Published online 17 December 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 4 **1**
doi: 10.1093/nargab/lqaa104

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

Adriana Arneson [1,2], Brooke Felsheim[2,3], Jennifer Chien[2,4] and Jason Ernst [1,2,5,6,7,8,9,*]

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, 90095, USA, [2]Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA, 90095, USA, [3]Department of Computer Science and Engineering, Washington University in St Louis, St Louis, MO, 63130, USA, [4]Department of Computer Science, Wellesley College, Wellesley, MA, 02481, USA, [5]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, Los Angeles, CA, 90095, USA, [6]Computer Science Department, University of California, Los Angeles, Los Angeles, CA, 90095, USA, [7]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA, [8]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA, 90095, USA and [9]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA, 90095, USA

Received March 01, 2020; Revised November 20, 2020; Editorial Decision November 23, 2020; Accepted November 24, 2020

ABSTRACT

ConsHMM is a method recently introduced to annotate genomes into conservation states, which are defined based on the combinatorial and spatial patterns of which species align to and match a reference genome in a multi-species DNA sequence alignment. Previously, ConsHMM was only applied to a single genome for one multi-species sequence alignment. Here, we apply ConsHMM to produce 22 additional genome annotations covering human and seven other organisms for a variety of multi-species alignments. Additionally, we extend ConsHMM to generate allele-specific annotations, which we use to produce conservation state annotations for every possible single-nucleotide mutation in the human genome. Finally, we provide a web interface to interactively visualize parameters and annotation enrichments for ConsHMM models. These annotations and visualizations comprise the ConsHMM Atlas, which we expect will be a valuable resource for analyzing a variety of major genomes and genetic variation.

INTRODUCTION

We recently introduced the ConsHMM method (1) to annotate reference genomes at single-nucleotide resolution into a number of different 'conservation states' based on the combinatorial and spatial patterns of which species have a nucleotide aligning to and/or matching the reference genome in a multi-species DNA sequence alignment. To do this, ConsHMM uses a multivariate hidden Markov model (HMM), building off the ChromHMM approach for modeling epigenomic data (2), without making any explicit phylogenetic modeling assumptions. Each nucleotide in the reference genome receives an annotation corresponding to the state of the HMM with the maximum posterior probability.

ConsHMM annotations are complementary to previous whole genome comparative genomic annotations, which have primarily focused on univariate scores or binary element calls of constraint (3–6). We previously applied ConsHMM to annotate one reference genome, human hg19, based on a 100-way vertebrate alignment (1). The conservation states had diverse and biologically meaningful enrichments for other genomic annotations, and were also able to isolate putative artifacts in the underlying multiple sequence alignment, which can confound some traditional constraint annotations.

Here, we report applying ConsHMM to produce an additional 22 genome annotations for different reference genomes and based on different multi-species DNA sequence alignments. In addition to human, seven other organisms are represented in these additional genome annotations. We have also extended the ConsHMM software to produce allele-specific annotations as opposed to only position-specific annotations based on the reference allele. We have applied this to produce annotations for each possible single-nucleotide mutation for every nucleotide in the human genome. To aid in the analysis of different ConsHMM models, we have created a web interface for interactive visualization of model parameters and annotation enrichments. These new annotations of the human genome and variation as well as model organism genomes along with a new visualization tool comprise the ConsHMM Atlas (https://ernstlab.biolchem.ucla.edu/ConsHMMAtlas/), which we expect to be a valuable resource to the community for analyzing various genomes and genetic variation.

*To whom correspondence should be addressed. Tel: +1 310 825 3658; Fax: +1 310 206 5272; Email: jason.ernst@ucla.edu

MATERIALS AND METHODS

Generating ConsHMM annotations for reference genomes

We used ConsHMM v1.0 as described in (1) to learn model parameters, to generate segmentations and annotations of reference genomes and to compute the enrichments for external annotations. We used the same parameters except the number of state parameters. The number of states we used for each alignment depended on the number of species in the alignment. Specifically, if the alignment had >50 species, then the number of states was equivalent to the number of species in the alignment; if the alignment had between 25 and 50 species, then the number of states was set to 50; and if the alignment had <25 species, then the number of states was set to 25. This set of rules allows for the number of states to be dependent on the number of species in the alignment, while also ensuring a sufficient, but not excessive, number of states for alignments with smaller number of species. Within each model, states are ordered by hierarchical clustering with optimal leaf ordering (7), but we note that there is no expected relationship between two states with same number from different models. For computing enrichments of states within each model for external enrichments, we used the OverlapEnrichment command of ConsHMM.

