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

Adriana Sperlea[1,2] and Jason Ernst[1,2,3,4,5,6]

[1] Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, California, USA.

[2] Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, California, USA.

[3] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of California, Los Angeles, Los Angeles, California, USA.

[4] Computer Science Department, University of California, Los Angeles, Los Angeles, California, USA.

[5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California, USA.

[6] Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California, USA.

Correspondence: J.E. (jason.ernst@ucla.edu)

Abstract

Comparative genomics sequence data is an important source of information for interpreting genomes. Genome-wide annotations based on this data have largely focused on univariate scores or binary calls of evolutionary constraint. Here we present a complementary whole genome annotation approach, ConsHMM, which applies a multivariate hidden Markov model to learn *de novo* different '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. We applied ConsHMM to a 100-way vertebrate sequence alignment to annotate the human genome at single nucleotide resolution into 100 different conservation states. These states have distinct enrichments for other genomic information including gene annotations, chromatin states, and phenotype associated genetic variation, supporting their biological significance. The ConsHMM method and conservation state annotations provide a resource for interpreting whole genomes and potential disease-associated variants.

Introduction

The large majority of phenotype-associated variants implicated by genome-wide association studies (GWAS) fall outside of protein coding regions (Hindorff et al. 2009). Identifying the causal variants and interpreting their biological role in these less well understood non-coding regions is a significant challenge (Ward and Kellis 2012). Large scale mapping of epigenomic data across different cell and tissue types has been one approach for annotating and interpreting the non-coding regions of the genome (ENCODE Project Consortium 2012; Ernst et al. 2011; Roadmap Epigenomics Consortium et al. 2015). Using comparative genomics data to identify regions of evolutionary constraint has been a complementary approach for these purposes (Kellis et al. 2014; Lindblad-Toh et al. 2011; Pollard et al. 2010; Siepel et al. 2005).

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 (Cooper and Shendure 2011; Weedon et al. 2014). 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 (Cooper et al. 2005; Davydov et al. 2010; Garber et al. 2009; Pollard et al. 2010; Siepel et al. 2005). 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 (Finucane et al. 2015). These scores and elements have also been highly informative features to integrative methods for prioritizing pathogenic variants (Kircher et al. 2014; Ritchie et al. 2014; Ionita-Laza et al. 2016; Huang et al. 2017). Further improvements for predicting pathogenic variants in coding regions have been made to the integrative scores by incorporating features defined directly from a multiple sequence alignment (Jagadeesh et al. 2016).

While highly 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 sequenced and incorporated into multi-species alignments such as a 100-way alignment to the human genome (Rosenbloom et al. 2015). Approaches have been developed to associate constrained elements, regions, or individual bases with specific branches in a phylogenetic tree (Lowe et al. 2011; Siepel et al. 2006; Kim and Pritchard 2007; Marnetto et al. 2018; Herrero et al. 2016; Cotney et al. 2013; Villar et al. 2015). While also useful, such directed approaches are biased to only representing certain types of patterns present in the alignment. An alternative approach used for comparative genomic based annotation learned patterns of different classes of mutations between human and orangutan (Don et al. 2013), but this approach was only applicable at a broad region level and only incorporated information from one non-human genome.

Analogous to the many sequenced genomes available for comparative analysis, many different datasets are available for annotating the genome based on epigenomic data. 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 (Ernst and Kellis 2010, 2012, 20; Ernst et al. 2011; Hoffman et al. 2012). Inspired

by the success of these approaches, here we develop a new method, ConsHMM, that extends ChromHMM (Ernst and Kellis 2012) to systematically annotate genomes into 'conservation states' at single nucleotide resolution. The conservation states assignments are based on the combinatorial and spatial patterns of which species align to and which match a reference genome at each nucleotide in a multiple species DNA sequence alignment. ConsHMM takes a relatively unbiased modeling approach that does not explicitly assume a specific phylogenetic relationship between species. The set of conservation patterns ConsHMM can infer are thus flexible and determined directly from the DNA sequence alignment.

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

Results

Annotating the human genome into conservation states

We developed a novel approach, ConsHMM, to annotate a genome into different conservation states based on a multiple species DNA sequence alignment (**Fig. 1a, Methods**). We model the combinatorial patterns within the alignment of which species align to and which match a reference genome, for which we used the human genome. Specifically, at each nucleotide in the human genome we encode one of three possible observations for each other species in the alignment: (1) aligns with a nucleotide present that is the same as the human reference genome, (2) aligns with a nucleotide present that is different than the human reference genome, or (3) does not have a nucleotide present in the alignment for that position. We further model these observations as being generated from a multivariate hidden Markov model (HMM), which probabilistically captures both the combinatorial patterns in the observations and their spatial context. Specifically, we assume that in each state the probability of observing a specific combination of observations is determined by a product of independent multinomial random variables. The parameters to the distributions of these multinomial random variables will differ between states and are learned from the data. After the model is learned, each nucleotide in the human genome is assigned to the state that had the maximum posterior probability of generating the observations.

ConsHMM builds on ChromHMM (Ernst and Kellis 2012), which has previously been applied to annotate genomes based on epigenomic data at 200-bp resolution (Ernst and Kellis 2010), to now annotate genomes at single nucleotide resolution based on a multiple species DNA sequence alignment (**Methods**). We applied ConsHMM to a 100-way Multiz vertebrate alignment with the human genome and focused our analysis here on a model learned using 100 states in order to balance recovery of additional biological features and model tractability (**Fig. 2**, **Supplementary Fig. 1-7, Supplementary Tables 1,2, Methods**). We illustrate the ConsHMM conservation state annotations at two different loci showing that different bases that are associated with calls of evolutionary constraint from existing approaches can have very

different underlying alignment patterns and conservation state assignments (**Fig. 1b, Supplementary Fig. 8**). Conservation state annotations genome-wide are available at https://github.com/ernstlab/ConsHMM.

Major groups of conservation states

We hierarchically clustered the conservation states based on their align and match probabilities, and then cut the resulting dendrogram to reveal eight notable groups of states or distinct individual states (**Fig. 2a**, **Supplementary Fig. 9, Supplementary Table 3, Methods**). The first of these subsets of states was a single state (State 1; AM_allVert) that showed high align and match probabilities through essentially all vertebrate species considered. The second subset showed relatively high align and match probabilities for all mammals and some non-mammalian vertebrates (States 2-4; AM_nonMam). The third subset showed relatively high align and match probabilities for most if not all mammals, but not non-mammalian vertebrates (States 5-22; AM_Mam). The fourth subset showed high align probabilities for many mammalian species, but had low align probabilities for notable species such as mouse and rat for many of the states in the group (States 23-46; AM_SMam). The combination of the absence of mouse and rat alignments with the presence of mammals that are assumed to have diverged earlier is consistent with the previously observed increased substitution rates for mouse and rat(Lindblad-Toh et al. 2011). The fifth subset showed high align probabilities for many mammalian species, but did not show high match probabilities (States 47-63; A_SMam). The sixth subset showed high align probabilities for most primates, but not for other species (States 64-89; AM_Prim). The seventh subset showed high align probabilities for at most a subset of primates (States 90-99; AM_SPrim). The final subset was a single state (State 100; artifact) that showed a noteworthy pattern of high align and match probabilities for most primates and non-mammalian vertebrates, but low probabilities for non-primate mammals, consistent with a previous observation that inclusion of non-mammalian vertebrates can be associated with increased presence of suspiciously aligned regions (Chen and Tompa 2010).

Conservation states exhibit distinct patterns of positional enrichments relative to gene annotations and regulatory motif instances

The conservation states showed strong and distinct positional enrichments relative to GENCODE annotated gene features including transcription start sites (TSS), transcription end sites (TES), exon start sites, and exon end sites for both protein coding genes and pseudogenes (**Fig. 3a-d**, **Supplementary Fig. 10**). Notable positional enrichments were also seen for regulatory motifs instances (**Fig. 3e,f**). Relative to starts of exons of protein coding genes seven of the states (States 1-4, 7, 28, and 54) had 13 fold or greater enrichment for some position within 20 base pairs of exon starts, both when considering all such exons and subsets of exons in specific coding phases (**Fig. 3a, Supplementary Fig. 10a-c**). All of these states showed alignability for some non-mammalian vertebrates in addition to most mammals. Within exons we saw the strongest enrichment for States 1-4 and 54, and among these state 1 showed the strongest enrichment as expected given its high match probabilities through all vertebrates (**Figs. 2b, 3a-b, Supplementary Fig. 10a-e**). Interestingly, state 1 showed very strong enrichment (>80 fold) in the two nucleotides immediately upstream of the exon start with the third upstream nucleotide also having high enrichment (46 fold) (**Supplementary Fig. 10c**).

