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

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

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

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

[3] Department of Computer Science and Engineering, Washington University in St. Louis, St Louis, Missouri, USA

[4] Department of Computer Science, Wellesley College, Wellesley, Massachusetts, USA

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

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

[7] Depart of Computational Medicine, University of California, Los Angeles, Los Angeles, California, USA.

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

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

Correspondence: J.E. (615 Charles E Young Dr South, 350B BSRB, Los Angeles, CA 90095; jason.ernst@ucla.edu)

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 have extended ConsHMM to generate allele specific annotations, which we used to produce conservation state annotations for every possible single nucleotide mutation in the human genome. Finally, we provide a web interface to interactively visualize parameters and annotation enrichments for ConsHMM models. These annotations and visualizations comprise the ConsHMM Atlas, which we expect will be a valuable resource for analyzing a variety of major genomes and genetic variation.

Introduction

We recently introduced the ConsHMM method(Arneson and Ernst 2019) 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(Ernst and Kellis 2012), 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(Cooper et al. 2005; Siepel et al. 2005; Garber et al. 2009; Pollard et al. 2010). We previously applied ConsHMM to annotate one reference genome, human hg19, based on a 100-way vertebrate alignment(Arneson and Ernst 2019). 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 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 (http://www.biolchem.ucla.edu/labs/ernst/ConsHMMAtlas), which we expect to be a valuable resource to community for analyzing various genomes and genetic variation.

Materials and Methods

Generating ConsHMM annotations for reference genomes

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

Creating allele-specific ConsHMM annotations

To generate allele specific ConsHMM annotations we used ConsHMM v1.1, containing the new updateInitialParams and ReassignVariantState commands. 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.

We investigated the effect of different choices of *W* by first sampling a set of 40,000 common variants from dbSNP(Sherry et al. 2001) that are further than 200kb 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(Arneson and Ernst 2019). We did this with the ConsHMM model for hg38 based on the 100-way vertebrate alignment. We then compared the agreement in the conservation state assignment when we applied ReassignVariantState with values of *W* between 1 and 10 and found that the agreement between the procedures plateaued at 99.6%. The final allele-specific annotations were generated using *W* = 10.

We did this 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.

Data and code availability

All alignments we used were obtained from the UCSC genome browser or Ensembl(Herrero et al. 2016; Haeussler et al. 2019). The UCSC multiple sequence alignments listed in **Supplementary Table S1** were downloaded from https://hgdownload.soe.ucsc.edu/downloads.html(Haeussler et al. 2019). 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/ (Herrero et al. 2016). SiPhy-omega, SiPhy-pi constrained element calls, and HAR calls were downloaded from https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info(Garber et al. 2009; Lindblad-Toh et al. 2011). Narrowpeak fetal brain DNase I Hypersensitivity Sites with Roadmap identifiers E081 and E082 were downloaded from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/(Roadmap Epigenomics Consortium et al. 2015). The E081 data was used in **fig. 2** and both E081 and E082 were used in **Supplementary fig. 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 http://www.biolchem.ucla.edu/labs/ernst/ConsHMMAtlas. 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 in addition to human genome include annotations for the mouse, rat, dog, zebrafish, fruit fly, C. elegans and S. 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(Arneson and Ernst 2019), except setting the number of states for a model based on the number of species in the alignment (**Materials and Methods**).

We highlight as an illustrative example of one of the new ConsHMM models that we learned, the model for annotating the mouse mm10 genome based on the 60-way Multiz alignment of 59 vertebrates to it (**fig. 1a**). In this model, which has 60 states, ConsHMM identified a number of noteworthy states showing enrichment for other external genomic annotations (**fig. 1b, 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 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, though still had a 12 fold enrichment for PhastCons constrained element calls. There were three other states (states 2, 3, and 13), which had similar though weaker versions of the state 1 alignment pattern and also enriched for PhastCons elements (3.8-6.4 fold). These four states were all also enriched for pseudogenes, 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(Arneson and Ernst 2019).