Creating and evaluating allele-specific ConsHMM annotations

To generate allele-specific ConsHMM annotations, we used ConsHMM v1.1, containing the new updateInitialParams and ReassignVariantState commands. ConsHMM v1.1 is built on top of ChromHMM v1.20. The updateInitialParams takes as input the parameters of a ConsHMM model and a genome-wide segmentation, and outputs an updated parameter set where the initial state parameters are replaced by the genome-wide frequency of each state in the segmentation, to better reflect the state assignment prior for a variant at any position in the genome. The ReassignVariantState command takes as input the ConsHMM model outputted by updateInitialParams, a file containing the multiple alignment on which the model is based and a parameter W. The file containing the multiple alignment is in a format processed by the parseMAF command of ConsHMM. The parameter W indicates to consider W flanking bases upstream and also W bases downstream of the allele when computing allele-specific conservation state assignments. We note that since ConsHMM uses an HMM, the state assignment at a position of interest can depend on the observations at neighboring positions. As W increases, the conservation state assignments are expected to more closely approximate those that would be obtained by directly applying ConsHMM to annotate all bases, but at the cost of a greater run-time.

We investigated the effect of different choices of W by first sampling a set of 10 000 common variants from dbSNP v150 (8) that are further than 200 kb apart, the segment size previously used with ConsHMM for genome segmentations. We then applied ConsHMM as previously done, but with the alternate allele for those common variants (1), and recorded which state the alternate allele was assigned using the ConsHMM model for hg19 based on the 100-way vertebrate alignment. We then compared the agreement in the conservation state assignment for the alternate allele when we applied ReassignVariantState with values of W between 0 and 10 and found that the agreement between the procedures plateaued at 99.9% (Supplementary Figure S1). The final allele-specific annotations were generated using $W = 10$ for each possible allele as the base in the center of the window, for every nucleotide in the human genome, and for both the hg19 and hg38 100-way vertebrate alignments. For variants in which the flanking region extends past the beginning or end of chromosomes, the missing bases upstream or downstream of the position of interest were marked as positions where the multiple sequence alignment is empty, which ConsHMM encodes as positions where no species align to the reference species.

For evaluating the major and minor allele state assignments, we used all single variants from dbSNP v150. The major allele was defined as the most frequent allele and the minor allele was the second most frequent allele for the variant. This was done for both hg19 and hg38.

For evaluating the significance of the number of alternate alleles assigned to state 5 of the hg38 100-way vertebrate alignment in a human accelerated region (HAR) for a given state of the reference allele, we used a hypergeometric distribution. For the hypergeometric distribution, the population size was the number of reference–alternate allele pairs with the reference allele in the given state. The number of successes in the population was the number of pairs that also had the alternate allele in state 5. The number of draws is the number of pairs with the reference allele in the given state and in a HAR. The number of observed successes is the number of those pairs with the alternate allele in state 5. The P-value corresponds to the probability of observing the observed number of successes or more. A Bonferroni correction was applied for testing 100 states.

Mapping between hg38 and hg19 100-way Multiz alignment states

The coordinates of the state assignments from the hg38 100-way vertebrate Multiz model were converted to hg19 coordinates using the liftOver tool from the UCSC Genome Browser (9). We then computed the enrichment of the original hg19 states for the liftOver hg38 states using ConsHMM OverlapEnrichment and mapped the hg38 state to the hg19 state that had the highest enrichment for it.

Data and code availability

All alignments we used were obtained from the UCSC Genome Browser or Ensembl (9,10). The UCSC multiple sequence alignments listed in Supplementary Table S1 were downloaded from https://hgdownload.soe.ucsc.edu/downloads.html (9). The Ensembl multiple sequence alignments listed in Supplementary Table S1 were downloaded from ftp://ftp.ensembl.org/pub/release-97/maf/ensembl-compara/ and ftp://ftp.ensembl.org/pub/release-75/emf/ensembl-compara/ (10).

