UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Comparative and Epigenomic Approaches to Improve Genome Interpretation

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Bioinformatics

by

Soo Bin Kwon

2021

ABSTRACT OF THE DISSERTATION

Computational Comparative and Epigenomic Approaches

to Improve Genome Interpretation

by

Soo Bin Kwon

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2021

Professor Jason Ernst, Chair

Systematic analysis of sequence or mappings of biochemical activities can reveal biologically relevant information that may be otherwise overlooked. Such information can be elusive in a large collection of genomic data from varied sources. We therefore propose and apply computational methods that detect complex relationships among data from different genomic loci within or across genomes and generate annotations that highlight notable patterns.

First, we focus on locating genomic regions with conserved properties by scoring cross-species similarity between two regions from different species based on their functional genomic datasets. To do so, we develop a method, Learning Evidence of Conservation from Integrated Functional genomic annotations (LECIF). When we apply LECIF to thousands of human and mouse datasets, we learn a score that highlights human and mouse loci with shared properties, which is expected to be useful in mouse model research.

Building on this work, we also develop a method that scores association between two regions within the same genome based on epigenomic and TF binding data. We apply this approach to thousands of human datasets and learn a score that highlights regions with similar

or associated properties within human, which we expect to be useful in studying multiple loci together.

Lastly, motivated by the COVID-19 pandemic, we apply an existing comparative genomics approach to coronavirus sequences and annotate the SARS-CoV-2 genome. Specifically, we apply ConsHMM, a hidden Markov model method that learns conservation states that capture recurring patterns in an alignment of sequences, to alignments of coronavirus sequences. We then analyze the learned state annotations using external annotations of genes, protein domains, SARS-CoV-2 mutations, and other regions of interest and demonstrate that the states reflect biologically relevant information for interpreting the SARS-CoV-2 genome.

Overall, our work aims to learn meaningful patterns in large genomic datasets from diverse sources and provide annotations for interpreting important DNA elements and their relationships. All methods we present are flexible and scalable, making them applicable to newer and larger datasets that will be made available in the future. We expect our methods and genomic annotations to be useful resources for studying various genomes.

The dissertation of Soo Bin Kwon is approved.

Jingyi Li

Stefan Horvath

Sriram Sankararaman

Jason Ernst, Committee Chair

University of California, Los Angeles

2021

DEDICATION

This dissertation is dedicated to my parents for their love and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 2

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my advisor Dr. Jason Ernst. He has been an extremely reliable and committed advisor. He taught me what it means to be a good scientist by example with his rigor, integrity, and thoroughness. I thank him for encouraging me to delve deeper into my projects and giving me the space to grow as a researcher.

I would like to thank the rest of my committee, Drs. Jessica Li, Steve Horvath, and Sriram Sankararaman. I learned a great deal from rotating in labs, taking their courses, and collaborating with their labs. I appreciate their encouragement and constructive feedback on my work. I am grateful to the past and present student affair officers of the UCLA Bioinformatics Interdepartmental Program for their assistance. I would also like to acknowledge UCLA's Graduate Division for providing me with the Dissertation Year Fellowship.

I am thankful to the former and current members of the Ernst lab for their friendship and thoughtful discussions. I also want to acknowledge Dr. Ahmet Ay for his invaluable mentorship during my college years. Lastly, I am deeply grateful to my parents and grandmothers for their support throughout my academic career. I cannot thank them enough for their sacrifices and prayers.

Chapter 2 is a version of Kwon, S. B. & Ernst, J. Learning a genome-wide score of human-mouse conservation at the functional genomics level. *Nat. Commun.* **12**, 2495 (2021). We thank the ENCODE, Mouse ENCODE, Roadmap Epigenomics, and FANTOM consortia for generating data and making it publicly available. We acknowledge funding from US National Institutes of Health (DP1DA044371 and U01MH105578 to J.E.); US National Science Foundation (CAREER Award #1254200 to J.E.); Kure It Cancer Research (Kure-IT award to J.E.), and a Rose Hills Innovator Award (J.E.).

Chapter 3 is a version of a manuscript in preparation for publication. It is authored by Kwon S.B. and Ernst J. We thank the ENCODE and Roadmap Epigenomics consortia for generating data and making it publicly available.

Chapter 4 is a version of Kwon, S. B. & Ernst, J. Single-nucleotide conservation state annotation of the SARS-CoV-2 genome. *Commun. Biol*. **4**, 698 (2021). We gratefully acknowledge all those who contributed to generating and sharing their SARS-CoV-2 sequence data via GISAID. 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, the US National Institutes of Health (DP1DA044371), and the National Science Foundation (2125664).

VITA

EDUCATION

2021	Ph.D. Candidate, Bioinformatics University of California, Los Angeles
2016	B.A. Computer Science, Music Colgate University, Hamilton, NY

RESEARCH EXPERIENCE

2016 – Present	Ph.D. trainee Department of Biological Chemistry University of California, Los Angeles
2020	Bioinformatics Intern Genomics Institute of the Novartis Research Foundation

ACADEMIC FELLOWSHIPS

2021	UCLA Dissertation Year Fellowship

RESEARCH PUBLICATIONS

Kwon, S. B. & Ernst, J. Single-nucleotide conservation state annotation of the SARS-CoV-2 genome. *Commun. Biol.* **4**, 698 (2021).

Kwon, S. B. & Ernst, J. Learning a genome-wide score of human-mouse conservation at the functional genomics level. *Nat. Commun.* **12**, 2495 (2021).

Grujic, O. *et al.* Identification and characterization of constrained non-exonic bases lacking predictive epigenomic and transcription factor binding annotations. *Nat. Commun.* **11**, 6168 (2020).

Ge, X. *et al.* EpiAlign: an alignment-based bioinformatic tool for comparing chromatin state sequences. *Nucleic Acids Res.* **47**, e77–e77 (2019).

Keskin, S. *et al.* Noise in the Vertebrate Segmentation Clock Is Boosted by Time Delays but Tamed by Notch Signaling. *Cell Rep.* **23**, (2018).

Liberman, A. R. *et al.* Circadian Clock Model Supports Molecular Link Between *PER3* and Human Anxiety. *Sci. Rep.* **7**, 9893 (2017).

Ingram, K. K. *et al.* Molecular insights into chronotype and time-of-day effects on decision-making. *Sci. Rep.* **6**, 29392 (2016).

REVIEWS AND COMMENTARIES

Palmer, R. H. C. et al. Integration of evidence across human and model organism studies: A meeting report. *Genes, Brain Behav*. **20**, e12738 (2021).

Kwon, S. B. & Ernst, J. Investigating enhancer evolution with massively parallel reporter assays. *Genome Biol*. **19**, 114 (2018).

MANUSCRIPTS IN REVIEW

Arneson, A. *et al.* A mammalian methylation array for profiling methylation levels at conserved sequences. *bioRxiv* (2021). doi:10.1101/2021.01.07.425637

MANUSCRIPTS IN PREPARATION

Kwon, S. B. & Ernst, J. Learning a pairwise epigenomic and TF binding association score across the human genome.

Chapter 1. Introduction

Large amounts of genomic sequences and genome-wide mappings of open chromatin, histone marks, transcription factor binding, and transcription have become available[1–4]. Identifying notable patterns in such data within a genome or across multiple genomes can provide novel insight for genome interpretation. This provides an opportunity to develop and apply systematic computational methods that can leverage large-scale genomic data from diverse sources and highlight notable patterns in an interpretable manner.

In **Chapter 2** we present Learning Evidence of Conservation from Integrated Functional genomic annotations (LECIF). We were motivated by the challenge of discovering genomic loci with properties shared by two species when there are large collections of functional genomic annotations for both species from diverse cell types and assays. Given experiments between human and mouse, for example, matching the experiments across species by origin and data type can be difficult and sometimes infeasible. Moreover, comparing experiments that do not necessarily match by their data type or source may reveal additional information. LECIF thus scores similarity in pairs of regions from two species by comparing their functional genomic datasets without requiring them to be matched across species. When applied to human and mouse, LECIF captures correspondence of similar human and mouse experiments without prior knowledge and highlights human and mouse loci with similar properties, which we expect to be useful in mouse model research.

In **Chapter 3** we modify LECIF to analyze epigenomic data within one species, focusing on within-species association rather than across-species conservation. When studying multiple genomic loci jointly for their role in gene regulation or contribution to disease risk, it is helpful to examine multiple epigenomic datasets to infer which regions are associated or similar. We thus modify LECIF to score pairwise association between two genomic windows within a species with their distances varying from 1 kb to 100 kb. When applied to human, the modified approach learns a score that highlights loci with similar or associated properties, such as loci within the same gene

or potential enhancer-gene pairs, which may be useful in understanding how linearly distal loci closely cooperate to influence gene regulation or disease risk.

In **Chapter 4** we apply an existing comparative genomics method to alignments of coronavirus sequences to annotate the SARS-CoV-2 genome. Ernst lab previously developed ConsHMM, which applies a hidden Markov model to multi-species sequence alignment to learn conservation states that summarize recurring patterns in the alignment. Motivated by the recent need to better understand the SARS-CoV-2 genome, we apply ConsHMM to alignments of coronaviruses and learn conservation state annotations. To understand the information captured by the state annotations, we study them with respect to external annotations and prior studies that were not provided as input to ConsHMM. We observe that many of the learned states capture information relevant to genes, host interaction, pathogenicity, and SARS-CoV-2 mutations accumulating in the current pandemic.

Chapter 2. Learning a genome-wide score of human-mouse conservation at the functional genomics level

Abstract

Identifying genomic regions with functional genomic properties that are conserved between human and mouse is an important challenge in the context of mouse model studies. To address this, we develop a method to learn a score of evidence of conservation at the functional genomics level by integrating information from a compendium of epigenomic, transcription factor binding, and transcriptomic data from human and mouse. The method, Learning Evidence of Conservation from Integrated Functional genomic annotations (LECIF), trains neural networks to generate this score for the human and mouse genomes. The resulting LECIF score highlights human and mouse regions with shared functional genomic properties and captures correspondence of biologically similar human and mouse annotations. Analysis with independent datasets shows the score also highlights loci associated with similar phenotypes in both species. LECIF will be a resource for mouse model studies by identifying loci whose functional genomic properties are likely conserved.

Introduction

Many studies interrogate human loci of interest, such as those implicated in genome-wide association studies (GWAS), by perturbing their homologous loci in mouse[5–8]. A key question in this context is the extent to which the homologous loci in mouse is expected to have similar roles to the human loci. Conversely, loci associated with phenotypes can be discovered in mouse first, raising the question of the degree to which their properties are shared with human[9].

A relatively large percentage of the human genome, approximately 40%, has a homologous locus in the mouse genome as determined by human-mouse pairwise sequence alignment[10]. However, a much smaller fraction of bases in these aligning pairs of loci are

constrained at the sequence level[11–14]. This is because many bases are within regions whose sequences are similar enough to be aligned between species, but not necessarily constrained, which is defined at a higher resolution and generally has even greater sequence similarity. In general, it is unclear to what extent human and mouse loci that align to each other have similar properties, in particular, functional genomic properties. With large-scale functional genomic resources of genome-wide maps of chromatin accessibility, transcription factor binding, histone modifications, gene expression data across diverse cell and tissue types that have become available in mouse[2,15] in addition to human[1,4,16], there is an opportunity to systematically and confidently detect evidence of conservation at the functional genomics level between these species.

Previous work comparing cross-species functional genomics data to infer conservation have largely focused on comparing pairs of matched experiments for the same assay in a corresponding cell or tissue type across species[17–22]. While useful, data from a pair of experiments from two species provides limited information for differentiating evidence of conservation from similarity observed by chance. Studies that jointly compare multiple pairs of experiments from different biological conditions have additional information available for inferring conservation of functional genomic properties[18,19,21,22]. However, such approaches have often relied on manually matching corresponding experiments and have not been scaled to leverage the vast amounts of diverse data available in both human and mouse. The challenge in taking advantage of such data is that many experiments do not have an obvious corresponding experiment, and even when one is assumed there could in practice be confounding differences. Previous work partly addressed some of these issues[2,23–28], but still limited their work to one data type at a time and thus only utilized a small fraction of the available data to find evidence of conservation. Given the increasingly diverse functional genomic resources available for human and mouse, there is a need for an integrative method to better leverage those resources to infer evidence of conservation at the functional genomics level between human and mouse.

Thus, here we develop Learning Evidence of Conservation from Integrated Functional genomic annotations (LECIF), a supervised learning approach that quantifies evidence of conservation based on large-scale functional genomic data from a pair of species, which we apply to human and mouse. While LECIF leverages data from diverse cell types collected by various assays, it does not require explicit matching of experiments from different species by biological source or data type. LECIF uses pairwise sequence alignment data only to label training examples, inferring conservation from functional genomics data and not from DNA sequence. We apply LECIF to a compendium of thousands of human and mouse functional genomic annotations and learn the LECIF score for every pair of human and mouse regions that align at the sequence level. The score captures correspondence of biologically similar annotations between human and mouse, even though LECIF was not explicitly given such information. While the LECIF score is moderately correlated with sequence constraint scores, it captures distinct information on conserved properties. The LECIF score is preferentially higher in regions previously shown to have similar phenotypic properties in human and mouse at the genetic and epigenetic level. Overall, we observe that the score can complement sequence conservation annotations in capturing human-mouse conservation and contribute to locating pairs of sequence-aligning regions whose functional genomic properties are likely conserved. We thus expect the human-mouse LECIF score will be an important resource for studies using mouse as a model organism.

Results

Overview of LECIF

LECIF quantifies evidence of conservation between human and mouse genomic regions at the functional genomics level based on a large and diverse set of functional genomic annotations (**Fig. 2.1**). LECIF uses functional genomic features as input to an ensemble of neural networks where sequence alignment information is used to label training data, but not as features

(**Methods**). For training data, positive examples are pairs of human and mouse regions that align at the sequence level while negative examples are randomly mismatched pairs of human and mouse regions that do not align to each other (**Fig. 2.1a**). All human and mouse regions included in negative examples align somewhere in the mouse and human genomes, respectively, which allows LECIF to learn pairwise characteristics of aligning human and mouse regions instead of the characteristics of regions that align to the other genome in general. LECIF assumes that positive examples are more likely to be conserved at the functional genomics level than negative examples. Since neighboring bases are likely annotated by the same annotations and for computational considerations, training examples and predictions were generated at every 50 bp within each pairwise alignment block (**Methods**). As a result, we provided the classifier with more than >2 positive and >2 negative training examples, which covered up to 90 Mb of the human and mouse genomes.

For each example, there were >8000 human and >3000 mouse functional genomic features defined. Among these features were binary features corresponding to whether a genomic base overlapped with peak calls from DNase-seq experiments, ChIP-seq experiments of transcription factors (TF), histone modifications and histone variants, and Cap Analysis of Gene Expression (CAGE) experiments. Additionally, there were binary features corresponding to each state and tissue combination of ChromHMM[29] chromatin state annotations and numerical features corresponding to normalized signals from RNA-seq experiments. These data covered a wide range of cell and tissue types and were generated by the ENCODE[1], Mouse ENCODE[2], Roadmap Epigenomics Project[4], or FANTOM5[30] consortia (**Methods**; **Supplementary Data 2.1**). We did not provide pairwise alignment or DNA sequence information as features to the classifier so that LECIF infers conservation specifically at the functional genomics level rather than at the sequence level.

After training, we used the classifier to make genome-wide predictions at 50 bp resolution or finer, annotating the 40% of the human genome that aligns to mouse and those

aligning regions in the mouse genome with the LECIF score **(Figs. 2.1b and 2.2a**). We weighted negative examples 50 times more than positive examples during training because we wanted the LECIF score to highlight regions with strong evidence of conservation at the functional genomics level. As a result, a small fraction of the aligning regions was highlighted with high LECIF score whereas most aligning regions would have scored high if the score was learned with positive and negative examples weighted equally (**Fig. 2.2b; Supplementary Fig. 2.1a**).

Comparative evaluation of LECIF's predictive performance

We evaluated LECIF at predicting whether pairs of regions that were held out from training align at the sequence level. LECIF had strong predictive power for this with an area under the receiver operating characteristic curve (AUROC) of 0.87 and an area under the precision-recall curve (AUPRC) of 0.23 compared to a random expectation of 0.50 and 0.02, respectively (**Fig. 2.2c,d**). Additionally, scores that were trained on non-overlapping sets of chromosomes had strong agreement with each other with a Pearson correlation coefficient (PCC) of 0.90 (**Methods**).

We compared LECIF to alternative methods that used random forest (RF), canonical correlation analysis (CCA), deep canonical correlation analysis (DCCA), or logistic regression (LR) instead of an ensemble of neural networks (**Fig. 2.2c,d**). When classifying held-out test examples, LECIF outperformed these methods with statistically significantly better AUROC and AUPRC values (RF AUROC: 0.82; CCA AUROC: 0.81; DCCA AUROC: 0.81; RF AUPRC: 0.13; CCA AUPRC: 0.06; DCCA AUPRC 0.07; LR AUROC: 0.50; AUPRC: 0.02; Wilcoxon signed-rank test $P<0.0001$). LR had no predictive power as expected, since it only considers features marginally and the positive and negative examples were defined such that each feature has an identical marginal distribution in positive and negative data.

We next evaluated LECIF design choices by comparing the LECIF score to predictions based on alternative choices. We first compared the LECIF score with a score computed at a single base resolution and confirmed they were strongly correlated (PCC: 0.99; **Methods**). We also compared the LECIF score to scores learned with different weightings of positive and negative examples and confirmed that relative ranking of predictions and predictive power for aligning regions were robust (**Supplementary Fig. 2.1**). We used LECIF with an ensemble of 100 neural networks and confirmed it led to better performance than using fewer networks, although fewer networks could be used to save computational cost with a small decrease in performance (**Supplementary Figs. 2.2-2.3**). We also compared the LECIF score to scores learned separately for the coding and non-coding genomes and observed that the scores were relatively well-correlated with the original LECIF score in the coding (PCC: 0.71) and non-coding (PCC: 0.95) genomes (**Supplementary Fig. 2.4**; **Methods**).

In addition, we evaluated the effect of the number of mouse features on LECIF's performance by learning two models with fewer mouse features (**Methods**). A score learned with 10% of the mouse features had strong agreement with the original LECIF score (PCC: 0.88; Spearman correlation coefficient (SCC): 0.80) and slightly weaker predictive performance (AUROC: 0.83 vs. 0.86; AUPRC: 0.16 vs. 0.21; **Supplementary Fig. 2.5**). However, a score learned with 1% of the mouse features had substantially weaker agreement with the original LECIF score (PCC: 0.66; SCC: 0.18) and weaker predictive performance for aligning pairs (AUROC: 0.66; AUPRC: 0.07).

Predictive power when including adjacent non-aligning mouse regions

The LECIF method can also score pairs of human and mouse regions that do not align at the sequence level. Previous comparative studies have reported movements of regulatory elements during evolution, where homologous regulatory activity of a human region is found in a region near the aligning region in another species instead of the aligning region[17,31,32]. We thus

investigated whether it is advantageous for LECIF to consider also the scores at non-aligning mouse regions proximal to the mouse region aligning to human. Specifically, for a given aligning pair of human and mouse regions we took the maximum LECIF score from pairs consisting of the human region and any mouse region located within a window centered around the aligning mouse region (**Methods**; **Supplementary Fig. 2.6**). We varied window sizes and repeated the same AUROC evaluations for predicting aligning regions as above (**Supplementary Fig. 2.7**).

We found that as we expanded the window size the predictive power decreased overall. We saw similar results when we repeated the evaluation with pairs stratified by the LECIF score at the aligning regions except for pairs with the lowest LECIF score (**Supplementary Fig. 2.8**). When we trained LECIF with an alternative set of negative examples selected from a genome background and repeated the evaluations (**Methods**), the expanded window still had decreased predictive power overall (**Supplementary Fig. 2.7**). These results suggested that applying LECIF to non-aligning regions would result in a substantial increase in false positive predictions, which indicates that sequence alignment provides strong prior information in detecting evidence for conservation at the functional genomics level. Moreover, non-aligning regions in general tend to be less conserved and exhibit different properties at the functional genomics level than aligning regions on which LECIF was trained[2], making LECIF relatively less applicable to such regions. We thus focused our initial application of LECIF to aligning regions. We note that because of the resolution at which the LECIF score is defined, even without explicitly expanding the window the score may still be capturing small movements of regulatory sites, which cannot be explicitly detected in the coarse-resolution functional genomics data currently available to LECIF.

Distribution of LECIF score in chromatin states

To characterize DNA elements highlighted by LECIF, we investigated the distribution of the LECIF score overlapping the chromatin state annotations that were provided to LECIF as

input features. When we computed the mean LECIF score for each chromatin state across epigenomes[4] (**Fig. 2.2e; Methods**), chromatin states associated with strong regulatory or transcriptional activity tended to have a higher mean LECIF score than other states, with the highest of 0.71 for an active transcription start site (TSS) state and the lowest of 0.07 and 0.08 for the heterochromatin and quiescent states, respectively. Candidate enhancer states outside of transcribed regions had an intermediate mean LECIF score ranging from 0.18 to 0.32, which was lower than the mean scores of promoter associated states, 0.53 to 0.71, and consistent with previous findings that enhancers tend to evolve faster than promoters[20]. We also observed similar trends with other input features and external gene annotations in both human and mouse (**Supplementary Figs. 2.9-2.11**).

LECIF highlights shared functional genomic activity

To validate that the LECIF score reflects expected cross-species similarity in functional genomic features, we investigated the LECIF score in relation to human and mouse genomic annotations jointly. We first matched a subset of human and mouse ChIP-seq experiments of H3K27ac by their tissue of origin for 14 tissue type groups (**Methods**). We then quantified the cross-species similarity of the peak calls for each pair of regions jointly across the 14 tissue type groups using a weighted Jaccard similarity coefficient (**Methods**). We saw that the LECIF score was positively correlated with the weighted Jaccard similarity coefficient (PCC: 0.45; **Fig. 2.3a**). This is despite LECIF not being given any information regarding tissue of origin of the experiments in the compendium of functional genomic annotations.

To provide further evidence that the LECIF score reflects expected cross-species similarity in functional genomic annotations, we examined the LECIF score in relation to the chromatin state annotations of pairs of human and mouse regions. We used the state annotations from a concatenated model of ChromHMM[29] where a shared set of states were learned for human and mouse[2]. For different ranges of the LECIF score, we correlated the

chromatin state frequency between human and mouse across regions in that score range (**Methods**). High-scoring pairs of regions tended to be annotated with similar sets of states in human and mouse epigenomes (**Fig. 2.3b,c, Supplementary Fig. 2.12**). Low-scoring pairs of regions were annotated with dissimilar sets of states in human and mouse and the quiescent state more frequently than high-scoring pairs (**Fig. 2.3b,d, Supplementary Figs. 2.12-2.14**).

We also investigated the LECIF score at topologically associated domain (TAD) boundaries that were previously identified in human and mouse cell types[33] as they represent an important regulatory genomic feature not provided to LECIF. Human regions overlapping a TAD boundary in any human cell type had a mean LECIF score of 0.17 compared to the genome-wide mean of 0.14 (Mann-Whitney U test $P<0.0001$). Pairs with human and mouse regions both overlapping a TAD boundary in a matched cell type had an even higher mean of 0.20, scoring significantly higher than pairs with either human or mouse region or neither regions overlapping a TAD boundary in the cell type (**Supplementary Fig. 2.15**; Mann-Whitney U test $P<0.0001$).

We also verified the advantage of integrating human and mouse data by generating a human-only baseline score. The score was learned using human functional genomics data with human regions that align to mouse as positive examples and the rest as negative examples (**Methods**). The human-only baseline score was weakly correlated with the human-mouse LECIF score with a PCC of 0.13 and did not reflect cross-species similarity in functional genomic features as strongly as the LECIF score (**Supplementary Figs. 2.12, 2.15, 2.16**). These results support the contribution of mouse data to identifying conserved functional genomic properties.

Relationship to sequence-based conservation annotations

We next analyzed the relationship between the LECIF score and various sequence-based annotations of conservation within aligning regions. We note that while human regions

that align to mouse at the sequence level do show some increase in sequence constraint relative to the entire genome, the majority of aligning regions do not show high levels of sequence constraint (**Supplementary Fig. 2.18**). We found that human regions overlapping sequence constrained elements had a greater average LECIF score, ranging from 0.19 to 0.22 across different element sets, than the mean among human regions that align to mouse in general (0.14) (**Fig. 2.4a**). When compared to five sequence constraint scores and additionally the percent identity between human and mouse, the LECIF score was moderately correlated with PCCs ranging from 0.18 to 0.25 for 50-bp windows with each score averaged across 50 bases (**Fig. 2.4c, Supplementary Fig. 2.19**; **Methods**). This moderate correlation may reflect biological difference between sequence conservation and functional genomics conservation[34], although potentially also the coarse resolution and incompleteness of functional genomics data.

To provide evidence that most high LECIF scores observed in regions with low sequence constraint scores are unlikely LECIF's false positives, we analyzed human and mouse chromatin state annotations in regions where the two scores strongly disagreed. Specifically, for pairs of regions where the LECIF score was high and the PhyloP score[12] was low in all bases within 500 bp of the human region, we computed the correlation of chromatin state frequencies as described above (**Fig. 2.4d, Supplementary Fig. 2.20**). We found that such pairs had strong cross-species similarity for all states, often as strong as pairs that scored high in both scores. In comparison, pairs of regions with low LECIF score and high PhyloP score had weaker cross-species similarity of frequency in all states. This suggests that the LECIF score can capture conservation at the functional genomics level even in regions that align, but have limited sequence constraint among aligning regions, potentially detecting signatures of conservation not captured by sequence constraint scores defined from multi-species sequence alignments.

To further understand the differences between the LECIF score and constraint scores, we next identified patterns within a multi-species sequence alignment that may correspond to

those differences. To do this, we leveraged the ConsHMM[35] conservation state annotation of the human genome, which annotates each human genomic base based on alignment and matching patterns with vertebrate genomes in a 100-way sequence alignment (**Supplementary Fig. 2.21**). Among a hundred conservation states, the state with the highest average LECIF score corresponded to human bases that align and match to many vertebrate genomes with a moderate probability, indicating signatures of conservation across many vertebrates. This state was previously shown to most strongly enrich for promoter and CpG islands out of all conservation states. In contrast, this state had only the 12th highest average PhyloP score. This suggests that the disagreement between the LECIF score and constraint scores could be partly explained by constraint scores not capturing signatures of conservation that are actually present in the multi-species sequence alignment, and further supports that the LECIF score can provide complementary information to sequence constraint scores about conservation.

Since the LECIF score prioritized the conservation state most enriched for CpG islands, which are known to have varying evolutionary dynamics at the sequence level, we analyzed the LECIF score of human CpG islands previously grouped by their distinct regimes during primate sequence evolution[36] (**Fig. 2.4b**). CpG islands in general scored high with a mean LECIF score of 0.53, and the score positively correlated with the likelihood of a CpG island being classified as slowly evolving as opposed to quickly evolving (**Supplementary Fig. 2.22;** PCC: 0.50). Slowly evolving CpG islands characterized by low rate of C-to-T deamination had higher LECIF scores with a mean of 0.65. In contrast, quickly evolving CpG islands had lower LECIF scores with a mean of 0.35. Although LECIF scores CpG islands higher than the rest of the genome in general, the score reflects the distinct evolutionary dynamics among them.

Relationship to phenotype-associated variation

To investigate if the LECIF score enriches for biologically important genomic loci linked to phenotype, we analyzed the relationship between the LECIF score and phenotype-associated

genetic variation (**Fig. 2.5a**). We observed that regulatory disease variants from Human Gene Mutation Database (HGMD)[37] enriched for regions with high LECIF score. In contrast, we saw small depletions for common variants[38] in those high-scoring regions. We saw that high-scoring regions also exhibited enrichment of Genome-wide Association Studies (GWAS) Catalog[39] variants and expression quantitative trait loci (eQTLs) from GTEx[40].

We also conducted a heritability partitioning analysis with the LECIF score for 12 complex traits[41]. Specifically, we applied heritability partitioning with an annotation of bases with a LECIF score in the top 5% in the context of a baseline set of annotations[42], which we extended to also include annotations of human regions that align to mouse and top 5% regions based on the human-only baseline score. We note that the baseline annotation set includes multiple sequence constraint annotations. We observed that the top 5% regions based on the LECIF score resulted in enrichments of heritability with statistical significance for several traits (**Fig. 2.5b**). Furthermore, we observed overall stronger enrichments for the LECIF annotation than the human-only baseline annotation and the annotation of human regions that align to mouse.

LECIF highlights regions in mouse QTL relevant to disease

To demonstrate how LECIF could be applied to translating biological findings, particularly in mapping trait-associated loci between mouse and human, we analyzed mouse insulin secretion quantitative trait loci (QTL) and human diabetes GWAS variants[43]. Previously, it was shown that human regions syntenic to the mouse insulin secretion QTL were enriched for the human diabetes GWAS variants. However, mouse QTL in general can span several megabases, making it difficult to identify likely causal variants within the loci for the trait of interest[9]. We thus mapped the mouse insulin secretion QTL to the human genome based on sequence alignment and asked whether the LECIF score could provide information in locasting

regions within the mapped mouse insulin secretion QTL that correspond to human diabetes GWAS variants.

We observed that human genomic windows within the mapped mouse insulin secretion QTL that overlap the human GWAS variants had a statistically higher distribution of mean LECIF scores than windows within the mouse QTL not overlapping the variants or windows overlapping the variants (Mann-Whitney U test $P<0.0001$; **Fig. 2.6a**, **Supplementary Fig. 2.23b,c**). Additionally, we saw that the human diabetes GWAS variants that lie within the mapped mouse QTL had a higher distribution of mean LECIF scores than human GWAS variants outside the mouse QTL in addition to human bases within the mouse QTL that are not the human GWAS variants (Mann-Whitney U test $P<0.0001$; **Supplementary Fig. 2.23a**). These results indicate LECIF's potential value in locating regions within mouse QTL that are more likely relevant to a given trait in human.

LECIF highlights conserved methylation patterns linked to phenotype

To further illustrate potential applications of LECIF, we also evaluated the ability of the LECIF score to prioritize epigenetic features conserved between human and mouse in a disease relative context. Specifically, we considered data from an epigenetic study on differential methylation in diabetic phenotypes in human and mouse[44], which was independent of the data provided to LECIF. The study identified conserved differentially methylated regions (DMRs) associated with obesity by first finding DMRs in high-fat-fed and low-fat-fed mice and then testing their homologous human regions for differential methylation between obese and lean patients. The LECIF score was significantly higher in conserved DMRs in comparison to mouse-specific DMRs (Mann-Whitney U test $P<0.01$; **Fig. 2.6b**). This supports the potential value of the LECIF score for prioritizing among all loci with epigenetic associations with phenotype in one species the specific loci whose associations are more likely to be shared in the other species.

Discussion

We presented LECIF, a method that scores evidence for conservation between human and mouse based on a compendium of functional genomic annotations from each species. To do so, LECIF trains neural networks to differentiate aligning pairs of regions from mismatched pairs of the same set of regions based on their functional genomic annotations without using sequence information as features. The functional genomic annotations include maps of open chromatin, transcription factor binding, gene expression signals, and chromatin state annotations. The resulting score captures evidence of conservation at the functional genomics level that is based on a diverse set of annotations and thus not specific to one class of DNA elements.

We applied LECIF with more than 10,000 functional genomic annotations from human and mouse to learn the human-mouse LECIF score. The LECIF score had greater predictive power than several baseline scores at discriminating pairs of human and mouse regions that align to each other from mismatched pairs of aligning regions. Using H3K27ac samples matched by their tissue of origin and separately using chromatin state annotations learned jointly between human and mouse, we showed that the LECIF score reflects the relationships between biologically similar human and mouse functional genomic annotations. LECIF was able to do this without any explicit information provided about the relationship between different features within or across species. Furthermore, LECIF was able to do so even in regions where sequence constraint was low, supporting that the LECIF score provides complementary information to sequence constraint annotations. Regions with high LECIF score were enriched for phenotype-associated variants from curated databases and also for heritability of complex traits. Using matched DNA methylation samples between human and mouse and separately using matched GWAS and QTL data sets, both in the context of a diabetes trait, we showed that

the LECIF score has preference for human and mouse regions with shared associations with the trait.

These results support the potential value of the LECIF score in various applications in the context of model organism research. For example, given a set of phenotypic-associated loci identified in a mouse model, which are increasingly available through efforts like the Mouse Phenome Database[45], the highest-scoring loci could be prioritized for experimental validation in human cells if possible. Conversely, given human genomic variants or candidate regulatory elements with known associations with a trait, those with the highest LECIF scores could be prioritized for testing in mouse models. In addition, when loci exhibit signals of interest in both species, those with the highest LECIF scores could be prioritized for follow-up experiments.

While we expect LECIF to be useful, we do note a few limitations. LECIF only scores evidence of conservation at the functional genomics level. There thus could be regions that are conserved at the functional genomics level, but have a low LECIF score, since the evidence was not present in the data currently available to LECIF. This makes it difficult to distinguish the case of human-specific regulatory activity from insufficient evidence in the aligning mouse region's annotations based on a low LECIF score. Fortunately, the interpretation of high LECIF scores is less ambiguous. We also note that the LECIF score's resolution is limited by the resolution of the input functional genomic annotations and thus does not have the base resolution that sequence-based conservation annotations can have. Additionally, LECIF is designed to aggregate information across multiple tissues and cell types and thus does not provide direct information about a particular tissue.

In addition, we note that currently the LECIF score is only available for pairs of regions that align to each other. While in principle LECIF can be applied to score any pairs of regions, more false positive predictions are expected as a result, compared to our presented strategy of restricting to regions that align at the sequence level. Although we explored an alternative strategy that considered non-aligning regions in a neighborhood of each pair of aligning regions,

this did not lead to improvements in our evaluations over considering only the aligning regions. However, future work could develop other strategies that lead to improvements.

While here we focused on human and mouse, as mouse is a widely used model organism for human and there is substantial data available for both, LECIF can be applied to compare human to any species with a genome-wide pairwise sequence alignment to human and functional genomics data. Applying LECIF to human and mouse with mouse features down-sampled demonstrated that a few hundred annotations from the non-human species may be sufficient to capture a large portion of conservation at the functional genomics level, although the quality of the score will depend on the coverage of the data available for the non-human species. As functional genomics data from a more diverse set of species, cell types, and assays continues to become available, the utility of LECIF will continue to grow for identifying regions conserved at the functional genomics level and transferring findings from mouse and other model organism research to human biology.

Methods

Pairwise sequence alignment

For the pairwise sequence alignment, we used the chained and netted alignment[10] between the human genome (hg19) and the mouse genome (mm10), with human as the reference genome for the alignment. Given multiple mouse genome segments that map to a single human genome segment, we chose the mouse segment with the highest alignment score. This alignment was obtained from the UCSC Genome Browser[3].

