Single-nucleotide conservation state annotation of the SARS-CoV-2 genome

Soo Bin Kwon[1,2], Jason Ernst[1,2,3,4,5,6,7]

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA 90095, USA

[2]Department of Biological Chemistry, University of California, Los Angeles, CA 90095, USA

[3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of California, Los Angeles, CA 90095, USA

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

[5]Department of Computational Medicine, University of California, Los Angeles, CA 90095, USA

[6]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, USA

[7]Molecular Biology Institute, University of California, Los Angeles, CA 90095, USA

Correspondence: jason.ernst@ucla.edu

Abstract

Given the global impact and severity of COVID-19, there is a pressing need for a better understanding of the SARS-CoV-2 genome and mutations. Multi-strain sequence alignments of coronaviruses (CoV) provide important information for interpreting the genome and its variation. We apply a comparative genomics method, ConsHMM, to the multi-strain alignments of CoV to annotate every base of the SARS-CoV-2 genome with conservation states based on sequence alignment patterns among CoV. The learned conservation states show distinct enrichment patterns for genes, protein domains, and other regions of interest. Certain states are strongly enriched or depleted of SARS-CoV-2 mutations, which can be used to predict more likely consequential mutations. We expect the conservation states to be a resource for interpreting the SARS-CoV-2 genome and mutations.

Introduction

With the urgent need to better understand the genome of SARS-CoV-2, multi-strain sequence alignments of coronaviruses (CoV) have become available[1] where multiple sequences of CoV are aligned against the SARS-CoV-2 reference genome. These alignments provide important information on the

evolutionary history of different genomic bases. Such information can be useful in interpreting mutations, as for example bases with high level of sequence constraint or accelerated evolution in certain lineages have been shown to be enriched for phenotype-associated variants[2,3]. While existing systematic annotations that quantify sequence constraint from these alignments[4,5] are informative, they are limited in the information they convey on which genomic bases align and match between the reference and each strain in the alignment, which may be useful in interrogating genetic variation[6].

As a complementary approach, ConsHMM was recently introduced to systematically annotate a given genome with conservation states that capture combinatorial and spatial patterns in multi-species sequence alignment[6]. ConsHMM specifically models whether bases from non-reference sequences align and match to the reference. ConsHMM extends ChromHMM, a widely used method that uses a multivariate hidden Markov model (HMM) to learn patterns in epigenomic data *de novo* and annotate genomes based on them[7]. Previous work applying ConsHMM to multi-species alignment of other genomes have shown that the conservation states learned by ConsHMM capture various patterns in the alignment overlooked by previous methods and are useful for interpreting DNA elements and phenotype-associated variants[6,8].

Motivated by the current need to better understand the SARS-CoV-2 genome and strain mutations, here we apply ConsHMM to two multi-strain sequence alignments of CoV that were recently made available and learn two sets of conservation states (**Fig. 1**). The first is a 44-way alignment of Sarbecoviruses, a subgenus under genus Betacoronavirus, which is part of the family of Coronavirdae. The 44-way alignment consists of SARS-CoV and 42 other Sarbecoviruses which infect bats aligned to the SARS-CoV-2 genome. The second sequence alignment to which we apply ConsHMM consists of 56 CoV that infect various vertebrates aligned to the SARS-CoV-2 genome. The vertebrate hosts include various mammals (e.g. human, bat, pangolin, mouse) and birds.

Given the two sets of conservation states learned by ConsHMM from these two alignments, we annotated the SARS-CoV-2 genome with the states and analyze the states' relationship to external annotations to understand their properties. We observed that the states capture distinct patterns in the input alignment data, some of which are in agreement with recent findings. Using external annotations of genes, regions of interest, and mutations observed among SARS-CoV-2 sequences, we observed that

the states have distinct enrichment patterns for various annotated regions. We evaluated the predictive power of the conservation state annotations for nonsingleton SARS-CoV-2 mutations and used the states to analyze and interpret previously identified mutation hotspots. Overall, our analysis suggests that the ConsHMM conservation states highlight genomic bases with distinct evolutionary patterns in the input sequence alignments and potential biological significance. The learned ConsHMM conservation states are a resource for interpreting the SARS-CoV-2 genome and variation.

Result

Annotating SARS-CoV-2 with conservation states learned from the alignment of Sarbecoviruses

First, we annotated the SARS-CoV-2 genome with 30 conservation states learned from the Sarbecovirus sequence alignment, ranging from states S1 to S30 (**Fig. 2**; **Supplementary Table 1**; **Methods**). The states capture distinct patterns of which strains align and match to the SARS-CoV-2 genome (**Fig. 2a**) and show notable enrichment patterns for external annotations of genes, proteins, and regions of interest within them (**Fig. 2b**, **Supplementary Table 4**). One state corresponds to bases where all strains align and match to SARS-CoV-2 with high probability and appears in the genome most frequently, covering 48% of the genome (S17). In contrast, another state corresponds to bases where only the strain closest to SARS-CoV-2, bat CoV RaTG13, aligns and matches to SARS-CoV-2 with high probability, covering 1% of the genome (S28). This state highlights bases that distinguish SARS-CoV-2 and bat CoV RaTG13 from other Sarbecoviruses. Notably, the state is highly enriched for human ACE2 binding domain (22 fold; $P<0.0001$; **Fig. 2b**), consistent with recent work suggesting that this binding domain is under strong positive selective pressure due to its critical role in host infection[9,10]. This state also annotates a region, known as the PRRA motif, that may have been inserted into SARS-CoV-2 potentially resulting in stronger infectiousness[11–13]. We note that this state also annotates the first five and the last seventeen bases of the genome, which may reflect technical issues with sequencing the genome ends in some strains[14]. In addition, a state corresponds to bases where all strains align to the reference with high probability, but only a subset of the strains have the same nucleotide as SARS-CoV-2 with high probability (S13; **Fig. 2a**). This subset of strains includes Sarbecoviruses that are relatively distal to SARS-CoV-2 while excluding strains that are closer to SARS-CoV-2, corresponding to a deviation along a

specific branch of the phylogenetic tree (**Supplementary Fig. 1**). Another state, (S29), shows strong enrichment of intergenic bases (35 fold; $P<0.0001$) and gene ORF10 (59 fold; $P<0.0001$), which is consistent with recent work suggesting that ORF10 may not be a protein-coding gene based on gene expression[15] and phylogenetic codon modeling[16].

Annotating SARS-CoV-2 with conservation states learned from the alignment of Coronaviruses infecting vertebrates

In addition to the 30-state model learned from the Sarbecovirus sequence alignment, we learned another 30-state model by applying ConsHMM to the alignment of CoV from vertebrate hosts (V1~V30; **Fig. 3**; **Supplementary Table 2**; **Methods**). The vertebrate CoV alignment in addition to including some Sarbecovirus included a diverse set of CoV that are evolutionarily more diverged from SARS-CoV-2. We therefore applied ConsHMM separately to the vertebrate CoV alignment, instead of combining the two alignments.

The resulting conservation states correspond to bases with distinct probabilities of aligning and matching to various strains of vertebrate CoV and exhibit notable enrichment patterns for previously annotated regions within genes (**Fig. 3a**, **Supplementary Table 4**). A state (V27) annotates bases that align and match to all 56 CoV with a genome coverage of 9%. Another state (V19) corresponds to bases that align and match specifically to four strains most closely related to SARS-CoV-2 based on phylogenetic distance, which include two bat CoV (RaTG13 and BM48-31/BGR/2008), pangolin CoV, and SARS-CoV. A state (V20) has a high align and match probabilities primarily for CoV with bat or pangolin as hosts and is enriched for the spike protein's receptor binding motif (RBM), where a recombination event between a bat CoV and a pangolin CoV might have occurred[11] (6.9 fold enrichment). Additionally, a state (V29) with high align and match probabilities specifically for bat CoV RaTG13 annotates the PRRA motif mentioned in the previous section, which is consistent with the possibility that the motif was recently introduced to the SARS-CoV-2 genome.

Conservation states' relationship to nonsingleton SARS-CoV-2 mutations observed in the pandemic

We next investigated how the learned conservation states relate to nonsingleton SARS-CoV-2 mutations observed in the current pandemic (**Fig. 2b, 3b**). Specifically, we analyze the state enrichment patterns for mutations observed at least twice in about 4,000 SARS-CoV-2 sequences from GISAID (Global Initiative on Sharing All Influenza Data)[17]. To focus on reliable calls of mutations, we limited our analysis to nonsingleton mutations and also with genomic positions with known issues[14] filtered out (**Methods**). In the Sarbecovirus model, as expected, the state that corresponds to bases where all strains align and match to SARS-CoV-2 (S17) is significantly depleted of mutations observed in the current pandemic (0.7 fold enrichment; $P<0.0001$) while multiple states (S6, S26, S29) are significantly enriched for mutations (2.0~3.1 fold; $P<0.0001$).

We observed some similarities, but also notable differences in the relationship between the vertebrate CoV model's conservation states and nonsingleton SARS-CoV-2 mutations. The vertebrate CoV model learns states that are depleted of mutations with a minimum fold enrichment of 0.2 ($P<0.0001$; V11), which is a stronger depletion than the minimum enrichment of 0.7 observed in the Sarbecovirus model. This is expected as the vertebrate CoV alignment contains a more diverse set of strains and thus is likely to capture deeper constraint than those learned from the Sarbecovirus alignment (**Fig. 3c**). Moreover, while the only state that is statistically significantly depleted of mutations in the Sarbecovirus model has high align and match probabilities for all Sarbecoviruses (S17), states significantly depleted of mutations in the vertebrate CoV model include not only an analogous state with high align and match probabilities for all vertebrate CoV (V27; 0.3 fold enrichment; $P<0.0001$), but also V11 which has high align and match probabilities for only a subset of vertebrate CoV. This subset excludes strains in a specific subtree in the phylogeny of CoV, largely consisting of CoV from avian hosts (**Supplementary Fig. 2**). This indicates that bases constrained among a specific subset of vertebrate CoV, which appear to have diverged in some of the avian CoV genomes, may be as important to SARS-CoV-2 as those constrained across all vertebrate CoV. In addition, the vertebrate CoV model learns states that are enriched for mutations (1.7~1.8 fold; $P<0.01$; V3, V14, V20, V30) although not as strongly enriched as some from the Sarbecovirus model which had fold enrichments between 2.0 and 3.1 (S6, S26, S29). The enrichment patterns for nonsingleton mutations reported here are largely consistent when we include all

observed mutations or control for the nucleotide composition of each base being mutated (**Supplementary Table 3**).

Using conservation states to predict SARS-CoV-2 mutations

Since we observed strong enrichment patterns of nonsingleton mutations in certain conservation states, we next used the states to predict genomic bases without nonsingleton mutations, which are more likely to be important bases than other bases. Specifically, we used the state annotations to distinguish positive bases, which are bases with no mutations or with a singleton, from negative bases, which are bases with mutations observed in at least two SARS-CoV-2 sequences. The positive and negative bases were matched by their nucleotide composition such that the proportions of each nucleotide in positive and negative bases were the same. As a comparison, we also used existing sequence constraint annotations learned from the sequence alignment identical to or containing the same strains as those provided to ConsHMM as input, which include PhastCons scores[4], PhyloP scores[5], and a five-way annotation based on mutation type and codon constraint in Sarbecoviruses[16] (**Methods**). For PhastCons and PhyloP scores, we additionally generated discrete versions of the annotation by binning the scores and evaluated these bins in an analogous way to state annotations (**Methods**). Additionally, we used gene annotations and annotations of possible intergenic, synonymous, missense and nonsense mutations to prioritize bases without nonsingleton mutations by treating each gene or separately each type of mutation as a conservation state to see if the predictive power of ConsHMM states was also obtainable in these other annotations (**Methods**).

