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, 90095, 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, 90095, USA.

[4] Computer Science Department, University of California, Los Angeles, Los Angeles, California, 90095, USA.

[5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California, 90095, USA.

[6] Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California, 90095, USA.

*Correspondence: Jason Ernst (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 repeat families, which were used to characterize their biological significance. Conservation states have greater or complementary predictive information than standard constraint based measures for a variety of genome annotations. Bases in constrained elements have distinct heritability enrichments depending on the conservation state assignment, demonstrating their relevance to analyzing phenotypic associated variation. The conservation states also highlight differences in the conservation patterns of bases prioritized by a number of scores used for variant prioritization. The ConsHMM method and conservation state annotations provide a valuable resource for interpreting genomes and genetic variation.

Introduction

The large majority of phenotype-associated variants implicated by genome-wide association studies (GWAS) fall outside of protein coding regions.[1] Identifying the causal variants and interpreting their biological role in these less well understood non-coding regions is a significant challenge.[2] 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 genomes.[3–5] Using comparative genomics data to identify regions of evolutionary constraint has been a complementary approach for these purposes.[6–9]

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.[10,11] 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.[8,9,12–14] 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.[15] These scores and elements have also been highly informative features to

integrative methods for prioritizing pathogenic variants.[16–19] 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.[20]

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.[21] Approaches have been developed to associate constrained elements, regions, or individual bases with specific branches in a phylogenetic tree.[22–28] 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[29], 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.[4,30–32] Inspired by the success of these approaches, here we develop a method, ConsHMM, that extends ChromHMM[31] 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. We demonstrate that the conservation state annotations capture additional information that is not represented by scores or binary calls of constraint. We also show how the conservation states enable a deeper understanding of types of bases prioritized by a number of different scores used for variant prioritization, including those scores that integrate constraint information with a diverse set of other genomic annotations. Overall, 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.

Material and 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 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[31] 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 $P(y_{i,j}=1 \mid s_i=k)$. $g_{k,j}$ 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 $P(z_{i,j} = 1 \mid y_{i,j}=1 \text{ and } s_i=k)$. 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.[21,33]

154

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

156 Since for our application ConsHMM needs to run ChromHMM at single base resolution ('-b
157 1' flag) with 198 features after our binary encoding (2 for each non-human species in the 100-
158 way alignment), we had to address scalability issues in terms of both memory and time. To
159 address the memory issue we modified ChromHMM to support only loading in main memory
160 input for chromosomes it is actively processing, as previously ChromHMM would only support
161 loading all data into main memory upfront. This option can now be accessed in ChromHMM
162 through the '-lowmem' flag. To reduce the time required we used 12-parallel processors ('-p 12'
163 flag) and we trained on a different random subset of the human genome on each iteration of the
164 Baum-Welch algorithm. We divided each chromosome into 200kb segments (with the exception
165 of the last segment of each chromosome which was less than this) in order to form random
166 subsets of the human genome. We modified ChromHMM to allow training for each iteration on a
167 randomly selected subset of 150 of these segments ('-n 150' flag), corresponding to 30MB per
168 iteration. We ran this for 200 iterations by adding the '-d -1' flag, which removed one of
169 ChromHMM's default stopping criterion based on computed likelihood change on the sampled
170 data, since the likelihood is now expected to both increase and decrease between iterations as
171 different sequences are sampled. These new options were included in version 1.13 of
172 ChromHMM. The unique code to ConsHMM is written in Python. The code of ConsHMM shared
173 with ChromHMM is written in Java and included with ConsHMM.

174

175 *Generating genome-wide annotations*

176 After learning a 100-state model, we used it to segment and annotate the genome at
177 base-pair resolution into one of the 100 conservation states. Each base in the human genome is
178 classified into the state with the highest posterior probability. ConsHMM does this by running the
179 MakeSegmentation command of ChromHMM. Due to computational constraints, the

180 segmentation could not be generated for entire chromosomes at once. Instead, we ran
181 MakeSegmentation on the same 200kb partitioning made for learning the model. We then
182 merged the resulting files together using ConsHMM's mergeSegmentation.py command with
183 slice size parameter set to 200,000 ('-s 200000' flag) and the number of states parameter set to
184 100 ('-n 100 flag').
185
186 *Computing enrichments for external annotations*
187 All overlap enrichments for external annotations were computed using the ChromHMM
188 OverlapEnrichment command. OverlapEnrichment computes enrichments for an external
189 annotation in each state assuming a uniform background distribution. Specifically the fold
190 enrichment of a state for an external annotation is
191

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

192
193 Positional enrichments of states relative to an anchor point from an external annotation
194 were computed using the ChromHMM NeighborhoodEnrichment command at single base
195 resolution ('-b 1' flag), single base spacing from the anchor point ('-s 1') and using the '-l' and '-r'
196 flags to specify the size of the region of interest around the anchor point. The '-lowmem' flag
197 was also used for computing the enrichments for OverlapEnrichment and
198 NeighborhoodEnrichment.
199
200 *External data sources for enrichment analyses*
201 The external annotations of repeat elements were obtained from the UCSC genome
202 browser RepeatMasker track.[21,34] We generated an annotation for whether a base overlapped
203 any repeat element, as well as separate annotations for bases falling in each class and family of

204 repeat elements. The gene annotations were obtained from GENCODE v19 for hg19[35]. CpG
205 Island annotations were obtained from the UCSC genome browser. Annotations of SNPs with
206 >=1% minor allele frequency were obtained from the commonSNP147 track from the UCSC
207 genome browser, which is based on dbSNP build 147. GWAS catalog variants were obtained
208 from the NHGRI-EBI Catalog, accessed on Dec 5, 2016.[36] For annotations of DNase I
209 Hypersensitive Sites (DHS) processed by the Roadmap Epigenomics Consortium, we used
210 Macs2 narrowPeak calls.[5] The Fetal Brain and HepG2 DHS used were of epigenome samples
211 E082 and E118 respectively. For the median non-exonic DHS enrichments and ranking of
212 states in the heritability partitioning analysis we used narrowPeak calls from the ENCODE
213 consortium.[3] In the cases where ENCODE provided more than one replicate for a cell or tissue
214 type, we used the first replicate.

215 PhyloP and PhastCons scores and constrained element calls were obtained from the
216 UCSC genome browser. Assembly gap annotations were obtained from the Gap track from the
217 UCSC genome browser. The context-dependent tolerance score (CDTS) used was that based
218 on a cohort of 7784 unrelated individuals, following the analyses in Iulio et al.[37], which focused
219 on this version of the score. The CDTS and variants from this cohort were both lifted from hg38
220 to hg19 using the liftOver tool from the UCSC genome browser.[21]

221

222 *Choice of number of states*

223 We learned models with each number of states between 2 and 100 states. We set 100
224 as the maximum number of states we would consider for computational tractability and
225 maintaining a manageable number of states for analysis. The choice of a maximum of 100 also
226 corresponds to the number of species used and allows for the possibility of each state to cover
227 1% of the genome. We analyzed the Bayesian Information Criterion (BIC) for models with each
228 number of states between 2 and 100, and found that the BIC generally decreases as the
229 number of states increases in the range considered (**Figure S1**). The BIC was calculated using

230 the BIC_HMM function from the HMMpa R package.[38] Analyzing the 100-state model's internal
231 confidence estimate of its state assignments also supported a larger number of states.
232 Specifically, for each state in the 100-state model we computed the average posterior
233 probability of that state at each base in the genome assigned to it, and confirmed consistently
234 high average posterior probability values in the range [0.92,1.00] with a median of 0.97 (**Figure**
235 **S2**). The posterior probabilities were computed by running the MakeSegmentation command in
236 ChromHMM with the '-printposterior' flag. We also investigated if additional states in models
237 with larger number of states were biologically relevant. Specifically, we computed enrichments
238 for various external annotations for models with each number of states between 2 and 100 to
239 determine if biologically relevant enrichments were only robustly observed in models with more
240 than a certain number of states. In the case of CpG islands, we observed that only models with
241 at least 87 states consistently obtained >15 fold enrichment and only models with at least 95
242 states consistently obtained >30 fold enrichment (**Figure S3**). We saw a similar pattern of
243 increasing enrichments for annotated transcription start sites (TSSs) for models with large
244 number of states. We therefore decided to analyze the largest model, 100 states, that we were
245 considering. We note that annotations based on chromatin states used fewer number of states,
246 but were also defined on fewer features at a coarser resolution and had a less uniform genome
247 coverage.[4,30,39]
248
249 *State clustering*
250 We clustered the states based on the correlation of vectors containing the values $f_{k,j}$ and
251 $f_{k,j} \times g_{k,j}$ for each species j defined above. State clustering was performed using the hclust
252 hierarchical clustering function from the cba R package.[40] The leaves of the resulting
253 hierarchical tree were ordered according to the optimal leaf ordering algorithm[41] implemented in
254 the order.optimal R function from the cba package. We then cut the tree such that the 8 major
255 groups of states were designated. The full tree is provided in **Figure S4**.

256

257 *Genome segmentation using uniform transition probabilities*

258 For analyzing the effect of the transition probabilities on the genome segmentation, we
259 created a separate model, which was the same model we used in the main analyses, except we
260 set all transition probabilities to 0.01, corresponding to each state having an equal probability of
261 transitioning to any state including itself. We then created a new genome segmentation by
262 running the MakeSegmentation command in ChromHMM with this new model. For each state,
263 we counted how many of the bases assigned to it in the original annotation were also assigned
264 to it in the annotation created with the uniform transitions and divided this number by the
265 number of bases in the state in the original annotation. This calculation provided a fraction from
266 0 to 1. We also reported the number of segments produced by each model, where a segment is
267 defined to be one or more consecutive bases all assigned to the same state, such that any
268 immediately adjacent bases are assigned to a different state or states.

269

270 *Gene Ontology enrichments*

271 For each state and each protein coding gene based on GENCODE we computed the
272 number of bases in that state that are within +/- 2kb of the gene's TSS. In the case of genes
273 with multiple annotated TSSs, we used the outermost TSS. We then created a ranking of genes
274 for every state by sorting the genes in descending order of this number of bases. For each state
275 we then created a set of 969 genes that represent the top 5% of genes in the state among the
276 19,397 genes we considered. We performed a Gene Ontology (GO) enrichment analysis
277 (ontology and annotations files from Nov. 24th, 2016) for the top 5% genes in each state using
278 the STEM software in batch mode with default options and the set of all genes considered as
279 background.[42] STEM computed an uncorrected p-value based on the hypergeometric
280 distribution for each term displayed in the figures summarizing the analysis. STEM also reported

281 corrected p-values for testing multiple GO terms for a single state based on randomization to
282 three significant digits, which was less than 0.001 for all p-values mentioned in the main text.
283
284 *Transcription factor binding site motif enrichments*
285 We computed the enrichment of the conservation states within 15 bases upstream and
286 downstream of the center point of the POU5F1 and STAT known transcription factor-binding site
287 motifs.[43] The enrichment was computed relative to the background regions of the genome that
288 were used to identify the motifs, which excluded repeat elements, coding sequence, and 3'
289 untranslated regions (UTRs). The *known1* version of the motifs was used for both motifs.
290
291 *Clustering of cell-type specific DNase I hypersensitive site enrichments*
292 For the clustering of DHS analysis, we first computed the enrichments of all conservation
293 states for DHS for 53 samples processed by the Roadmap Epigenomics consortium[5], of which
294 16 were originally generated by the ENCODE project consortium.[3] We then selected the subset
295 of states that had a fold enrichment of at least 2 in at least one sample, leading to a subset of 21
296 conservation states. To more directly focus on each state's relative enrichments across
297 samples, we $\log_2$ transformed each enrichment value, and then normalized the enrichments for
298 each state by subtracting the mean enrichment across samples and dividing by the standard
299 deviation. We then hierarchically clustered the states based on the correlation of their
300 enrichments across samples and hierarchically clustered the samples based on their
301 correlations across states using the pheatmap R package.[44] We also computed for each sample
302 the fold enrichment of DHS bases for bases in CpG islands, as the ratio between the percent of
303 DHS bases in CpG islands and the percent of the genome falling in CpG islands.
304
305 *Precision recall analysis for recovery of gene annotations and DHS*