Allele Specific ConsHMM annotations

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

To address this challenge we extended ConsHMM to be able to compute conservation state assignments for any alternate allele with high accuracy under two assumptions. The first assumption is that it is sufficient to assume an alternate allele would not cause changes to the multi-species alignment except for the nucleotide present in the reference genome. The second assumption is that it is sufficient to consider a small local window around each variant to derive a state annotation opposed to segmenting 200kb at time as previously done(Arneson and Ernst 2019). We empirically verified this second assumption by considering a range of window sizes upstream and downstream of a variant and showing that a window of size 21 (10 bases upstream and 10 bases downstream) obtained 99.6% agreement in the conservation state assignments compared to applying ConsHMM as previously applied for a set of 40,000 common variants (**Materials and Methods**, **Supplementary fig. S1**).

Using this extended version of ConsHMM we produce allele specific conservation state annotations for each possible single nucleotide alternate allele for both the hg19 and hg38 human reference genomes based on ConsHMM models trained on 100-way vertebrate alignments. To demonstrate the additional information in having allele specific conservation state assignments beyond just the reference genome we consider the set of positions that were assigned to a state in the human hg38 model associated with high probability of aligning through mammals, but a high probability of matching in a only a few primates, state 36. We then analyzed for different subsets of positions the frequency at which an alternate allele result in a conservation state assignment to a very different state that had high probability of both aligning and matching in many mammals, state 5 (**fig. 2**).

For only 0.8% 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 Consortium et al. 2015), the percentage was 2% and for those in GERP++ constrained elements(Davydov et al. 2010) it was 5% (**fig. 2b**). The percentage increased to 7% for those annotated as both. The percentage increased even further to 12% for previously annotated bases in human accelerated regions (HAR)(Lindblad-Toh et al. 2011). Similar percentages were found when using other sets of conserved elements and another Fetal Brain DNase I hypersensitivity data set (**Supplementary fig. S2**). These results highlight how allele specific conservation state assignments provide additional information beyond the conservation state assignment from the reference allele.

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

We created a web interface built on an R shiny app in which one can browse a representation of emission parameters of ConsHMM models and annotation enrichments (**fig. 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 for that species aligning the reference genome, and also the probability of having a nucleotide matching the human reference. The interface allows a user to select a subset of states and/or species in the alignment to display, for ease of visualization. Lastly, the interface allows users to display precomputed enrichments of states for external annotations. These include enrichments for existing annotations of gene bodies, exons, transcription start and end sites, and the PhastCons elements called on the same alignment, when available.

Funding

This work was supported by funding from the US National Institutes of Health [grant numbers DP1DA044371, R01ES024995, U01HG007912, U01MH105578 to J.E., T32CA201160 to A.A., R25MH109172 to B.F. and J.C.]; US 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.].

Acknowledgements

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

References

Arneson A, Ernst J. 2019. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. Commun. Biol. 2:1–14.

Bar-Joseph Z, Gifford DK, Jaakkola TS. 2001. Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 17:S22–S29.

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

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

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

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

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. 2019. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 47:D853–D858.

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

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

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

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.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29:308–311.

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

Figure Legends

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 by hierarchical clustering, using optimal leaf ordering (Bar-Joseph et al. 2001) 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 2kb window around transcription start sites (**Materials and Methods**). 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**.

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 **fig. 1**, with the species ordered by phylogenetic distance to human in this case. **(b)** Bar graph showing for different subset 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 shows 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 human accelerated regions. Error bars represent a 95% binomial confidence interval computed using a normal approximation of the error around the estimate.

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 webpage allows 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 **fig. 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 (Bar-Joseph et al. 2001) 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 (Herrero et al. 2016). The checkboxes in the ‘state selection’ area of the app allow users to subset the heatmap to certain states of interest.