In general, the conservation states provided greater information for predicting bases without nonsingleton mutations than other annotations (**Fig. 4**, **Supplementary Fig. 3**). States from the Sarbecovirus model had an area under the receiver operating characteristic curve (AUROC) of 0.62 , which was significantly better than PhastCons scores learned from Sarbecovirus sequence alignments, gene annotations, and annotations of possible intergenic, synonymous, missense and nonsense mutations (AUROC: 0.54-0.57; DeLong test $P<0.0001$) and marginally better than PhyloP scores learned from the same alignment (AUROC: 0.61; **Fig. 4a**, **Supplementary Fig. 3a**). States from the vertebrate CoV model result in higher precision than those from the Sarbecovirus model at low (<0.5) recall rates

whereas states from the Sarbecovirus model result in higher precision at high (>0.5) recall rates (**Fig. 4b**), suggesting that the two state annotations may provide complementary information to each other in prioritizing important genomic bases.

We thus integrated states from the two ConsHMM models and observed a stronger predictive power in differentiating bases without nonsingleton mutations. To integrate the state annotations, given two states from different ConsHMM models annotating a base of interest, we annotated the base with the state that is more depleted of nonsingleton mutations among a subset of bases that excluded the base of interest (**Methods**). By integrating the two state annotations, we obtained better areas under the curve than using annotations from a single model (**Fig. 4a, b**). Overall, we observe that the conservation states provide more information than existing annotations in finding bases without nonsingleton mutations. When mutations are observed in states with strong depletion of mutations, they may be more consequential than mutations observed in other states. To allow researchers to utilize this information in studying SARS-CoV-2 mutations, we share a genome-wide track that scores each base by the degree of depletion or enrichment of nonsingleton mutations genome-wide among those observed with statistical significance in a combined set of conservation states (**Fig. 4c**, **Methods**).

Analyzing SARS-CoV-2 mutation hotspots with conservation states

Lastly, we used the conservation states to analyze bases that are SARS-CoV-2 mutation hotspots. Specifically we analyzed thirteen mutation hotspots identified by a recent study[18] in the context of their spatial and temporal trends of mutation occurrences (**Table 1**). We observe that all but two of the eight hotspots specific to Europe and North America, most of which emerged in February, are annotated by the same state from the Sarbecovirus model, state S17, whereas only one of the five hotspots shared with Asia and Oceania, which all emerged in January, is annotated by that state (hypergeometric test $P<0.09$). This raises the possibility that mutations in state S17, which has high align and match probability for all Sarbecoviruses and is most depleted of SARS-CoV-2 mutations in general among states from the same model, are more consequential when they do occur.

Several of the mutations that are specific to Europe and North America and annotated with state S17 have been previously hypothesized to be consequential. For example, one of the late mutation

hotspots observed in Europe (position 14408) was suggested to contribute to worsening the virus's proofreading mechanism, making it easier for the virus to adapt and harder for its hosts to gain immunity[18]. This silent mutation lies in a coding region for RNA-dependent RNA polymerase (RdRp), an enzyme involved in the virus's proofreading machinery, and viruses with this mutation tend to have more subsequent mutations than others[18]. We also note that the D614G mutation in the spike protein (position 23403) that was recently implicated to result in increased infectivty[19] is also one of the late hotspots in Europe annotated by state S17. Another hotspot annotated by state S17, position 17857 in gene ORF1ab, which emerged in North America in March, is notable in that it is the only hotspot that is annotated by a state with high align and match probabilities for all vertebrate CoV (V27), suggesting strong constraint among not only Sarbecoviruses, but also vertebrate CoV. This mutation is a missense mutation in protein nsp13 helicase. The conservation state annotation suggests this mutation has a possibly stronger functional consequence than other hotspots with otherwise similar geographic and temporal properties given the strong depletion of mutations in this state in general.

Discussion

Here we applied a comparative genomics method ConsHMM to two sequence alignments of CoV, one consisting of Sarbecoviruses that infect human and bats and the other consisting of a more diverse collection of CoV that infect various vertebrates. The conservation states learned by ConsHMM capture combinatorial and spatial patterns in the multi-strain sequence alignments. The states show associations with various other annotations not used in the model learning. The conservation state annotations are complementary to constraint scores, as they capture a more diverse set of evolutionarily patterns of bases aligning and matching, enabling one to group genomic bases by states and study each state's functional relevance. Understanding patterns of conservation across different strains can be important potentially for understanding cross-immunity from other coronaviruses[20–22].

We showed that the conservation states have predictive information on whether a nonsingleton SARS-CoV-2 mutation is observed in a genomic base and that this predictive power is greater than that of other annotations considered. Based on state enrichments and depletions for bases without nonsingleton mutations, we generated a genome-wide score track that can be used to prioritize mutations

of potentially greater consequence based on evolutionary information of both the Sarbecovirus and vertebrate CoV alignments. We note that the score is generated in a transparent way directly from the fold enrichment values for nonsingleton mutations observed in the conservation states. Overall, we expect the two sets conservation state annotations along with this score track to be resources for locating bases with distinct evolutionary patterns and analyzing and interpreting mutations that are currently accumulating among SARS-CoV-2 sequences.

Methods

Sequence alignments

We obtained the 44-way Sarbecovirus sequence alignment from UCSC Genome Browser[1] (**http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/multiz44way/**). We obtained the vertebrate CoV sequence alignment by first downloading the 119-way vertebrate CoV sequence alignment from UCSC Genome Browser (**http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/multiz119way/**) and then removing the SARS-CoV-2 sequences from the alignment, except the reference sequence, wuhCor1. This resulted in 56 CoV aligned against the reference.

External annotations

Mutations found in SARS-CoV-2 sequences were point mutations identified by Nextstrain[23] (accessed on May 22, 2020) from sequences available on GISAID[17]. For our analysis, we filtered out mutations if their ancestral alleles did not match the reference genome used by Nextstrain, MN908947.3. All the other annotations, including the annotations of genes, codons, and UniProt protein products and regions of interest, were accessed through UCSC Genome Browser[1].

Choice of number of ConsHMM conservation states

Given the two input sequence alignments, we first learned multiple ConsHMM models from each alignment with varying numbers of states ranging from 5 to 100 with increments of 5 and then chose a number of states that is applicable to both alignments. Specifically, we aimed to find a number of states

that results in states few enough to easily interpret and generalize, but specific enough to capture distinct patterns in the alignment data.

To do so, for each model, we considered whether the model's states had sufficient coverage of the genome to avoid having states that annotate too few bases (e.g. 10 bp). We additionally considered whether the model's states exhibited distinct emission parameters to ensure that they were different enough to capture distinct patterns in the alignment data. Lastly, we considered whether the model's states showed distinct enrichment patterns for external annotations of genes, protein domains, and mutations in SARS-CoV-2 and showed strong predictive power for bases without mutations to ensure that the different states annotate bases with potentially different biological roles. As a result, we chose 30 as the number of conservation states for both the Sarbecovirus and vertebrate CoV ConsHMM models because the resulting states were sufficiently distinct in their emission parameters and association with external annotations and most of the states covered more than 1% of the genome.

PhastCons and PhyloP scores

We obtained the 44-way PhastCons and PhyloP scores learned from the Sarbecovirus sequence alignment from UCSC Genome Browser (**http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/**). We additionally used the PHAST software[24] to learn PhastCons and PhyloP scores from the vertebrate CoV sequence alignment that we generated from the 119-way alignment as described above. To do so, we first ran 'tree_doctor' to prune out SARS-CoV-2 sequences except the reference from the phylogenetic tree generated for the 119-way alignment. We then followed the procedure used to generate the 44-way and 119-way scores as described on UCSC Genome Browser. Specifically, to learn the vertebrate CoV PhastCons score, we used the following arguments to run 'phastCons': --expected-length 45 –target-coverage 0.3 –rho 0.3. To learn the vertebrate CoV PhyloP score, we used the following arguments to run 'phyloP': --wig-scores –method LRT –mode CONACC.

Fold enrichment for external annotations

When computing fold enrichments for annotations of genes, positions within codons, and regions of interest, we considered whether a genomic base is annotated or not. Because multiple mutations could

be observed in the same genomic base, when computing fold enrichments for mutations, we first generated all possible point mutations in the SARS-CoV-2 genome and then considered whether each of the possible mutations was observed or not. We focused on mutations observed in at least two SARS-CoV-2 sequences. Additionally, when computing enrichment for mutations, we masked the following genomic positions based on a prior analysis[14] as they are likely affected by sequencing errors, low coverage, contamination, or hypermutability: 1-55, 29804-29903, 187, 1059, 2094, 3037, 3130, 4050, 6990, 8022, 10323, 10741, 11074, 11083, 13402, 13408, 14786, 15324, 19684, 20148, 21137, 21575, 24034, 24378, 25563, 26144, 26461, 26681, 28077, 28826, 28854, 29700. For all fold enrichment values, we also conducted a two-sided binomial test to report statistical significance. We applied a Bonferroni correction by setting the significance threshold to 0.05 divided by 30, the number of states.

Prediction of genomic bases without nonsingleton mutations

Our positive genomic bases were defined as bases with no mutations or a singleton among mutations observed during the current pandemic identified by Nextstrain[23]. Our negative genomic bases were the remaining bases in the genome, which are bases with mutations observed in at least two SARS-CoV-2 sequences. As explained above, we masked multiple positions to avoid including erroneous mutation calls. We also downsampled the positive bases such that the fractions of each nucleotide in positive and negative bases were the same as a way of controlling for nucleotide composition bias. This resulted in 9589 positive and 1410 negative bases.

We split the positive and negative bases randomly into 10 groups to conduct 10-fold cross-validation (CV). To make predictions on one group, we used bases in the other nine groups as training bases. We then followed the precision recall analysis procedure used in Arneson et al.[6] Specifically, given the training bases, we ordered the conservation states in decreasing order of their enrichment for positive training bases. Based on the ordering, we assigned a ranking to test bases in each bin, which was then used as a score for prioritizing bases without mutations. We repeated this for all 10 folds.

For any other categorical predictors, we used the same procedure, treating each category as a conservation state. Specifically, for annotation of genes, genes were ranked based on their enrichment for positive training bases, which was later used to rank test bases. The same procedure was applied to

an annotation of possible intergenic, synonymous, missense, and nonsense mutations and separately to a five-way annotation based on mutation type and codon constraint in Sarbecoviruses[16]. In each of these categorical predictors, if a test base was assigned to multiple categories, the category with the stronger enrichment for positive training bases was assigned to the test base. For PhastCons and PhyloP scores, we used the score itself to rank test bases without considering training bases. In addition, to make the scores more analogous to the 30 conservation states, we applied the above procedure based on enrichment of positive training bases to the constraint scores by generating 30 bins based on the scores. Specifically, for each score, we ranked the training bases based on the score, partitioned the training bases into 30 bins based on the ranking, and applied the procedure based on enrichment of positive training bases as explained above by treating each score bin as a conservation state.

When integrating conservation states learned from different alignments, we aggregated the states and ranked them based on their enrichment for positive training bases. Then, given two states from different models annotating each test base, we annotated the test base with the state that had greater enrichment for positive training bases than the other state and then used the ranking of that state to score the test base. To integrate sequence constraint scores learned from different alignments, we trained a logistic regression classifier with features consisting of two sequence constraint scores, one learned from the Sarbecovirus alignment and the other from the vertebrate CoV alignment. We considered the probability of a test base being classified as a positive base to be a predictor for bases without mutations. To train a logistic regression classifier, we used liblinear[25] with L1 penalty and regularization which was tuned via grid search. Specifically, for each of the ten outer CV folds, we conducted an inner five-fold CV where we optimized the inverse of regularization strength, given a grid of ten values in a logarithmic scale ranging from 0.0001 to 10,000. AUROC was used to choose the best regularization value in each outer fold. We also applied the binning approach described above to the integrated score learned by the logistic regression models, generating 60 bins based on the scores and ranking them by their enrichment for positive training bases.

Generating a browser track of depletion of nonsingleton SARS-CoV-2 mutations

Based on the procedure of computing state enrichment of SARS-CoV-2 mutations, we selected states from both ConsHMM models that exhibited statistically significant enrichment or depletion of nonsingleton mutations at a Binomial test p-value threshold of 0.05 after Bonferroni correction. For each base annotated with any of the selected states, we scored the base with $-\log_2(v)$ where v is the fold enrichment value of the state annotating the base, such that stronger depletion of mutations corresponded to a higher score above 0 and stronger enrichment to a lower score below 0. If a base was annotated with two of the selected states, each from different ConsHMM models, and the two states agreed in the enrichment direction (enriched or depleted), we annotated the base with the $-\log_2(v)$ from the states that had a higher absolute value of $-\log_2(v)$. If a base was annotated with two of the selected states but the states disagreed in the enrichment direction, we annotated the base with score of 0. Bases not annotated by any of the selected states were assigned a score of 0 as well.