306 We randomly split the 200kb genome segments used for training the model and
307 segmentation into two halves corresponding to training and testing data. For each target set in
308 the precision-recall analyses we ordered the ConsHMM states in decreasing order of their
309 enrichment for the target among the training set bases. We then used that ordering to iteratively
310 add the testing set bases in each state to form cumulative sets of bases predicted to be of the
311 target set and computed the precision and recall for them. For each constraint score we
312 computed the precision-recall curve for predicting the target set in the test data using two
313 methods. For the first method, we directly ordered bases in descending order of their assigned
314 score. For the second method, we split the sorted scores into 400 bins such that each bin
315 contains on average 0.25% of the genome, which was the size of the smallest state of the
316 ConsHMM model (0.25% of the genome in state 100). Specifically, we assigned all bases in the
317 genome where the score was not defined to one bin and then divided the remaining bases
318 uniformly among the 399 other bins based on their score. In some cases score increments were
319 at the boundary between two bins at their provided floating-point precision, or overlapped
320 multiple bins. In these cases we uniformly split the target bases assigned to that score
321 increment into multiple bins proportionally to the overall percentage of the score increment
322 falling in each bin. We then treated the 400 bins as 400 states and followed the same procedure
323 described for the ConsHMM states. We also computed the precision and recall of bases in each
324 constrained element set for predicting the target set on the testing data. For the DHS analyses,
325 we also separately evaluated recovery of DHS bases when restricting the analysis to non-
326 exonic regions. Additionally, both genome-wide and within non-exonic regions, we evaluated the
327 recovery of DHS bases when restricting the analysis to bases distal to a TSS, defined as more
328 than 2kb from a TSS.
329
330 *Precision recall analysis for recovery of DHS bases aggregated across cell and tissue types*

331 For the analysis of the recovery of DHS aggregated across cell and tissue types we
332 concatenated DHS from 53 cell or tissue types processed by the Roadmap Epigenomics
333 Consortium into one annotation in which each combination of chromosome and cell or tissue
334 type effectively becomes a new chromosome. We then split the concatenated data into training
335 and testing sets as described above. We computed the enrichments of the ConsHMM states
336 and scores split into bins as detailed above, but multiplying the size of each state and bin by the
337 number of DNase I hypersensitivity data sets. The precision and recall values for the ConsHMM
338 states, constraint scores considered directly, constraint scores split into bins, and constrained
339 element sets were then computed on the testing data.
340
341 *Enrichment analysis for constrained non-exonic elements assigned to phylogenetic branches*
342 We lifted over the constrained non-exonic elements (CNEEs) from *Lowe et al.*[22] from
343 hg18 coordinates to hg19, using the liftOver tool from the UCSC genome browser with default
344 settings.[21] These elements were previously partitioned into subsets based on the inferred
345 branch point of origin in a phylogenetic tree.[22] We computed the enrichments of the
346 conservation states for all the CNEEs and for each subset of the CNEEs separately, using the
347 OverlapEnrichment command from ChromHMM at single nucleotide resolution (`-b 1` flag) and
348 using the low memory option (`-lowmem`). We also computed analogous enrichments for
349 CNEEs overlapping PhastCons elements called on the same 100-way alignment that the
350 conservation states were annotated based on. To compute the enrichments of CNEEs for bases
351 in CpG islands we created an annotation consisting of a state for each CNEE subset and one
352 additional state for bases not assigned to any CNEE. We then ran the same OverlapEnrichment
353 command as above to compute enrichments of CNEE bases for non-exonic CpG islands, and
354 non-exonic bases in general. The reported enrichment of CpG islands is the ratio of these two
355 enrichments, effectively computing an enrichment relative to the non-exonic background. The

356 set of non-exonic bases for the enrichment analysis was generated by excluding all bases
357 annotated as an exon in GENCODE v19.
358
359 *Heritability partitioning analysis*
360 The heritability partitioning was performed using the LD-score regression ldsc software.
361 [15] We partitioned the PhastCons constrained elements into two halves based on a ranking of the
362 conservation states. We focused on the PhastCons constrained elements for this analysis, since
363 it was the only element set defined on the same alignments as our conservation states. We
364 focused on halves since the LD-score regression estimates can be unstable for annotations
365 covering too small of a percentage of the genome.[15] To determine the two halves we ranked the
366 conservation states in descending order of median fold-enrichment of non-exonic bases for
367 DHS from 123 experiments from the University of Washington ENCODE group.[3] We then
368 divided bases in PhastCons elements between the top 7 ranked states (1-5, 8 and 28), which
369 contain 51.9% of bases in PhastCons elements, and the bottom 93 states, which contain the
370 other 48.1% of bases in PhastCons elements. We applied ldsc to these two sets for 8 traits (age
371 at menarche, body mass index (BMI), coronary artery disease, educational attainment, height,
372 low-density lipoprotein (LDL) levels, schizophrenia and smoking behavior), all of which were
373 previously considered in heritability partitioning analysis.[15] We followed the procedure for
374 partitioning heritability as done in *Finucane et al.*[15], including using the baseline annotation set
375 and 500 base-pair windows around annotations to dampen the artificial inflation of heritability in
376 neighboring regions caused by linkage disequilibrium. The baseline annotation set contains a
377 range of annotations including DHS. For our analysis, we first removed the constrained element
378 set already included in the baseline annotation set, then added our two halves of PhastCons
379 elements and finally ran the ldsc software on the full set of annotations.
380
381 *Enrichment analysis for variant prioritization scores*

382 For each variant prioritization score included in the conservation state enrichment
383 analysis of prioritized bases, we extracted the top 1%, 5% and 10% of all the bases ranked by
384 each score, both genome-wide and just in non-coding regions. The non-coding regions were
385 defined as the intersection of where the LINSIGHT and FunSeq2 scores provided a value, as
386 these two scores were only defined on non-coding regions. This intersection results in a set of
387 bases covering 90% of the genome that excludes coding regions in addition to other regions
388 filtered for technical reasons by either of the two methods.[19,45] For each score we chose the
389 score threshold that gave us a size for the top set that was as close as possible to the target
390 percentage, which did not always exactly match the target percentage due to the precision of
391 the scores. If a score did not provide a value for a particular base being considered, then that
392 base was assigned to the lowest value of that score, but would still be counted when
393 establishing the percentage thresholds. For the scores that provided separate score values for
394 alternate alleles at a certain position, we used the maximum of the values for all alleles. The
395 state enrichments were then computed using the OverlapEnrichment command from
396 ChromHMM at single base resolution ('-b 1' flag) and with the low memory option ('-lowmem'
397 flag). For the analysis restricted to non-coding regions, we also computed the enrichment of the
398 states for this background region using the same command. The enrichment for each score in a
399 state was then divided by the enrichment of the background region for the state. For the Eigen
400 and Eigen-PC scores we used version 1.1, for FunSeq2 we used version 2.1.6, and for CADD
401 we used both v1.0 and v1.4.
402
403 **Results**
404
405 *Annotating the human genome into conservation states*
406 We developed an approach, ConsHMM, to annotate a genome into different
407 conservation states based on a multiple species DNA sequence alignment (**Figure 1A,**

408 **Methods**). We model the combinatorial patterns within the alignment of which species align to
409 and which match a reference genome, for which we used the human genome. Specifically, at
410 each nucleotide in the human genome we encode one of three possible observations for each
411 non-reference species in the alignment: (1) aligns with a nucleotide present that is the same as
412 the human reference genome, (2) aligns with a nucleotide present that is different than the
413 human reference genome, or (3) does not have a nucleotide present in the alignment for that
414 position. We further model these observations as being generated from a multivariate hidden
415 Markov model (HMM), which probabilistically captures both the combinatorial patterns in the
416 observations and their spatial context. Specifically, we assume that in each state the probability
417 of observing a specific combination of observations is determined by a product of independent
418 multinomial random variables. The parameter values of these multinomial random variables will
419 differ between states and are learned from the data. After the model is learned, each nucleotide
420 in the human genome is assigned to the state that had the maximum posterior probability of
421 generating the observations.

422 ConsHMM builds on ChromHMM[31], which has previously been applied to annotate
423 genomes based on epigenomic data at 200-bp resolution[30], to now annotate genomes at single
424 nucleotide resolution based on a multiple species DNA sequence alignment (**Methods**). We
425 applied ConsHMM to a 100-way Multiz vertebrate alignment with the human genome and
426 focused our analysis here on a model learned using 100 states in order to balance recovery of
427 additional biological features and model tractability (**Figures 2** and **S1-S8, Tables S1** and **S2,**
428 **Methods**). We note that HMMs have previously been used to provide local smoothing of signal
429 for the task of identifying constrained elements.[9,14] We verified that the HMM had a smoothing
430 effect in our application by comparing to a segmentation derived from a model with the same
431 emission parameters as our learned model, but that ignored the information in the transition
432 parameters (**Methods**). We saw an increase in the number of segments from 889 million to 1.06
433 billion when not using the transition information, though a large majority of the state

434 assignments to individual bases were the same (**Figure S9**). We illustrate the ConsHMM
435 conservation state annotations at two different loci showing that different bases that are
436 associated with calls of evolutionary constraint from existing approaches can have very different
437 underlying alignment patterns and conservation state assignments (**Figures 1B** and **S10**).
438 Conservation state annotations genome-wide are available online (**Web Resources**).
439
440 *Major groups of conservation states*
441 We hierarchically clustered the conservation states based on their align and match
442 probabilities, and then cut the resulting dendrogram to reveal eight notable groups of states or
443 distinct individual states (**Figures 2A** and **S4, Table S3, Methods**). We named the resulting
444 groups based on the aligning and matching properties of major subsets of species for most
445 states in each group. We also summarized for each individual state the most distal species to
446 human that had a majority of positions aligning and the closest one that did not, and similarly for
447 matching (**Table S3**). The first of these subsets of states was a single state (State 1;
448 AM_allVert) that showed high align and match probabilities through essentially all vertebrate
449 species considered. The second subset showed relatively high align and match probabilities for
450 all mammals and some non-mammalian vertebrates (States 2-4; AM_nonMam). The third
451 subset showed relatively high align and match probabilities for most if not all mammals, but not
452 non-mammalian vertebrates (States 5-22; AM_Mam). The fourth subset showed high align
453 probabilities for many mammalian species, but had low align probabilities for notable species
454 such as mouse and rat for many of the states in the group (States 23-46; AM_SMam). The
455 combination of the absence of mouse and rat alignments with the presence of mammals that
456 are assumed to have diverged earlier is consistent with the previously observed increased
457 substitution rates for mouse and rat.[7] The fifth subset showed high align probabilities for many
458 mammalian species, but did not show high match probabilities (States 47-63; A_SMam). The
459 sixth subset showed high align probabilities for most primates, but not for other species (States

460 64-89; AM_Prim). The seventh subset showed high align probabilities for at most a subset of
461 primates (States 90-99; AM_SPrim). The final subset was a single state (State 100; artifact) that
462 showed a noteworthy pattern of high align and match probabilities for most primates and non-
463 mammalian vertebrates, but low probabilities for non-primate mammals, consistent with a
464 previous observation that inclusion of non-mammalian vertebrates can be associated with
465 increased presence of suspiciously aligned regions.[46]
466
467 *Conservation states exhibit distinct patterns of positional enrichments relative to gene*
468 *annotations and regulatory motif instances*
469 The conservation states showed strong and distinct positional enrichments relative to
470 GENCODE[35] annotated gene features including transcription start sites (TSS), transcription end
471 sites (TES), exon start sites, and exon end sites for both protein coding genes and
472 pseudogenes (**Figures 3A-D** and **S11**). Notable positional enrichments were also seen for
473 regulatory motifs instances (**Figure 3E** and **3F**). Relative to starts of exons of protein coding
474 genes seven of the states (States 1-4, 7, 28, and 54) had 13 fold or greater enrichment for some
475 position within 20 base pairs of exon starts, both when considering all such exons and subsets
476 of exons in specific coding phases (**Figures 3A** and **S11A-C**). These seven states were the only
477 states that had a majority of positions aligning for at least some non-mammalian vertebrates,
478 while still having a majority of positions aligning for all primates (**Figure 2A** and **Table S3**).
479 Within exons we saw the strongest enrichment for states 1-4 and 54, and among these state 1
480 showed the strongest enrichment, as expected given its high match probabilities through all
481 vertebrates (**Figures 2B, 3A-B,** and **S11A-E**). Interestingly, state 1 showed very strong
482 enrichment (>80 fold) in the two nucleotides immediately upstream of the exon start, with the
483 third upstream nucleotide also having high enrichment (46 fold) (**Figure S11C**). These three
484 nucleotide positions correspond to the positions of the canonical 3' splice site sequence that is
485 highly conserved throughout vertebrates.[47] At the ends of exons of protein coding genes

486 (**Figure 3B**), state 1 maintained a >40 fold enrichment for six nucleotides past the end of coding
487 sequence corresponding to positions of the known canonical 5' splice site sequence.[47]
488 Downstream of the start of protein-coding exons, the enrichment profile for state 1 showed a 3-
489 bp oscillation period, with a dip of enrichment at each 3rd base corresponding to codon wobble
490 positions. In contrast, states 3 and 54, which were both associated with high align probabilities
491 through many vertebrates and lower match probabilities, showed the inverse oscillation pattern
492 to state 1 (**Figures 3A** and **S11A-C**).