Functional genomics data used for input features

ChromHMM[29] chromatin state annotations for human were from the 25-state model learned for 127 cell and tissue types based on imputed data from the Roadmap Epigenomics Project[4] and for mouse from the 15-state model learned for 66 cell and tissue types from

ENCODE[46]. Peak calls for DNase-seq and ChIP-seq experiments of transcription factors, histone modifications, and histone variants were from Roadmap Epigenomics[4], ENCODE[1], and Mouse ENCODE[2]. Peak calls for Cap Analysis Gene Expression (CAGE) experiments were from FANTOM5[30]. RNA-seq signal data were from ENCODE[1] and Mouse ENCODE[2]. For ENCODE and Mouse ENCODE data, we used the uniformed processed version available from the ENCODE portal. Additional information including the specific source of each dataset used is listed in **Supplementary Data 2.1**.

Defining pairs of human and mouse regions for training and prediction

To define pairs of human and mouse regions for training and prediction for LECIF, we first identified alignment blocks from the pairwise alignment. We defined alignment blocks as pairs of human and mouse genomic segments without any alignment gap, meaning the human and mouse genomic segments both had a nucleotide present at each base in the block. We then for each alignment block defined non-overlapping windows of 50 bp starting from the first base in the alignment block. Each 50-bp window defined a region. If the alignment block ended within the 50-bp window, we truncated the window to the end of the block to define the region. This resulted in some regions being shorter than 50 bp. To define negative examples, we randomly paired up human and mouse regions included in the positive examples. With this procedure, all human regions included in the negative examples aligned somewhere else in the mouse genome, and all mouse regions in the negative examples aligned somewhere else in the human genome.

Defining subsets of pairs of regions for training and evaluation

All human and mouse chromosomes, except for Y and mitochondrial chromosomes, were used. X chromosomes were excluded from training, validation, and test, but included for prediction and downstream analyses. To generate predictions for all pairs of human and mouse regions that included a human region from an even chromosome or X chromosome, we trained LECIF on

pairs of human and mouse regions such that both the human and mouse regions came from a subset of odd chromosomes for its respective species (**Supplementary Data 2.2**). To form a validation set, which we used for hyper-parameter tuning and early stopping during training, we used pairs of regions such that the human region came from a subset of odd chromosomes not used in training and likewise for mouse. To form a test set, which we used to generate the receiver operating characteristic (ROC) and precision-recall (PR) curves, we used all pairs of regions such that both the human and mouse region were from an even chromosome. To generate predictions for all pairs that included a human region from an odd chromosome, we took an analogous approach as above (**Supplementary Data 2.2**). There was no overlap in genomic regions used for training, validation, and test. To assess the agreement between a model trained on odd chromosomes and a model trained on even chromosomes, we used pairs of regions that were from a subset of chromosomes not used in training or validation of either model (**Supplementary Data 2.2**).

Feature representations

For each pair of human and mouse regions, we generated two feature vectors. The two vectors were based on annotations overlapping the first base of the human and mouse regions, respectively, which were at most 50 bp. For computational considerations, we only used the first base of each region to provide the LECIF score for all bases in the region. To evaluate the effect of this, we computed the Pearson correlation coefficient (PCC) between a score defined at base resolution for 1 million randomly sampled pairs of human and mouse bases that align to each other and the LECIF score, which was defined at every 50 bp within each alignment block, for the same set of 1 million pairs.

Each peak call corresponded to one binary feature. If a base overlapped a peak call for an experiment, the corresponding value in the feature vector was encoded as a 1, otherwise it was encoded as a 0. While real-valued signals are also available for these experiments with peak

calls, we used the binary peak calls for improved scalability and reduced potential for overfitting. Chromatin state annotations were one hot-encoded such that there was a separate binary feature representing the presence of each chromatin state in each cell or tissue type. Each RNA-seq experiment corresponded to one continuous feature. For human RNA-seq experiments, to also have the features in the range 0 to 1, we first computed the maximum and minimum signal value at any base in any of the human RNA-seq experiments. We then normalized values by subtracting the minimum signal value and dividing by the difference between the maximum and minimum signal values. We separately did the same normalization for mouse RNA-seq experiments. In total, we used 8,824 human features and 3,113 mouse features. Number of features from each data type are reported in **Supplementary Data 2.1**.

LECIF Classifier

The classifier that LECIF uses is an ensemble of neural networks where each neural network had a pseudo-Siamese architecture[47] (**Supplementary Fig. 2.24**). A Siamese neural network consists of two identical sub-networks followed by a final sub-network that combines the output from the two sub-networks to generate a final prediction[48]. A pseudo-Siamese network is similar except it uses two distinct sub-networks instead of identical sub-networks. In LECIF, the two sub-networks corresponded to human and mouse. Human and mouse feature vectors were given to the human and mouse sub-networks, respectively, as input. We also evaluated using a fully-connected neural network, but found that it led to highly similar predictions (PCC: 0.95) while taking longer to train.

Hyper-parameters of a neural network consisted of number of layers in each sub-network and the final sub-network, number of neurons in each layer, batch size, learning rate, and dropout rate. To set the values of the hyper-parameters, we conducted a random search, where we generated 100 neural networks, each with different randomly selected combinations of hyper-parameters (**Supplementary Data 2.3**). Each neural network was trained on the same set of

randomly selected 1 million positive and 1 million negative training examples. We applied 50 times more weight to our negative examples than positive examples during training so that a high LECIF score corresponds to strong evidence of conservation. We identified the best-performing combination of hyper-parameters based on maximizing the AUROC on the validation examples.

With the best-performing combination of hyper-parameters, we then trained a new set of 100 neural networks each provided with different subsets of 1 million positive and 1 million negative training examples randomly selected from a pool of all training examples (>2.2 million positive and >2.2 million negative). While the same genomic regions in each species appear in both positive and negative examples given all available training examples, a single neural network may not necessarily encounter the same set of regions in its positive and negative examples due to random sampling. We applied the same increased weighting of negative examples as above. The final prediction of the ensemble was the average of the predictions from the 100 trained neural networks.

For both hyper-parameter search and training, we stopped training if there were no improvements in AUROC evaluated on the validation examples over three epochs. We saved the classifier from the epoch with the highest AUROC on the validation examples. The maximum number of epochs we allowed during training was 100 and the maximum training time we allowed was 24 hours.

We also generated a version of the LECIF classifier, LECIF-GB, which was trained in the same way as LECIF except the negative examples were pairs of human and mouse regions that were both randomly selected from anywhere in their respective genomes as opposed to being constrained to aligning regions.

We used PyTorch (version 0.3.0.post4)[49] for implementation of the neural networks.

Random forest baseline

We trained, applied, and evaluated random forest using the same procedure as explained above, except we used a decision tree in place of a neural network. We also did hyper-parameter selection as explained above, but for a set of hyper-parameters unique to decision trees (**Supplementary Data 2.3**)**.** We used Scikit-learn (version 0.19.1)[50] for implementation.

Canonical correlation analysis baseline

We trained an ensemble of canonical correlation analysis (CCA) mappings using the same procedure as above, except using a CCA mapping in place of a neural network and positive examples only. We applied and evaluated the ensemble using the same procedure as explained above. We also did hyper-parameter selection as explained above, but for a set of hyper-parameters unique to CCA mapping (**Supplementary Data 2.3**) and through a grid search instead of random search. We used Pyrcca[51] for implementation. Similarly, we also trained an ensemble of deep canonical correlation analysis (DCCA) mappings[52]. We did hyper-parameter selection as done for CCA, but for a set of hyper-parameters unique to DCCA mapping and through a random search (**Supplementary Data 2.3**). We used a MATLAB implementation of DCCA from prior work[53].

Logistic regression baseline

We trained, applied, and evaluated an ensemble of logistic regression classifiers using the same procedure as above, except we used a logistic regression classifier in place of a neural network. We also did hyper-parameter selection as for the neural networks, but for a set of hyper-parameters unique to logistic regression models (**Supplementary Data 2.3**) and through a grid search instead of random search. We used Scikit-learn (version 0.19.1)[50] for implementation.

LECIF scores separately learned for coding and non-coding bases

We trained, applied, and evaluated two models using separate training data from coding and non-coding bases. Training examples used to learn the original LECIF score were grouped into coding and non-coding examples based on whether each example's human region overlapped GENCODE annotation of coding sequence (CDS). Given non-coding training examples, the same learning procedure used to learn the original LECIF score was used to learn a score from non-coding regions. For coding training examples, all available training and validation examples (~40,000 training and ~20,000 validation examples) were used for hyperparameter search. Given optimized parameters, each classifier was trained on 10,000 positive and 10,000 negative training examples, instead of 1 million for each. These adjustments were made specifically for training a model on coding regions because there were much fewer regions to use.

LECIF scores with fewer mouse features

We trained, applied, and evaluated two models that used the same human features as LECIF, but used fewer mouse features. One of the models used 10% of the original set of mouse features and the other used 1%. Except for down-sampling features, model training and hyper-parameter search were done the same way as LECIF with the full set of features. To select mouse features for the 10% model, we first randomly selected 6 out of 66 epigenomes in the 15-state mouse ChromHMM chromatin state annotations, resulting in 90 one-hot encoded features corresponding to chromatin states. We then randomly sampled 221 features from features corresponding to mouse DNase-seq, ChIP-seq, RNA-seq, and CAGE annotations, resulting in 331 mouse features in total. For the 1% model, we randomly sampled 31 features from those corresponding to mouse DNase-seq, ChIP-seq, RNA-seq, and CAGE annotations. We did not use any features corresponding to chromatin state annotations in the 1% model. This allowed us to simulate LECIF's application to a non-human species with limited functional genomic data where chromatin state annotations are not available. As in training, only the selected mouse

features along with the full set of human features were used for prediction and evaluation for these scores based on fewer mouse features.

Human-only baseline

We trained, applied, and evaluated a human-only baseline, which used the same human features as LECIF, but did not use any mouse features and used a different set of positive and negative examples for training. The positive examples were human regions that align to the mouse genome and the negative examples were human regions that do not align to the mouse genome. We otherwise used the same procedure for training, prediction, and evaluation as for LECIF except we used an ensemble of fully-connected neural networks. We also did hyper-parameter selection as for LECIF, but for a set of hyper-parameters of a fully-connected neural network (**Supplementary Data 2.3**). We used PyTorch (version 0.3.0.post4)[49] for implementation.

Area under the ROC and PR curves

To compute each classifier's classification performance based on area under the receiver operating characteristic (ROC) curve and precision-recall (PR) curve, we used Scikit-learn's implementation[50].

Defining LECIF score including adjacent non-aligning mouse regions

To generate a LECIF score for each pair of a human region and its aligning mouse region with adjacent non-aligning mouse regions also considered, we computed LECIF scores for additional pairs that consisted of the same human region and distinct 50-bp mouse regions located within a neighborhood of W bases centered around the aligning mouse region **(Supplementary Fig. 2.3)**. The non-aligning mouse regions were defined by sliding a 50-bp window from the first base of the aligning mouse region in both the 5' and 3' directions. We then took the maximum over these LECIF scores to produce a score which we refer to as the region-

neighborhood LECIF score. We varied *W* between 0 and 20kb. We note that *W* of 0 corresponds to the original LECIF score.

Computing mean LECIF score for chromatin states

To compute the mean LECIF score for each chromatin state in the 25-state ChromHMM annotation across 127 human epigenomes[4,29], for every pair of chromatin state and epigenome, we first averaged the LECIF score in all aligning regions annotated by the state in the epigenome. We then for each chromatin state computed the average of 127 mean scores, each coming from an epigenome.

H3K27ac activity similarity

To define the H3K27ac activity similarity between human and mouse based on known biology, we took all human and mouse H3K27ac experiments used for features and manually grouped them into the following 14 tissue type groups based on available annotations of the experiments: adipose, bone element, brain, embryo, heart, intestine, kidney, limb, liver, lung, lymph node, spleen, stomach, thymus. **Supplementary Data 2.1** specifies which experiment was assigned to which group, but we note that information about these groups were not used in learning the LECIF score. The 14 groups listed above were represented in at least one H3K27ac experiment in both species. For the analysis, we discarded experiments that did not belong to any of the tissue groups.

For each pair of human and mouse regions, we then defined vectors *h* and *m* of length 14 where h_i and m_i correspond to the fraction of experiments in the *i*-th group with peak calls that overlapped the human and mouse regions, respectively. Finally, for each pair of human and mouse regions, we computed the weighted Jaccard similarity coefficient[54] between these two vectors. The weighted Jaccard similarity coefficient is defined as:

$$J(h,m) = \frac{\sum_i \min(h_i, m_i)}{\sum_i \max(h_i, m_i)} \qquad (2.1)$$

Any pair with an undefined similarity coefficient due to the denominator summing up to zero was removed from the analysis.

Chromatin state frequency correlation

To analyze cross-species agreement of chromatin state frequencies as a function of the LECIF score, we first grouped pairs of human and mouse regions based on their LECIF score. When binning based on either score, five or ten equal-width bins were used with varying numbers of pairs in each bin. We repeated the procedure when using the human-only baseline score in place of the LECIF score. We also binned based on the percentile rank of scores, where either five or ten bins were used with nearly the same number of pairs in each bin.

To compute the chromatin state frequency correlation across a set of pairs of human and mouse regions defined as described above, we used a chromatin state model jointly learned from both human and mouse genome[2]. For each of the seven chromatin states, we defined vectors for human and mouse. An element of a vector for human corresponds to the fraction of epigenomes in which one of the human regions is annotated with the state, and similarly for the mouse vector and regions. We then computed the PCC between the two vectors for each chromatin state, resulting in seven PCC values.

Correlation between the LECIF score and sequence constraint scores

To compute the correlation between the LECIF score and sequence constraint scores, we slid a 50-bp genomic window in 10-bp increment across the human genome. For each window, we computed the mean of each score (LECIF or sequence constraint). For each sequence constraint score, we computed the PCC and SCC between the LECIF score and the sequence constraint score for windows with at least *n* bases annotated by the two scores, with *n* ranging

from 1 to 50. The two scores were not required to be defined on the same set of bases within the 50-bp window.

Heritability partitioning analysis

To perform the heritability partitioning analysis, we used the LD-score regression software ldsc (v1.0.0)[41]. We generated an annotation of all human regions that align to the mouse genome and have a LECIF score above the 95th percentile. We used this annotation in the context of the baseline annotation set (v2.1) from Gazal et al.[42] along with another annotation generated based on the human-only baseline score instead of the LECIF score as well as an annotation of human regions that align to the mouse genome. We also included 500-bp windows around each annotation to dampen the inflation of heritability in neighboring regions due to linkage disequilibrium, following the procedure in Ref. [41].

We applied ldsc to this extended set of 60 annotations for the following 12 traits[41]: age at menarche, body mass index (BMI), coronary artery disease, education attainment, HDL cholesterol level, height, LDL cholesterol level, rheumatoid arthritis, schizophrenia, smoking, triglyceride level, and type 2 diabetes.

Data availability

The human-mouse LECIF score is available at https://github.com/ernstlab/LECIF. Links to data files used to generate input features to LECIF are listed in **Supplementary Data 2.1**. The human-mouse pairwise alignment is available at http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsMm10/axtNet/. For TSS, gene body, intron, exon, coding exon, 5' UTR, and 3' UTR annotations, we used GENCODE annotations V31lift37 for human and VM23 for mouse. We downloaded these annotations along with classification of evolutionary dynamics of CpG islands[36] and common SNPs (dbSNP v7)[38] from the UCSC Table Browser[3]. The HGMD variants that we used were variants annotated as 'regulatory mutations' in

the April 2012 public release of HGMD database[37,55]. The following URLs contain data sets that were used in the heritability partitioning analysis: Baseline annotation set[42]: https://storage.googleapis.com/broad-alkesgroup-public/LDSCORE/1000G_Phase3_baselineLD_v2.1_ldscores.tgz; Age at menarche[56]: https://www.reprogen.org; Body mass index, height[57]: http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; Coronary artery disease[58]: http://www.cardiogramplusc4d.org/data-downloads; Education attainment[59]: https://www.thessgac.org/data; HDL cholesterol level, LDL cholesterol level, triglyceride level[60]: http://csg.sph.umich.edu/willer/public/lipids2010; Rheumatoid arthritis[61]: http://plaza.umin.ac.jp/yokada/datasource/software.htm; Schizophrenia[62], smoking[63]: www.med.unc.edu/pgc/downloads; Type 2 diabetes[64]: http://www.diagram-consortium.org/downloads.html.

Code availability

The LECIF software is available at https://github.com/ernstlab/LECIF.

Figures

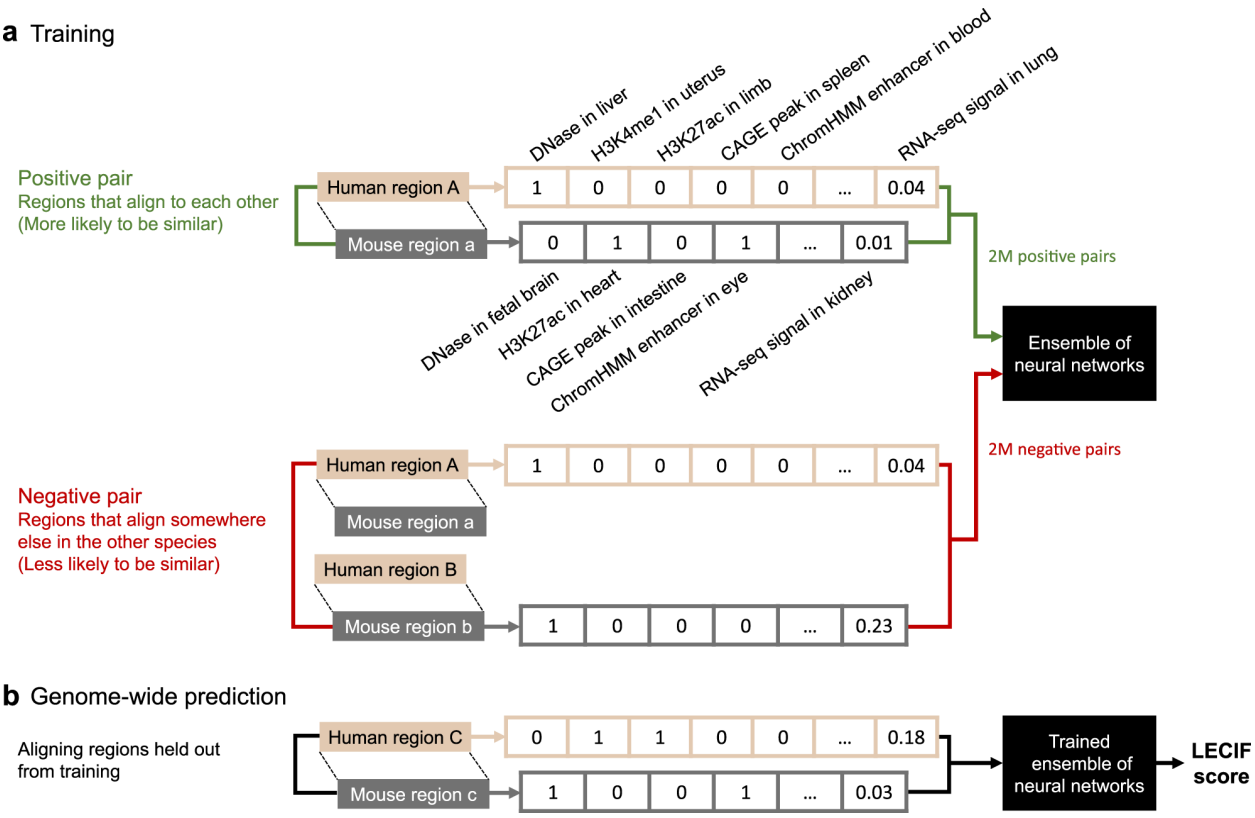
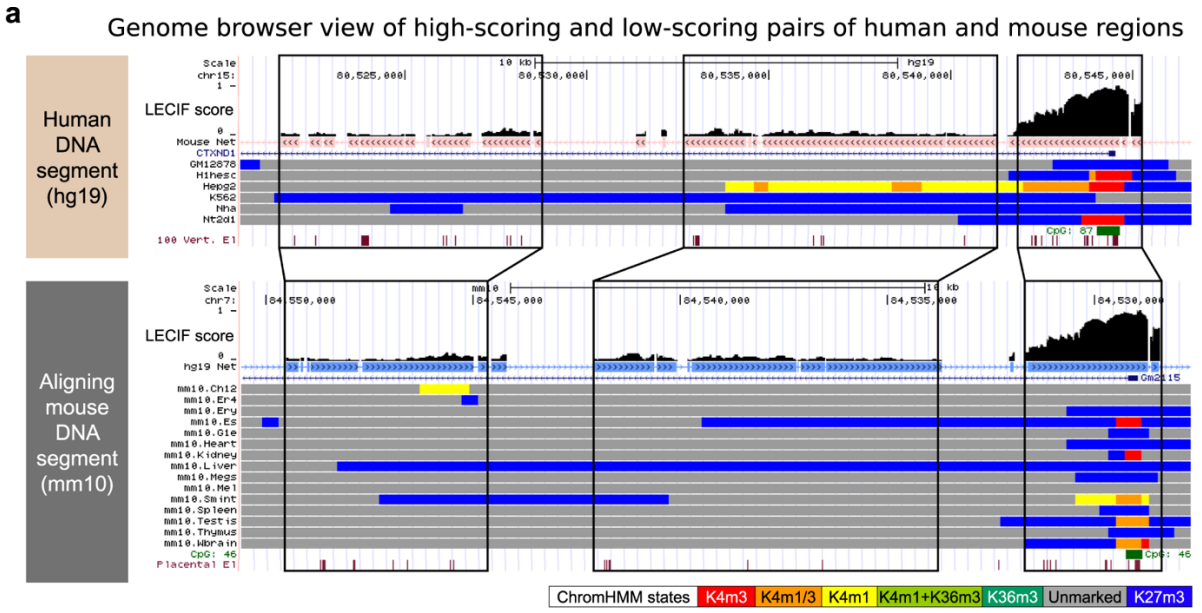
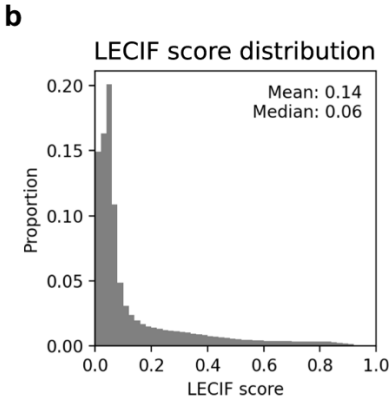
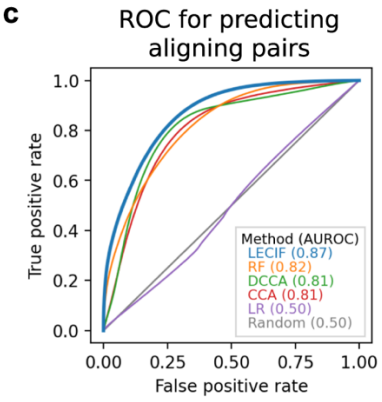
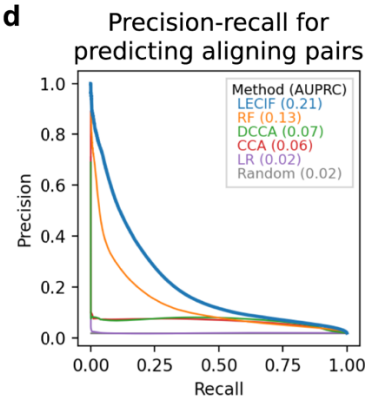


Figure 2.1. Overview of the LECIF method.
a. Supervised learning procedure of LECIF. For every pair of human and mouse genomic regions, two feature vectors are generated from their functional genomic annotations, one vector for the human region (beige) and the other vector for the mouse region (gray). Each feature vector consists of thousands of functional genomic annotations, as listed in **Supplementary Data 2.1**. Only a subset of the features are shown here. These two species-specific feature vectors are given to an ensemble of neural networks (ENN). The ENN is trained to distinguish positive pairs (green), which are aligning human and mouse regions, from negative pairs (red), which are randomly mismatched human and mouse regions that do not align to each other, but somewhere else in the other species. Here we provide about two million positive and two million negative training examples. Feature labels (e.g. DNase in liver) and matching of features across species are not provided to LECIF.
b. Genome-wide prediction procedure of LECIF. Once trained as illustrated in **a**, the ENN can estimate the probability of any given pair of human and mouse regions being classified as a positive pair. We consider this probability, the LECIF score, to represent the evidence of conservation observed in the functional genomics data annotating the given pair. Here we generate the LECIF score for all pairs of aligning human and mouse regions. Although not shown here, for model evaluation we also generate predictions for randomly mismatched negative pairs held out from training. When generating a prediction for a pair, LECIF uses an ENN trained on data excluding the pair as described in **Methods** and **Supplementary Data 2.2**.


a
Genome browser view of high-scoring and low-scoring pairs of human and mouse regions
Human DNA segment (hg19)
Aligning mouse DNA segment (mm10)
LECIF score
ChromHMM states
K4m3
K4m1/3
K4m1
K4m1+K36m3
K36m3
Unmarked
K27m3


b
LECIF score distribution
Mean: 0.14
Median: 0.06
Proportion
LECIF score


c
ROC for predicting aligning pairs
True positive rate
False positive rate
Method (AUROC)
LECIF (0.87)
RF (0.82)
DCCA (0.81)
CCA (0.81)
LR (0.50)
Random (0.50)


d
Precision-recall for predicting aligning pairs
Precision
Recall
Method (AUPRC)
LECIF (0.21)
RF (0.13)
DCCA (0.07)
CCA (0.06)
LR (0.02)
Random (0.02)

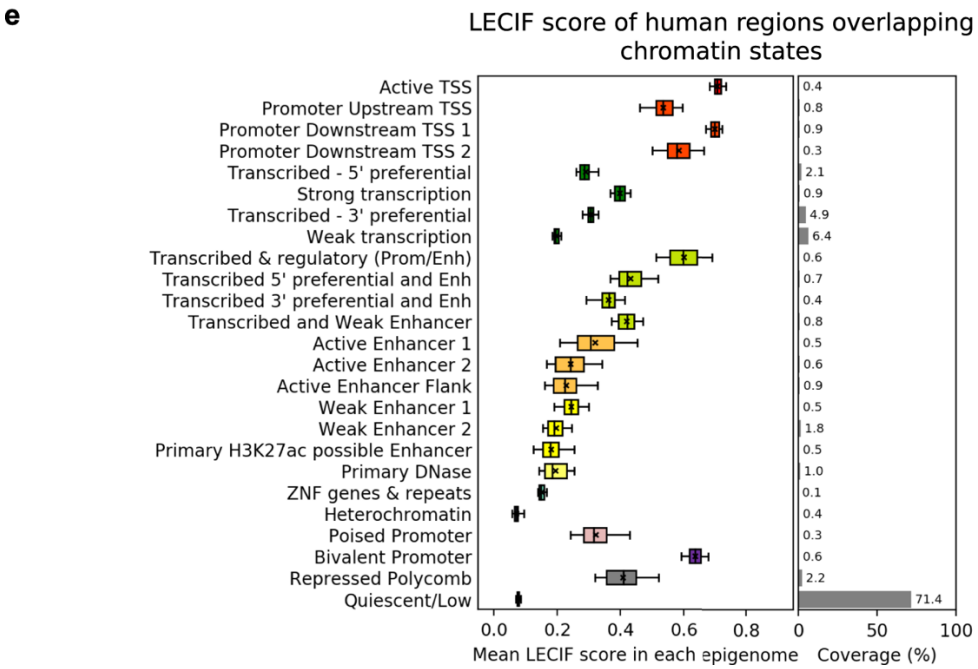

e
LECIF score of human regions overlapping chromatin states
Active TSS
Promoter Upstream TSS
Promoter Downstream TSS 1
Promoter Downstream TSS 2
Transcribed - 5' preferential
Strong transcription
Transcribed - 3' preferential
Weak transcription
Transcribed & regulatory (Prom/Enh)
Transcribed 5' preferential and Enh
Transcribed 3' preferential and Enh
Transcribed and Weak Enhancer
Active Enhancer 1
Active Enhancer 2
Active Enhancer Flank
Weak Enhancer 1
Weak Enhancer 2
Primary H3K27ac possible Enhancer
Primary DNase
ZNF genes & repeats
Heterochromatin
Poised Promoter
Bivalent Promoter
Repressed Polycomb
Quiescent/Low
Mean LECIF score in each epigenome
Coverage (%)

Figure 2.2. Characteristics of the human-mouse LECIF score.
a. Genome Browser[3] views with the LECIF score annotating human gene *CTXND1* (top) and its mouse ortholog *Gm2115* (bottom). In each view, LECIF score is shown in the top, followed by net alignment annotation[10] marking regions that align with colored boxes. Below the net annotation are RefSeq gene annotation[65] and ChromHMM chromatin state annotations[29] for different epigenomes from a model learned jointly for human and mouse[2]. State legend is in the bottom right. Below the state annotations are CpG island and PhastCons element[11] annotations. Black lines highlight segments that largely align. The mouse genome browser view is shown in the reverse direction (3'-5').
b. Distribution of the LECIF score. Fifty equal-width bins were used.
c. Receiver operating characteristic (ROC) curve comparing LECIF, random forest (RF), canonical correlation analysis (CCA), deep CCA (DCCA), and logistic regression (LR) for classifying pairs of regions that align at the sequence level, evaluated on a common set of held-out test data. Legend indicates color and mean area under the ROC curve (AUROC) for each method. The curve of each method was obtained by classifying 100,000 positive and 100,000 negative examples sampled with replacement from all test examples 100 times. Negative examples were weighted 50 times more than positive examples. For each method, standard deviation of the 100 AUROC values was under 0.005.
d. Similar to **c** except showing precision-recall (PR) instead of ROC. Standard deviation of the 100 area under the PR curve (AUPRC) values was under 0.005 for all methods.
e. Left panel shows for each human chromatin state as described previously[4,66] the distribution of mean LECIF score over different epigenomes (n=127). Mean LECIF score for a state in an epigenome is computed by averaging the score across regions overlapping the state in the epigenome. Each distribution is represented by a boxplot with median (black vertical line), mean (black 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). Right panel shows mean coverage of each state across human regions that align to mouse. Source data are provided as a Source Data file. A mouse version of this plot is in **Supplementary Fig. 2.10**.

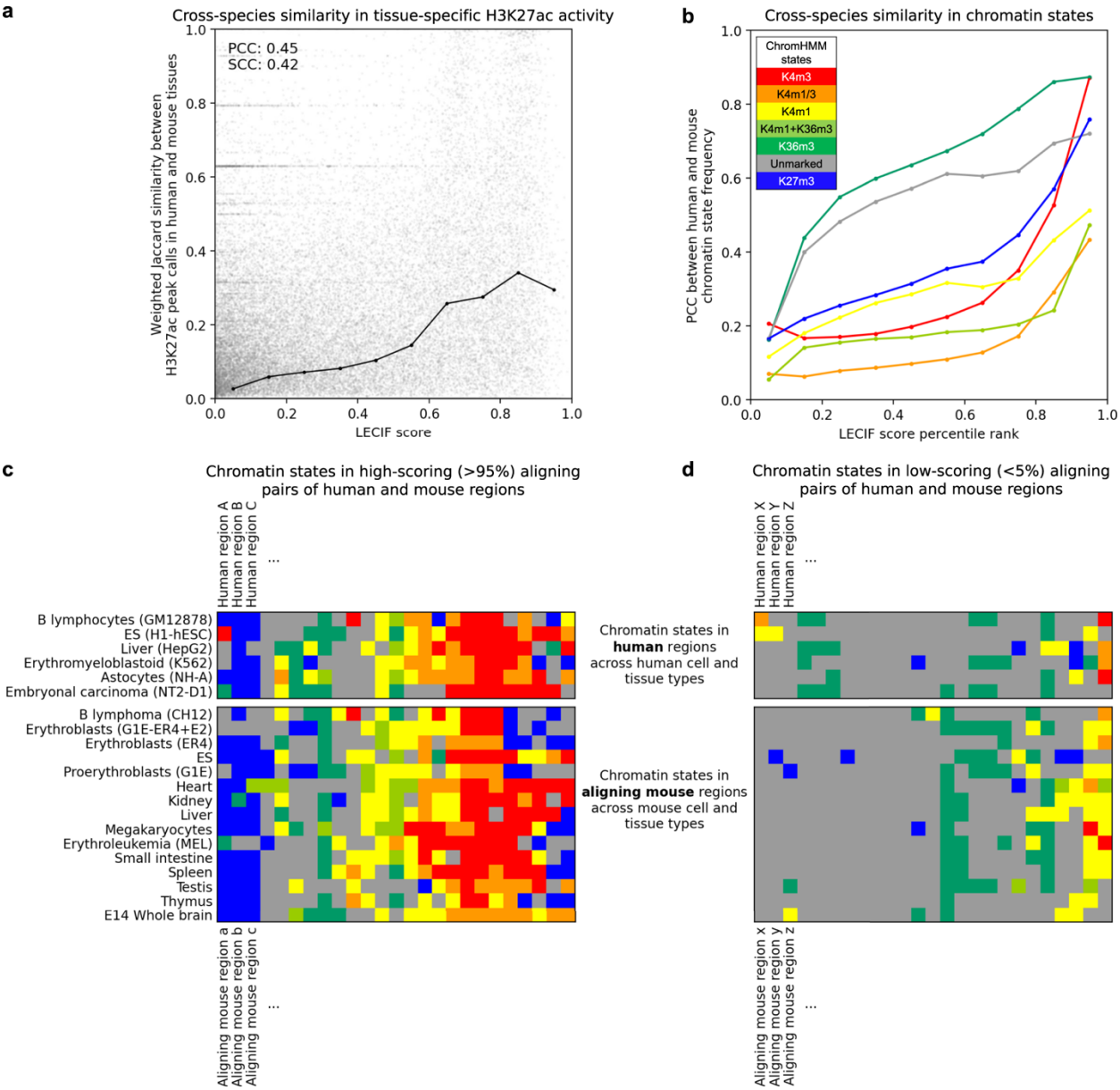


Figure 2.3. Correspondence of LECIF score to matched human and mouse annotations.
a. Scatter plot showing with a gray dot for each aligning pair of human and mouse regions the LECIF score (x-axis) and cross-species similarity of H3K27ac activity (y-axis). H3K27ac activity for a region in a tissue type is quantified as the fraction of experiments in the tissue type with peak calls overlapping the region. Its cross-species similarity is quantified as the weighted Jaccard similarity coefficient over 14 matched tissue types (**Methods**). One hundred thousand random pairs are shown. PCC and SCC, computed from all regions, are shown in the top left. Black circles show the mean coefficient of pairs binned by the LECIF score using ten equal-width bins. The circles are connected by piecewise linear interpolation. Source data are provided as a Source Data file. A version of this figure for the human-only baseline score is in **Supplementary Fig. 2.16**.
b. Cross-species agreement in chromatin state[2,29] frequency in aligning human and mouse regions for a ChromHMM model learned jointly for both species. Pairs were binned by LECIF score percentile rank using ten bins with similar number of pairs. For each state and percentile

rank bin, we computed PCC between the human and mouse state frequencies across all pairs in the bin (**Methods**). The values are shown with circles colored according to the top left legend from Ref. [2], which are connected by piecewise linear interpolation. Source data are provided as a Source Data file. Alternative versions of this plot with different binning schemes are in **Supplementary Fig. 2.12**.

c. ChromHMM chromatin state[2,29] annotations in high-scoring pairs of aligning human and mouse regions. Each row in top and bottom sub-panels corresponds to human and mouse epigenomes, respectively. Each column is a random pair of regions with high LECIF score (>95th percentile). Each cell shows the color of the state with which the region (column) is annotated in an epigenome (row) based on the same model as in **b**. Pairs (columns) were ordered based on hierarchical clustering applied to state annotations using Ward's linkage with optimal leaf ordering[67]. A version of this figure using mismatched non-aligning pairs is in **Supplementary Fig. 2.17**.

d. Same as **c,** but with pairs with low LECIF score (<5th percentile).