Computing and comparing areas under the curve

Area under the precision-recall curve (AUPRC) was computed conservatively without linear interpolation using Scikit-learn's implementation of computing average precision[26]. To evaluate the significance of the AUROC difference of a given ConsHMM conservation state annotation to that of other predictors, we conducted a DeLong test[27] with Bonferroni correction where we set the significance threshold to 0.05 divided by the number of comparisons. We compared both ConsHMM annotations to these six predictors predictors: gene annotation, annotation of mutation type, and four sequence constraint annotations which were learned from the same alignment as the ConsHMM annotation, as explained above. In addition, for the Sarbecovirus ConsHMM annotation, we compared to a five-way annotation based on mutation type and codon constraint[16].

Data access

ConsHMM conservation state annotation based on the Sarbecovirus and vertebrate CoV alignments are available at **https://github.com/ernstlab/ConsHMM_CoV/**. Track annotation of depletion of mutations observed in conservation states from both Sarbecovirus and vertebrate CoV ConsHMM models are available from the same URL.

Acknowledgements

We gratefully acknowledge all those who contributed to generating and sharing their SARS-CoV-2 sequence data via the GISAID Initiative. We thank those at Nextstrain.org who made their processed mutation data publicly available. We also thank Adriana Arneson for assistance on using ConsHMM. We thank Sriram Sankararaman for comments on the manuscript. This research was supported by the UCLA David Geffen School of Medicine – Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Award Program and the US National Institutes of Health (DP1DA044371).

Reference

1. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
2. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).
3. Xu, K., Schadt, E. E., Pollard, K. S., Roussos, P. & Dudley, J. T. Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. *Mol. Biol. Evol.* **32**, 1148–1160 (2015).
4. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, (2005).
5. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
6. Arneson, A. & Ernst, J. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun. Biol.* **2**, 248 (2019).
7. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
8. Arneson, A., Felsheim, B., Chien, J. & Ernst, J. ConsHMM Atlas: conservation state annotations for major genomes and human genetic variation. *bioRxiv* 2020.03.01.955443 (2020). doi:10.1101/2020.03.01.955443

9. Armijos-Jaramillo, V., Yeager, J., Muslin, C. & Perez-Castillo, Y. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. *Evol. Appl.* (2020). doi:doi:10.1111/eva.12980

10. Frank, H. K., Enard, D. & Boyd, S. D. Exceptional diversity and selection pressure on SARS-CoV and SARS-CoV-2 host receptor in bats compared to other mammals. *bioRxiv* (2020). doi:10.1101/2020.04.20.051656

11. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* (2020). doi:10.1126/sciadv.abb9153

12. Xiao, C. *et al.* HIV-1 did not contribute to the 2019-nCoV genome. *Emerg. Microbes Infect.* **9**, 378–381 (2020).

13. Wang, Q. *et al.* A Unique Protease Cleavage Site Predicted in the Spike Protein of the Novel Pneumonia Coronavirus (2019-nCoV) Potentially Related to Viral Transmissibility. *Virol. Sin.* 1–3 (2020). doi:10.1007/s12250-020-00212-7

14. De Maio, N. *et al.* Issues with SARS-CoV-2 sequencing data. *Virological.org* (2020).

15. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* (2020). doi:https://doi.org/10.1016/j.cell.2020.04.011

16. Jungreis, I., Sealfon, R. & Kellis, M. Sarbecovirus comparative genomics elucidates gene content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations. *bioRxiv* (2020). doi:10.1101/2020.06.02.130955

17. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. challenges (Hoboken, NJ)* **1**, 33–46 (2017).

18. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 179 (2020).

19. Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* 2020.06.12.148726 (2020). doi:10.1101/2020.06.12.148726

20. Le Bert, N. *et al.* SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* (2020). doi:10.1038/s41586-020-2550-z

21. Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans.

Science (80-.). (2020). doi:10.1126/science.abd3871

22. Grifoni, A. *et al.* Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. *Cell* **181**, 1489-1501.e15 (2020).
23. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
24. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).
25. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
26. Pedregosa, F. *et al.* Scikit-learn : Machine Learning in Python. *J. Mach. Learn. Res.* (2011). doi:https://dl.acm.org/citation.cfm?id=2078195
27. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845 (1988).
28. Bar-Joseph, Z., Gifford, D. K. & Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering . *Bioinformatics* **17**, S22–S29 (2001).
29. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).
30. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).
31. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

	Conservation states		*Geographic area*		*Time period*		
Locus	*Learned from the Sarbecovirus alignment*	*Learned from the vertebrate CoV alignment*	*Asia / Oceania*	*Europe / North America*	*Jan*	*Feb*	*March*
1397	S13	V20					
2891	S19	V15					
3036	S17	V22					
8782	S26	V25					
11083	S7	V9					
14408	S17	V3					
17746	S17	V7					
17857	S17	V27					
18060	S12	V8					
23403	S17	V21					
26143	S17	V21					
28144	S9	V30					
28881	S17	V13					

Table 1. ConsHMM conservation state assignment, sequence constraint scores, geographic area, and period of time of mutation hotspots.

Each row corresponds to a mutation hotspot reported by Pachetti et al[18]. First column contains the genomic position of the mutation. Second and third columns contain the ConsHMM conservation state learned from the Sarbecovirus alignment and the vertebrate CoV alignment, respectively, assigned to each mutation. The next two columns correspond to whether each mutation was observed in Asia or Oceania and whether each mutation was observed in Europe or North America, respectively, where grey is shown if the row's mutation was observed in the area and white otherwise. The last three columns correspond to whether each mutation was observed in a particular month, where grey is shown if the row's mutation was observed during the period and white otherwise.

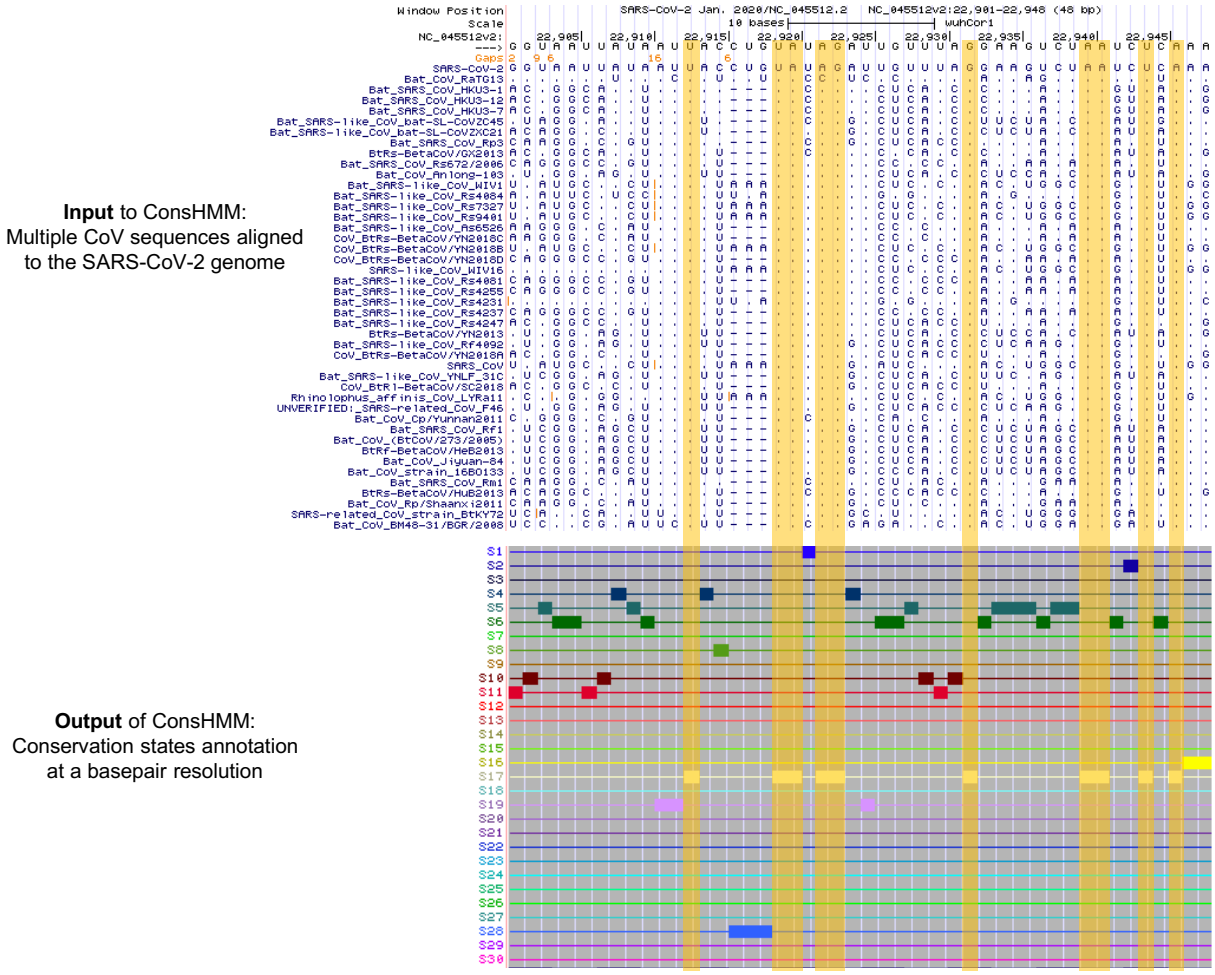


Figure 1. Genome browser view of ConsHMM input and output for a portion of the SARS-CoV-2 genome.

Shown is an example portion of the Sarbercovirus sequence alignment input to ConsHMM and ConsHMM's conservation state annotation of the SARS-CoV-2 genome as viewed in the UCSC Genome Browser. The top row of the alignments shows the reference sequence, the SARS-CoV-2 genome. This is followed by 43 rows corresponding to different Sarbecovirus sequences aligned against the reference, representing the 44-way Sarbecovirus sequence alignment. In each of these rows, a horizontal dash is shown if the row's sequence has no base that aligns to the reference base shown in the top row. A dot is shown if the sequence has the same nucleotide as the reference. A specific letter is shown if for that particular base the row's sequence has a different nucleotide than the reference. Below the alignment are 30 ConsHMM conservation states learned from the alignment.

Each row corresponds to a state. To demonstrate how bases with similar alignment patterns in the input data are annotated with the same state, bases annotated with state S17 are highlighted in yellow boxes, which have most Sarbecoviruses aligning and matching to the reference with high probabilities.