These three nucleotide positions correspond to the positions of the canonical 3' splice site sequence that is highly conserved throughout vertebrates (Zhang 1998). At the ends of exons of protein coding genes (**Fig. 3b**), state 1 maintained a >40 fold enrichment for six nucleotides past the end of coding sequence corresponding to positions of the known canonical 5' splice site sequence (Zhang 1998). 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 each 3rd base corresponding to codon wobble positions. In contrast, states 3 and 54, which were both associated with high align probabilities through many vertebrates and lower match probabilities, showed the inverse oscillation pattern to state 1 (**Fig. 3a, Supplementary Fig. 10a-c**).

Relative to TSS of protein coding genes, state 28 had the strongest enrichment reaching a maximum enrichment 30 fold at the TSS (**Fig. 3c**). State 28 was associated with moderate align and match probabilities for almost all the species present in the alignment. Consistent with its enrichment for TSSs state 28 also had the greatest enrichment for CpG islands (32 fold). However, state 28 also showed a 20 fold enrichment of CpG islands >2kb away from any TSS of protein coding genes and a 10 fold enrichment for TSS of protein coding genes >2kb away from a CpG island, suggesting the possibility that both of these features are making a partially independent contribution to the association, or the presence of additional unannotated TSSs that are associated with CpG islands(Sarda et al. 2017). Relative to TES of protein coding gene we saw the enrichment peak for state 2 at almost 12 fold (**Supplementary Fig. 10f**), which had high align and match probabilities for almost all vertebrates except for fish.

Relative to pseudogene exon starts and ends, states 100 and 82, both associated with alignability to distal vertebrates without many mammals closer to human, had strong enrichments peaking at greater than 100 and 38 fold respectively (**Supplementary Fig. 10g,h**). These two states also showed the greatest enrichment relative to TSSs of pseudogenes peaking at 184 and 68 fold for states 100 or 82 respectively (**Fig. 3d**) and for TESs of pseudogenes peaking at 199 and 61 fold respectively (**Supplementary Fig. 10i**).

Relative to instances of regulatory motifs different conservation states showed single nucleotide enrichment variation, often associated with variation in the amount of information in the positional-weight matrix (**Fig. 3e-f** and **Methods**) (Kheradpour and Kellis 2014). For example, in the case of the POU5F1 and STAT motifs we saw state 2 from the AM_nonMam group and state 5 from the AM_Mam group respectively reach 1.8 fold enrichments but have lower enrichments (1.4-1.5) at some nucleotides with lower information content. For the STAT motif, states 55-57, associated with high align probabilities for most mammals, but high match probabilities only for a few primates, enrichments peaked at the CG dinucleotide in the center of the motif consistent with their genome-wide enrichments for CG dinucleotides (**Fig. 3e, Supplementary Fig. 11**).

Enrichment of conservation states for different gene classes

The previous analyses demonstrated that different conservation states have distinct enrichments in promoter regions of genes. We next investigated whether different conservation states also exhibit distinct enrichments for different classes of genes after controlling for the state's relative preference for promoter regions. Specifically, for each state we determined the 5% of genes with the greatest presence of the state in its promoter region and evaluated Gene Ontology (GO) enrichments for those genes, revealing distinct enrichment patterns (**Fig. 4b,**

Supplementary Fig. 12, Methods). For example, even among states 1-3, all of which had high alignability through at least birds and matching through mammals, we observed substantial differences in their gene preferences. Out of these three states, state 1 (the AM_allVert group) was the only one enriched for nucleosomes ($p<10^{-41}$; 10.5 fold), while state 3, which had high matching only through mammals, was the only one with a significant enrichment for a set of genes related to sensory perception of smell ($p<10^{-300}$; 15.5 fold). State 2, which had high align and match probabilities through all vertebrates except fish, was the state most enriched for cellular developmental processes ($p<10^{-30}$; 1.8 fold), which did not show enrichment in state 3. We also observed notable enrichments for states with overall lower align or match probabilities. For example, state 89, associated with high alignability and low matching in primates as well as some alignability and low matching in non-primate mammals, was the state most enriched for antigen binding ($p<10^{-14}$; 6.7 fold). This is consistent with antigen binding being associated with many species, but fast evolving (Litman et al. 1999).

Enrichments for repeat elements in conservation states

The conservation states showed a wide range of enrichments and depletions (from 2 fold enrichment to 133 fold depletion) for bases overlapping any repeat element (**Fig. 2b, Fig. 4d, Supplementary Fig. 13**). Different states had distinct patterns of enrichments for different repeat classes and families, even though in some cases the difference in state parameters was subtle. For instance, among states in the AM_Prim group, which primarily differed in terms of the specific combinations of primates with high align and match probabilities, we found distinct enrichments. Notably, four different states from the group AM_Prim, 74, 86, 76, and 77, showed maximal enrichments for the DNA, LINE, LTR, and SINE repeat classes respectively (**Fig. 4d**). State 74, which is characterized by high align and match probabilities for all primates, had an enrichment of 5.6 fold for DNA repeats, while the enrichment for the other three classes were between 1.0 and 1.8 fold. On the other hand, state 86, which lacked alignability of a subset of primates, had a 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 3.3 and 4.5 fold enrichments for LTR and SINE respectively compared to 1.1 and 2.1 fold for SINE and LTR respectively. State 76 and state 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 the alignment probabilities, these states had substantial differences in their repeat enrichment profiles.

Relationship of conservation states to chromatin states

We compared our conservation states to annotations of the genome based on a 25-chromatin state imputation based model defined on 127 samples of diverse cell and tissue types (Ernst and Kellis 2015; Roadmap Epigenomics Consortium et al. 2015) (**Fig**. **4a, Supplementary Fig. 14**). For each conservation state we determined the median enrichment of each chromatin state across the 127 samples. Eleven different conservations states were maximally enriched for at least one of the 25-chromatin states. Conservation state 28 showed the greatest enrichment for any chromatin state, with a 35 fold enrichment for a chromatin state associated with active promoters, and was maximally enriched for four other promoter associated states. Conservation state 1 was maximally enriched for five chromatin states all

associated with transcribed and exonic regions (Ernst and Kellis 2015) (3.8-8.7 fold), which is consistent with this conservation state being most enriched for exons. Conservation state 2 had the maximal enrichment for five enhancer associated states (3.1-4.7 fold), while conservation state 5 high enrichments for these states and also had the greatest enrichment of any conservation state for a state primarily associated with just DNase I hypersensitivity (2.5 fold). These and other distinct enrichments of the conservation states for the different chromatin states highlight that conservation states are able to capture multi-dimensional information in the genome.

Conservation states capture enrichment patterns of DNase I hypersensitive sites across cell and tissue types

The previous analysis demonstrated that conservation states can exhibit different enrichment patterns for different chromatin states. We next investigated whether different conservation states also capture distinct enrichment patterns for a chromatin mark across cell and tissue types. For this we analyzed DNase I hypersensitive sites (DHS) from 53 of the 127 samples considered above for which maps of experimentally observed DHS were available (Roadmap Epigenomics Consortium et al. 2015). We focused on the 21 conservation states that exhibited at least 2 fold enrichment in at least one sample (**Fig. 4c**). We then row normalized the enrichments in order to focus on the relative enrichment patterns across cell and tissue types (**Methods**). Hierarchical clustering of the enrichment patterns revealed two major clusters of states (**Fig. 4c**). One of these clusters contained 14 of the 21 states and was associated with strong enrichments for fetal related samples. Ten of the states in this cluster have maximum enrichment for a fetal sample, while the remaining four state have maximum enrichment for the cell type Human Umbilical Vein Cells (HUVEC). The second major cluster consisted of seven states, all of which were enriched for CpG islands (**Fig. 2b**, **Supplementary Fig. 4**). The samples that showed the greatest enrichment in states in these clusters also had the greatest enrichment of CpG islands (**Fig. 4c**), but were biologically diverse in terms of the type of cell or tissue and could potentially reflect technical experimental differences.