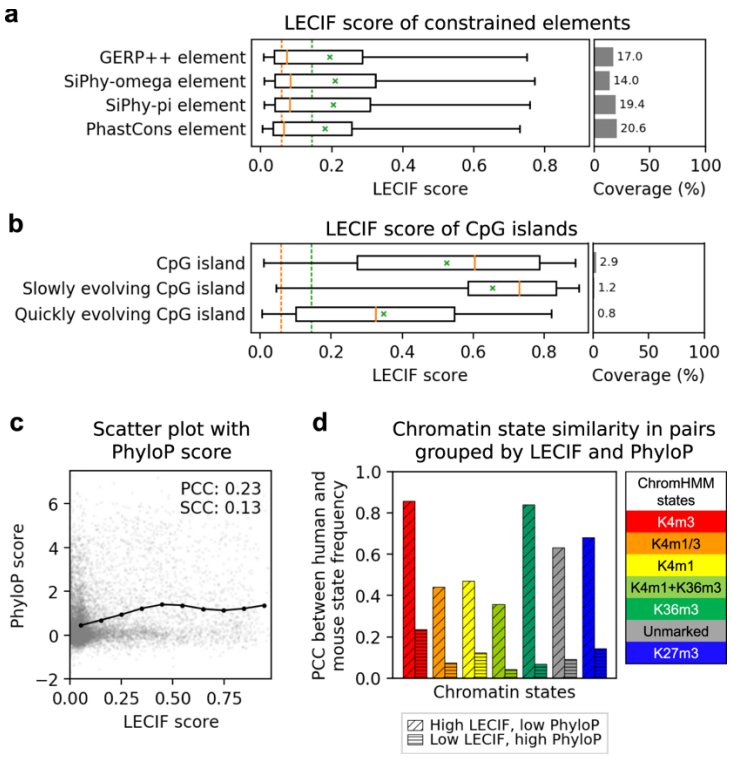


Figure 2.4. Relationship of LECIF score to sequence constraint annotations
a. Distribution of LECIF score in human regions overlapping constrained elements called by GERP++, SiPhy-omega, SiPhy-pi, and PhastCons (n=5,500,681, 4,515,990, 6,277,929, and 6,634,667 human regions, respectively)[11,13,14,68]. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). Orange and green dashed vertical lines denote genome-wide median and mean, respectively. Right sub-panel shows coverage of each annotation across all human regions aligning to mouse.
b. Similar to **a**, except showing LECIF score of human regions overlapping CpG islands (n=950,523) as well as subsets of regions overlapping slowly and quickly evolving CpG islands (n=399,280 and 260,132, respectively) as defined based on primates[36].
c. Scatter plot showing with gray dots the LECIF score and PhyloP score based on a 100-way vertebrate alignment[12]. The plot displays 100,000 random human regions that align to mouse with all bases annotated by both scores. PCC and SCC, computed from all applicable regions, are shown in the top right. Mean PhyloP score of all applicable regions binned by the LECIF score with ten equal-width bins are shown in black circles, connected by piecewise linear interpolation.
d. Cross-species agreement in chromatin state[2,29] frequency in pairs where the LECIF score is high and PhyloP score is low or vice versa. The PhyloP score is the same as in **c**. The states are the same as in **Fig. 2.3b-d**. Diagonally hatched bars show PCC from pairs with high LECIF score (>90th percentile) and low PhyloP score (<10th) in all bases within 500 bp of the human region. Horizontally hatched bars show PCC from pairs with low LECIF score (<10th) and high mean PhyloP score (>90th) in the human region. Bars are colored according to the legend on the right. Similar plots with different percentile cutoffs and also including pairs with both scores above or below the cutoffs are in **Supplementary Fig. 2.20**.
Source data for **a-d** are provided as Source Data files.

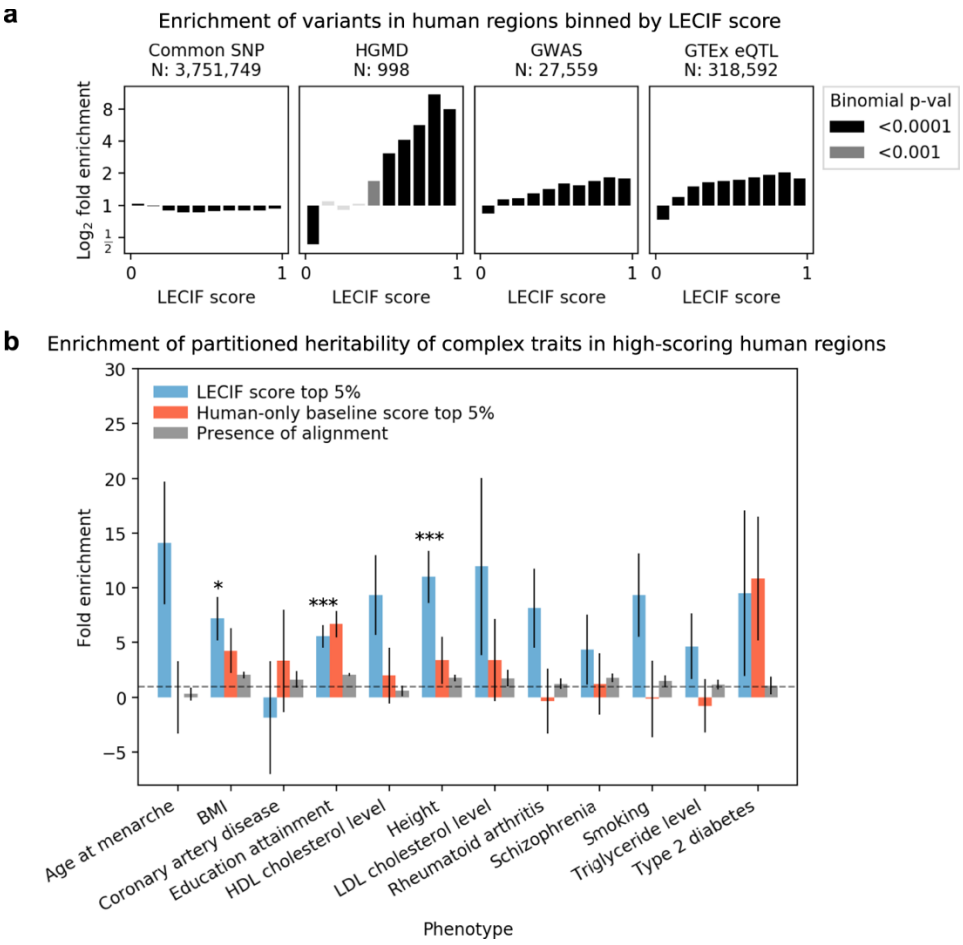


Figure 2.5. Relationship of LECIF score to genetic variants and heritability.
a. Shown from left to right are plots of $\log_2$ fold enrichment for variants based on four different sets, (i) common SNPs[38], (ii) HGMD regulatory variants[37], (iii) GWAS catalog SNPs[39], and (iv) GTEx cis-eQTLs[40] across tissues, within human regions binned by the LECIF score with ten equal-width bins. Analysis was restricted to human regions that align to mouse, and a uniform background within these regions was used. Displayed above each subplot is the number of regions overlapping the variants from the corresponding set included in the analysis. Black and dark gray bars denote $\log_2$ fold enrichments that resulted in *P* values below 0.0001 and 0.001, respectively, based on one-sided bionomial tests.
b. Fold enrichments for partitioned heritability of 12 phenotypes[41] in human regions with high LECIF score. Enrichments are shown for human regions with high human-mouse LECIF score (>95th percentile) (blue) and additionally for comparison regions with high human-only baseline score (>95th percentile) (orange) and human regions that align to mouse (gray). Heritability partitioning[41] for the LECIF score was applied in the context of a baseline set of annotations[42], which included sequence constraint annotations and was extended to include additional annotations generated based on the human-only baseline score and sequence alignment (**Methods**). Error bars denote standard error around the enrichment estimates. Horizontal dashed line denote no enrichment (fold enrichment of 1). * and *** denote Bonferroni-corrected one-sided *P* values for the LECIF score annotation's enrichment below 0.05 and 0.001, respectively. *P* values and standard errors were calculated using a block jackknife over SNPs with 200 equally sized blocks of adjacent SNPs as described in Ref. [41].
Source data for **a** and **b** are provided as Source Data files.

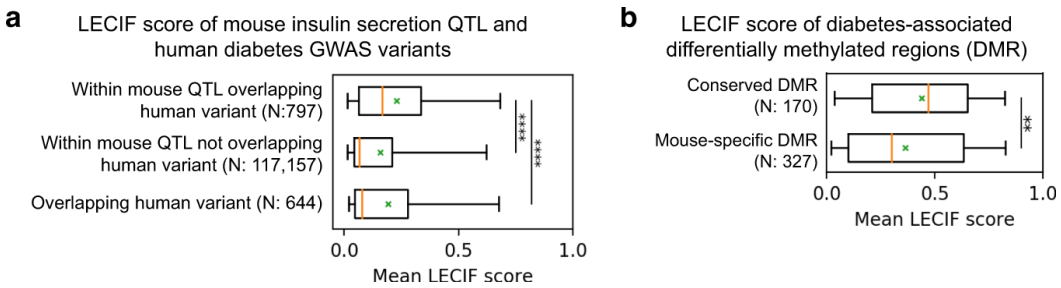
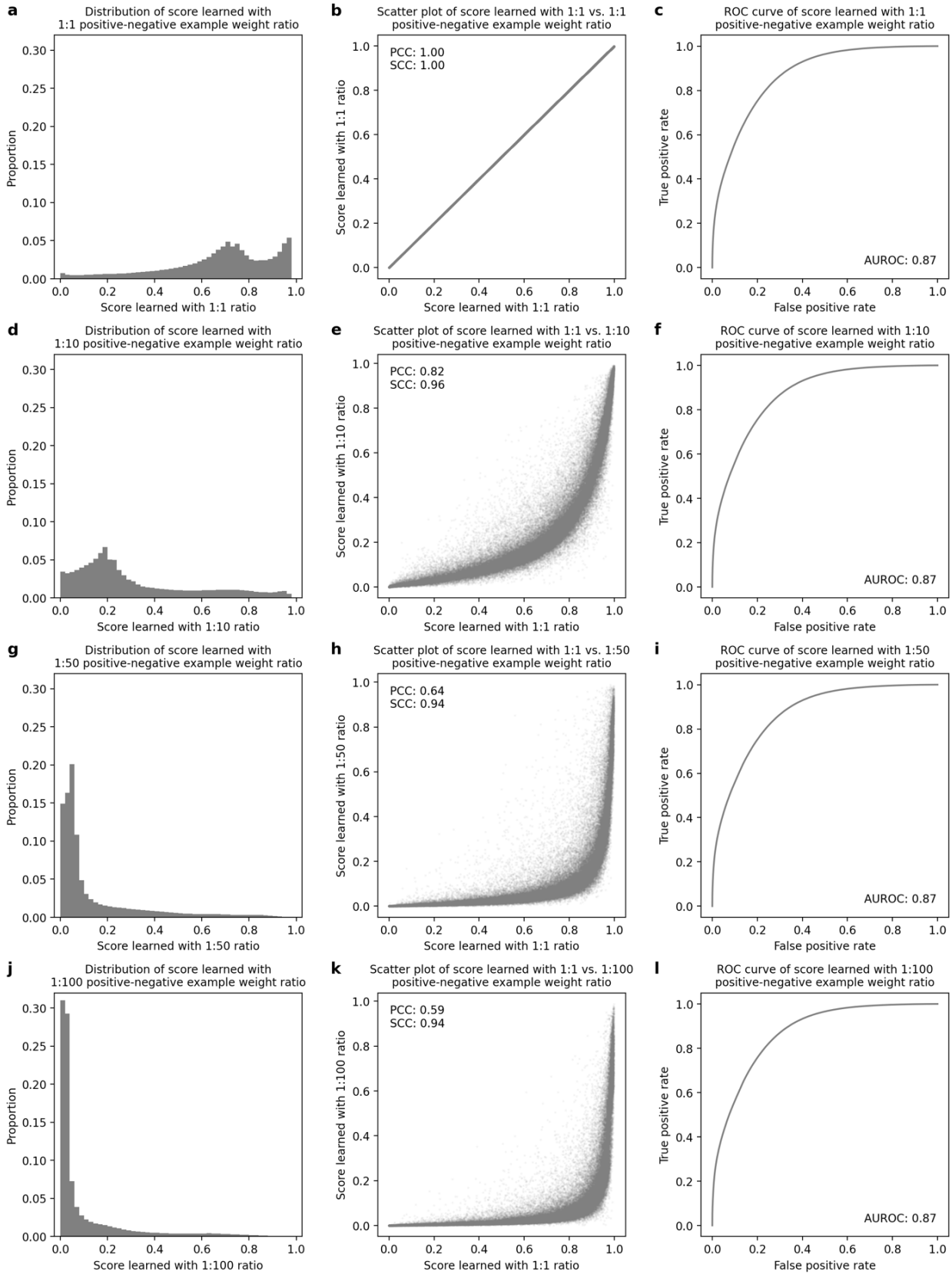


Figure 2.6. Relationship of LECIF score to genetic and epigenetic variation associated with phenotypes.

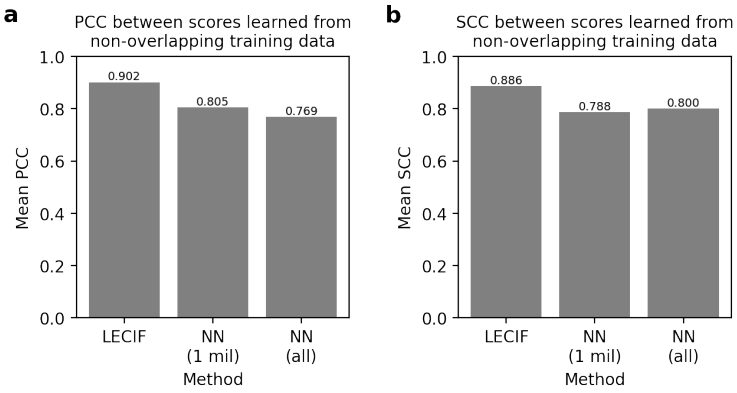
a. Distribution of mean LECIF score of non-overlapping 1-kb human genomic windows identified as lying within a mapped mouse insulin scretion QTL or containing a human diabetes GWAS variant or both[43]. 'Within mouse QTL overlapping human variant' refers to windows that lie within the mouse QTL mapped to human and overlap the human diabetes GWAS variant. 'Within mouse QTL not overlapping human variant' refers to windows within the mapped mouse QTL that do not overlap any human diabetes GWAS variant. 'Overlapping human variant' refers to windows that overlap the human diabetes GWAS variant and lie in loci obtained by randomly permuting the locations of the mapped mouse QTL. All windows were obtained by sliding a fixed window aross the QTL, and any window with less than half of its bases annotated with the LECIF score was excluded from this analysis. Displayed after each label is the number of qualified windows corresponding to that label. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). **** denotes *P* value below 0.0001 based on a two-sided Mann-Whitney U test. Similar plots generated using different window sizes are shown in **Supplementary Fig. 2.23**.

b. Distribution of mean LECIF score in conserved differentially methylated regions (DMRs) and mouse-specific DMRs with respect to a diabetic phenotype[44]. 'Conserved DMR' refers to regions with significant differential methylation (*P* value<0.05) in both human and mouse and the same directionality with respect to the phenotype. 'Mouse-specific DMR' refers to regions with significant differential methylation in mouse, but either lacking significant differential methylation in human or showing inconsistent direction of methylation change between human and mouse. The study in which the DMRs were reported did not provide human-specific DMRs because it first identified mouse DMRs and then tested those in human and not vice versa. Displayed below each label is the number of DMRs corresponding to that label. Boxplots are formatted as in **a.** ** denotes *P* value below 0.01 based on a two-sided Mann-Whintey U test (*P*=0.003).

Source data for **a** and **b** are provided as Source Data files.

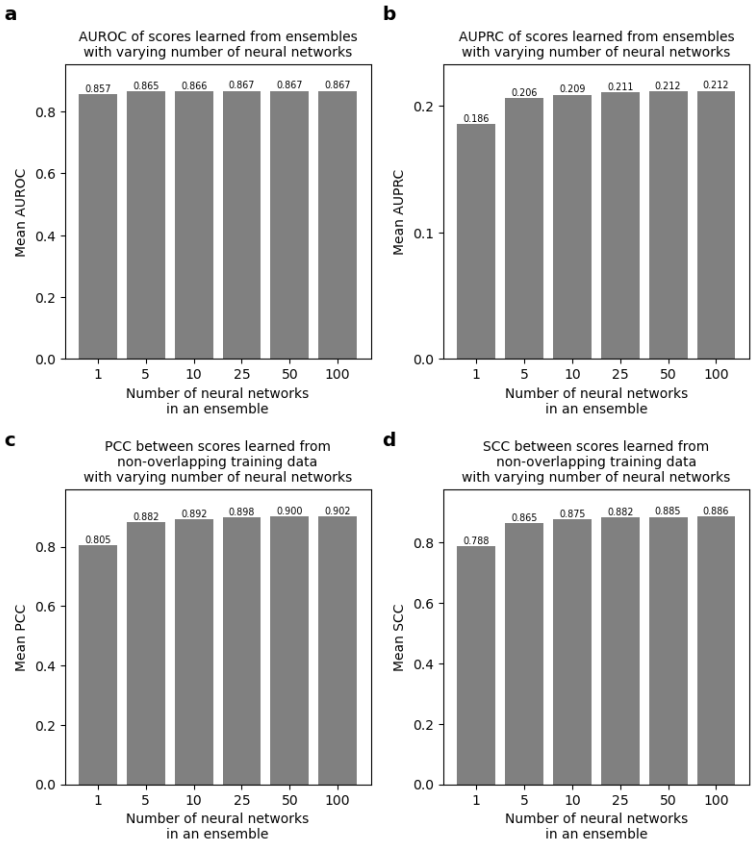

a
Distribution of score learned with 1:1 positive-negative example weight ratio
Proportion
Score learned with 1:1 ratio
b
Scatter plot of score learned with 1:1 vs. 1:1 positive-negative example weight ratio
PCC: 1.00
SCC: 1.00
Score learned with 1:1 ratio
Score learned with 1:1 ratio
c
ROC curve of score learned with 1:1 positive-negative example weight ratio
True positive rate
False positive rate
AUROC: 0.87
d
Distribution of score learned with 1:10 positive-negative example weight ratio
Proportion
Score learned with 1:10 ratio
e
Scatter plot of score learned with 1:1 vs. 1:10 positive-negative example weight ratio
PCC: 0.82
SCC: 0.96
Score learned with 1:10 ratio
Score learned with 1:1 ratio
f
ROC curve of score learned with 1:10 positive-negative example weight ratio
True positive rate
False positive rate
AUROC: 0.87
g
Distribution of score learned with 1:50 positive-negative example weight ratio
Proportion
Score learned with 1:50 ratio
h
Scatter plot of score learned with 1:1 vs. 1:50 positive-negative example weight ratio
PCC: 0.64
SCC: 0.94
Score learned with 1:50 ratio
Score learned with 1:1 ratio
i
ROC curve of score learned with 1:50 positive-negative example weight ratio
True positive rate
False positive rate
AUROC: 0.87
j
Distribution of score learned with 1:100 positive-negative example weight ratio
Proportion
Score learned with 1:100 ratio
k
Scatter plot of score learned with 1:1 vs. 1:100 positive-negative example weight ratio
PCC: 0.59
SCC: 0.94
Score learned with 1:100 ratio
Score learned with 1:1 ratio
l
ROC curve of score learned with 1:100 positive-negative example weight ratio
True positive rate
False positive rate
AUROC: 0.87

Supplementary Figure 2.1. Effect of different weight ratios between positive and negative examples.
Comparisons of the LECIF score, which was learned with negative examples weighted 50 times more than positive examples, to alternative versions of the score learned with different weighting schemes. To generate each alternative version, we repeated the hyper-parameter search and prediction procedures with the same dataset, but with different weigthing scheme.
a,d,g,j. Distribution of a score learned with positive-negative example weight ratio of 1:1, 1:10, 1:50, and 1:100, respectively. Fifty equal-width bins were used to plot this histogram.
b,e,h,k. Scatter plot showing with a gray dot for each aligning pair of human and mouse regions a score learned with positive and negative examples weighted equally (x-axis) and a score learned with positive-negative example weight ratio of 1:1, 1:10, 1:50, and 1:100, respectively (y-axis). Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) between the two scores are shown in the top left. One hundred thousand pairs of human and mouse regions were randomly selected to be included in the scatter plot.
c,f,i,l. ROC curve of a score learned with positive-negative example weight ratio of 1:1, 1:10, 1:50, and 1:100, respectively, for differentiating positive and negative pairs. Mean ROC curve was obtained by classifying 100,000 positive and 100,000 negative examples randomly sampled with replacement from all available test examples 100 times Mean area under the ROC curve (AUROC) is shown in the bottom right corner. Standard deviation of the 100 AUROC values was under 0.001 for any weight ratio.



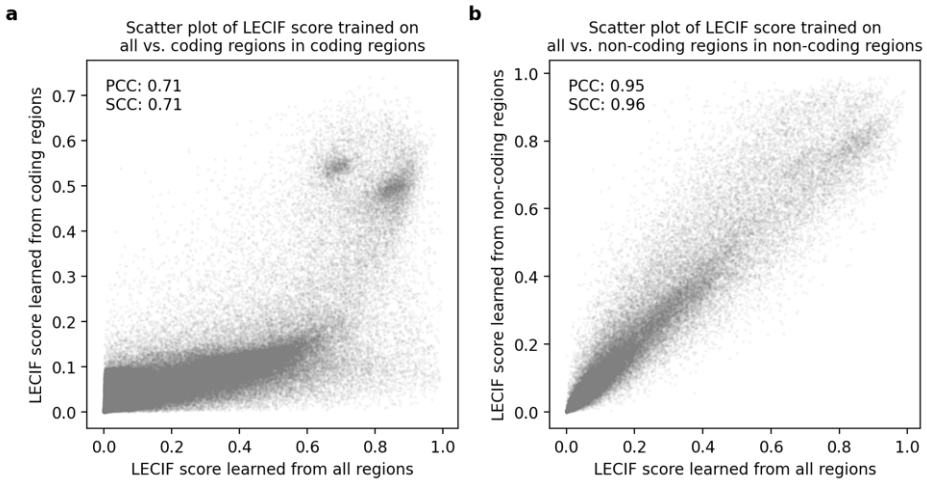
Supplementary Figure 2.2. Effect of ensembling and sampling training data on robustness.

Analysis of the effect of the ensembling strategy of LECIF, which trains an ensemble of 100 neural networks, where each neural network (NN) is given 1 million positive and 1 million negative examples that are randomly sampled from all available training data, on the robustness of predictions. To evaluate the effect of ensembling, LECIF's robustness is compared to the average robustness of individual NN in the ensemble. Additionally, to evaluate the effect of sampling training data instead of using all available training data, LECIF's robustness is compared to the robustness of a single NN trained on all available training data (>2.2 million positive and >2.2 million negative examples). We measure the robustness by computing the **a.** PCC and **b.** SCC between scores generated by classifiers that were trained on non-overlapping set of chromosomes (**Methods**). 'NN (1 mil)' refers to the mean of 100 PCC or SCC computed from pairs of NN where the two NN are trained on different data. The NN were paired up randomly. 'NN (all)' refers to the PCC or SCC from two NN where each NN was trained on non-overlapping data, but without any down-sampling as done in 'NN (1 mil)'. The scores we compare here were generated for pairs of human and mouse regions held out from training, validation, and test (**Supplementary Data 2.2**). The same set of scores were used to compute both PCC and SCC. These results confirm that LECIF leads to more robust predictions than any individual NN in the ensemble or a NN trained on all available data.



Supplementary Figure 2.3. Effect of the number of ensembled neural networks on predictive power and robustness.
Analysis of the effect of the number of neural networks used in LECIF, which trains an ensemble of 100 neural networks (NN), on classification performance and robustness of predictions. LECIF's predictive performance and robustness are compared to those of ensembles with fewer NN.
a. Effect of the number of NN in an ensemble on the area under the receiver operating characteristic curve (AUROC). Given 100 individual NN trained in LECIF, for each number of NN shown in the x-axis, *x*, we select at most 100 different ensembles, each of which is a combination of *x* NN. If there are 100 or fewer possible combinations, all are used. Otherwise, 100 combinations are randomly selected from all possible combinations. For each ensemble, we generated its prediction for test data by averaging the predictions from its NN. This test data was held out from training and validation of the NN. We finally computed AUROC for each ensemble and obtain the mean AUROC for each *x* by averaging the AUROCs over all ensembles consisting of *x* neural networks. Negative examples were weighted 50 times more than positive examples when computing AUROC.
b. Similar to **a** except showing area under the precision-recall curve (AUPRC) instead of AUROC. The same procedure and test data were used as **a**.
c. Similar to **a** except showing PCC between scores learned from different training data instead of AUROC. Ensembles were selected as done in **a** except we generated their predictions for held-out data that was excluded from all training, validation, and test (**Methods**). Given the ensembles generated for each number of NN shown in the x-axis, we computed PCC between scores predicted by two ensembles, each trained on non-overlapping training data. If there are multiple ensembles trained on different data, but with the same number of NN, then the two

ensembles are matched randomly. We then computed the mean PCC for each number of NN by averaging the PCCs over the pairs of ensembles.
d. Similar to **c** except showing SCC instead of PCC.

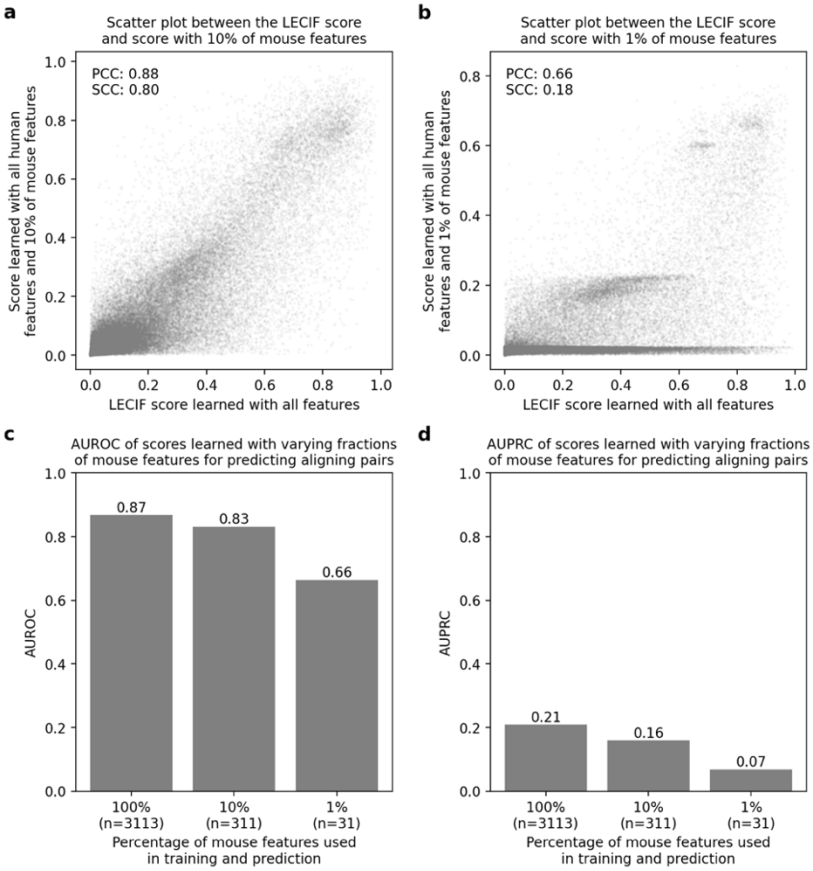


Supplementary Figure 2.4. Comparison of the LECIF score to scores learned with training data from either non-coding or coding regions.

To evaluate the effect of splitting training examples into coding and non-coding, we learned two separate scores, one from coding examples and the other from non-coding examples (**Methods**). A pair of human and mouse regions was considered coding if the human region overlapped any coding sequence and considered non-coding otherwise. The same procedure for learning the LECIF score was applied to learn a score from non-coding examples. The same procedure for learning the LECIF score was also done for coding examples, except, due to limited number of coding examples, all available training and tuning examples were used for hyperparameter search and then each classifier with optimized parameters was trained on 10,000 positive and 10,000 negative training examples. The scores learned separately on coding and non-coding regions are largely similar to the original LECIF score.

a. Scatter plot showing with a gray dot for a coding region its LECIF score learned from all regions (x-axis) and its score learned from coding regions (y-axis). Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) between the scores are shown in the top left. One hundred thousand pairs of human coding regions were randomly selected to be included in the scatter plot.

b. Similar to **a** except showing with a gray dot for a non-coding region its LECIF score learned from all regions (x-axis) and score learned from non-coding regions (y-axis).



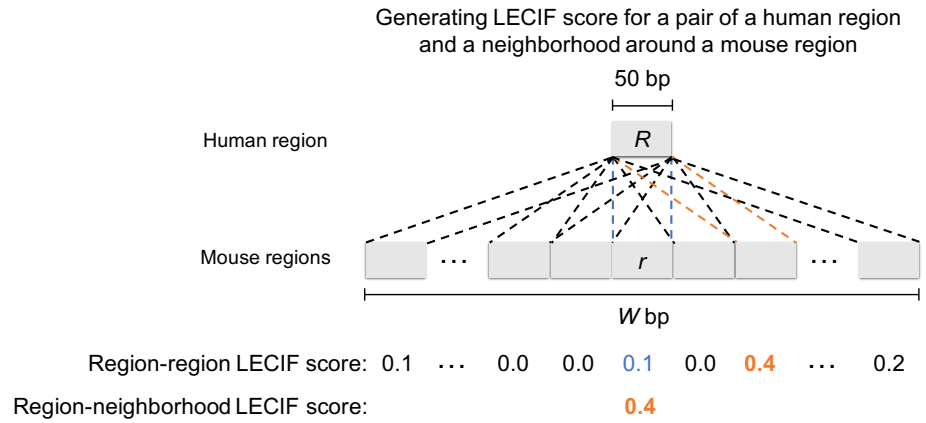
Supplementary Figure 2.5. Effect of using fewer mouse functional genomic features. To examine the contribution of mouse data to LECIF, we learned two alternative scores using LECIF, one with 10% of the original mouse features and the other with 1% (**Methods**). Specifically, to sample 10% of the mouse features, we randomly selected 6 out of 66 epigenomes in the 15-state ChromHMM chromatin state annotations, selecting 90 chromatin state features. We then additionally sampled 221 features from those corresponding to mouse DNase-seq, ChIP-seq, RNA-seq, and CAGE experiments. To sample 1% of the mouse features, we randomly selected 31 features from those corresponding to mouse DNase-seq, ChIP-seq, RNA-seq, and CAGE experiments. Both scores were learned with all human features originally used in LECIF.

a. Scatter plot showing with a gray dot for each aligning pair of human and mouse regions the LECIF score learned with all features (x-axis) and the alternative score learned with 10% of mouse features. Pearson correlation coefficient (PCC) and Spearman correlation coefficient (SCC) between the two scores are shown in the top left. One hundred thousand pairs of human and mouse regions were randomly selected to be included in the scatter plot.

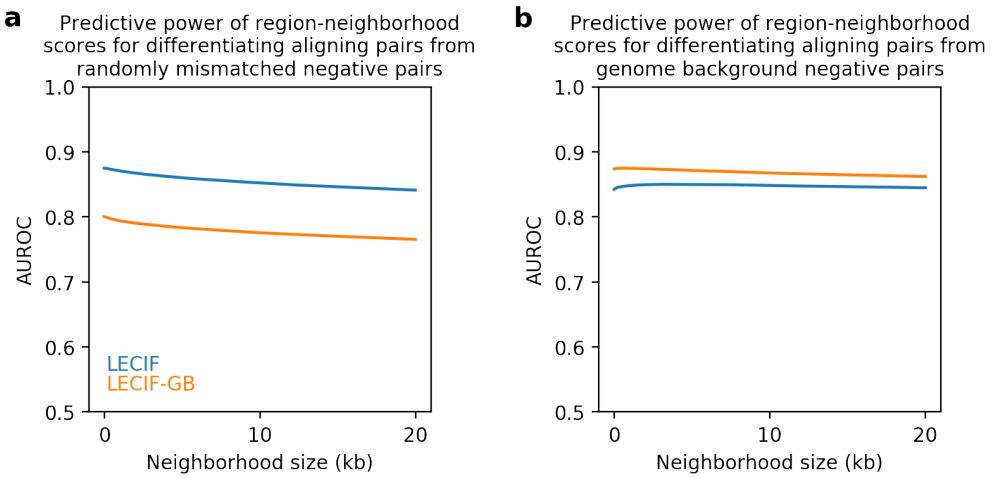
b. Similar to **a** except showing the alternative score learned with 1% of mouse features in the y-axis.

c. Bar plot showing mean AUROC of the LECIF score learned with all features and the alternative scores learned with all human features and 10% or 1% of mouse features for differentiating aligning pairs from randomly mismatched pairs. One hundred AUROCs were obtained by classifying 100,000 positive and 100,000 negative examples randomly sampled with replacement from all available test examples 100 times, as done in **Fig. 2.2c**. Mean AUROC is shown above each bar. Standard deviation of the 100 AUROC values was under 0.001 for all scores.