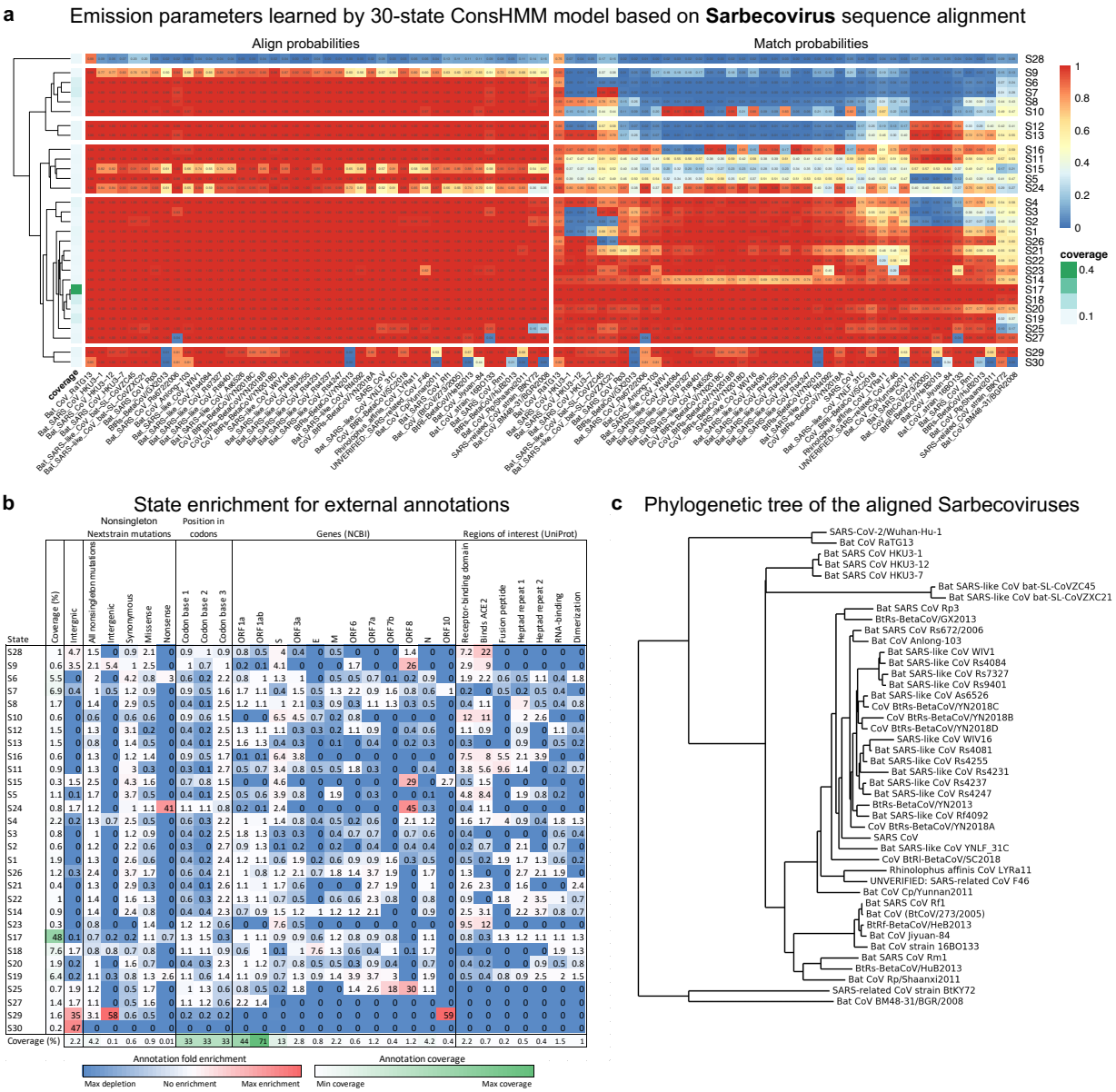


Figure 2. ConsHMM Conservation state learned from the Sarbecovirus alignment.

a. State emission parameters learned by ConsHMM. The left half of the heatmap shows the probability of a base in a CoV strain aligning to the reference, which is SARS-CoV-2. The right half shows the probability of a base in a CoV strain aligning to and matching (having the same nucleotide) the reference. In both halves, each row in the heatmap corresponds to a ConsHMM conservation state with its number on the right side of the heatmap. Rows are ordered based on hierarchical clustering and optimal leaf ordering[28]. In both halves, each column corresponds to SARS-CoV or one of the 42 CoV that infect bats. Columns are ordered based on each strain's phylogenetic distance to SARS-CoV-2,

with closer strains on the left. The column on the left shows the genome-wide coverage of each state colored according to a legend labeled “coverage” on the right.

b. State enrichment for external annotations of nonsingleton mutations, codons, genes, and regions of interest. The first column of the heatmap corresponds to each state’s genome coverage, and the remaining columns correspond to fold enrichments of conservation states for external annotations of intergenic regions, nonsingleton mutations occurring among SARS-CoV-2 sequences, position within codons, NCBI gene annotations[29], and UniProt regions of interests[30]. Each row, except the last row, corresponds to a conservation state, ordered based on the ordering shown in **a**. The last row shows the genome coverage of each annotation. Each cell corresponding to an enrichment value is colored based on its value with blue as 0 (annotation not overlapping the state), white as 1 to denote no enrichment (fold enrichment of 1), and red as the global maximum enrichment value. Each cell corresponding to a genome coverage percentage value is colored based on its value with white as 0 and green as the maximum. All annotations were accessed through UCSC Genome Browser[1], except for the mutations which were obtained from Nextstrain[23].

c. Phylogenetic tree of the Sarbecoviruses included in the alignment. Each leaf corresponds to a Sarbecovirus strain included in the 44-way Sarbecovirus alignment. This tree was obtained from the UCSC Genome Browser[1] and plotted using Biopython[31]. SARS-CoV-2/Wuhan-Hu-1, the reference genome of the alignment, is at the top.

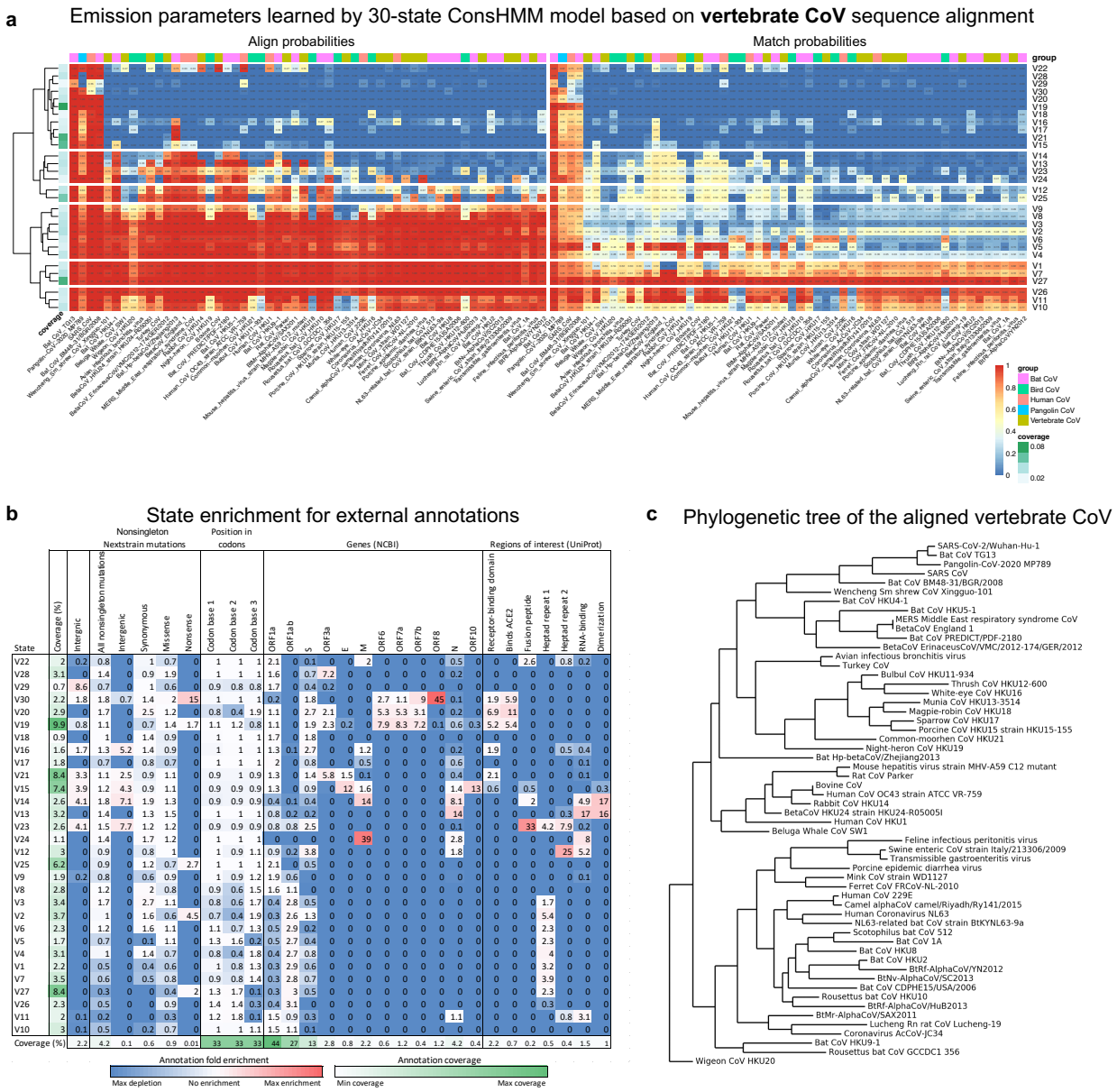


Figure 3. ConsHMM Conservation states learned from the vertebrate CoV alignment.

a. State emission parameters learned by ConsHMM. The left half of the heatmap shows the probability of a base in a CoV strain aligning to the reference, which is SARS-CoV-2. The right half shows the probability of a base in a CoV strain aligning to and matching (having the same nucleotide) the reference. In both halves, each row in the heatmap corresponds to a ConsHMM conservation state with its number on the right side of the heatmap. Rows are ordered based on hierarchical clustering and optimal leaf ordering[28]. In both halves, each column corresponds to SARS-CoV or one of the 56 CoV

that infect vertebrates. Columns are ordered based on each strain's phylogenetic distance to SARS-CoV-2, with closer strains on the left. Cells in the top row above the heatmap is colored according to the color legend on the bottom right to highlight specific groups CoV with common vertebrate hosts. The column on the left shows the genome-wide coverage of each state colored according to a legend in the bottom right.

b. State enrichment for external annotations of nonsingleton mutations, codons, genes, and regions of interest. The first column of the heatmap corresponds to each state's genome coverage, and the remaining columns correspond to fold enrichments of conservation states for external annotations of intergenic regions, nonsingleton mutations occurring among SARS-CoV-2 sequences, position within codons, NCBI gene annotations[29], and UniProt regions of interests[30]. Each row, except the last row, corresponds to a conservation state, ordered based on the ordering shown in **a**. The last row shows the genome coverage of each annotation. Each cell corresponding to an enrichment value is colored based on its value with blue as 0 (annotation not overlapping the state), white as 1 to denote no enrichment (fold enrichment of 1), and red as the global maximum enrichment value. Each cell corresponding to a genome coverage percentage value is colored based on its value with white as 0 and green as the maximum. All annotations were accessed through UCSC Genome Browser[1], except for the mutations which were obtained from Nextstrain[23].

c. Phylogenetic tree of the vertebrate CoV included in the alignment. Each leaf corresponds to a vertebrate CoV strain included in the vertebrate CoV. This tree was generated by pruning out SARS-CoV-2 genomes except the reference from the phylogenetic tree of the 119-way vertebrate CoV alignment obtained from the UCSC Genome Browser[1] (**Methods**) and was plotted using Biopython[31]. SARS-CoV-2/Wuhan-Hu-1, the reference genome of the alignment, is at the top.