SiPhy-omega, SiPhy-pi constrained element calls and HAR calls were downloaded from https://www.broadinstitute.org/mammals-models/29-

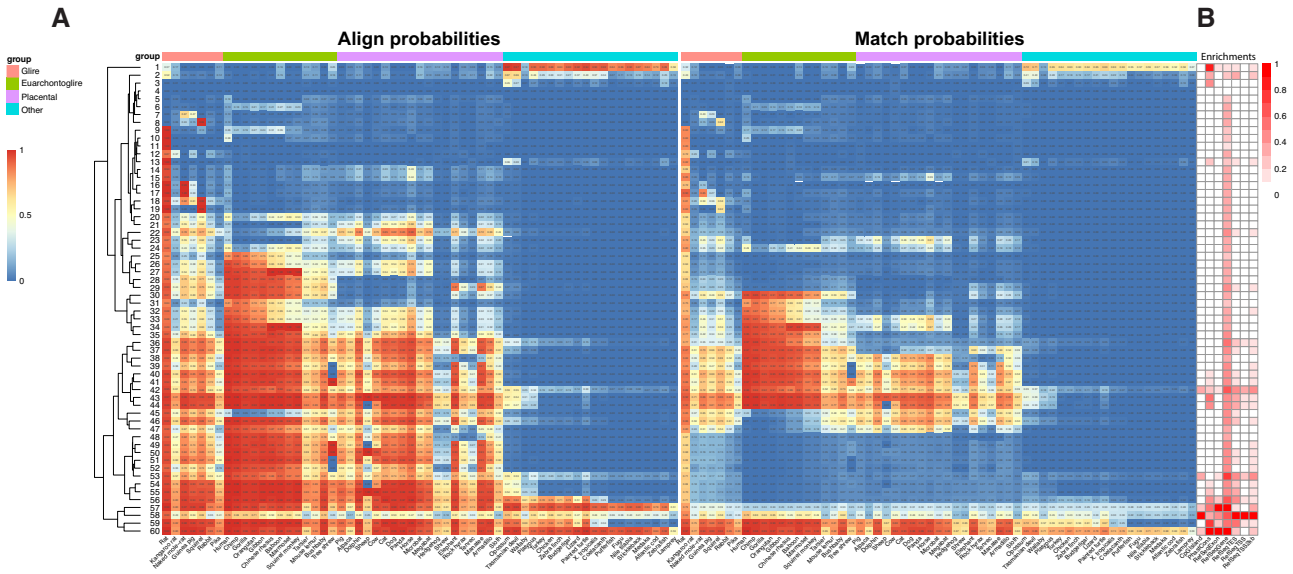

Figure 1. Conservation state emission parameters of a ConsHMM model based on a 60-way alignment of vertebrates to mouse and enrichments for other genomic annotations. **(A)** The rows of the heatmap correspond to conservation states and the columns of the heatmap correspond to species. For each state and species, the left half of the heatmap contains the probability of that species aligning to the mouse genome (mm10) at the position, which means there is a non-indel nucleotide present at the position in the alignment for the species (one minus the probability of the not aligning observation). The right half of the heatmap contains the probability of observing a species matching the mouse genome at the position, which means there is a nucleotide present in the alignment at the position for the species that is the same as in mouse. Species are ordered by phylogenetic distance to mouse and grouped by major clades. States are ordered by hierarchical clustering, using optimal leaf ordering (7) implemented in ConsHMM. **(B)** The columns of the heatmap indicate the relative enrichments of conservation states for CpG islands, PhastCons elements, and RefSeq exons, genes, transcription end sites, transcription start sites and a 2-kb window around transcription start sites (see the 'Materials and Methods' section). Each column of the enrichment heatmap was normalized by subtracting the minimum value of the column and dividing by its range. The values in the enrichment heatmap can be found in the Supplementary Data.

mammals-project-supplementary-info (4,11). Narrow-peak fetal brain DNase I hypersensitivity sites with identifiers E081 and E082 were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/ [Roadmap Epigenomics (12)]. The E081 data are used in Figure 2 and both E081 and E082 are used in Supplementary Figure S2. The mouse pseudogene annotations were obtained from https://www.gencodegenes.org/mouse/.

PhastCons constrained element calls, RefSeq and CpG island annotations, and dbSNP v150 variants were obtained from the UCSC Genome Browser. The ConsHMM model parameters and the corresponding genomic segmentations and annotations are available at https://ernstlab.biolchem.ucla.edu/ConsHMMAtlas/. The genome segmentations and annotations are available in a plain bed format and in dense and expanded bed formats that can be viewed in a genome browser. The allele-specific state annotations for the human genome and link to the R Shiny app can also be found through the same URL. The ConsHMM software is available at https://github.com/ernstlab/ConsHMM.