493 Relative to TSS of protein coding genes, state 28 had the strongest enrichment reaching
494 a maximum enrichment 30 fold at the TSS (**Figure 3C**). State 28 was associated with moderate
495 align and match probabilities for almost all the species present in the alignment. Consistent with
496 its enrichment for TSSs state 28 also had the greatest enrichment for CpG islands (32 fold).
497 However, state 28 also showed a 20 fold enrichment of CpG islands >2kb away from any TSS
498 of protein coding genes and a 10 fold enrichment for TSS of protein coding genes >2kb away
499 from a CpG island, suggesting the possibility that both of these features are making a partially
500 independent contribution to the association, or the presence of additional unannotated TSSs
501 that are associated with CpG islands.[48] Relative to TES of protein coding genes we saw the
502 enrichment peak for state 2 at almost 12 fold (**Figure S11F**), which had high align and match
503 probabilities for almost all vertebrates except for fish.

504 Relative to pseudogene exon starts and ends, states 100 and 82, both associated with
505 alignability to distal vertebrates without many mammals closer to human (**Figure 2B** and **Table**
506 **S3**), had strong enrichments peaking at greater than 100 and 38 fold respectively (**Figure S11G**
507 and **S11H**). These two states also showed the greatest enrichment relative to TSSs of
508 pseudogenes peaking at 184 and 68 fold for states 100 or 82 respectively (**Figure 3D**) and for
509 TESs of pseudogenes peaking at 199 and 61 fold respectively (**Figure S11I**).

510 Relative to instances of regulatory motifs, different conservation states showed single
511 nucleotide enrichment variation, often associated with variation in the amount of information in

512 the positional-weight matrix (**Figure 3E** and **3F, Methods**).[43] For example, in the case of the
513 POU5F1 and STAT motifs we saw state 2 from the AM_nonMam group and state 5 from the
514 AM_Mam group respectively reach 1.8 fold enrichments, but have lower enrichments (1.4-1.5)
515 at some nucleotides with lower information content. For the STAT motif, states 55-57,
516 associated with high align probabilities for most mammals, but high match probabilities only for
517 a few primates, showed enrichments that peaked at the CG dinucleotide in the center of the
518 motif, consistent with their genome-wide enrichments for CG dinucleotides (**Figures 3E and**
519 **S12**).
520
521 *Enrichment of conservation states for different gene classes*
522 The previous analyses demonstrated that different conservation states have distinct
523 enrichments in promoter regions of genes. We next investigated whether different conservation
524 states also exhibit distinct enrichments for different classes of genes after controlling for the
525 state's relative preference for promoter regions. Specifically, for each state we determined the
526 5% of genes with the greatest presence of the state in its promoter region and evaluated Gene
527 Ontology (GO) enrichments for those genes, revealing distinct enrichment patterns (**Figures 4B**
528 and **S13, Methods**). For example, even among states 1-3, all of which had high alignability
529 through at least birds and matching through mammals, we observed substantial differences in
530 their gene preferences. Out of these three states, state 1 (the AM_allVert group) was the only
531 one enriched for nucleosomes ($p<10^{-41}$; 10.5 fold), while state 3, which had high matching only
532 through mammals, was the only one with a significant enrichment for a set of genes related to
533 sensory perception of smell ($p<10^{-300}$; 15.5 fold). State 2, which had high align and match
534 probabilities through all vertebrates except fish, was the state most enriched for cellular
535 developmental processes ($p<10^{-30}$; 1.8 fold), which did not show enrichment in state 3. We also
536 observed notable enrichments for states with overall lower align or match probabilities. For
537 example, state 89, associated with high alignability and low matching in primates as well as

538 some alignability and low matching in non-primate mammals, was the state most enriched for
539 antigen binding ($p<10^{-14}$; 6.7 fold). This is consistent with antigen binding being associated with
540 many species, but fast evolving.[49]
541
542 *Enrichments for repeat elements in conservation states*
543 The conservation states showed a wide range of enrichments and depletions (from 2
544 fold enrichment to 133 fold depletion) for bases overlapping any repeat element (**Figures 2B**
545 and **S8**).[21,34] Of the 25 states that did not have any species outside of primates with a majority of
546 positions aligning, all but two had an enrichment of 1.55 or greater for repeat elements, while
547 the other 75 states all either had an enrichment below that or did not show enrichment (**Table**
548 **S3**). The two exceptions were state 89 and state 96, neither of which showed enrichment for
549 repeat elements. As noted above, state 89 is likely associated with fast evolving bases shared
550 with some non-primate mammals, as opposed to bases new to primates. State 96 is associated
551 with assembly gaps (**Figure S8**). Different repeat classes and families had distinct patterns of
552 enrichments for different states, even though in some cases the difference in state parameters
553 was subtle (**Figures 4D** and **S14**). For instance, among states in the AM_Prim group, which
554 primarily differed in terms of the specific combinations of primates with high align and match
555 probabilities, we found distinct enrichments. Notably, four different states from the group
556 AM_Prim, 74, 86, 76, and 77, showed maximal enrichments for the DNA, LINE, LTR, and SINE
557 repeat classes respectively (**Figure 4D**). State 74, which is characterized by high align and
558 match probabilities for all primates, had an enrichment of 5.6 fold for DNA repeats, while the
559 enrichment for the other three classes were between 1.0 and 1.8 fold. On the other hand, state
560 86, which lacked alignability of a subset of primates, had a 3.0 fold enrichment for LINE repeats,
561 while the enrichment for the other classes were between 0.6 and 1.6 fold. States 76 and 77 had
562 3.3 and 4.5 fold enrichments for LTR and SINE respectively compared to 1.1 and 2.1 fold for
563 SINE and LTR respectively. State 76 and state 77 both had high align probabilities through

564 primates up to and including squirrel monkey, with the exception that state 77 lacked alignability
565 to gorilla. Despite these subtle differences in the alignment probabilities, these states had
566 substantial differences in their repeat enrichment profiles.
567
568 *Relationship of conservation states to chromatin states*
569 We compared our conservation states to annotations of the genome based on a 25-
570 chromatin state model defined on 127 samples of diverse cell and tissue types using imputed
571 data (**Figures 4A** and **S15**).[5,39] For each conservation state we determined the median
572 enrichment of each chromatin state across the 127 samples. Eleven different conservations
573 states were maximally enriched for at least one of the 25-chromatin states. Conservation state
574 28 showed the greatest enrichment for any chromatin state, with a 35 fold enrichment for a
575 chromatin state associated with active promoters, and was maximally enriched for four other
576 promoter associated chromatin states. Conservation state 1 was maximally enriched for five
577 chromatin states all associated with transcribed and exonic regions[39] (3.8-8.7 fold), which is
578 consistent with this conservation state being most enriched for exons. Conservation state 2 had
579 the maximal enrichment for five enhancer associated chromatin states (3.1-4.7 fold), while
580 conservation state 5 had high enrichments for these states and also had the greatest
581 enrichment of any conservation state for a chromatin state primarily associated with just DNase
582 I hypersensitivity (2.5 fold). These and other distinct enrichments of the conservation states for
583 the different chromatin states highlight that conservation states are able to capture multi-
584 dimensional information in the genome.
585
586 *Conservation states capture enrichment patterns of DNase I hypersensitive sites across cell and*
587 *tissue types*
588 The previous analysis demonstrated that conservation states can exhibit different
589 enrichment patterns for different chromatin states. We next investigated whether different

590 conservation states also capture distinct enrichment patterns for a chromatin mark across cell
591 and tissue types. For this we analyzed DNase I hypersensitive sites (DHS) from 53 of the 127
592 samples considered above for which maps of experimentally observed DHS were available from
593 the Roadmap Epigenomics Consortium.[5] We focused on the 21 conservation states that
594 exhibited at least 2 fold enrichment in at least one sample (**Figure 4C**). We then row normalized
595 the enrichments in order to focus on the relative enrichment patterns across cell and tissue
596 types (**Methods**). Hierarchical clustering of the enrichment patterns revealed two major clusters
597 of states (**Figure 4C**). One of these clusters contained 14 of the 21 states and was associated
598 with strong enrichments for fetal related samples. Ten of the states in this cluster have
599 maximum enrichment for a fetal sample, while the remaining four states have maximum
600 enrichment for the cell type Human Umbilical Vein Endothelial Cells (HUVEC). The second
601 major cluster consisted of seven states, all of which were enriched for CpG islands (**Figures 2B**
602 and **S8**). The DHS from samples that showed the greatest enrichments in states in these
603 clusters also had the greatest enrichment of CpG islands (**Figure 4C, Methods**), but were
604 biologically diverse in terms of the type of cell or tissue and could potentially reflect technical
605 experimental differences.
606
607 *Relationship of conservation states to constraint based annotations*
608 We next investigated the relationship of our conservation state annotations with calls
609 and univariate scores of evolutionary constraint. Specifically, we considered constrained
610 element sets based on four methods (GERP++, SiPhy-omega, SiPhy-pi, and PhastCons) and
611 constraint scores based on three methods (GERP++, PhastCons, and PhyloP) publicly available
612 for hg19 and also defined on Multiz alignments. The PhastCons and PhyloP scores and
613 elements we compared to were defined on the same 100-way vertebrate alignment. The
614 available GERP++, SiPhy-omega, and SiPhy-pi score and elements were derived from different
615 versions of Multiz alignments and only considered mammals.

616 We consistently found conservation states 1-5 to be highly enriched (>9 fold) for all
617 constrained element sets (**Figures 2B** and **S16A**). These states were also among the top six
618 states in terms of mean score for constraint scores considered (**Figure S16B**). Consistent with
619 this, states 1-5 were the states that had the highest average matching probability across
620 mammals. Two other states exhibited at least 6 fold enrichment for at least one constrained
621 element set: states 54 and 100. State 100, associated with putative artifacts, showed high
622 enrichments for PhastCons elements (15 fold) and high average scores for PhastCons and
623 PhyloP. This is consistent with this state having high aligning and matching probabilities
624 primarily in non-mammalian vertebrates and these elements and scores being defined using
625 such species. State 54 was consistently enriched for all the constrained elements (4-7 fold), but
626 did not show high mean base-wise scores particularly for the GERP++ and PhyloP scores. This
627 difference of high enrichment in constrained elements but not base-wise scores is consistent
628 with state 54 having high alignability through most vertebrates, but low matching outside
629 primates. More generally, we found that constrained element calls did not have the resolution to
630 exhibit biologically relevant single nucleotide variation in enrichments around regulatory motifs
631 and exon start and ends as we saw with our conservation state annotations, with the exception
632 of those from PhastCons (**Figures 3** and **S17**).

633 The objective of our conservation state annotations is different than that of binary calls
634 and univariate scores of evolutionary constraint, which have a more specific and complementary
635 goal. However, to better understand their relative biologically relevant information we compared
636 their ability to recover annotated starts and ends of exons and TSS and TES of genes
637 separately for protein coding and pseudogenes, as there are well established genome
638 annotations of these features (**Figures 5A-C** and **S18**). In almost all cases the conservation
639 states had greater information available for recovering annotated gene features. The only
640 exceptions were that PhyloP scores could achieve higher precision at low recall levels for

641 protein coding exon starts and ends, and that SiPhy-pi elements had slightly higher precision for
642 TSS of protein coding genes at their one recall point.

643 We also compared the ability of conservation states to recover bases covered by DHS,
644 both genome-wide and restricted to non-exonic bases, and repeated these analyses when also
645 restricting to bases distal to a TSS (**Figures S19** and **S20, Methods**). When considering DHS
646 bases in aggregate over 53 cell and tissue types both genome-wide and restricting to non-
647 exonic bases, we found that at the same recall level the conservation states could identify bases
648 in a DHS at greater precision than all constraint scores considered and PhastCons constrained
649 elements. GERP++, SiPhy-pi and SiPhy-Omega elements did have higher precision at their
650 single recall point (**Figure S19**). Similar results were seen when just considering distal regions,
651 except for some of the scores in the non-exonic comparison at very low recall levels. The
652 relative precision at the same recall levels between conservation states and the GERP++,
653 SiPhy-pi and SiPhy-Omega constrained element sets did not hold for all cell types (**Figure S20**).
654 The increase in precision for those constrained element sets in the aggregate evaluation over
655 constraint scores, PhastCons elements, and ConsHMM annotations might be related to the
656 coarser resolution at which they were defined (**Figure S17**). We note that the information about
657 DHS in the conservation states was complementary to that in constrained element sets, as
658 evidenced by the substantial variation in DHS enrichments of bases within constrained elements
659 depending on their conservation state (**Figures 5D, S21** and **S22**). For example, bases in
660 PhastCons constrained elements falling in 35 different states were depleted for Fetal Brain DHS
661 in non-exonic regions, covering 10% of PhastCons bases, while bases in PhastCons elements
662 in 12 other states were over 5-fold enriched, covering 37% of PhastCons bases. Additionally,
663 we saw cases where certain states had greater enrichments for DHS for their bases not in a
664 constrained element compared to bases in a constrained element in other states. On the other
665 hand, constrained element calls offered additional information, as we observed that in most

666 cases, for a given conservation state, bases that were in a constrained element call had greater
667 enrichment for DHS than those that were not.