Relationship of conservation states to constraint based annotations

We next investigated the relationship of our conservation state annotations with calls and univariate scores of evolutionary constraint. Specifically, we considered constrained element sets based on four methods (GERP++, SiPhy-omega, SiPhy-pi, and PhastCons) and constraint scores based on three methods (GERP++, PhastCons, and PhyloP) publicly available for hg19 and also defined on Multiz alignments (**Fig. 2b**, **Supplementary Fig. 15**). The PhastCons and PhyloP scores and elements we compared to were defined on the same 100-way vertebrate alignment. The available GERP++, SiPhy-omega, and SiPhy-pi score and elements were derived from different versions of Multiz alignments and only considered mammals.

We consistently found conservation states 1-5 to be highly enriched (>9 fold) for all constrained element sets (**Fig. 2b**, **Supplementary Fig. 15a**). These states were also among the top six states in terms of mean score for constraint scores considered (**Supplementary Fig. 15b**). Consistent with this, states 1-5 were the states that had the highest average matching probability across mammals. Two other states exhibited at least 6 fold enrichment for at least

one constrained element set: states 54 and 100. State 100, associated with putative artifacts, showed high enrichments for PhastCons elements (15 fold) and high average scores for PhastCons and PhyloP. This is consistent with this state having high aligning and matching probabilities primarily in non-mammalian vertebrates and these elements and scores being defined using such species. State 54 was consistently enriched for all the constrained elements (4-7 fold), but did not show high mean base-wise scores. The difference of high enrichment in constrained elements but not base-wise scores is consistent with state 54 having high alignability through most vertebrates, but low matching outside primates. More generally, we found that constrained element calls did not have the resolution to exhibit biologically relevant single nucleotide variation in enrichments around regulatory motifs and exon start and ends as we saw with our conservation state annotations, with the exception of those from PhastCons (**Fig. 3, Supplementary Fig. 16**).

The objective of our conservation state annotations is different than that of binary calls and univariate scores of evolutionary constraint, which have a more specific and complementary goal. However, to better understand their relative biologically relevant information we compared their ability to recover annotated starts and ends of exons and TSS and TES of genes separately for protein coding and pseudogenes (**Fig. 5a-c**, **Supplementary Fig. 17**). In almost all cases the conservation states had greater information available for recovering annotated gene features. The only exceptions were that PhyloP scores could achieve 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 compared the ability of conservation states to recover bases covered by DHS sites in specific cell types both genome-wide and restricted to non-exonic bases (**Supplementary Fig. 18**). For these analyses we generally found that at the same recall level the conservation states could identify bases in a DHS at greater precision than constraint scores. Compared to constrained elements the relative ability depended both on the specific constrained element set being compared and the specific DHS experiment. Importantly, the information about DHS in the conservation states was complementary to that in the constrained element sets as evidenced by large variation in DHS enrichments of bases within constrained elements depending on their conservation state (**Fig. 5d, Supplementary Fig. 19**). For example, the enrichment of bases in PhastCons constrained elements for Fetal Brain DHS in non-exonic regions ranged from 0.3 to 10.1 fold depending on the conservation state. Additionally, we saw cases where certain states had greater enrichments for DHS for their bases not in a constrained element compared to bases in a constrained element in other states. On the other hand, constrained element calls offered additional information as we observed that in most cases, for a given conservation state, bases that were in a constrained element call had greater enrichment for DHS than those that were not.

We also analyzed the enrichments of our conservation states for previously defined nine-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 (**Supplementary Fig. 20a**) (Lowe et al. 2011). This demonstrated in some cases the heterogeneous nature of the resulting assignments when relying on directed phylogenetic partitioning approaches. For example bases in elements assigned to originating at the branch point of the Tetrapod clade showed a 37 fold enrichment for state 2, as would be expected since

state 2 is associated with aligning and matching through all vertebrates except fish, but an even greater enrichment (51 fold) for state 100, associated with putative artifacts. A similar pattern of enrichments was observed when considering only the CNEEs overlapping a PhastCons element called on the same alignment as the conservation states (**Supplementary Fig. 20b**).

Enrichment of conservation states for human genetic variation

Previous analyses have found a depletion of human genetic variation in evolutionarily constrained elements (Lindblad-Toh et al. 2011). Consistent with that, the greatest depletion (3.3 fold depletion) of common SNPs is in state 1, the state most enriched for constrained elements. Interestingly, six states, A_SMam states 55-57 and AM_Prim states 87-89, had enrichments in the range 5 to 8 fold for common SNPs. These were also the six states with greatest enrichment of CG dinucleotides (**Supplementary Fig. 11**). These six states have in common that they show high align probabilities for most primates, but low match probabilities for some of those same primates. These states are thus associated with substantial variation both among primates and among humans.

When analyzing the enrichment of GWAS catalog variants (Welter et al. 2014) relative to the background of common SNPs we saw opposite enrichment patterns for these states (**Fig. 6a,b**). For example, relative to this background, state 1 was most enriched for GWAS catalog variants, which is consistent with previous observations of constrained elements enriching for GWAS variants (Lindblad-Toh et al. 2011). On the other hand, states 55-57 and 87-89 showed the greatest depletion. These results suggest that common variants are less likely to be phenotypically significant if they fall in conservation states most enriched for common genetic variation.

Constrained element enrichment for partitioned heritability of complex traits depends on conservation state

Previous analyses have suggested a strong enrichment of constrained elements for phenotype heritability (Finucane et al. 2015). As we saw large differences in DHS enrichments of constrained elements depending on the conservation state, we investigated the extent to which constrained elements in conservation states most enriched for DHS enriched for phenotype heritability compared to the remaining states. Specifically, we ranked the conservation states in descending order of their median enrichment for DHS from a compendium of 125 experiments within the non-exonic portion of the state (**Fig. 2b, Methods**). We then partitioned bases in PhastCons constrained elements into two almost equal size sets based on whether they overlapped one of the top seven ranked conservation states (states 1-5, 8, 28) or not. We then computed the heritability for these two sets for eight phenotypes in the context of a set of baseline annotations (Finucane et al. 2015) (**Methods**). For seven of the phenotypes, we found that bases in constrained element overlapping the top seven states had greater enrichment than those in the remaining 93 states, often substantially so (**Fig. 6c**). These results suggest additional value in the conservation state annotations for isolating more likely disease relevant variants.

Discussion

We presented a new framework for genome annotation based on comparative genomics sequence data. Our approach learns a set of 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. We applied this approach to annotate the human genome at single nucleotide resolution into one of 100 conservation states.

Conservation state annotations exhibited substantial enrichments for a wide range of other genomic annotations, thus supporting their biological significance. Specific conservation states exhibited strong enrichments for various gene annotations including exons and TSS and TES of genes. Conservation states showed differential enrichment patterns for various classes of genes and DHS from multiple cell types, even though they were defined independently of any functional genomics data. Specific conservation states exhibited enrichments for common human variants, while a different set of states exhibited enrichments for variants identified by GWAS relative to common variation.

Our modeling approach does not explicitly use a specific phylogenetic tree, except to the extent to which it was used to generate the input multi-species sequence alignment. This leads to relatively unbiased, simple and interpretable models. However, many state patterns discovered are consistent with expected observations from commonly assumed phylogenetic relationships of the species. While states' parameters often decreased with divergence time from human, there were a number of exceptions. Some of these exceptions corresponded to missing specific sub-clades of species, particularly those with long branch lengths. For example, a number of states were not represented by mouse and rat while being represented by more distally diverged mammals. Other exceptions isolated putative artifacts in the alignments that might otherwise confound analyses, as we saw for two states heavily enriched for pseudogenes.