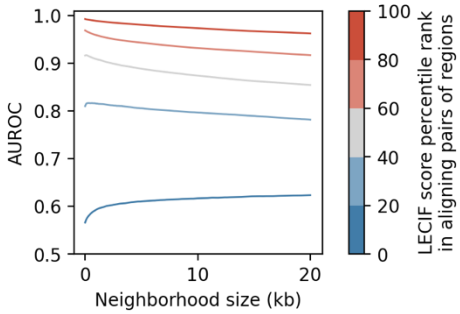
d. Similar to **c** except showing AUPRC instead of AUROC.



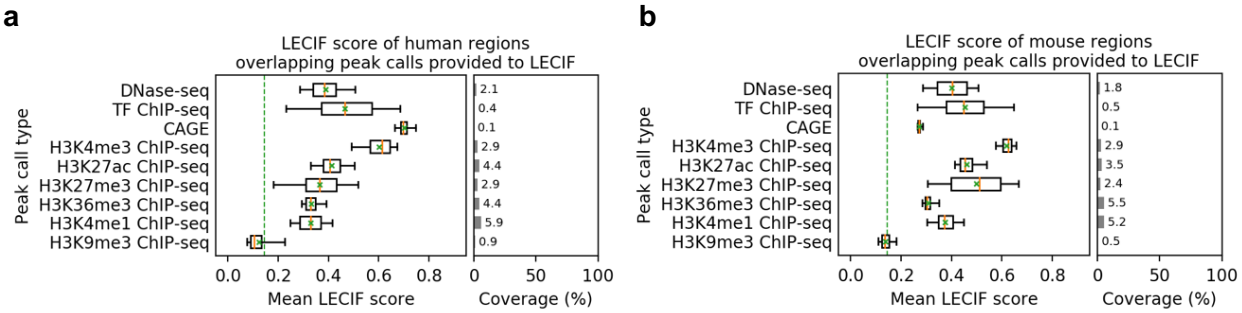
Supplementary Figure 2.6. Overview of generating region-neighborhood LECIF score for pairs of human regions and extended mouse regions.
Illustration of how LECIF is used to generate region-neighborhood LECIF score for a pair of a human region and a neighborhood of a mouse region. A given 50-bp human region *R* is compared to a set of multiple 50-bp mouse regions in a neighborhood of length *W* bp centered around a mouse region *r*. Each comparison (pair of dashed lines) results in a region-region LECIF score. For a pair of human region and a neighborhood in mouse, we define the region-neighborhood LECIF score as the maximum of all the region-region LECIF scores. In this example, the region-region LECIF score of the aligning human and mouse regions (blue; *R* and *r*) is 0.1. The maximum region-region LECIF score, 0.4, comes from the human region paired up with a mouse region near the aligning mouse region (orange). As a result, in this example, the region-neighborhood LECIF score is 0.4. We evaluated using the region-neighborhood LECIF score to predict aligning pairs, as an alternative to using the region-region LECIF score of the aligning human and mouse regions. Results of the evaluation are shown in **Supplementary Fig. 2.7**.



Supplementary Figure 2.7. Predictive power of region-neighborhood LECIF score for aligning pairs as a function of neighborhood size around each pair's mouse region. We evaluate the predictive power of the region-neighborhood LECIF score of aligning human and mouse regions as a function of neighborhood size. We also evaluate using a LECIF-Genome Background (LECIF-GB) score in place of LECIF score in this analysis. LECIF-GB was trained with 'genome background' negative examples, which are pairs of human and mouse regions randomly selected from the entire human and mouse genomes (**Methods**). Shown for LECIF (blue) and LECIF-GB (orange) is the area under the ROC curve (AUROC) for differentiating positive examples from negative examples as a function of the size of the neighborhood centered around each pair's mouse region. Positive examples are pairs of human and mouse regions that align to each other. Negative examples are either **a.** randomly mismatched human and mouse regions that align somewhere in the other species (equivalent to the negative examples provided to LECIF) or **b.** genome background (equivalent to the negative examples provided to LECIF-GB). The neighborhood size varies from 0 to 20 kb with increments of 100 bp. Given a particular neighborhood size of *W*, the region-neighborhood score for each pair of human and mouse regions was the maximum region-region scores of any pair consisting of the human region and any mouse region within 0.5**W* bp from the aligning mouse region of the pair (**Methods**; **Supplementary Fig. 2.6**). This region-neighborhood LECIF score was then used to predict aligning pairs. We note that a neighborhood size of 0 gives region-region LECIF and LECIF-GB scores. For each comparison, the same set of 100,000 positive and 100,000 negative test examples, which were on chromosomes excluded from training and validation, were used to compute the AUROC. In this analysis, there was no advantage in using the region-neighborhood LECIF score, as defined, compared to using the region-region LECIF score and similarly for LECIF-GB.



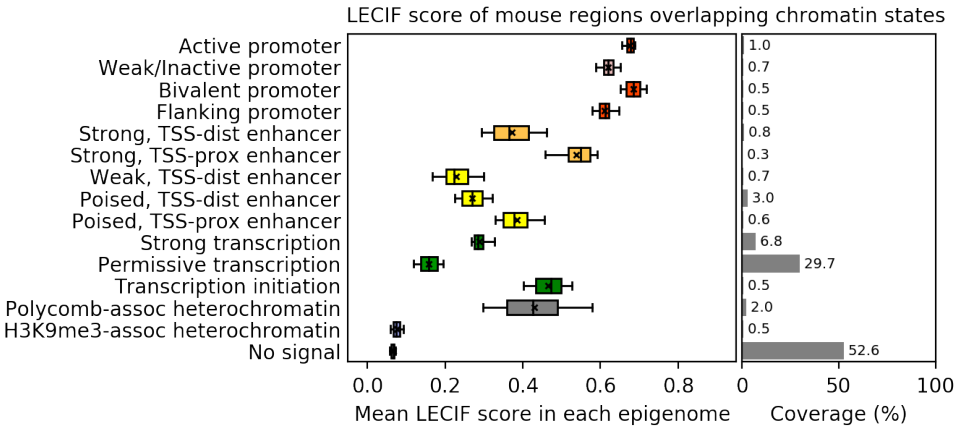
Supplementary Figure 2.8. Predictive power of region-neighborhood LECIF score for aligning pairs binned by score percentile as a function of neighborhood size around each pair's mouse region.
Similar to **Supplementary Fig. 2.7a**, except we first bin aligning pairs into five bins based on their region-region LECIF score percentile rank at the aligning regions. For each bin, we evaluate the predictive power of the region-neighborhood LECIF score of aligning human and mouse regions as a function of neighborhood size. Each line corresponds to a percentile rank bin and is colored based on the color bar on the right. When measuring AUROC, for every positive example falling into a percentile rank bin, we provide a negative example that consists of the same human region of the positive example and a randomly chosen mouse region that aligns somewhere else in the human genome. While extending the neighborhood around each pair's mouse region does not improve predictive power in general, it does help when the aligning regions are scoring low and hard to distinguish from randomly mismatched pairs.



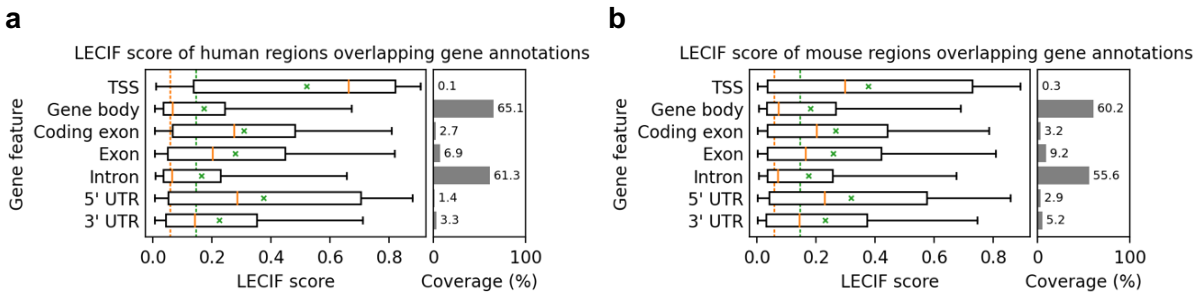
Supplementary Figure 2.9. Distribution of mean LECIF score of peak calls provided to LECIF.

a. Left panel shows for each type of functional genomic experiments listed the distribution of mean LECIF score over experiments of that type in human. The mean LECIF score for an experiment is computed based on averaging the LECIF score of regions overlapping a peak call from the experiment. The set of experiments are the same as provided to LECIF as input features. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). Green dashed vertical line across the entire left panel denotes the genome-wide mean LECIF score. Right panel shows mean coverage of each type of peak call across all human regions that align to the mouse genome. Human regions in all aligning pairs of human and mouse regions (n=32,285,361) as defined in **Methods** were used to generate this plot. The number of experiments for each peak call type is reported in **Supplementary Data 2.1**. Source data are provided as a Source Data file.

b. Similar to **a** except for mouse experiments instead of human**.** Mouse regions in all aligning pairs of human and mouse regions (n=32,285,361) were used to generate this plot. Source data are provided as a Source Data file.



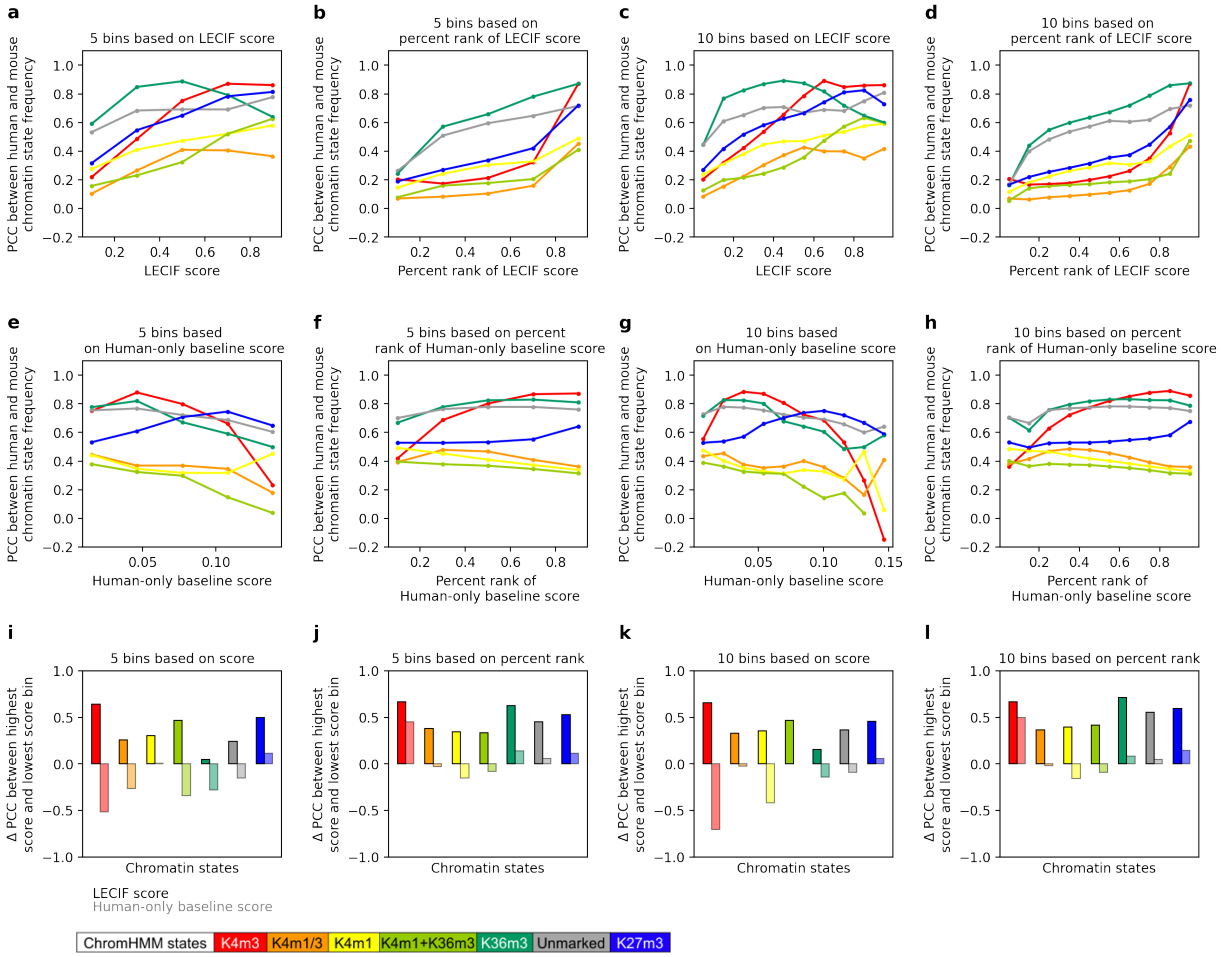
Supplementary Figure 2.10. Distribution of mean LECIF score in different mouse chromatin states.
Similar to **Fig. 2.2e** except for mouse chromatin state annotations[29,46] instead of human**.** Left panel shows for each chromatin state from a model learned in mouse the distribution of mean LECIF score over different epigenomes (n=66). The mean LECIF score for a chromain state in an epigenome is computed by averaging the LECIF score of regions overlapping the chromatin state in the epigenome. Each distribution is represented by a boxplot with median (black vertical line), mean (black 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). Right panel shows mean coverage of each state across mouse regions in all aligning pairs of human and mouse regions. Mouse regions in all aligning pairs of human and mouse regions (n=32,285,361) as defined in **Methods** were used to generate this plot. State colors were assigned to match the state colors of the 25-state human ChromHMM model[4] shown in **Fig. 2.2e** based on state descriptions. Source data are provided as a Source Data file.



Supplementary Figure 2.11. Distribution of LECIF score of GENCODE gene feature annotations.

a. Left panel shows the distribution of LECIF score in human regions overlapping indicated GENCODE gene feature annotations. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). Dashed vertical lines in orange and green across the entire left panel denote the genome-wide median and mean LECIF scores, respectively. Right panel shows coverage of each annotation across all human regions that align to the mouse genome. Human regions in all aligning pairs of human and mouse regions (n=32,285,361) were used to generate this plot. TSS: transcription start site; CDS: coding sequence; UTR: untranslated region.

b. Similar to **a** except for mouse regions overlapping mouse gene feature annotations instead of human. Mouse regions in all aligning pairs of human and mouse regions (n=32,285,361) as defined in **Methods** were used to generate this plot.



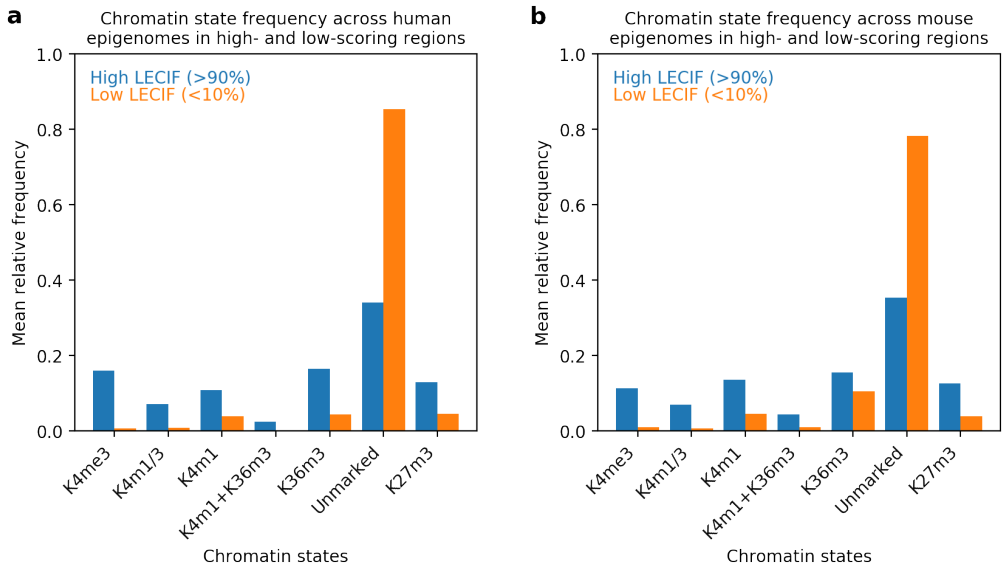
Supplementary Figure 2.12. Cross-species similarity in chromatin states in pairs binned by LECIF score or human-only baseline score.

Extended version of the analysis in **Fig 2.3b**. **a-d.** Cross-species agreement in chromatin state[2,29] frequency in pairs of aligning human and mouse regions binned by the LECIF score for a ChromHMM model learned jointly between human and mouse. Pairs are binned using **a.** 5 equal-width bins based on the LECIF score, **b.** 5 bins based on the percentile rank of the LECIF score, **c.** 10 equal-width bins based on the LECIF score, or **d.** 10 bins based on the percentile rank of the LECIF score. Binning based on the percentile rank results in similar number of pairs in each bin, whereas binning based on the score results in varying number of pairs in each bin. For each state and aligning region, we computed the frequency of the state across cell and tissue types for human and mouse separately. We then, for each state and bin, computed the PCC between the corresponding human and mouse frequencies for that state across all aligning pairs within the bin (**Methods**). The values are shown with colored circles according to the chromatin state legend on the bottom from Ref. [2]. The circles for the same state are connected with lines based on piecewise linear interpolation. **d** is identical to **Fig. 2.3b**. Source data are provided as a Source Data file.

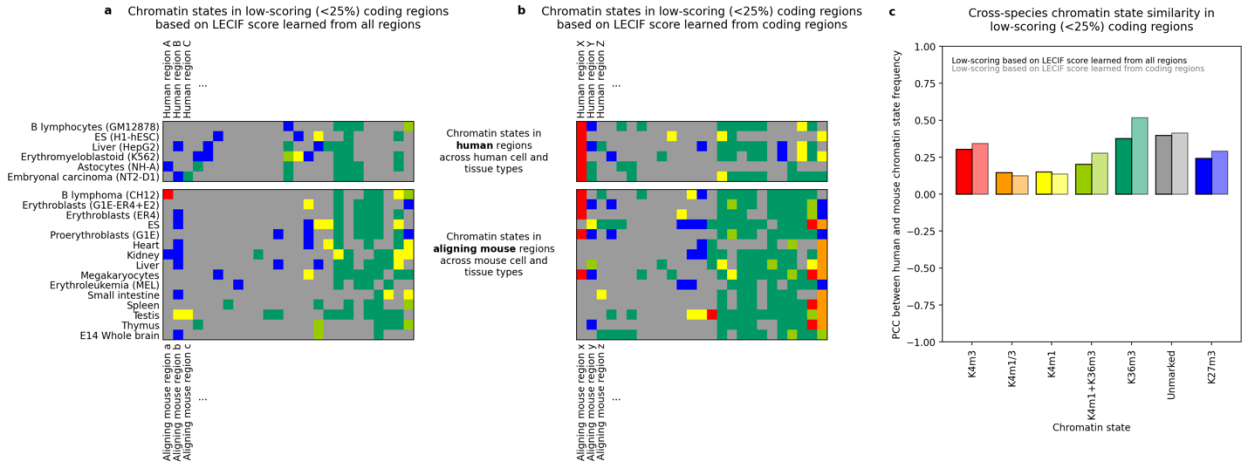
e-h. Similar to **a-d**, respectively, except using the human-only baseline score instead of the LECIF score.

i-l. Shown for each chromatin state (x-axis) is the difference in the chromatin state's PCC between pairs from the highest score and lowest score bin Δ(PCC), based on either the LECIF score (bold-colored bars) or human-only baseline score (light-colored bars). Each panel corresponds to the two panels above it in the same column. The ΔPCC values are shown with

colored bars according to the chromatin state legend on the bottom from Ref. [2] and the score used for binning the pairs (bold for LECIF score, light for human-only baseline score).
This figure illustrates that pairs of human and mouse regions with high LECIF score show stronger cross-species agreement in chromatin state frequency than pairs with low LECIF score. It also highlights that pairs with high human-only baseline score do not consistently show stronger cross-species agreement than pairs with low human-only baseline score.



Supplementary Figure 2.13. Relative frequency of chromatin states in regions with low or high LECIF score.
Comparing relative frequency of chromatin states[2,29] for a seven state ChromHMM model learned jointly between human and mouse in high LECIF score (>90th percentile; blue) and low LECIF score (<10th percentile; orange) regions. The comparison is shown both for **a.** human and **b.** mouse regions. The chromatin states are the same as in **Fig. 2.3b** and **Supplementary Fig. 2.12**. For a species, the mean relative frequency of a chromatin state in a set of regions satisfying the LECIF score threshold was computed by averaging over epigenomes the fraction of those regions overlapping the chromatin state in each epigenome. These figures illustrate that regions with low LECIF score are more likely to be annotated with the 'Unmarked' chromatin state in both human and mouse than regions with high LECIF score.

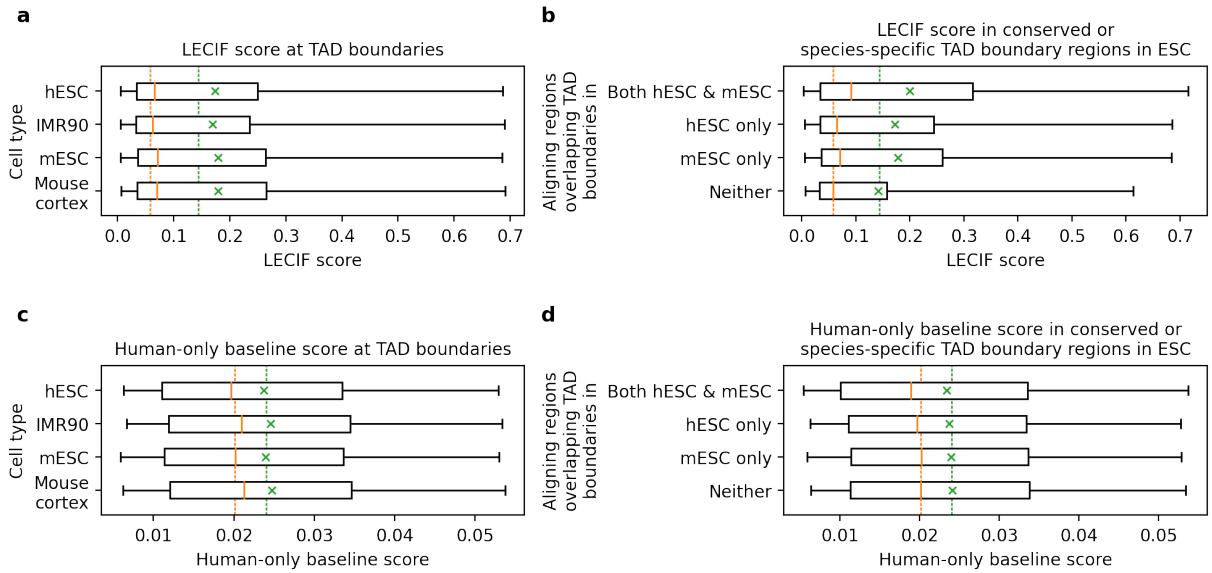


Supplementary Figure 2.14. Chromatin state similarity in low-scoring coding regions. As described in **Supplementary Figure 2.4**, we learned an alternative score using pairs with coding human regions. Here we examine human and mouse chromatin states in low-scoring coding regions based on either the original LECIF score learned from all regions or the alternative score learned from coding regions. Coding regions that score low according to either scores exhibit weak cross-species similarity in their chromatin states as expected.

a. ChromHMM chromatin state[2,29] annotations in randomly selected pairs that include a human coding region with low LECIF score. The pairs were selected based on whether their human regions overlapped GENCODE annotation of coding sequence (CDS). Each row in the top sub-panel corresponds to a human cell or tissue type. Each row in the bottom sub-panel corresponds to a mouse cell or tissue type. Each column is a randomly selected pair with a human coding region with low LECIF score among all pairs with a human coding region (<25th percentile among coding regions). Each cell shows the color of the chromatin state with which the human or mouse region (column) is annotated in a specific cell or tissue type (row). The chromatin state model and state coloring are the same as in **Fig. 2.3b** and **Supplementary Fig. 2.12**. Pairs (columns) were ordered based on hierarchical clustering applied to their chromatin state annotations using Ward's linkage with optimal leaf ordering[67].

b. Same as **a**, but with pairs selected based on the alternative score learned from coding regions instead of the LECIF score.

c. Shown for each chromatin state (x-axis) is the state's PCC in low-scoring pairs with a human coding region based on the LECIF score (<25th percentile among coding regions; bold-colored bars) or the alternative score learned from coding training data (light-colored bars). Each state's PCC was computed as explained in **Fig. 2.3b** and **Supplementary Fig. 2.12** where the correlation is computed between the state's frequencies in human cell or tissue types and its frequencies in mouse cell or tissue types across all low-scoring pairs restricted to human coding regions.

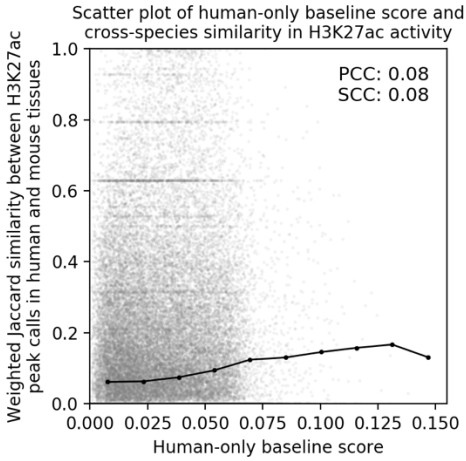


Supplementary Figure 2.15. **LECIF score and human-only baseline score in topologically associated domain (TAD) boundaries**[33]**.**

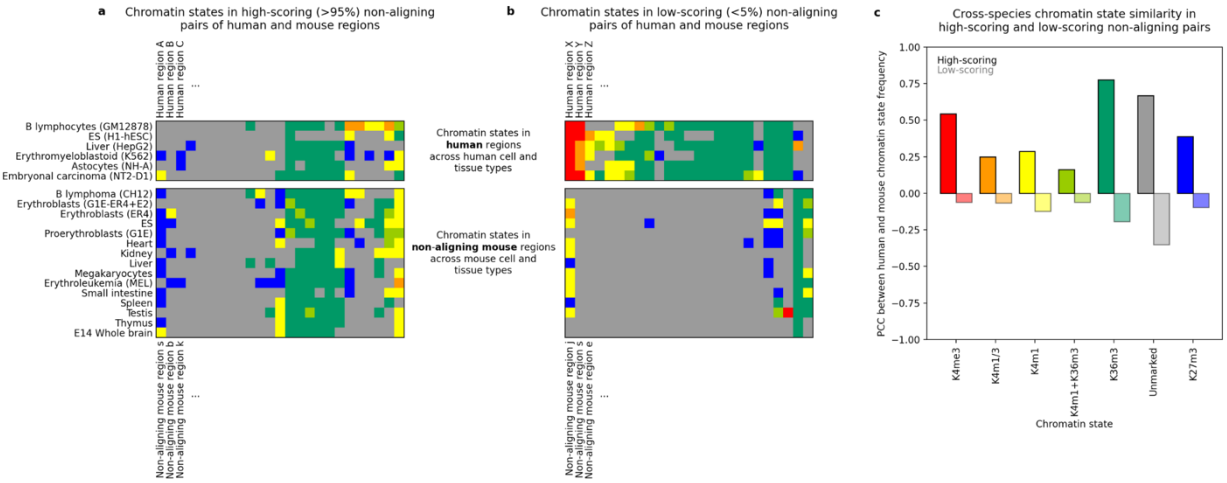
a. Box plot showing the distribution of LECIF score of pairs with a human or mouse genomic region overlapping TAD boundaries in different human or mouse cell types. Top two cell types listed along the y-axis are human cell types, and the other two are mouse cell types. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). Orange and green dashed lines vertical lines across the entire panel denote the genome-wide median and mean LECIF scores, respectively. There were 1,488,669, 1,344,362, 1,731,487, and 1,995,527 pairs of human and mouse regions as defined in **Methods** overlapping TAD boundaries in human embryonic stem cells (hESC), IMR90, mouse embryonic stem cells (mESC), and mouse cortex, respectively.

b. Similar to **a** but showing the distribution of LECIF score of pairs with human and mouse regions with respect to their overlap with TAD boundaries in embryonic stem cells (ESC). Top distribution corresponds to aligning human and mouse regions overlapping TAD boundaries in both hESC and mESC ('Both hESC & mESC'; n=82,075). Second and third distributions correspond to aligning pairs with either human or mouse region overlapping TAD boundaries in ESC ('hESC only' and 'mESC only'; n=1,406,056 and 1,234,447, respectively). Bottom distribution corresponds to the remaining pairs which are those with neither region overlapping TAD boundaries in ESC (n=29,552,172).

c-d. Similar to **a-b**, respectively, except for human-only baseline score instead of LECIF score. These results show that LECIF score is higher at TAD boundaries than average, which are known to be highly conserved between human and mouse, and also higher in conserved TAD boundary regions than in species-specific TAD boundary regions. These patterns are not consistently observed with human-only baseline score.



Supplementary Figure 2.16. **Scatter plot of the human-only baseline score and cross-species similarity in tissue-specific H3K27ac activity.**
Similar scatter plot to **Fig. 3a** except for the human-only baseline score (**Methods**) instead of the LECIF score. The scatter plot shows with a gray dot for each aligning pair of human and mouse regions the human-only baseline score (x-axis) and cross-species similarity of matched tissue-specific H3K27ac activity (y-axis). The H3K27ac activity for a region in a tissue and species is quantified as the fraction of experiments in the tissue type of the species with peak calls overlapping the region. The cross-species similarity of the tissue-specific H3K27ac activity is quantified as the weighted Jaccard similarity coefficient over 14 matched tissue types (**Methods**). PCC and SCC computed from all aligning pairs are shown in the top right. In black circles the mean similarity coefficient of pairs binned by the LECIF score with ten equal-width bins spanning from the minimum to maximum of the human-only baseline score is shown. These circles are connected with lines determined based on piecewise linear interpolation. One hundred thousand random aligning pairs were sampled to plot the scatter plot. This analysis shows that the human-only baseline score exhibits weaker agreement with cross-species similarity in tissue-specific H3K27ac activity compared to the LECIF score (PCC: 0.08 vs 0.45 and SCC: 0.08 vs 0.42).

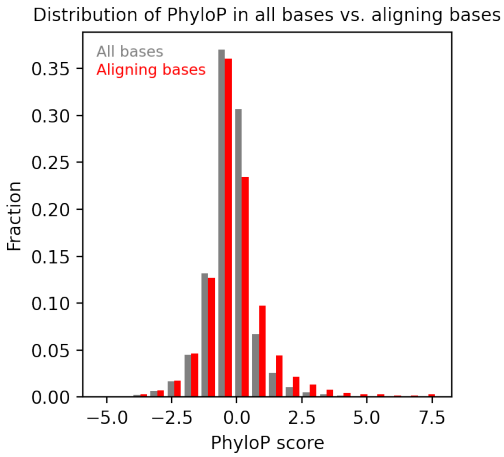


Supplementary Figure 2.17. Chromatin states in non-aligning pairs with high or low LECIF scores.

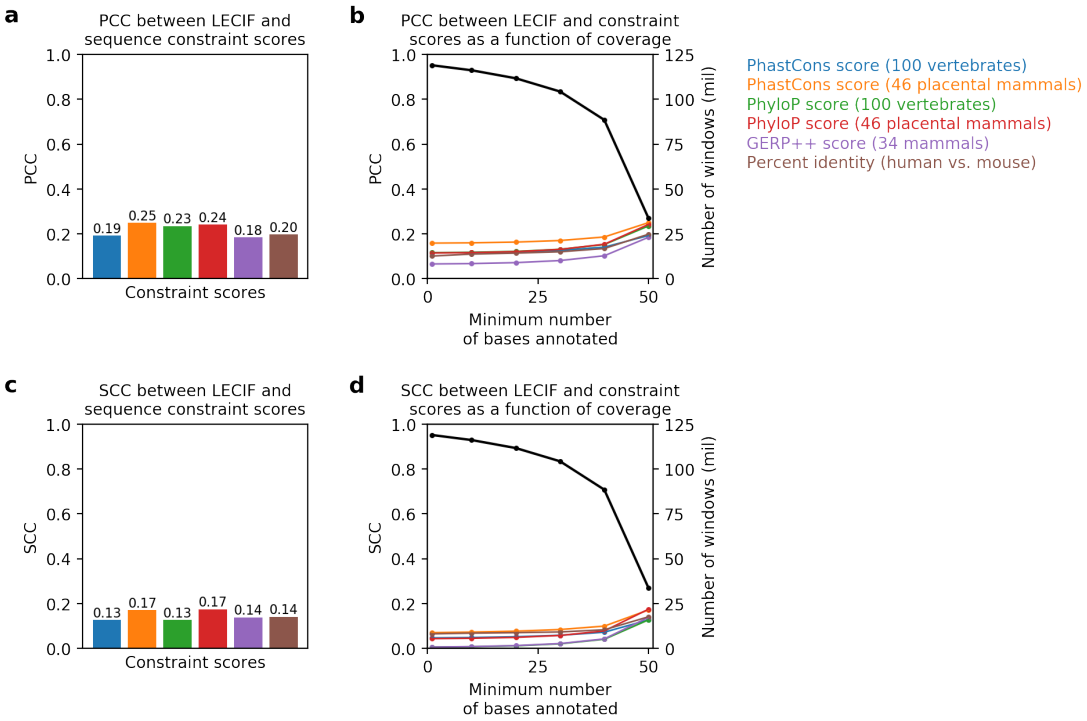
a. ChromHMM chromatin state[2,29] annotations in randomly selected pairs of non-aligning human and mouse regions with high LECIF score. The pairs were selected from negative test examples which consist of randomly mismatched pairs of human and mouse regions that do not align to each other (**Methods**). All human and mouse regions included in these pairs do align somewhere in the other species. Each row in the top sub-panel corresponds to a human cell or tissue type. Each row in the bottom sub-panel corresponds to a mouse cell or tissue type. Each column is a randomly selected non-aligning pair with high LECIF score among all non-aligning pairs (>95th percentile). Each cell shows the color of the chromatin state with which the human or mouse region (column) is annotated in a specific cell or tissue type (row). The chromatin state model and state coloring are the same as in **Fig. 2.3b** and **Supplementary Fig. 2.12**. Pairs (columns) were ordered based on hierarchical clustering applied to their chromatin state annotations using Ward's linkage with optimal leaf ordering[67].

b. Same as **a**, but with randomly selected non-aligning pairs with low LECIF score (<5th percentile).

c. Shown for each chromatin state (x-axis) is the state's PCC in non-aligning pairs with high (>95th percentile; bold-colored bars) or low (<5th percentile; light-colored bars) LECIF score. Each state's PCC was computed as explained in **Fig. 2.3b** and **Supplementary Fig. 2.12** where the correlation is computed between the state's frequencies in human cell or tissue types and its frequencies in mouse cell or tissue types across 100,000 pairs with either high or low LECIF scores. Pairs were randomly sampled from negative test examples as done in **a** and **b**.



