Title
Genome variation over multiple timescales and dimensions

Author
Keough, Kathleen Coll

Publication Date
2020

Peer reviewed|Thesis/dissertation

Genome variation over multiple timescales and dimensions

by
Kathleen Coll Keough

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Nadav Ahituv
77BA96E7DAE34F4...

Nadav Ahituv

Chair

DocuSigned by:
K S Pollard
DocuSigned by:
E6...

Katherine S. Pollard

DocuSigned by:
Bruce Conklin
A2A860066965458...

Bruce Conklin

Committee Members

Acknowledgements

I would like to first thank my advisor, Katie Pollard, for supporting me over the past five years, giving me lots of exciting ideas to pursue, and encouraging me to think and work creatively. I would also like to thank my co-advisor, Bruce Conklin, for support, ideas and feedback over the years. It was an honor to be part of both the Pollard and Conklin labs, and I am grateful to all lab members from each for feedback, advice, and help over the years. In addition, I thank the chair of my thesis committee, Nadav Ahituv, for welcoming me into the graduate program at UCSF five years ago, and continuing to be a source of support, feedback, and collaboration through my graduate education. I have collaborated with many wonderful scientists over the years, and am grateful to everyone for teaching me new things and contributing towards making research rewarding.

I would like to thank my friends, from the west coast to the midwest and beyond, for helping ensure I balanced out my research with ample skiing, mountain biking, hiking, and other adventures. My dog, Frodo, for recognizing when we both need a break and requesting timely walks. I have so much appreciation for my partner, Reeve Dunne, for supporting and encouraging me through this journey, particularly by cooking lots of tasty food and encouraging fun adventures.

Finally, thank you to my family; particularly my mom and dad, who have supported and encouraged me throughout my life.

Contributions

Previously published material:

Basis for chapter 1:

Keough, Kathleen C., Svetlana Lyalina, Michael P. Olvera, Sean Whalen, Bruce R. Conklin, and Katherine S. Pollard. “AlleleAnalyzer: A Tool for Personalized and Allele-Specific GRNA Design.” *Genome Biology* 20, no. 1 (August 15, 2019): 167. https://doi.org/10.1186/s13059-019-1783-3.

Contribution towards chapter 2:

Ryu, Hane, Fumitaka Inoue, Sean Whalen, Alex Williams, Martin Kircher, Beth Martin, Beatriz Alvarado, et al. “Massively Parallel Dissection of Human Accelerated Regions in Human and Chimpanzee Neural Progenitors.” *BioRxiv*, January 29, 2018, 256313. https://doi.org/10.1101/256313.

Basis for chapter 3:

Damas, J.*, Hughes, G.M.*, Keough, K.C.*, Painter, C.A.*, Persky, N.S.*, Corbo, M., Hiller, M., Koepfli, K.-P., Pfenning, A.R., Zhao, H., et al. (2020). Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. *BioRxiv* 2020.04.16.045302. https://www.biorxiv.org/content/10.1101/2020.04.16.045302v1

**these authors contributed equally to this work*

Genome variation over multiple timescales and dimensions

Kathleen Keough

Abstract

Genomic variation does not only include nucleotide changes, it also comprises changes in DNA shape, structure, epigenetic marks, and expression, all of which can occur over generations, cellular differentiation, the span of a few hours or a few millennia. This doctoral thesis explores the implications and opportunities presented by these multiple forms of genomic variation for genome editing, cellular differentiation, genome regulation and comparative genomics, all towards improving our understanding of genome evolution and development and benefiting human health.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1KGP: 1000 Genomes Project

ChIP: chromatin immunoprecipitation

CRISPR genome editing: genome editing system based on the clustered regularly interspaced palindromic repeats adaptive immunity system in bacteria

eQTL: expression quantitative trait locus

ESC/ES: embryonic stem cell (human if not prefixed with “m”)

GO: gene ontology

gRNA: guide RNA, used in CRISPR genome editing

GWAS: genome-wide association study

HAR: human accelerated region

iPSC: induced pluripotent stem cells

kb: kilobase (1,000 basepairs of DNA)

KDD: H3K9me2-associated domain

LAD: lamina-associated domain

LB1: lamin B1

MAF: minor allele frequency

Mb: megabase (1,000,000 basepairs of DNA)

mESC: mouse embryonic stem cell

MPRA: massively parallel reporter assay

PAM: protospacer-adjacent motif

SNP: single nucleotide polymorphism

SV: structural variant

TAD: topological associating domain

TFBM: transcription-factor binding motif

WTC: iPSC cell line developed in the Conklin lab and widely used in many contexts

1 Introduction

Genomic variation exists on all scales from DNA sequence to genome structure. Variations between individual genomes render us unique, while also rendering us susceptible to disease. Inter-individual variation is not always random, it is often inherited, passed from parent to child, generating networks of genomic variation tracing us back to the first humans (Sudmant et al., 2015). These patterns of shared genomic variation can then be used to assign us to groups, either at a high level with clades and species, within a species such a population groups, or at an even finer resolution, linking us to previously unknown family members. We can also be grouped genetically based on our susceptibility to various diseases. However, in this work I use shared genetic variation as a tool, using those same networks of genetic sharedness to design genomic tools specific to individuals with a particular disease that aim to amend the cause of the disease at its genetic source (Keough et al., 2019).

On a longer timescale, genomic variation, or the lack thereof, reminds us how very similar we actually are to the other species that share this planet with us. Many of our basic physiological processes, breathing, pumping blood, digestion, are common throughout mammals and further, and this similarity is reflected throughout our genomes. Most of our genome is exactly the same as our closest living relative, the chimpanzee, and yet there are clearly many differences between our species. The

genetic cause of these differences likely lies in the 98% of our genomes that are noncoding. This was shown in a seminal paper that compared the similarity between human and chimpanzee for the sequences that code for genes , finding that they are highly identical, suggesting that noncoding sequences have a major impact for differences between organisms (King and Wilson, 1975). To try and understand these differences subtly encoded in our genomes, we hone in on the fastest evolving loci known in humans (Pollard et al., 2006a, 2006b). To these loci we add further dimensions of information, made possible by technological advances since their discovery such as high-throughput sequencing, new methods for computational analysis of genomes, and techniques to disentangle the meaningful organization of the genome in 3D at a high resolution. These new data around these intriguing loci enable us to develop and pursue new hypotheses for their origin and function, and get closer to determining what makes us human.

Despite our complexities as humans, we are still susceptible to some of the most basic pathogens: RNA viruses, fast-evolving predators waging war against the entire phylogenetic tree of life. As I wrote this thesis, SARS-CoV2 brought the world to a halt (Zhou et al., 2020). This insidious foe was able to pass between species and individuals, often without symptoms, frequently deadly, without a cure or treatment. Scientists around the world dropped everything to help where they could, and I joined them to identify how our genomic similarities and variation may render some species more or less vulnerable to this virus, to enable more informed conservation efforts, potentially inform model animal selection, essential for vaccine and treatment

development, and inform other sources of risk, such as agriculture (Damas et al., 2020). That work continues beyond what is included here.

Finally, variation in genomic structure is important for health and development. I explore some of the basic foundations of this via analysis of lamina-associated and H3K9me2-associated domains. These generally repressive, heterochromatic regions comprise a huge portion of our genomes, and yet our understanding of how they're generated, what is their function, and how they interact with other genomic elements is nascent (Guelen et al., 2008; Meuleman et al., 2013; Peric-Hupkes et al., 2010; Poleshko et al., 2017). Disruption of these elements disrupts differentiation and causes disease, and so a better understanding is necessary. By applying a previously unused type of statistical model to the data, I uncover evidence that these regions are not monolithic, but comprise multiple subtypes, each of which vary between different cell types and encode cell-type-specific information.

Genomic variation defines us; it is both our boon and our bane, generating both beneficial traits and disease. The advances in this doctoral thesis contribute to our understanding of the forces of genomic variation in sequence and shape over cellular differentiation and evolutionary time, within and between individuals and species, and provides a tool to use genomic variation against genomic variation-based disease.

2 AlleleAnalyzer: a tool for personalized and allele-specific gRNA design

The work described in this chapter was previously published in *Genome Biology* (Keough et al., 2019). The end result of this study was a more thorough understanding of the implications and opportunities presented by naturally occurring genomic variation (single nucleotide variants and small insertions or deletions) in the field of CRISPR genome editing guide RNA (gRNA) design, as well as an open-source software tool (https://github.com/keoughkath/AlleleAnalyzer) to design and optimize gRNA combinations based on genetic variation to maximize coverage of a cohort.

2.1 Rationale for considering genomic variation in CRISPR gRNA design

CRISPR genome editing success depends on the efficiency and specificity of the gRNA design. Current gRNA design tools primarily predict efficiency and specificity of gRNAs using features such as prevalence of off-target sites, epigenetic marks and chromatin accessibility (Doench et al., 2014; Haeussler et al., 2016; Horlbeck et al., 2016). Generally, gRNAs are designed using reference genomes, such as the hg38 assembly for human or the GRCm38 assembly for mouse. However, these gRNAs are used on cell lines or organisms with many nucleotide differences from the reference (e.g., on average 0.1% of a human genome (National Institutes of Health, 2007)). While gRNAs can sometimes tolerate a single base-pair mismatch, frequently these mismatches negatively impact gRNA efficiency and render imprecise the results of specificity

prediction (Scott and Zhang, 2017; Yang et al., 2014), with potentially serious effects when gRNAs are deployed.

Previous work analyzing data from Exome Aggregation Consortium (ExAc) (Lek et al., 2016) and the 1000 Genomes project (1000 Genomes Project Consortium et al., 2015) determined that genetic variants could have a large impact on gRNA efficiency and specificity, demonstrating the need for a tool to design gRNAs using genetic variation and to identify gRNAs that could work in many people to facilitate regulatory approval for therapeutic use . The solution implemented in this previous work was to avoid negative effects of genetic variation by identifying universal gRNAs located in sites with little to no genetic variation and possessing few predicted off-targets (Scott and Zhang, 2017). However, many loci that may need to be edited lack variation-free regions for designing such gRNAs (see below). We propose personalized gRNA design, which uses the genetic variants in a genome or population, as a second approach that offers more flexibility in guide design. We further note that genetic variation is not only a challenge for gRNA design, but also an opportunity. Specifically, the use of CRISPR in research areas such as haploinsufficiency, genomic imprinting, and dominant negative diseases requires allele-specific gRNA design, which may be accomplished using heterozygous variants.

To address these needs, I developed AlleleAnalyzer, an open-source Python software tool that designs personalized and allele-specific gRNAs for individual genomes, identifies pairs of gRNAs to generate excisions likely to block expression of a gene, and

leverages patterns of shared genetic variation across thousands of publicly available genomes to design gRNA pairs that will have the greatest utility in a target population.

2.2 Incorporating genomic variation into gRNA design

Incorporating genetic variation into gRNA design enables personalized and allele-specific CRISPR experiments. A personalized gRNA is defined here as an gRNA designed to incorporate the genetic variants of the research subject. A genetic variant can impact gRNA sites by being located in or near a protospacer adjacent motif (PAM site), potentially generating or eliminating gRNA sites in an individual in a heterozygous or homozygous manner. Beyond being an impediment to designing effective gRNAs, these variants enable the design of personalized, non-allele-specific gRNAs (incorporating homozygous variants and avoiding heterozygous variants to match both alleles) and allele-specific gRNAs (incorporating heterozygous variants). The way in

which genetic variation impacts or is incorporated into gRNA design depends on the use case for the gRNA and variant zygosity (Figure 1a).

Because Cas nucleases have different PAM sequences, a variant may impact an gRNA site for one Cas but not another. We analyzed 11 Cas types (**Error! Reference source not found.**),

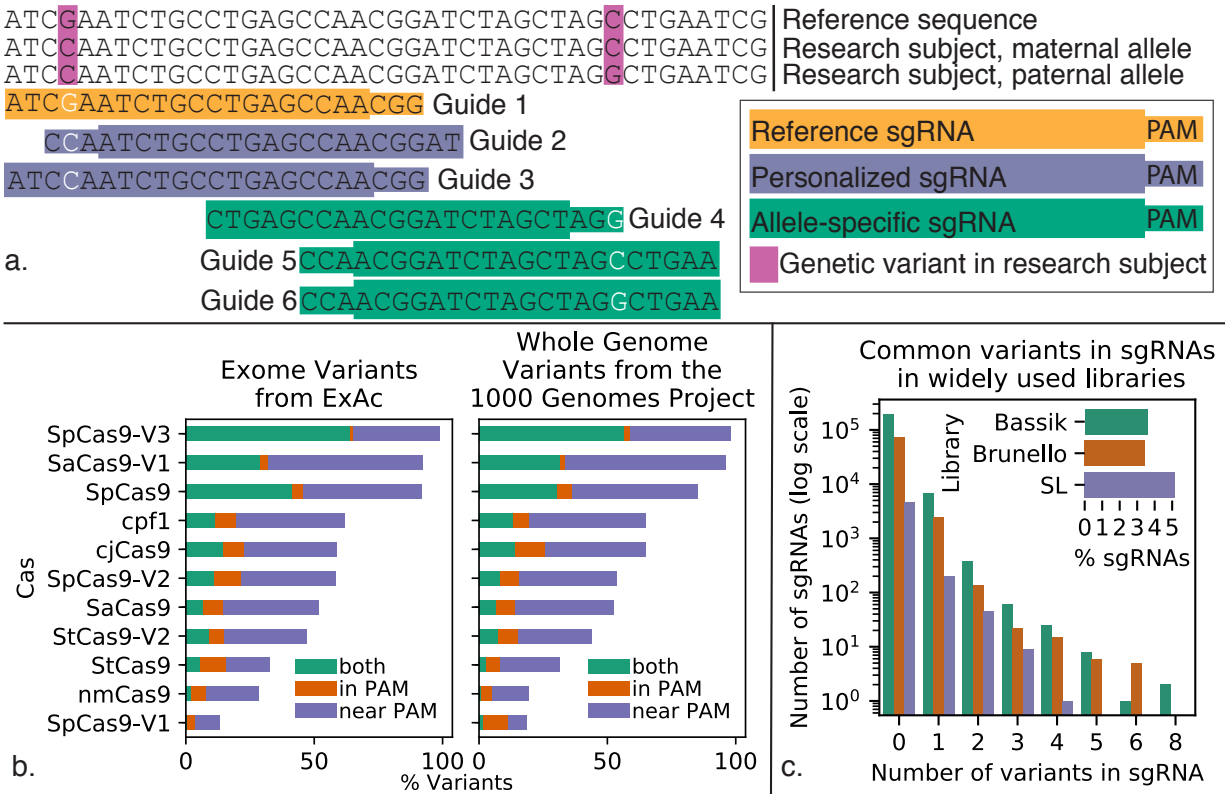

Figure 1: Analysis of allele specific gRNA sites

A) In a sample genome, tools designing gRNAs for the reference genome are imperfect matches due to genetic variants, exemplified by guide 1. AlleleAnalyzer designs personalized gRNAs, as demonstrated by guides 2 and 3, which incorporate homozygous and avoid heterozygous variants, thus designing a guide perfectly matched to both alleles in a subject. It also designs personalized allele-specific gRNAs based on incorporation of heterozygous variants, shown by guides 4-6. Guides 4 and 6 target the paternal allele, while guide 5 targets the maternal allele. B) Most variants annotated by the 1000 Genomes Project (1KGP) and the Exome Aggregation Consortium (ExAc) are in or near a PAM site. C) Analysis of common variants (minor allele frequency (MAF) greater than 5% in 1KGP), and all variants in an individual cell line (WTC) within commonly used gRNA libraries.

Table 1: Cas types

11 types of Cas enzyme were evaluated, each of which has a distinct PAM site.