Our conservation state annotation is complementary to existing binary calls and scores of evolutionary constraint based on phylogenetic modeling. Both locations called as constrained and those called as non-constrained are heterogeneous in their assigned conservation state. Our annotations thus provide additional descriptive information about the conservation patterns at each base. In terms of information for predicting external annotations, we found that in many cases the conservation states had greater information than constraint scores or elements. For other cases, such as DHS, the relative information depended on the constrained element set or score being compared. Importantly, we observed that DHS information provided by the states was complementary to information in the constrained element calls. Additionally, we observed that bases in constrained elements showed substantially different enrichments for phenotype-associated heritability, depending on their conservation state.

The conservation states are both inspired by, and provide complementary information to, existing chromatin state annotation approaches. While the states from the two approaches are based on very different data and have fundamental differences, they also exhibited substantial cross-enrichments. In general, conservation states have the advantage of providing information at single nucleotide resolution, which we demonstrated by showing enrichments patterns in and around coding exons and regulatory motifs. Conservation states can also provide information

about bases in the genome even if the relevant cell type has not been experimentally profiled, while chromatin states have the advantage of directly providing cell type specific information.

We expect many applications for the methodology and annotations we have presented here. While we applied ConsHMM here to one multiple species alignment, a 100-way Multiz human alignment, the methodology is general, and thus can be readily applied to alignments to other species or alignments generated by other methods (Herrero et al. 2016). The annotations we produced serve as a resource to directly interpret other genomic datasets. They can also be integrated with complementary functional genomics or comparative information in methods that aim to better prioritize disease relevant variants (Huang et al. 2017; Ionita-Laza et al. 2016; Kichaev et al. 2014; Kircher et al. 2014; Pickrell 2014; Ritchie et al. 2014). This work represents a step in the direction of improving whole genome annotations, which will continue to be of increasing importance towards understanding health and disease as the availability of whole genome sequencing data increases.

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 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 species j at position i. The three possible observations are: (1) the other species aligns with a non-indel nucleotide symbol present matching the reference nucleotide, (2) the other species aligns with a non-indel nucleotide symbol present, but does not match the reference nucleotide, or (3) the other 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 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), 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 (Ernst and Kellis 2012) software originally designed to solve an analogous problem of annotating the 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 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 $P(z_{i,j} = 1| y_{i,j}=1 \text{ and } s_i=k)$. 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

Our method and software 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. The alignment was obtained from the UCSC Genome Browser at:
http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/ (Blanchette et al. 2004; Rosenbloom et al. 2015).

Scaling-up ConsHMM to single base resolution with hundreds of features

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 a random subset 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 sequences are sampled. These new options were included in version 1.13 of ChromHMM.

Generating genome-wide annotations

After learning a 100-state model, we used it to segment and annotate the genome at base-pair resolution into one of the 100 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 (Smit et al. 2013; Rosenbloom et al. 2015). 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 (https://www.gencodegenes.org/releases/19.html). 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. Chromatin state annotations were the 25-state imputation based model obtained from http://compbio.mit.edu/roadmap (Ernst and Kellis 2015). GWAS catalog variants were obtained from https://www.ebi.ac.uk/gwas/, accessed on Dec 5, 2016. For annotations of DHS in the analysis in **Fig. 4c**, **Fig. 5d** and **Supplementary Fig. 18-19**, we used the Macs2 narrowPeak calls from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/. For **Fig. 5d** and **Supplementary Fig. 18-19** the Fetal Brain DHS were of epigenome sample E082 while the HepG2 DHS were of epigenome sample E118. For the median non-exonic DHS enrichments and ranking of states in the heritability partitioning analysis in **Fig. 2b**, **Fig. 6c**, **Supplementary Fig. 4,** and **Supplementary Table 2** we used narrowPeak calls from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/. PhyloP and PhastCons scores and constrained element calls were obtained from the UCSC genome browser. GERP++ scores and constrained element calls were obtained from http://mendel.stanford.edu/SidowLab/downloads/gerp/. SiPhy-omega and SiPhy-pi constrained element calls were the hg19 liftover version obtained from https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info. CNEEs from Lowe et al. (2011) used in **Supplementary Fig. 20** were obtained from http://www.stanford.edu/~lowec/data/threePeriods/hg19cnee.bed.gz. Assembly gap annotations were obtained from the Gap track from the UCSC genome browser.

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 (**Supplementary Fig. 5**). The BIC was calculated using the BIC_HMM function from the HMMpa R package (Witowski and Foraita 2014). 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 (**Supplementary Fig. 6**). 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. We saw a similar pattern of increasing enrichments for annotated TSSs 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 (Ernst and Kellis 2010, 2015; Ernst et al. 2011).

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 (Hahsler 2017). The leaves of the resulting hierarchical tree were ordered according to the optimal leaf ordering algorithm (Bar-Joseph et al. 2001) 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 **Supplementary Fig. 9**.

Gene Ontology enrichments

For each state and each protein coding gene based on GENCODE v19 (Harrow et al. 2012) 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 TSSs, 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 Gene Ontology enrichment analysis (ontology and annotations files from Nov. 24th, 2016) for the top genes in each state using the STEM software in batch mode with default options and the set of all genes considered as background (Ernst and Bar-Joseph 2006). STEM computed an uncorrected p-value based on the hypergeometric distribution for each term that is displayed in **Fig. 4b** and **Supplementary Fig. 12**. 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 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 (**Fig. 3e,f**). 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' UTRs. The motifs instances and background were obtained from http://compbio.mit.edu/encode-motifs/ (Kheradpour and Kellis 2014). The *known1* version of the motifs was used for both motifs.

Precision recall analysis for recovery of gene annotations and DHS

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 graphs in **Fig. 5a-c** and **Supplementary Fig. 17, 18** 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.

Clustering of cell-type specific DNase I hypersensitive site enrichments

For the clustering of DHS analysis, we first computed the enrichments of all conservation states for DHS for 53 samples processed by the Roadmap Epigenomics consortium (Roadmap Epigenomics Consortium et al. 2015), of which 16 were originally generated by the ENCODE project consortium (ENCODE Project Consortium 2012). 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 (Kolde 2015).

Heritability partitioning analysis

The heritability partitioning was performed using the LD-score regression ldsc software (Finucane et al. 2015). 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 our 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 (Finucane et al. 2015). To determine the two halves we ranked the conservation states in descending order of median fold-enrichment of non-exonic bases for DHS from 125 experiments from the University of Washington ENCODE group (ENCODE Project Consortium 2012). 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, BMI, coronary artery disease, educational attainment, height, LDL levels, schizophrenia and smoking behavior), all of which were previously considered in heritability partitioning analysis (Finucane et al. 2015). We followed the procedure for partitioning heritability as done in Finucane et al. (2015), including using a 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.

Software availability

The ConsHMM v1.0 software is available through https://github.com/ernstlab/ConsHMM. The unique code to ConsHMM is written in Python. The code of ConsHMM shared with ChromHMM is written in Java and included with ConsHMM.

Data availability

The ConsHMM conservation state annotations of hg19 are available at https://github.com/ernstlab/ConsHMM.

Author Contributions

A.S. and J.E. developed the method, analyzed the results, and wrote the paper.

Acknowledgements

We thank Ewan Birney and members of the Ernst lab for useful discussions. We acknowledge funding from US National Institutes of Health grants DP1DA044371, R01ES024995, U01HG007912 and U01MH105578 (J.E.), and T32CA201160 (A.S.), US National Science Foundation CAREER Award #1254200 (J.E.), a Kure-IT award and an Alfred P. Sloan Fellowship (J.E.).

References

Bar-Joseph Z, Gifford DK, Jaakkola TS. 2001. Fast optimal leaf ordering for hierarchical clustering. *Bioinforma Oxf Engl* **17 Suppl 1**: S22-29.

Blanchette M, Green ED, Miller W, Haussler D. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res* **14**: 2412–2423.

Chen X, Tompa M. 2010. Comparative assessment of methods for aligning multiple genome sequences. *Nat Biotechnol* **28**: 567–572.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**: 628–640.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**: 901–913.

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP. 2013. The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. *Cell* **154**: 185–196.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++ ed. W.W. Wasserman. *PLoS Comput Biol* **6**: e1001025.