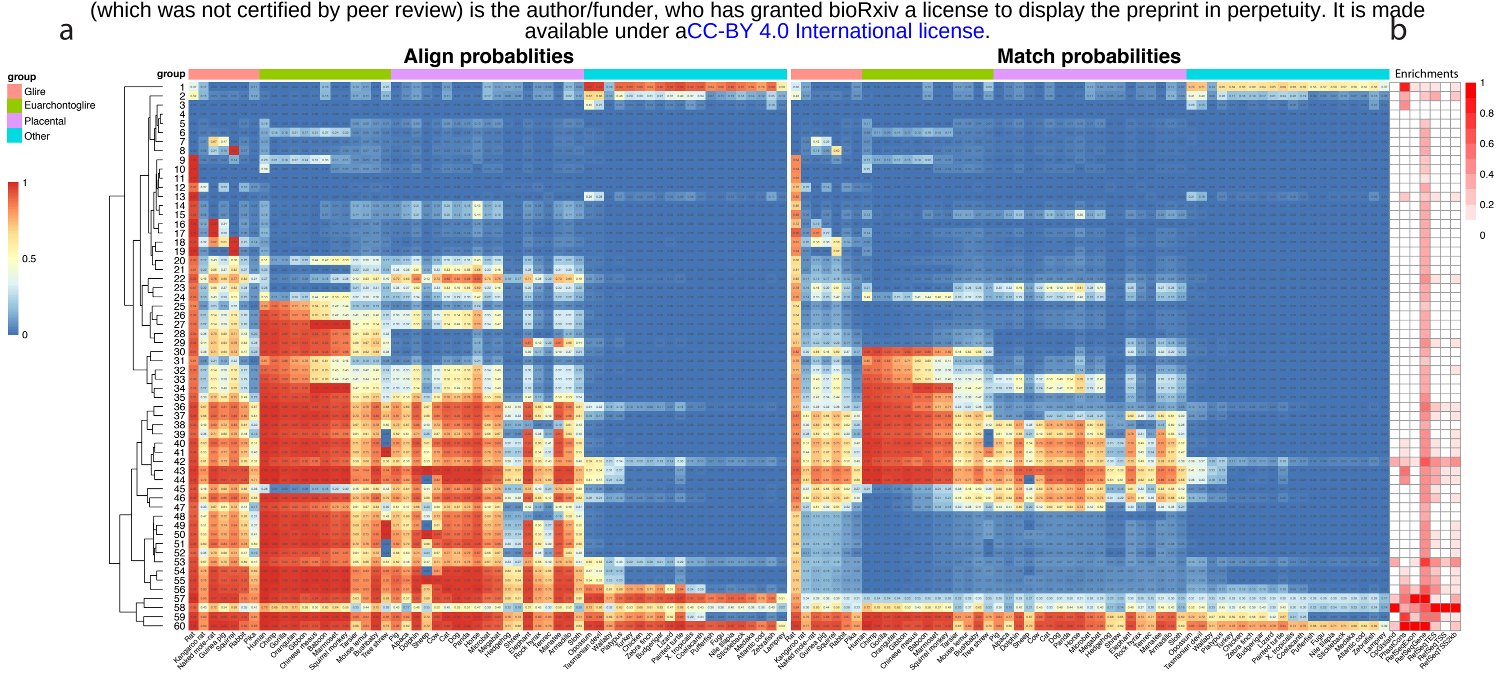

a
b
Align probablities
Match probabilities
Enrichments
group
group
Glire
Euarchontoglire
Placental
Other
1
0.5
0
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
Rat
Kangaroo rat
Naked mole–rat
Guinea pig
Squirrel
Rabbit
Pika
Human
Chimp
Gorilla
Orangutan
Gibbon
Chinese rhesus
Baboon
Marmoset
Squirrel monkey
Tarsier
Mouse lemur
Bushbaby
Tree shrew
Pig
Alpaca
Dolphin
Sheep
Cow
Cat
Dog
Panda
Horse
Microbat
Megabat
Hedgehog
Shrew
Elephant
Rock hyrax
Tenrec
Manatee
Armadillo
Sloth
Opossum
Tasmanian devil
Wallaby
Platypus
Turkey
Chicken
Zebra finch
Budgerigar
Lizard
Painted turtle
X. tropicalis
Coelacanth
Pufferfish
Fugu
Nile tilapia
Stickleback
Medaka
Atlantic cod
Zebrafish
Lamprey
CpGIsland
PhastCons
RefSeqExon
RefSeqGene
RefSeqTES
RefSeqTSS
RefSeqTSS2kb