Supplementary Figure 2.18. **Distribution of PhyloP score in aligning bases.** Comparison of the distribution of PhyloP score (100 vertebrate) in human genomic bases in general (gray) and bases that align to mouse (red). 1 million bases annotated by PhyloP score were randomly sampled from the genome. Shown in gray is the distribution of PhyloP score of all 1 million bases. Shown in red is the distribution of PhyloP score of bases that align to mouse among the 1 million bases. Twenty equal-width bins ranging from -5 to 8 were used to plot the histogram, covering more than 99% of the score distribution. Bins outside the range are not shown. This comparison demonstrates that although aligning bases have a slightly higher distribution of sequence constraint than all bases they still have a wide distribution of constraint.



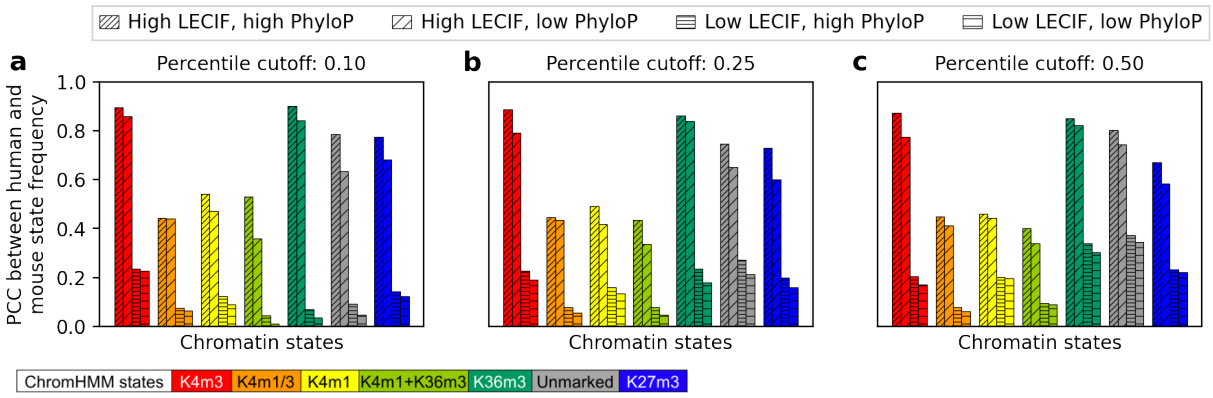
Supplementary Figure 2.19. Correlation between LECIF score and sequence constraint scores.

a. Shown for a set of sequence constraint scores[11–13] is the PCC computed between the LECIF score and a constraint score. For a given constraint score, to compute the PCC, we first slid a non-overlapping 50-bp genomic window across the human genome and selected windows with all 50 bases annotated by both the LECIF score and the given constraint score. We then computed the mean LECIF score and mean constraint score for each selected window. The PCC for a constraint score is the PCC between those two sets of values. Each resulting PCC is shown with a bar colored according to the legend on the right. Percent identity is defined as the number base-pairs with matching nucleotides (e.g. G in human and G is mouse) within a given window divided by 50. Source data are provided as a Source Data file.

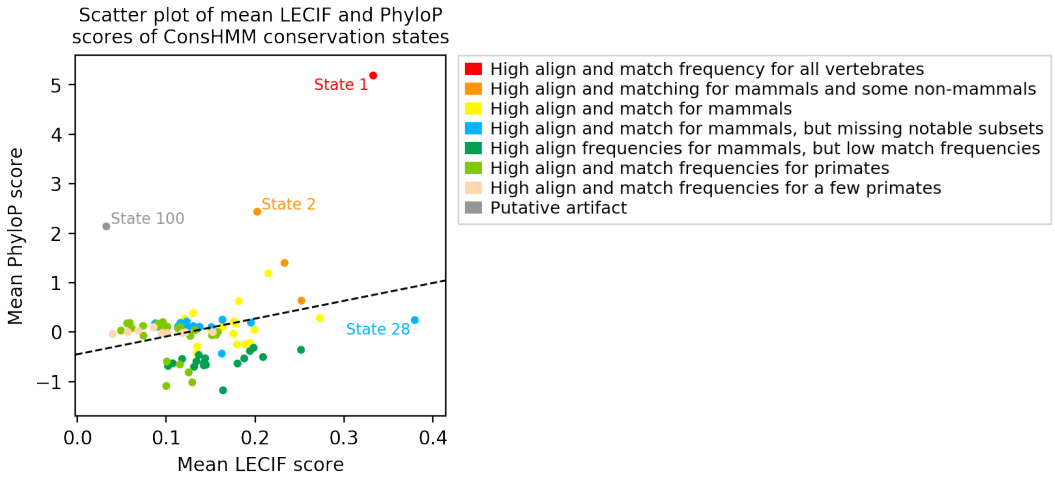
b. PCC between the LECIF score and constraint scores as a function of the minimum number of bases required to be annotated in the genomic windows. Also shown is the number of windows selected to compute the PCC. The PCC for a constraint score is computed as described in **a,** except windows with at least *n* bases annotated by the LECIF score and the constraint score of interest are selected, where *n* varies from 1 to 50. The two scores being compared need not annotate the same set of bases in each window. The PCC are shown with colored circles according to the y-axis on the left and legend in the top right. The circles for the same constraint score are connected with lines based on piecewise linear interpolation. The rightmost values where the minimum number of bases annotated equals 50 correspond to the PCC shown in **a**. Black circles show the number of windows in millions that had at least *n* bases (x-axis) annotated by the LECIF score and constraint scores according to the y-axis on the right. These circles are connected with lines based on piecewise linear interpolation. All six comparisons of the LECIF score to constraint scores had the same number of selected genomic windows. Source data are provided as a Source Data file.

c-d. Similar to **a-b**, respectively, except for SCC instead of PCC.

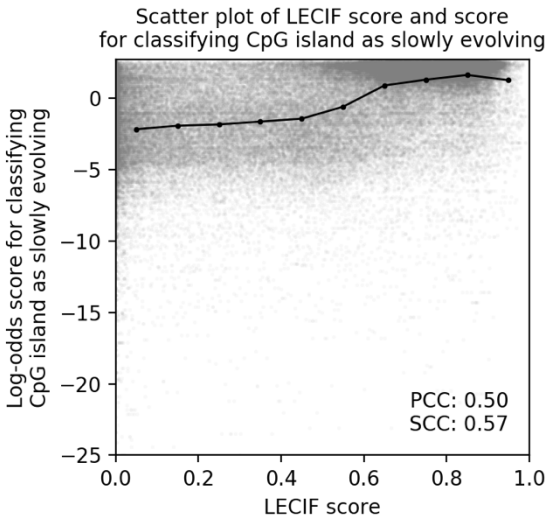
These results show that the LECIF score is moderately correlated with sequence constraint scores, and that the correlations are weaker as we include windows with fewer bases annotated by the scores within each genomic window.



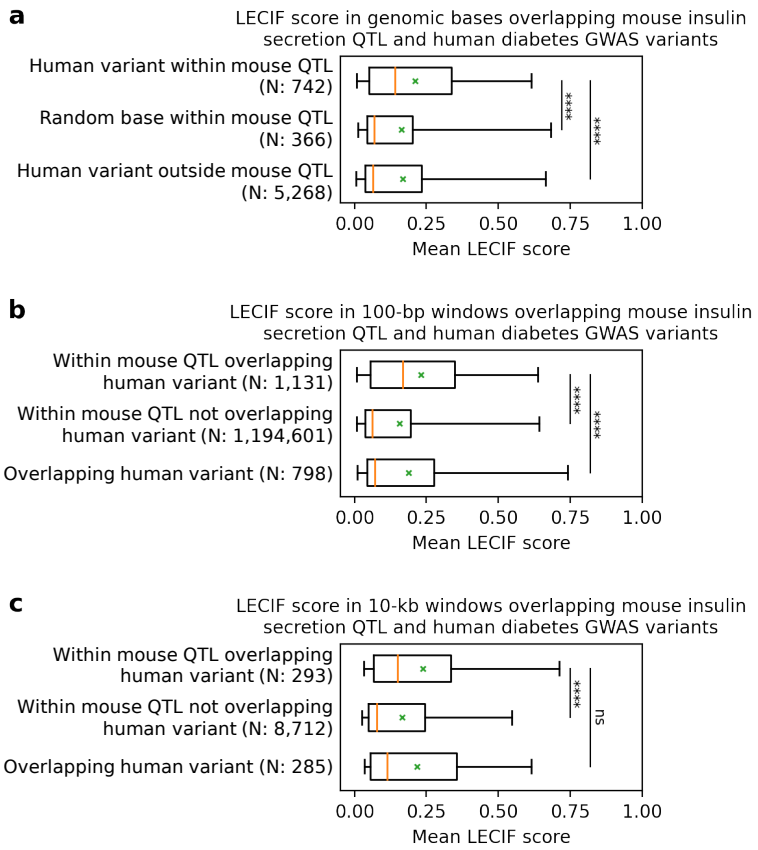
Supplementary Figure 2.20. Cross-species agreement in chromatin state frequency in pairs grouped based on LECIF score and PhyloP score.
a. ChromHMM chromatin state[2,69] frequency correlation between human and mouse in pairs of aligning human and mouse regions grouped based on whether their LECIF score and human PhyloP score[12] (defined based on a 100-way vertebrate alignment) were high (>90th percentile) or low (<10th percentile). The chromatin states are the same as in **Fig. 2.3b** and **Supplementary Fig. 2.12**. Separate bars are shown for each combination of high or low score of LECIF or PhyloP as indicated based on the legend at top. For the low PhyloP case, we required that there be a low (<10th percentile) score at all annotated bases within 500 bp to ensure the low score was not driven by the higher resolution at which sequence conservation is defined. The frequency correlation for each state and a set of aligning pairs is quantified as the PCC between the human and mouse frequencies for that state across the pairs, as done in **Fig. 2.3b** and **Supplementary Fig. 2.12** (**Methods**). Any region that did not have a PhyloP score for all bases was discarded from this analysis. Bars for each state are colored according to the bottom legend, as previously defined in Ref. [2]. Source data are provided as a Source Data file.
b. Similar to **a** except using a percentile cutoff of 0.25 instead of 0.05. Scores above the 75th percentile are considered high, and scores below the 25th percentile are considered low.
c. Similar to **a** except using a percentile cutoff of 0.50 instead of 0.05. Scores above the median are considered high, and scores below the median are considered low.
These results demonstrate that pairs with high LECIF score exhibit strong cross-species agreement in chromatin state frequency even when there is a low PhyloP score in the region. In contrast, pairs with a high PhyloP score and a low LECIF score did not exhibit strong correlations.



Supplementary Figure 2.21. Relationship of LECIF score and PhyloP score in ConsHMM conservation states.
We use a ConsHMM 100-conservation-state annotation of the human genome based on a 100-way vertebrate sequence alignment[35] to understand the relationship between the LECIF score and sequence constraint scores. The scatter plot shows with a dot for each ConsHMM conservation state the mean LECIF score (x-axis) and mean human PhyloP score[12] (y-axis; defined based on a 100-way vertebrate alignment). For each conservation state, the mean LECIF or PhyloP score is computed by averaging the score of bases overlapping the conservation state. Each dot is colored according to the eight major groups of conservation states listed in the legend on the right, as previously defined in Ref. [35]. Dashed line is a linear regression fit applied to the 100 data points. We label four noteworthy conservation states. State 28 (blue), which is the promoter enriched state, has the highest mean LECIF score and the 12th highest mean PhyloP score. State 1 (red), which is the most enriched state for exons, has the 2nd highest mean LECIF score and the highest mean PhyloP score. State 2 (orange), which is the state most enriched for enhancer chromatin states, has the 8th highest mean LECIF score and the 2nd highest mean PhyloP score. State 100 (gray), which is characterized by pseudogenes and putative artifacts in the multi-species sequence alignment[35], has the lowest mean LECIF score, while having the 3rd highest mean PhyloP score. Source data are provided as a Source Data file.

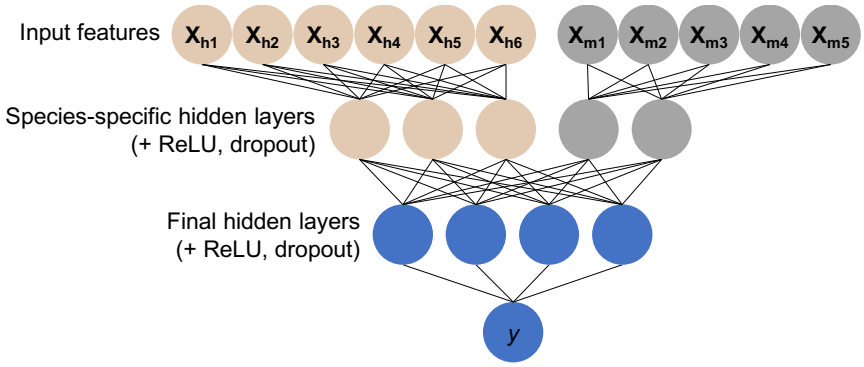


Supplementary Figure 2.22. Relationship of LECIF score and log-odds score for CpG island being classified as slowly evolving.
Scatter plot showing with a gray dot for each human CpG island the mean LECIF score (x-axis) and the log-odds score for classifying the CpG island as slowly evolving as opposed to quickly evolving (y-axis) from a previous study on primate CpG island sequence evolution[36]. In black circles the mean log-odds score for CpG islands binned by the LECIF score with ten equal-width bins is shown. These circles are connected with lines based on piecewise linear interpolation. One hundred thousand random human CpG islands annotated with the LECIF score were sampled to plot this scatter plot. PCC and SCC computed between the two scores across all CpG islands annotated with the LECIF score are shown in the bottom right. This illustrates that the LECIF score is positively correlated with the likelihood of a human CpG island being classified as slowly evolving as opposed to quickly evolving. Source data are provided as a Source Data file.



Supplementary Figure 2.23. Distribution of mean LECIF score of human genomic windows overlapping mouse insulin secretion QTL and human diabetes GWAS variant. **a.** Distribution of mean LECIF score in genomic bases identified as human diabetes GWAS variant or overlapping a mapped mouse insulin secretion QTL or both[43]. The top group refers to human GWAS variants that lie within the mouse QTL mapped to human. The middle groups refers to random genomic bases that overlap the mapped mouse QTL where the bases were obtained by randomly permutating the locations of the human diabetes GWAS variants. The bottom groups refers to human GWAS variants that do not overlap any mapped mouse QTL. Displayed after each label is the number of bases corresponding to that group. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). **** denotes *P* value below 0.0001 based on a two-sided Mann-Whitney U test. Specifically, the *P* values for comparing the top vs. middle groups and top vs. bottom groups were 2e-6 and 1e-15, respectively. Source data are provided as a Source Data file.
b. Similar to **Fig. 2.6a**, but showing the distribution of mean LECIF score in non-overlapping 100-bp genomic windows, instead of 1-kb windows, identified as containing a human diabetes GWAS variant or overlapping a mapped mouse insulin secretion QTL or both[43]. The top group refers to windows that lie within the mouse QTL mapped to human and overlap the human GWAS variant. The middle group refers to windows within the mouse QTL that do not overlap the human GWAS variant. The bottom group refers to windows from the human genome that overlap the GWAS variant where the windows lie in loci obtained by randomly permuting the locations of the mapped mouse QTL. Displayed after each label is the number of windows corresponding to that group. Each distribution is represented by a boxplot with median (orange solid line), mean (green 'x'), 25th and 75th percentiles (box), and 5th and 95th percentiles (whisker). All windows were obtained by sliding a fixed window across the QTL, and any window

with less than half of the bases annotated with the LECIF score was excluded. **** denotes *P* value below 0.0001, and ns denotes *P* value above 0.05 based on a two-sided Mann-Whitney U test. Specifically, the *P* values for comparing the top vs. middle groups and the top vs. bottom groups were 4e-54 and 7e-11, respectively. Source data are provided as a Source Data file.
c. Similar to **b**, but showing the distribution of mean LECIF score in non-overlapping 10-kb genomic windows, instead of 100-bp genomic windows. The *P* values for comparing the top vs. middle groups and top vs. bottom groups were 5e-11 and 0.10, respectively. Source data are provided as a Source Data file.
a shows that human diabetes GWAS variants that overlap mouse insulin secretion QTL tend to have a higher LECIF score than the GWAS variants outside of the mouse QTL or bases that are not GWAS variants, but within the mouse QTL. **b** and **c** show the result of **Fig. 2.6a,** that human genomic windows that overlap both mouse insulin secretion QTL and human diabetes GWAS variant tend to have a higher LECIF score than windows that overlap only one of them and that this result also holds for other window sizes.



Supplementary Figure 2.24. A schematic of a pseudo-Siamese neural network. A pseudo-Siamese neural network consists of two distinct sub-networks that do not share any weights[47]. The sub-network on the left (beige) takes in human feature vectors, $\mathbf{X_h}$, and the sub-network on the right (gray) takes in mouse feature vectors, $\mathbf{X_m}$. Each feature vector consists of multiple features, denoted as $\mathbf{X}_{hi}$ or $\mathbf{X}_{mi}$, with *i* ranging from 1 to total number of features. A final network (blue) takes in concatenated output vectors from the two sub-networks and generates the final prediction, y. Each layer within a sub-network is followed by a rectified linear unit (ReLU) and dropout is used in the training[70]. We only show a small number of input features, layers, and neurons here. **Supplementary Data 2.3** lists the hyper-parameters that define this architecture.

Chapter 3. Learning a pairwise epigenomic and TF binding association score across the human genome

Abstract

According to maps of chromatin contact, quantitative trait loci, and disease-associated variants, genomic loci are often associated with each other in the context of gene regulation or disease risk. There is a growing collection of epigenomic and TF binding data, which may be useful in understanding such pairwise relationships. We thus develop an approach that quantifies evidence of association for pairs of genomic windows based on large-scale epigenomic and TF binding data. The approach, Learning Evidence of Pairwise Association from Epigenomic and TF binding data (LEPAE), trains for each distance a Siamese neural network with pairs of windows with the distance apart as positives and randomly mismatched pairs of the same windows as negatives. We apply LEPAE to thousands of human datasets and learn the LEPAE score for pairs of windows with up to 100 kb between them. Using chromatin contact and gene annotations, we validate that the score highlights loci with associated or similar properties and may complement existing annotations. We expect LEPAE to be a resource for studying groups of genomic loci.

Introduction

Genomic loci are better understood in the context of other loci than alone. Maps of chromatin interactions from experiments such as Hi-C demonstrate that distal DNA elements can be in close contact in three dimensional space, for example, giving rise to promoter-enhancer interactions. Similarly, quantitative trait loci (QTL) studies show that a variant can influence expression of a distal gene and genome-wide association studies (GWAS) show that multiple variants spread across the genome can jointly contribute to one's risk for disease. Studying groups of variants or DNA elements together can lead to better biological context and also statistical power. When doing so, it can be advantageous to determine whether pairs of genomic

loci of interest exhibit any associated properties. Given the increasingly diverse genome-wide maps of histone modifications and variants, chromatin accessibility, chromatin state annotations, and transcription factor (TF) binding, there is an opportunity to estimate how associated pairs of genomic loci are based on such data.

We thus develop a supervised method, Learning Evidence of Pairwise Association from Epigenomic and TF binding data (LEPAE), that scores evidence of association for pairs of windows across the human genome. LEPAE is a variant of Learning Evidence of Conservation from Integrated Functional genomics data (LECIF)[71] from **Chapter 2**. LEPAE leverages a compendium of epigenomic and TF binding data from various assays and cell types. For each distance, LEPAE uses a Siamese neural network classifier trained to distinguish pairs of windows with the distance apart from randomly mismatched pairs of the same set of windows. As a result, we learn the LEPAE score for every pair of 1-kb windows with pairwise distances ranging from 1 kb to 100 kb. Using external annotations not provided as input, we validate that the learned score reflects associated or similar properties in pairs of loci. We expect the LEPAE score to be a resource for understanding complex relationships among genomic loci in studying gene regulation or disease risk.

Results

Overview of LEPAE

LEPAE quantifies evidence of association between two distinct human genomic loci based on epigenomic and TF binding data. Specifically, given 1-kb nonoverlapping genomic windows, epigenomic and TF binding data annotating those windows, and a specific pairwise distance, a Siamese neural network classifier is trained with pairs of windows in the same chromosome and with the pairwise distance apart as positive pairs and randomly mismatched pairs of the same set of windows as negative pairs (**Methods**). It is assumed that positive pairs represent windows with associated or similar properties given their proximity unlike the negative pairs.

For each pair, we provide two feature vectors corresponding to the pair's two1-kb genomic windows. Each vector contains binary features corresponding to whether a window overlaps with peak calls from DNase-seq experiments and ChIP-seq experiments of histone modifications, histone variants, and TFs, all from a wide variety of human cell and tissue types and generated by ENCODE[1] and Roadmap Epigenomics Project[4]. Because the same set of windows are used in positive and negative pairs, the classifier learns the pairwise characteristics of genomic windows at the specified pairwise distance based on epigenomic and TF binding data. Once trained, the classifier makes predictions for pairs of genomic windows from chromosomes that were held out from training and also at the pairwise distance for which the classifier was trained. This training and prediction procedure is repeated for pairwise distances ranging from 1 kb to 100 kb with increments of 1 kb. As a result, we annotate more than 302 million pairs of 1-kb genomic windows with the LEPAE score (**Fig. 3.1a**).

LEPAE's predictive power and relationship to distance

LEPAE has strong predictive power when differentiating positive pairs from negative pairs, all held out from training, particularly for pairs with small pairwise distances. We observe the strongest predictive power for pairs with the smallest pairwise distance, 1 kb, with a mean area under the receiver operating characteristic curve (AUROC) of 0.91 (**Fig. 3.1b**). For pairs with larger distances, LEPAE has weaker predictive performance as expected, with a minimum mean AUROC of 0.71.

To ensure the LEPAE score is distinct from pairwise distance, we characterize the score with respect to 1D distance (**Fig. 3.1c**). The score is negatively correlated with pairwise distance with a Pearson correlation coefficient (PCC) of -0.15. Pairs above the 99th percentile span up to 100 kb, the maximum pairwise distance, indicating that LEPAE can highlight pairs of windows that are distal but exhibit sufficient evidence of association in their epigenomic and TF binding data.

To validate that LEPAE learns information beyond similarity in input features, we compare the LEPAE score to Jaccard index, which was computed for every pair of windows to which LEPAE was applied (**Methods**). Specifically, for each pairwise distance we compute the correlation between the LEPAE score and Jaccard index for pairs of windows with the pairwise distance. We observe moderate correlation with a mean PCC of 0.13 (**Fig. 3.3a**), indicating that LEPAE may capture information beyond agreement in features. When differentiating positive pairs from negative pairs, LEPAE achieves better predictive power (LEPAE AUROC: 0.71~0.91; Jaccard index AUROC: 0.57~0.76; **Fig. 3.1b**).

Relationship to chromatin contact frequency

We further compare the LEPAE score to chromatin interaction data to understand its relationship to 3D distance. Specifically, we compare to normalized 1-kb resolution chromatin interaction frequencies collected in a Hi-C experiment in GM12878 (**Methods**). As done above with Jaccard index, when we compute correlation between the LEPAE score and normalized Hi-C matrix for pairs with the same pairwise distance, we observe a low mean PCC of 0.05 (**Fig. 3.3a**). We observe similar results when applying logarithmic transformation to Hi-C data with a mean PCC of 0.06. While further validation is needed, this weak agreement suggests that LEPAE may provide distinct information and could potentially complement Hi-C data in studying pairwise relationships.

High-scoring pairs highlight similar or associated loci based on external annotation of chromatin states and genes

We next study the LEPAE score for pairs of DNA elements using external annotations of chromatin states that were not provided as input features. We specifically use a universal annotation of 100 chromatin states learned from integrated datasets from more than 100 human cell types[72]. In general, pairs of states associated with transcription ('TxEx', 'Tx', 'TxEnh') tend to

score high with mean LEPAE score ranging from 0.66 To 0.77 (**Fig. 3.2**). Using external genic annotations and correcting for varying gene lengths, we validate that pairs of windows within genes score higher than those crossing gene boundaries (**Fig 3.3b**; Wilcoxon signed-rank test $P<0.0001$; **Methods**). We further validate that this difference in stronger than the difference observed with Jaccard index or Hi-C data (Mann-Whitney U test $P<0.0001$; **Methods**).

In addition, a state associated with enhancer activity in blood and thymus ('EnhA9') scores high when paired up with transcription-associated states. While the LEPAE score is 0.78 on average when this state is paired up with itself, similarly high scores are observed when the state is paired up with a state associated with transcription and enhancer activity in blood ('TxEnh6') and an active promoter state ('PromF2') with a mean of 0.75 and 0.74, respectively (**Fig. 3.2**). This suggests that LEPAE may highlight not only similar loci but also biologically meaningful relationships such as promoter-enhancer interactions.

Discussion

Here we presented LEPAE, a method that scores evidence for pairwise association between genomic windows based on a large collection of epigenomic and TF binding annotations. LEPAE is a variant of LECIF[71], a comparative genomics approach that scores evidence of conservation between two species based on functional genomics data. Instead of comparing loci from two different species, LEPAE compares loci within the same species using an integrated collection of maps of open chromatin, histone modifications and variants, transcription factor binding, and chromatin state annotations from various tissue and cell types.

We applied LEPAE with more than 3000 annotations from the human genome and learned the LEPAE score. The score had greater predictive power for differentiating pairs of windows at a fixed distance from randomly mismatched pairs of the same windows than a naïve approach of computing the Jaccard index of input features. Using external annotations not provided as input features, we validated that the LEPAE score reflects information beyond epigenomic similarity

and demonstrated evidence that it may capture biologically meaningful associations. Given these results, we expect the score to be useful in various applications such as grouping rare variants in burden tests and prioritizing biologically relevant Hi-C interactions[73] or eQTL pairs.

While here we focused on learning a score with maximum pairwise distance of 100 kb, we plan to use LEPAE to generate scores with larger pairwise distances to interrogate additional long-range relationships. Moreover, although here we applied LEPAE to epigenomic and TF binding data from the human genome, LEPAE can integrate functional genomics data such as RNA-seq data and is applicable to other widely studied species such as mouse or rat. With abundant and diverse genomic data, we expect LEPAE to be useful in leaverging the data to find interesting and relevant pairs of loci to study.

Methods

Defining genomic windows

We segmented all autosomal chromosomes and X chromosome into non-overlapping 1-kb windows. We used hg38 as the genome assembly.

Input features

Each pair of windows was assigned two feature vectors, one corresponding to the upstream window and the other corresponding to the downstream window. For each window, each peak call corresponded to a binary feature. If a genomic window overlapped a peak call in an experiment, the corresponding value in the feature vector was set to 1, otherwise it was set to 0. The state annotations were one-hot encoded such that each binary feature corresponded to the presence of a chromatin state in a cell or tissue type.

Peak calls for DNase-seq and ChIP-seq experiments of histone modifications, histone variants, and TFs were from ENCODE4[1]. ChromHMM chromatin state annotations[29] were from

the 25-state model learned from imputed data for 127 cell and tissue types from the Roadmap Epigenomics Project[4] that were lifted over from hg19.

Defining positive and negative pairs

For each pairwise distance *d*, ranging from 1 kb to 100 kb, pairs of windows with *d* bases between their first bases were defined as positive pairs. To generate negative pairs, we randomly shuffled the pairing of the positive pairs within the same chromosome, resulting in pairs of windows that are not necessarily *d* bases apart from each other but from another window in the same chromosome.

Defining subset of pairs for training, validation, and test

To generate predictions for all pairs of genomic windows from an odd chromosome or X chromosome, we first randomly selected three even chromosomes and defined a random subset (n=5000) of pairs of windows from those three chromosomes as a validation set. We defined a random subset (n=50,000) of pairs of windows from the remaining even chromosomes as a training set. To form a test set, we used a random subset (n=5000) of pairs of windows from odd chromosomes. To generate predictions for all pairs of windows from an even chromosome, we took an analogous approach as above. There was no overlap in genomic regions used for training, validation, and test. X chromosome was excluded from training, validation, and test, but included for prediction and downstream analyses.

Classifier

For each set of training pairs and a pairwise distance, one neural network classifier was trained to generate prediction for held-out pairs with the pairwise distance. The neural network had a Siamese architecture[48] consisting of two identical sub-networks, which share their weights, followed by a final sub-network that combines the output from the two sub-networks to generate

a final prediction. The two input feature vectors were fed into the two sub-networks. All sub-networks had a single hidden layer, resulting in two hidden layers in total. To set the hyper-parameters of a classifier, we conducted a random search, where we generated 5 classifiers, each with different randomly selected combination of hyper-parameters. Each classifier was trained on the same set of training pairs and evaluated on the same set of validation pairs. Hyper-parameters were varied as follows during random search:

- Batch size: 16, 32, 64
- Learning rate: 1e-8, 1e-6, 1e-4
- Dropout rate: 0.1, 0.3, 0.5
- Number of neurons in the initial layer: 32, 64, 128
- Number of neurons in the final layer: 32, 64, 128

A flipped version of each pair where its two windows are flipped, such that the upstream window is downstream and the downstream window is upstream, was also provided along with the original version. This doubled the number of input pairs provided. We identified the best-performing combination of hyper-parameters by maximizing the AUROC on the validation pairs.

With the best-performing combination of hyper-parameters, we then trained a new classifier, which was finally used to generate the score for pairs with the pairwise distance. For each pair, the trained classifier outputted two values, the probability of the pair being classified as a positive pair and the same value but for the flipped pair. The final LEPAE score of a pair was the average of these two values. This procedure of hyper-parameter search, training the final classifier, and prediction was repeated for each set of training pairs and each pairwise distance, resulting in 200 separately trained classifiers.

We used PyTorch (version 0.3.0.post4)[49] for implementation.

Random forest baseline

We trained, applied, and evaluated random forest using the same procedure as explained above, except we used a random forest in place of a neural network. We also did hyper-parameter selection as explained above, but for the following set of set of hyper-parameters unique to random forests:

- Maximum tree depth; 16, 32, 64, 128, 256
- Minimum fraction of samples required to split at an internal node: 0.0005, 0.001, 0.002, 0.005, 0.01
- Minimum fraction of samples required to be at a leaf node: 0.0005, 0.001, 0.002, 0.005, 0.01

Maximum number of features to consider when looking for the best split was set to square root of the total number of features. Bootstrap samples were used when building trees. We used Scikit-learn (version 0.19.1)[50] for implementation.

Filtering pairs

In all analyses, except for reporting predictive performance, all pairs that overlap any assembly gap annotations in either window were excluded since most input features do not map well to these pairs. We downloaded the assembly gap annotation from the UCSC Genome Browser[3].

Computing Jaccard index

For each pair, given its two binary input feature vectors, we defined *A* as the set of features set to 1 in the first feature vector and *B* as the set of feature set to 1 in the other feature vector. The two vector's Jaccard index was defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (3.1)$$

If the numerator was zero, the pair was eliminated from our analysis.

Accessing Hi-C data

We downloaded in situ Hi-C data for GM12878 (experiment 4DNES3JX38V5)[74] from the 4DN Nucleome Data Portal[75] using this link: https://4dn-open-data-public.s3.amazonaws.com/fourfront-webprod/wfoutput/a98ca64a-861a-4a8c-92e9-586af457b1fb/4DNFI1UEG1HD.hic. Within the downloaded file, we specifically used values with square root of vanilla coverage (VC_SQRT) normalization applied. We used software straw[76] to extract the values for the pairs of our interest from the file. If no data was found for a pair in the file, we discarded the pair from our analysis.

Computing weighted mean LEPAE score for pairs of chromatin states

For each state pair, s_u and s_d, its weighted mean LEPAE score was computed as follows. For each pair of windows, w_u and w_d, annotated by states s_u and s_d, respectively, the pair's weight was the product of the fraction of bases in window w_u annotated by state s_u and the fraction of bases in window w_d annotated by state s_d. The pair's LEPAE score was multiplied by this weight. The overall mean of the state pair was the sum of these weighted scores from all applicable pairs of windows divided by the sum of the weights.

Gene analysis

For each protein-coding gene with length l, we set quartile length q to l divided by 4. Given the position of the transcription start site (TSS) and transcription end site (TES) of the gene, s_1 and s_2, we defined three sets of bases for the gene as follows:

- Upstream of TSS: s_1-4q, s_1-3q, s_1-2q, s_1-q
- Within gene: s_1, s_1+q, s_1+2q, s_1+3q, s_2
- Downstream of TES: s_2+q, s_2+2q, s_2+3q, s_2+4q

We defined pairs of these bases where at least one base in the pair is within the gene (e.g. s_1-3q vs. s_1+q). We then compared two pairs that had the same pairwise distance and shared a base within the gene but one pair crossed a gene boundary, TSS or TES, while the other pair did not (e.g. s_1-q vs. s_1+q compared to s_1+q vs. s_1+3q). This procedure allowed us to evaluate whether the LEPAE score, Jaccard index, or Hi-C contact frequency favors pairs of loci within genes over those crossing gene boundaries while correcting for varying gene lengths.