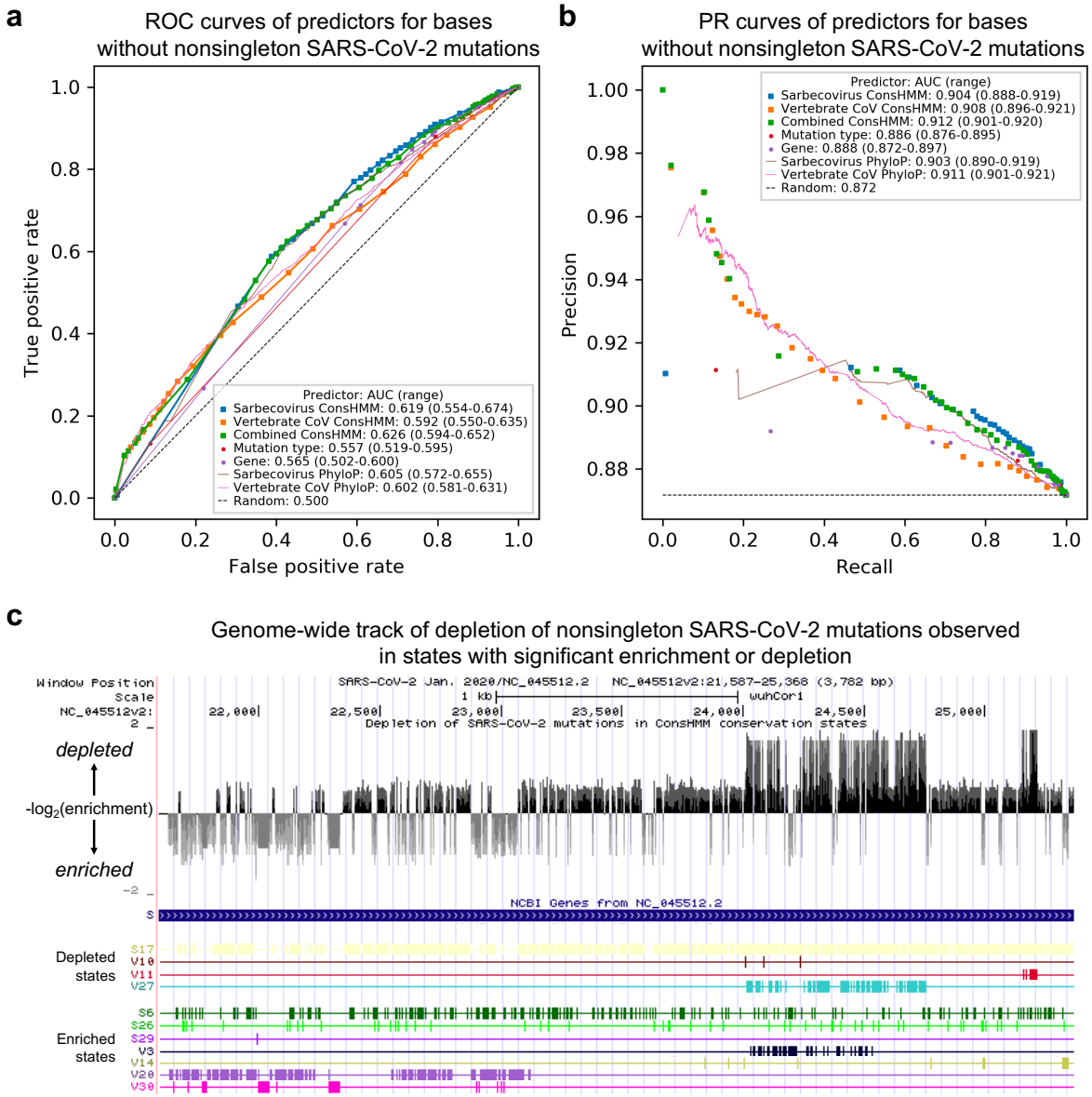

a
ROC curves of predictors for bases without nonsingleton SARS-CoV-2 mutations
True positive rate
False positive rate
Predictor: AUC (range)
Sarbecovirus ConsHMM: 0.619 (0.554-0.674)
Vertebrate CoV ConsHMM: 0.592 (0.550-0.635)
Combined ConsHMM: 0.626 (0.594-0.652)
Mutation type: 0.557 (0.519-0.595)
Gene: 0.565 (0.502-0.600)
Sarbecovirus PhyloP: 0.605 (0.572-0.655)
Vertebrate CoV PhyloP: 0.602 (0.581-0.631)
Random: 0.500
b
PR curves of predictors for bases without nonsingleton SARS-CoV-2 mutations
Precision
Recall
Predictor: AUC (range)
Sarbecovirus ConsHMM: 0.904 (0.888-0.919)
Vertebrate CoV ConsHMM: 0.908 (0.896-0.921)
Combined ConsHMM: 0.912 (0.901-0.920)
Mutation type: 0.886 (0.876-0.895)
Gene: 0.888 (0.872-0.897)
Sarbecovirus PhyloP: 0.903 (0.890-0.919)
Vertebrate CoV PhyloP: 0.911 (0.901-0.921)
Random: 0.872
c
Genome-wide track of depletion of nonsingleton SARS-CoV-2 mutations observed in states with significant enrichment or depletion
Window Position
Scale
NC_045512v2:
SARS-CoV-2 Jan. 2020/NC_045512.2
NC_045512v2:21,587-25,368 (3,782 bp)
1 kb
wuhCor1
Depletion of SARS-CoV-2 mutations in ConsHMM conservation states
depleted
-log2(enrichment)
enriched
NCBI Genes from NC_045512.2
Depleted states
Enriched states

Figure 4. Prediction of genomic bases without nonsingleton mutations in the current pandemic.

a. Receiver operating characteristic (ROC) curves of ConsHMM conservation state annotations and PhyloP scores predicting genomic bases in the SARS-CoV-2 genome without nonsingleton mutations. Positive bases are defined as bases with no mutation observed or only observed once bases, and negative bases are the remaining bases in the genome. Positive and negative bases are matched by their nucleotide by downsampling the positive bases in the reference genome. Enrichment of positive bases in each ConsHMM conservation state is computed among training bases, which is then used to rank held-out test bases through ten-fold cross-validation (Sarbecovirus ConsHMM, Vertebrate CoV ConsHMM; **Methods**). We also generate a predictor where the enrichment of positive bases in 60 conservation states from both ConsHMM models is considered to rank held-out test bases (Combined ConsHMM; **Methods**). The same procedure of using enrichment of positive training bases to make predictions on test bases is applied for an annotation of possible intergenic, synonymous, missense, and nonsense mutations and a one-hot encoded annotation of all genes, where a mutation type or gene is considered a conservation state. For PhyloP scores, which are learned from the two alignments from which the conservation states are learned (**Methods**), the score itself is used to rank test bases without any consideration of training bases. True and false positive rates computed at different thresholds for each predictor are shown in colored circles for conservation states and a colored line for PhyloP scores according to the legend on the bottom right. The colored circles are connected with lines of the same color based on piecewise linear interpolation. The legend also reports mean and range of areas under the ROC curve obtained from ten cross validation folds. Black diagonal dashed line denotes random expectation. **Supplementary Figure 3** shows additional comparisons.

b. Precision-recall (PR) curves of ConsHMM conservation state annotations and PhyloP scores predicting genomic bases in the SARS-CoV-2 genome without nonsingleton mutations. Predictions are made as explained in **a**. As mentioned in **a**, positive bases are downsampled such that the proportions of each nucleotide are the same in positive and negative bases. Precision and recall values computed at different thresholds for each predictor are shown with colored circles or a colored line according to the legend on the top right. The legend also reports mean and range areas under the PR curve, which correspond to average precision scores, obtained from ten cross validation folds (**Methods**). Black

horizontal dashed line denotes random expectation. **Supplementary Figure 3** shows additional comparisons.

c. Genome browser view of gene S with a score of depletion of nonsingleton mutations in conservation states and annotations of states from which the score is generated. Top row in black and grey vertical bars correspond to the score, which is a negative $\log_2$ of fold enrichment value of a state selected from either ConsHMM models that annotates a given base and is statistically significantly enriched or depleted of nonsingleton mutations at a genome-wide level (**Methods**). The following rows correspond to the states with significant enrichment or depletion.

Description	*State*	*Aligns to*	*Matches to*	*Notable enrichments*
Unique to SARS-CoV-2 and RaTG13	S28	RaTG13	RaTG13	Most enriched for human ACE2 binding domain
Aligns to most and matches to Sarbecoviruses closely related to SARS-CoV-2	S9	All Sarbecoviruses	RaTG13	
	S6			Enriched for mutations
	S7		Small subset of close strains including RaTG13	
	S8			Most enriched for heatpad repeat 1
	S10		Subset of strains including RaTG13 and SARS-CoV	Most enriched for spike protein's receptor binding motif (RBM)
Deviation along a branch of the Sarbecovirus phylogeny	S12	All Sarbecoviruses	Subset of strains corresponding to a subtree in the phylogeny (**Supplementary Fig. 1**)	
	S13			
Aligns to most and matches to a subset of Sarbecoviruses	S16	All Sarbecoviruses	Distinct subsets of strains with varying distance to SARS-CoV-2	Most enriched for heatpad repeat 2
	S11			Most enriched for fusion peptide
	S15			Enriched for mutations
	S5			
	S24	Most except several distal strains		Most enriched for gene ORF8
Aligns and matches to most Sarbecoviruses	S4	All Sarbecoviruses	Most except several strains	
	S3			
	S2			
	S1			
	S26			Enriched for mutations
	S21			Most enriched for dimerization-associated region
	S22			
	S14		All Sarbecoviruses	
	S23		Most except a distal strain	Most enriched for gene S
	S17		All Sarbecoviruses	Most depleted of mutations
	S18			Most enriched for gene E
	S20			
	S19		Most except several distal strains	Most enriched for gene ORF6
	S25	Most except two distal strains	Most except two distal strains	
	S27	Most except two strains	Most except two strains	
Non-coding or putative artifact	S29	Most except several close and distal strains	Most except several close and distal strains	Most enriched for intergenic bases, gene ORF10, and mutations
	S30			Most enriched for intergenic bases; No overlapping mutations

Supplementary Table 1. Summary of grouping, align and match probabilities, and notable enrichments of ConsHMM conservation states learned from the Sarbecovirus alignment.

First column contains each group's description, where a group consists of one or more states based on the hierarchical clustering of emission parameters as explained in **Fig. 2a**. Second column contains the state identifiers. Third and fourth columns describe the strains for which each state has align and match probabilities greater than 0.5, respectively. The last column summarizes notable enrichment of external annotations, as shown in **Fig. 2b**. RaTG13 refers to a bat CoV most closely related to SARS-CoV-2. All

mutations mentioned in this table are nonsingleton mutations observed in SARS-CoV-2 sequences. All enrichment and depletion reported here have a two-sided binomial test p-value significant at a 0.05 after Bonferroni correction.

Description	*State*	*Aligns to*	*Matches to*	*Notable enrichments*
Aligns and matches to four closest strains --two bat CoV (RaTG13 and BM48-31/BGR/2008), pangolin CoV, and SARS-CoV	V22	Four closest strains and several others	Four closest strains and several others	
	V28	Four closest strains except pangolin CoV	Four closest strains except pangolin CoV	Enriched for mutations
	V29	RaTG13 and SARS-CoV	RaTG13	Most enriched for intergenic bases
	V30	RaTG13 and pangolin CoV	RaTG13 and pangolin CoV	Enriched for mutations; Most enriched for gene ORF8
	V20	Four closest strains	RaTG13 and pangolin CoV	Enriched for mutations; Most enriched for human ACE2 binding domain and receptor binding motif (RBM)
	V19		Four closest strains	
	V18	Four closest strains and several others	Four closest strains	
	V16			
	V17		Four closest strain and a bat CoV	
	V21			Enriched for mutations
	V15		Four closest strains	Most enriched for gene ORF10
Aligns and matches to about half of the strains, particularly to four closest strains	V14	Up to half of strains, most close to SARS-CoV-2	Up to half of strains, most close to SARS-CoV-2	Enriched for mutations; Most enriched for dimerization-associated region
	V13			
	V23			Most enriched for fusion peptide
	V24			Most enriched for gene M
	V12			Most enriched for heatpad repeat 2
	V25			
Aligns to most and matches to some vertebrate CoV	V9	Most except several distal strains	Four closest strains	
	V8			
	V3	All vertebrate CoV		Most enriched for mutations
	V2			
	V6		Four closest strains and several distal strains, most of which are from birds	
	V5		Four closest strains with several others	
	V4			
Aligns to all and matches to most vertebrate CoV	V1	All vertebrate CoV	Most except several close strains	
	V7		All vertebrate CoV	
	V27			Depleted of mutations
Aligns and matches to most except some CoV with avian hosts	V26	All vertebrate CoV	Most except several CoV, most of which are from birds	Depleted of mutations
	V11	Most except several CoV, most of which are from birds		Most depleted of mutations
	V10			Depleted of mutations

Supplementary Table 2. Summary of grouping, align and match probabilities, and notable enrichments of ConsHMM conservation states learned from the vertebrate CoV alignment.

a Enrichment for SARS-CoV-2 mutations in states learned from the Sarbecovirus alignment