Don PK, Ananda G, Chiaromonte F, Makova KD. 2013. Segmenting the human genome based on states of neutral genetic divergence. *Proc Natl Acad Sci* **110**: 14699–14704.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Ernst J, Bar-Joseph Z. 2006. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**: 191.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

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

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

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**: 1228–1235.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**: i54–i62.

Hahsler CB and M. 2017. *cba: Clustering for Business Analytics*. http://cran.us.r-project.org/web/packages/cba/index.html.

Harrow J, A F, Jm G, E T, M D, F K, Bl A, D B, A Z, S S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. 2016. Ensembl comparative genomics resources. *Database J Biol Databases Curation* **2016**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4761110/.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. 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.

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

Huang Y-F, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* **49**: 618–624.

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

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. 2016. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**: 1581–1586.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci* **111**: 6131–6138.

Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**: 2976–2987.

Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. 2014. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies ed. A. Di Rienzo. *PLoS Genet* **10**: e1004722.

Kim SY, Pritchard JK. 2007. Adaptive Evolution of Conserved Noncoding Elements in Mammals. *PLOS Genet* **3**: e147.

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

Kolde R. 2015. *pheatmap: Pretty Heatmaps*. http://cran.us.r-project.org/web/packages/pheatmap/index.html.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–482.

Litman GW, Anderson MK, Rast and JP. 1999. Evolution of Antigen Binding Receptors. *Annu Rev Immunol* **17**: 109–147.

Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. *Science* **333**: 1019–1024.

Marnetto D, Mantica F, Molineris I, Grassi E, Pesando I, Provero P. 2018. Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion. *Am J Hum Genet* **102**: 1–12.

Pickrell JK. 2014. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *Am J Hum Genet* **94**: 559–573.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat Methods* **11**: 294–296.

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

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**: D670-681.

Sarda S, Das A, Vinson C, Hannenhalli S. 2017. Distal CpG islands can serve as alternative promoters to transcribe genes with silenced proximal promoters. *Genome Res* **27**: 553–566.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Siepel A, Pollard KS, Haussler D. 2006. New Methods for Detecting Lineage-Specific Selection. In *Research in Computational Molecular Biology*, *Lecture Notes in Computer Science*, pp. 190–205, Springer, Berlin, Heidelberg https://link.springer.com/chapter/10.1007/11732990_17 (Accessed January 17, 2018).

Smit A, Hubley R, Green P. 2013. *RepeatMasker Open-4.0*.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**: 554–566.

Ward LD, Kellis M. 2012. Interpreting non-coding variation in complex disease genetics. *Nat Biotechnol* **30**: 1095–1106.

Weedon MN, Cebola I, Patch A-M, Flanagan SE, De Franco E, Caswell R, Rodríguez-Seguí SA, Shaw-Smith C, Cho CH-H, Allen HL, et al. 2014. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**: 61–64.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001-1006.

Witowski V, Foraita DR. 2014. *HMMpa: Analysing accelerometer data using hidden Markov models*. http://cran.us.r-project.org/web/packages/HMMpa/index.html.

Zhang MQ. 1998. Statistical features of human exons and their flanking regions. *Hum Mol Genet* **7**: 919–932.

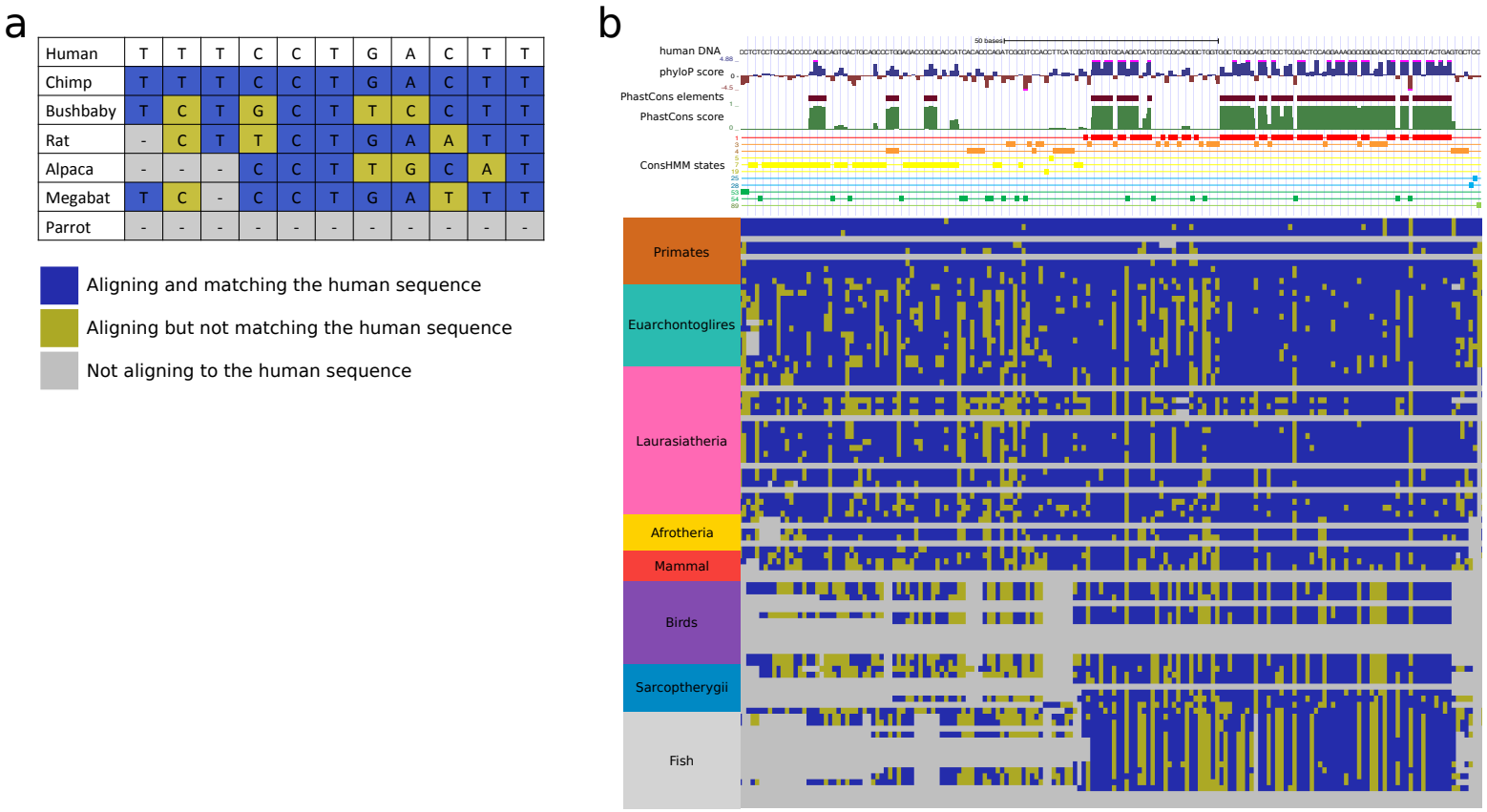

Figure 1: Illustration of ConsHMM modeling approach. (a) The input to ConsHMM is a multi-species alignment, which is illustrated for a subset of 6 species aligned to the human sequence. At each position and for each species ConsHMM represents the information as one of three observations: (1) aligns with 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 a locus chr22:25,024,640-25,024,812. 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 sequence alignment according to panel (a). The major clade of species as annotated on the UCSC genome browser21 are labeled and ordered based on divergence from human. Above the conservation state assignments are PhastCons constrained elements and scores and PhlyoP constraint scores. The figure and **Supplementary Fig. 8** 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.