Common name(s)	**Abbreviation**	**PAM**	**Properties**
SpCas9	SpCas9	NGG	*Streptococcus pyogenes* (Sp) Cas9., most widely used version with dozens of variants using same PAM, e.g. eSpCas9, SpCas9-HF1, eSpCas9 1.1 and more (Jinek et al. 2012)
SpCas9 VRER Variant	SpCas9-V1	NGCG	Version of SpCas9 with alternative targeting range (Kleinstiver et al. 2015)
SpCas9 EQR Variant	SpCas9-V2	NGAG	Version of SpCas9 with alternative targeting range (Kleinstiver et al. 2015)
SpCas9 VQR Variant	SpCas9-V3	NGAN or NGNG	Version of SpCas9 with wider targeting range (Kleinstiver et al. 2015)
SaCas9	SaCas9	NNGRRT	*Staphylococcus aureus* (Sa) Cas9. Small relative to SpCas9, (Horvath et al. 2008, Jiang et al. 2013)
SaCas9 KKH Variant	SaCas9-V1	NNNRRT	Version of SaCas9 with 2 to 4-fold increased targeting range relative of SaCas9 (Kleinstiver et al. 2015)
nmCas9	nmCas9	NNNNGATT	*Neisseria meningitidis* (Nm) Cas9, with different PAM site (Hou et al. 2013)
cpf1	cpf1	TTTN	Multiple variations, notably opposite orientation system and sticky-end cut rather than blunt. Multiple species exist, including from *Acidaminucoccus* and *Lachnospiraceae.* (Zetsche et al. 2015)
StCas9 1	StCas9-V1	NNAGAA	*Streptococcus thermophilus* (St) Cas9. Smaller relative of SpCas9. Increased specificity. (Kleinstiver et al. 2015, Muller et al. 2016)
StCas9 2	StCas9-V2	NGGNG	*Streptococcus thermophilus* (St) Cas9. Smaller relative of SpCas9. Increased specificity. (Muller et al 2016)
cjCas9	cjCas9	NNNNACA	*Campylobacter jejuni* Cas9. Smallest Cas9 ortholog to date, easy to package (Kim et al. 2017)

genome-wide variants from >2500 individuals from the 1000 Genomes Project(1000 Genomes Project Consortium et al., 2015) (1KGP), and exome variants from >60,000 individuals in ExAc. From these analyses we discovered that most variants impact gRNA sites for at least one Cas type, even when considering only variants in PAMs, which are putatively more allele-specific (Christie et al., 2017) (Figure 1b). The likelihood that a variant impacts an gRNA site differs across Cas nucleases (1KGP: range 19-98%, ExAc: range 13-99%), is positively correlated with PAM frequency in the reference genome (1KGP: Pearson rho=0.89, p=0.0002, ExAc: Pearson rho=0.84, p=0.0011, Figure 4a), and is negatively correlated with PAM size (1KGP: Pearson rho=-0.71, p=0.014, ExAc: Pearson rho=-0.74, p=0.0094). In fact, >3% of gRNAs in each of three widely used gRNA libraries (Doench et al., 2016; Morgens et al., 2017; Park et al., 2016) contain at least one common genetic variant (minor allele frequency > 5% in the 1KGP cohort), and >2% of these gRNAs contain a variant in the individual human genome of an induced pluripotent stem cell (iPSC) line WTC, commonly used for disease modeling (Drubin and Hyman, 2017) (Figure 1c, Figure 2). Failing to account for variants can reduce the efficacy of gRNAs and also

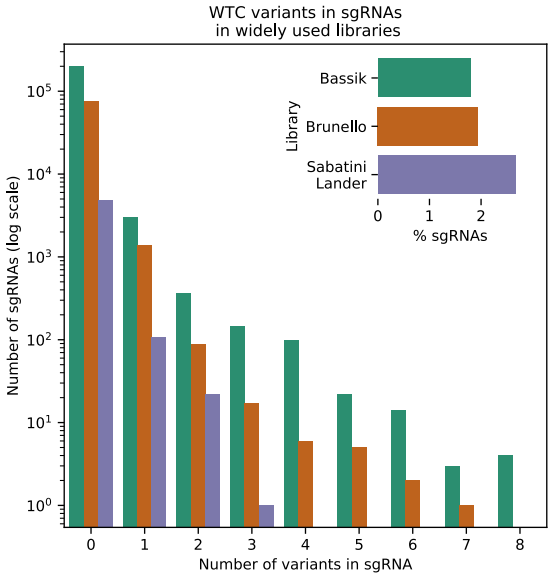

Figure 2: gRNA variants in WTC iPSC

Genomic variants in commonly-used gRNAs libraries in the WTC iPSC line.

generate unexpected off-target effects (Lessard et al., 2017). These results emphasize the importance of designing gRNAs using the personal genome of the patient or cell line where they will be deployed, or at least accounting for both heterozygous and homozygous genetic variants when interpreting results using gRNA libraries designed for the reference genome.

Heterozygous genetic variants can be leveraged to establish new therapeutic and research possibilities with allele-specific genome editing. Questions that allele-specific editing could help address include haploinsufficiency, imprinting, and allele-specific gene regulation, as well as discovery and correction of heterozygous disease variants. One promising example is genome surgery to treat dominant negative disease by excising only the disease causing copy of a gene, an approach which rescues healthy phenotypes in cell and animal models of dominant negative diseases including Huntington's disease (Shin et al., 2016a) and retinitis pigmentosa (Bakondi et al., 2015; Gao et al., 2018).

The strategy of allele-specific gene editing genome-wide was assessed by identifying pairs of allele-specific gRNA sites for each human protein-coding gene that could generate a genomic excision and eliminate protein production from just one allele. Given a Cas nuclease, an estimated maximum distance between the two gRNAs on the haplotype to be excised, and allele-specific gRNA sites based on the individual's genetic variants, it is possible to classify genes–or other genomic elements such as enhancers–as putatively targetable or not (Figure 3).

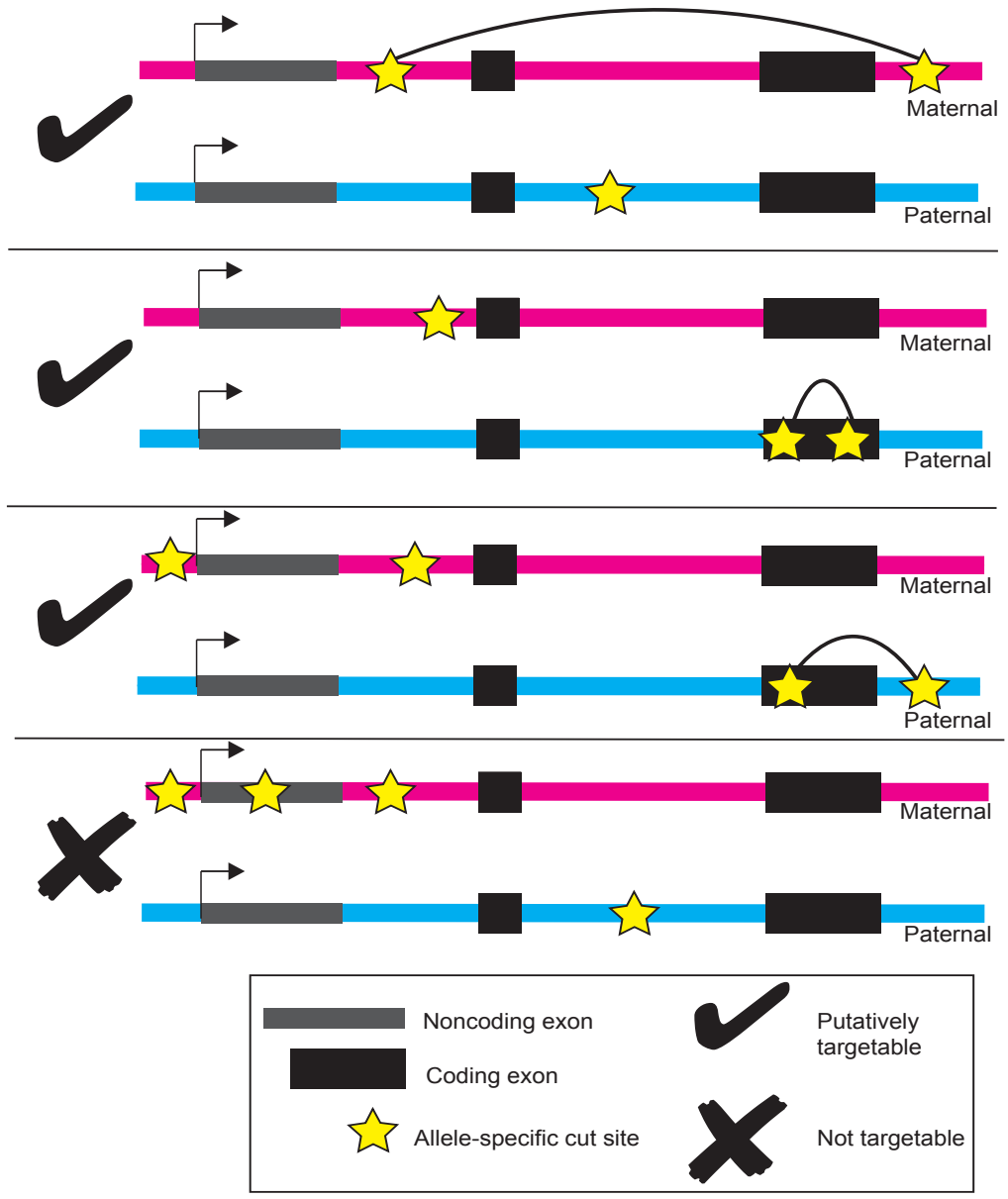

Figure 3: Allele-specific gene targeting with paired gRNAs

Strategy to determine whether a gene is targetable for dual-gRNA allele-specific excision-based knockout.

The term putatively targetable is used here when a pair of allele-specific gRNAs exists

but has not yet been tested, because it will not always be possible to cut specifically at a

site and coding exon excision will not always stop expression. Previous work indicates that excision of large genomic fragments (>10 kilobases) is feasible, and that excision of coding exons via gRNAs targeted to flanking noncoding regions, such as promoter or intronic regions, can mediate gene knockout (Chen et al., 2014; Shin et al., 2016b; Tabebordbar et al., 2016).

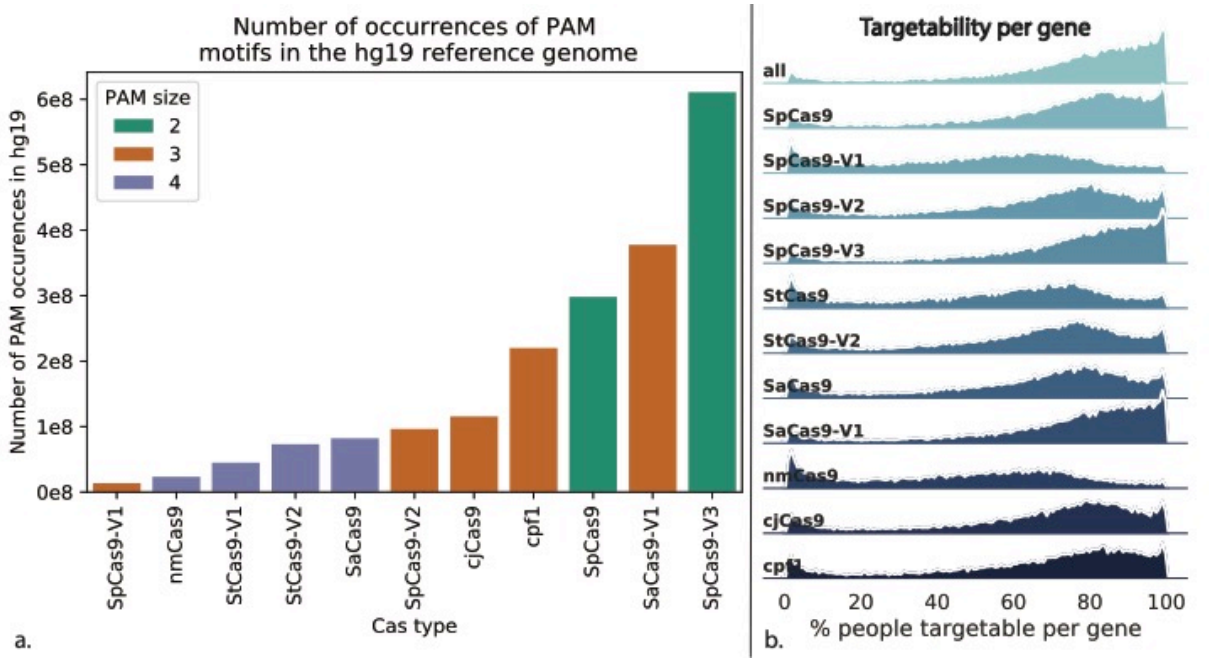

Figure 4: Target availability by Cas enzyme

A) PAM frequencies in the human reference genome hg19, colored by size of the PAM site (number of non-"N" nucleotides in motif). B) In this faceted density plot, height of the colored portion indicates the proportion of genes where the specified percentage (on the x-axis) of the 1000 genomes cohort is putatively targetable.

As an example, suppose we choose a maximum distance of 10 kilobases (kb) between gRNAs, require the gRNAs to be within the gene including introns, and consider 11 Cas varieties (Table 1). Then the average individual from 1KGP is putatively targetable for allele-specific excision at 64% of protein-coding genes (Shin et al., 2016a). The rate of

putatively targetable individuals per gene is evenly distributed across chromosomes but varies by Cas nuclease and gene (Figure 4b). For genes that are not putatively targetable, additional allele-specific gRNA sites may be found by leveraging non-coding variants up- and down-stream of the gene, or even in distal enhancers for the gene (Shin et al., 2016a). As a second example, I found that by simply including the 5 kb flanking regions of each gene, we can increase the mean proportion of putatively targetable protein-coding genes per 1KGP individual to 75%. A caveat to this is that specificity of each gRNA pair will vary greatly, potentially even between gRNAs targeting the same pair of heterozygous variants. Therefore, we conclude that allele-specific excision may be applicable to the vast majority of genes in most human genomes, but extensive experimental optimization for efficiency and specificity will be needed.

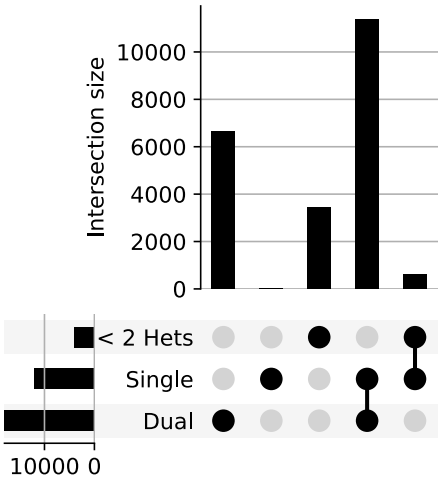

Figure 5: Genes targetable in WTC with single- or paired-gRNA approach

Genes were evaluated using variants from the WTC iPSC line for targetability based on a single- or paired-gRNA approach.

Since some genes in a given individual do not have a pair of allele-specific gRNAs, we asked if gene silencing with a single allele-specific gRNA within the coding sequence (single-guide strategy) makes more genes putatively targetable. I compared paired-guide and single-guide strategies for allele-specific gene knockout in the individual human genome of the WTC iPSC line (Drubin and Hyman, 2017) and found that more than twice as many

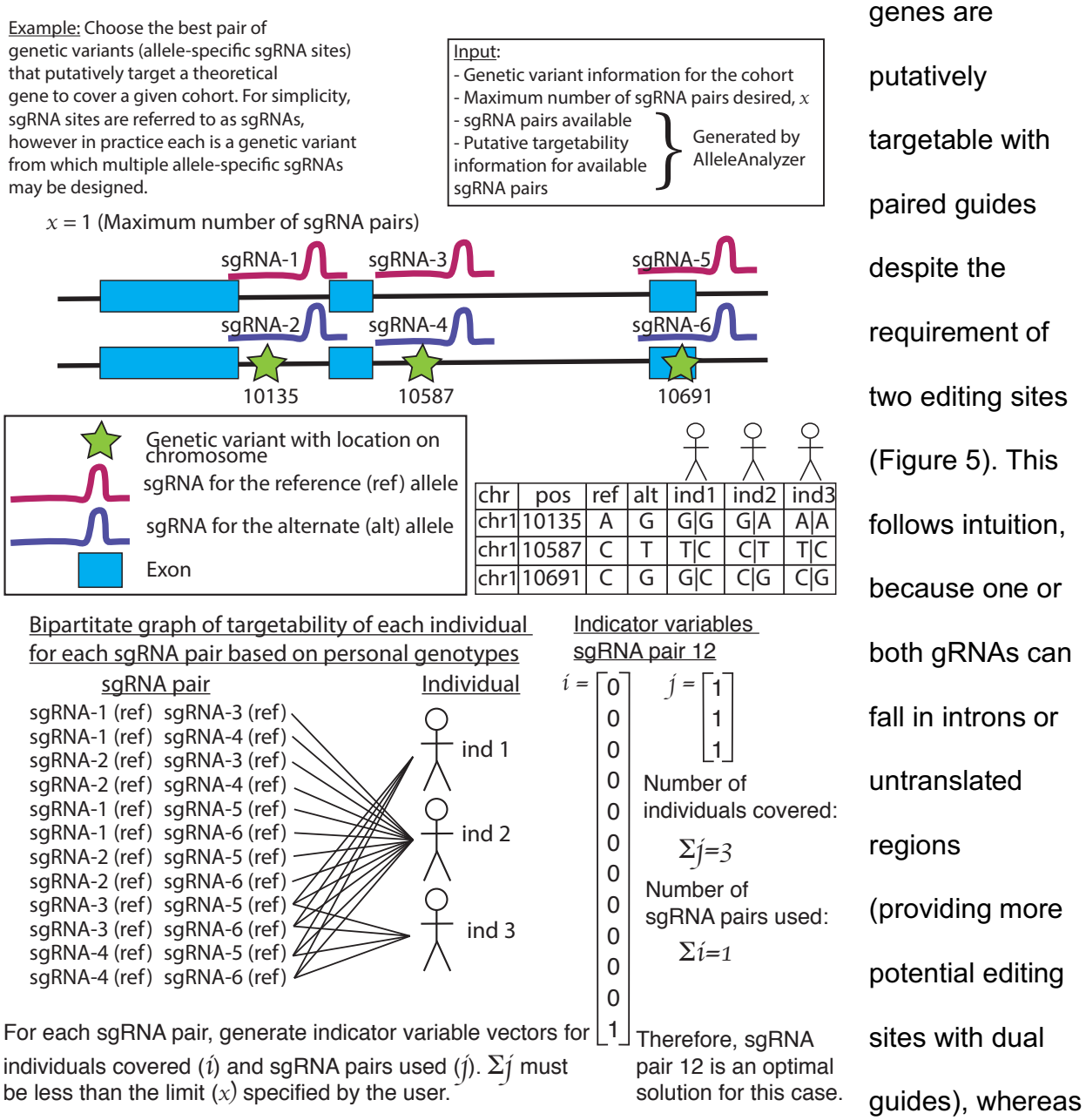

Figure 6: Set cover approach to maximize population coverage by variant-informed gRNA selection

Our approach, based on the set cover problem, to maximize coverage of a population group based on informed variant selection for gRNA design.

genes are putatively targetable with paired guides despite the requirement of two editing sites (Figure 5). This follows intuition, because one or both gRNAs can fall in introns or untranslated regions (providing more potential editing sites with dual guides), whereas individual gRNAs in the single-guide strategy are limited to coding regions. Genes that are putatively targetable with a single- and not paired-guide

approach tend to have less than two heterozygous variants in the gene, indicating that a lack of multiple variants is the primary reason a paired-guide strategy fails. These genes could be putatively targetable with a paired-guide strategy by incorporating flanking, promoter, or other regulatory regions. Again, putative editing sites and gRNAs need to be experimentally validated. This suggests that in most cases allele-specific gene targeting may be greatly enhanced by including paired-guides in the experimental approach.