Figures

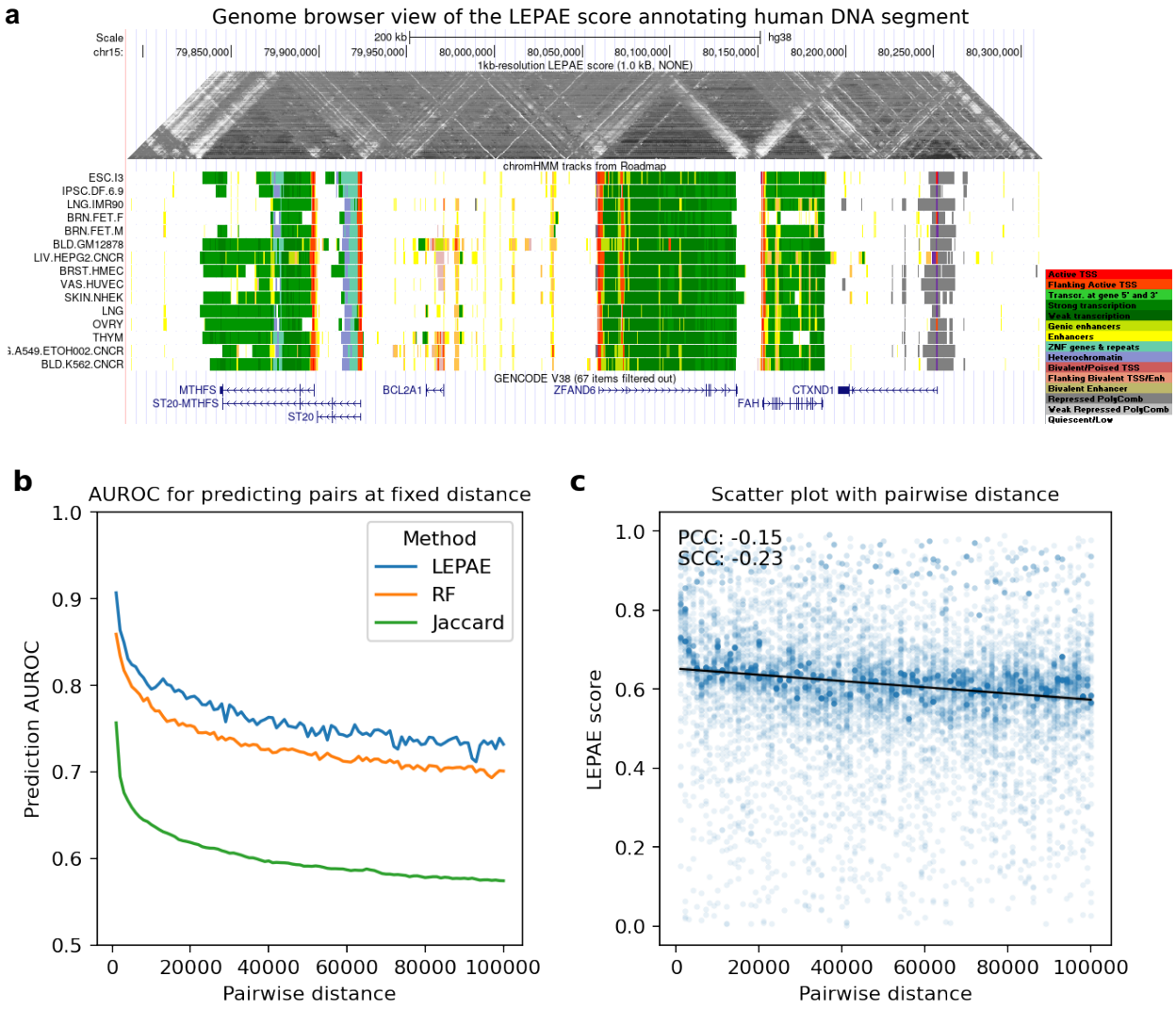


Figure 3.1. Characteristics of the LEPAE score

a. Genome Browser[3] view of the LEPAE score annotating multiple genes and its neighboring regions in human chromosome 15. The score is shown in the top as a heatmap in a format primarily used for visualizing Hi-C contact matrices[76]. Darker color corresponds to higher LEPAE score. Below the score are ChromHMM chromatin state annotations[29] for different epigenomes from the Roadmap Epigenomics Project[4], which were provided as input. While chromatin state annotations from 127 epigenomes were used as input features, here we show a subset of the epigenomes. State legend is on the bottom right. The state annotations are followed by GENCODE V38 gene annotation[77].

b. Relationship between pairwise distance and prediction AUROC. For each pairwise distance (*x*-axis), mean prediction AUROC of the LEPAE score for distinguishing pairs of windows at the distance from randomly mismatched pairs of the same windows is shown in blue. The mean is computed from two classifiers trained on non-overlapping training sets (**Methods**). Mean AUROC values computed for a baseline model where random forest instead of neural network was used as supervised classifier are shown in orange (**Methods**). Mean AUROC values computed using Jaccard index instead of the LEPAE score to perform the same classification task are shown in green. Values belonging to the same method are connected by piecewise linear interpolation.

c. Scatter plot showing with a blue dot for each pair of windows the pairwise distance (*x*-axis) and the LEPAE score (*y*-axis). Ten thousand random pairs are shown. A linear regression line fitted to the ten thousand random pairs is shown in black and its 95% confidence interval shown in grey shaded area. PCC and SCC, computed from 1 million randomly selected pairs, are shown in the top left.

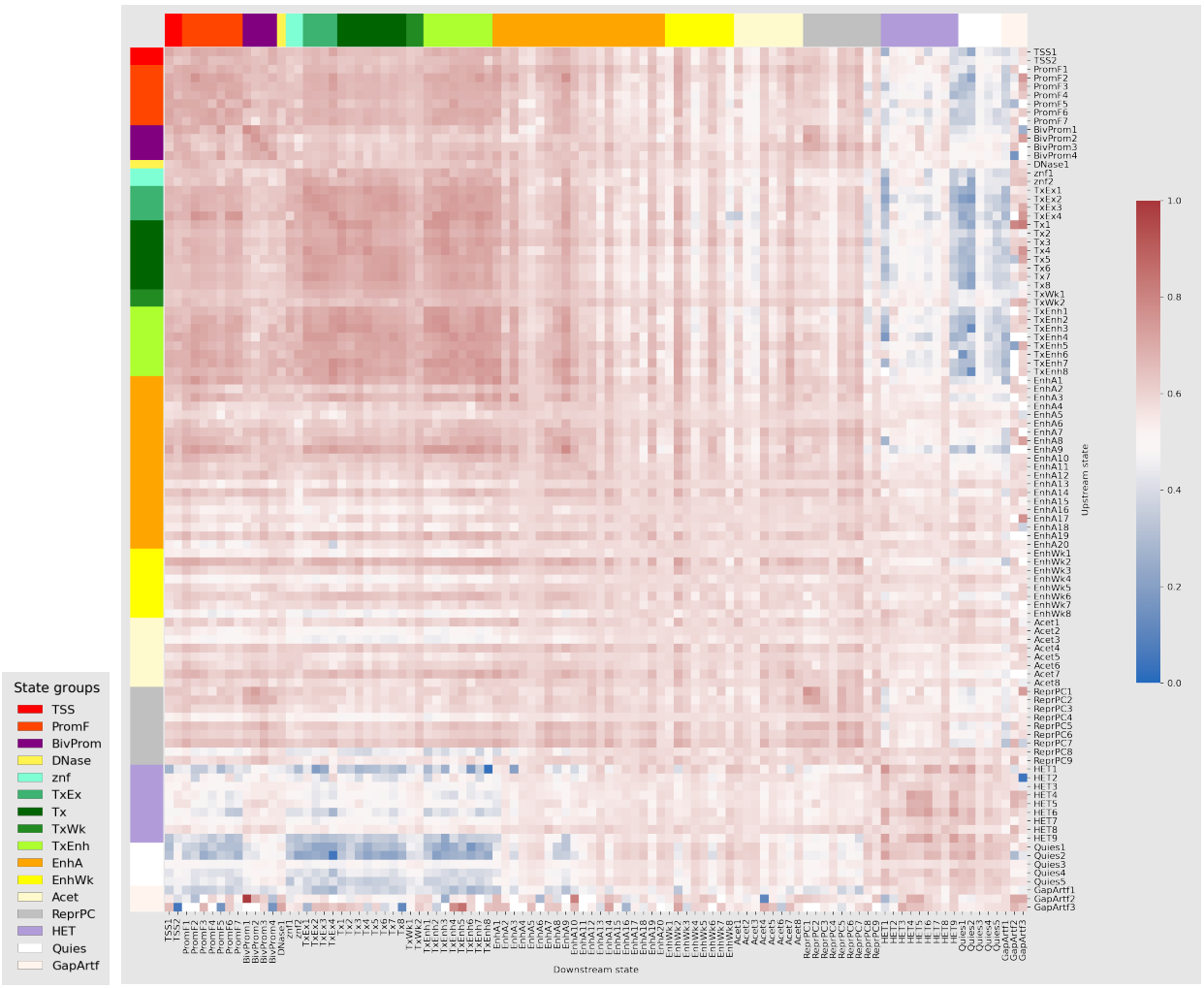


Figure 3.2. Heatmap of mean LEPAE score for pairs of chromatin states
Each cell in the heatmap corresponds to a state pair, one annotating the upstream window of a pair of windows (row) and the other annotating the downstream window of the same pair (column). The states are from a universal chromatin state annotation based on more than 1000 epigenomic datasets from more than 100 cell types[72]. The ordering of states in the rows and columns are the same. Color shown next to the topmost row or leftmost column corresponds to the state group of each state along the column or row, respectively, according to the legend on the bottom left. Color shown in each cell corresponds to the weighted mean LEPAE score of pairs of windows annotated by the states specified in the row and column. For each state pair, its weighted mean LEPAE score was computed such that pairs of windows with more bases annotated by the states contribute more to the mean than windows with fewer bases annotated by the same states (**Methods**). Color legend for the score is shown on the right. One million randomly selected pairs of windows were included in this analysis.

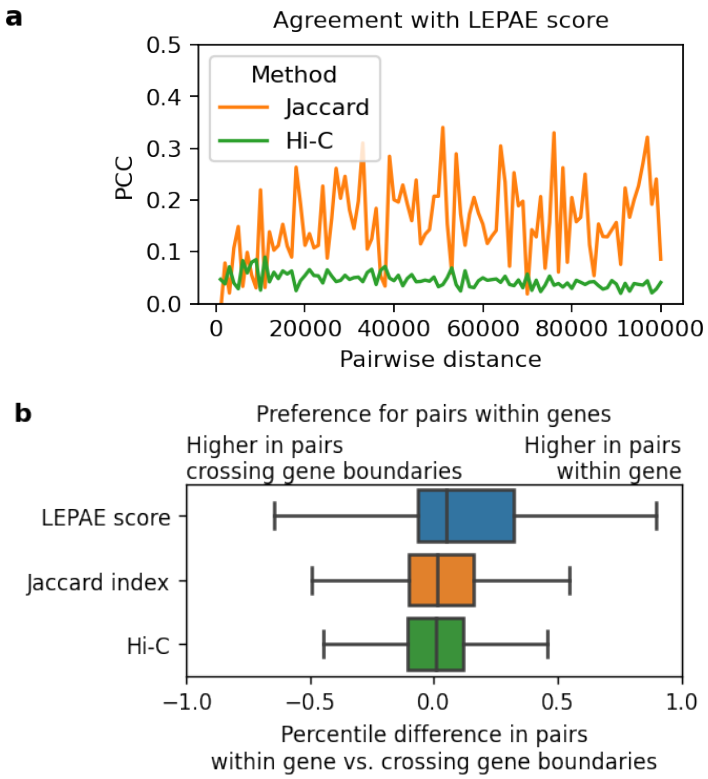


Figure 3.3. LEPAE score's relationship to Hi-C contact frequency and genic annotations.
a. Shown for each pairwise distance (*x*-axis) is PCC of the LEPAE score with either Jaccard index (orange) or normalized Hi-C contact frequency (green) for pairs of windows with the specified distance between them. Values belonging to the same method are connected by piecewise linear interpolation.
b. Shown for the LEPAE score, Jaccard index, or normalized Hi-C contact frequency is the distribution of percentile differences between pairs with both bases within a gene and pairs with only one base within a gene. The difference is computed by taking the value for a pair of bases within a gene and subtracting it by the value for a pair of bases crossing a gene boundary. The two pairs are matched by their pairwise distance and share one base located within a gene (**Methods**). A positive percentile difference indicates a preference for pairs within genes over pairs crossing gene boundaries.

Chapter 4. Single-nucleotide conservation state annotation of the SARS-CoV-2 genome

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 potentially 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 and mutations of SARS-CoV-2, multi-strain sequence alignments of coronaviruses (CoV) have become available[78] where multiple sequences of CoV are aligned against the SARS-CoV-2 reference genome. Sequence 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 strong sequence constraint or accelerated evolution have been shown to be enriched for phenotype-associated variants[41,79]. While existing systematic annotations that quantify sequence constraint from alignments[11,12] are informative, they reduce the information in the underlying alignment to a single univariate or binary value and thus are limited in the information they convey. Additional information about patterns of which sequences align to and match the SARS-CoV-2 genome at each base may be useful in analyzing the SARS-CoV-2 genome and mutations.

As a complementary approach to sequence constraint scoring methods, ConsHMM was recently introduced to systematically annotate a given genome with conservation states that capture combinatorial and spatial patterns in a multi-species sequence alignment[35]. ConsHMM specifically models whether bases from non-reference sequences align to and match each base in the reference genome. 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 the learned patterns[29]. Apart from the input alignments which were generated using phylogenetic trees, ConsHMM does not explicitly use any phylogenetic information and therefore does not make any strict assumptions on the phylogenetic relationship among sequences. This allows ConsHMM to be more flexible in capturing various patterns within alignments than the more commonly used comparative genomics approaches that define a single constraint score or binary calls of constrained elements based on phylogenetic modeling. 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[35,80].

Motivated by the current need to better understand the SARS-CoV-2 genome and mutations, here we apply ConsHMM to two multi-strain sequence alignments of CoV that were recently made available[78] and learn two sets of conservation states. The first alignment consists of Sarbecoviruses, a subgenus under genus Betacoronavirus in the family of Coronavirdae[81]. This alignment consists of SARS-CoV and other Sarbecoviruses that infect bats aligned to the SARS-CoV-2 genome. The second alignment consists of CoV that infect various vertebrates (e.g. human, bat, pangolin, mouse, birds) aligned to the SARS-CoV-2 genome.

Given the two sets of conservation states learned by ConsHMM from these two alignments, we annotate the SARS-CoV-2 genome with the states and analyze the states' relationship to external annotations to understand their properties. We observe that the states

capture distinct patterns in the input alignment data. Using external annotations of genes, regions of interest, and mutations observed among SARS-CoV-2 sequences, we observe that the states also have distinct enrichment patterns for various annotated regions. We generate genome-wide tracks that score each nucleotide based on state depletions and enrichments for observed mutations, which can be used to prioritize bases where mutations are more likely to be consequential. 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 ConsHMM conservation state annotations and tracks of state depletion of mutations are resources for interpreting the SARS-CoV-2 genome and mutations.

Results

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 a Sarbecovirus sequence alignment, labeled as states S1 to S30 (**Figs. 4.1-4.2**; **Supplementary Table 4.1**; **Methods**). The Sarbecovirus alignment consists of SARS-CoV and 42 other Sarbecoviruses that infect bats aligned to the SARS-CoV-2 genome (**Fig. 4.2c**). The states capture distinct patterns of which Sarbecovirus strains align to and match the SARS-CoV-2 genome (**Fig. 4.2a**) and show notable enrichment patterns for external annotations of genes, proteins, and regions of interest within them (**Fig. 4.2b**, **Supplementary Fig. 4.1**). State S17 corresponds to bases where all strains align to and match SARS-CoV-2 with high probability and appears in the genome most frequently, covering 48% of the genome. Similarly, state S18 annotates bases with high align and match probabilities except it has slightly reduced probability of matching two strains that are most distal from SARS-CoV-2 (SARS-related CoV strain BtKY72 and Bat CoV BM48-31/BGR/2008). Unlike state S17, state S18 is strongly enriched for a region in RNA-dependent RNA polymerase (RdRp) that is known to interact with

the antiviral drug remdesivir (10-fold; P<0.0001). State S6 annotates bases where all strains align to SARS-CoV-2 with high probability but only the strain closest to SARS-CoV-2, bat CoV RaTG13, matches SARS-CoV-2 with high probability, highlighting bases with alleles unique to SARS-CoV-2 and bat CoV RaTG13 with respect to other Sarbecoviruses. As expected, state S6 is enriched for the third codon position (2.2-fold; P<0.0001) where derived alleles are less likely to alter the amino acid. In contrast to state S6, state S28 corresponds to bases where bat CoV RaTG13 both aligns to and matches SARS-CoV-2 with high probability but has a low probability of aligning to other Sarbecoviruses. State S28 covers 1% of the genome and highlights bases unique to SARS-CoV-2 and bat CoV RaTG13 with respect to other Sarbecoviruses. Notably, state S28 is highly enriched for human ACE2 binding domain (22-fold; P<0.0001), which is consistent with recent work suggesting that this binding domain is under strong positive selective pressure due to its critical role in host infection[82,83]. State S28 also annotates a region, known as the PRRA motif, that may have been inserted into the SARS-CoV-2 genome, potentially resulting in increased infectiousness[84,85]. We note that state S28 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[86]. A different state, state S13, corresponds to bases where all strains align to the reference with high probability, but only a specific subset of the strains have the same nucleotide as SARS-CoV-2 with high probability (**Fig. 4.2a**). 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. 4.2**). State S29 shows strong enrichment of intergenic bases (36-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[87] and phylogenetic codon modeling[81].

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 56 CoV from vertebrate hosts against SARS-CoV-2 (states V1 to V30; **Fig. 4.3**; **Supplementary Table 4.2**; **Methods**). The vertebrate CoV alignment consisted of a diverse set of CoV that included not only Sarbecoviruses, but also CoV that are evolutionarily more diverged from SARS-CoV-2 than Sarbecoviruses (**Fig. 4.3c**). 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 various strains of vertebrate CoV aligning to and matching SARS-CoV-2 and exhibit notable enrichment patterns for previously annotated regions within genes (**Fig. 4.3a**, **Supplementary Fig. 4.1**). State V27 annotates bases in which all 56 CoV align to and match SARS-CoV-2, with a genome coverage of 8%. State V19 corresponds to bases in which specifically the 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, align to and match SARS-CoV-2 with high probabilities. State V20 has both high align and match probabilities for bat CoV RaTG13 and pangolin CoV and is enriched for the spike protein's receptor binding domain (RBD), where a recombination event between a bat CoV and a pangolin CoV might have occurred[84] (6.9-fold enrichment; $P<0.0001$). Additionally, 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.

Since the input vertebrate CoV alignment includes several CoV infecting human, the states learned from this alignment can be used to investigate the varying pathogenicity among human CoV. State V14 corresponds to bases shared among pathogenic human CoV, including

SARS-CoV-2, SARS-CoV, and Middle East respiratory syndrome-related CoV (MERS-CoV), but not shared among less pathogenic human CoV which are associated with common cold (OC43, HKU1, 229E, and NL63). Bases annotated by this state are candidates for contributing to the shared pathogenicity of SARS-CoV, SARS-CoV-2, and MERS-CoV (**Supplementary Table 4.3**). We compared bases annotated by this state to positions identified in a previous study that located indels differentiating pathogenic CoV from common-cold-associated CoV using an alignment of 944 human CoV sequences under a supervised learning framework[88]. State V14 overlapped with two insertions identified in that study, one of which is in the nucleocapsid protein and was suggested to contribute to the virus's pathogenicity by enhancing its nuclear localization signals[88] (overlapping positions: 29116-29124). Moreover, using state V14 we identify additional loci potentially unique to pathogenic CoV that were not reported in the previous study (**Supplementary Table 4.3**). While this could be explained mostly by the different sequences included in the alignments used here and in the previous study, we find among the additional loci those that are shared among all pathogenic sequences, but missing in all common-cold-associated sequences according to the previous study's human CoV alignment (**Supplementary Table 4.3**; **Methods**). Among such additional loci that are unique to pathogenic sequences, but not previously reported, is an 8-bp region (positions 28416-28423) in the nucleocapsid protein. This protein was shown to enrich for indels specific to pathogenic CoV in the previous study. Overall, this demonstrates the conservation state annotations learned using an unsupervised approach identified additional genomic bases that may contribute to the pathogenicity of CoV infecting humans.

Conservation states' relationship to standard sequence constraint annotations

To establish that conservation states contain additional information relative to standard sequence constraint scores, we compared to constraint scores generated by PhastCons[11] and PhyloP[12] and binary constrained elements called by PhastCons using the same alignments

provided to ConsHMM in their ability to predict genes and regions of interest (**Methods**). When predicting bases overlapping genes or regions of interest within them, in most cases at least one of the conservation states achieves substantially greater precision at the same recall levels than PhastCons and PhyloP annotations (**Supplementary Fig. 4.3**). This suggests that when compared to existing constraint annotations based on the same alignments, ConsHMM conservation states capture additional biologically relevant information. Consistent with this, while some states have distinct distributions of PhastCons and PhyloP scores and fractions of constrained bases, many states have largely overlapping distributions of them (**Supplementary Fig. 4.4**).

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. 4.4a,c**). Specifically, we analyzed the state enrichment patterns for mutations observed at least twice in about four thousand SARS-CoV-2 sequences from GISAID (Global Initiative on Sharing All Influenza Data)[89]. To focus on reliable calls of mutations, we limited our analysis to nonsingleton mutations and masked genomic positions with known technical issues[86] (**Methods**). In the Sarbecovirus model, as expected, states with high probabilities that all strains align to and match SARS-CoV-2 (S17, S18) are significantly depleted of mutations observed in the current pandemic (0.6-0.7-fold enrichment; $P<0.0001$) while several states (S6, S12, S19, S26, S28, S29) are significantly enriched for mutations (1.3-2.4-fold; $P<0.001$).

The vertebrate CoV model's conservation states exhibit additional enrichment patterns for nonsingleton SARS-CoV-2 mutations. The model learns several 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 0.6 observed in the Sarbecovirus model. This is expected as

the vertebrate CoV alignment contains a more diverse set of strains and is thus likely to capture deeper constraint than the Sarbecovirus alignment (**Fig. 4.3c**). Moreover, while the states significantly depleted of mutations in the Sarbecovirus model have high align and match probabilities for all strains (S17, S18), 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; 0.2-fold enrichment; $P<0.0001$), but also several states that have high align and match probabilities for only a specific subset of vertebrate CoV (0.2-0.4-fold; $P<0.0001$; V10, V11). This subset excludes strains in a specific subtree in the phylogeny of CoV, largely consisting of CoV from avian hosts (**Supplementary Fig. 4.5**). 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 significantly enriched for mutations (1.5-1.8-fold; $P<0.0001$; V3, V13, V20, V30). 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 Fig. 4.6**). These patterns are also largely consistent when we control for whether each mutation is intergenic, synonymous, missense, or nonsense, indicating that the observed state enrichment patterns are not simply driven by mutation type (**Supplementary Fig. 4.6**).

To understand the state annotation's relationship to positive selection, we next examined state enrichment patterns for homoplastic mutations (**Fig. 4.4b,d**). Specifically, we examined 198 stringently identified homoplastic mutations from a previous study[90]. These mutations were independently and repeatedly observed in separate SARS-CoV-2 lineages and are therefore more likely to be under positive selection than other mutations. State S6, which annotates bases with high align probability for all Sarbecoviruses, but high match probability specifically for bat CoV RaTG13 only, is enriched for homoplastic mutations (2.3-fold; $P<0.001$). Similarly, state

V13 is significantly enriched for homoplastic mutations (2.7-fold; $P<0.001$), significantly more so than for nonsingleton mutations (1.5-fold; binomial $P<0.05$). This state corresponds to bases that align to and match about a third of the vertebrate CoV, which excludes CoV with avian hosts and others. The state is also enriched for the nucleocapsid protein, particularly its dimerization and RNA-binding regions which are highlighted by UniProt[91] (14-, 16-, and 17-fold, respectively; $P<0.0001$).

Notably, state S17, which has high align and match probabilities for all Sarbecoviruses, is strongly depleted of nonsingleton mutations and homoplastic mutations (0.7- and 0.6- fold enrichment, respectively; $P<0.0001$). Interestingly, specific mutations that were previously suggested to be consequential to SARS-CoV-2 are also in this state. For example, in state S17 is a frequently observed missense mutation (position 14408) in the coding region of RdRp that was previously 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[92]. The D614G mutation in the spike protein that was implicated to disrupt a Sarbecovirus-conserved residue[81] and result in increased infectivity[93] is also annotated by state S17. These occurrences of potentially consequential mutations in a state depleted of mutations are consistent with the notion that the state is experiencing negative selection and new mutations that do occur in the state are more likely to have stronger consequences than mutations introduced elsewhere. This depletion of potentially more consequential mutations is also seen with mutation type annotations, where 4% of all possible synonymous mutations are observed as nonsingleton mutations whereas only 0.3% of all possible nonsense mutations are observed as nonsingletons, reflecting their well-established difference in deleteriousness, though as noted above the conservation states show distinct enrichments for observed mutations even when conditioned on mutation type.

Genome-wide tracks based on state depletion of SARS-CoV-2 mutations

We next generated genome-wide tracks that reflect state depletion of mutations to highlight bases where new mutations are more likely to be consequential (**Fig. 4.4e**). Specifically, for each ConsHMM model, we generated a track that scores each genomic base by its state's statistically significant depletion or enrichment of nonsingleton mutations, reflecting the mutation frequency among bases that likely share a common evolutionary history. To merge distinct information captured by the two ConsHMM models, we also generated an integrated genome-wide track, where given two states from different ConsHMM models annotating a base of interest that are both either depleted or enriched for nonsingleton mutations we annotated the base with the state with stronger depletion or enrichment (**Methods**).

We analyzed these tracks based on state depletion of mutations with respect to experimentally measured mutational effect on RBD from a previous study that conducted a deep mutational scanning of RBD[94]. The study specifically measured changes in RBD expression and binding affinity due to each possible amino acid change within RBD, where a positive value denoted increased expression or affinity and a negative value denoted decreased expression or affinity. We observe that all three tracks based on state depletion of mutations are negatively correlated with measured expression changes caused by single nucleotide mutations (Pearson's *r*: -0.24~-0.18, *P*<0.0001; **Fig. 4.4g**, **Supplementary Fig. 4.7c**), which is consistent with our expectation that mutations at bases depleted of observed mutations in general are likely to be more deleterious than other mutations. Furthermore, we observe significant negative correlation between the track based on the vertebrate CoV state annotations and binding affinity changes (Pearson's *r*: -0.12; *P*<0.0001; **Supplementary Fig. 4.7d**).

We further compared the state-based tracks to four sequence constraint scores that were learned from either alignment provided to ConsHMM using PhastCons[11] or PhyloP[12] (**Methods**). Specifically, we examined the constraint scores' correlation with our tracks based on state depletion of mutations and also with measured mutational effect on RBD expression and binding affinity. The constraint scores are moderately correlated with our state-based

genome-wide tracks (**Fig. 4.4f**; **Supplementary Fig. 4.7a-b**; Pearson's *r*: 0.25-0.63). For the evaluation on measured mutational effect on RBD expression, we see a statistically significant difference with constraint scores, with two out of four constraint scores having statistically significantly weaker correlation than our tracks' correlations with the mutational effect (**Fig. 4.4g**, **Supplementary Fig. 4.7c**; $P<0.004$; **Methods**).

Overall, our genome-wide tracks based on significant depletion of mutations in conservation states show expected agreement with measured mutational effect. This suggests that our genome-wide tracks based on depletion of mutations could help prioritize mutations with strong impact on the virus's protein expression and binding affinity or potentially other functionalities, but we note that this analysis does not provide direct evidence for other parts of the genome or other phenotypes of the virus.

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 evolutionary patterns of bases aligning and matching, enabling one to group genomic bases by states and study each state's functional relevance. Identifying patterns of conservation across different strains can be important potentially for understanding the relative pathogenicity of different coronaviruses and cross-immunity from prior infections[95–97]. It should be noted, however, that ConsHMM does not consider where bases in the reference strain align to in non-reference strains and is therefore not expected to capture large-scale rearrangements.

We showed that certain conservation states are strongly enriched or depleted of nonsingleton SARS-CoV-2 mutations. Based on this information, we generated three genome-wide tracks that can be used to prioritize mutations of potentially greater consequence based on evolutionary information of the Sarbecovirus and vertebrate CoV alignments. We note that these tracks are 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 of conservation state annotations along with these tracks based on state depletion of mutations to be resources for locating bases with distinct evolutionary patterns and analyzing mutations that are currently accumulating among SARS-CoV-2 sequences.

Methods

Sequence alignments

We obtained the 44-way Sarbecovirus sequence alignment from the UCSC Genome Browser[78] (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 the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/multiz119way/) and then removing SARS-CoV-2 sequences from the alignment, except the reference sequence, wuhCor1. This resulted in 56 CoV aligned against the reference. Both sequence alignments were generated by the alignment tool Multiz[98].

External annotations

Mutations found in SARS-CoV-2 sequences were point mutations identified by Nextstrain[99] (accessed on Sept 7, 2020) from sequences available on GISAID[89]. For our analysis, to minimize putative false calls we filtered out mutations if their ancestral alleles did not match the reference genome used by Nextstrain, MN908947.3, such as C>T at a base where T

is the reference allele. All the other annotations, including the annotations of genes, codons, and UniProt protein products and regions of interest, were accessed through the UCSC Genome Browser (accessed on Sept 7, 2020)[78].

Learning ConsHMM conservation states and choice of number of 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 be relatively easy to interpret, 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. 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 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 0.5% of the genome.

PhastCons and PhyloP scores

We obtained the 44-way PhastCons and PhyloP scores learned from the Sarbecovirus sequence alignment from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/). We additionally used the PHAST

software[100] 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 the 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.

Masking bases

For all but one downstream analysis, we masked problematic genomic positions listed in the UCSC Genome Browser track 'Problematic Sites' (accessed on Sept 7, 2020) as they are likely affected by sequencing errors, low coverage, contamination, homoplasy, or hypermutability[86,101,102]. As a result, we masked 228 bases, analyzing 29,675 out of 29,903 bases (99.2%). The one exception was when we computed state enrichment for homoplastic mutations from a prior study[90]. For this analysis only, we masked all problematic positions except for those described as homoplastic or highly homoplastic. As a result we masked 175 bases instead of 228 bases, analyzing 29,728 bases (99.4%).

Fold enrichments 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 by the external annotations. To compute the fold enrichment for each external annotation and each state, we divided the fraction of the state's bases in the external annotation out of all bases in the state by the fraction of bases in the external annotation genome-wide. 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. Thus, to compute fold enrichment for mutations in an external annotation for each state, we divided the fraction of observed mutations in the external annotation among possible mutations occurring at bases in the state by the fraction of observed mutations in the external annotation out of all possible mutations genome-wide. We defined nonsingleton mutations as mutations observed in at least two SARS-CoV-2 sequences. For homoplastic SARS-CoV-2 mutations, we used all 198 mutations reported in a prior study[90]. 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.

Correction of state enrichments for SARS-CoV-2 mutations by nucleotide composition or mutation type

To show that the conservation state fold enrichment values for nonsingleton mutations are not simply driven by nucleotide composition or mutation type (i.e. intergenic, synonymous, missense, nonsense), we corrected state enrichment values by nucleotide composition or mutation type as follows. To control for nucleotide composition, for each nucleotide *i*, we first computed the genome-wide fraction f_i of observed nonsingleton mutations out of all possible mutations with nucleotide *i* as the reference base. Then for each state and for each nucleotide *i*, we multiplied the genome-wide fraction f_i and the number of possible mutations in the state with nucleotide *i* as the reference base. For each state, we summed up these values across the nucleotides to obtain the expected number of nonsingleton mutations based on nucleotide composition. Finally, the enrichment corrected by nucleotide composition for each state was computed as the ratio of actual and expected number of observed nonsingleton mutations.

Similarly, to control for mutation type, for each type *j* we computed the genome-wide fraction f_j of observed nonsingleton mutations out of all possible mutations belonging to mutation

type *j*. Then for each state and for each mutation type *j*, we multiplied the genome-wide fraction f_j with the number of possible mutations in the state belonging to mutation type *j*. We then followed the same procedure as above.

Identifying bases unique to pathogenic human CoV and missing in less pathogenic human CoV

We first identified bases annotated by state V14, which corresponds to high align probability for pathogenic human CoV (SARS-CoV, MERS-CoV) and low align probability for less pathogenic human CoV (OC43, HKU1, 229E, and NL63) in the vertebrate CoV sequence alignment. Among these bases, we then identified bases that appeared among all pathogenic human CoV but missing in all less pathogenic human CoV in an alignment of 944 human CoV sequences generated by a prior study. All the 944 sequences come from the seven human CoV including SARS-CoV-2[88].

Precision-recall analysis for recovery of annotated genes and regions of interest

For each NCBI gene[65] or UniProt region of interest[91], we predicted bases in each state from both models to be in the gene or region and computed precision and recall, resulting in 60 pairs of precision and recall values. Similarly, we predicted all bases annotated as a PhastCons element[11] to be in each gene or region and computed precision and recall. With PhastCons and PhyloP scores[12], we computed precision-recall curve for predicting the bases in each gene or region using each score.

Generating browser tracks of depletion of nonsingleton SARS-CoV-2 mutations

Based on the procedure of computing state enrichment of SARS-CoV-2 mutations, for each ConsHMM model, we selected states that exhibited statistically significant enrichment or depletion of nonsingleton mutations at a binomial test p-value threshold of 0.05 after Bonferroni correction. To generate a track for each ConsHMM model, we scored each base overlapping

any of the selected states in the model 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. Bases not annotated by any of the selected states were assigned a score of 0.

We generated an integrated track of mutation depletion in states from both ConsHMM models as follows. If a base was annotated with two states with statistically significant enrichment or depletion of nonsingleton mutations, 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 state 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 a score of 0. If a base was annotated by one state with statistically significant enrichment or depletion of nonsingleton mutations, we annotated the base with the $-\log_2(v)$ value from that state. Bases not annotated by any of the selected states were assigned a score of 0.

Comparing correlation to mutational effect on RBD expression and binding affinity

For each of the three aforementioned genome-wide tracks based on state depletion of mutations, we computed its Pearson's *r* with mutational effect on RBD expression measured by a previous study[94]. For each of the four sequence constraint scores, we also computed its correlation with mutational effect on RBD expression and then compared it to the correlations computed using our genome-wide tracks, using Zou's confidence interval test[103] implemented in the R package cocor[104]. The four sequence constraint scores included PhyloP and PhastCons scores learned from either the Sarbecovirus or vertebrate CoV alignment. When reporting the significance of correlations, we applied a Bonferroni correction by setting the significance threshold to 0.05 divided by 7, the total number of computed correlations. When comparing correlations using Zou's confidence interval test, we compared a state-based track's correlation

to a constraint score's correlation if at least one of the two correlations was negative and statistically significant and applied a Bonferroni correction by setting the confidence level to 1 – 0.05 / *n* where *n* is the total number of pairwise comparisons, which was at most 12. The same procedure was applied to compute correlations with measured mutational effect on RBD binding affinity.

Statistics and Reproducibility

All statistical tests performed are described in detail above. In general, Bonferroni correction was applied and a threshold of 0.05 was used to discern statistical significance.

Data availability

ConsHMM conservation state annotation based on the Sarbecovirus and vertebrate CoV alignments are available at https://github.com/ernstlab/ConsHMM_CoV/. Track annotations of depletion of mutations observed in conservation states from both Sarbecovirus and vertebrate CoV ConsHMM models or each model are available from the same URL. All annotations are also included in **Supplementary Data 4.1**. Source data for **Fig. 4.2a-b, 4.3a-b, 4.4a-d,f-g**, and **Supplementary Fig. 4.7** are provided in **Supplementary Data 4.2**.