668 We also analyzed the enrichments of our conservation states for previously defined
669 nine-subsets of PhastCons constrained non-exonic elements (CNEEs) based on a directed
670 phylogenetic approach that assigned each element to a phylogenetic branch point of origin
671 (**Figure S23A**).[22] This demonstrated the heterogeneous nature of some of the resulting
672 assignments when relying on directed phylogenetic partitioning approaches. For example,
673 bases in elements assigned to originating at the branch point of the Tetrapod clade showed a
674 high enrichment (37 fold) for state 2, as would be expected since state 2 is associated with
675 aligning and matching through all vertebrates except fish, but an even greater enrichment (51
676 fold) for state 100, associated with putative artifacts (**Figure S23A**). We also evaluated whether
677 within non-exonic regions any subset of CNEE assigned to a specific clade exhibits enrichments
678 comparable to enrichments seen with the conservation states for CpG islands within non-exonic
679 regions (**Figure S23C**). The most enriched subset of CNEE bases was only 6.7 fold enriched
680 compared to the 37.6 fold enrichment observed for state 28 in non-exonic regions, and only
681 covered 1.9% of non-exonic CpG island bases compared to 12.8% of such bases covered in
682 state 28. A similar pattern of enrichments was observed when considering only the CNEEs
683 overlapping a PhastCons element called on the same alignment as the conservation states
684 (**Figure S23B** and **S23D**). These results highlight that the conservation states are able to
685 capture additional biologically relevant information present in the alignment that is not captured
686 by directed phylogenetic branch assignments of constrained elements.

687

688 *Bases prioritized by different variant prioritization scores have distinct conservation state*
689 *enrichment patterns*

690 In addition to scores defined based on just interspecies constraint, a variety of other
691 scores have been proposed to prioritize variants, including some based on intra-species

692 constraint or integrating inter-species constraint with other genomic annotations. A number of
693 these scores are widely used, even though a systematic understanding of different types of
694 bases prioritized by various scores is generally lacking. We leveraged the conservation state
695 annotations to more systematically understand the bases prioritized by a variety of different
696 scores in terms of their underlying pattern of conservation.

697 Specifically, we analyzed the conservation states' genome-wide enrichments of bases
698 prioritized by 12-different scores (CADD (v1.4), CDTS, DANN, Eigen, Eigen-PC, FATHMM-XF,
699 FIRE, fitCons, GERP++, PhastCons, PhyloP, and REMM) to be in the top 1, 5, and 10% of the
700 genome as well as the enrichment specifically in non-coding regions for those 12-scores and
701 two additional ones defined only on non-coding regions, LINSIGHT and FunSeq2 (**Figures 6A,**
702 **6B** and **S24-S27**).[8,9,13,16,18,19,37,45,50–54] We observed an overall strong enrichment for bases
703 prioritized by most scores for a specific set of conservation states. For example, the top 1%
704 CADD bases showed a 77.2 fold enrichment for state 1, amounting to 46% of the top 1% CADD
705 bases falling in this state. This enrichment was greater than that observed for any interspecies
706 constraint score, despite the CADD score being defined on a diverse set of genomic
707 annotations, including many non-conservation based annotations. There was a general
708 consistency in states with higher enrichment across the various measures. For example, when
709 considering the top 1% bases for the genome-wide analysis, the set of states that were among
710 the top five most enriched by at least one of the 12 scores contained only 13 states. Nine of
711 these 13 states (states 1-5, 7, 28, 54, 100) were in the top five for at least three scores.
712 However, there were important differences for the scores in the relative enrichment among
713 these top states, and in several cases a single score prioritized other states.

714 One interesting result was the wide disagreement among the scores of the relative
715 importance of state 2, the most enhancer enriched state, and state 28, the most promoter
716 enriched state, particularly in non-coding regions. For example, when considering top 1% bases
717 in non-coding regions, state 2 was the second or third most enriched state for CADD, Eigen,

718 FATHMM-XF, GERP++, LINSIGHT, PhastCons, PhyloP, and REMM prioritized variants, with
719 fold enrichments in the range of 24.9-47.2. On the other hand, state 28 was not one of the top
720 five most enriched states for any of those scores and its enrichments ranged between 0.3-6.2.
721 In contrast, for CDTS, DANN, and Eigen-PC state 2 only had enrichments between 0.8-2.1,
722 while state 28 was the first or second most enriched state for each score, with enrichments
723 ranging from 7.6-18.6. Also surprising was that DANN showed a depletion for state 2, which
724 showed high matching through all vertebrates except fish, but enriched for states that were only
725 associated with subsets of primates. For example, states 92, 93, and 96 did not have an
726 alignment frequency greater than 0.05 for any species past Gibbon, but were among the top five
727 states with the greatest enrichments for DANN prioritized variants, with enrichments in the
728 range 4.2-5.8. None of the other 13 scores considered showed enrichment for these states. This
729 is despite DANN using the same overall framework as CADD except using a deep neural
730 network, and previously reporting to be better able to predict the simulated variants used to train
731 CADD[51]. We verified that this difference with CADD also held for the original version of the
732 CADD score that used the same features as DANN (**Figure S25**). FitCons and FunSeq2 had
733 more balanced and relatively lower maximum enrichments for states 2 and 28. The FIRE score
734 was an outlier in that the maximum enrichment it had for any conservation state was only 2.8.
735 States for which FIRE prioritized bases showed the greatest enrichment included states such as
736 77, 80, and 85, which only showed substantial alignments among primates. The FIRE score
737 was trained based on predicting expression quantitative trait loci (eQTL) in lymphoblastoid cell
738 lines, which is a very different training objective than the other scores considered. It was
739 previously noted that this led to background selection being the most important feature to this
740 score.[50]

741 There were also strong disagreements about the relative importance of other states
742 across scores. State 100, associated with likely alignment artifacts, was one such state. For
743 example, at the top 1% threshold for the non-coding genome analysis, the state was among the

744 most enriched states for FATHMM-XF, FitCons, PhastCons and PhyloP with enrichments in the
745 range 14.7-34.5, while other scores showed more modest enrichments or even depletions,
746 highlighting differences in the vulnerability of each score to likely alignment artifacts. State 54,
747 which associated with high aligning through most vertebrates but not matching, was another
748 state with wide disagreement among scores in the importance of those bases, particularly in the
749 genome-wide analysis. At the top 1% threshold in the genome-wide analyses, this was the third
750 most enriched state based on CDTS, Eigen-PC, FIRE, and fitCons, with the enrichment for
751 fitCons reaching 21.1. In contrast, state 54 was depleted for the top GERP++ and REMM bases.
752 The high enrichment of fitCons for these bases is expected, as the features it considers lack the
753 resolution to differentiate the third codon position from the more conserved first and second
754 codon positions when scoring coding regions. There were also differences in the relative
755 importance given to state 1. For example, when considering variants genome-wide at the top
756 1% threshold, nine scores had the strongest enrichment for this state, but EIGEN, EIGEN-PC,
757 and FIRE did not. EIGEN-PC did show the strongest enrichment for state 1 at other thresholds
758 and EIGEN did when restricting to non-coding genome at all thresholds. However, their
759 inconsistent ranking of state 1 is likely reflective of the unsupervised prioritization scheme used
760 by these scores. Overall, these results show that the ConsHMM state annotations provides
761 insights into key differences in variants prioritized by various scores by systematically and in an
762 unbiased way capturing biologically diverse classes of nucleotides at single nucleotide
763 resolution.
764
765 *Enrichment of conservation states for human genetic variation*
766 Previous analyses have found a depletion of human genetic variation in evolutionarily
767 constrained elements.[7] Consistent with that, the greatest depletion (3.3 fold depletion) of
768 common SNPs from dbSNP is in state 1, the state most enriched for constrained elements.
769 Interestingly, six states, A_SMam states 55-57 and AM_Prim states 87-89, had enrichments in

770 the range 5 to 8 fold for common SNPs. These were also the six states with greatest enrichment
771 of CG dinucleotides (**Figure S12**). These six states have in common that they show high align
772 probabilities for most primates, but low match probabilities for some of those same primates.
773 These states are thus associated with substantial variation both among primates and among
774 humans. We observed similar patterns of enrichment for variants identified from whole genome
775 sequencing (WGS) of a cohort of 7784 unrelated individuals[37], with the levels of state
776 enrichments and depletions increasing with the minor allele frequency (**Figure S28**).
777 When analyzing the enrichment of GWAS catalog variants[36] relative to the background
778 of common SNPs we saw opposite enrichment patterns for these states (**Figure 6C** and **6D**).
779 For example, relative to this background, state 1 was most enriched for GWAS catalog variants,
780 which is consistent with previous observations of constrained elements enriching for GWAS
781 variants.[7] On the other hand, states 55-57 and 87-89 showed the greatest depletion. These
782 results suggest that common variants are less likely to be phenotypically significant if they fall in
783 conservation states most enriched for common genetic variation.
784
785 *Constrained element enrichment for partitioned heritability of complex traits depends on*
786 *conservation state*
787 Previous analyses have suggested a strong enrichment of constrained elements and
788 DHS for phenotype heritability.[15,55] As we saw large differences in DHS enrichments of
789 constrained elements depending on the conservation state, we investigated the extent to which
790 constrained elements in conservation states most enriched for DHS enriched for phenotype
791 heritability compared to the remaining states. Specifically, we ranked the conservation states in
792 descending order of their median enrichment for DHS from a compendium of 123 experiments
793 from the ENCODE consortium, within the non-exonic portion of the state (**Figure 2B,**
794 **Methods**).[3] We then partitioned bases in PhastCons constrained elements into two almost
795 equal size sets based on whether they overlapped one of the top seven ranked conservation

796 states (states 1-5, 8, 28) or not. We then computed the heritability for these two sets for eight
797 phenotypes in the context of a set of baseline annotations that include DHS annotations
798 (**Methods**).[15] For seven of the phenotypes, we found that bases in constrained elements
799 overlapping the top seven states had greater enrichment than those in the remaining 93 states,
800 often substantially so (**Figure 6E**). These results suggest additional value in the conservation
801 state annotations for isolating more likely disease relevant variants.
802
803 **Discussion**
804 We presented a framework for genome annotation based on comparative genomics
805 sequence data. Our approach learns a set of conservation states *de novo* using a multivariate
806 HMM based on the combinatorial and spatial patterns of which species align and match a
807 reference genome in a multi-species DNA sequence alignment. We applied this approach to
808 annotate the human genome at single nucleotide resolution into one of 100 conservation states.
809 Conservation state annotations exhibited substantial enrichments for a wide range of other
810 genomic annotations that were not provided to the model in training, thus supporting their
811 biological significance. Specific conservation states exhibited strong enrichments for various
812 gene annotations including exons, TSS and TES of genes, while others showed strong
813 enrichments for specific types of repeat elements. Conservation states showed differential
814 enrichment patterns for various classes of genes and DHS from multiple cell types, even though
815 they were defined independently of any functional genomics data. Specific conservation states
816 exhibited enrichments for common human variants, while a different set of states exhibited
817 enrichments for variants identified by GWAS relative to common variation.
818 ConsHMM differs from other comparative genomics based annotation approaches in
819 several respects. One difference is that it takes an unsupervised approach that does not
820 explicitly use a phylogenetic tree, except to the extent to which a phylogenetic tree was used to
821 generate the input multi-species sequence alignment. This leads to relatively unbiased, simple

822 and interpretable models. However, many state patterns discovered are consistent with
823 expected observations from commonly assumed phylogenetic relationships of the species.
824 While states' parameters often decreased with divergence time from human, there were a
825 number of exceptions. Some of these exceptions corresponded to missing specific sub-clades
826 of species, particularly those with long branch lengths. For example, a number of states were
827 not represented by mouse and rat, while being represented by more distally diverged mammals.
828 Other exceptions isolated putative artifacts in the alignments that might otherwise confound
829 analyses, as we saw for two states heavily enriched for pseudogenes. ConsHMM also differs
830 from other commonly used modeling approaches in how it explicitly differentiates non-aligning
831 bases from aligning non-matching bases, which allowed it, for example, to identify states
832 particularly associated with third codon positions. Another difference between the ConsHMM
833 annotations and standard constraint measures is that the ConsHMM annotations are defined
834 directly relative to the variant present in the genome being annotated. When applying
835 ConsHMM to annotate the human genome, a mutation unique to human would be expected to
836 have a much larger effect on the ConsHMM annotations than a mutation unique to a single
837 other species. This would not in general be expected to be the case for constraint measures
838 that do not differentiate the target genome for annotation from other genomes in an alignment.
839 An interesting future direction would be to produce and analyze individual specific ConsHMM
840 annotations.

