequence representing every possible combination of those degenerate bases. For Infinium I probes, which have both a methylated and an unmethylated version of the probe sequence, only the methylated version

was used as BSBolt's version of the genome treats all CG sites as methylated. Thus, the initial 37550 probe sequences resulted in a set of 184,352 sequences to be aligned against the various species genomes. We then ran BSBolt with parameters Align -M 0 –DB [path to bisulfite-treated genome] -BT2 bowtie2 -BT2-p 4 -BT2-k 8 -BT2-L 20 -F1 [Probe Sequence File] –O [Alignment Output File] –S to align the enlarged set of probe sequences to each prepared genome.

As we were not interested in the final BSBolt style output, we made a small modification to the code to retain its temporary output of alignment results in sam format. From these files, we collected only alignments where the entire length of the probe perfectly matched to the genome sequence (i.e. the CIGAR string '50M' and flag XM=0"). Then, for each genome we collapsed all the sequence variant alignments for each probeID down to a list of loci for that genome and for that probe. We report probes whose variants only mapped to one unique locus in a particular genome in **Supplementary Data 4**. We report the full list of loci for each probe, including the ones with unique mappings in **Supplementary Data 5**.

Dimensionality reduction of samples from 10 mammalian species

We normalized the probes within each sample using the SeSaMe R package[73]. We then performed Principal Component Analysis using the prcomp R function. We computed the variance explained by each principal component (PC), and found that the top 23 PCs amounts to 80% of the total variance in the data. We then peformed tSNE dimensionality reduction with default parameters from the Rtsne R package, substituting the 23 PCs as features.

4.5 Tables

Species	No. CpGs	Species	No. CpGs	Species	No. CpGs	Species	No. CpGs	Species	No. CpGs
Aardvark	20549	Cow	24817	Guinea pig	18931	Pika	16512	Weddell seal	25716
Alpaca	24455	Crab-eating macaque	32629	Hedgehog	14924	Platypus	4867	White rhinoceros	24888
Armadillo	19462	David's myotis bat	19441	Horse	23823	Prairie vole	18536		
Bactrian camel	23058	Dog	25305	Killer whale	24170	Rabbit	19492		
Big brown bat	20555	Dolphin	23396	Lesser Egyptian jerboa	16851	Rat	18440		
Black flying-fox	23546	Domestic goat	23913	Manatee	19960	Rhesus	31134		
Brush-tailed rat	19180	Elephant	19584	Marmoset	27075	Sheep	24652		
Bushbaby	23249	Ferret	25384	Megabat	21250	Shrew	16776		
Cape elephant shrew	18125	Gibbon	30196	Microbat	19984	Squirrel	24393		
Cape golden mole	18673	Golden hamster	18699	Mouse	22231	Squirrel monkey	28045		
Cat	25252	Gorilla	32157	Naked mole-rat	19856	Star-nosed mole	21577		
Chimp	32809	Green monkey	32375	Opossum	8160	Tasmanian devil	7962		
Chinchilla	21020	Green monkey	32189	Orangutan	30812	Tenrec	14521		
Chinese hamster	18615	Cow	24817	Pacific walrus	26570	Tibetan antelope	24011		
Chinese tree shrew	22903	Crab-eating macaque	32629	Pig	22880	Wallaby	6032		

Table 4.1 Expected number of CpGs targeted by probes on the mammalian array for 62 mammalian species.

The species in the table are the 61 mammalian species present in the 100-way Multiz alignment of 99 vertebrates to the human genome[64], which were considered by the CMAPS algorithm. A CpG is counted for a certain species if the choice of degenerate bases by the CMAPS algorithm lead to sequence identity for that species' genome.

Species common name	Species latin name	Number of samples
Human	Homo sapiens	1781
Mouse	Mus musculus	1425
Dog	Canis lupus familiaris	577
Sheep	Ovis aries	432
Olive baboon	Papio hamadryas	336
Horse	Equus caballus	336
Naked mole-rat	Heterocephalus glaber	261
Vervet monkey	Chlorocebus aethiops sabaeus	243
Pig	Sus scrofa domesticus	243
Killer whale	Orcinus orca	214

Table 4.2 Summary of data set processed using mammalian array for which more than 200 samples were available.