Genome editing gRNAs do not need to be designed one genome at a time. Variants that impact gRNA sites are often shared among large proportions of the individuals within and sometimes between populations due to haplotype structure. Previous work had a similar goal of developing gRNAs for broad use (Scott and Zhang, 2017). However that work focused on targeting invariant (or low variation) segments of the genome towards homozygous, single-gRNA-based CRISPR editing while AlleleAnalyzer focuses on taking advantage of genome variation for allele-specific editing with individual gRNAs, or pairs of gRNAs. Allele sharing varies by population and locus, as individuals with common ancestry will share haplotypes that harbor specific sets of variants. We therefore developed an algorithm to identify allele-specific gRNA guide pairs for a given gene that cover the maximum number of individuals in a population; these have the broadest therapeutic potential, similar to designing a drug to treat as many people as possible (**Error! Reference source not found.**). Specifically, this method seeks to cover the most people with the fewest gRNA pairs using their shared heterozygous variants; this is similar to the “set cover” problem in that the

algorithm identifies an optimal combination rather than simply selecting most shared gRNA pairs, which could disproportionately favor one group over another (Clarkson, 1993). Our algorithm generates optimized pairs of gRNAs that can be used to study or

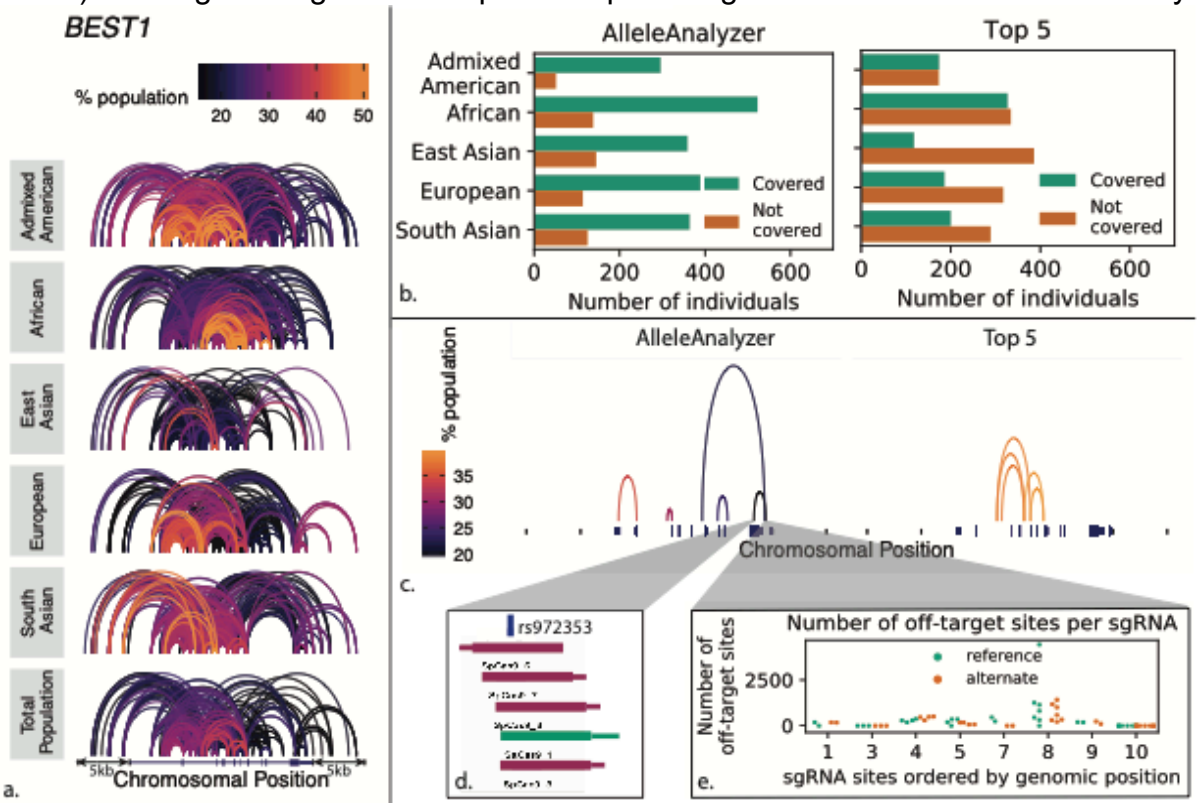

Figure 7: Targeting pairs of allele specific polymorphisms

A) Common shared targetable variant pairs for SpCas9 and SaCas9 vary greatly by population, as demonstrated in the gene *BEST1* including the 5kb flanking regions in the five 1000 Genomes superpopulations. B) AlleleAnalyzer optimizes gRNA pair combinations to best cover a cohort, which performs much better compared to the naïve approach of selecting the most highly shared pairs ("Top 5"). C) The pairs identified by the AlleleAnalyzer and Top 5 approaches demonstrate disparate patterns of sharing among the entire 1KGP population. Height of the arcs is only for visualization purposes, and is not otherwise meaningful. D) AlleleAnalyzer designs gRNAs, colored by Cas variety, here with SpCas9 represented by purple and SaCas9 by green. E) For each variant, or gRNA site, multiple gRNAs can be designed on both the reference and alternate alleles, depending which is to be targeted. Each gRNA, then, has its own set of off-target sites, predicted using the incorporation of the CRISPOR tool in AlleleAnalyzer.

treat genetic diseases in large groups, potentially eliminating the need to develop new gRNA pairs for each patient or cell line, with practical implications for the development of genome surgery as a field. Our algorithm can also be used to identify gRNA pair

combinations applicable to a custom cohort; this enables researchers to design guides that are maximally shared among multiple cell lines, for example, which would improve experimental efficiency. Optimized gRNAs can then be validated for each individual via targeted genotyping, reducing sequencing and gRNA synthesis costs.

As a case study, I investigated the feasibility of excising at least one coding exon of bestrophin 1 (*BEST1),* which can cause dominant negative macular degeneration (Yang et al., 2015). Because mutations in this gene can cause macular degeneration by a dominant negative mechanism, a strategy that eliminates or silences the disease allele would be therapeutically desirable. Considering the gene plus 5 kb of flanking sequence on either side, and allowing 10 kb between each gRNA in a pair, there are 563 pairs of allele-specific gRNA sites for SpCas9 that are shared by >10% of all 1KGP individuals, with the number and composition of these pairs varying across 1KGP populations (Figure 3a). The goal was to identify an optimal combination of five allele-specific gRNA pairs to potentially target the majority of the 1KGP cohort. The result was that a combination of five allele-specific gRNA pairs could putatively excise at least one coding allele of *BEST1* while leaving the other allele intact in ~78% of the overall 1KGP population. This compares to only 48% that would be covered by the naïve approach of selecting a combination of the top 5 most highly shared pairs (Figure 3b, c). At each gRNA site, multiple gRNAs are possible for both the reference and alternate alleles (Figure 3d) depending on which is being targeted in the research subject. Each of these gRNAs has a unique off-target profile (Figure 3e), which we identified by integrating the tool CRISPOR into AlleleAnalyzer (Haeussler et al., 2016). Previous studies have

predicted that genetic variation may have a large impact on the off-target landscape (Lessard et al., 2017; Scott and Zhang, 2017). One of these produced a set of "platinum" gRNAs for all coding genes identified based on the target sites having low genetic variation and predicted off-targets, including off-targets generated by genetic variation (Scott and Zhang, 2017). Using the WTC genome, I compared these gRNAs to those produced by AlleleAnalyzer in the gene proprotein convertase subtilisin/kexin type 9 (*PCSK9),* a gene involved in various cardiovascular diseases and susceptibility to HIV infection (Dixon et al., 2016).

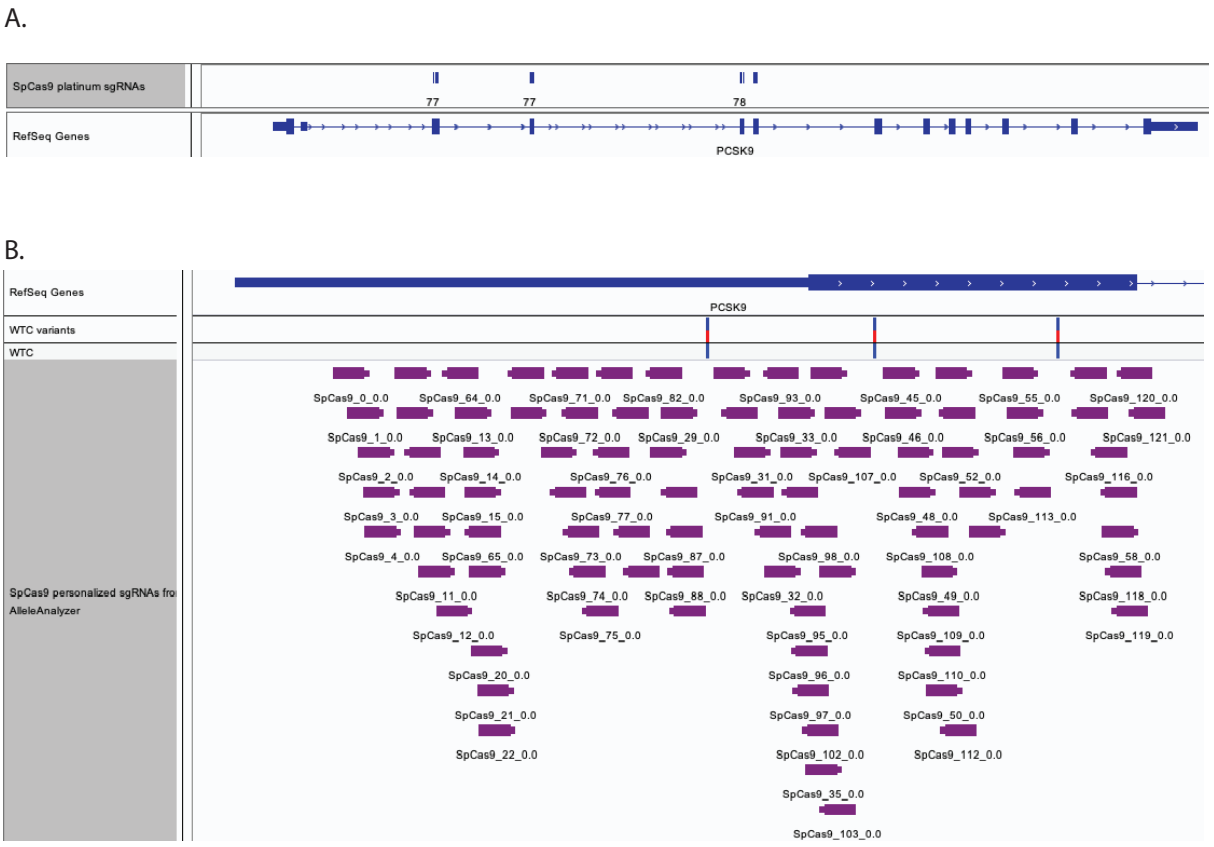

Figure 8: Comparison of gRNAs from AlleleAnalyzer to platinum gRNAs at PCSK9

Demonstration of the ability of AlleleAnalyzer to design gRNAs where other tools cannot due to incapability to incorporate genetic variants.

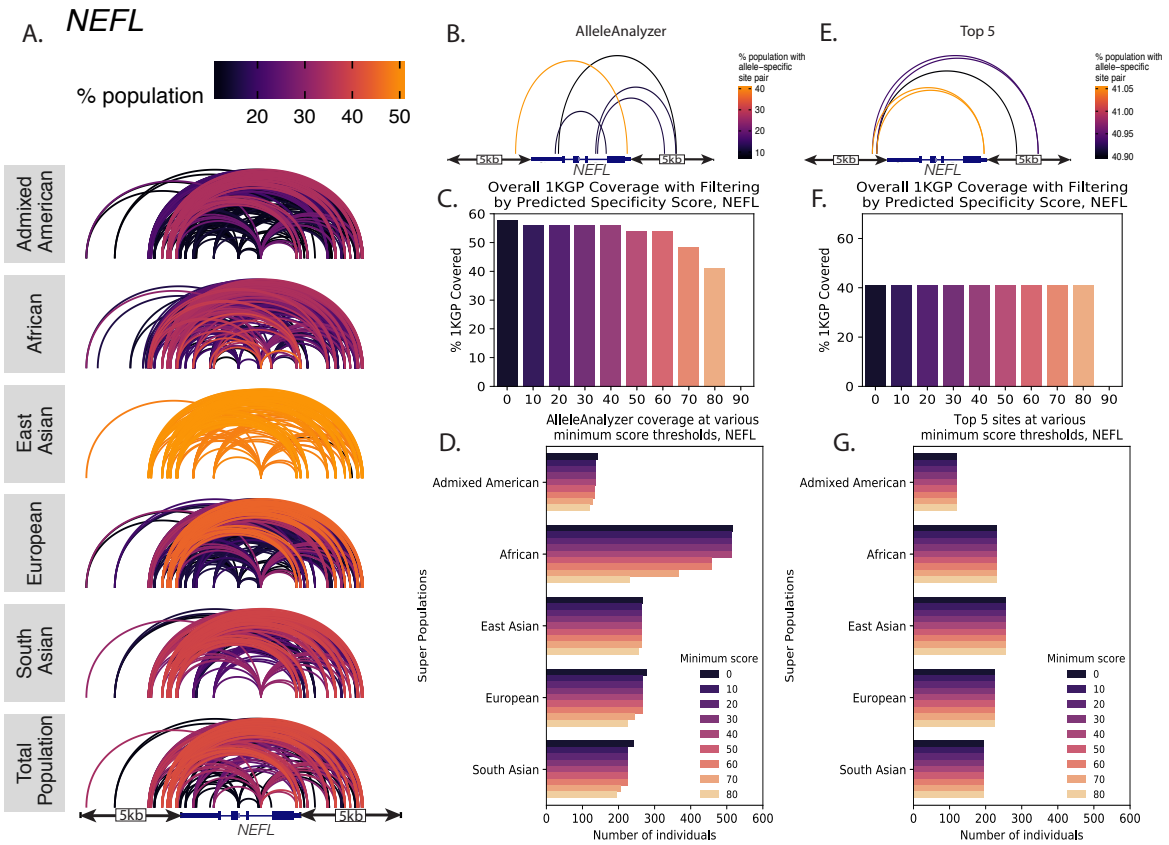

Figure 9: gRNA pair optimization for coverage of groups

A.) Variant pairs in *NEFL* and the flanking 5kb that are shared by at least 10% of the 1KGP cohort. These are pairs of variants, not pairs of gRNAs, so reflect potential dual-guide editing sites prior to designing or filtering gRNAs. 10% was chosen for visualization purposes. B.) 5 variant pairs identified by AlleleAnalyzer to achieve greatest possible coverage of the 1KGP cohort. C.) Coverage of the 1KGP cohort with the AlleleAnalyzer set of 5 pairs at various minimum predicted specificity score thresholds. D.) Coverage of each super population in the 1KGP cohort with the AlleleAnalyzer set of 5 pairs at various minimum predicted specificity score thresholds. E.) 5 top shared variant pairs in the 1KGP cohort. F.) Coverage of the 1KGP cohort with the "Top 5" set of pairs at various minimum predicted specificity score thresholds. G.) Coverage of each super population in the 1KGP cohort with the "Top 5" set of pairs at various minimum predicted specificity score thresholds.

I determined that the set of platinum gRNAs indeed has high predicted sensitivity and specificity in WTC, but some loci lack platinum gRNAs; AlleleAnalyzer is able to design personalized gRNAs in these loci, making it a flexible option that we expect will be

useful in practice (Figure 8). CRISPOR specificity scoring will be robust to most variation as it searches for all similar sites in the genome to an gRNA with up to four mismatches. Additionally, the predictive power of these scores is low in general (Haeussler et al., 2016).

AlleleAnalyzer allows the user to filter gRNAs for predicted specificity, and doing so can impact relative coverage using either the AlleleAnalyzer or top 5 pairs methods, as demonstrated here in six therapeutically relevant genes including neurofilament light gene (*NEFL)* (**Error! Reference source not found.**), a gene in which dominant negative mutations can cause Charcot-Marie-Tooth disease (Miltenberger-Miltenyi et al., 2007). Therefore, particularly in cases of therapeutic development, we recommend rigorous experimental whole-genome off-target analysis. Together, these results demonstrate important considerations for allele-specific gRNA design.