841 Our conservation state annotation is complementary to existing binary calls and scores
842 of evolutionary constraint based on phylogenetic modeling. Both locations called as constrained
843 and those called as non-constrained are heterogeneous in their assigned conservation state.
844 Our annotations thus provide additional descriptive information about the conservation patterns
845 at each base. In terms of information for predicting external annotations, we found that in many
846 cases the conservation states had greater information than constraint scores or elements.
847 Notably, our modeling approach identified a conservation state, state 28, associated with a

848 pattern of aligning and matching some mammalian and non-mammalian vertebrates, but not
849 with high probability for any one species. This conservation state strongly enriched for
850 transcription start sites and CpG islands, and was not well captured by phylogenetic modeling
851 approaches. For other cases, such as DHS, the relative information depended on the
852 constrained element set or score being compared. Importantly, we observed that DHS
853 information provided by the states was complementary to information in the constrained element
854 calls. We also used the conservation state annotations to systematically understand key
855 similarities and differences in the patterns of conservation in bases prioritized by a large number
856 of different variant prioritization scores, including scores based on integrating diverse features
857 (**Figure 6A** and **6B**). The conservation state annotations provide a powerful framework for
858 gaining such an understanding, since the corresponding conservation patterns are defined
859 systematically in an unbiased way, at single nucleotide resolution and capture a diverse set of
860 biological features. Furthermore, we observed that bases in constrained elements showed
861 substantially different enrichments for phenotype-associated heritability, depending on their
862 conservation state.

863 The conservation states are both inspired by, and provide complementary information to,
864 existing chromatin state annotation approaches. While the states from the two approaches are
865 based on very different data and have fundamental differences, they also exhibited substantial
866 cross-enrichments. In general, conservation states have the advantage of providing information
867 at single nucleotide resolution, which we demonstrated by showing enrichments patterns in and
868 around coding exons and regulatory motifs. Conservation states can also provide information
869 about bases in the genome even if the relevant cell type has not been experimentally profiled,
870 while chromatin states have the advantage of directly providing cell type specific information.

871 We expect many applications for the methodology and annotations we have presented
872 here. While we applied ConsHMM here to one multiple species alignment, a 100-way Multiz
873 human alignment, the methodology is general, and thus can be readily applied to alignments to

874 other species or alignments generated by other methods.[26] The annotations we produced serve
875 as a resource to directly interpret other genomic datasets or variant prioritization scores. They
876 could also potentially be integrated with other complementary genomic annotations in methods
877 that produce integrated variant prioritization scores. This work represents a step in the direction
878 of improving whole genome annotations, which will continue to be of increasing importance
879 towards understanding health and disease as the availability of whole genome sequencing data
880 increases.
881
882 **Supplemental Data**
883 Supplemental Data include twenty-eight figures and three tables.
884
885 **Conflicts of Interest**
886 The authors declare that they have no conflicts of interest.
887
888 **Acknowledgements**
889 We thank Ewan Birney and members of the Ernst lab for useful discussions. We
890 acknowledge funding from US National Institutes of Health grants DP1DA044371,
891 R01ES024995, U01HG007912 and U01MH105578 (J.E.), and T32CA201160 (A.S.), US
892 National Science Foundation CAREER Award #1254200 (J.E.), a Kure-IT award and an Alfred
893 P. Sloan Fellowship (J.E.).
894
895 **Web Resources**
896 25-state chromatin state annotations:
897 http://compbio.mit.edu/roadmap
898 CADD score v1.0:
899 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 Lowe et al.[22]:

http://www.stanford.edu/~lowec/data/threePeriods/hg19cnee.bed.gz

ConsHMM v1.0 software and ConsHMM state annotations:

https://github.com/ernstlab/ConsHMM

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:

LINSIGHT score:

http://compgen.cshl.edu/~yihuang/tracks/LINSIGHT.bw

Motif instances and background:

http://compbio.mit.edu/encode-motifs/

https://www.ebi.ac.uk/gwas/

Multiz 100-way alignment to hg19 reference:

http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/

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

References

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (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.

2. Ward, L.D., and Kellis, M. (2012). Interpreting non-coding variation in complex disease genetics. Nat. Biotechnol. *30*, 1095–1106.

953 3. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in
954 the human genome. Nature *489*, 57–74.

955 4. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B.,
956 Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of
957 chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

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

961 6. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward,
962 L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA
963 elements in the human genome. Proc. Natl. Acad. Sci. *111*, 6131–6138.

964 7. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S.,
965 Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of
966 human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

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

969 9. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K.,
970 Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily
971 conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*,
972 1034–1050.

973 10. Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding
974 disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

975 11. Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., De Franco, E., Caswell, R.,
976 Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., Allen, H.L., et al. (2014).
977 Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis.
978 Nat. Genet. *46*, 61–64.

979 12. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.
980 (2005). Distribution and intensity of constraint in mammalian genomic sequence.
981 Genome Res. *15*, 901–913.

982 13. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S.
983 (2010). Identifying a High Fraction of the Human Genome to be under Selective
984 Constraint Using GERP++. PLoS Comput. Biol. *6*, e1001025.

985 14. Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009).
986 Identifying novel constrained elements by exploiting biased substitution patterns.
987 Bioinformatics *25*, i54–i62.

988 15. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R.,
989 Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional
990 annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–
991 1235.

992 16. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J.
993 (2014). A general framework for estimating the relative pathogenicity of human genetic
994 variants. Nat. Genet. *46*, 310–315.

995 17. Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation
996 of noncoding sequence variants. Nat. Methods *11*, 294–296.

997 18. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral
998 approach integrating functional genomic annotations for coding and noncoding variants.
999 Nat. Genet. *48*, 214–220.

1000 19. Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of
1001 deleterious noncoding variants from functional and population genomic data. Nat.
1002 Genet. *49*, 618–624.

1003 20. Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper,
1004 D.N., Bernstein, J.A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants
1005 of uncertain significance in clinical exomes at high sensitivity. Nat. Genet. *48*, 1581–
1006 1586.

1007 21. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans,
1008 M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2015). The UCSC
1009 Genome Browser database: 2015 update. Nucleic Acids Res. *43*, D670-681.

1010 22. Lowe, C.B., Kellis, M., Siepel, A., Raney, B.J., Clamp, M., Salama, S.R., Kingsley,
1011 D.M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regulatory innovation
1012 during vertebrate evolution. Science *333*, 1019–1024.

1013 23. Siepel, A., Pollard, K.S., and Haussler, D. (2006). New Methods for Detecting
1014 Lineage-Specific Selection. In Research in Computational Molecular Biology, (Springer,
1015 Berlin, Heidelberg), pp. 190–205.

1016 24. Kim, S.Y., and Pritchard, J.K. (2007). Adaptive Evolution of Conserved Noncoding
1017 Elements in Mammals. PLOS Genet. *3*, e147.

1018 25. Marnetto, D., Mantica, F., Molineris, I., Grassi, E., Pesando, I., and Provero, P.
1019 (2018). Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome
1020 Expansion. Am. J. Hum. Genet. *102*, 1–12.

1021 26. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella,
1022 A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics
1023 resources. Database J. Biol. Databases Curation *2016*,.

1024 27. Cotney, J., Leng, J., Yin, J., Reilly, S.K., DeMare, L.E., Emera, D., Ayoub, A.E.,
1025 Rakic, P., and Noonan, J.P. (2013). The Evolution of Lineage-Specific Regulatory
1026 Activities in the Human Embryonic Limb. Cell *154*, 185–196.

1027 28. Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park,
1028 T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer Evolution
1029 across 20 Mammalian Species. Cell *160*, 554–566.

1030 29. Don, P.K., Ananda, G., Chiaromonte, F., and Makova, K.D. (2013). Segmenting the
1031 human genome based on states of neutral genetic divergence. Proc. Natl. Acad. Sci.
1032 *110*, 14699–14704.

1033 30. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states
1034 for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

1035 31. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery
1036 and characterization. Nat. Methods *9*, 215–216.

1037 32. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S.
1038 (2012). Unsupervised pattern discovery in human chromatin structure through genomic
1039 segmentation. Nat. Methods *9*, 473–476.

1040 33. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M.,
1041 Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning Multiple
1042 Genomic Sequences With the Threaded Blockset Aligner. Genome Res. *14*, 708–715.

1043 34. Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.

1044 35. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F.,
1045 Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference
1046 human genome annotation for The ENCODE Project.Genome Res. *22*,1760–1774.

1047 36. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A.,
1048 Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated
1049 resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001-1006.

1050 37. Iulio, J. di, Bartha, I., Wong, E.H.M., Yu, H.-C., Lavrenko, V., Yang, D., Jung, I.,
1051 Hicks, M.A., Shah, N., Kirkness, E.F., et al. (2018). The human noncoding genome
1052 defined by genetic diversity. Nat. Genet. *50*, 333-337.

1053 38. Witowski, V., and Foraita, D.R. (2014). HMMpa: Analysing accelerometer data using
1054 hidden Markov models.

1055 39. Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for
1056 systematic annotation of diverse human tissues. Nat. Biotechnol. *33*, 364–376.

1057 40. Hahsler, M. and Buchta, C.(2017). cba: Clustering for Business Analytics.

1058 41. Bar-Joseph, Z., Gifford, D.K., and Jaakkola, T.S. (2001). Fast optimal leaf ordering
1059 for hierarchical clustering. Bioinforma. Oxf. Engl. *17 Suppl 1*, S22-29.

1060 42. Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time
1061 series gene expression data. BMC Bioinformatics *7*, 191.

1062 43. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of
1063 regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–
1064 2987.

1065 44. Kolde, R. (2015). pheatmap: Pretty Heatmaps.

1066 45. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein,
1067 M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in
1068 cancer. Genome Biol. *15*, 480.

1069 46. Chen, X., and Tompa, M. (2010). Comparative assessment of methods for aligning
1070 multiple genome sequences. Nat. Biotechnol. *28*, 567–572.

1071 47. Zhang, M.Q. (1998). Statistical features of human exons and their flanking regions.
1072 Hum. Mol. Genet. *7*, 919–932.

1073 48. Sarda, S., Das, A., Vinson, C., and Hannenhalli, S. (2017). Distal CpG islands can
1074 serve as alternative promoters to transcribe genes with silenced proximal promoters.
1075 Genome Res. *27*, 553–566.

1076 49. Litman, G.W., Anderson, M.K., and Rast, and J.P. (1999). Evolution of Antigen
1077 Binding Receptors. Annu. Rev. Immunol. *17*, 109–147.

1078 50. Ioannidis, N.M., Davis, J.R., DeGorter, M.K., Larson, N.B., McDonnell, S.K., French,
1079 A.J., Battle, A.J., Hastie, T.J., Thibodeau, S.N., Montgomery, S.B., et al. (2017). FIRE:
1080 functional inference of genetic variants that regulate gene expression. Bioinforma. Oxf.
1081 Engl. *33*, 3895–3901.

1082 51. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for
1083 annotating the pathogenicity of genetic variants. Bioinforma. Oxf. Engl. *31*, 761–763.

1084 52. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R., and Campbell, C.
1085 (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended
1086 features. Bioinforma. Oxf. Engl. *34*, 511–513.

1087 53. Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating
1088 probabilities of fitness consequences for point mutations across the human genome.
1089 Nat. Genet. *47*, 276–283.

1090 54. Smedley, D., Schubach, M., Jacobsen, J.O.B., Köhler, S., Zemojtel, T., Spielmann,
1091 M., Jäger, M., Hochheiser, H., Washington, N.L., McMurry, J.A., et al. (2016). A Whole-

1092 Genome Analysis Framework for Effective Identification of Pathogenic Regulatory
1093 Variants in Mendelian Disease. Am. J. Hum. Genet. *99*, 595–606.

1094 55. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C.,
1095 Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning Heritability of
1096 Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. Am. J. Hum.
1097 Genet. *95*, 535–552.