RESULTS AND DISCUSSION

ConsHMM annotations for additional organisms and multiple sequence alignments

We generated an additional 22 ConsHMM genome annotations that include annotations for the human, mouse, rat, dog, zebrafish, fruit fly, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* genomes (Supplementary Table S1, Supplementary Data). For some species, we generated multiple different genome annotations that corresponded to different sets of species in the multi-species alignment, different alignment methods used to generate the alignment or different assemblies of the reference genome. We applied ConsHMM as previously described (1), except setting the number of states for a model based on the number of species in the alignment (see the 'Materials and Methods' section). For each ConsHMM genome annotation, we computed enrichments for external annotations (Supplementary Data). Additionally, for the human hg38 model based on the Multiz 100-way vertebrate alignment we determined its best matching state in the previously characterized hg19 version (1) (Supplementary Table S2).

We highlight as an illustrative example of one of the new ConsHMM models that we learned for annotating the mouse mm10 genome based on the 60-way Multiz alignment of 59 vertebrates to mouse (Figure 1A and Supplementary Figure S3). In this model, which has 60 states, ConsHMM identified a number of noteworthy states showing enrichment for other external genomic annotations (Figure 1B and Supplementary Data). For example, a state that showed high aligning and matching probabilities in all the species in the alignment, state 60, was the most enriched state for exons (34.9-fold). A different state, state 58, showed a pattern of moderate probabilities of aligning and matching for almost all species, and showed strong enrichment for CpG islands (50.4-fold) and transcription start sites (34-fold). Another state, state 1, had high aligning probabilities only in distal species to mouse, which is likely capturing alignment artifacts. Despite state 1 likely capturing alignment artifacts, the state still had a 12-fold enrichment for