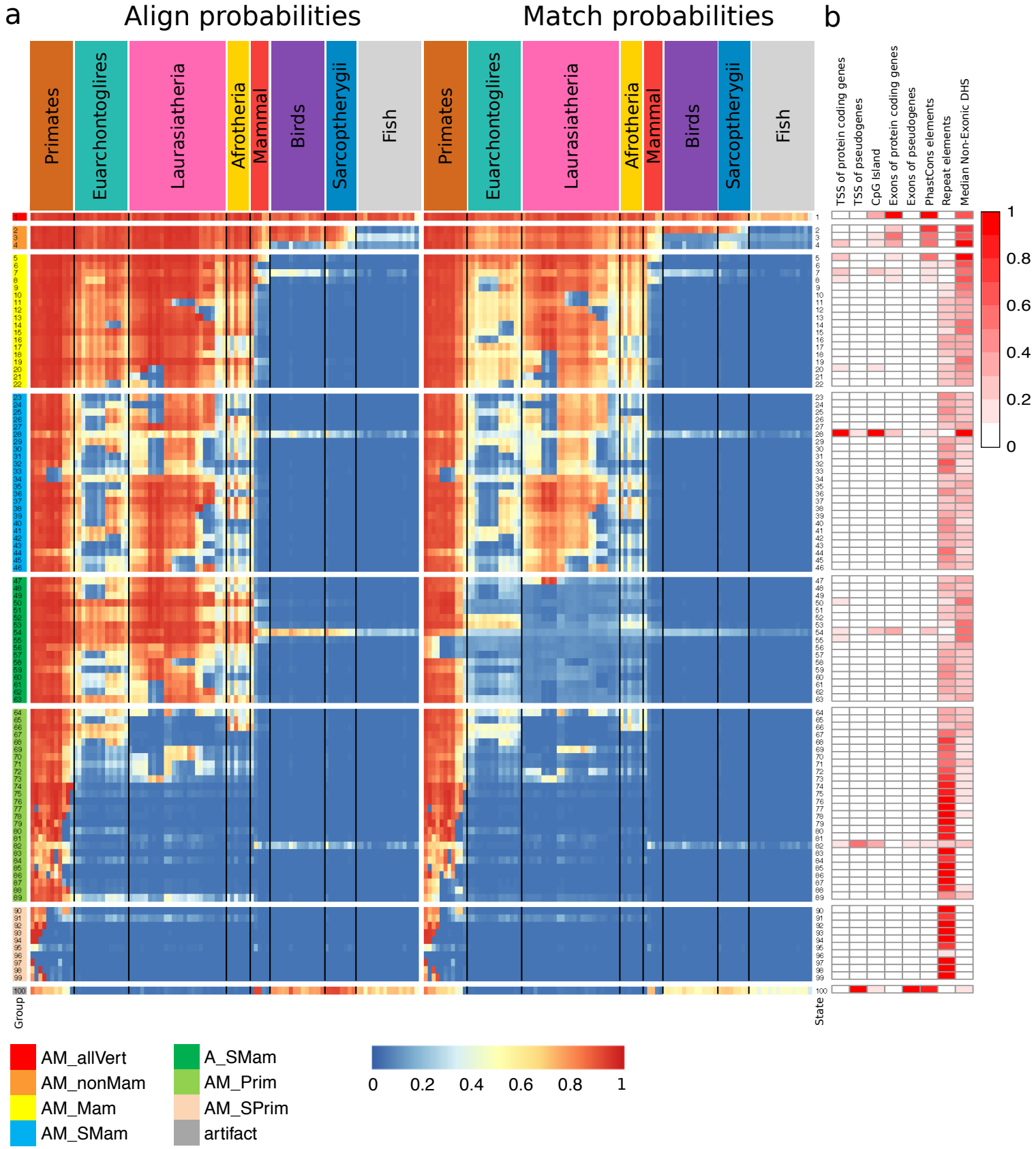

Figure 2: Conservation state emission parameters learned by ConsHMM and enrichments for other genomic annotations. (a) Each row in the heatmap corresponds to a conservation state. For each state and species, the left half of the heatmap gives the probability of aligning to the human sequence, which is one minus the probability of the not aligning emission. Analogously, the right half of the heatmap gives the probability of observing the matching emission. Each individual column corresponds to one species with the individual names displayed in **Supplementary Fig. 1**. 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 (Rosenbloom et al. 2015). The conservation states are ordered based on the results of applying hierarchical clustering and optimal leaf ordering (Bar-Joseph, Gifford, and Jaakkola 2001). The states are divided into eight major groups based on cutting the dendrogram of the clustering. The groups are indicated by color bars on the left hand side and a white row between them. The groups are AM_allVert (state 1, colored red), AM_nonMam (states 2-4, orange), AM_Mam (states 5-22; yellow), AM_SMam (states 23-46; blue), A_SMam (states 47-63; dark green), AM_Prim (states 64-89; light green), AM_SPrim (states 90-99; apricot), artifact (state 100; gray). Transition parameters between states of the model can be found in **Supplementary Fig. 2**. **(b)** From left to right, the columns of the heatmap indicate the relative enrichments of conservation states for: TSS of protein coding genes, TSS of pseudogenes, CpG islands, exons of protein coding genes, exons of pseudogenes, PhastCons elements, bases covered by any type of repeat element, and median enrichment of DHS in the non-exonic portion of the state from 125 experiments (**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. Values for these enrichments and additional enrichments can be found in **Supplementary Fig. 4** and **Supplementary Table 2** and enrichments for individual repeat classes and families can be found in **Supplementary Fig. 13**.