Code availability

We used ConsHMM v1.1 obtained from https://github.com/ernstlab/ConsHMM_CoV/.

Figures

Input to ConsHMM:
Multiple CoV sequences aligned to the SARS-CoV-2 genome

Output of ConsHMM:
Conservation states annotation at a basepair resolution

Figure 4.1. Genome browser view of ConsHMM input and output for a portion of the SARS-CoV-2 genome. Shown is an example portion of the Sarbecovirus sequence alignment input to ConsHMM and ConsHMM's conservation state annotation of the SARS-CoV-2 genome as viewed in the UCSC Genome Browser[78]. 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 at a position if the row's sequence has no base that aligns to the reference base at the position 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 to and matching the reference with high probabilities.

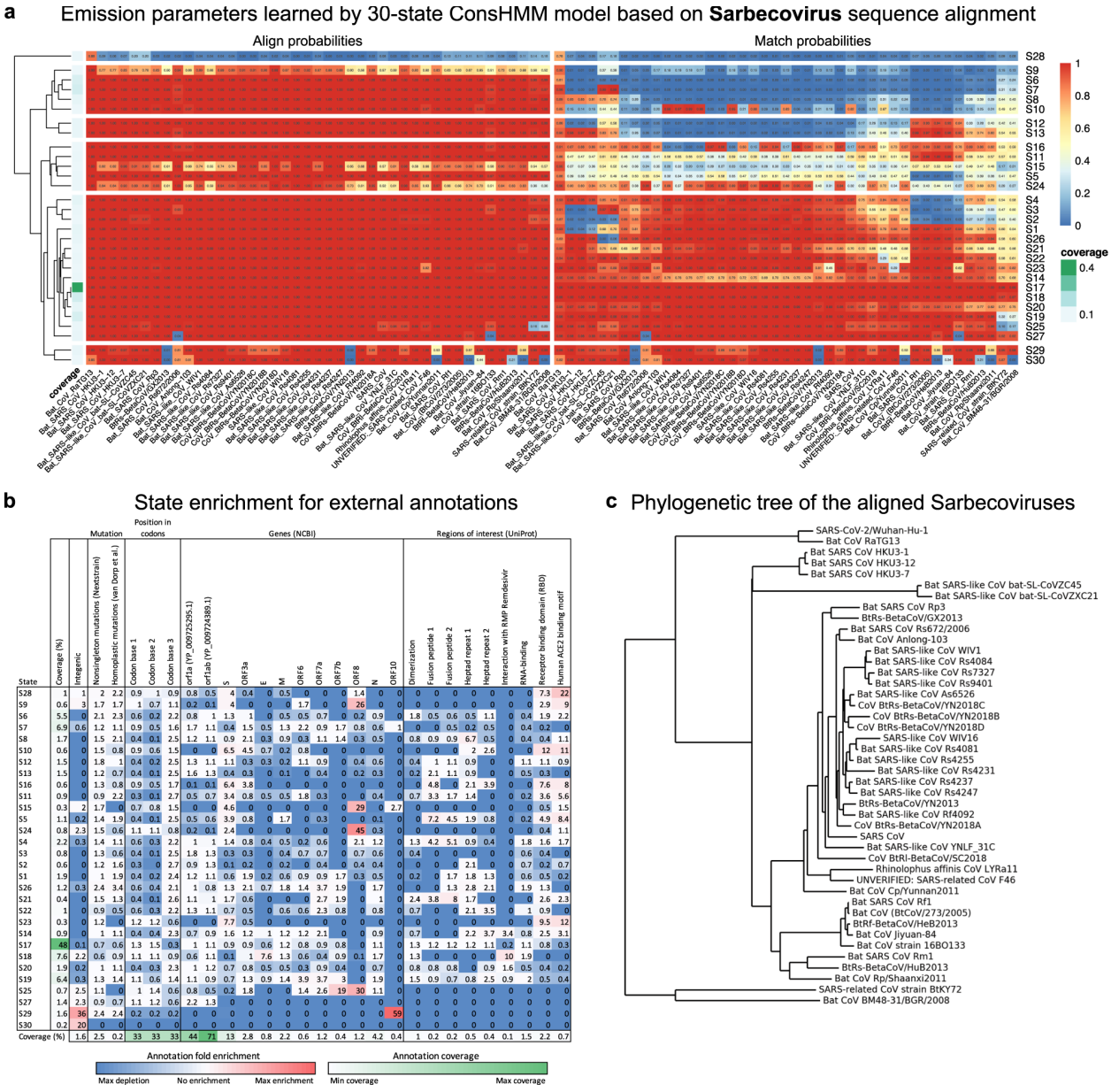


Figure 4.2. ConsHMM conservation states learned from the Sarbecovirus alignment.
a. State emission parameters learned by ConsHMM. The left half of the heatmap shows for each state the probability of each CoV strain having a base aligning to a base in the reference, which is SARS-CoV-2. The right half shows for each state the probability of each CoV strain having a base aligning to and matching (having the same nucleotide) a base in 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[67]. 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 divergence from SARS-CoV-2 according to the phylogenetic tree shown in **c**, 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 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, mutations, position within codons, NCBI gene annotations[65], and UniProt regions of interest[91]. 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 external 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 the minimum and green as the maximum. All annotations were accessed through the UCSC Genome Browser[78] except for nonsingleton mutations from Nextstrain[99] and homoplastic mutations from a prior study[90].

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[78] and plotted using Biopython[105]. SARS-CoV-2/Wuhan-Hu-1, the reference genome of the alignment, is at the top.

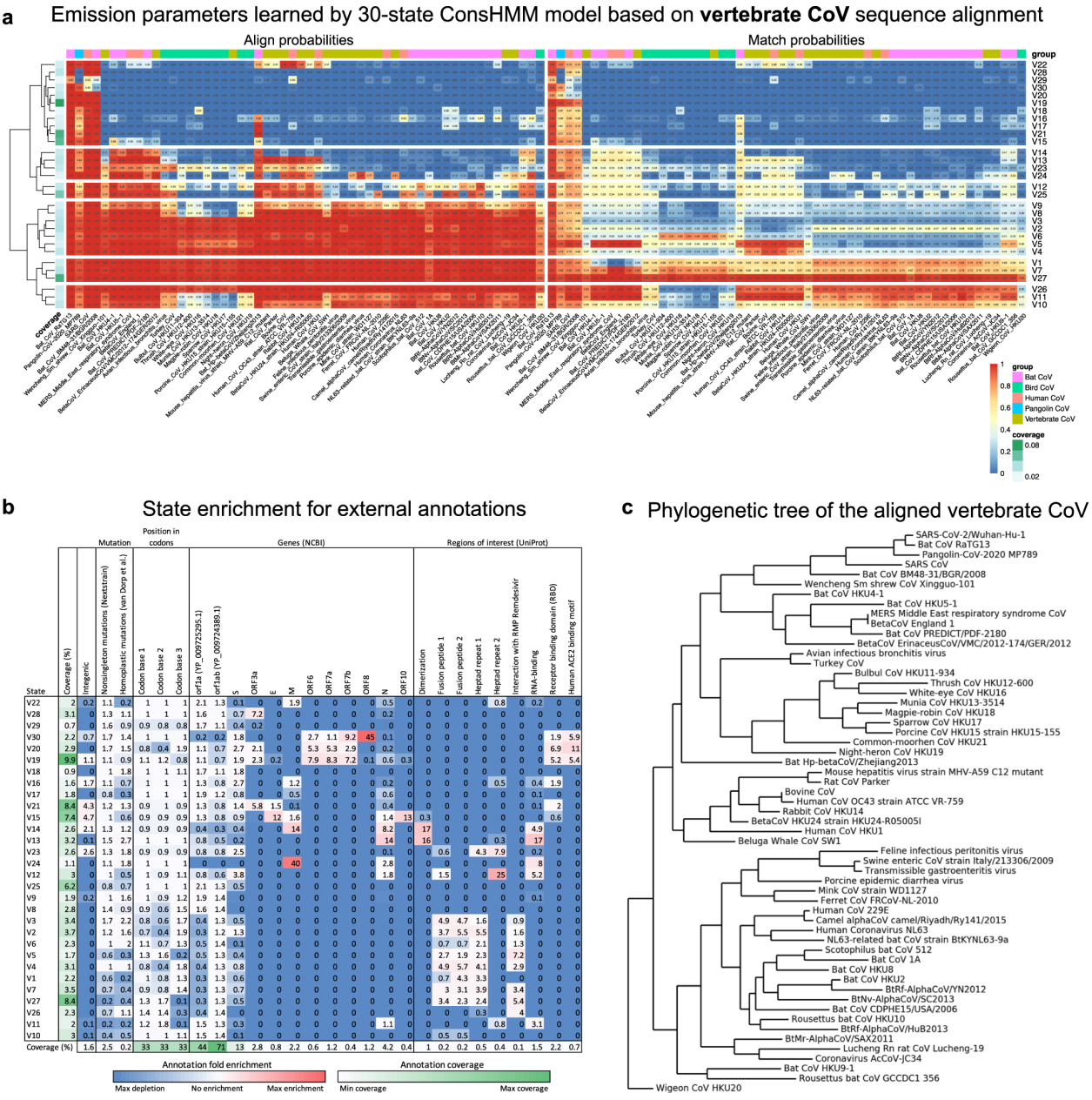


Figure 4.3. ConsHMM conservation states learned from the vertebrate CoV alignment. **a.** State emission parameters learned by ConsHMM. The left half of the heatmap shows for each state the probability of each CoV strain having a base aligning to a base in the reference, which is SARS-CoV-2. The right half shows for each state the probability of each CoV strain having a base aligning to and matching (having the same nucleotide) a base in 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[67]. In both halves, each column corresponds to one of the 56 CoV that infect vertebrates, excluding SARS-CoV-2. Columns are ordered based on each strain's phylogenetic divergence from SARS-CoV-2 according to the phylogenetic tree shown in **c**, with closer strains on the left. Cells in the top row above the heatmap are colored according to the color legend on the bottom right to highlight specific groups of 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 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, mutations, position within codons, NCBI gene annotations[65], and UniProt regions of interest[91]. 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 external 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 the minimum and green as the maximum. All annotations were accessed through the UCSC Genome Browser[78] except for nonsingleton mutations from Nextstrain[99] and homoplastic mutations from a prior study[90].
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[78] (**Methods**) and was plotted using Biopython[105]. SARS-CoV-2/Wuhan-Hu-1, the reference genome of the alignment, is at the top.

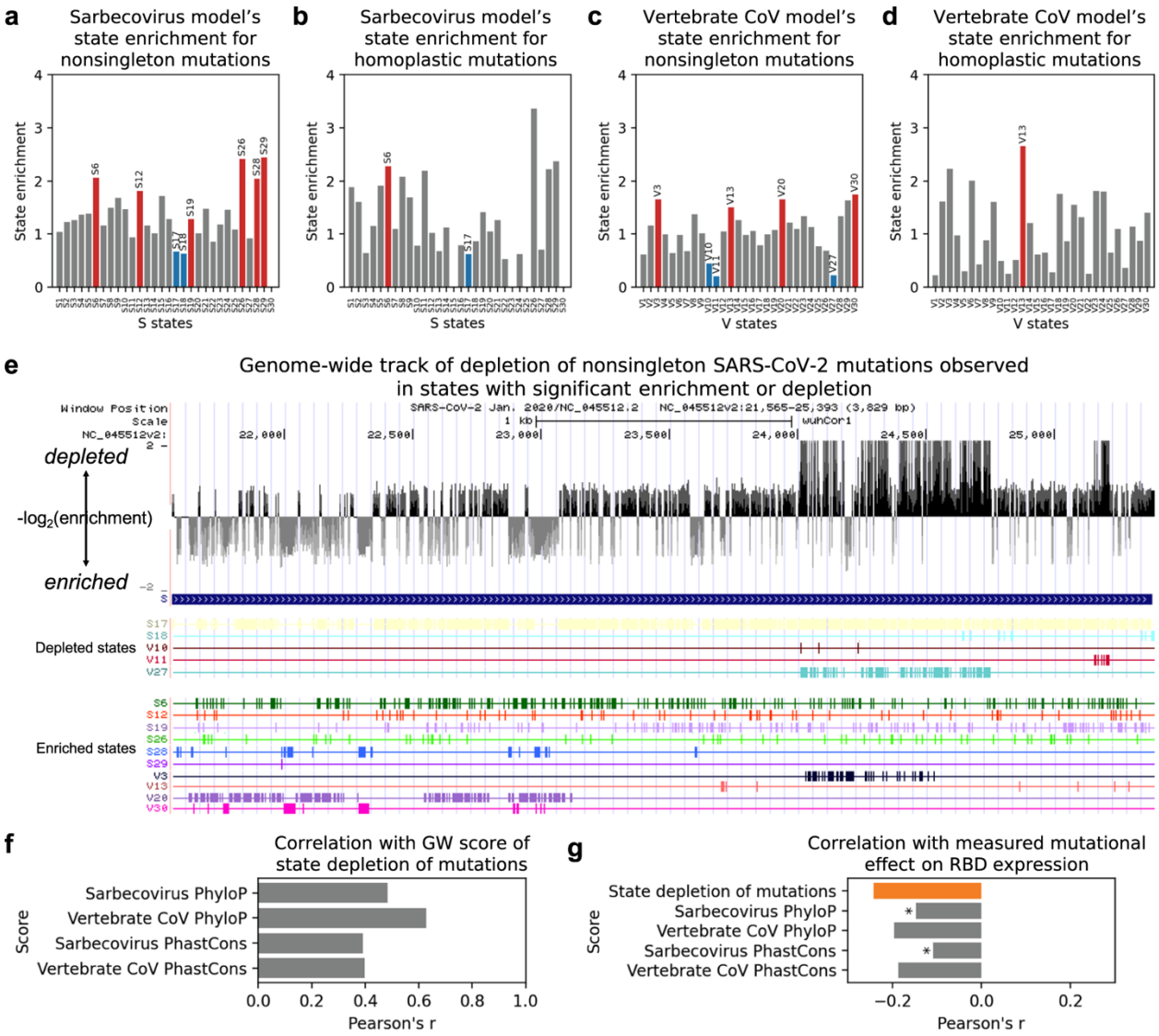


Figure 4.4. State enrichment patterns for nonsingleton mutations in the current pandemic and their relation to other annotations.

a. Bar graph showing enrichment values of states S1-S30 learned from the Sarbecovirus sequence alignment for nonsingleton mutations (n=2,201; **Methods**). Red and blue bars correspond to states that enriched and depleted, respectively, with statistical significance after Bonferroni correction (**Methods**). Above each red or blue bar is the state ID. Grey bars correspond to states for which the enrichment was not statistically significant. Nonsingleton mutations were identified from Nextstrain mutations[99].

b. Similar to **a** but showing state enrichment values for homoplastic mutations (n=198) instead of nonsingleton mutations in states S1-S30. Homoplastic mutations are mutations independently and repeatedly observed in separate SARS-CoV-2 lineages and were previously stringently identified through maximum parsimony tree reconstruction and homoplasy screen using thousands of SARS-CoV-2 sequences[90].

c. Similar to **a** but showing state enrichment values of states V1-V30 learned from the vertebrate CoV sequence alignment instead of states S1-S30.

d. Similar to **b** but showing state enrichment values of states V1-V30 learned from the vertebrate CoV sequence alignment instead of states S1-S30.

e. Genome browser view of gene *S* with an integrated score of depletion of nonsingleton mutations in conservation states derived from both ConsHMM models and annotations of states from which the score is generated. Top row with black and grey vertical bars corresponds to the score, which is a negative $\log_2$ of the fold enrichment value of a state selected from one of the 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.
f. Bar graph showing correlation between our genome-wide (GW) score of state depletion of mutations shown in **e** and four sequence constraint scores listed along the y-axis. The sequence constraint scores were based on either the Sarbecovirus or vertebrate CoV sequence alignment provided to ConsHMM using either PhastCons or PhyloP as the scoring method (**Methods**). Similar plots using scores of mutation depletion in states from each ConsHMM model separately instead of both models together are shown in **Supplementary Fig. 4.7a-b**.
g. Bar graph showing correlation between measured mutational effect on RBD expression and five scores which include our genome-wide score based on state depletion of mutations and the four sequence constraint scores from **f**. Correlation computed with our state-based score is shown in orange. Correlations computed with sequence constraint scores are shown in grey. All correlations were statistically significant after Bonferroni correction (**Methods**). Asterisk is shown next to a grey bar if its corresponding correlation was statistically significantly different than the correlation with our state-based score based on Zou's confidence interval test[103] with Bonferroni correction (**Methods**)**.** The null hypothesis is rejected if the confidence interval (99.6% after correction) of a difference between two correlations excludes 0. The confidence intervals corresponding to the top and bottom asterisks are (-0.18, -0.01) and (-0.22, -0.05), respectively. Mutational effect on RBD expression was measured by a study that conducted a deep mutational scanning of 3,819 nonsynonymous mutations in RBD[94]. To compute the correlations, we restricted to the 1,215 mutations that were caused by single nucleotides and free of experimental measurements that were not determined (n.d.). A positive value indicates increased expression due to mutation and a negative value indicates decreased expression. An extended version of this plot that includes two genome-wide scores based on mutation depletion in states from each ConsHMM model separately is shown in **Supplementary Fig. 4.7c**.

a

Enrichment for protein products in states learned from the Sarbecovirus alignment

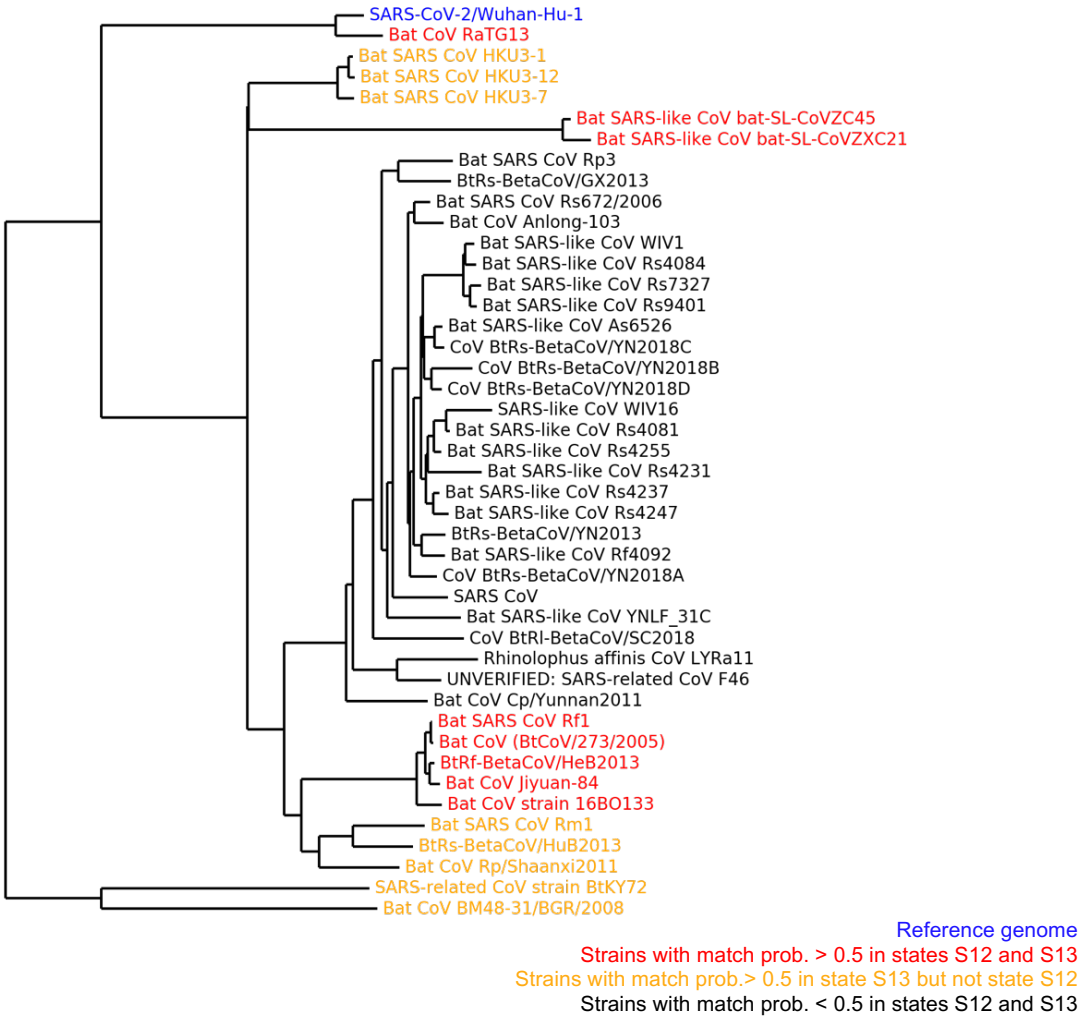
State	nsp1	pp1a	pp1ab	nsp2	nsp3	nsp4	3CL-PRO	nsp6	nsp7	nsp8	nsp9	nsp10	Pol	nsp11	Hel	ExoN	nsp15	nsp16	S glycoprotein	Spike protein S1	Spike protein S2	Spike protein S2'	ORF3a	E protein	M protein	ns6	ORF7a	ns7b	ns8	NC	ORF9b	ORF14
S1	0.6	1.2	1.1	1	1.4	1	0.9	0.8	0.2	0.3	0.9	0.8	1.3	0	1	0.9	1	1	0.6	0.4	0.9	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	1.4	0.4	0	0.5	0.5	0.8	2.1	0	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	1.9	1.3	0.1	0.1	0.4	0	0.1	0	0.1	0.2	1.1	0.8	0.3	0.2	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	0.7	0.5	0.3	0.1	0.5	1.4	0.7	0	0.9	0.8	1.4	1.3	1.4	1.3	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	0.4	0.3	0	0.1	1.1	0.2	0.5	0	0.7	0.5	0.9	2.1	3.9	5.2	2.4	2	0.8	0	1.7	0	0.3	0	0	0.1	0	0
S6	0.5	0.8	1	0.6	0.8	0.6	0.5	0.7	0.5	0.4	1.7	1.3	1.4	2.9	1.1	1.9	1.9	0.6	1.3	1.6	1	0.9	1	0	0.5	0.5	0.7	0	0.3	0.9	0.6	0.3
S7	1.6	1.7	1.1	2.7	2	2	1.1	1.5	0	0	0	0	0	0	0	0	0.2	0.8	0.4	0.3	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	1.1	1	0.4	0.1	1.4	0.3	0.6	0	1	1	0.8	0.8	1	0.8	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	3.8	7.2	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	6.6	9.3	3.4	2.3	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	0.1	0.3	0.2	0	0.6	0.5	0.4	0	1.1	1.3	1.4	1.3	3.4	4.4	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	1	1.1	0.5	0.2	0.4	0.7	0.8	0	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	1.3	1.7	0.2	0.3	1.2	1.3	0.7	0	1.1	0.8	1	0.7	0.4	0.2	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	1.2	0.7	0.3	0.4	0.8	0.3	0.1	0.3	0.5	0.9	0	1	1.2	2	2.5	1.5	1.8	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	4.4	8.2	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	6.5	7.3	5.5	3.8	3.8	0	0	0	0	0	0	0	0	0
S17	1	1	1.1	0.9	0.9	1.2	1.2	1.2	0.3	0.3	1	1.1	1	0.4	1.3	1.2	1.2	1.3	0.9	0.7	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	1.1	0.9	0.3	0.8	3.5	3.1	3.4	7.6	1.1	1.1	0.7	0.3	0.1	0	0.2	0.3	0.9	7.5	1.3	0.4	0.2	0.9	0	1.7	2.2	1.9
S19	1.7	1.1	0.9	1.5	1.1	0.9	0.6	0.9	0.1	0.1	0.5	0.4	0.7	0.8	1	0.7	0.9	0.6	0.7	0.5	1	0.9	1.3	0.9	1.4	3.9	3.8	3.1	0	1.9	2	2.6
S20	1.2	1	1.2	1	1	0.9	0.8	1	0.5	0.3	0.3	1.4	1.5	0	1.3	1.8	1	1.2	0.7	0.5	0.8	1	0.8	0.5	0.6	0.3	1	0.4	0	0.7	0.7	0.2
S21	2.3	1.1	1	1.4	0.9	2	0.3	0.3	0	0.3	0	1.8	0.3	6.6	1.4	0.6	0.7	1.4	1.7	2.4	1	1.3	0.6	0	0	0	2.4	2	0	1	0.9	0
S22	1.9	1.3	1.1	2.2	1.1	1.3	1.5	0.5	0.4	0.4	0	1.3	0.5	0	1.2	0.9	1.2	0.8	0.7	0.5	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	7.3	13	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	2.3	4.3	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	0.2	0.3	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	1.6	0.8	0.1	0	0.7	0	0.2	0	0.6	0.2	0.8	1.7	1.3	1	1.6	1.6	2.1	0.7	1.8	1.4	3.2	2	0	1.7	2.9	1.5
S27	0	2.2	1.3	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	3.7	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	0.1	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	19	5	3.1	2.9	3.3	7.9	1.1	1.4	9.3	0.1	6	5.3	3.5	3	13	6.7	5.9	4.6	2.8	0.8	2.2	0.6	1.1	0.4	1.1	4.2	1	0.7

b

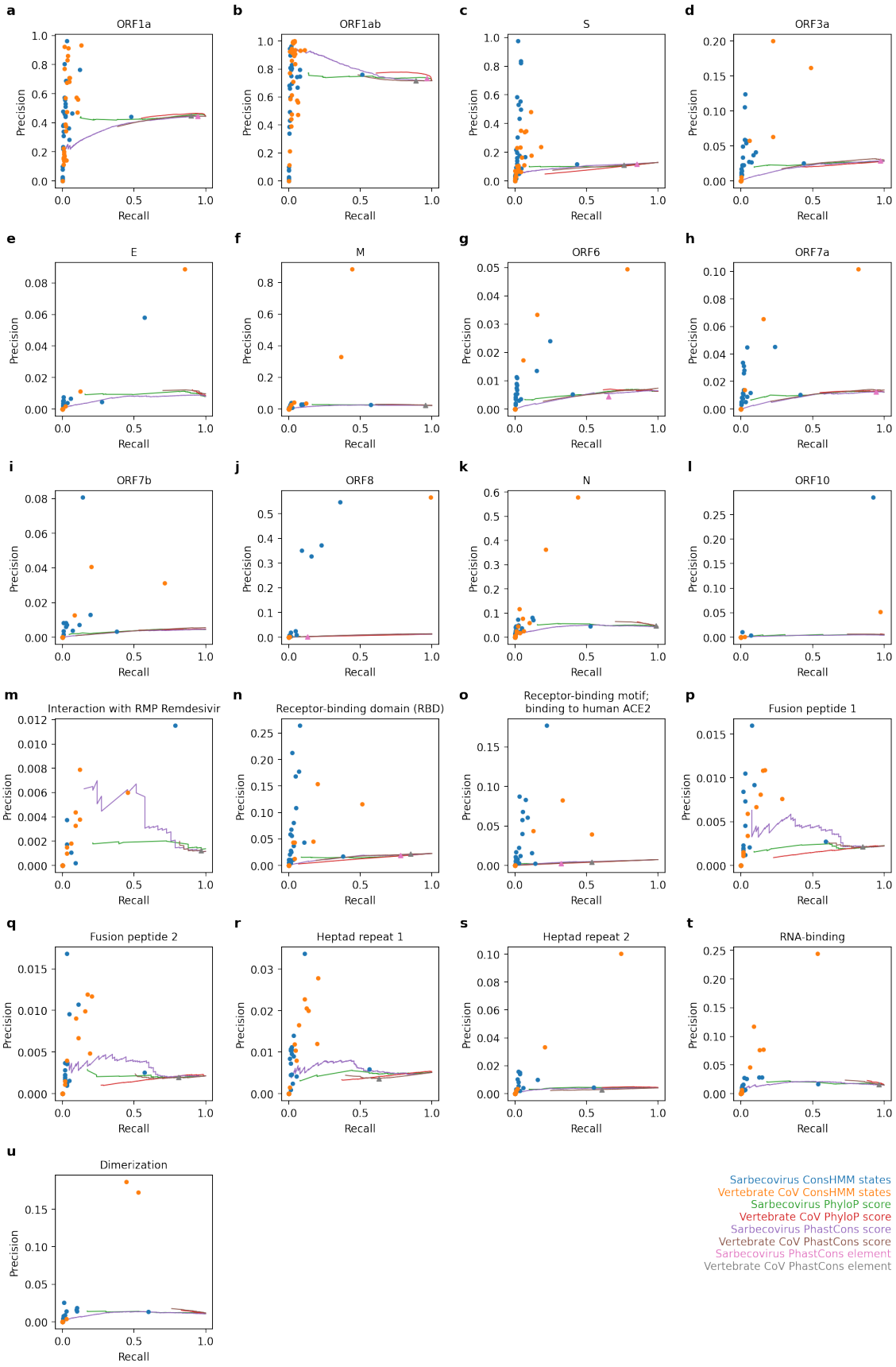
Enrichment for protein products in states learned from the vertebrate CoV alignment

State	nsp1	pp1a	pp1ab	nsp2	nsp3	nsp4	3CL-PRO	nsp6	nsp7	nsp8	nsp9	nsp10	Pol	nsp11	Hel	ExoN	nsp15	nsp16	S glycoprotein	Spike protein S1	Spike protein S2	Spike protein S2'	ORF3a	E protein	M protein	ns6	ORF7a	ns7b	ns8	NC	ORF9b	ORF14
V1	0	0.3	1.3	0	0	0	1.9	0.6	0	0.2	0.4	3.6	2.9	0	3.1	3.1	2.6	2.4	0.6	0	1.2	1.5	0	0	0	0	0	0	0	0	0	0
V2	0	0.3	1.2	0	0	0	1.7	0.6	0.1	0.1	0.7	3.2	2.7	0	2.8	2.3	2.7	1.9	1.3	0	2.8	3.6	0	0	0	0	0	0	0	0	0	0
V3	0	0.4	1.3	0	0	0	2	1.2	0	0.1	0.9	3.6	3.3	1.6	2.9	2.8	2.3	2.1	0.5	0	1.1	1.4	0	0	0	0	0	0	0	0	0	0
V4	0	0.4	1.3	0	0	0	1.7	1.3	0	0.2	0.9	3.3	2.8	1.7	3	2.8	2.6	2	0.8	0	1.8	2.3	0	0	0	0	0	0	0	0	0	0
V5	0	0.5	1.3	0	0	0.8	2	0.9	0.5	0.5	0.5	1.9	3	1.6	3	2.1	2.6	2.2	0.4	0	0.9	1.2	0	0	0	0	0	0	0	0	0	0
V6	0	0.5	1.4	0	0	0	3.4	0.4	0.1	0.3	1.2	3.2	2.6	0	3.2	2.8	4.2	1.7	0.1	0	0.3	0.4	0	0	0	0	0	0	0	0	0	0
V7	0	0.3	1.3	0	0	0	2	0.7	0.1	0.3	0.8	2.8	3	0.7	2.8	2.6	2.8	2.5	0.7	0	1.5	2	0	0	0	0	0	0	0	0	0	0
V8	0	1.6	1.4	0	0.3	1.4	2.9	3.7	2.7	2.6	7.2	0.2	0.6	0	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	4.7	15	0.7	0.3	4.9	0.3	0.3	11	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	3.8	4.2	1.9	2	8	0.1	0.5	1.7	0.6	1.5	1.4	3.2	0.1	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0
V11	0	1.5	1.3	0	1	1.6	3.1	1.6	1.8	1.5	4.9	0.6	0.4	2.6	0.2	1.3	0.7	3.6	0.3	0	0.6	0.7	0	0	0	0	0	0	0	1.1	0	0
V12	0	0.8	0.6	0	1.6	0.1	0.1	1.2	0.3	0.1	0.9	0	0.1	0	0.1	0.1	1.3	0	3.8	0	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	0.4	0.6	0	0	0.1	0	0.2	4.9	0.1	0.3	0.3	0.3	0.2	0	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	0.2	0.1	0	0	0.1	0.3	0	0	0	0.2	0.4	0	0.4	0	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	0.9	1.4	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	0.1	0.5	0	0	0	0	0	0	0.1	0	0.4	0	2.2	2.8	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	0.8	1.5	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	1.8	3.4	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	0.5	0	0	0	0	1.9	3.5	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	2.8	5.2	0	0	2.1	0	0	5.4	5.1	3	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	1.4	2.2	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	0.1	0	0.3	0.1	0	0	1.9	0	0	0	0	0.5	0	0.7
V23	0	0.8	0.8	0	0.1	0.4	1.7	6.3	1	0.3	2.8	1	0.3	0	0.6	1.1	2.5	0.1	2.6	0	5.5	3.1	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	0.5	0	1.1	0	0	0	0	0	0	0	0	0	0	0
V26	0	0.4	1.4	0	0	0	1.8	0.7	0.2	0.3	1	3	3.1	1.1	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	1.6	0.6	0.1	0.2	0.8	2.9	3.4	3.5	3.4	2.6	2	2.2	0.5	0	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	0.7	1.2	0.1	0.1	7.2	0	0	0	0	0	0	0.2	0	0
V29	0	1.7	1.1	2.8	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4	0.8	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	1.8	3.4	0	0	0	0	0	2.7	1.3	8.3	45	0.1	0.5	0
Coverage	1.8	44	71	6.4	19	5	3.1	2.9	3.3	7.9	1.1	1.4	9.3	0.1	6	5.3	3.5	3	13	6.7	5.9	4.6	2.8	0.8	2.2	0.6	1.1	0.4	1.1	4.2	1	0.7

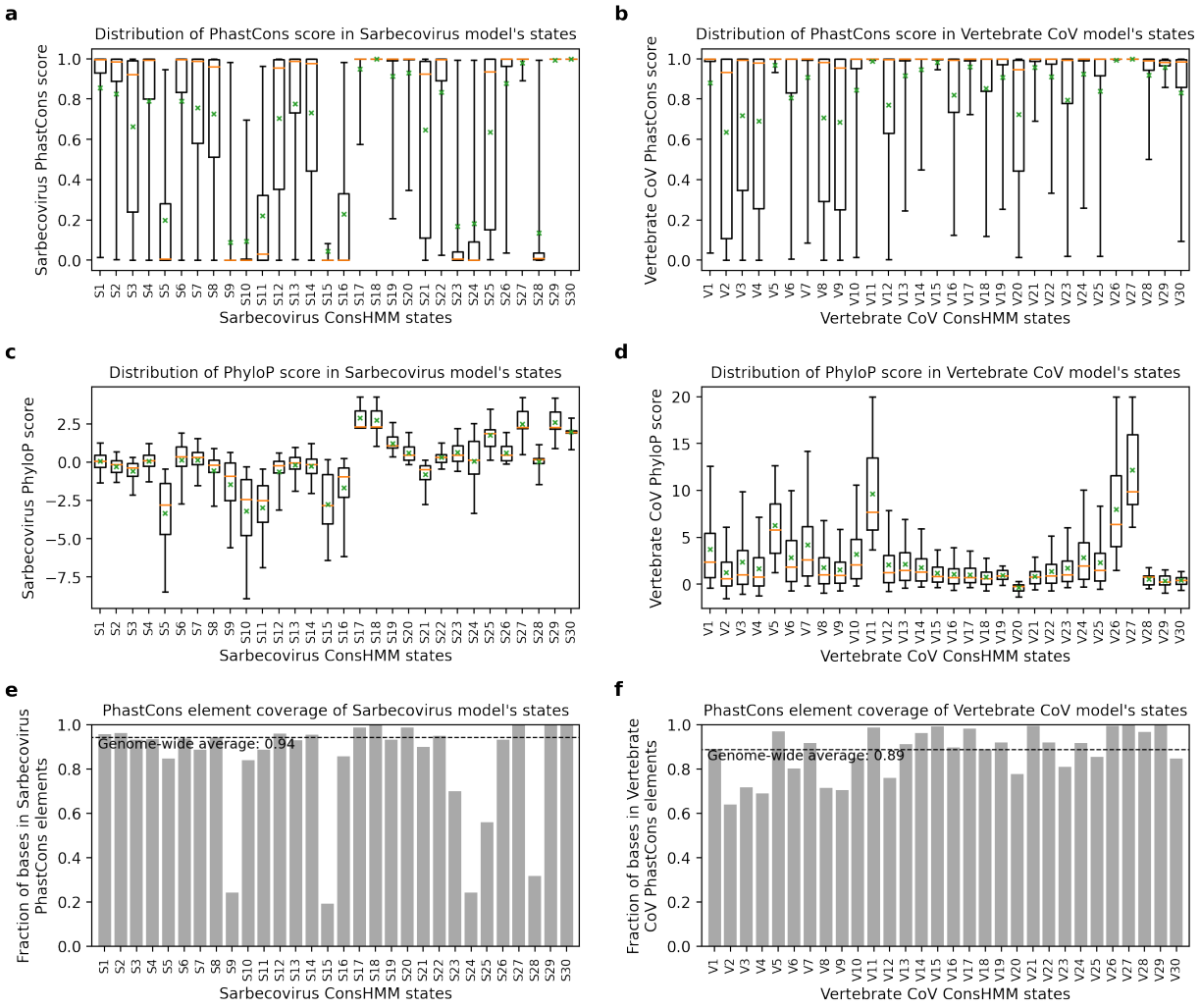
Supplementary Figure 4.1. Conservation state enrichment for protein products.
a. Fold enrichment for protein products in conservation states learned from the Sarbecovirus model. Each row corresponds to a state. First column contains the state ID. The following columns contain fold enrichment values for different protein products listed at the top of each column. Protein product coordinates and names were from UniProt Protein Product annotation[91]. 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 4.2. Sarbecoviruses associated with states S12 and S13 in the phylogenetic tree of the 44-way Sarbecovirus alignment. Similar to **Fig. 4.2c** 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.


a
ORF1a
b
ORF1ab
c
S
d
ORF3a
e
E
f
M
g
ORF6
h
ORF7a
i
ORF7b
j
ORF8
k
N
l
ORF10
m
Interaction with RMP Remdesivir
n
Receptor-binding domain (RBD)
o
Receptor-binding motif; binding to human ACE2
p
Fusion peptide 1
q
Fusion peptide 2
r
Heptad repeat 1
s
Heptad repeat 2
t
RNA-binding
u
Dimerization
Precision
Recall
Sarbecovirus ConsHMM states
Vertebrate CoV ConsHMM states
Sarbecovirus PhyloP score
Vertebrate CoV PhyloP score
Sarbecovirus PhastCons score
Vertebrate CoV PhastCons score
Sarbecovirus PhastCons element
Vertebrate CoV PhastCons element

Supplementary Figure 4.3. Precision-recall plots for predicting genes and regions of interest. Shown in each subplot is a precision-recall plot for predicting bases that overlap external genomic annotations using ConsHMM conservation states and sequence constraint annotations. Above each subplot is the target annotation, which is either a gene (**a-l**) or a region of interest defined by UniProt[91] (**m-u**). In each subplot, prediction based on ConsHMM conservation states for bases overlapping the target annotation is shown with circles. Prediction based on sequence constraint scores is shown with continuous lines. Prediction based on PhastCons element is shown with triangles. Circles, lines, and triangles are colored according to the bottom right legend. Y-axis varies from subplot to subplot because the target annotations have different genome coverage. In most cases, at least one of the ConsHMM states have substantially greater precision at the same recall level than other sequence constraint annotations, suggesting that it has greater correspondence with the annotated bases.