State	Enrichment for nonsingleton mutations based on genome-wide expectation (**Fig. 2b**)	Enrichment for nonsingleton mutations corrected by nucleotide composition	Enrichment for all observed mutations based on genome-wide expectation	Enrichment for all observed mutations corrected by nucleotide composition
S1	1.3	1.5	1.1	1.3
S2	1.2	1.4	1.3	1.4
S3	1.0	1.1	1.2	1.3
S4	1.3	1.4	1.3	**1.4**
S5	1.7	1.7	1.2	1.3
S6	**2.0**	**1.9**	**1.6**	**1.6**
S7	1.0	1.1	1.1	**1.2**
S8	1.4	1.5	1.3	1.4
S9	2.1	2.0	1.6	1.6
S10	0.6	0.6	0.7	0.7
S11	1.3	1.6	1.2	1.4
S12	1.3	1.3	1.3	1.4
S13	0.8	1.0	0.9	1.1
S14	1.4	1.6	1.3	1.4
S15	2.5	**3.0**	**2.3**	**2.6**
S16	1.3	1.3	1.0	1.1
S17	**0.7**	**0.7**	**0.8**	**0.7**
S18	0.8	0.7	0.9	0.8
S19	1.1	1.1	1.1	1.1
S20	1.0	1.3	1.0	1.2
S21	1.3	1.4	1.5	1.6
S22	1.4	1.6	1.0	1.2
S23	0.8	1.0	1.2	1.4
S24	1.2	1.2	1.2	1.2
S25	1.2	1.2	1.3	1.3
S26	**2.4**	**2.1**	**1.8**	**1.7**
S27	1.1	1.1	1.0	1.0
S28	1.5	1.3	**1.7**	**1.5**
S29	**3.1**	**2.8**	**2.5**	**2.3**
S30	0.0	0.0	0.0	0.0

b Enrichment for SARS-CoV-2 mutations in states learned from the vertebrate CoV alignment

State	Enrichment for nonsingleton mutations based on genome-wide expectation (**Fig. 3b**)	Enrichment for nonsingleton mutations corrected by nucleotide composition	Enrichment for all observed mutations based on genome-wide expectation	Enrichment for all observed mutations corrected by nucleotide composition
V1	0.5	0.7	0.7	0.8
V2	1.0	0.9	0.9	0.9
V3	**1.7**	1.2	**1.6**	1.2
V4	1.0	1.1	1.0	1.1
V5	0.7	0.6	0.8	0.7
V6	1.2	1.1	0.9	0.9
V7	0.6	0.8	0.8	0.9
V8	1.2	1.1	1.2	1.1
V9	0.8	0.7	0.8	0.8
V10	**0.5**	0.6	**0.7**	0.8
V11	**0.2**	**0.2**	**0.3**	**0.3**
V12	0.9	0.9	1.0	1.0
V13	1.4	1.2	1.3	1.2
V14	**1.8**	**1.6**	**1.4**	1.3
V15	1.2	1.2	1.1	1.1
V16	1.3	1.4	1.3	1.3
V17	0.7	0.7	0.8	0.9
V18	1.0	1.0	1.3	1.3
V19	1.1	1.1	1.1	1.1
V20	**1.7**	**1.5**	**1.5**	**1.4**
V21	1.1	1.1	**1.2**	**1.2**
V22	0.8	0.8	0.8	0.8
V23	1.5	1.4	1.2	1.1
V24	1.4	1.3	1.1	1.1
V25	0.9	0.9	0.9	0.9
V26	0.5	0.5	**0.6**	**0.7**
V27	**0.3**	**0.3**	**0.4**	**0.4**
V28	1.4	1.4	**1.4**	**1.4**
V29	0.7	0.7	1.1	1.0
V30	**1.8**	**1.8**	**1.6**	**1.6**

Supplementary Table 3. Conservation state enrichment for SARS-CoV-2 mutations.

a. Fold enrichment for SARS-CoV-2 mutations in conservation states learned from the Sarbecovirus model. Each row corresponds to a state. First column contains the state ID. Second column contains fold enrichment values for nonsingleton mutations currently observed in SARS-CoV-2 mutations where the enrichment is computed as the ratio between the fraction of observed mutations among possible mutations in each state and the genome-wide fraction of observed mutations among possible mutations,

as done in **Fig. 2b** (**Methods**). Third column contains fold enrichment values for the same set of nonsingleton mutations except the enrichment is corrected by the nucleotide composition of the bases annotated by each state. This corrected enrichment is computed as the ratio between the fraction of observed mutations among possible mutations in each state and the expected fraction of observed mutations among possible mutations based on relative frequencies of each nucleotide in bases annotated by the state. Fourth and fifth columns are similar to second and third columns except the enrichment values are computed based on all observed mutations instead of nonsingleton mutations. Each cell corresponding to an enrichment value is colored based on its value with blue as 0 (annotation not overlapping the state), white as 1 to denote no enrichment (fold enrichment of 1), and red as the maximum enrichment value in this table. A value is shown in bold if the associated two-sided binomial test p-value was significant at a 0.05 threshold after Bonferroni correction.

b. Similar to **a**, except based on states learned from the vertebrate CoV model. Row order in this table do not have any correspondence to row order in **a**.

a

Enrichment for protein products in states learned from the Sarbecovirus alignment

State	nsp1	pp1a	pp1ab	nsp2	nsp3	nsp4	3CL-PRO	nsp6	nsp7	nsp8	nsp9	nsp10	nsp11	Pol	Hel	ExoN	nsp15	nsp16	Spike protein S1	S glycoprotein	Spike protein S2	Spike protein S2'	ORF3a	E protein	M protein	ns6	ORF7a	ns7b	ns8	NC	ORF-9b	ORF14
S1	0.6	1.2	1.1	1	1.4	1	0.2	0.8	0.6	1.4	0.9	0.8	0	1.3	1	0.9	1	1	0.4	0.6	0.9	1.1	1.9	0.2	0.6	0.9	0.8	1.7	0.3	0.5	0.6	0.7
S2	0.9	0.9	1.3	0.3	0.9	1	0.3	0.4	0	2.2	0.5	0.8	0	2.1	2.3	3.1	1.5	0.4	0.1	0.1	0.2	0.2	0.2	0	0.2	0	0.5	0	0.5	0.4	1.6	0
S3	0.9	1.8	1.3	2.2	2.4	1.1	0.5	1.3	0.5	0.4	0.4	0	0	0.1	0.1	0.2	1.1	0.8	0.2	0.3	0.4	0.3	0.3	0	0.4	0.7	0.8	0	0.8	0.6	0.9	1.2
S4	1.5	1	1	1.2	1.1	0.9	0.2	0.5	1.3	0.5	0.5	1.4	0	0.7	0.9	0.8	1.4	1.3	1.3	1.4	1.5	1.6	0.8	0.4	0.5	0.2	0.7	0	2.2	1.2	1.4	1.7
S5	0.2	0.5	0.6	0.7	0.6	0.4	0.1	0.3	0	0.3	1.1	0.2	0	0.5	0.7	0.5	0.9	2.1	5.1	3.9	2.4	2	0.8	0	1.9	0	0.3	0	0	0.1	0	0
S6	0.5	0.8	1	0.6	0.8	0.7	0.1	0.7	2.1	1.6	1.7	1.3	2.8	1.4	1.1	1.9	1.9	0.6	1.6	1.3	1	0.9	1	0	0.5	0.5	0.7	0.1	0.3	0.9	0.6	0.3
S7	1.6	1.7	1.1	2.7	2	2	0.3	1.5	0	0	0	0	0	0	0	0	0.2	0.8	0.3	0.4	0.5	0.5	1.5	0.5	1.3	2.2	1	1.7	0.7	0.6	0.5	1.3
S8	0.7	1.2	1.1	1.5	1.3	1.3	0.3	1	1.4	0.4	1.4	0.3	0	0.6	1	1	0.8	0.8	0.8	1	1.2	1.3	2.1	0.3	0.9	0.3	0.9	1.4	0.4	0.5	0.4	0
S9	0	0.2	0.1	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	7.2	3.9	0	0	0	0	0	0.9	0	0	27	0	0	0
S10	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0.5	9.4	6.6	3.4	2.2	4.5	0.7	0.2	0.8	0	0	0	0	0	0
S11	0.4	0.5	0.7	1	0.5	0.4	0	0.3	0.9	0.2	0.6	0.5	0	0.4	1.1	1.3	1.4	1.3	4.4	3.5	2.4	2.2	0.8	0.5	0.5	1.8	0.3	0	0	0.4	0.4	0
S12	1	1.3	1.1	1.7	1.3	1.4	0.3	1.1	2.2	0.8	0.4	0.7	0	0.8	0.5	1.4	1	0.7	1.1	1.1	1.2	1	0.3	0.3	0.2	1.1	0.9	0	0.2	0.6	0.9	0.3
S13	0.4	1.6	1.3	1.4	1.8	1.8	0.3	1.7	0.8	1.1	1.2	1.3	0	0.7	1.1	0.8	1	0.7	0.2	0.4	0.5	0.5	0.3	0	0.1	0	0	0	0.2	0.3	0.5	0.6
S14	1	0.7	0.9	1.3	0.7	0.3	0.1	0.8	1.3	0.4	0.3	0.5	0	0.9	1	1.2	2	2.5	1.8	1.5	1.2	1.1	1.2	1	1.2	1.2	2.1	0	0	0.9	0.4	2
S15	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	8.2	4.4	0	0	0	0	0	0	0	0	33	0	0	0
S16	0	0.1	0.1	0	0.1	0	0	0	0	0	0	0.4	0	0	0	0	0	1.6	7.3	6.4	5.4	3.8	3.8	0	0	0	0	0	0	0	0	0
S17	1	1	1.1	0.9	0.9	1.2	0.3	1.2	1.2	1	1	1.1	0.4	1	1.3	1.2	1.2	1.3	0.7	0.9	1.2	1.2	0.9	0.6	1.2	0.9	0.9	0.8	0	1.1	1	0.9
S18	1.2	0.6	1	0	0.1	0.4	0.3	0.9	1	3.1	3.5	3.2	7.8	3.4	1.1	1.1	0.7	0.3	0	0.1	0.2	0.3	0.9	7.5	1.3	0.4	0.2	1	0	1.7	2.2	1.9
S19	1.7	1.1	0.9	1.5	1.1	0.9	0.2	0.9	0.4	0.4	0.5	0.4	0.8	0.6	1	0.7	0.9	0.6	0.5	0.7	1	0.9	1.3	0.9	1.4	3.9	3.8	3	0	1.9	2	2.6
S20	1.2	1	1.2	1	1	0.9	0.2	1	2.1	1.2	0.3	1.4	0	1.5	1.3	1.8	1	1.2	0.5	0.7	0.9	1	0.8	0.5	0.5	0.3	1	0.4	0	0.7	0.7	0.2
S21	2.3	1.1	1	1.4	0.9	2	0.1	0.3	0	1.3	0	1.8	6.4	0.3	1.4	0.6	0.7	1.4	2.4	1.7	1	1.3	0.6	0	0	0	2.4	1.9	0	1	0.9	0
S22	1.9	1.3	1.1	2.2	1.1	1.3	0.4	0.5	1.7	1.6	0	1.2	0	0.5	1.2	0.9	1.2	0.8	0.5	0.7	1.1	1.1	0.5	0	0.6	0.6	1.6	0.8	0	0.8	1.1	1.4
S23	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	13	7.3	0.4	0.3	0.5	0	0	0	0	0	0	0	0	0
S24	0	0.2	0.1	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	4.2	2.3	0	0	0	0	0	0	0	0	47	0.3	0	0
S25	0	0.8	0.5	0.7	1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3	0.2	0.2	0.3	1.8	0	0	1.5	3	16	26	1.1	1.4	1.2
S26	0.8	1	0.8	1	1.2	1.6	0.4	0.8	0.3	0	0.7	0	0	0.1	0.6	0.2	0.8	1.7	1	1.3	1.6	1.6	2.1	0.7	1.8	1.4	3.2	1.9	0	1.7	2.9	1.5
S27	0	2.2	1.4	0	4.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S28	0	0.8	0.5	0.6	1.6	0	0	0	0	0	0	0	0	0	0	0	0	0	7	3.7	0	0	0.4	0	0.5	0	0	0	1.6	0	0	0
S29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
S30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Coverage	1.8	44	71	6.4	20	5	12	2.9	0.8	2	1.1	1.4	0.1	9.3	6	5.3	3.5	3	6.8	13	5.9	4.6	2.8	0.8	2.2	0.6	1.1	0.4	1.1	4.2	1	0.7