2.3 Open-source software tool for genetic-variation-aware gRNA design

The bioinformatics methods from this study have been implemented in AlleleAnalyzer, an open-source Python software tool (**Error! Reference source not found.**). This tool designs personalized and allele-specific gRNAs for unique individuals and cohorts, given their genetic variants, and optimizes gRNA pairs to cover many individuals based on shared variants. To our knowledge, this is the first computational resource that designs personalized and allele-specific CRISPR gRNAs. AlleleAnalyzer accounts for single nucleotide variants and short insertions and deletions, and currently supports eleven Cas proteins while

providing user options to add new Cas proteins, thus expanding and building upon the existing repertoire of gRNA design tools. The AlleleAnalyzer toolkit and tutorials are

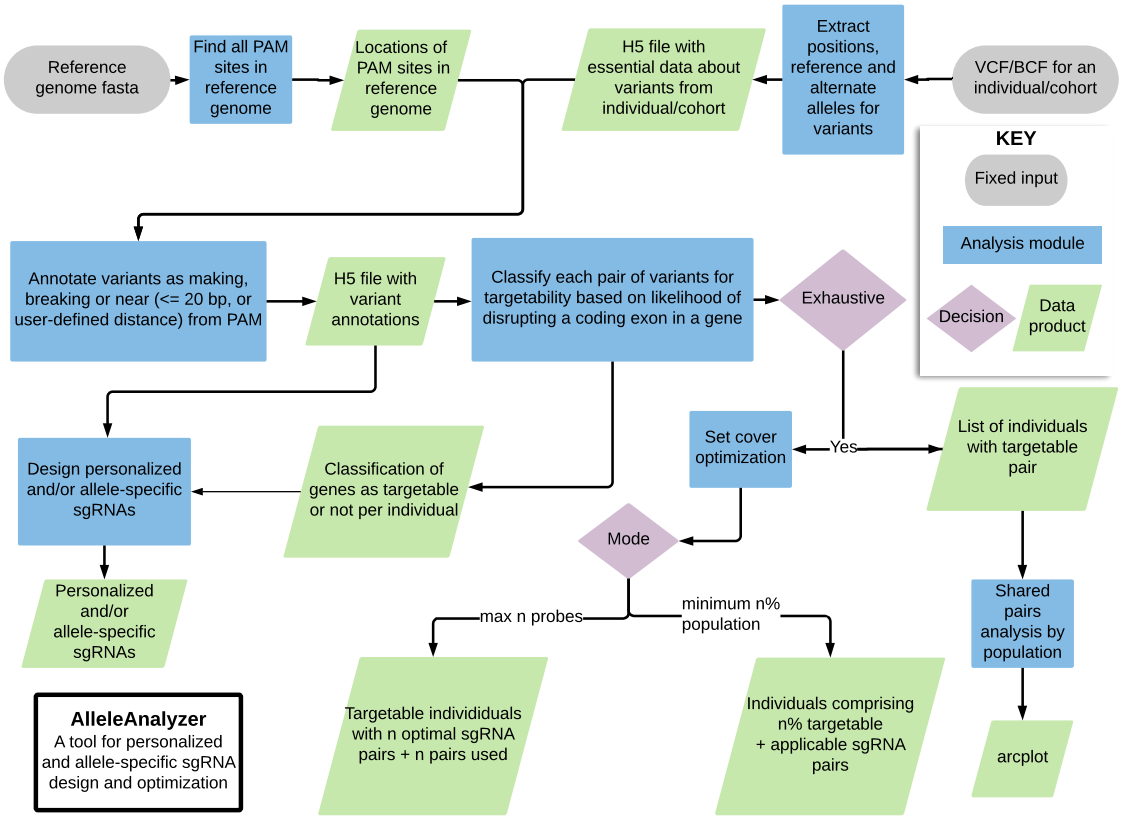

Figure 10: AlleleAnalyzer tool overview

Overview of designing gRNAs with AlleleAnalyzer.

available along with the database of annotated 1KGP variants at: https://github.com/keoughkath/AlleleAnalyzer under the MIT license (DOI: 10.5281/zenodo.3354488)

2.4 Methods

PAM occurrence in the human reference genome

PAM frequency

The AlleleAnalyzer tool includes a script enabling scanning of a reference genome fasta file for existing PAM sites. This was used to identify PAM sites for 11 Cas types (Figure 4, **Error! Reference source not found.**) in the reference human genomes hg19 and hg38. These are viewable in publically accessible UCSC Genome Browser sessions (hg19: https://bit.ly/2GB9cXK, hg38: https://bit.ly/2BZAmVh).

PAM size

PAM sizes were equated as the sum of non-N (A, C, G or T) bases in a PAM site. Thus "NGG" for SpCas9 would have size 2, and "NNGRRT" for SaCas9 would have size 4.

Analysis of variants in commonly used gRNA libraries

For each gRNA library, genomic coordinates for the protospacer regions were obtained from the relevant supporting manuscript. These were converted into BED files including the protospacer and PAM sites. Bcftools(Danecek et al., 2014) then was used to extract variants with a minor allele frequency (MAF) > 5% from the 1000 Genomes data, or variants from WTC with no MAF restriction. Variants that fell in the "N" position of the PAM were removed.

AlleleAnalyzer analyses

Annotation of variants

Genetic variants were determined to generate or destroy an allele-specific gRNA site if they were proximal to or in a PAM site (Figure 1a). Sufficient proximity to a PAM site was defined for this study as 20 base pairs based on the common length of gRNA recognition sequences. For all Cas varieties this was the 20 base pairs 5' of the PAM, except for cpf1 (Cas12a) for which it was 3' of the PAM. The gRNA design tools that are part of AlleleAnalyzer allow different user-defined gRNA lengths and addition of Cas enzymes and PAMs. There is evidence to suggest that genetic variants that generate or destroy a PAM are more likely to lead to allele-specific Cas activity compared to those in the seed sequence (Doench et al., 2014); AlleleAnalyzer thus provides options to differentiate between CRISPR sites in a PAM site versus the gRNA recognition sequence. All variants genome-wide were annotated for the 1KGP cohort for reference genomes hg19 and hg38. All variants in the ExAc dataset were annotated for the reference genome hg19 only, as that dataset is not available in hg38.

Generation of gene set

The analyzed gene set was compiled using the canonical transcripts for RefSeq gene annotations for human reference genome hg19 and hg38 downloaded using the UCSC table browser(Karolchik et al., 2004). Values reported in the text are for hg19 unless stated otherwise, but 1KGP analyses were conducted for both reference genomes with similar results.

Allele-specific putative gene targetability genome-wide

Putative allele-specific targetability of a gene is defined here as whether a gene contains a pair of allele-specific gRNA sites for at least one of the 11 Cas enzymes evaluated that are less than 10 kb apart on the same haplotype in an individual that will disrupt a coding exon (Figure 3). This metric was calculated for each gene for all 2,504 1KGP individuals. It was not calculated for the ExAc cohort as that dataset contains only exome rather than whole-genome variants.

Set cover analysis

In order to find the optimal set of gRNAs, two vectors of indicator variables were initialized that are constrained to be binary, one for gRNAs and one for individuals. When these indicator variables are set to 1, this means a gRNA is chosen or a person is covered, respectively. The objective function was specified to maximize the sum of person indicator variables. Next, the constraint was set on maximum value allowed for the sum of gRNA indicator variables. Finally, the constraints deduced from the data were assembled into the bipartite graph of gRNAs and patients targetable by them. This graph gets translated to multiple inequality constraints that specify that if a person indicator is 1, then at least one of its connected gRNA indicators must also be 1. Having specified all these elements of the problem, one may solve it with any number of integer linear programming solvers; here the Python package PuLP was used (Mitchell et al., 2011). The final values of the indicator variables were extracted from the solution with

the set of gRNAs that fulfill the chosen objective. The specific python implementation of the constraints and objective function and subsequent call to an integer linear programming solver can be seen in the GitHub repository for this tool. This is visualized in Figure 3.

Comparison of AlleleAnalyzer to platinum gRNAs from Scott & Zhang 2017

Platinum gRNAs for SpCas9 were obtained from the supplementary materials of their paper (Scott and Zhang, 2017). Personalized non-allele-specific gRNAs were designed for *PCSK9* exon 1 in WTC using AlleleAnalyzer. This analysis was done in reference genome hg19.

WTC sequencing

The genome for the iPSC line WTC (Drubin and Hyman, 2017) was sequenced by the Allen Institute for Cell Science. Analysis and variant calls in the reference genome hg19 were done according to GATK version 3.7 best practices (Van der Auwera et al., 2013) and phased using Beagle version 4.1 with default settings (Browning and Browning, 2007).

WTC targetability analysis

Variant annotation procedures were the same as in the 1KGP analysis and ExAc.

Packages used

Python

Docopt was used for handling of command-line arguments. Pandas (McKinney, 2010) version 0.21.0 and NumPy (Stéfan van der Walt, 2011) version 1.13.3 and elements of the standard Python distribution sys, os, and regex were used for multiple aspects of data analysis. PuLP (Mitchell et al., 2011) version 1.6.8 was used for set cover analysis. PyTables (Francesc Alted) was used for data management. Biopython (Cock et al., 2009) and pyfaidx (Shirley et al., 2015) were used for Fasta processing. Scripts from CRISPOR (Haeussler et al., 2016) were integrated into AlleleAnalyzer to facilitate specificity scoring of gRNAs. Seaborn (Waskom et al., 2018) and matplotlib (Hunter, 2007) were used for plotting.

R

Packages used to generate arcplots included viridis version 0.5.1, viridisLite version 0.3.0, igraph version 1.1.2, ggraph version 1.0.0, ggplot2 version 2.2.1, reshape2 version 1.4.3, dplyr version 0.7.4, tidyr version 0.7.2, and readr version 1.1.1.

Bioinformatics

Bcftools version 1.9 were used to manipulate VCF and BCF files.

Code Availability and Scripts

All data processing and analysis scripts as well as the gRNA design tool are located at github.com/keoughkath/AlleleAnalyzer, available under the MIT license (DOI: 10.5281/zenodo.3354488). Scripts were written in Python version 3.6.1, R version 3.3.2 and Bash version 3.2.57.

Availability of Data and Materials

1KGP phase 3 data were downloaded from the 1KGP website (http://www.internationalgenome.org/). ExAc data were downloaded from the ExAc website (http://exac.broadinstitute.org/). The reference hg19 and hg38 genome data were downloaded from the UCSC genome browser. The 1KGP and ExAc analysis datasets have been made available for public access online at UCSF Dash (https://datashare.ucsf.edu/stash/dataset/doi:10.7272/Q63F4MSR). Additionally, PAM sites identified in reference genomes hg19 and hg38 are viewable in UCSC Browser sessions (hg19: https://bit.ly/2GB9cXK or https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A61717368%2D61717468&hgsid=743058527_XLIEJrwnSVsZQLgeXUfU7NKQWeNn
, hg38: https://bit.ly/2BZAmVh or https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A61957117-

61957165&hgsid=710108079_SecTcyDrgBPU4AocIPTRF2Uq4Omd). WTC whole-genome sequencing data is made available by the Allen Institute at (https://www.allencell.org/genomics.html). In addition to the Github repository for AlleleAnalyzer (github.com/keoughkath/AlleleAnalyzer, available under the MIT license), an archived release of the software is available under DOI:10.5281/zenodo.3354488 provided through Zenodo.

3 Investigation of the driving forces behind accelerated evolution in humans

Comparing genomic sequences within and among species enables the identification of regions of the genome that are conserved or accelerated, indicating what types of selective pressures may be acting on those loci (Pollard et al., 2010). In addition to defining evolutionary pressures causing evolution of protein-coding genes, evidence of selective pressures in the noncoding genome help us develop theories about the various functions of DNA that does not produce proteins. Elements that are conserved in the noncoding genome can indicate the presence of functional elements such as enhancers, promoters, noncoding RNAs, transcription factor binding sites and motifs important for the proper folding of the genome (for a review of these elements, see (Chatterjee and Ahituv, 2017)). Accelerated evolution can help us identify where on a phylogenetic tree the selective pressure has changed on these elements, which can give us insight on how these features contribute towards the unique features of each species.

Evolution happens both at the sequence and the organizational level, as evidenced by differences in 3D genomic structure between various species, such as human and closely related primates (Eres et al., 2019). These changes in genome structure can rewire regulatory circuits, for example by placing enhancers in contact with genes they did not previously regulate, termed “enhancer hijacking” (Northcott et al., 2014). Within humans, enhancer hijacking has been demonstrated to cause various types of polydactyly and be involved in cancer (Lupiáñez et al., 2016). In this chapter, I describe various projects I contributed to relating to genome evolution in 1D (sequence-based) and 3D (structure-based).

3.1 Human accelerated regions in psychiatric disease

Human accelerated regions (HARs) are genomic loci that are conserved in many species but demonstrate uniquely accelerated sequence evolution in humans (Franchini and Pollard, 2017; Hubisz and Pollard, 2014; Pollard et al., 2006a). Many of these loci function as enhancers, while some function as noncoding RNAs and other have currently unknown functions. During my tenure as a graduate student, I contributed to multiple projects exploring the function of these regions, and initiated a new project to investigate the genomic forces driving their accelerated sequence evolution.

In one project, I contributed to we used massively parallel reporter assays (MPRAs) to assess the enhancer activity of hundreds of HARs by attaching them to reported genes and infecting them into neural progenitor cells from human and chimpanzees (Ryu et

al., 2018). In order to predict putative target genes for HARs that demonstrated strong enhancer activity I used Hi-C data from biologically relevant cell types (human fetal brain germinal zone and cortical plate tissue (de la Torre-Ubieta et al., 2018)) to identify genes in the same topological associating domains (TADs) with HARs. TADs are genomic regions that interact highly with each other in 3D space (Dixon et al., 2012; Nora et al., 2012). This information was combined with RNA sequencing (RNA-seq) data in order to assign putative activate cell types for HARs based on expression of genes in the same TAD in different cell types. GWAS SNPs were also assigned to HARs based on co-occurrence in the same TADs, thereby providing new hypotheses about potential roles for HARs in the associated disease. This approach enabled more informed discovery of target genes, whereas previous approaches relied on “nearest gene” approach, although we now know that many enhancers do not act on their nearest gene.

Through these combined analyses, I discovered that 2XHAR.170 contains a SNP associated with schizophrenia by genome-wide association study (GWAS), a technique that identifies genomic variants significantly associated with a particular phenotype in a large cohort of individuals. Because HARs are highly conserved for the most part among humans, it is unlikely to find many variants that occur at a high enough allele frequency so as to be genome-wide significant for a disease association. Thus, finding rs2434531, a variant that is associated with schizophrenia in the sequence of the 2xHAR.170, is notable, even though at a p-value of 7.47e-8, this SNP just misses the cutoff for genome-wide significance (Pouget et al., 2016). Notably, rs2434531 is in

an LD block with a genome-wide significant SNP, rs11740474 (Forrest et al., 2017). SNP rs2434531 has minor allele frequency ~23% in 1000 Genomes and TOPMED, with the human-derived nucleotide (C) being more common than the ancestral (i.e., matching chimp) nucleotide (T). This HAR resides in the first intron of the gene *GALNT10*. Beyond its accelerated evolution, 2xHAR.170 displays other markers of a potential enhancer in many cell types, including neuronal, based on ChromHMM (Ernst and Kellis, 2017). It also binds FOXP2, a transcription factor widely associated with schizophrenia. 2xHAR.170 drives significantly higher expression with the human as opposed to the chimp sequence in our assay. This matches results from eQTL studies of rs2434531 for *GALNT10* [http://eqtl.rc.fas.harvard.edu/eqtlbrowser/mrcau133list/19443]. Interestingly, *GALNT10* has been shown to be significantly more highly expressed in cases of schizophrenia compared to controls (Voisey et al., 2017). Therefore, while these relationships may indicate a previously unknown link between 2xHAR.170 and schizophrenia, it is clear that the link is complex, consistent with its associated disease.

3.2 Methodological considerations for calling HARs in an era of many genomes

HARs were first identified using a multiple alignment of 29 vertebrate genomes (Pollard et al., 2006a, 2006b). Since then, new sequencing technologies have led to an explosion in available species genomes, enabling the assessment of factors in the pipeline that impact which genomic loci are identified as HARs. Indeed, HARs have been identified with multiple pipelines and datasets, and tend to show low levels of

overlap. I assembled a pipeline to call HARs that enables easy tuning of parameters and analysis of the impact of various decision made throughout the pipeline.