4.6 Figures

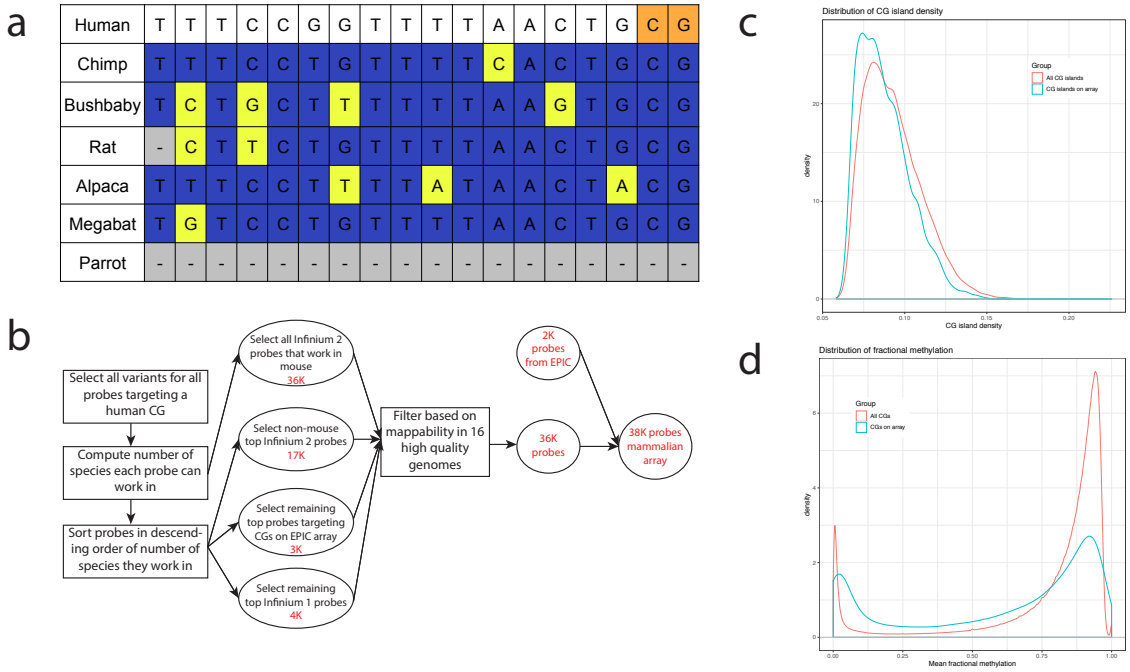


Figure 4.1 Overview of mammalian array design process and resulting distribution of genomic characteristics.

(a) Toy example of multiple sequence alignment at a CpG site considered by the CMAPS algorithm. The orange coloring highlights the CpG being targeted. Positions where other species have alignment that matches the human sequence are in dark blue; positions where other species have alignment that does not match the human sequence are in neon yellow; positions where other species have no alignment are in grey. **(b)** Flowchart detailing the selection of probes on the array by the CMAPS algorithm. **(c)** Distribution of CpG island density overlapping a probe on the mammalian array (blue) and all CpG islands in the human genome (red). **(d)** Distribution of average fractional methylation across 37 cell and tissue types at CpG sites on the array (blue) and all sites in the genome (red).