1098

1099
1100
1101
1102 **Figure Legends**

1103 **Figure 1: Illustration of ConsHMM modeling approach. (A)** The input to ConsHMM is a
1104 multi-species alignment, which is illustrated for a subset of 6 species aligned to the human
1105 sequence. At each position and for each species ConsHMM represents the information as one
1106 of three observations: (1) aligns with a non-indel nucleotide matching the human sequence
1107 shown in blue, (2) aligns with a non-indel nucleotide not matching the human sequence shown
1108 in yellow, or (3) does not align with a non-indel nucleotide shown in gray. **(B)** Illustration of
1109 conservation state assignments at a locus chr22:25,024,640-25,024,812. Only states assigned
1110 to at least one nucleotide in the locus are shown. Below the conservation state assignments is a
1111 color encoding of the input multiple sequence alignment according to panel (A). The major clade
1112 of species as annotated on the UCSC genome browser[21] are labeled and ordered based on
1113 divergence from human. Above the conservation state assignments are PhastCons constrained
1114 elements and scores and PhlyoP constraint scores. This figure and **Figure S9** together illustrate
1115 that positions of nucleotides that have the same status in terms of being in a constrained
1116 element or not or have similar constraint scores can be assigned to different conservation states
1117 depending on the patterns in the underlying multiple-species alignment.

1118

1119 **Figure 2: Conservation state emission parameters learned by ConsHMM and enrichments**
1120 **for other genomic annotations. (A)** Each row in the heatmap corresponds to a conservation
1121 state. For each state and species, the left half of the heatmap gives the probability of aligning to
1122 the human sequence, which is one minus the probability of the not aligning emission.
1123 Analogously, the right half of the heatmap gives the probability of observing the matching
1124 emission. Each individual column corresponds to one species with the individual names
1125 displayed in **Figure S5.** For both halves, species are grouped by the major clades and ordered
1126 based on the hg19.100way.nh phylogenetic tree from the UCSC genome browser, with species
1127 that diverged more recently shown closer to the left.[21] The conservation states are ordered
1128 based on the results of applying hierarchical clustering and optimal leaf ordering.[41] The states
1129 are divided into eight major groups based on cutting the dendrogram of the clustering. The
1130 groups are indicated by color bars on the left hand side and a white row between them.
1131 Transition parameters between states of the model can be found in **Figure S6**. **(B)** The columns
1132 of the heatmap indicate the relative enrichments of conservation states for external genomic
1133 annotations (**Methods**). For each column, the enrichments were normalized to a [0,1] range by
1134 subtracting the minimum value of the column and dividing by the range and colored based on
1135 the indicated scale. Values for these enrichments and additional enrichments can be found in
1136 **Figure S8** and **Table S2** and enrichments for individual repeat classes and families can be
1137 found in **Figure S14.**
1138
1139 **Figure 3: Conservation state positional enrichments.** Plots of positional fold enrichments of
1140 conservation states relative to **(A)** start of exons of protein coding genes in phase 0, **(B)** end of
1141 exons of protein coding genes and **(C,D)** TSS of **(C)** protein coding, and **(D)** pseudogenes
1142 genes. Positive values represent the number of bases downstream in the 5' to 3' direction of
1143 transcription, while negative values represent the number of bases upstream. Enrichments
1144 relative to gene annotations are based on a genome-wide background. The subset of states

1145 included in panels (A)-(D) were the states that had at least a 3 fold enrichment at some position
1146 within +/-2kb from the anchor point. **(E,F)** Also shown are positional plots relative to the central
1147 nucleotide of a set of instances of **(E)** STAT and **(F)** POU5F1 motifs. The subset of states
1148 included in (E), (F) are the states that had an enrichment of at least 1.5 for some position within
1149 +/-15bp from the center nucleotide of either motif. Enrichments for motif instances were
1150 computed relative to the portion of the genome scanned for regulatory motifs in Kheradpour and
1151 Kellis (2014), which excludes coding, 3'UTRs, and repeat elements. Additional position
1152 enrichment plots can be found in **Figure S11**.
1153
1154 **Figure 4: Conservation states enrichment for chromatin states, GO terms, DHS and**
1155 **repeat elements. (A)** Median fold enrichment of conservation states (rows) for one of 25
1156 chromatin states from a previously defined chromatin state model defined across 127 samples
1157 of diverse cell and tissue types (columns).[15] Only conservation states that had the maximum
1158 value for at least one chromatin state are shown, and those values are boxed. See **Figure S15**
1159 for the enrichments of all conservation states. **(B)** $-\log_{10}$ p-value (uncorrected) of the
1160 conservation states (rows) for the GO term (columns) where each conservation state is
1161 associated with its top 5% genes based on promoter regions (**Methods**). Only GO terms which
1162 were the most enriched term for some conservation state are shown, restricted to the top 10
1163 terms based on the significance of the enrichment. Only conservation states that had the most
1164 significant enrichment for one of the displayed GO terms are shown, with the maximal
1165 enrichments boxed. The full set of conservation states with additional GO terms are in **Figure**
1166 **S13**. **(C)** Relative enrichments of conservation states for DHS across cell and tissue types. Only
1167 conservation states with at least a 2 fold enrichment in one sample considered are shown.
1168 Enrichment values were $\log_2$ transformed and then row normalized by subtracting the mean
1169 (right heatmap) and dividing by the standard deviation. States and experiments were then
1170 hierarchically clustered and revealed two major clusters. In the top cluster conservation states

1171 showed the greatest enrichment for experiments in which the DHS also strongly enriched for
1172 CpG islands (top heatmap). In the bottom cluster conservation states generally had the
1173 strongest relative preference for a number of fetal related samples. **(D)** Enrichment of
1174 conservation states with the maximal enrichment for LINE, SINE, LTR or DNA repeats next to
1175 the state align probabilities for primates. These states all had low align probabilities outside of
1176 primates, but their differences among primates corresponded to substantial differences in repeat
1177 enrichments.
1178
1179 **Figure 5**: **Relationship of conservation states with constrained elements and scores.**
1180 Precision-recall plots for recovery of **(A)** TSS of protein coding genes, **(B)** TES of protein coding
1181 genes, and **(C)** the start of exons of protein coding genes. Recovery based on ordering
1182 ConsHMM conservation states for their enrichment for the target set in the training data, then
1183 cumulatively adding the states in that ranked order and evaluating on the test data is shown with
1184 a series of blue dots (**Methods**). The first few conservation states added are labeled with their
1185 state number. Recovery based on ranking from highest to lowest value of constraint scores is
1186 shown with continuous lines. Recovery based on score partitioning into 400 bins and
1187 subsequent ordering based on enrichment for the target set in the training data, then
1188 cumulatively adding bins in that ranked order and evaluating on the test data is shown in a
1189 series of dots of the same color as the continuous line corresponding to the score. Recovery of
1190 target test bases by a constrained element set is shown with a single dot for each constrained
1191 element set. See **Figure S18-20** for plots based on additional targets. **(D)** The graph shows the
1192 fold enrichment for Fetal Brain DHS[5] within the non-exonic portion of each conservation state,
1193 separately for those bases in a PhastCons constrained element (pink) and bases not in such an
1194 element (blue). Enrichments within constrained elements varied substantially depending on the
1195 conservation state. For a given conservation state, bases in a constrained element had greater
1196 enrichments than bases not in a constrained element, illustrating complementary information of

1197 conservation states and constrained elements. See **Figure S21** for graphs based on different
1198 element sets or DHS data and **Figure S22** for these enrichments plotted against the size of the
1199 set.
1200
1201 **Figure 6: Conservation states and association with human genetic variation. (A)** Fold
1202 enrichments of bases ranked in the top 1% genome-wide by 12 variant prioritization scores.
1203 Only states that were among the top five most enriched states for at least one score are shown.
1204 The enrichment of the top five ranking states for each score is colored according to the ranking
1205 and the color scale shown on right. **(B)** Enrichments of bases ranked in the top 1% of the non-
1206 coding genome by 14 variant prioritization scores. The criteria for selecting states to display and
1207 coloring enrichments was the same as in panel (A). Enrichments at additional thresholds and for
1208 all states both genome-wide and for the non-coding genome are in **Figure S24-S27**. The
1209 enrichments for CADD shown here are based on v1.4, while enrichments based on the original
1210 version of CADD are also shown in **Figure S25-S27**. (**C**) The panel displays the $\log_2$ fold
1211 enrichment of each state for common SNPs (pink) and GWAS catalog variants relative to
1212 common SNPs (blue). State 1, associated with high alignability and matching through all
1213 vertebrates, showed the greatest depletion of common SNPs and the highest enrichment for
1214 GWAS variants relative to common SNPs. States 55-57 and 87-89 exhibited the opposite
1215 pattern having the greatest enrichment for common SNPs and the greatest depletion of GWAS
1216 variants relative to this background. The second most depleted state for common SNPs, which
1217 did not show enrichment for GWAS catalog SNPs, was state 96 which captured large gaps in
1218 the assembly (**Figure S8**). (**D**) Panel shows the representation of state emission parameters
1219 from **Figure 2A** for the subset of states highlighted in panel (C). The states with the greatest
1220 depletion of GWAS variants all had relatively high alignability at least through primates, but low
1221 matching probabilities for almost all species except a few closely related primates. (**E**) Applying
1222 the heritability partitioning enrichment method of Finucane et al.[15] on two disjoint subsets of

1223 bases in PhastCons elements, with eight phenotypes previously analyzed with heritability
1224 partitioning in the context of a baseline set of annotations (**Methods**).[15] One set of bases are
1225 those in PhastCons elements that are also in one of the seven conservation states showing the
1226 greatest enrichment for DHS in its non-exonic portion (States 1-5, 8, and 28) covering 51.9% of
1227 PhastCons bases (pink). The other set are those bases in PhastCons elements overlapping the
1228 remaining 93 states covering 48.1% of PhastCons bases (blue).

A

Human	T	T	T	C	C	T	G	A	C	T	T
Chimp	T	T	T	C	C	T	G	A	C	T	T
Bushbaby	T	C	T	G	C	T	T	C	C	T	T
Rat	-	C	T	T	C	T	G	A	A	T	T
Alpaca	-	-	-	C	C	T	T	G	C	A	T
Megabat	T	C	-	C	C	T	G	A	T	T	T
Parrot	-	-	-	-	-	-	-	-	-	-	-