Using a 100-way vertebrate alignment from UCSC, I implemented the HAR pipeline laid out in the original HAR publication using Nextflow, a pipeline management software that enables greater modularity and reproducibility for multi-step analyses (Di Tommaso et al., 2017; Pollard et al., 2006a, 2006b). This pipeline was applied to a 100-way alignment of vertebrate species from UCSC. When applied to all 100 species in the alignment, 7,153 HARs were found. However, it is known that errors in assembly can be influential when identifying accelerated sequence evolution, because assembly errors and acceleration can look similar. Therefore, using assembly quality metrics and optimizing representation throughout the vertebrate tree, I defined a set of fifty-two "high-quality" sequences and ran the HAR identification pipeline on an alignment filtered for only those species. This resulted in a smaller set of 110 HARs (Benjamini-Hochberg-corrected p-value < 0.01) that demonstrated greater proportional overlap with previous sets (Figure 11: HAR set comparisons), such as from an analysis of 29 mammals (Broad Institute Sequencing Platform and Whole Genome Assembly Team et al., 2011)

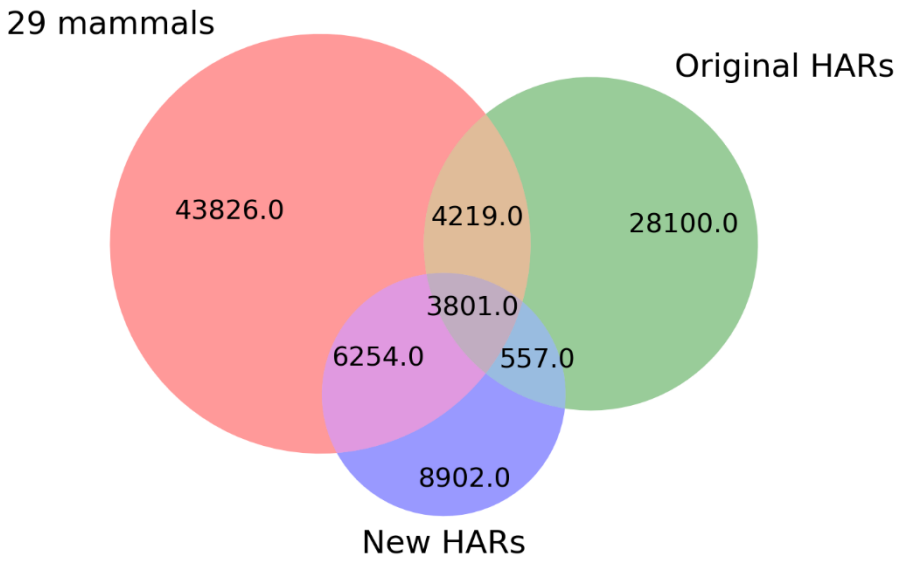

Figure 11: HAR set comparisons

Analysis of basepairs of overlap between various HAR sets.

and the original HAR set (Pollard et al., 2006a, 2010). Additionally, a greater proportion of the HARs from high-quality species assemblies had evidence for positive selection using a method optimized for analysis in the noncoding genome, indicating that I may have filtered out more instances of loss of negative selection by focusing on species with higher quality data (Kostka et al., 2012). Therefore, I learned that sequence and assembly quality are important in identifying regions undergoing species-specific accelerated evolution, and identified an updated set of HARs along with principles to guide identification of sequence evolution when the number of genomes is not the limiting factor.

3.3 Changes in the 3D genome may influence the evolutionary rate of HARs

Structural variants can change the 3D structure of the genome, for example by removing TAD boundaries. This in turn can rewire regulatory networks by generating interactions between enhancers and genes that had not previously interacted. This phenomenon, termed “enhancer hijacking”, has been implicated in various cancers and

polydactyly-related diseases (Lupiáñez et al., 2016). Structural variation also occurs over evolutionary time, inserting, rearranging, and removing loci, and meanwhile changing the 3D and regulatory networks of the genome. Recent work generated high quality assemblies for chimpanzee and orangutan, and compared these along with a high-quality gorilla genome to the human genome, identifying many human-specific insertions and deletions (Kronenberg et al., 2018).

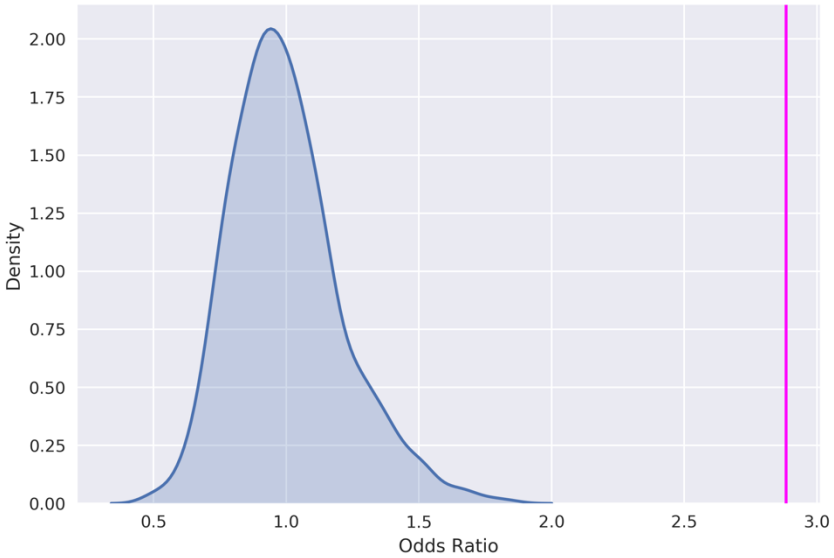

Figure 12: HAR enrichment in TADs with human-specific SVs

Blue shaded region indicates the null distribution generated by calculating the odds ratios of the number of HARs sized sets of phastCons elements co-occurring in the same TAD with human-specific SVs, randomly selected 1000 times. The magenta line indicates the actual odds ratio for HARs.

Our hypothesis is that in some cases human-specific structural variants put human accelerated regions in contact with genes they did not previously regulate that were important for human-specific traits, thus generating the selective pressure to induce accelerated rates of evolution at these loci. To investigate this, I assessed whether TADs that contain human-specific structural variants are enriched for HARs. Using TADs called in N2-neural progenitor cells, I compared the odds ratios of HARs being in TADs with human-specific SVs compared to that of sets of randomly drawn phastCons elements equal to

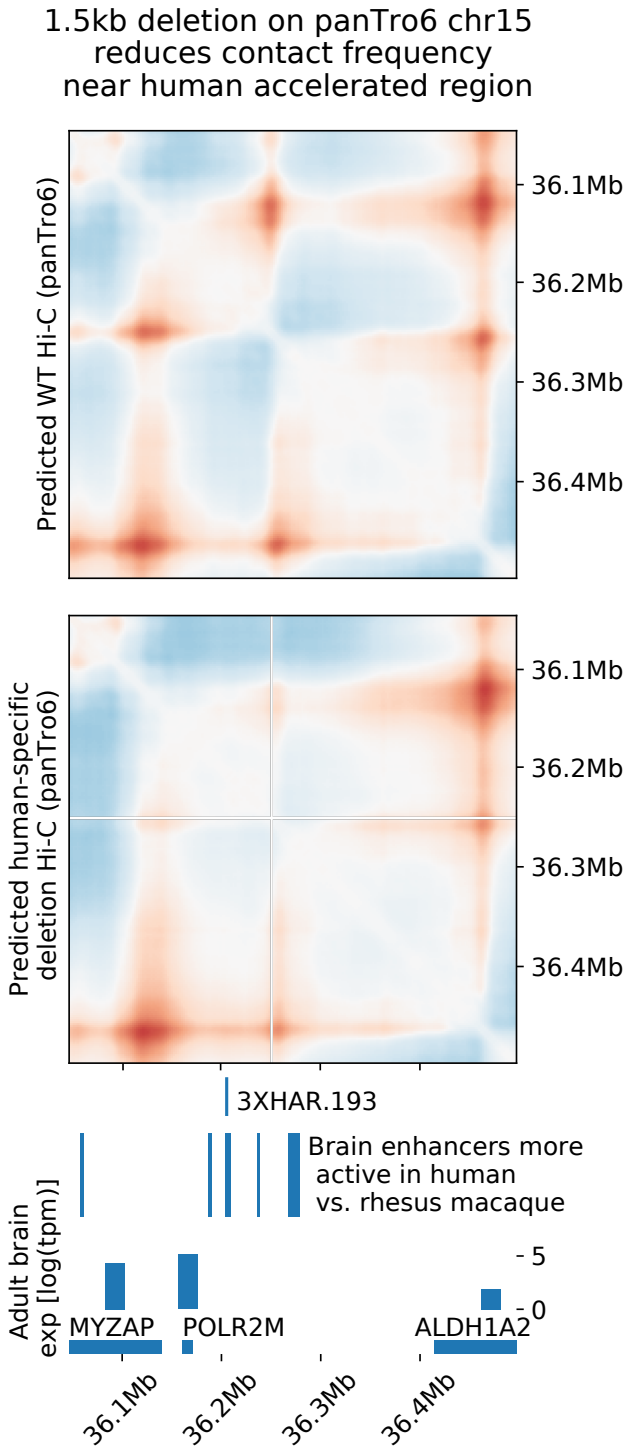

Figure 13: Predicted impact of a human-specific SV on 3D genome conformation

Akita-predicted 3D genome changes due to a human-specific deletion in the chimpanzee genome, with enhancer-related epigenetic and gene data for the locus.

the number of HARs, drawn 1000 times as a null model. Based on this, I found that TADs with human-specific structural variants are significantly enriched for HARs (Figure 12). This supports our hypothesis that human-specific SVs and HARs may be impacting each other.

3.4 Human-specific SVs near HARs are predicted to alter 3D genome structure

Our hypothesis that a similar mechanism to enhancer hijacking may have contributed to the acceleration of HARs depends on human-specific SVs altering 3D genome structure, for instance by removing or altering TAD boundaries. This is difficult to study via comparative analysis in relevant tissues and cell types in human and chimpanzee due to scarcity of samples and technical challenges, although this

is something we are working towards. However, deep learning models have recently been developed that are capable of predicting 3D genome structure based on DNA sequence alone. One of these models, Akita, was developed in the Pollard lab and has been able to accurately reconstruct the impact of SVs in the human genome (Fudenberg et al., 2019). Using this model, I was able to identify human-specific SVs that are predicted to significantly alter 3D genomes structure near a HAR. For example, 3XHAR.193 is located near a human-specific deletion on chromosome 15. Using Akita, I predicted and visualized the impact of this deletion in the chimpanzee genome, finding that the deletion is predicted to disrupt a sub-TAD structure within a larger TAD, decreasing its contact frequency to that of the surrounding TAD (Figure 13). This disrupts chromatin patterns that previously insulated 3XHAR.193 from nearby genes, such as aldehyde dehydrogenase 1 family member A2 (*ALDH1A2*), a gene with some expression in the adult human brain (GTex), and important for neural tube development. This HAR has been predicted to be a brain enhancer and has epigenetic marks consistent with higher enhancer activity in human relative to rhesus macaque (Capra et al., 2013; Vermunt et al., 2016). Additionally, our comparative analyses indicate this HAR as being under positive selection, supporting the idea that this locus gained increased enhancer activity specifically in the human lineage, with potential beneficial effects towards human brain development.

Overall, more work is needed to prove or refute the hypothesis of a role for human-specific SVs or enhancer hijacking in HAR acceleration, but these analyses lay the groundwork for future hypothesis testing and experimentation.

4 Comparative genomics to identify the host range for SARS-CoV2

The work described in this chapter is currently available as a preprint on BioRxiv and is under peer review (Damas et al., 2020). I was able to apply the skills I had learned during projects focused mainly on human evolution and disease to co-lead a manuscript to predict host species for SARS-CoV2, the virus that upended the world beginning in 2019 and caused me to write this thesis entirely sequestered at home. My main contributions to the paper were the PHAST and phyloP-based selection analyses (Hubisz et al., 2011; Pollard et al., 2010; Ramani et al., 2019). The paper represented a sprint effort by 19 authors from universities, zoos, and research institutions all over the world.

The novel coronavirus SARS-CoV-2 is the cause of COVID-19, a major pandemic that threatens millions of human lives and the global economy, with infections now reported in other species as well. We identified a large number of mammals that can potentially be infected by SARS-CoV-2 through their ACE2 proteins, which can assist identification of an intermediate host(s) for SARS-CoV-2 and hence reduce the opportunity of a future outbreak of COVID-19. Among the highest risk species for SARS-CoV-2 infection using ACE2 are wildlife species and endangered species. These species represent an opportunity for spillover of SARS-CoV-2 from humans to other susceptible animals, and

should thus be a focus of surveillance and conservation efforts. The impact of this work is likely to inform COVID-19-related conservation efforts for endangered species, protective measures between humans and other species, and model-animal selection for COVID vaccine and therapeutic development.

4.1 Justification for a comparative analysis of *ACE2* in vertebrates

The 2019-novel coronavirus (2019-nCoV, also, SARS-CoV-2 and COVID-19 virus) is the cause of Coronavirus Disease-2019 (COVID-19), a major pandemic that threatens millions of lives and the global economy (Zhou et al., 2020). Comparative analysis of SARS-CoV-2 and related coronavirus sequences has shown that SARS-CoV and SARS-CoV-2 likely originated in bats, followed by transmission to an intermediate host, and that both viruses may have an extended host range that includes primates and other mammals (Lu et al., 2015; Shan et al.; Zhou et al., 2020). However, the immediate source population/species for SARS-CoV and SARS-CoV-2 viruses has not yet been identified. Several mammalian species host coronaviruses, and these infections are frequently associated with severe clinical diseases, such as respiratory and enteric disease in pigs and cattle (Laude et al., 1993; Saif, 2010). Molecular phylogenetics revealed that at least one human coronavirus (HCov-OC43), may have originated in cattle or swine (Chen et al., 2005), and that this virus was associated with a human pandemic that emerged in the late 19th century (Vijgen et al., 2005). Recent data indicate that coronaviruses can move from bats to other wildlife species and humans (Lam et al., 2020) and from humans to tigers (United States Department of Agriculture

Animal and Plant Health Inspection Service) and pigs (Qian et al., 2013). Therefore, understanding the host range of SARS-CoV-2 and related coronaviruses is essential for improving our ability to predict and control future pandemics. It is also crucial for protecting populations of wildlife species in native habitats and under human care, particularly non-human primates, who may also be susceptible to COVID-19 (Sun et al., 2020a).

The angiotensin I converting enzyme 2 (ACE2) serves as a functional receptor for the spike protein (S) of SARS-CoV and SARS-CoV-2 (Lan et al., 2020; Li et al., 2003). Under normal physiological conditions, ACE2 is a dipeptidyl carboxypeptidase that catalyzes the conversion of angiotensin I into angiotensin 1-9, a peptide of unknown function, and angiotensin II, a vasoconstrictor that is important in the regulation of blood pressure (Patel et al., 2016). ACE2 also converts angiotensin II into angiotensin 1-7, a vasodilator that affects the cardiovascular system (Patel et al., 2016) and may regulate other components of the renin-angiotensin system (Feng et al., 2008). The host range of SARS-CoV-2 may be extremely broad due to the conservation of ACE2 in mammals (Lan et al., 2020; Lu et al., 2015) and its expression on ciliated bronchial epithelial cells and type II pneumocytes (Qian et al., 2013). While coronaviruses related to SARS-CoV-2 use ACE2 as a primary receptor, coronaviruses may use other proteases as receptors, such as CD26 (DPP4) for MERS-CoV (Raj et al., 2013), thus limiting or extending their host range.

In humans, ACE2 may be a cell membrane protein or it may be secreted (Patel et al., 2016). The secreted form is created primarily by enzymatic cleavage of surface-bound ACE2 by ADAM17 and other proteases (Patel et al., 2016). Sequence variation in *ACE2* affects the protein's functions. *ACE2* is polymorphic in humans, with many synonymous and nonsynonymous mutations identified, although most are rare at the population level (Karczewski et al., 2020) and few are believed to affect cellular susceptibility to human coronavirus infections (Stawiski et al., 2020). Site-directed mutagenesis and co-precipitation of SARS-CoV constructs have revealed critical residues on the ACE2 tertiary structure that are essential for binding to the virus receptor binding domain (RBD) (Li, 2013). These findings have been strongly supported by co-crystallization and the structural determination of the SARS-CoV and SARS-CoV-2 S proteins with human ACE2 (Lan et al., 2020; Li et al., 2005; Shang et al., 2020), as well as binding-affinity with heterologous ACE2 (Li, 2013). The RBD of human coronaviruses may mutate to change the binding affinity of S for ACE2, and thus lead to adaptation in humans or other hosts. The best studied example is the palm civet, believed to have been the intermediate host between bats and humans for SARS-CoV(Lu et al., 2015). To date, an intermediate host for SARS-CoV-2 has not been identified definitively, although Malayan pangolins (*Manis javanica*) have been proposed as a possible reservoir (Zhang et al., 2020).

Comparative analysis of ACE2 nucleotide and protein sequences can predict their ability to bind SARS-CoV-2 S and therefore will yield important insights into the biology and potential zoonotic transmission of SARS-CoV-2 infection. Recent work has

examined ACE2 from different vertebrate species and predicted its ability to bind SARS-CoV-2 S, but phylogenetic sampling was extremely limited (Liu et al., 2020; Sun et al., 2020a). Here, we made use of sequenced genomes of 410 vertebrates and protein structural analysis, to identify ACE2 homologs in all vertebrate classes (fishes, amphibians, birds, reptiles, and mammals) that have the potential to serve as a receptor for SARS-CoV-2, and to understand the evolution of ACE2 SARS-CoV-2 S binding sites. Our results reinforce earlier findings on the natural host range of SARS-CoV-2, and predict a broader group of species that may serve as a reservoir or intermediate host for this virus. Importantly, many threatened and endangered species were found to be at potential risk for SARS-CoV-2 infection, suggesting that as the pandemic spreads, humans could inadvertently introduce a potentially devastating new threat to these already vulnerable populations, especially for great apes and other primates.