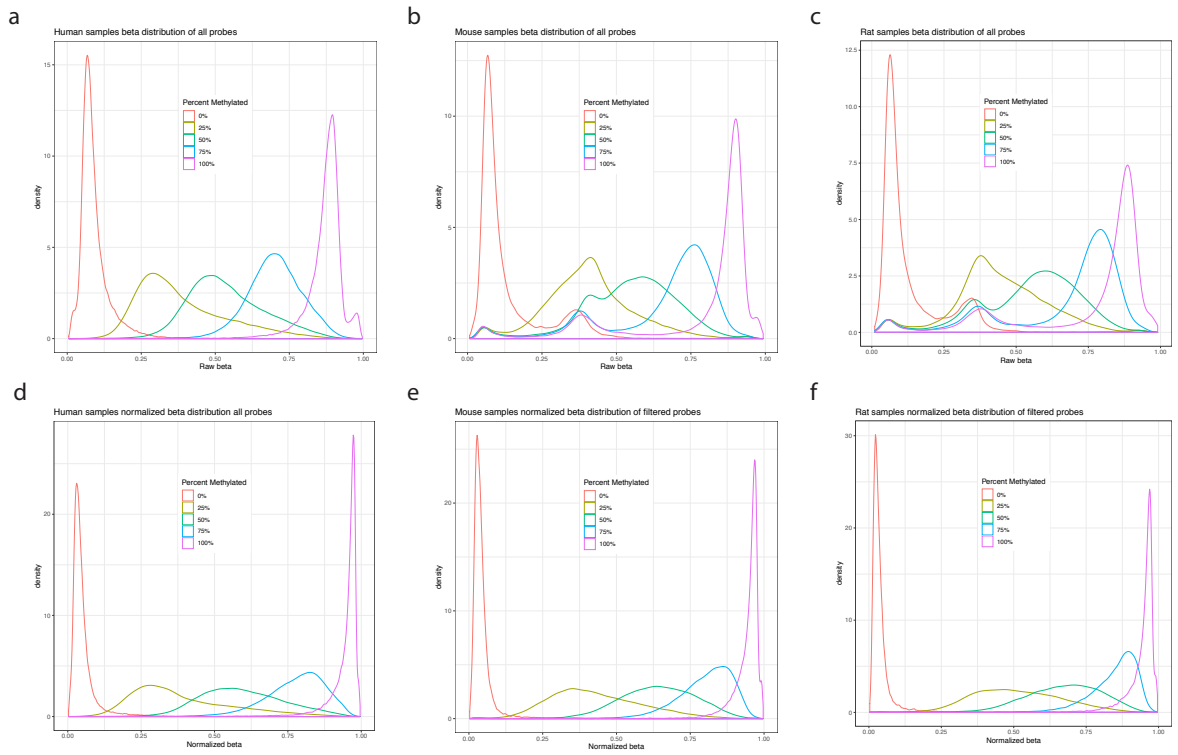


Figure 4.2 Distribution of probe intensities within sample, colored by the expected percentage of methylation at each site.

(a-c) Distribution of probe intensity of all probes on the array before normalization for **(a)** human samples, **(b)** mouse samples, and **(c)** rat samples. **(d-f)** Distribution of probe intensity after normalization and restricting probes to those that map to **(d)** the human genome in human samples, **(e)** the mouse genome in mouse samples, and **(f)** the rat genome in rat samples.

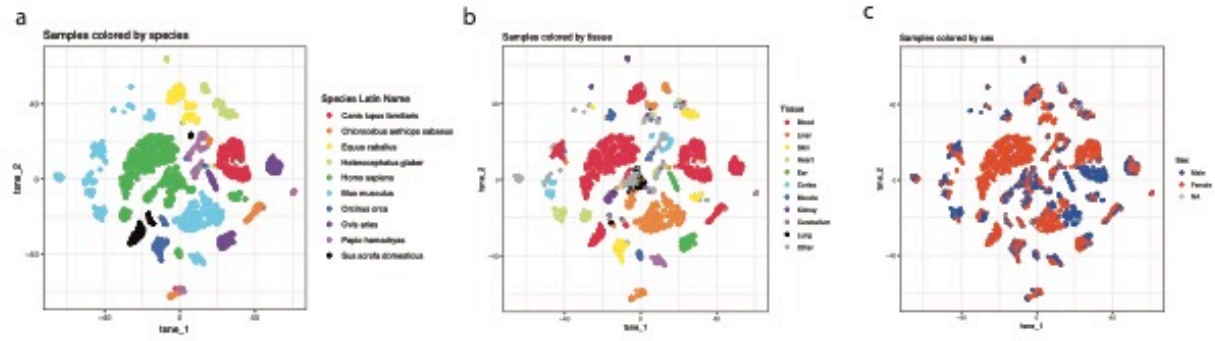


Figure 4.3 tSNE representation of samples profiled with the mammalian methylation array.

Each subpanel represents the same data points, colored by **(a)** species, **(b)** tissue and **(c)** sex.

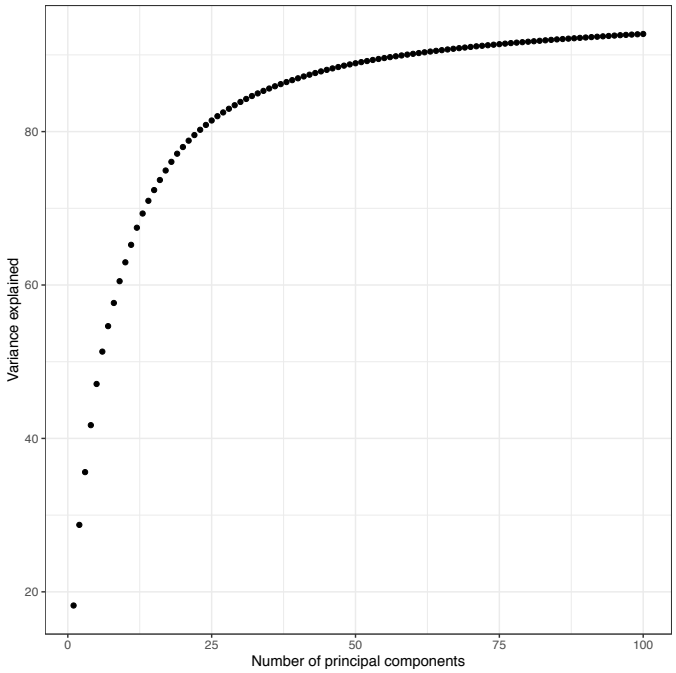


Figure 4.4 Cumulative variance explained by principal components.

The amount of variance explained rises sharply with the first components and then plateaus. 80% of the variance is explained by the top 23 components.