b

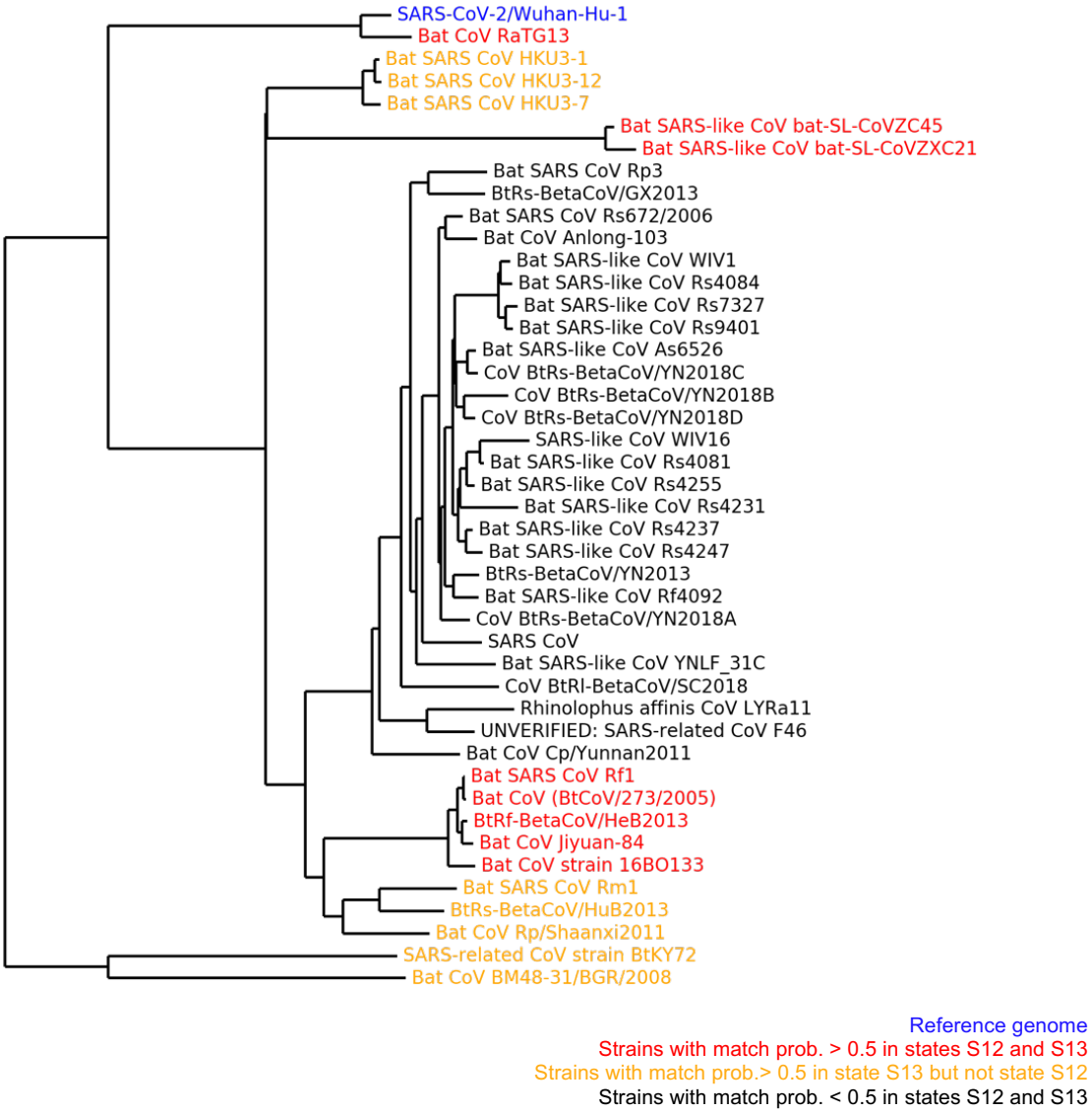
Enrichment for protein products in states learned from the vertebrate CoV alignment

State	nsp1	pp1a	pp1ab	nsp2	nsp3	nsp4	3CL-PRO	nsp6	nsp7	nsp8	nsp9	nsp10	nsp11	Pol	Hel	ExoN	nsp15	nsp16	Spike protein S1	S glycoprotein	Spike protein S2	Spike protein S2'	ORF3a	E protein	M protein	ns6	ORF7a	ns7b	ns8	NC	ORF-9b	ORF14
V1	0	0.3	1.3	0	0	0	0.5	0.6	0	1	0.4	3.7	0	2.9	3.1	3.1	2.5	2.4	0	0.6	1.2	1.5	0	0	0	0	0	0	0	0	0	0
V2	0	0.3	1.2	0	0	0	0.4	0.6	0.2	0.6	0.7	3.2	0	2.7	2.8	2.3	2.7	1.9	0	1.3	2.8	3.6	0	0	0	0	0	0	0	0	0	0
V3	0	0.4	1.3	0	0	0	0.5	1.2	0.1	0.5	0.9	3.6	2.3	3.3	2.9	2.8	2.3	2.1	0	0.5	1.1	1.4	0	0	0	0	0	0	0	0	0	0
V4	0	0.4	1.3	0	0	0	0.4	1.3	0.1	0.8	0.9	3.3	1.7	2.8	3	2.8	2.6	2	0	0.8	1.8	2.3	0	0	0	0	0	0	0	0	0	0
V5	0	0.5	1.3	0	0	0.7	0.5	0.9	1.9	2.1	0.5	1.8	1.5	3	3	2.1	2.6	2.2	0	0.4	0.9	1.2	0	0	0	0	0	0	0	0	0	0
V6	0	0.5	1.4	0	0	0	0.9	0.4	0.5	1.1	1.2	3.2	0	2.6	3.2	2.8	4.2	1.7	0	0.2	0.3	0.4	0	0	0	0	0	0	0	0	0	0
V7	0	0.3	1.3	0	0	0	0.5	0.7	0.2	1	0.8	2.8	0.7	3	2.8	2.6	2.8	2.5	0	0.7	1.5	2	0	0	0	0	0	0	0	0	0	0
V8	0	1.6	1.4	0	0.3	1.4	0.7	3.7	11	10	7.2	0.2	0	0.6	0.3	1.9	0.1	3.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V9	0	1.9	1.4	0	0.7	0.1	1.2	15	3	1.3	4.9	0.3	11	0.3	0.1	0.7	1.5	1.6	0	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0
V10	0	1.5	1.4	0	0.4	0.9	0.9	4.2	7.6	7.8	8	0.1	1.7	0.5	0.6	1.5	1.4	3.3	0	0.1	0.1	0.1	0	0	0	0	0	0	0	0	0	0
V11	0	1.5	1.3	0	1	1.6	0.8	1.6	7.1	6.1	4.9	0.6	2.5	0.4	0.2	1.3	0.7	3.6	0	0.3	0.6	0.8	0	0	0	0	0	0	0	1.1	0	0
V12	0	0.9	0.6	0	1.6	0.1	0	1.2	1.3	0.3	0.9	0	0	0.1	0.1	0.1	1.3	0	0	3.8	8.2	10	0	0	0	0	0	0	0	1.8	0.2	0
V13	0	0.8	0.5	0	0.8	2.8	0.1	0.6	0	0	0.1	0	4.8	0.2	0.1	0.3	0.3	0.3	0	0.2	0.5	0.2	0	0	0	0	0	0	0	14	10	26
V14	0	0.4	0.3	0	0.5	1.1	0.1	0.1	0	0	0.1	0.3	0	0	0	0.2	0.4	0	0	0.4	0.9	1.1	0	0	14	0	0	0	0	8.2	4.8	6.2
V15	0	1.3	0.8	1.3	2.5	0	0	0	0	0	0	0	0	0	0	0	0	0	1.4	0.9	0.3	0.3	0	11	1.6	0	0	0	0	1.4	1	0
V16	0	1.3	0.8	0	2.7	0.5	0	0.5	0	0	0	0	0	0	0.1	0	0.4	0	2.8	2.1	1.4	1.5	0	0	1.2	0	0	0	0	0.2	0.7	0
V17	0	1.9	1.2	0	4.4	0	0	0	0	0	0	0	0	0	0	0	0	0	1.5	0.8	0	0	0	0	0.5	0	0	0	0	0.6	0	0
V18	0	1.7	1.1	2.1	3.3	0	0	0	0	0	0	0	0	0	0	0	0	0	3.4	1.8	0	0	0	0	0	0	0	0	0	0	0	0
V19	8.7	1.1	0.7	3.6	0.4	0	0	0	0	0	0	0	0.5	0	0	0	0	0	3.5	1.9	0	0	2.3	0.2	0	7.9	8.3	7.4	0	0.6	1.5	0
V20	4.8	1.1	0.7	4.4	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	2.7	0	0	2.1	0	0	5.4	5.1	3.2	0	0.2	0.6	0
V21	0	1.3	0.8	1.6	2.4	0	0	0	0	0	0	0	0	0	0	0	0	0	2.2	1.4	0.5	0	5.8	1.5	0.1	0	0	0	0	0.4	1.7	0
V22	0	2.1	1.3	0	4.7	0	0	0	0	0	0.3	0	0	0	0	0	0	0	0	0.1	0.3	0.1	0	0	2	0	0	0	0	0.5	0	0.7
V23	0	0.8	0.8	0	0.1	0.4	0.4	6.3	4.1	1.2	2.8	1	0	0.3	0.6	1.1	2.5	0.1	0	2.6	5.5	3	0	0	0	0	0	0	0	0.1	0.3	0
V24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	2.8	12	0
V25	0	2.1	1.3	0	1.7	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	1.1	0	0	0	0	0	0	0	0	0	0	0
V26	0	0.4	1.4	0	0	0	0.4	0.7	0.7	1.2	1	3	1.1	3.1	3.2	3.1	2.8	3.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V27	0	0.3	1.3	0	0	0	0.4	0.6	0.3	0.8	0.8	2.9	3.4	3.4	3.4	2.6	2	2.2	0	0.5	1.1	1.4	0	0	0	0	0	0	0	0	0	0
V28	0	1.6	1	8.2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1.2	0.7	0.1	0.1	7.2	0	0	0	0	0	0	0.2	0	0
V29	0	1.7	1.1	2.8	2.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0.4	0	0	0.2	0	0	0	0	0	0	0	0	0
V30	0	0.2	0.2	0.1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	3.4	1.8	0	0	0	0	0	2.7	1.3	8.1	45	0.1	0.5	0
Coverage	1.8	44	71	6.4	20	5	12	2.9	0.8	2	1.1	1.4	0.1	9.3	6	5.3	3.5	3	6.8	13	5.9	4.6	2.8	0.8	2.2	0.6	1.1	0.4	1.1	4.2	1	0.7

Supplementary Table 4. Conservation state enrichment for protein products.