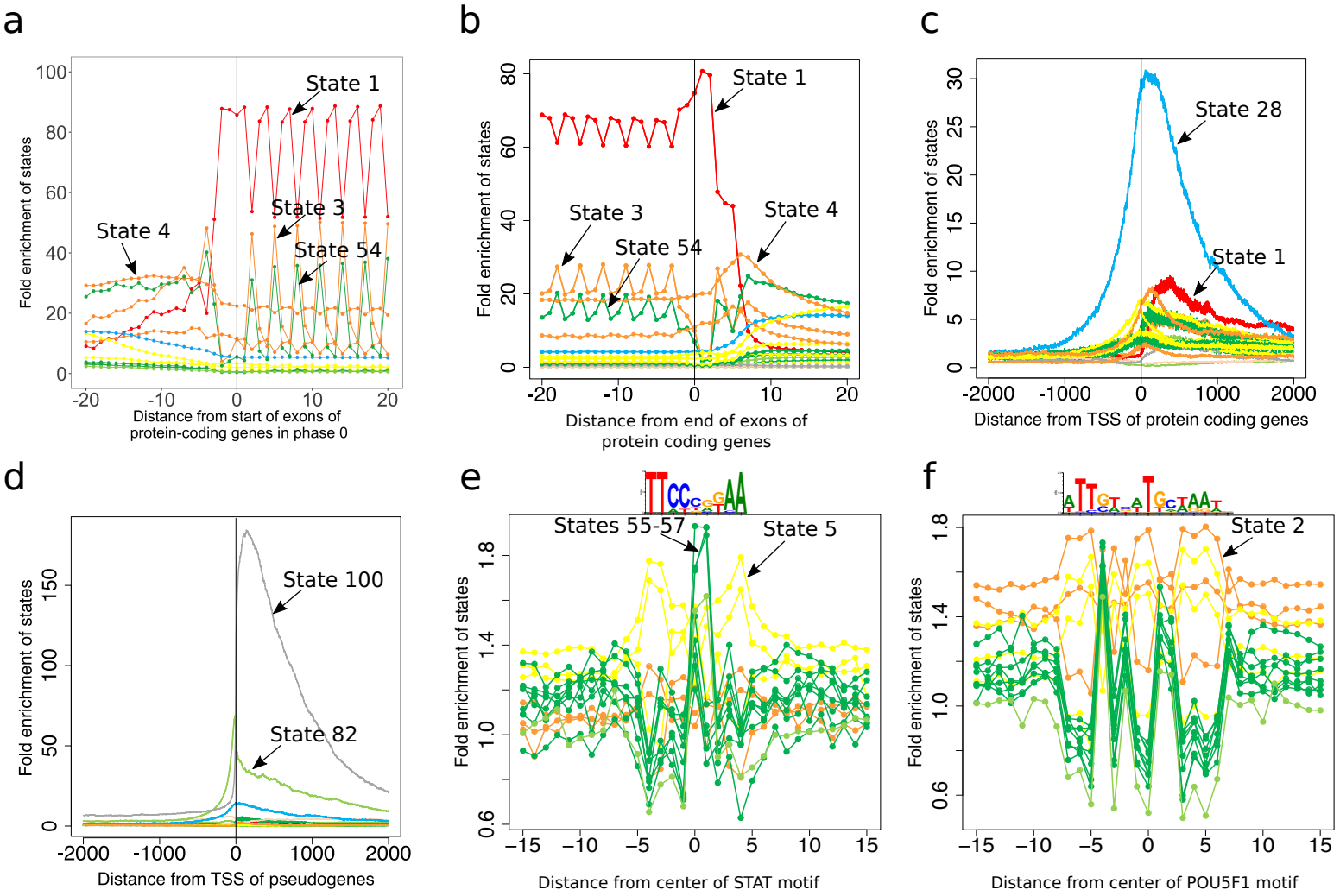

Figure 3: Conservation state positional enrichments. Plots of positional fold enrichments of conservation states relative to **(a)** start of exons of protein coding genes in phase 0, **(b)** end of exons of protein coding genes and TSS of **(c)** protein coding, and **(d)** pseudogenes genes. Positive values represent the number of bases downstream in the 5' to 3' direction of transcription, while negative values represent the number of bases upstream. Enrichments relative to gene annotations are based on a genome-wide background. The subset of states included in panels (a)-(d) were the states that had at least a 3 fold enrichment at some position within +/-2kb from the anchor point. Also shown are positional plots relative to the central nucleotide of a set of instances of **(e)** STAT and **(f)** POU5F1 motifs. The subset of states included in (e), (f) are the states that had an enrichment of at least 1.5 for some position within +/- 15bp from the center nucleotide of either motif. Enrichments for motif instances were computed relative to the portion of the genome scanned for regulatory motifs in Kheradpour and Kellis (2014), which excludes coding, 3'UTR, and repeat elements. Additional position enrichment plots can be found in **Supplementary Fig. 10**.

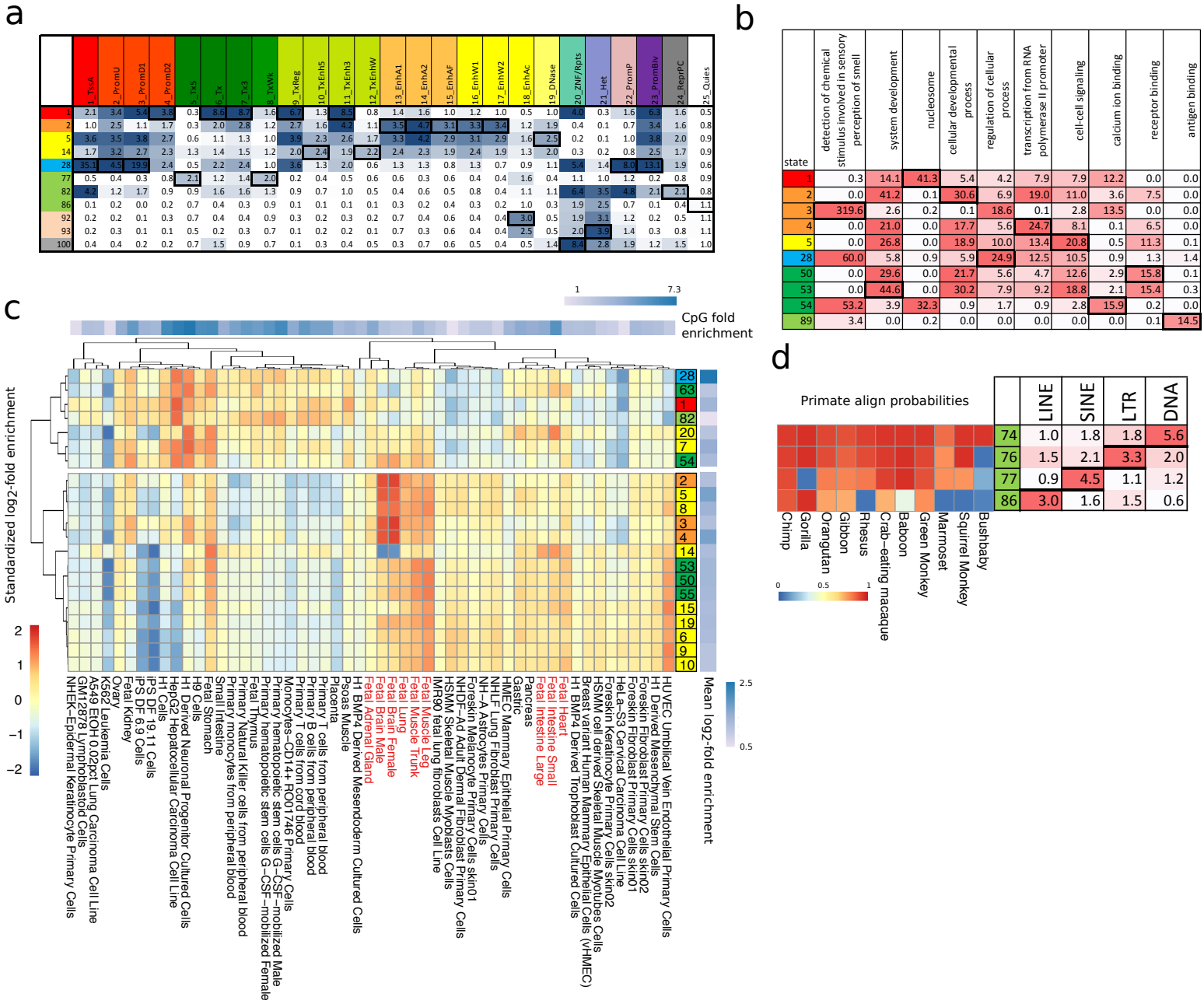

Figure 4: Conservation states enrichment for chromatin states, GO terms, DHS and repeat elements. (a) Each column corresponds to one of 25 chromatin states from a previously defined chromatin state model defined across 127 samples of diverse cell and tissue types (Ernst and Kellis 2015). The rows correspond to conservation states. Values correspond to the median fold enrichment of the conservation state for the chromatin state across all the samples. Only conservation states that had the maximum value for at least one chromatin state are shown, and those values are boxed. See **Supplementary Fig. 14** for the enrichments of all conservation states. **(b)** Columns correspond to GO terms and rows correspond to conservation states. Values correspond to the $-\log_{10}$ p-value (uncorrected) of the conservation state for the GO term where each conservation state is associated with its top 5% genes in its promoter region (**Methods**). Only GO terms which were the most enriched term for some conservation state 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 **Supplementary Fig. 12**. **(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 $\log_2$ transformed and then row normalized by subtracting the mean and dividing by the standard deviation. The subtracted mean values are shown on the right hand side. 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, indicated by the CpG island enrichments on top. In the bottom cluster conservation states generally had the strongest relative preference for a number of fetal related samples, labeled in red, with sub-clusters showing different relative preferences among those samples. **(d)** Conservation states with the maximal enrichment for LINE, SINE, LTR or DNA class repeats are shown with their enrichment for each of these classes on the right side. On the left side are 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.