4.2 Comparison of vertebrate ACE2 sequences and their predicted ability to bind SARS-CoV-2 based on sequence and structure homology

We identified 410 unique vertebrate species with *ACE2* orthologs that included representatives of all vertebrate taxonomic classes. Among these were 252 mammals, 72 birds, 65 fishes, 17 reptiles and 4 amphibians. Twenty-five amino acids corresponding to known SARS-CoV-2 S-binding residues (Lan et al., 2020; Shang et al., 2020; Sun et al., 2020a) were examined for their similarity to the residues in human ACE2 (Figure 14; **Error! Reference source not found.**). On the basis of known

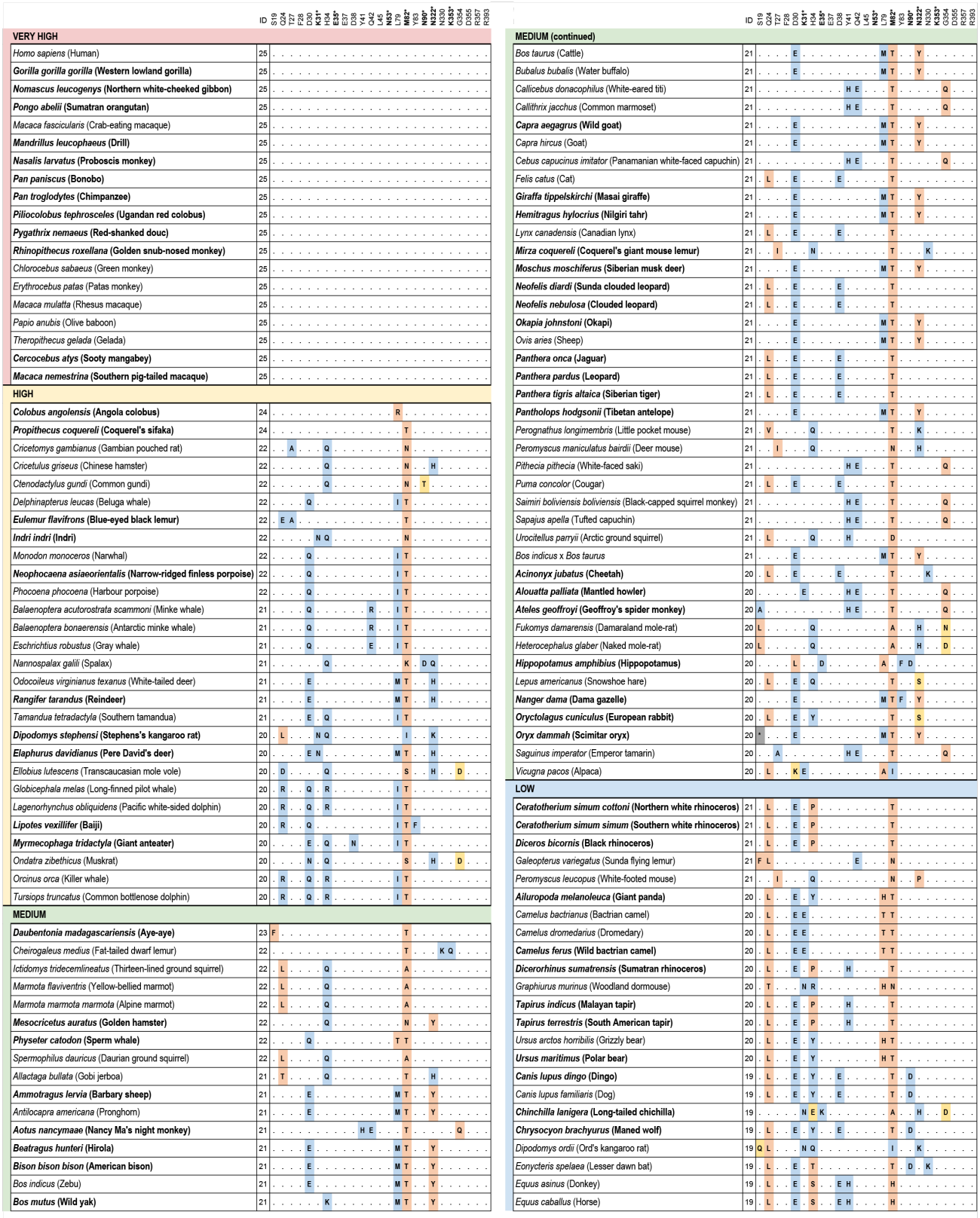

Figure 14: Protein sequence-based score high through low

Scores for each species based on homology between amino acids sequences for species scoring high through low, continued in Figure 14.

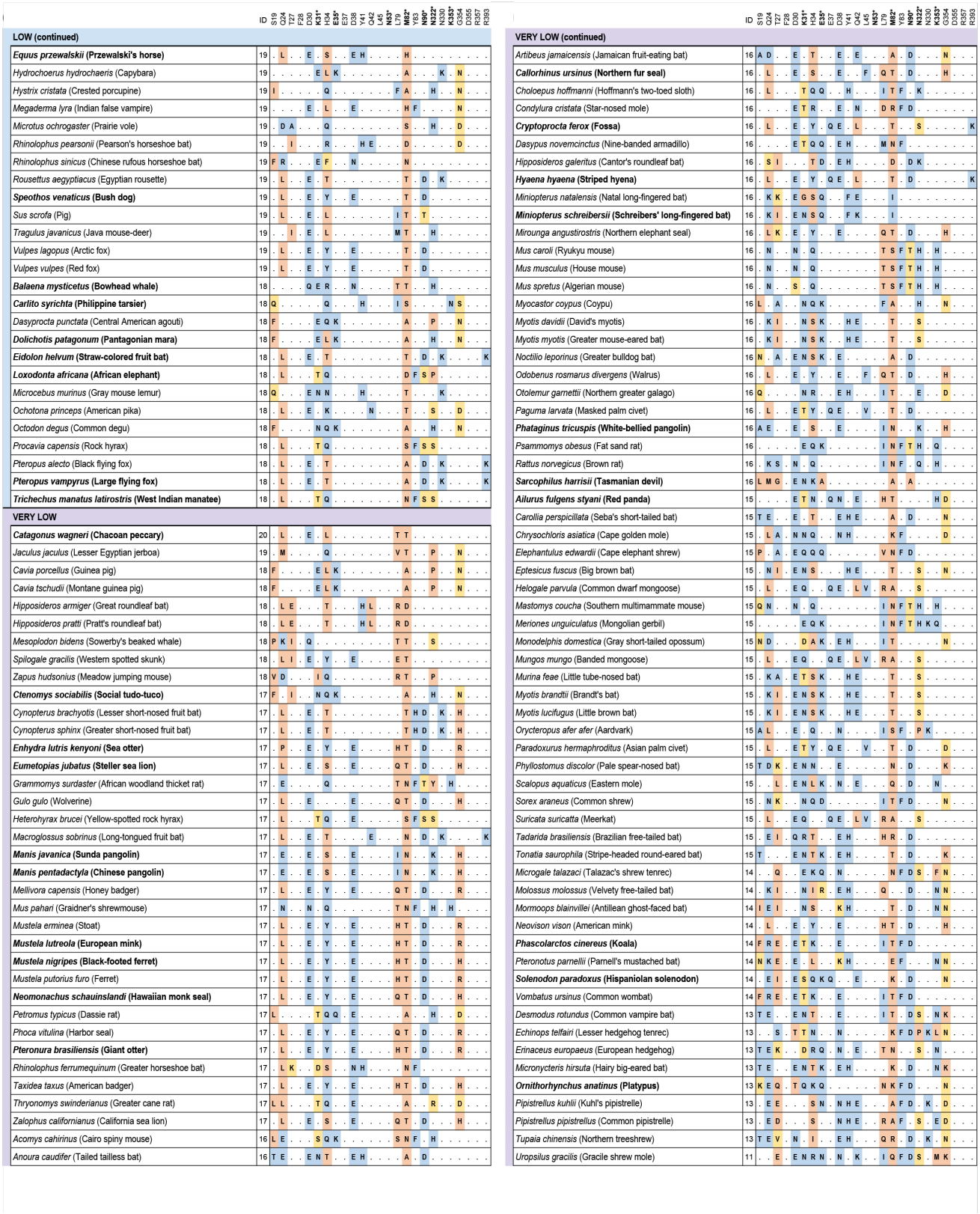

Figure 15: Protein sequence-based score low through very low

Scores for each species based on homology between amino acids sequences (continuation of Figure 14 for species scoring low through very low.

each species (see Materials and Methods). Five score categories were predicted: *very high*, *high*, *medium*, *low* and *very low*. Results for all species and all SARS-CoV-2 S binding scores are shown in Dataset S1, and results for mammalian species are also shown in Figure 14 and **Error! Reference source not found.**.

We complemented the sequence-identity based scoring scheme with a qualitative approach that combined structural homology modeling and best fit rotamer positioning.

	Species		Q24		T27		D30		K31		H34		E35		E37		D38		Y41		Q42		L45		L79		M82		Y83		K353		G354	
High	Odocoileus virginianus texanus	*White-tailed deer*	.	.	.	.	E	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	N	T	N	.	.	.	.	.	.
	Rangifer tarandus	*Reindeer*	.	.	.	.	E	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	N	T	N	.	.	.	.	.	.
	Eulemur flavifrons	*Blue-eyed black lemur*	E	N	A	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	N	.	.	.	.	.	.
	Propithecus coquereli	*Coquerel's sifaka*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	N	.	.	.	.	.	.
	Dipodomys stephensi	*Stephens's kangaroo rat*	L	U	.	.	.	.	N	W	Q	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	N	.	.	.	.	.	.
Medium	Bos taurus	*Cattle*	.	.	.	.	E	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	M	N	T	N	.	.	.	.	.	.
	Felis catus	*Cat*	L	U	.	.	E	N	.	.	.	.	.	.	.	.	E	N	.	.	.	.	.	.	.	.	T	N	.	.	.	.	.	.
	Panthera tigris altaica	*Siberian tiger*	L	U	.	.	E	N	.	.	.	.	.	.	.	.	E	N	.	.	.	.	.	.	.	.	T	N	.	.	.	.	.	.
	Mesocricetus auratus	*Golden hamster*	.	.	.	.	.	.	.	.	Q	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	N	U*	.	.	.	.	.	.
Low	Sus scrofa	*Pig*	L	U	.	.	E	N	.	.	L	U	.	.	.	.	.	.	.	.	.	.	.	.	I	N	T	N	.	.	.	.	.	.
	Ailuropoda melanoleuca	*Giant panda*	L	U	.	.	E	N	.	.	Y	U	.	.	.	.	.	.	.	.	.	.	.	.	H	W	T	N	.	.	.	.	.	.
	Canis lupus familiaris	*Dog*	L	U	.	.	E	N	.	.	Y	U	.	.	.	.	E	N	.	.	.	.	.	.	.	.	T	N	.	.	.	.	.	.
	Rhinolophus pearsonii	*Pearson's horseshoe bat*	.	.	I	N	.	.	.	.	R	N	.	.	.	.	.	.	H	W	E	W	.	.	.	.	D	U	.	.	.	.	D	U
	Dipodomys ordii	*Ord's kangaroo rat*	L	U	.	.	.	.	N	W	Q	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	I	N	.	.	.	.	.	.
Very Low	Catagonus wagneri	*Chacoan peccary*	L	U	.	.	E	N	.	.	L	U	.	.	.	.	.	.	.	.	.	.	.	.	T	W*	T	N	.	.	.	.	.	.
	Mustela putorius furo	*Ferret*	L	U	.	.	E	N	.	.	Y	U	.	.	.	.	E	N	.	.	.	.	.	.	H	W	T	N	.	.	.	.	R	U
	Paguma larvata	*Masked palm civet*	L	U	.	.	E	N	T	U	Y	U	.	.	Q	N	E	N	.	.	.	.	V	N	.	.	T	N	.	.	.	.	.	.
	Hipposideros armiger	*Great roundleaf bat*	L	U	E	U	.	.	.	.	T	U*	.	.	.	.	.	.	H	W	L	U*	.	.	R	W*	D	U	.	.	.	.	.	.
	Hipposideros galeritus	*Cantor's roundleaf bat*	S	U	I	N	.	.	.	.	T	U*	D	W*	.	.	E	N	H	W	.	.	.	.	.	.	D	U	.	.	.	.	.	.
	Hipposideros pratti	*Pratt's roundleaf bat*	L	U	E	U	.	.	.	.	T	U*	.	.	.	.	.	.	H	W	L	U*	.	.	R	W*	D	U	.	.	.	.	.	.
	Rhinolophus ferrumequinum	*Greater horseshoe bat*	L	U	K	U*	.	.	D	W	S	W*	.	.	.	.	N	N	H	W	.	.	.	.	.	.	N	U*	F	W	.	.	.	.
	Uropsilus gracilis	*Gracile shrew mole*	.	.	E	U	E	N	N	W	R	N	N	W	.	.	N	N	.	.	K	N/A	.	.	I	N	Q	U	F	W	M	W	K	U
	Manis javanica	*Sunda pangolin*	E	N	.	.	E	N	.	.	S	W*	.	.	.	.	E	N	.	.	.	.	.	.	I	N	N	U*	.	.	.	.	H	U
	Manis pentadactyla	*Chinese pangolin*	E	N	.	.	E	N	.	.	S	W*	.	.	.	.	E	N	.	.	.	.	.	.	I	N	N	U*	.	.	.	.	H	U
	Otolemur garnettii	*Northern greater galago*	.	.	.	.	.	.	N	W	R	N	.	.	.	.	E	N	H	W	.	.	.	.	I	N	T	N	.	.	.	.	D	U

Figure 16: Evaluation of binding contacts between host ACE2 and SARS-CoV-2

Evaluation of binding contacts between host ACE2 and SARS-CoV-2 in 28 representative species selected from *very low*, *low*, *medium* and *high* binding score groups, and for each residue in the ACE2 binding interface that varied from human (55 substitutions in 16 residues). For each residue, amino acid substitutions are shown on the left as white boxes, with sites matching human ACE2 shown in gray. For each residue, the evaluation of the binding contact is shown on the right as *neutral* (N; blue box), w*eakening* (W; orange box); or *unfavorable* (U; red box), with sites matching human ACE2 in blue. Evaluations discordant with Procko (5) are marked with an asterix and lighter background color.

We examined the 25 ACE2 binding residues in a subset of 28 representative species (Figure 16). First, we assessed the similarity of every contact at the binding interface between two recently solved crystal structures for the human ACE2/SARS-CoV-2 S RBD complex in humans, 6M0J and 6WV1 (Lan et al., 2020; Shang et al., 2020). We examined a total of 55 substitutions and assigned each to one of three types: *neutral* (N; likely to maintain similar contacts; 18 substitutions); *weaken* (W; likely to weaken the interaction; 14 substitutions); or *unfavorable* (U; likely to introduce unfavorable interactions; 23 substitutions). The structural homology binding assessments support the sequence identity analysis, with the fraction of residues ranked as U correlating very strongly with the substitution scoring scheme (Spearman correlation rho=0.76; p< 2.2e-16; Figure 17).

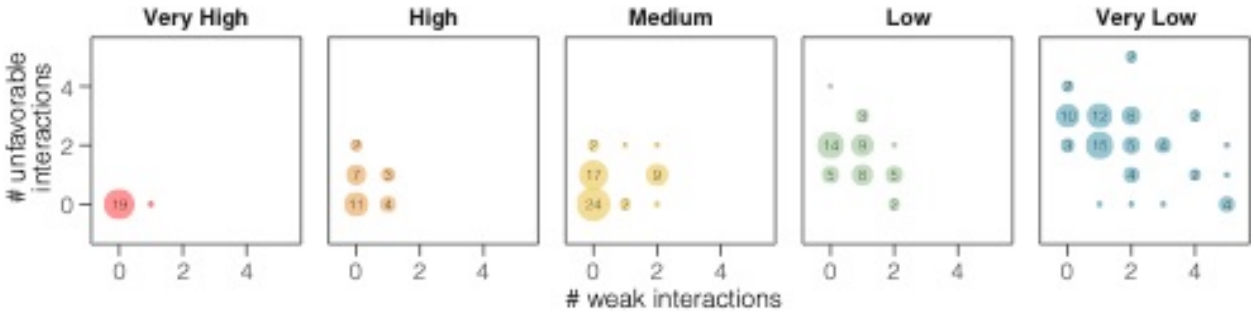

Figure 17: Congruence between binding score and structural homology analysis

Species classified by sequence identity to human ACE2 as *very high* (red) or *high* binding score (orange) have significantly fewer amino acid substitutions rated as potentially altering the binding interface between ACE2 and SARS-CoV-2 through protein structural analysis, as compared to *low* (green) or *very low* (blue) species. The more severe *unfavorable* variants are counted on y-axis and less severe *weaken* variants on the x-axis. Black numerical labels indicate species count.

4.4 Structural analysis of variation in human *ACE2*

We applied the same approach used to compare species, sequence identity and protein structural analysis, to examine the variation in ACE binding residues within humans,

some of which have been proposed to alter binding affinity (Cao et al., 2020; Hussain et al., 2020; Othman et al., 2020; Renieri et al., 2020; Stawiski et al., 2020). We integrated data from six different sources: dbSNP (Sherry, 2001), 1KGP (Voight et al., 2015), Topmed (NHLBI), UK10K (UK10K Consortium et al., 2015) and CHINAMAP (Cao et al., 2020), and identified a total of 11 variants in ten of the 25 ACE2 binding residues. All variants found are rare, with allele frequency less than 0.01 in any populations, and less than 0.0007 over all populations. Three of the 11 variants were synonymous changes, seven were conservative missense variants, and one, S19P, was a semi-conservative substitution. S19P has the highest allele frequency of the 11 variants, with a global frequency of 0.0003 (Karczewski et al., 2020). We evaluated, by structural homology, six missense variants. Four were *neutral* and two weakening (E35K, frequency=0.000016; E35D, frequency=0.000279799). Thus, with an estimated summed frequency of 0.001, genetic variation in the ACE2 S-binding interface is overall rare, and it is unclear whether the variation that does exist increases or decreases susceptibility to infection.