Aligning and matching the human sequence

Aligning but not matching the human sequence

Not aligning to the human sequence

B

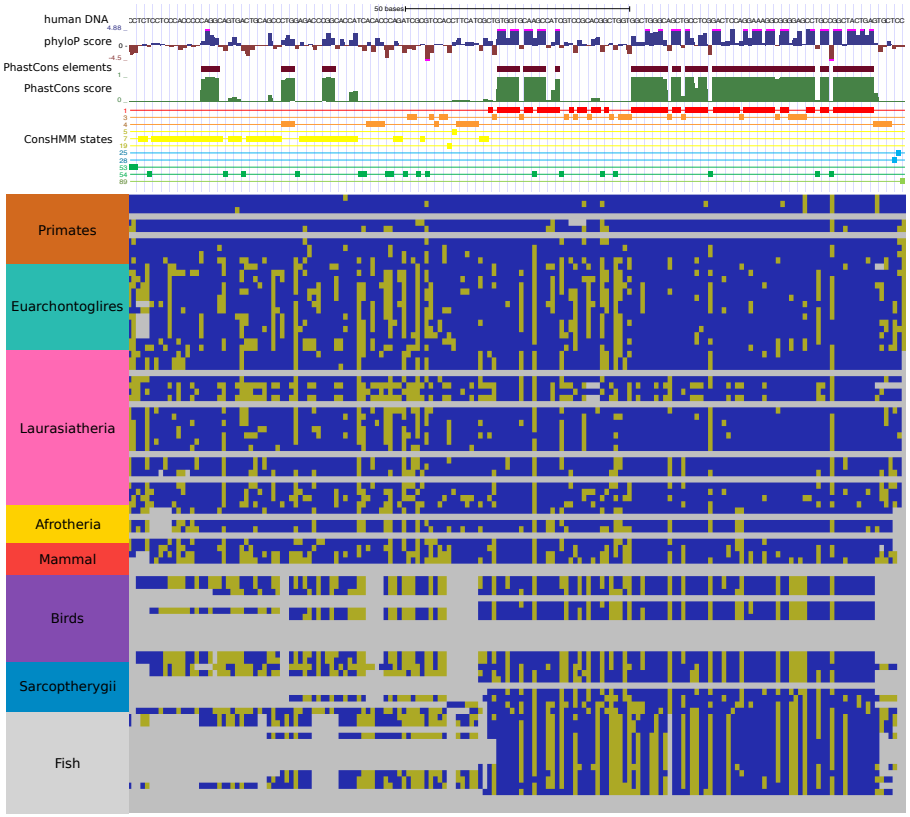


Figure 1

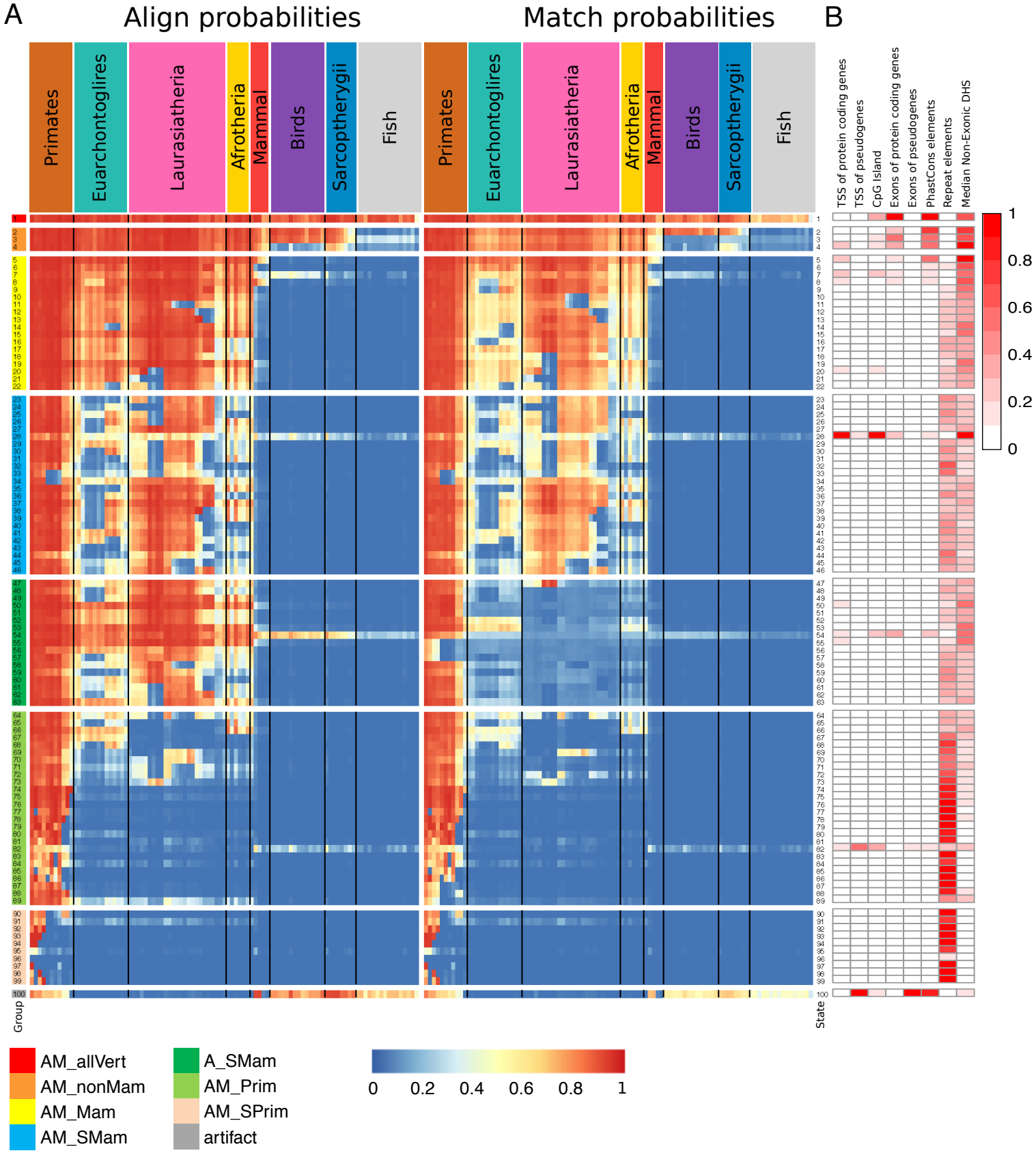

A
Align probabilities
Match probabilities
B
Primates
Euarchontoglires
Laurasiatheria
Afrotheria
Mammal
Birds
Sarcoptherygii
Fish
TSS of protein coding genes
TSS of pseudogenes
CpG Island
Exons of protein coding genes
Exons of pseudogenes
PhastCons elements
Repeat elements
Median Non-Exonic DHS
Group
State
AM_allVert
AM_nonMam
AM_Mam
AM_SMam
A_SMam
AM_Prim
AM_SPrim
artifact
0
0.2
0.4
0.6
0.8
1

Figure 2

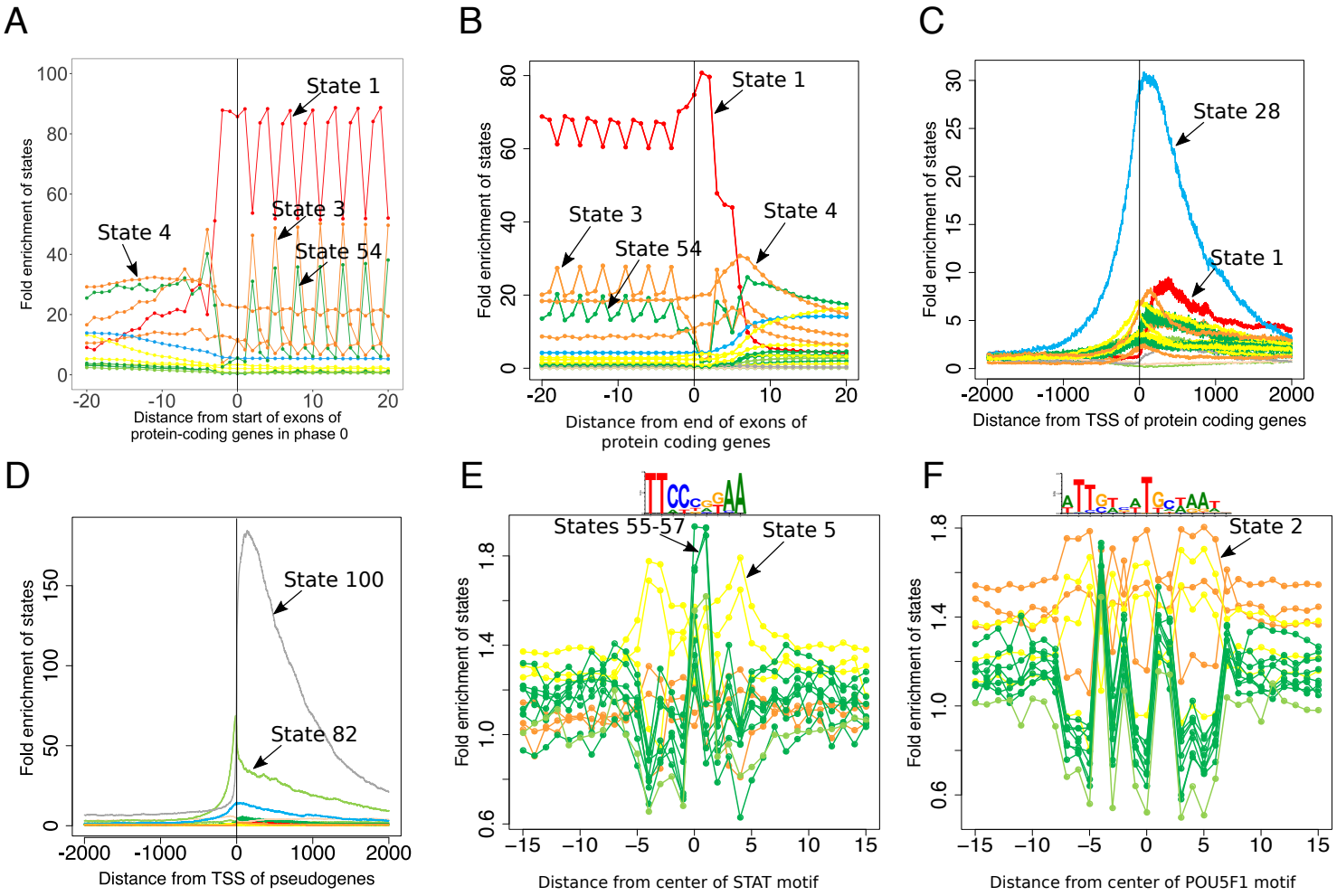

A
State 1
State 3
State 4
State 54
Fold enrichment of states
Distance from start of exons of protein-coding genes in phase 0
B
State 1
State 3
State 54
State 4
Fold enrichment of states
Distance from end of exons of protein coding genes
C
State 28
State 1
Fold enrichment of states
Distance from TSS of protein coding genes
D
State 100
State 82
Fold enrichment of states
Distance from TSS of pseudogenes
E
States 55-57
State 5
Fold enrichment of states
Distance from center of STAT motif
F
State 2
Fold enrichment of states
Distance from center of POU5F1 motif

Figure 3

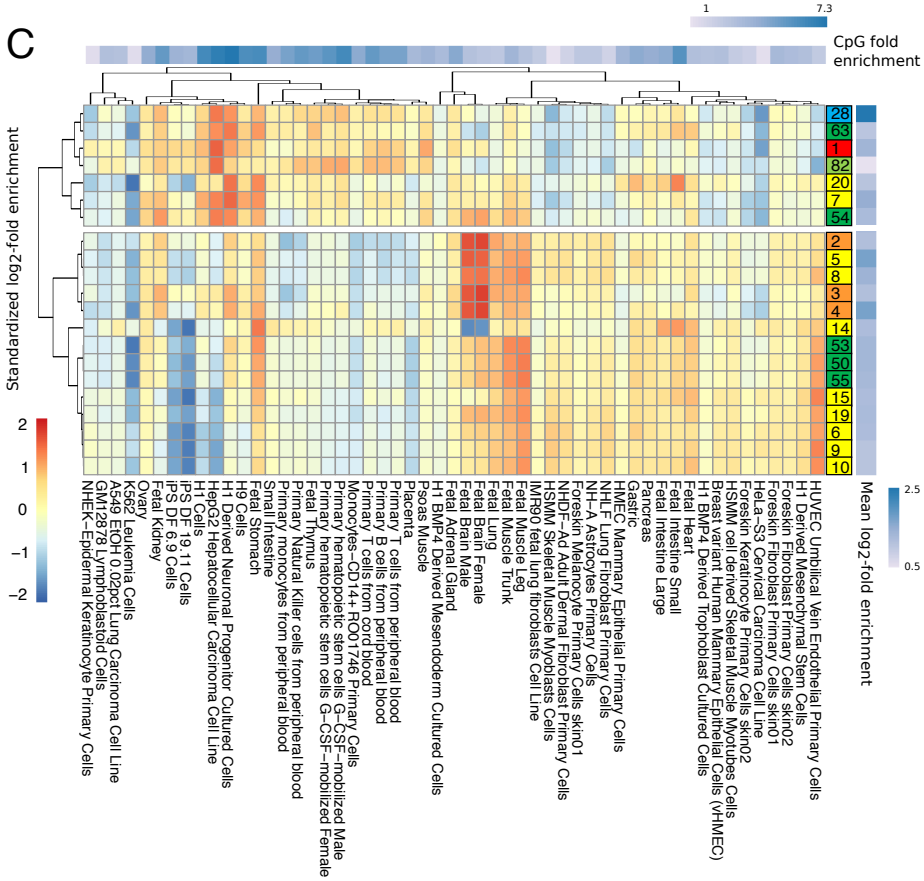
A

	1_TssA	2_PromU	3_PromD1	4_PromD2	5_Tx5	6_Tx	7_Tx2	8_TxWk	9_TxReg	10_TxEnh5	11_TxEnh3	12_TxEnhW	13_EnhA1	14_EnhA2	15_EnhAF	16_EnhW1	17_EnhW2	18_EnhAc	19_DNase	20_ZNF/Rpts	21_Het	22_PromP	23_PromBiv	24_ReprPC	25_Quies
1	2.1	3.4	5.4	3.8	0.3	8.6	8.7	1.6	6.7	1.3	8.3	0.8	1.4	1.6	1.0	1.2	1.0	0.5	1.0	4.0	0.3	1.6	6.3	1.6	0.5
2	1.0	2.5	1.1	1.7	0.3	2.0	2.8	1.2	2.7	1.6	4.3	1.1	3.5	4.7	3.1	3.3	3.4	1.2	1.9	0.4	0.1	0.7	3.4	1.6	0.8
5	3.6	3.5	3.8	2.7	0.6	1.1	1.3	1.0	3.9	2.3	2.6	1.7	3.3	4.2	2.9	3.1	2.9	1.6	2.5	0.2	0.1	1.0	3.8	2.0	0.9
14	1.7	3.2	2.7	2.3	1.3	1.4	1.5	1.2	2.0	2.4	1.9	2.2	2.4	2.3	1.9	2.4	1.9	1.3	2.0	0.3	0.2	0.8	2.5	1.7	0.9
28	35.1	4.5	19.9	2.4	0.5	2.2	2.4	1.0	3.6	1.3	2.0	0.6	1.3	1.3	0.8	1.3	0.7	0.9	1.1	5.4	1.4	8.0	13.1	1.9	0.6
77	0.5	0.4	0.3	0.8	2.1	1.2	1.4	2.0	0.2	0.4	0.5	0.6	0.6	0.3	0.6	0.4	0.4	1.6	0.4	1.1	1.0	1.2	0.2	0.8	0.9
82	4.2	1.2	1.7	0.9	0.9	1.6	1.6	1.3	0.9	0.7	1.0	0.5	0.4	0.6	0.5	0.8	0.4	1.1	1.1	6.4	3.5	4.8	2.1	2.1	0.8
86	0.1	0.2	0.0	0.2	0.7	0.4	0.3	0.9	0.1	0.1	0.1	0.1	0.3	0.2	0.4	0.3	0.2	1.0	0.3	1.9	2.5	0.7	0.1	0.4	1.1
92	0.2	0.2	0.1	0.3	0.7	0.4	0.4	0.9	0.3	0.2	0.3	0.2	0.7	0.4	0.6	0.4	0.4	3.0	0.5	1.9	3.1	1.2	0.2	0.5	1.1
93	0.2	0.3	0.1	0.1	0.5	0.2	0.3	0.7	0.1	0.2	0.2	0.1	0.7	0.3	0.6	0.4	0.4	2.5	0.5	2.0	3.9	1.4	0.3	0.8	1.1
100	0.4	0.4	0.2	0.2	0.7	1.5	0.9	0.7	0.1	0.3	0.4	0.5	0.3	0.7	0.6	0.5	0.4	0.5	1.4	8.4	2.8	1.9	1.2	1.5	1.0