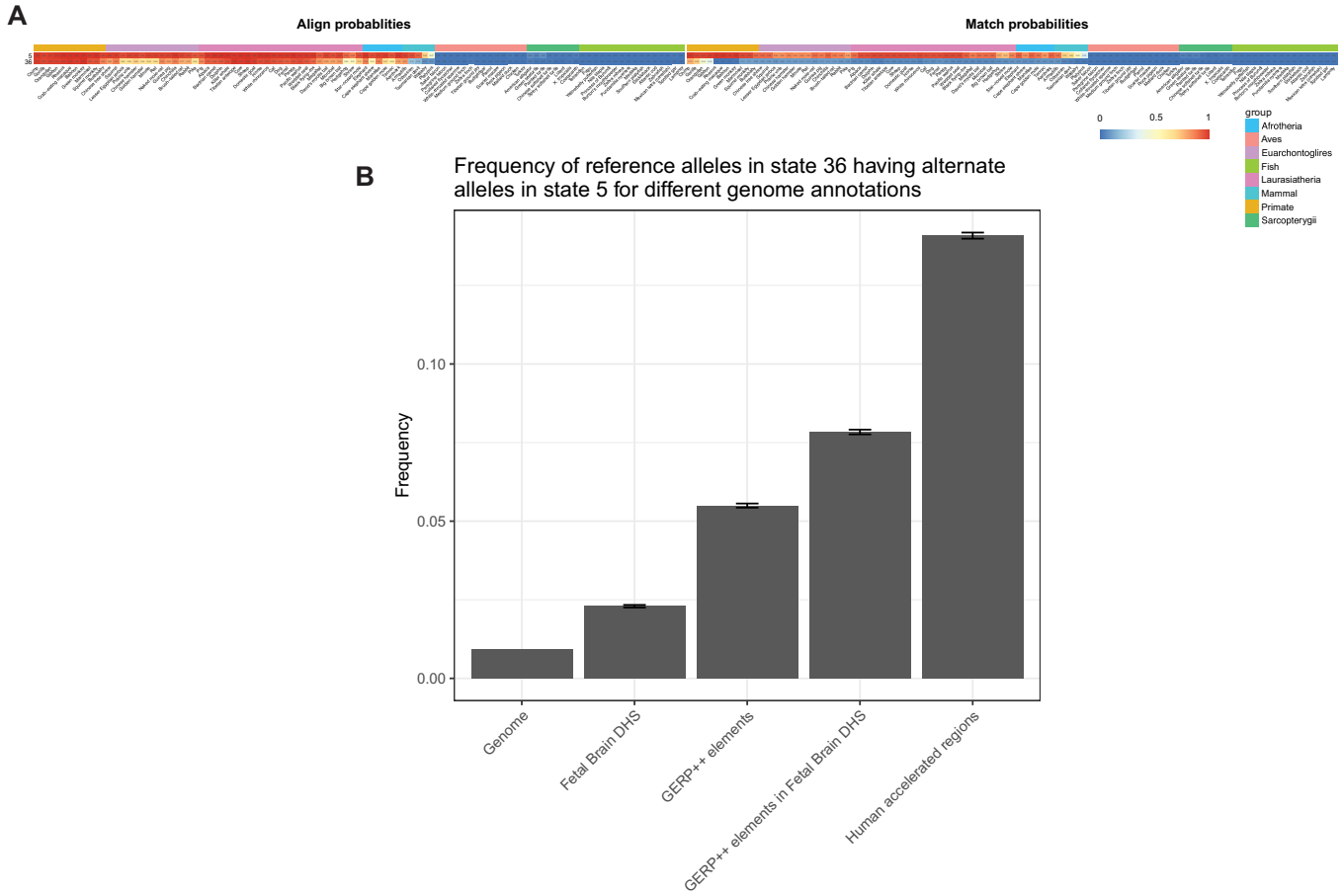

Figure 2. Example of additional information within allele-specific conservation state assignments. **(A)** Emission parameters of two states from a 100-state model based on a 100-way vertebrate alignment to the hg38 human genome, states 5 and 36. State 5 is associated with high frequency of aligning and matching through mammals, while state 36 is associated with high frequency of aligning through mammals, but only matching through a subset of primates. The heatmap is structured analogously to the heatmap in Figure 1, with the species ordered by phylogenetic distance to human in this case. **(B)** Bar graph showing for different subsets of variants assigned to state 36 based on the reference allele the frequency of assignment change to state 5 out of all possible alternate alleles. The 'Genome' category shows this frequency for all variants assigned to state 36 based on the reference allele. The rest of the columns show the frequency when restricting to subset of those variants positioned in a fetal brain DNase I hypersensitivity site (DHS), GERP++ element, the intersection of fetal brain DHS and GERP++ elements, and HARs where those annotations were the liftOver from hg19 to hg38. Error bars represent a 95% binomial confidence interval computed using a normal approximation of the error around the estimate.

PhastCons constrained element calls, suggesting many of the constrained element calls in this state are likely false positives. There were three other states (states 2, 3 and 13), which had similar though weaker versions of the state 1 alignment pattern and also enriched for PhastCons elements (3.8–6.4-fold). The alignment patterns suggestive of alignment artifacts in these states are likely due in part to pseudogenes, which all four of these states also enriched for, with enrichments ranging from 3.3- to 55.9-fold. These various state patterns and corresponding enrichment were similar to those found for a previously analyzed human conservation state annotation (1). We computed state enrichments for all other ConsHMM models (Supplementary Data). This showed as expected that the maximum state enrichments for exons were lower in genomes with a smaller portion of the genome in intergenic and intronic regions; for example, the maximum state enrichment for the ConsHMM annotation of the fruit fly genome (dm6) based on the Multiz 27-way alignment was 3.8-fold.

Allele-specific ConsHMM annotations

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

To address this challenge, we extended ConsHMM to be able to compute conservation state assignments for any alternate allele with high accuracy after learning the model parameters based on the reference genome under two assumptions. The first assumption is that it is sufficient to assume an alternate allele would not cause changes to the multi-species alignment except for the nucleotide present in the reference genome. The second assumption is that it