4.5 Evolution of ACE2 across mammals.

We next investigated the evolution of ACE2 variation in vertebrates, including how patterns of positive selection compare between bats, a mammalian lineage known to harbor a diversity of coronaviruses (Anthony et al., 2017), and other mammalian clades. We first inferred the phylogeny of *ACE2* using our 410-vertebrate alignment and IQTREE, using the best-fit model of sequence evolution (JTT+F+R7) and rooting the

topology on fishes. I assayed sequence conservation with PhyloP (36). The majority of ACE2 codons are significantly conserved across vertebrates and across mammals, likely reflecting its critical function in the renin-angiotensin system (Oudit et al., 2003), with ten residues in the ACE2 binding domain exceptionally conserved in Chiroptera and/or Rodentia.

We next used phyloP and CODEML to test for acceleration and positive selection with a co-author (Graham Hughes) leading the CODEML analyses and me leading the phyloP analyses (Pollard et al., 2010). PhyloP compares the rate of evolution at each codon to the expected rate in a model estimated from third nucleotide positions of the codon, and is agnostic to synonymous versus nonsynonymous substitutions (dN/dS). CODEML uses ω=dN/dS>1 and Bayes Empirical Bayes (BEB) scores to identify codons under

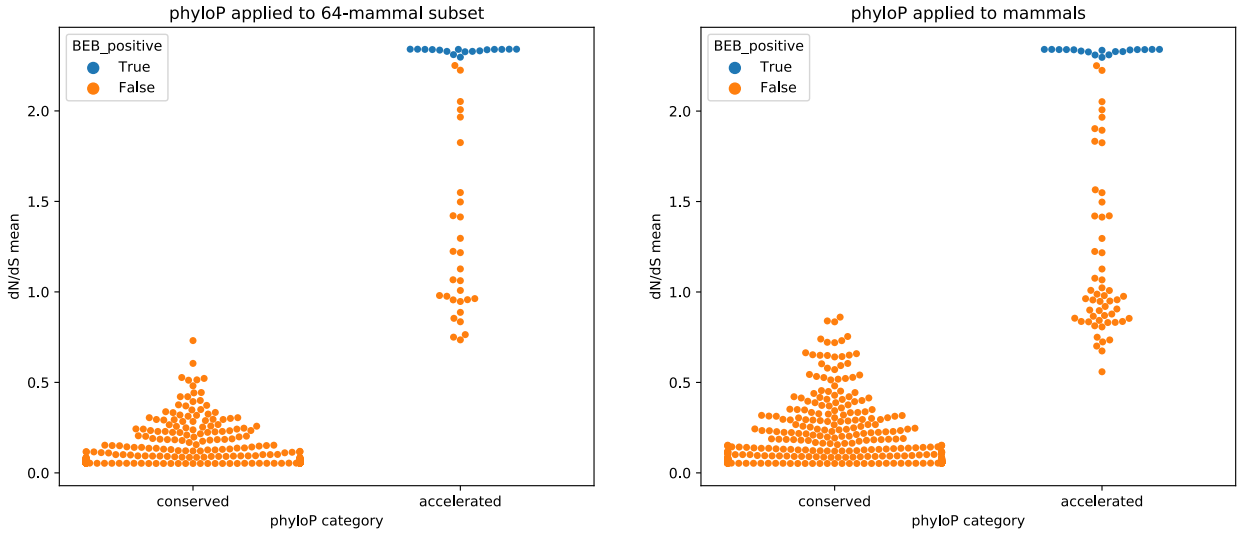

Figure 18: Significant results from phyloP, both conserved and accelerated, for ACE2 codons compared with CODEML BEB scores

Left panel shows phyloP results for the 64-mammals subset used in the mammal CODEML analysis. Right panel shows phyloP results for all mammals in the alignment. The y-axis represents dN/dS values calculated by CODEML, x-axis indicates whether the codons were classified as conserved or accelerated by phyloP. All dots are significant results from phyloP, blue dots are also significantly positively selected from CODEML.

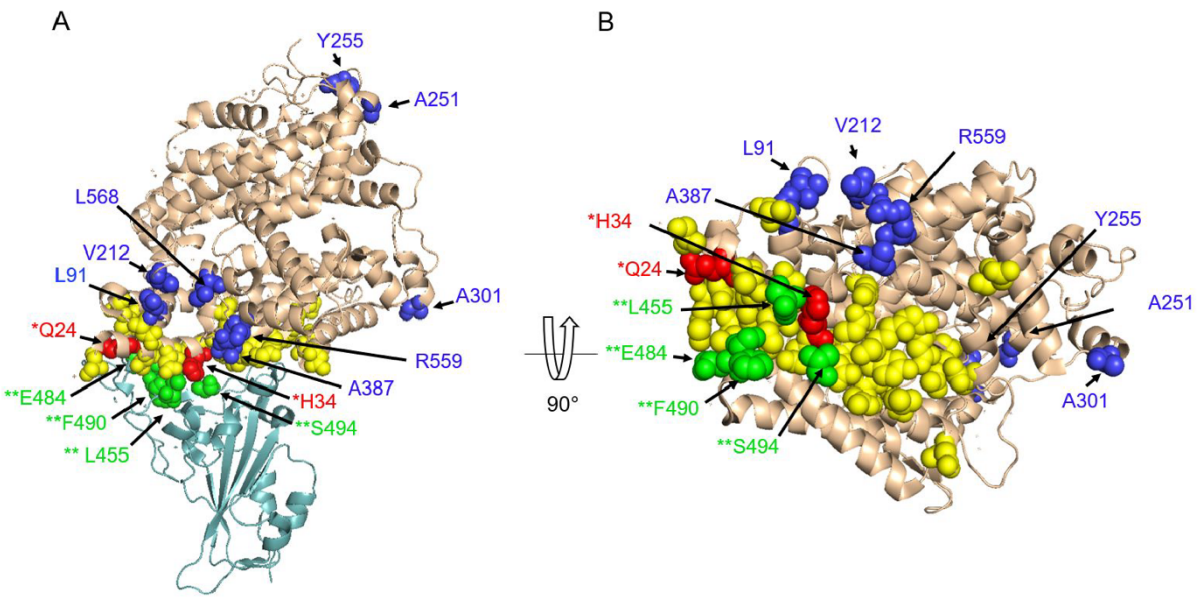

Figure 19: Residues under positive selection detected with CODEML and acceleration with phyloP in mammals

(A) ACE2 is represented in wheat cartoon with residues involved in the binding interface shown in yellow spheres. Dark blue and red spheres indicate residues in ACE2 that are accelerated and under positive selection. Red spheres represent residues that overlap with positions in the binding interface and are labeled with (*). The spike RBD is shown in light teal cartoon. Green spheres indicate residues on the SARS-CoV-2 spike protein under positive selection and are labeled with (**). (B) 90 degree rotation of the ACE2 protein.

positive selection, and was run on a subset of 64 representative mammals (see Materials and Methods).

ACE2 shows significant evidence of positive selection across mammals (ω=1.83, LRT=194.13, p<0.001). Almost 10% of codons (N=73; 9 near the RBD) are accelerated within mammals, and 18 of these have BEB scores greater than 0.95, indicating positively selected residues (Figure 18). Nineteen accelerated residues, including two positively-selected codons (Q24, H34), are critical for the binding of the ACE2 RBD and SARS-CoV-2 S (**Error! Reference source not found.**; **Error! Reference source not found.**).

Q24 has not been observed to be polymorphic within the human population, and H34 harbors a synonymous polymorphism (AF=0.00063) but no non-synonymous polymorphisms.

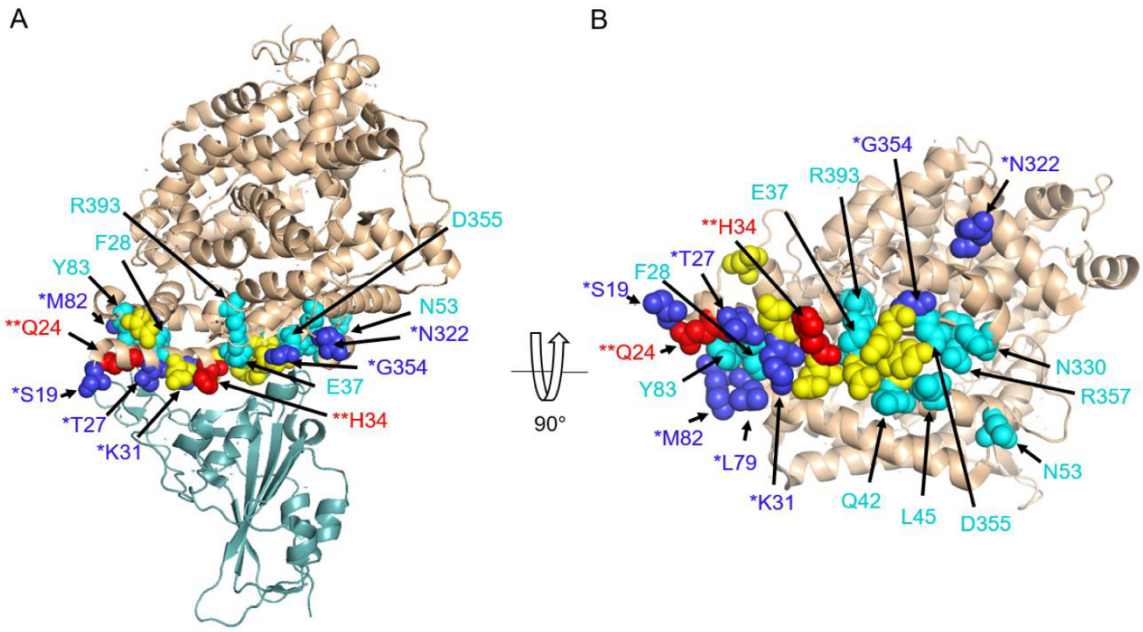

Figure 20: Residues under accelerated evolution in mammals, overlapping the binding interface, as detected using phyloP

(A) The SARS-CoV-2 spike RBD is shown in light teal cartoon. ACE2 is shown in wheat cartoon with residues involved in the binding interface shown in yellow spheres. (*) Dark blue and red spheres indicate ACE2 residues that are accelerated, under positive selection and overlapping the binding interface. Cyan spheres indicate ACE2 residues that are conserved. (**) Red spheres also demonstrate positive selection with CODEML. (B) 90 degree rotation of the ACE2 protein.

This pattern of acceleration and positive selection in *ACE2* also holds for individual mammalian lineages. Using CODEML, positive selection was detected within the orders Chiroptera (LRT=346.40, ω=3.44 p<0.001), Cetartiodactyla (LRT=92.86, ω=3.83, p<0.001), Carnivora (LRT=65.66, ω=2.27. p<0.001), Primates (LRT=72.33, ω=3.16,

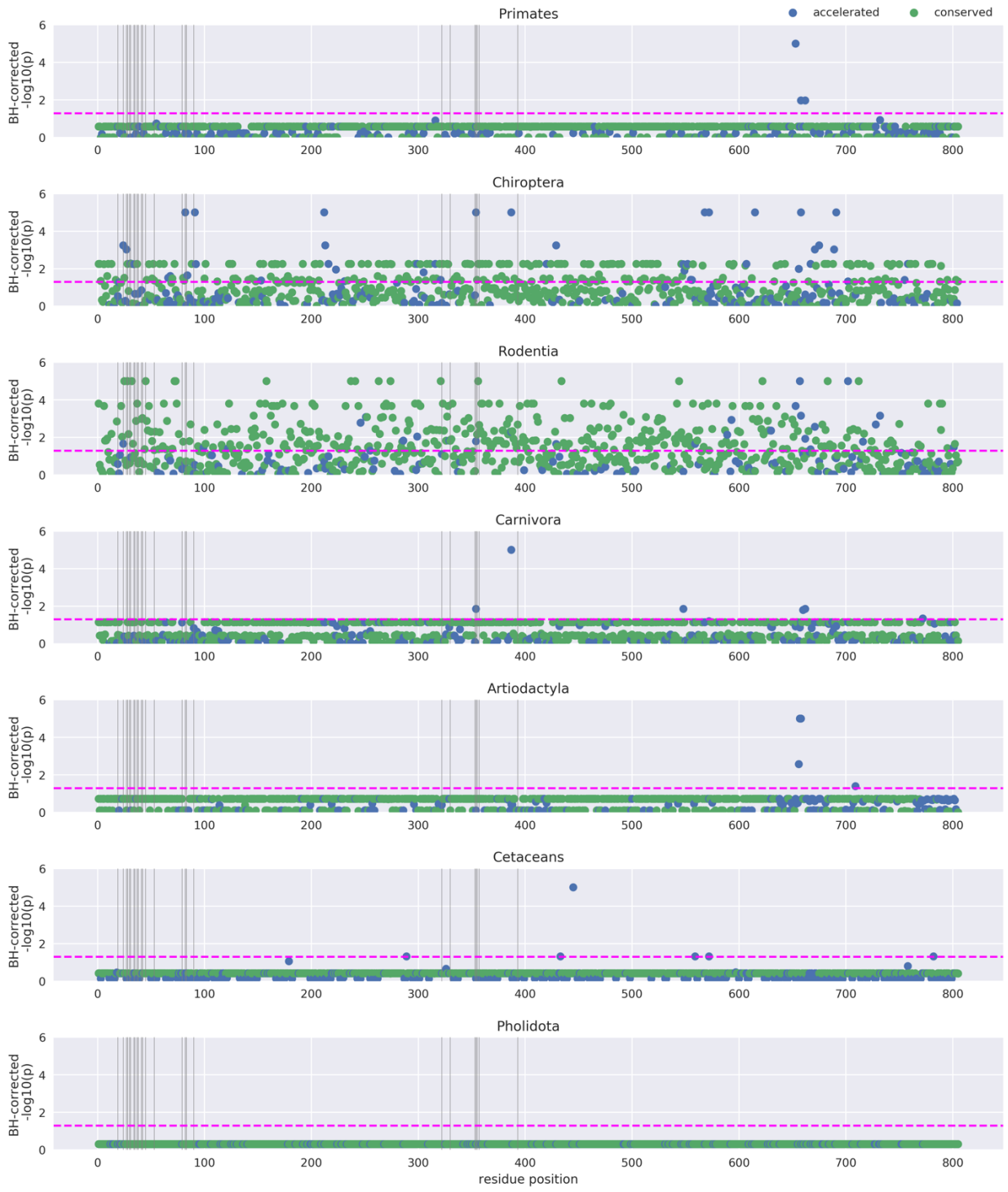

Figure 21: Intralineage phyloP results for all ACE2 codons

PhyloP signal was assessed at all ACE2 codons for various mammalian lineages against neutral models trained on those lineages, thereby identifying intralineage signals of shifts in evolutionary rate. Green dots indicate codons classified as conserved and blue dots accelerated. Vertical grey lines indicate important binding residues in ACE2. The x-axis indicates the corresponding position in the ACE2 protein for each codon, and the y-axis indicates the phyloP p-value for each codon.

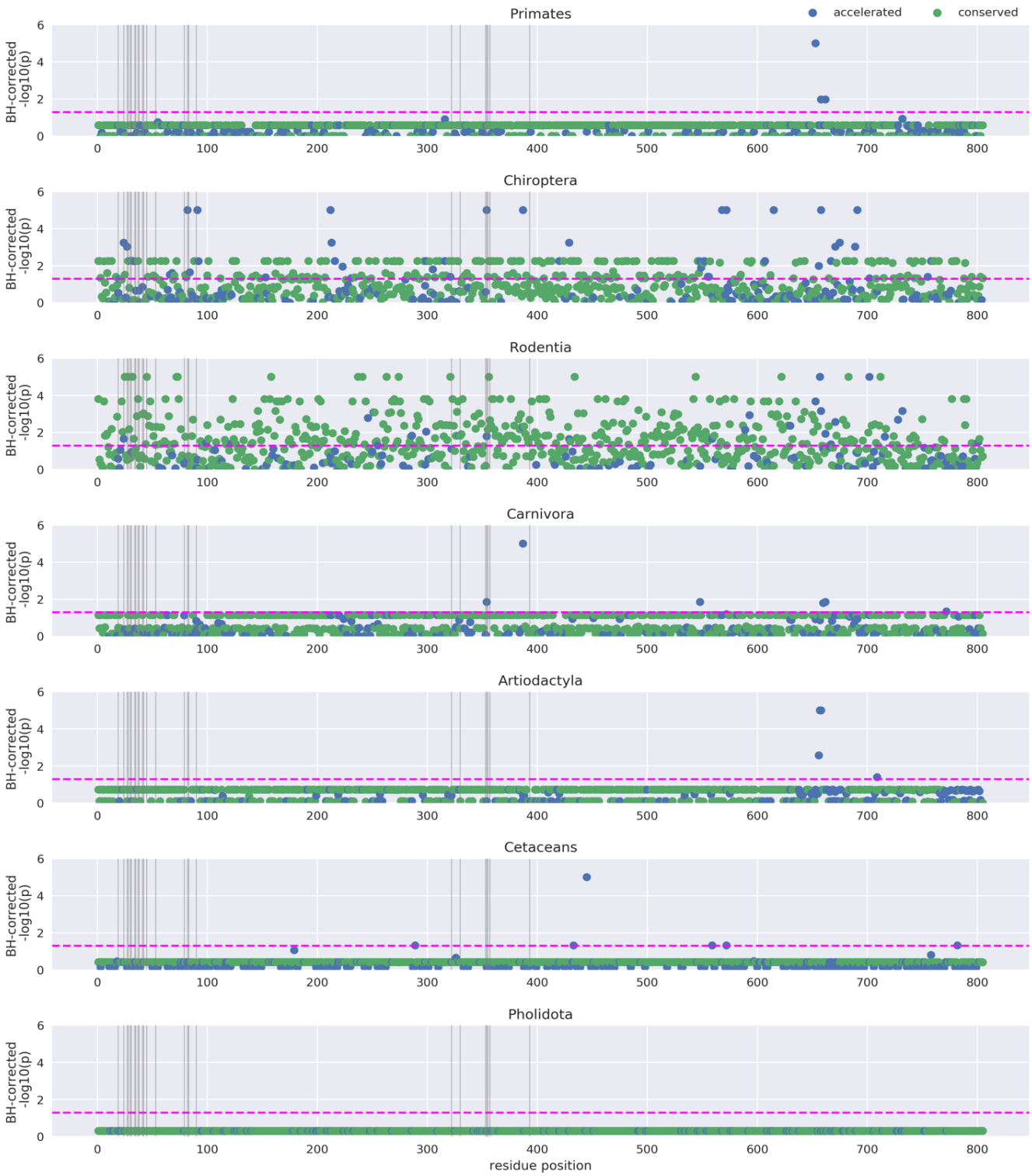

Figure 22: PhyloP results for mammalian lineages against a mammal neutral model

PhyloP signal was assessed at all ACE2 codons for various mammalian lineages against a neutral model trained on all mammalian species in the alignment. Green dots indicate codons classified as conserved and blue dots accelerated. Vertical grey lines indicate important binding residues in ACE2. The x-axis indicates the corresponding position in the ACE2 protein for each codon, and the y-axis indicates the phyloP p-value for each codon.