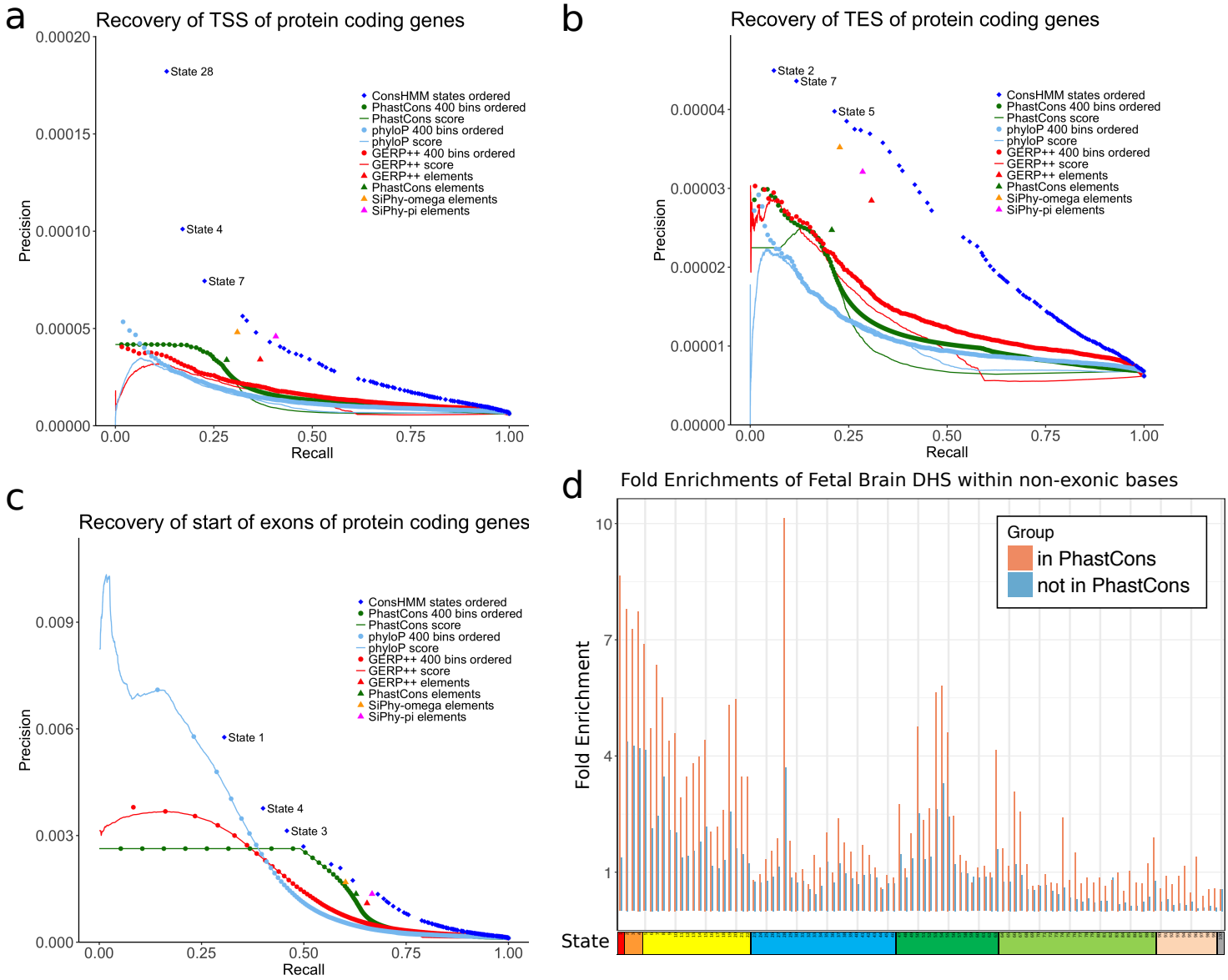

Figure 5: Relationship of conservation states with constrained elements and scores. Precision-recall plots for recovery of **(a)** TSS of protein coding genes, **(b)** TES of protein coding genes, and **(c)** the start of exons of protein coding genes. Recovery based on ordering ConsHMM conservation 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 **Supplementary Figs. 17-18** for plots based on additional targets. **(d)** The graph shows the fold enrichment for Fetal Brain DHS (Roadmap Epigenomics Consortium et al. 2015) 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 Supplementary Fig. 19** for graphs based on different element sets or DHS data.

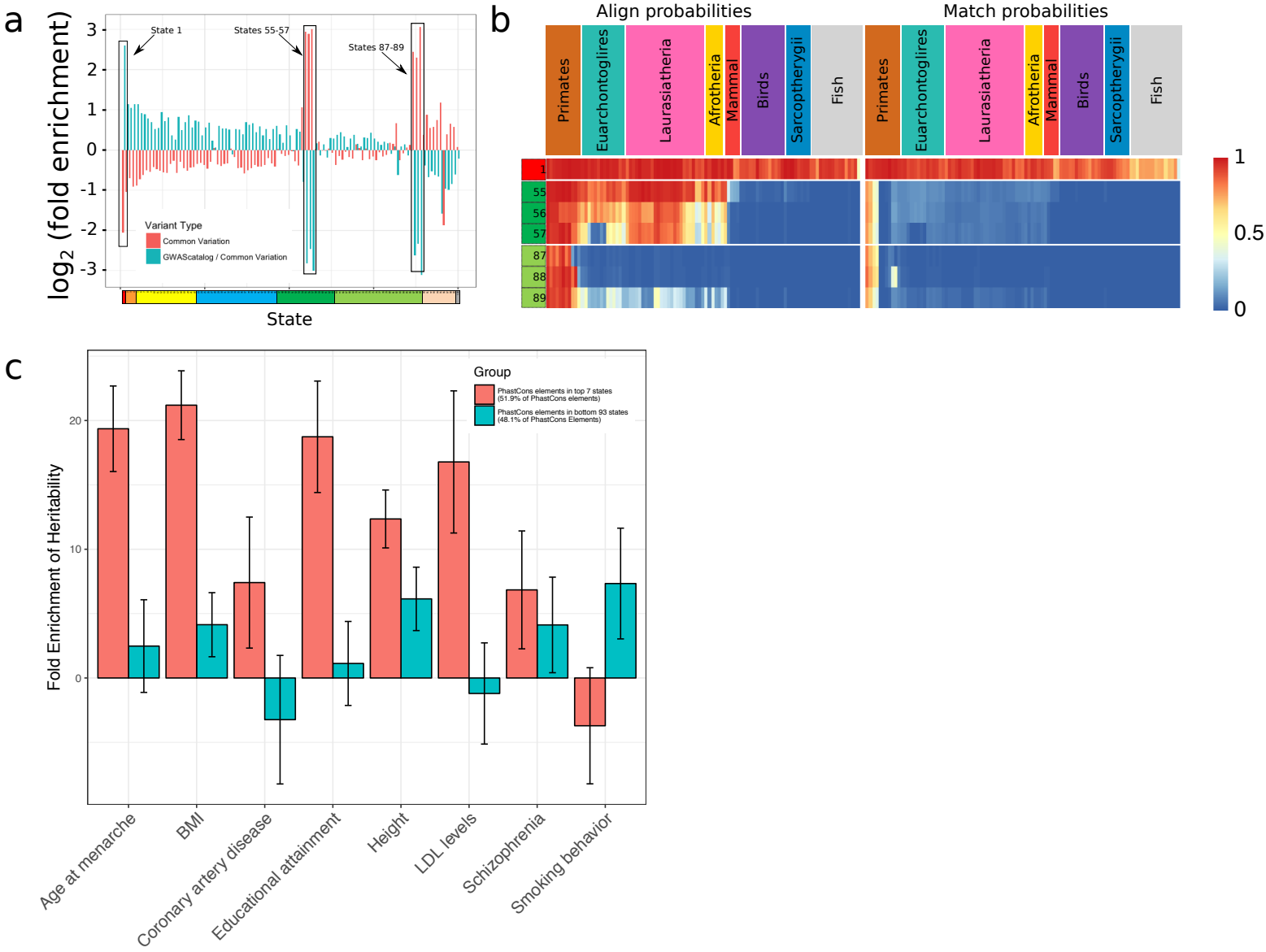

Figure 6: Conservation state enrichments for common and GWAS SNPs. (a) The panel displays the $\log_2$ fold enrichment of each state for common SNPs (pink) and GWAS catalog variants relative to common SNPs (blue). State 1, associated with high alignability and matching through all vertebrates, showed the greatest depletion of common SNPs and the highest enrichment for GWAS variants relative to common SNPs. States 55-57 and 87-89 exhibited the opposite pattern having the greatest enrichment for common SNPs and the greatest depletion of GWAS variants relative to this background. The second most depleted state for common SNPs, which did not show enrichment for GWAS catalog SNPs, was state 96 which captured large gaps in the assembly **(Supplementary Fig. 4)**. **(b)** Panel shows the representation of state emission parameters from **Fig. 2a** for the subset of states highlighted in panel (a). The states with the greatest depletion of GWAS variants all had relatively high alignability at least through primates, but low matching probabilities for almost all species except a few closely related primates. **(c)** Applying the heritability partitioning enrichment method of Finucane et al. (2015) on two disjoint subsets of bases in PhastCons elements, with eight phenotypes previously analyzed with heritability partitioning in the context of a baseline set of annotations (**Methods**) (Finucane et al. 2015). One set of bases are those in PhastCons elements that are also in 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). The other set are those bases in PhastCons elements overlapping the remaining 93 states covering 48.1% of PhastCons bases (blue).