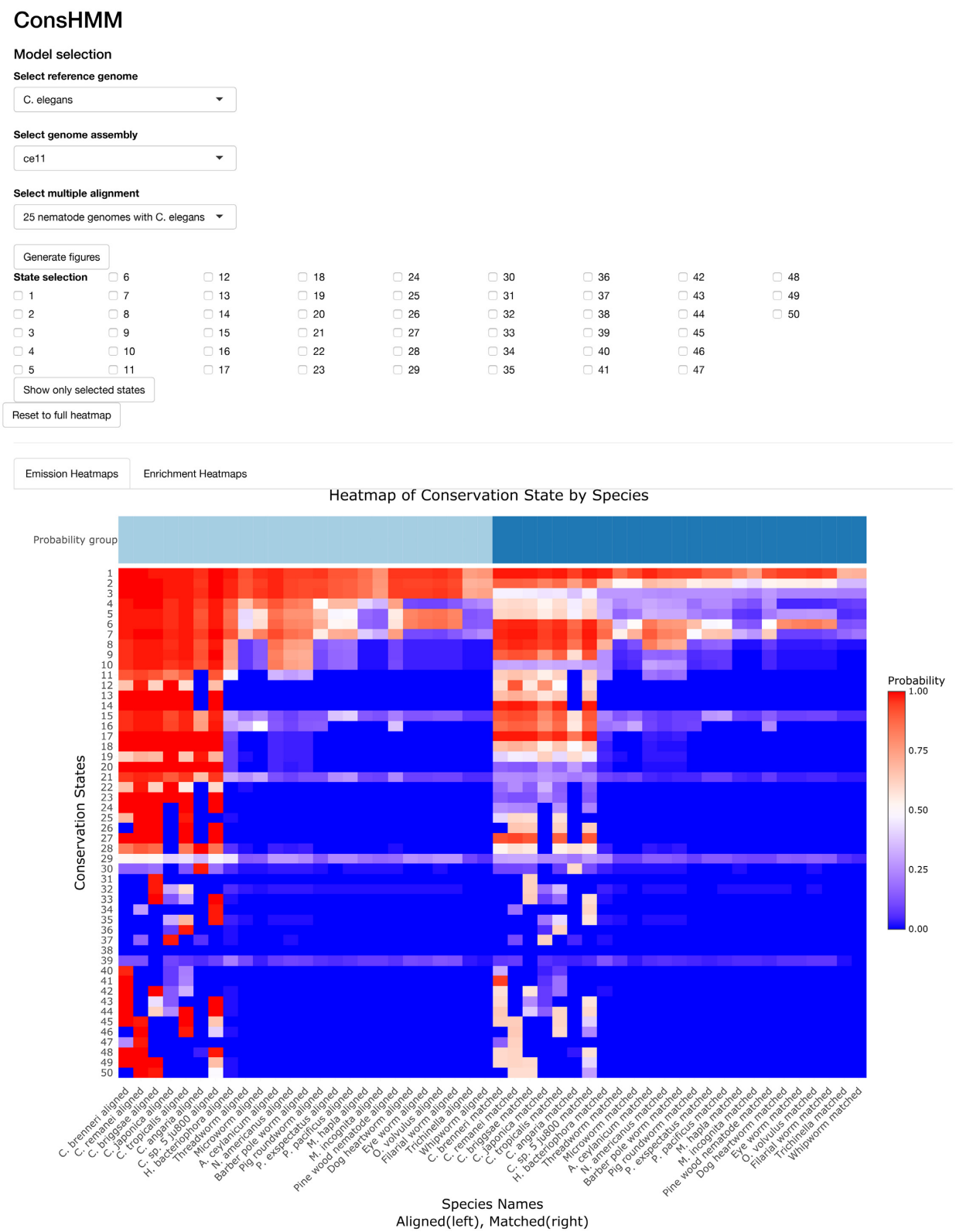

Figure 3. Screenshot of the ConsHMM R Shiny app. The screenshot captures a representation of the emission probabilities of a 50-state model based on a 26-way alignment of nematodes with *C. elegans*. The dropdown menus at the top of the web page allow users to select a different reference organism, genome or multiple sequence alignment for which to generate similar heatmaps. Each row in the heatmap corresponds to a state and each column corresponds to a species. As in Figure 1, the left half of the heatmap contains the probability of a species aligning the reference genome in the alignment, and the right half of the heatmap contains the probability of a species matching the reference genome in the alignment. The rows are sorted by hierarchical clustering, using optimal leaf ordering (7) implemented in ConsHMM, and the columns are sorted by phylogenetic distance to the reference genome in the alignment. The phylogenetic distance was extracted from the Ensembl species tree (10). The checkboxes in the 'state selection' area of the app allow users to subset the heatmap to certain states of interest.

is sufficient to consider a small local window around each variant to derive a state annotation as opposed to segmenting 200 kb at a time as previously done (1). We empirically verified this second assumption by considering a range of window sizes upstream and downstream of a variant and showing that a window of size 21 (10 bases upstream and 10 bases downstream) obtained 99.9% agreement in the conservation state assignments compared to applying ConsHMM as previously applied for a set of 10 000 common variants (see the 'Materials and Methods' section and Supplementary Figure S1). In comparison with a window of size 1, the agreement was lower, at 91.2%, though many of the state assignment differences were to similar states (Supplementary Figure S1 and Supplementary Table S3).