Supplementary Figure 4.4. Conservation states' relationship to PhastCons and PhyloP annotations.

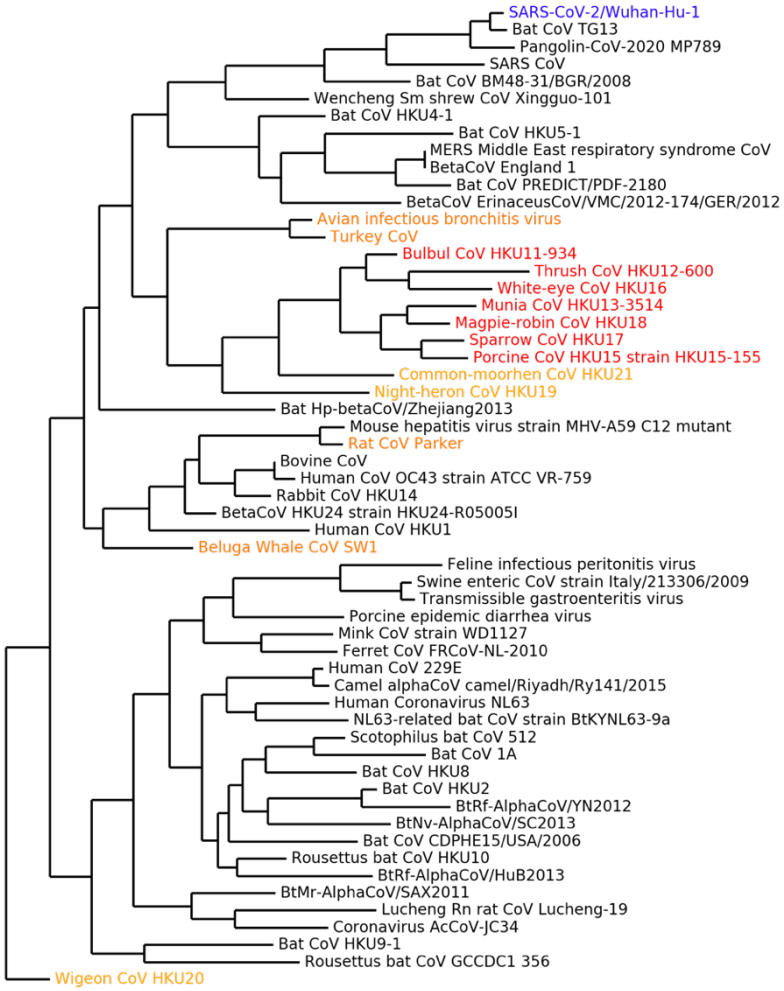
a. Shown for each conservation state learned from the Sarbecovirus alignment (x-axis) is the distribution of PhastCons score learned from the same alignment (y-axis) in bases overlapping the state. Each distribution is represented by a boxplot with median (orange horizontal line), mean (green 'x'), Q1 and Q3 (box), and Q1–1.5 IQR and Q3+1.5 IQR (whisker), where Q1 and Q3 represent 25th and 75th percentiles, respectively, and IQR (interquartile range) represent the difference between them.

b. Similar to **a** except showing conservation states and PhastCons score learned from the vertebrate CoV alignment.

c-d. Similar to **a-b**, respectively, except showing PhyloP score instead of PhastCons score.

e. Shown for each conservation state learned from the Sarbecovirus alignment (x-axis) is the fraction of bases overlapping PhastCons elements based on the same alignment (y-axis). Indicated by the horizontal dashed line is the genome-wide coverage of the PhastCons element annotation. The exact coverage is reported below the line.

f. Similar to **e** except showing conservation states and PhastCons elements learned from the vertebrate CoV alignment.



Reference genome
Strains with align and match prob. < 0.5 in states V10 and V11
Strains with align prob. > 0.5 and match prob. < 0.5 in state V10
Strains with align prob. > 0.5 and match prob. < 0.5 in state V10 and V11
Strains with align and match prob. > 0.5 in states V10 and V11

Supplementary Figure 4.5. Vertebrate CoV associated with states V10 and V11 in the phylogenetic tree of the vertebrate CoV alignment. Similar to **Fig. 4.3c** except strains colored according to their align and match probabilities in states V10 and 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 below 0.5 for both states V10 and V11, which include six CoV from avian hosts and a CoV from pig. The strains colored in orange are those with align probabilities above 0.5 and match probabilities below 0.5 for state V10. 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 both states.

a

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

State	Enrichment for *nonsingleton* mutations			Enrichment for *all observed* mutations		
	Based on GW expectation	Corrected by nucleotide composition	Corrected by mutation type	Based on GW expectation	Corrected by nucleotide composition	Corrected by mutation type
S1	1.0	1.2	0.8	1.1	1.3	0.9
S2	1.2	1.4	0.9	1.1	1.3	0.8
S3	1.3	1.4	1.0	1.3	1.5	1.1
S4	1.4	**1.6**	1.1	**1.3**	**1.5**	1.1
S5	1.4	1.4	1.1	1.3	1.4	1.1
S6	**2.1**	**1.9**	**1.6**	**1.7**	**1.7**	**1.4**
S7	1.2	**1.3**	1.0	**1.2**	**1.3**	1.0
S8	1.5	**1.6**	1.2	**1.4**	**1.6**	1.2
S9	1.7	1.6	1.6	**1.8**	**1.7**	**1.7**
S10	1.5	1.5	1.3	1.2	1.2	1.1
S11	0.9	1.2	0.7	1.1	1.3	0.9
S12	**1.8**	**1.9**	1.4	1.3	1.4	1.0
S13	1.2	1.5	0.9	1.0	1.2	0.8
S14	1.0	1.2	0.8	1.1	1.2	0.9
S15	1.7	2.0	1.5	1.5	1.7	1.4
S16	1.3	1.3	1.1	1.1	1.2	1.0
S17	**0.7**	**0.6**	**0.8**	**0.8**	**0.7**	**0.9**
S18	**0.6**	**0.6**	**0.6**	**0.8**	**0.8**	**0.8**
S19	**1.3**	1.2	1.2	1.2	1.1	1.1
S20	1.0	1.3	0.8	1.0	1.2	0.9
S21	1.5	1.7	1.2	1.5	1.7	1.2
S22	0.9	1.0	0.7	1.1	1.2	0.9
S23	1.2	1.5	1.4	1.0	1.2	1.2
S24	1.5	1.6	1.5	1.4	1.5	1.5
S25	1.1	1.1	1.2	1.1	1.2	1.2
S26	**2.4**	**2.2**	**2.0**	**1.6**	**1.6**	**1.4**
S27	0.9	0.9	1.0	1.0	1.0	1.0
S28	**2.0**	**1.8**	**2.0**	**1.7**	**1.6**	**1.7**
S29	**2.4**	**2.2**	1.3	**2.1**	**1.9**	1.3
S30	0.0	0.0	0.0	0.0	0.0	**0.0**

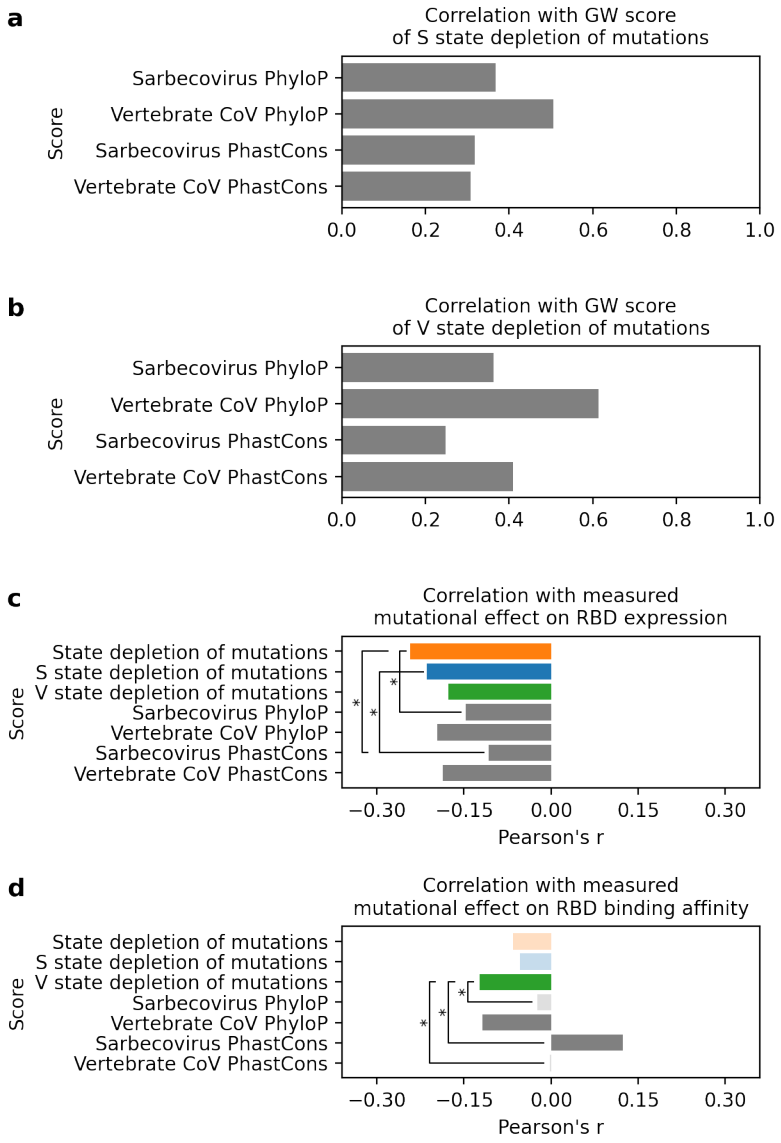
b

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

State	Enrichment for *nonsingleton* mutations			Enrichment for *all observed* mutations		
	Based on GW expectation	Corrected by nucleotide composition	Corrected by mutation type	Based on GW expectation	Corrected by nucleotide composition	Corrected by mutation type
V1	0.6	0.8	0.6	0.8	0.9	0.8
V2	1.2	1.1	1.0	1.1	1.1	1.0
V3	**1.7**	1.2	**1.4**	**1.4**	1.1	1.2
V4	1.0	1.2	0.9	1.0	1.2	0.9
V5	0.6	0.6	0.8	0.7	**0.6**	0.8
V6	1.0	0.9	0.9	1.0	0.9	1.0
V7	0.7	0.8	0.7	0.8	0.9	0.8
V8	1.4	1.2	1.3	1.2	1.1	1.1
V9	1.0	1.0	1.0	1.0	1.0	1.0
V10	**0.4**	**0.6**	**0.4**	**0.6**	**0.7**	**0.6**
V11	**0.2**	**0.2**	**0.3**	**0.3**	**0.3**	**0.4**
V12	1.0	1.1	1.1	1.0	1.0	1.0
V13	**1.5**	1.3	**1.5**	**1.4**	**1.3**	**1.4**
V14	1.2	1.1	1.2	1.2	1.1	1.2
V15	1.0	1.0	0.9	1.1	1.1	1.0
V16	1.1	1.1	1.1	1.1	1.2	1.1
V17	0.8	0.8	0.8	0.8	0.8	0.8
V18	1.0	0.9	1.0	1.0	0.9	1.0
V19	1.1	1.1	1.1	1.1	1.1	1.1
V20	**1.7**	**1.5**	**1.4**	**1.6**	**1.6**	**1.4**
V21	1.2	1.2	1.1	1.1	**1.1**	1.1
V22	1.1	1.1	1.1	0.9	1.0	1.0
V23	1.3	1.3	1.3	1.3	1.2	1.2
V24	1.1	1.1	1.2	0.9	0.9	1.0
V25	0.8	0.8	0.8	**0.8**	**0.8**	**0.8**
V26	0.7	0.7	0.9	**0.6**	**0.7**	0.8
V27	**0.2**	**0.2**	**0.3**	**0.4**	**0.4**	**0.5**
V28	1.3	1.3	1.3	**1.4**	**1.4**	**1.4**
V29	1.6	1.5	1.7	1.4	1.3	1.4
V30	**1.8**	**1.8**	**1.8**	**1.6**	**1.6**	**1.6**

Supplementary Figure 4.6. 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. State ID is shown in red if the state was significantly enriched for mutations in all six settings in which we computed enrichment, which are shown in the following six columns. State ID is shown in blue if the state was significantly depleted for mutations in all settings. Otherwise, state ID is shown in black. 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 (GW) fraction of observed mutations among possible mutations, as done in **Fig. 4.2b** (**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 (**Methods**). Similarly, fourth column contains enrichment values for nonsingleton mutations corrected by the type (i.e. intergenic, synonymous, missense, nonsense) of the mutations annotated by each state (**Methods**). Fifth, sixth, and seventh columns are similar to second, third, and fourth columns except the enrichment values are

computed based on all observed mutations instead of nonsingleton mutations. All mutations were reported by Nextstrain[99] based on sequences available on GISAID[89] (**Methods**). 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 figure do not have any correspondence to row order in **a**.



Supplementary Figure 4.7. Correlation with measured mutational effect for tracks based on state depletion of mutations and existing sequence constraint scores.
a. Bar graph showing correlation between our genome-wide (GW) score of depletion of mutations in conservation states from the Sarbecovirus model and four sequence constraint scores listed along the y-axis. The sequence constraint scores were based on either the Sarbecovirus or vertebrate CoV sequence alignment provided to ConsHMM using either PhastCons or PhyloP as the scoring method (**Methods**). A similar plot using the genome-wide score of depletion of mutations in states from both ConsHMM models instead of only the Sarbecovirus model is shown in **Fig. 4.4f**.
b. Similar to **a**, except using genome-wide (GW) score of depletion of mutations in conservation states from the vertebrate CoV model instead of the Sarbecovirus model.
c. Bar graph showing correlation between measured mutational effect on RBD expression and seven scores, which include three genome-wide scores based on conservation state depletion of mutations and four existing sequence constraint scores. Correlations computed with our scores based on both ConsHMM models, the Sarbecovirus model, and the vertebrate CoV

model are shown in orange, blue, and green bars, respectively. Correlations computed with sequence constraint scores are shown in grey bars. Correlations with no statistical significance after Bonferroni correction by the total number of scores ($p < 0.05/7$) are shown in lighter colors (**Methods**). Black connecting lines and an asterisk are shown for pairs of a state-based score (colored bars) and an existing constraint score (grey bars) if at least one of the two scores has a statistically significant negative correlation and if the two scores also exhibit statistically significant difference in their correlations. Statistically significant difference in correlation was determined based on Zou's confidence interval test[103]. The test's confidence level was set to 0.996 (1 – 0.05/12) after Bonferroni correction by the number of pairs of a state-based score and a constraint score, where at least one score in the pair has a statistically significant negative correlation with mutational effect on RBD expression (**Methods**). Mutational effect on RBD expression was measured by the study referenced in **Fig. 4.4g** that conducted a deep mutational scanning of 3,819 nonsynonymous amino acid mutations in RBD[94]. To compute the correlations we restricted to the 1,215 mutations that were caused by single nucleotides and free of experimental measurements that were not determined (n.d.). A positive value indicates increased expression due to mutation and a negative value indicates decreased expression. A subset of the correlations shown here are also shown in **Fig. 4.4g**.
d. Similar to **c**, except showing measured mutational effect on RBD binding affinity instead of expression and using confidence level of 0.992 (1 – 0.05/6) for Zou's confidence interval test given six pairs of correlations to compare (**Methods**).

Supplementary Tables

Description	*State*	*Aligns to*	*Matches to*	*Notable enrichments*
Unique to SARS-CoV-2 and RaTG13	S28	RaTG13	RaTG13	Most enriched for human ACE2 binding motif; Enriched for nonsingleton mutations
Aligns to most and matches to Sarbecoviruses closely related to SARS-CoV-2	S9	All Sarbecoviruses	RaTG13	
	S6			Enriched for nonsingleton mutations and homoplastic mutations
	S7		Small subset of close strains including RaTG13	
	S8			Most enriched for heptad repeat 1
	S10		Subset of strains including RaTG13 and SARS-CoV	Most enriched for spike protein's receptor binding domain (RBD) and gene ORF3a
Deviation along a branch of the Sarbecovirus phylogeny	S12	All Sarbecoviruses	Subset of strains corresponding to a subtree in the phylogeny (**Supplementary Fig. 4.2**)	Enriched for nonsingleton mutations
	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	
	S11			
	S15			
	S5			Most enriched fusion peptide 1
	S24	All except several distal strains		Most enriched for gene ORF8
Aligns and matches to most Sarbecoviruses	S4	All Sarbecoviruses	All except several strains	
	S3			
	S2			
	S1			
	S26			Enriched for nonsingleton mutations
	S21			
	S22		All except a strain	
	S23			Most enriched for gene S
	S14		All Sarbecoviruses	
	S17			Depleted of nonsingleton mutations and homoplastic mutations
	S18			Most enriched for gene E and region that interacts with RMP Remdesivir; Most depleted of nonsingleton mutations
	S20			
	S19		All except two distal strains	Most enriched for genes ORF6, ORF7a, and N and RNA-binding region; Enriched for nonsingleton mutations
	S25	All except two distal strains	All except two distal strains	Most enriched for gene ORF7b
	S27	All except two strains	All except two strains	Most enriched for genes orf1a (YP_009725295.1) and orf1ab (YP_009724389.1)
Non-coding or putative artifact	S29	All except several close and distal strains	All except several close and distal strains	Most enriched for intergenic bases, gene ORF10; Most enriched for nonsingleton mutations
	S30			

Supplementary Table 4.1. 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. 4.2a**. 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. 4.2b**. RaTG13 refers to a bat CoV most closely related to SARS-CoV-2. Nonsingleton mutations mentioned in this table are nonsingleton mutations observed in SARS-CoV-2 sequences based on Nextstrain's annotation of GISAID's SARS-CoV-2 sequences[89,99] (**Methods**). Homoplastic mutations mentioned in this table are stringently identified homoplastic mutations from a previous study[90]. 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	Most enriched for gene ORF3a
	V29	RaTG13 and SARS-CoV	RaTG13	
	V30	RaTG13 and pangolin CoV	RaTG13 and pangolin CoV	Most enriched for nonsingleton mutations; Most enriched for genes ORF7b and ORF8
	V20	Four closest strains	RaTG13 and pangolin CoV	Enriched for nonsingleton mutations; Most enriched for receptor binding domain (RBD) and human ACE2 binding domain motif
	V19		Four closest strains	Most enriched for genes ORF6 and ORF7a
	V18	Four closest strains and several others	Four closest strains	
	V16			
	V17		Four closest strain and a bat CoV	
	V21			
	V15		Four closest strains	Most enriched for intergenic bases and genes E and 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	Most enriched for dimerization-associated region
	V13			Most enriched for gene N and RNA-binding region; Enriched for nonsingleton mutations and homoplastic mutations
	V23			
	V24			Most enriched for gene M
	V12			Most enriched for gene S and heptad repeat 2
	V25			Most enriched for gene orf1a (YP_009725295.1)
Aligns to most and matches to some vertebrate CoV	V9	All except several strains	Four closest strains	
	V8			Most enriched for orf1ab (YP_009724389.1)
	V3	All vertebrate CoV		Most enriched for fusion peptide 1; Enriched for nonsingleton mutations
	V2			Most enriched for heptad repeat 1
	V6		Four closest strains and several distal strains, most of which are from birds	
	V5		Four closest strains with several others	
	V4			Most enriched for fusion peptide 2
Aligns to all and matches to most vertebrate CoV	V1	All vertebrate CoV	All except several close strains	
	V7		All vertebrate CoV	
	V27			Depleted of nonsingleton mutations
Aligns and matches to most except some CoV with avian hosts	V26	All vertebrate CoV	All except several CoV, most of which are from birds (**Supplementary Fig. 4.5**)	
	V11	All except several CoV, most of which are from birds (**Supplementary Fig. 4.5**)		Most depleted of nonsingleton mutations
	V10			Depleted of nonsingleton mutations

Supplementary Table 4.2. Summary of grouping, align and match probabilities, and notable enrichments of ConsHMM conservation states learned from the vertebrate CoV alignment. Similar to **Supplementary Table 4.1** except showing vertebrate CoV model's states instead of Sarbecovirus model's states.

start	end	gene	confirmed based on human CoV	Gussow et al.
7390	7450	orf1ab		
7807	7809	orf1ab		
7809	7816	orf1ab	TRUE	
7816	7825	orf1ab		
7868	7871	orf1ab		
7931	7933	orf1ab		
8575	8589	orf1ab		
8640	8647	orf1ab		
8658	8660	orf1ab		
8888	8892	orf1ab		
8892	8893	orf1ab	TRUE	
8893	8899	orf1ab		
8963	8968	orf1ab		
8969	8973	orf1ab		
10237	10238	orf1ab		
10797	10799	orf1ab		
10869	10871	orf1ab		
11074	11076	orf1ab		
11370	11371	orf1ab		
12912	12913	orf1ab		
13328	13331	orf1ab	TRUE	
16190	16193	orf1ab		
18171	18174	orf1ab		
18230	18231	orf1ab		
19131	19134	orf1ab		
19958	19961	orf1ab		
20351	20353	orf1ab		
20391	20397	orf1ab		
23843	23844	S		
23938	23941	S		
24001	24002	S		
24226	24227	S		
24227	24229	S	TRUE	TRUE
24775	24778	S		
24990	25000	S		
25322	25345	S		
26610	26611	M		
26874	26938	M		
26939	27041	M		
27043	27047	M		
27049	27067	M		
27078	27085	M		
27086	27135	M		
28396	28415	N		
28415	28423	N	TRUE	
28496	28500	N		
28561	28567	N		
28680	28686	N		
28704	28706	N		
28797	28809	N		
28857	28875	N		
28946	28966	N		
29001	29002	N		
29012	29014	N		
29024	29026	N		
29115	29116	N		TRUE
29116	29124	N	TRUE	TRUE
29218	29233	N		
29241	29362	N		
29374	29400	N		
29730	29731	non-coding		
29764	29771	non-coding		
29784	29803	non-coding		

Supplementary Table 4.3. Genomic segments unique to pathogenic human CoV and missing in less pathogenic human CoV identified by state V14.
Each row corresponds to a genomic segment annotated by state V14, which corresponds to bases with high (>0.5) align probabilities for SARS-CoV and MERS-CoV and low (<0.5) align probabilities for common-cold-associated human CoV. First and second columns denote 0-based genomic coordinates (BED format). Third column shows the gene in which the genomic segments are located if it is in a gene or "non-coding" if it is not a gene. Fourth column denotes whether the base is confirmed to be in all pathogenic human CoV and missing in all less pathogenic human CoV based on an alignment of 944 human CoV sequences. Last column denotes whether the genomic segment was identified as an insertion specific to pathogenic strains in a prior study[88].

References

1. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
3. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
4. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 (2015).
5. Pound, L. D. *et al.* Deletion of the mouse Slc30a8 gene encoding zinc transporter-8 results in impaired insulin secretion. *Biochem. J.* **421**, 371–376 (2009).
6. Church, C. *et al.* Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.* **42**, 1086–1092 (2010).
7. Nichols, C. E. *et al.* Lrp1 Regulation of Pulmonary Function: Follow-up of Human GWAS in Mouse. *Am. J. Respir. Cell Mol. Biol.* (2020).
8. Bi, X. *et al.* ILRUN, a human plasma lipid GWAS locus, regulates lipoprotein metabolism in mice. *Circ. Res.* **127**, 1347–1361 (2020).
9. Flint, J. & Eskin, E. Genome-wide association studies in mice. *Nat. Rev. Genet.* **13**, 807–817 (2012).
10. Schwartz, S. *et al.* Human–Mouse Alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
11. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, (2005).
12. 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).
13. Davydov, E. V *et al.* Identifying a High Fraction of the Human Genome to be under

Selective Constraint Using GERP++. *PLOS Comput. Biol.* **6**, e1001025 (2010).

14. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, (2011).
15. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).
16. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
17. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730–732 (2007).
18. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
19. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science (80-.).* **338**, 1593 LP – 1599 (2012).
20. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
21. Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, (2014).
22. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, (2014).
23. Le, H.-S., Oltvai, Z. N. & Bar-Joseph, Z. Cross-species queries of large gene expression databases. *Bioinformatics* **26**, 2416–2423 (2010).
24. Wise, A., Oltvai, Z. N. & Bar−Joseph, Z. Matching experiments across species using expression values and textual information. *Bioinformatics* **28**, i258–i264 (2012).
25. Li, W. V., Chen, Y. & Li, J. J. TROM: A Testing-Based Method for Finding Transcriptomic Similarity of Biological Samples. *Stat. Biosci.* **9**, 105–136 (2017).
26. Normand, R. *et al.* Found In Translation: a machine learning model for mouse-to-human

inference. *Nat. Methods* (2018). doi:10.1038/s41592-018-0214-9

27. Okamura, Y., Obayashi, T. & Kinoshita, K. Comparison of Gene Coexpression Profiles and Construction of Conserved Gene Networks to Find Functional Modules. *PLoS One* **10**, e0132039 (2015).

28. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411 (2018).

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

30. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).

31. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106- (2008).

32. Taher, L. *et al.* Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res.* **21**, 1139–1149 (2011).

33. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485**, 376–380 (2012).

34. Weirauch, M. T. & Hughes, T. R. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66–74 (2010).

35. Arneson, A. & Ernst, J. Systematic discovery of conservation states for single-nucleotide annotation of the human genome. *Commun. Biol.* **2**, 248 (2019).

36. Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection. *Cell* **145**, 773–786 (2011).

37. Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).

38. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

39. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2018).

40. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

41. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).

42. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421 (2017).

43. Keller, M. P. *et al.* Gene loci associated with insulin secretion in islets from nondiabetic mice. *J. Clin. Invest.* **129**, 4419–4432 (2019).

44. Multhaup, M. L. *et al.* Mouse-Human Experimental Epigenetic Analysis Unmasks Dietary Targets and Genetic Liability for Diabetic Phenotypes. *Cell Metab.* **21**, 138–149 (2015).

45. Bogue, M. A. *et al.* Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data. *Nucleic Acids Res.* **48**, D716–D723 (2020).

46. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).

47. Hughes, L. H., Schmitt, M., Mou, L., Wang, Y. & Zhu, X. X. Identifying Corresponding Patches in SAR and Optical Images With a Pseudo-Siamese CNN. *IEEE Geosci. Remote Sens. Lett.* **15**, 784–788 (2018).

48. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. Signature verification using a 'Siamese' time delay neural network. in *Proccedings of the Advances in Neural*

Information Processing Systems 737–744 (1994).

49. Paszke, A. *et al.* Automatic differentiation in PyTorch. in *Proceedings of Neural Information Processing Systems Audiodiff Workshop* (2017). doi:10.1017/CBO9781107707221.009

50. Pedregosa, F. *et al.* Scikit-learn : Machine Learning in Python. *J. Mach. Learn. Res.* (2011). doi:https://dl.acm.org/citation.cfm?id=2078195

51. Bilenko, N. Y. & Gallant, J. L. Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging. *Front. Neuroinform.* **10**, 49 (2016).

52. Andrew, G., Arora, R., Bilmes, J. & Livescu, K. Deep Canonical Correlation Analysis. in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* III–1247–III–1255 (JMLR.org, 2013).

53. Wang, W., Arora, R., Livescu, K. & Bilmes, J. On Deep Multi-View Representation Learning. in *Proceedings of the 32nd International Conference on Machine Learning* (2015).

54. Ioffe, S. Improved Consistent Sampling, Weighted Minhash and L1 Sketching. in *2010 IEEE International Conference on Data Mining* 246–255 (2010). doi:10.1109/ICDM.2010.80

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

56. Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).

57. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

58. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci

for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).

59. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

60. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

61. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

62. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

63. Rietveld, C. A. *et al.* GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science (80-.).* **340**, 1467 LP – 1471 (2013).

64. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).

65. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

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

67. Bar-Joseph, Z., Gifford, D. K. & Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**, S22–S29 (2001).

68. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).

69. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

70. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

71. Kwon, S. B. & Ernst, J. Learning a genome-wide score of human–mouse conservation at

the functional genomics level. *Nat. Commun.* **12**, 2495 (2021).

72. Vu, H. & Ernst, J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *bioRxiv* (2021). doi:10.1101/2020.11.17.387134

73. Liu, N. *et al.* Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. *Epigenetics Chromatin* **14**, 41 (2021).

74. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

75. Reiff, S. B. *et al.* The 4D Nucleome Data Portal: a resource for searching and visualizing curated nucleomics data. *bioRxiv* (2021). doi:10.1101/2021.10.14.464435

76. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).

77. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

78. Fernandes, J. D. *et al.* The UCSC SARS-CoV-2 Genome Browser. *Nat. Genet.* **52**, 991–998 (2020).

79. 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).

80. Arneson, A., Felsheim, B., Chien, J. & Ernst, J. ConsHMM Atlas: conservation state annotations for major genomes and human genetic variation. *NAR Genomics Bioinforma.* **2**, lqaa104 (2020).

81. Jungreis, I., Sealfon, R. & Kellis, M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nat. Commun.* **12**, 2642 (2021).

82. 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.* **13**, 2168–2178 (2020).

83. 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

84. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* **6**, eabb9153 (2020).

85. 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.* **35**, 337–339 (2020).

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

87. Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e10 (2020).

88. Gussow, A. B. *et al.* Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci.* **117**, 15193–15199 (2020).

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

90. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).

91. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

92. 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).

93. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat. Commun.* **11**, 6013 (2020).

94. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295-1310.e20 (2020).

95. Le Bert, N. *et al.* SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* **584**, 457–462 (2020).

96. Mateus, J. *et al.* Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science (80-.).* **370**, 89–94 (2020).

97. 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).

98. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).

99. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

100. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).

101. De Maio, N. *et al.* Updated analysis with data from 12th June 2020. *Virological.org* (2020).

102. Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLoS Genet.* **16**, e1009175 (2020).

103. Zou, G. Y. Toward using confidence intervals to compare correlations. *Psychol. Methods* **12**, 399–413 (2007).

104. Diedenhofen, B. & Musch, J. Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One* **10**, (2015).

105. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).