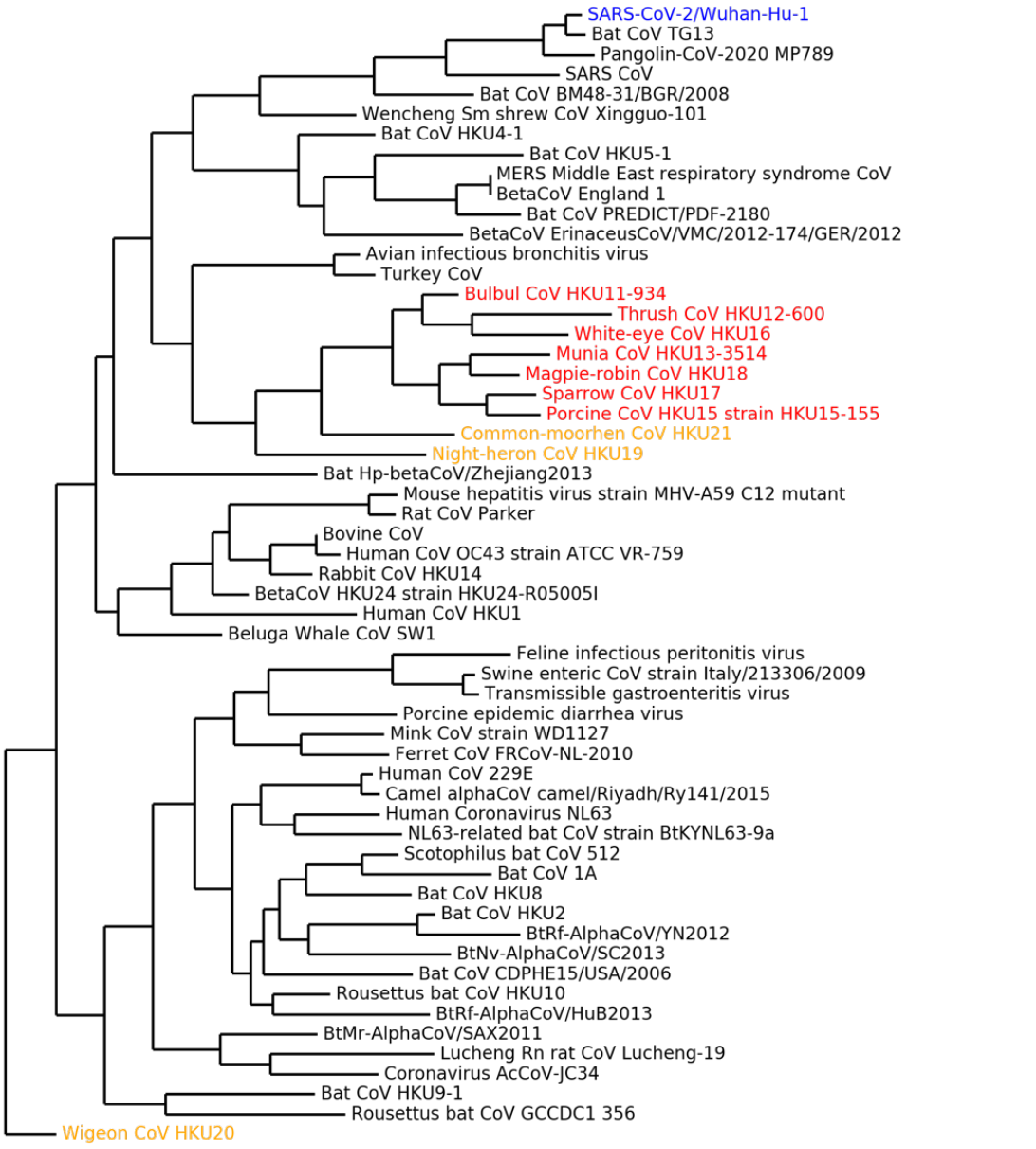
a. Fold enrichment for protein products in conservation states learned from the Sarbecovirus model. Each row corresponds to a state. First column contains the state ID. The following columns contain fold enrichment values for different protein products listed at the top of each column. Protein product coordinates and names were from UniProt Protein Product annotation[30]. Last row reports genome coverage percentage of each protein. Each cell corresponding to an enrichment value is colored based on its value with blue as 0 (annotation not overlapping the state), white as 1 to denote no enrichment (fold enrichment of 1), and red as the maximum enrichment value in this table. Each cell corresponding to a coverage percentage is colored based on its value with white as minimum and green as maximum.

b. Similar to **a**, except based on states learned from the vertebrate CoV model.



Supplementary Figure 1. Sarbecoviruses associated with states S12 and S13 in the phylogenetic tree of the 44-way Sarbecovirus alignment. Similar to **Fig. 2c** except strains colored according to their align and match probabilities in states S12 and S13. The strain colored in blue is the reference SARS-CoV-2 strain of the alignment, SARS-CoV-2/Wuhan-Hu-1. Strains colored in black are those that have match probabilities below 0.5 for both states S12 and S13. Strains colored in red are those with match probabilities above 0.5 for both states S12 and S13. Strains colored in yellow are those with match probabilities above 0.5 for state S13 but not for state S12. All strains have high

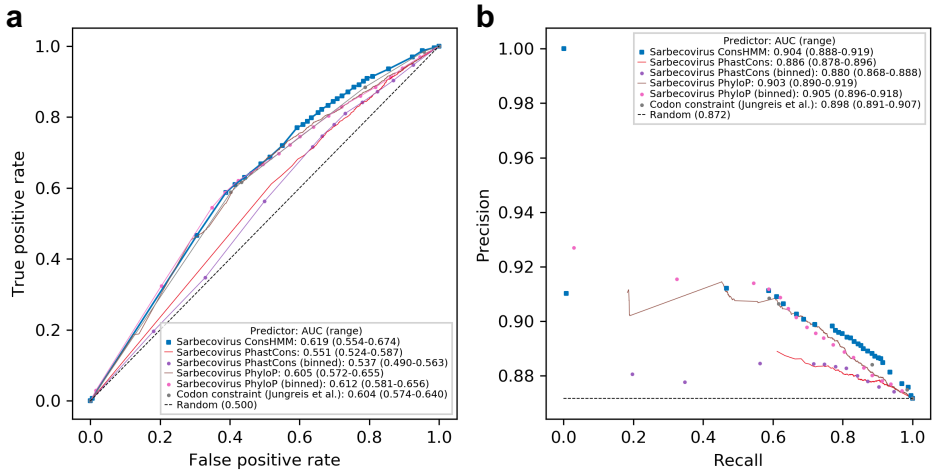
(>0.95) align probabilities for states S12 and S13. States S12 and S13 are likely to correspond to a deviation along the branch preceding all strains colored in black.



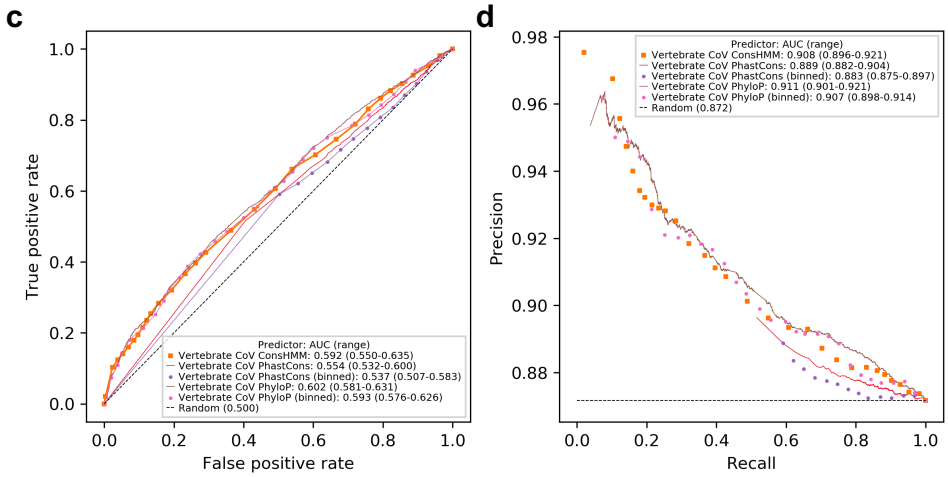
Supplementary Figure 2. Vertebrate CoV associated with state V11 in the phylogenetic tree of the vertebrate CoV alignment. Similar to **Fig. 3c** except strains colored according to their align and

match probabilities in state V11. The strain colored in blue is the reference SARS-CoV-2 strain of the alignment, Wuhan-Hu-1. The strains colored in red are those with both align and match probabilities above 0.5 for state V11, which include six CoV from avian hosts and a CoV from pig. The strains colored in yellow are those with align probabilities above 0.5 and match probabilities below 0.5 for state V11. The remaining strains in black are those with align and match probabilities above 0.5 for state V11.

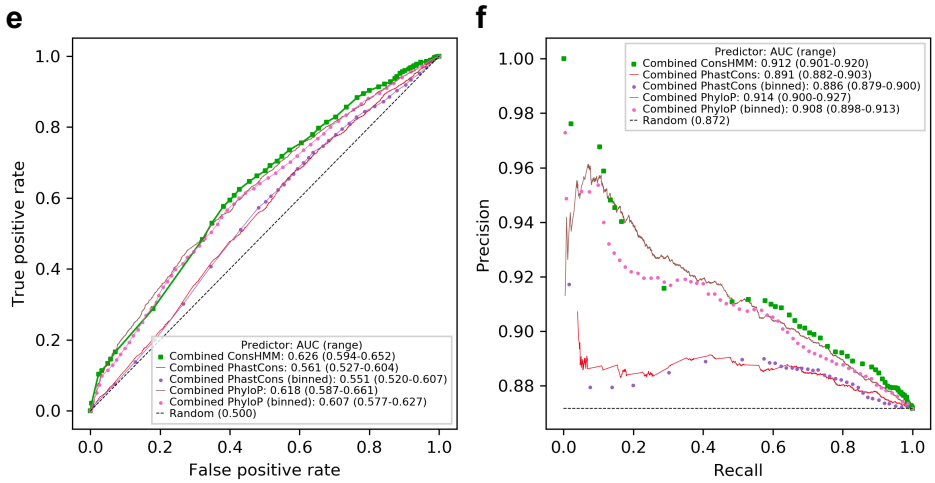
Predictive power of annotations learned from **Sarbecovirus** alignment for bases without nonsingleton SARS-CoV-2 mutations



Predictive power of annotations learned from **vertebrate CoV** alignment for bases without nonsingleton SARS-CoV-2 mutations



Predictive power of predictors that merge annotations from both alignments for bases without nonsingleton SARS-CoV-2 mutations



Supplementary Figure 3. Comparison of predictive power of ConsHMM conservation states and sequence constraint scores for genomic bases without nonsingleton mutations in the current pandemic.

a. Receiver operating characteristic (ROC) curves of ConsHMM conservation states and sequence constraint scores, all of which are learned from the 44-way Sarbecovirus sequence alignment, predicting genomic bases in the SARS-CoV-2 genome without nonsingleton mutations. Positive bases are defined as genomic bases with a mutation found in at most one SARS-CoV-2 strain and thus include bases with no mutation observed at all and also bases with a singleton. Negative bases are the remaining bases of the genome. Positive and negative bases are matched by their nucleotide in the reference genome. Enrichment of positive bases in each ConsHMM conservation state is computed among training bases, which is then used to rank held-out test bases through ten-fold cross-validation (**Methods**). For PhastCons and PhyloP scores, the score itself is used to rank test bases without any consideration of training bases (Sarbecovirus PhastCons, Sarbecovius PhyloP). Additionally, for each score, we rank the training bases based on the score and partition the training bases into 30 equal-sized bins based on the ranking. We then treat the bins as conservation states, using the enrichment of positive training bases in each bin to rank the test bases (Sarbecovirus PhastCons (binned), Sarbecovirus PhyloP (binned); **Methods**). Lastly, we also include as categorical predictor five categories of mutations from Jungreis et al.[16], which were defined based on mutation type and codon constraint learned from a different sequence alignment of the same set of Sarbecoviruses. Each ROC curve is colored according to the legend in the bottom right. The legend also reports mean and range of areas under the ROC curve obtained from ten cross validation folds. Black diagonal dashed line denotes random expectation. A subset of the curves shown here are the same as those shown in **Figure 4a**.

b. Precision-recall (PR) curves of ConsHMM conservation states and sequence constraint scores, all of which are learned from the 44-way Sarbecovirus sequence alignment, predicting genomic bases in the SARS-CoV-2 genome without nonsingleton mutations. Predictions are made as explained in **a**. Precision and recall values computed at different thresholds for each predictor are shown with colored circles or a colored line according to the legend on the top right. The legend also reports mean and

range of areas under the PR curve obtained from ten cross validation folds. Black horizontal dashed line denotes random expectation. A subset of the curves shown here are the same as those shown in **Figure 4b**.

c-d. Similar to **a-b**, respectively, except showing results based on conservation states and sequence constraint scores learned from the vertebrate CoV alignment instead of the Sarbecovirus alignment.

e-f. Similar to **a-b**, respectively, except showing results based on predictors that combine annotations learned from the Sarbecovirus and vertebrate CoV alignments.