Using this extended version of ConsHMM, we produced allele-specific conservation state annotations for each possible single-nucleotide alternate allele for both the hg19 and hg38 human reference genomes based on ConsHMM models trained on 100-way vertebrate alignments. For both assemblies, we analyzed the distribution of the major allele and minor allele of common variants and their ratios across conservation states (Supplementary Table S4). We note that the major allele is not necessarily the reference allele. We observed that states that had a strong relative preferences for the major or minor allele tended to be in conservation states with high or low matching frequency for species that align, respectively.

To further demonstrate the additional information in having allele-specific conservation state assignments beyond just the reference genome, we considered the set of positions that were assigned to a state in the human hg38 model associated with high probability of aligning through mammals, but a high probability of matching in only a few primates, state 36. We then analyzed for different subsets of positions the frequency at which an alternate allele results in a conservation state assignment to a very different state that had high probability of both aligning and matching in many mammals, state 5 (Figure 2).

For only 0.9% (492 177 out of 55 324 866) of possible alternate alleles for reference allele in state 36, we saw the conservation state assignment change to state 5. Interestingly, we saw this percentage increase substantially for subsets of positions with other unique annotations. Among positions in fetal brain DNase I hypersensitive sites [Roadmap Epigenomics (12)], the percentage was 2.3% (51 833 out of 2 261 211) and for those in GERP++ constrained elements (13) it was 6.1% (348 875 out of 5 707 080) (Figure 2B). The percentage increased to 7.8% (40 423 out of 515 856) for those annotated as both. The percentage increased even further to 14.1% (455 out of 3231) for previously annotated bases in HARs (11). Similar percentages were found when using other sets of conserved elements and another fetal brain DNase I hypersensitivity dataset (Supplementary Figure S2). We also repeated these analyses with each other state as the reference state in place of state 36 (Supplementary Table S5), and along with state 36 observed states 30, 32, 33 and 35 have a significantly greater percentage of bases with alternate alleles in state 5 within HARs than for bases in the state in general ($P < 0.01$). These results highlight how allele-specific conservation state assignments provide additional information beyond the conservation state assignment from the reference allele.

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

To facilitate the process of interpreting different states, we created a web interface built on an R Shiny app in which one can browse a representation of emission parameters of ConsHMM models and annotation enrichments (Figure 3). Users can access the models trained on each of the reference genomes and multiple sequence alignments listed in Supplementary Table S1. The app generates an interactive heatmap containing for each state and species the model probability of that species aligning the reference genome, and also the probability of having a nucleotide matching the reference genome. The interface allows a user to select a subset of states and/or species in the alignment to display, for ease of visualization. Lastly, the interface allows users to display precomputed enrichments of states for external annotations. These include enrichments for existing annotations of gene bodies, exons, transcription start and end sites, and the PhastCons elements called on the same alignment, when available.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank the members of the Ernst lab for useful discussions.

FUNDING

National Institutes of Health [DP1DA044371, R01E S024995, U01HG007912 and U01MH105578 to J.E., T32CA201160 to A.A., R25MH109172 to B.F. and J.C.]; National Science Foundation [CAREER Award #1254200 to J.E.]; Kure It Cancer Research [Kure It Award to J.E.]; Alfred P. Sloan Foundation [Alfred P. Sloan Fellowship to J.E.].

Conflict of interest statement. None declared.

REFERENCES

1. Arneson,A. and Ernst,J. (2019) Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun. Biol.*, **2**, 248.
2. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
3. Cooper,G.M., Stone,E.A., Asimenos,G., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
4. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.
5. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
6. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
7. Bar-Joseph,Z., Gifford,D.K. and Jaakkola,T.S. (2001) Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, **17**, S22–S29.

8. Sherry,S.T., Ward,M.-.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
9. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.* (2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
10. Herrero,J., Muffato,M., Beal,K., Fitzgerald,S., Gordon,L., Pignatelli,M., Vilella,A.J., Searle,S.M.J., Amode,R., Brent,S. *et al.* (2016) Ensembl comparative genomics resources. *Database (Oxford)*, **2016**, bav096.
11. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.
12. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
13. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.