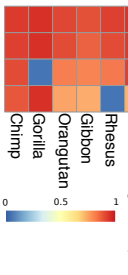
B

state	detection of chemical stimulus involved in sensory perception of smell	system development	nucleosome	cellular developmental process	regulation of cellular process	transcription from RNA polymerase II promoter	cell-cell signaling	calcium ion binding	receptor binding	antigen binding
1	0.3	14.1	41.3	5.4	4.2	7.9	7.9	12.2	0.0	0.0
2	0.0	41.2	0.1	30.6	6.9	19.0	11.0	3.6	7.5	0.0
3	319.6	2.6	0.2	0.1	18.6	0.1	2.8	13.5	0.0	0.0
4	0.0	21.0	0.0	17.7	5.6	24.7	8.1	0.1	6.5	0.0
5	0.0	26.8	0.0	18.9	10.0	13.4	20.8	0.5	11.3	0.1
28	60.0	5.8	0.9	5.9	24.9	12.5	10.5	0.9	1.3	1.4
50	0.0	29.6	0.0	21.7	5.6	4.7	12.6	2.9	15.8	0.1
53	0.0	44.6	0.0	30.2	7.9	9.2	18.8	2.1	15.4	0.3
54	53.2	3.9	32.3	0.9	1.7	0.9	2.8	15.9	0.2	0.0
89	3.4	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	14.5

C



D



	LINE	SINE	LTR	DNA
74	1.0	1.8	1.8	5.6
76	1.5	2.1	3.3	2.0
77	0.9	4.5	1.1	1.2
86	3.0	1.6	1.5	0.6

Figure 4

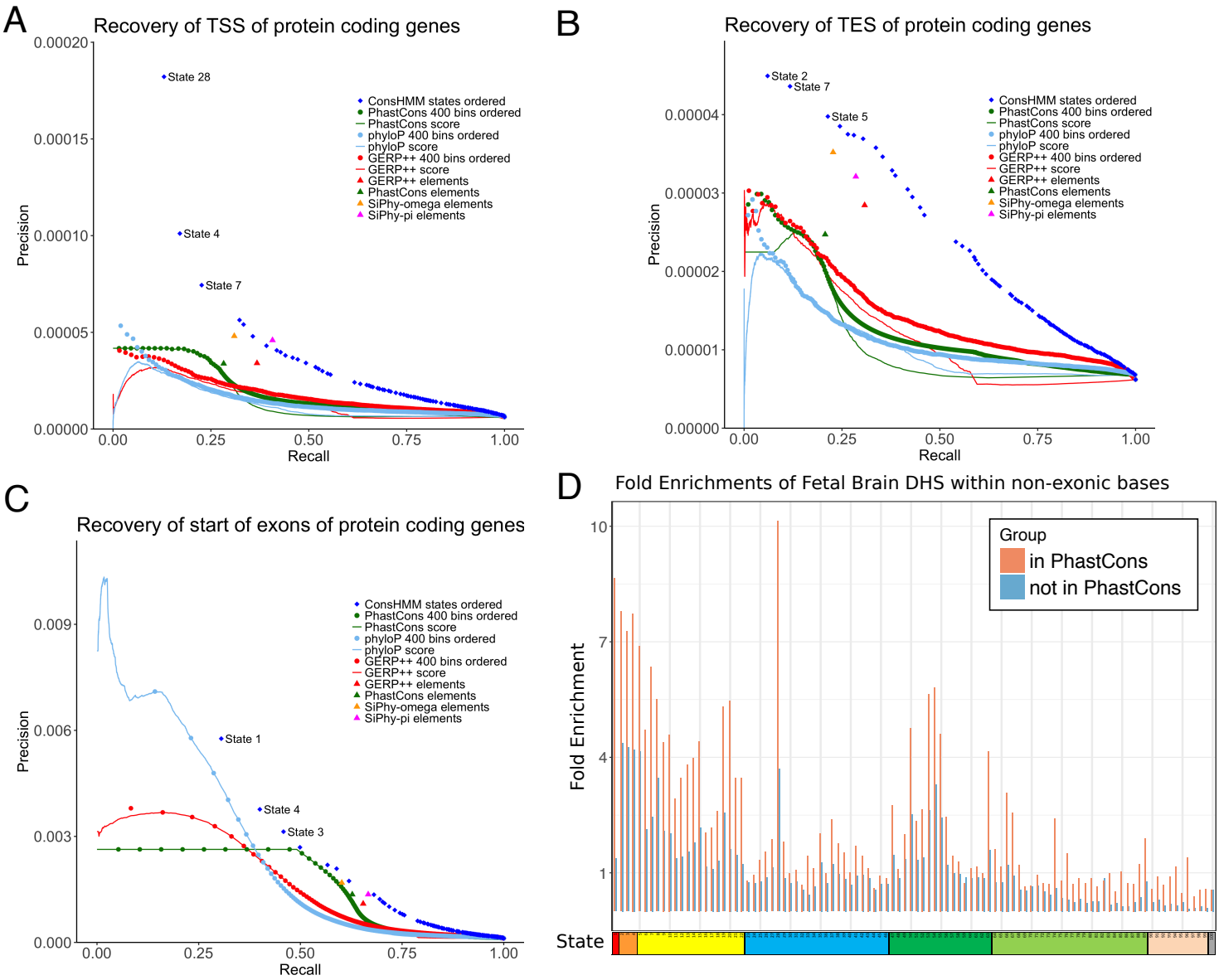

A
Recovery of TSS of protein coding genes
Precision
Recall
State 28
State 4
State 7
ConsHMM states ordered
PhastCons 400 bins ordered
PhastCons score
phyloP 400 bins ordered
phyloP score
GERP++ 400 bins ordered
GERP++ score
GERP++ elements
PhastCons elements
SiPhy-omega elements
SiPhy-pi elements
B
Recovery of TES of protein coding genes
State 2
State 7
State 5
C
Recovery of start of exons of protein coding genes
State 1
State 4
State 3
D
Fold Enrichments of Fetal Brain DHS within non-exonic bases
Fold Enrichment
Group
in PhastCons
not in PhastCons
State

Figure 5

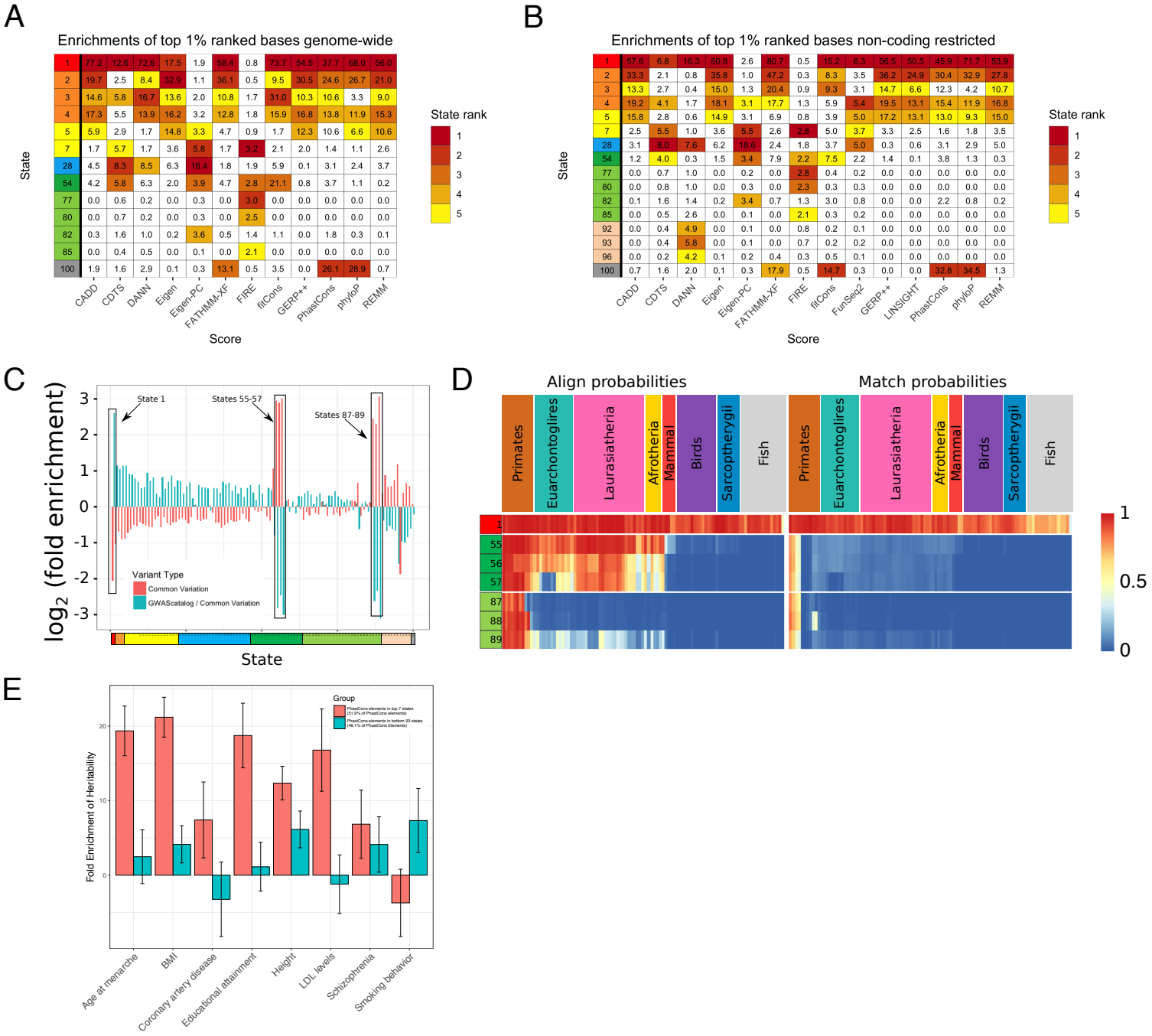

A
Enrichments of top 1% ranked bases genome-wide
State
Score
CADD
CDTS
DANN
Eigen
Eigen-PC
FATHMM-XF
FIRE
fitCons
GERP++
PhastCons
phyloP
REMM
State rank
1
2
3
4
5
B
Enrichments of top 1% ranked bases non-coding restricted
State
Score
CADD
CDTS
DANN
Eigen
Eigen-PC
FATHMM-XF
FIRE
fitCons
FunSeq2
GERP++
LINSIGHT
PhastCons
phyloP
REMM
State rank
1
2
3
4
5
C
log2 (fold enrichment)
State 1
States 55-57
States 87-89
Variant Type
Common Variation
GWAScatalog / Common Variation
State
D
Align probabilities
Match probabilities
Primates
Euarchontoglires
Laurasiatheria
Afrotheria
Mammal
Birds
Sarcopterygii
Fish
1
55
56
57
87
88
89
1
0.5
0
E
Group
Fold Enrichment of Heritability
Age at menarche
BMI
Coronary artery disease
Educational attainment
Height
LDL levels
Schizophrenia
Smoking behavior

Figure 6