Figure 1

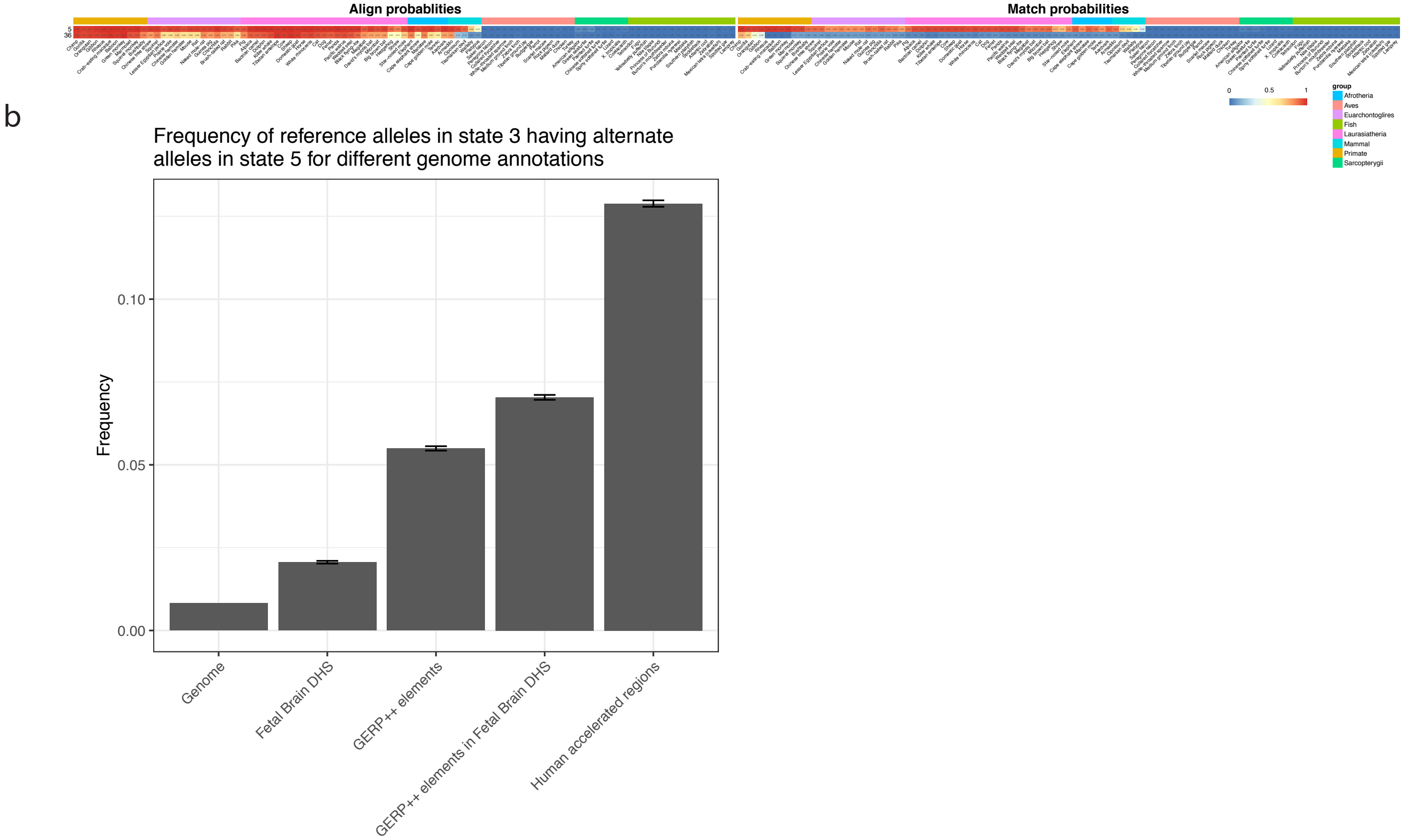
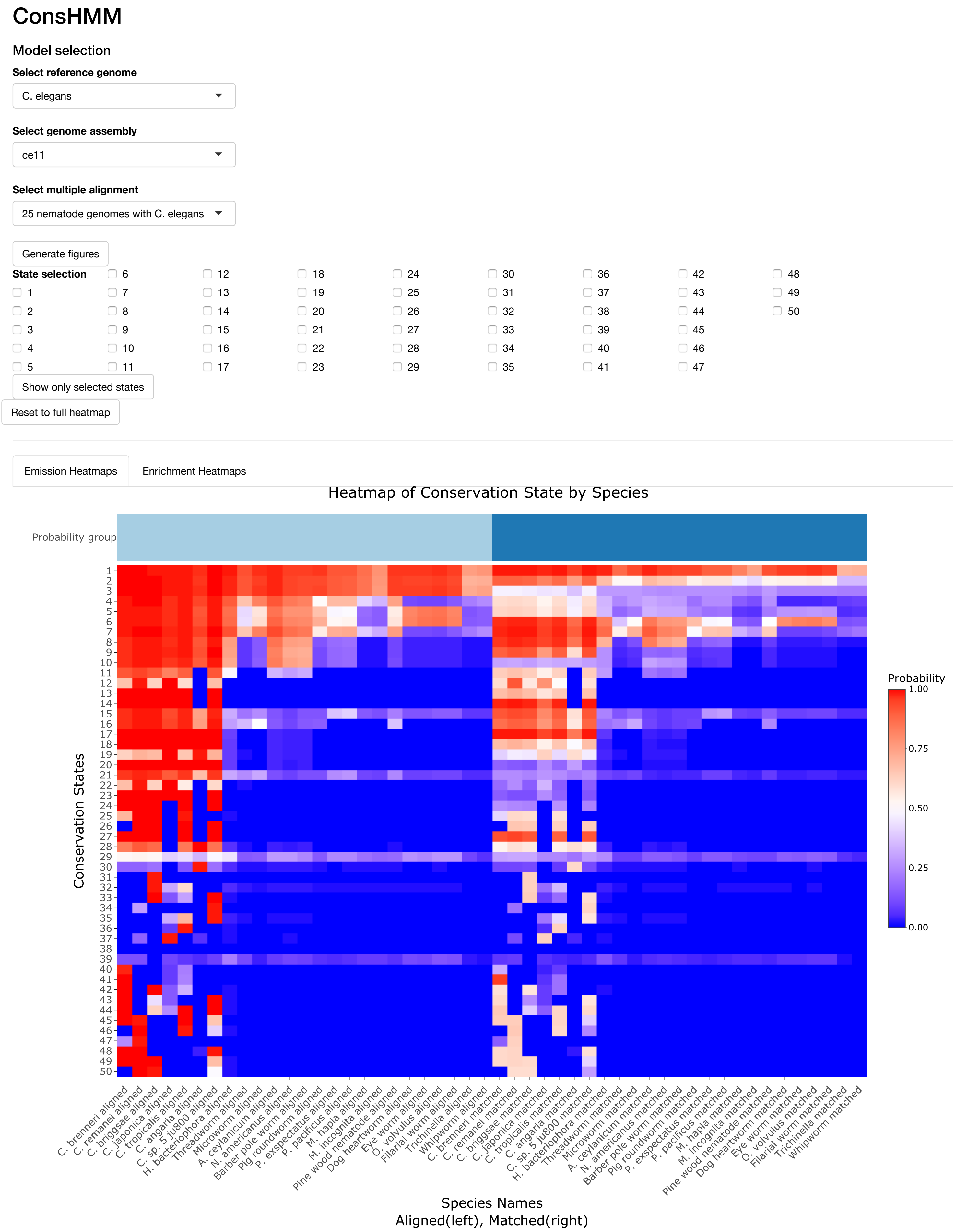