REFERENCES

1. Alföldi, J. & Lindblad-Toh, K. Comparative genomics as a tool to understand evolution and disease. *Genome Res.* **23**, 1063–1068 (2013).
2. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
3. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
4. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
5. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
6. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
7. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
8. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
9. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
10. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Mol. Cell* **71**, 882–895 (2018).

11. Leygo, C. *et al.* DNA Methylation as a Noninvasive Epigenetic Biomarker for the Detection of Cancer. *Disease Markers* https://www.hindawi.com/journals/dm/2017/3726595/ (2017) doi:https://doi.org/10.1155/2017/3726595.
12. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).
13. Ward, L. D. & Kellis, M. Interpreting non-coding variation in complex disease genetics. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
14. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
15. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111**, 6131–6138 (2014).
16. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).
17. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
18. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
19. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
20. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).

21. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
22. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
23. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
24. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670-681 (2015).
25. Lowe, C. B. *et al.* Three periods of regulatory innovation during vertebrate evolution. *Science* **333**, 1019–1024 (2011).
26. Siepel, A., Pollard, K. S. & Haussler, D. New Methods for Detecting Lineage-Specific Selection. in *Research in Computational Molecular Biology* 190–205 (Springer, Berlin, Heidelberg, 2006). doi:10.1007/11732990_17.
27. Kim, S. Y. & Pritchard, J. K. Adaptive Evolution of Conserved Noncoding Elements in Mammals. *PLOS Genet.* **3**, e147 (2007).
28. Marnetto, D. *et al.* Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion. *Am. J. Hum. Genet.* **102**, 1–12 (2018).
29. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database J. Biol. Databases Curation* **2016**, (2016).
30. Cotney, J. *et al.* The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. *Cell* **154**, 185–196 (2013).

31. Villar, D. *et al.* Enhancer Evolution across 20 Mammalian Species. *Cell* **160**, 554–566 (2015).
32. Don, P. K., Ananda, G., Chiaromonte, F. & Makova, K. D. Segmenting the human genome based on states of neutral genetic divergence. *Proc. Natl. Acad. Sci.* **110**, 14699–14704 (2013).
33. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
34. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
35. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
36. Blanchette, M. *et al.* Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.* **14**, 708–715 (2004).
37. Chen, X. & Tompa, M. Comparative assessment of methods for aligning multiple genome sequences. *Nat. Biotechnol.* **28**, 567–572 (2010).
38. Zhang, M. Q. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**, 919–932 (1998).
39. Sarda, S., Das, A., Vinson, C. & Hannenhalli, S. Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal promoters. *Genome Res.* **27**, 553–566 (2017).
40. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).

41. Litman, G. W., Anderson, M. K. & Rast, and J. P. Evolution of Antigen Binding Receptors. *Annu. Rev. Immunol.* **17**, 109–147 (1999).

42. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-4.0*. (2013).

43. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33**, 364–376 (2015).

44. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).

45. Ioannidis, N. M. *et al.* FIRE: functional inference of genetic variants that regulate gene expression. *Bioinforma. Oxf. Engl.* **33**, 3895–3901 (2017).

46. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinforma. Oxf. Engl.* **31**, 761–763 (2015).

47. Rogers, M. F. *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinforma. Oxf. Engl.* **34**, 511–513 (2018).

48. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).

49. Smedley, D. *et al.* A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. *Am. J. Hum. Genet.* **99**, 595–606 (2016).

50. Iulio, J. di *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333 (2018).

51. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-1006 (2014).
52. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. *Mol. Biol. Evol.* **30**, 1159–1171 (2013).
53. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
54. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
55. Witowski, V. & Foraita, D. R. *HMMpa: Analysing accelerometer data using hidden Markov models*. (2014).
56. Hahsler, C. B. and M. *cba: Clustering for Business Analytics*. (2017).
57. Bar-Joseph, Z., Gifford, D. K. & Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**, S22–S29 (2001).
58. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**, 191 (2006).
59. Kolde, R. *pheatmap: Pretty Heatmaps*. (2015).
60. Arneson, A. & Ernst, J. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun. Biol.* **2**, 1–14 (2019).
61. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinforma. Oxf. Engl.* **25**, i54-62 (2009).

62. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

63. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database J. Biol. Databases Curation* **2016**, (2016).

64. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).

65. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

66. Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714–717 (2007).

67. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

68. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).

69. Schumacher, A. *et al.* Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res.* **34**, 528–542 (2006).

70. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208 (2016).

71. Bibikova, M. *et al.* Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* **1**, 177–200 (2009).

72. Guo, W. *et al.* BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* **14**, 774 (2013).

73. Zhou, W., Triche, T. J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* **46**, e123–e123 (2018).