a
Align probablities
Match probabilities
0
0.5
1
group
Afrotheria
Aves
Euarchontoglires
Fish
Laurasiatheria
Mammal
Primate
Sarcopterygii
b
Frequency of reference alleles in state 3 having alternate alleles in state 5 for different genome annotations
Frequency
0.00
0.05
0.10
Genome
Fetal Brain DHS
GERP++ elements
GERP++ elements in Fetal Brain DHS
Human accelerated regions

Figure 2



ConsHMM

Model selection

Select reference genome

C. elegans ▼

Select genome assembly

ce11 ▼

Select multiple alignment

25 nematode genomes with C. elegans ▼

Generate figures

State selection

□ 1
□ 2
□ 3
□ 4
□ 5
□ 6
□ 7
□ 8
□ 9
□ 10
□ 11
□ 12
□ 13
□ 14
□ 15
□ 16
□ 17
□ 18
□ 19
□ 20
□ 21
□ 22
□ 23
□ 24
□ 25
□ 26
□ 27
□ 28
□ 29
□ 30
□ 31
□ 32
□ 33
□ 34
□ 35
□ 36
□ 37
□ 38
□ 39
□ 40
□ 41
□ 42
□ 43
□ 44
□ 45
□ 46
□ 47
□ 48
□ 49
□ 50

Show only selected states

Reset to full heatmap

Figure 3