10, 8, 7 and 15 sites in Cetartiodactyla, Carnivora, Primates and Rodentia, respectively). Positive selection at key sites for the binding of ACE2 and SARS-CoV-2 was only found in the bat-specific alignment. PhyloP was used to assess shifts in evolutionary rate within mammalian lineages, for each assessing signal relative to a neutral model trained on species from the specified lineage (Figure 21). We discovered six important binding residues, five of which showed evidence for positive selection, that are accelerated in one or more of Chiroptera, Rodentia, or Carnivora, with G354 accelerated in all of these lineages.

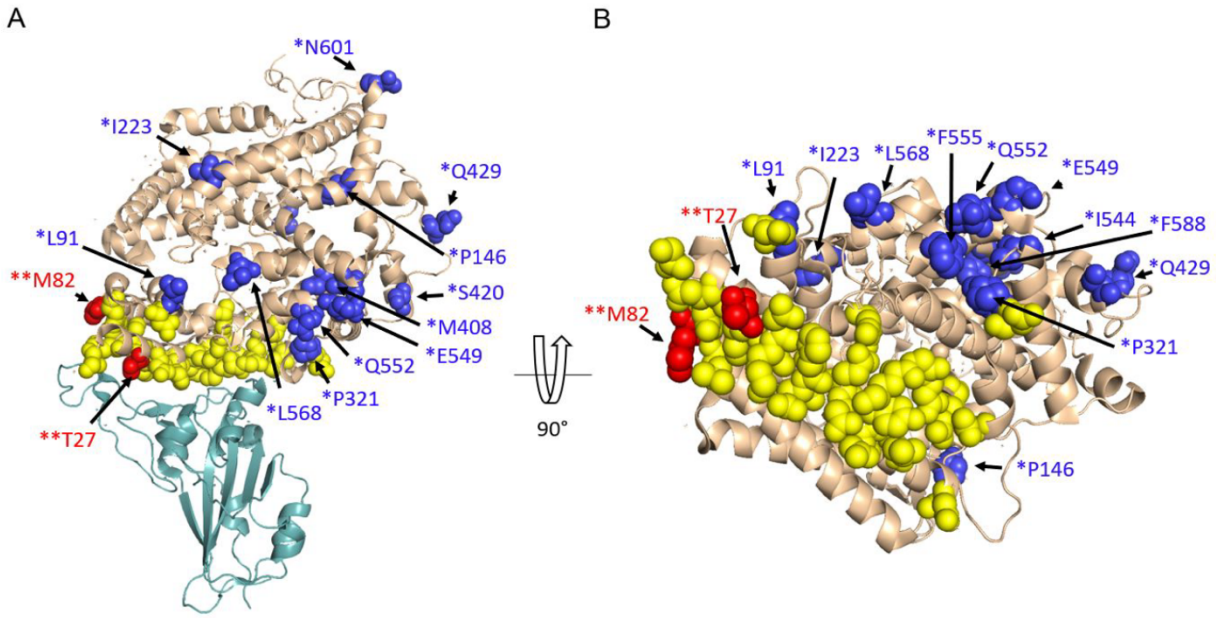

Figure 23: Residues under acceleration with phyloP in chiroptera relative to mammals

(**A**) The SARS-CoV-2 spike RBD is shown in light teal cartoon. ACE2 is shown in wheat cartoon with residues involved in the binding interface shown in yellow spheres. Dark blue and red spheres indicate residues that are accelerated in bats relative to mammals. Red spheres also overlap the binding interface. (**B**) 90 degree rotation of the ACE2 protein.

Given pervasive signatures of adaptive evolution in *ACE2* across mammals, we next sought to test if any mammalian lineages are evolving particularly rapidly compared to the others. CODEML branch-site tests identified positive selection in both the ancestral Chiroptera branch (1 amino acid, ω=26.7, LRT= 4.22, p=0.039) and ancestral Cetartiodactyla branch (2 amino acids, ω=10.38, LRT= 7.89, p=0.004) using 64 mammals. These residues did not correspond to known viral binding sites. We found no evidence for lineage-specific positive selection in the ancestral primate, rodent or carnivore lineages. PhyloP identified lineage-specific acceleration in Chiroptera, Carnivora, Rodentia, Artiodactyla and Cetaceans relative to mammals (Figure 22). Bats have a particularly high level of accelerated evolution (18 codons; $p<0.05$).

Of these accelerated residues, T27 and M82 are known to be important for binding SARS-CoV-2, with some bat subgroups having amino acids predicted to lead to less favorable binding of SARS-CoV-2 (**Error! Reference source not found.**; **Error! Reference source not found.**). Surprisingly, a residue that is conserved overall in our 410 species alignment and in the mammalian subset, Q728, is perfectly conserved in all 37 species of bats except for fruit bats (Pteropodidae), which have a substitution from Q to E. These results support the theory that *ACE2* is under lineage-specific selective pressures in bats relative to other mammals.

Positive selection was found using CODEML at sites L455, E484, F490 and S494 in the SARS-CoV-2 S sequence (ω=1.15, LRT=116.7, $p<0.001$); however, this signal was not particularly high, possibly due to the small sample size (N=8). All of these sites lie within

or near the ACE2 SARS-CoV-2 S RBD binding sites (**Error! Reference source not found.**) (Andersen et al., 2020).

4.6 Evolution of *ACE2*

Variation of *ACE2* in the human population is rare (Karczewski et al., 2020). We examined a large set of *ACE2* variants for their potential differences in binding to SARS-CoV-2 S and their relationship to selected and accelerated sites. We found rare variants that would result in missense mutations in 7 out of the 25 binding residues. Some of those (e.g. E35K with an AF of 0.00001636) could reduce the virus binding affinity, thus potentially lowering the susceptibility to the virus in a very small fraction of the population. The analysis suggests that some variants (e.g. D38E) might not affect the binding while others (e.g. S19P) have uncertain effects. Further studies are needed to confirm and correctly address recent discoveries (Cao et al., 2020; Hussain et al., 2020; Stawiski et al., 2020) and the data presented here, investigating the possible effect of these rare variants in specific populations.

When exploring patterns of codon evolution in ACE2, we found that a number of sites are evolving at different rates in the different lineages represented in our 410-species vertebrate alignment. Multiple ACE2 RBD residues important for the binding of SARS-CoV-2 are evolving rapidly across mammals, with two (Q24 and H34) under positive selection (**Error! Reference source not found.**; **Error! Reference source not found.**). Relative to other lineages analyzed, Chiroptera has a greater proportion of accelerated

versus conserved residues, particularly at the SARS-CoV-2 S RBD, suggesting the possibility of selective forces on these codons in Chiroptera driven by their interactions with SARS-CoV-2-like viruses (Figure 21). Indeed, distinct signatures of positive selection found in bats and in the SARS-CoV S protein support this hypothesis that bats are evolving to tolerate SARS-CoV-2-like viruses.

4.7 Relationship of the ACE2 binding score to known infectivity of SARS-CoV-2

Data on susceptibility of wild animals to SARS-CoV-2 is still very limited. It has been reported that a captive Malayan tiger was infected by SARS-CoV-2 (United States Department of Agriculture Animal and Plant Health Inspection Service) and that domestic cats, ferrets (Shi et al., 2020), rhesus macaques (Munster et al., 2020) and Syrian golden hamsters (Chan et al., 2020) are susceptible to experimental infection by SARS-CoV-2. These results agree with our predictions of ACE2 binding ability to SARS-CoV-2 S (Figure 14; **Error! Reference source not found.**); 4/5 five species with demonstrated susceptibility to SARS-CoV-2 score *very high* (Rhesus macaque) or *medium* (domestic cat, tiger and Golden hamster). The only inconsistency was observed for ferrets, which had a *low* ACE2 binding score. This inconsistency could be related to the high infectivity dose used for experimental infection that likely does not correspond to virus exposure in nature. Dogs have low susceptibility to SARS-CoV-2 under experimental conditions (Shi et al., 2020), and score *low* for binding of their ACE2 to SARS-CoV-2 S. However, kidney cell lines derived from dog showed ACE2-dependent SARS-CoV-2 S entry, suggesting that *in vitro* experiments may be

overestimating true infectivity potential (Hoffmann et al., 2020; Jebb et al., 2019). Pigs (*low*), ducks (*very low*) and chickens (*very low*) were similarly exposed to SARS-CoV-2 and showed no susceptibility (Shi et al., 2020), providing further support of our methodology. A recent publication reporting that SARS-CoV-2 could use pig, masked palm civet and Chinese rufous horseshoe bat ACE2 expressed in HeLa cells were inconsistent with our predictions, while data for mouse was in agreement (Zhou et al., 2020). Indeed, while mouse ACE2 scored *very low* in our analysis, pig and Chinese rufous horseshoe bat score *low*, while the masked palm civet scored *very low*. As for the ferret, high-level exposure to the virus *in vitro* could potentially result in infection via low affinity interactions with ACE2. Another possibility is that other cellular machinery present in the human HeLa cells is facilitating the infection, and that infectivity does not relate directly to ACE2 differences in these species. Confirmation of *in vitro* and *in vivo* susceptibility of these species under physiological conditions and with proper controls is clearly necessary. In addition, the expression of ACE2 varies across animal age, cell types, tissues and species (Sun et al., 2020b; Xie et al., 2006), which may lead to discrepancies between SARS-CoV-2 susceptibility gleaned from experimental infections or laboratory experiments and predictions made from the ACE2-based binding score.

4.8 Mammals with high predicted risk of SARS-CoV-2 infection

Of the 19 catarrhine primates analyzed, 18/19 scored *very high* for binding of their ACE2 to SARS-CoV-2 S and one scored *high* (the Angola colobus); the 18 species scoring *very high* had 25/25 identical binding residues to human ACE2, including rhesus

macaques (*Macaca mulatta*), which are known to be infected by SARS-CoV-2 and develop COVID-19-like clinical symptoms (Munster et al., 2020; Shan et al.). Our analysis predicts that all Old World primates are susceptible to infection by SARS-CoV-2 via their ACE2 receptors. Thus, many of the 21 primate species native to China could be a potential reservoir for SARS-CoV-2. The remaining primate species were scored as *high* or *medium*, with only the Gray mouse lemur and the Philippine tarsier scoring as *low*.

We were surprised to find that all three species of Cervid deer and 12/14 cetacean species have *high* scores for binding of their ACE2s to SARS-CoV-2 S. There are 18 species of Cervid deer found in China. Therefore, Cervid deer cannot be ruled out as an intermediate host for SARS-CoV-2. While coronavirus sequences have been found in white tailed deer (Alekseev et al., 2008) and gammacoronaviruses have been found in beluga whales (Mihindukulasuriya et al., 2008; Schütze, 2016) and bottlenose dolphins (Woo et al., 2014) and are associated with respiratory diseases, the cellular receptor used by these viruses is not known.

4.9 Other artiodactyls

A relatively large fraction (21/30) of artiodactyl mammals were classified with *medium* score for ACE2 binding to SARS-CoV-2 S. These include many species that are commonly found in Hubei Province and around the world, such as domesticated cattle, sheep and goats, as well as many species commonly found in zoos and wildlife parks

(e.g., Masai giraffe, okapi, hippopotamus, water buffalo, scimitar horned oryx, and Dama gazelle). Although cattle MDBK cells were shown in one study to be resistant to SARS-CoV-2 *in vitro* (Hoffmann et al., 2020), we propose immediate surveillance of common artiodactyl species for SARS-CoV-2 and studies of cellular infectivity, given our predictions. If ruminant artiodactyls can serve as a reservoir for SARS-CoV-2, it would have significant epidemiological implications as well as implications for food production and wildlife management (see below). It is noteworthy that camels and pigs, known for their ability to be infected by coronaviruses (Anthony et al., 2017), both score *low* in our analysis. These data are consistent with results (discussed above) indicating that pigs cannot be infected with SARS-CoV-2 both *in vivo* (Shi et al., 2020) and *in vitro* (Hoffmann et al., 2020).

4.10 Rodents

Among the rodents, 7/46 species score *high* for ACE2 binding to SARS-CoV-2 S, with the remaining 11, 10 and 18 scoring *medium*, *low* or *very low*, respectively. Brown rats *(Rattus norvegicus*) and the house mouse (*Mus musculus*), scored *very low*, consistent with infectivity studies (Hoffmann et al., 2020; Zhou et al., 2020). Given that wild rodent species likely come in contact with bats as well as with other predicted high risk species, we urge surveillance of *high* and *medium* binding likelihood rodents for the presence of SARS-CoV-2.

4.11 Bats and other species of interest

Chiroptera (bats) represent a clade of mammals that are of high interest in COVID-19 research because several bat species are known to harbor coronaviruses, including those most closely related to the betacoronavirus SARS-CoV-2 (Zhou et al., 2020). We analyzed ACE2 from 37 bat species of which 8 and 29 scored *low* and *very low*, respectively. These results were unexpected because the three *Rhinolophus* spp. including the Chinese rufous horseshoe bat are major suspects in the transmission of SARS-CoV-2, or a closely related virus, to humans (Zhou et al., 2020). Globally, bats have been shown to harbour the highest diversity of betacoronaviruses in mammals tested (Anthony et al., 2017) and show little pathology carrying these viruses (Banerjee et al., 2020). We found evidence for accelerated evolution at six RBD binding domain residues within the bat lineage, which is more than in any other lineage tested. Bats also had far more sites showing evidence of positive selection, including four binding domain residues, compared to other mammalian orders. This suggests that the diversity observed in bat ACE2 sequences may be driven by selective pressure from coronaviruses. Our results suggest that SARS-CoV-2 is not likely to use the ACE2 receptor in bats, which challenges a recent study showing that SARS-CoV-2 can infect HeLa cells expressing *Rhinolophus sinicus* ACE2 (Zhou et al., 2020). If bats can be infected with SARS-CoV-2, the virus likely uses a different receptor. For example, the MERS-CoV, a betacoronavirus, uses CD26/DPP4 (Raj et al., 2013) while the porcine transmissible enteritis virus, an alphacoronavirus uses aminopeptidase N (ANPEP) (Delmas et al., 1992). As detailed above, further *in vitro* and *in vivo* infectivity studies

are required to fully understand the mode of transmission of susceptibility of bats to SARS-CoV-2.

4.12 Carnivores

Recent reports of a Malayan tiger and a domestic cat infected by SARS-CoV-2 suggest that the virus can be transmitted to other felids (Shi et al., 2020; United States Department of Agriculture Animal and Plant Health Inspection Service). Our results are consistent with these studies; 9/9 felids we analyzed scored *medium* for ACE2 binding of SARS-CoV-2 S. However, the masked palm civet (*Paguma larvata*), a member of the Viverridae family that is related to but distinct from Felidae, scored as *very low*. These results are inconsistent with transfection studies using civet ACE2 receptors expressed in HeLa cells (Zhou et al., 2020), although these experiments have limitations as discussed above. While carnivores closely related to dogs (dingos, wolves and foxes) all scored *low*, experimental data supporting infection in dogs were inconsistent (Hoffmann et al., 2020; Shi et al., 2020; Temmam et al., 2020) so no conclusions can be drawn.

4.13 Pangolins

Considerable controversy surrounds reports that pangolins can serve as an intermediate host for SARS-CoV-2. Pangolins were proposed as a possible intermediate host (Zhang et al., 2020) and have been shown to harbor related

coronaviruses. In our study, ACE2 of Chinese pangolin (*Manis pentadactyla*), Sunda pangolin (*Manis javanica*), and white bellied pangolin (*Phataginus tricuspis*) had *low* or *very low* binding score for SARS-CoV-2 S. Neither experimental infection nor *in vitro* infection with SARS-CoV-2 has been reported for pangolins. As for ferrets and bats, if SARS-CoV-2 infects pangolins it may be using a receptor other than ACE2, based on our analysis.

4.14 Other vertebrates

Our analysis of 29 orders of fishes, 29 orders of birds, 3 orders of reptiles and 2 orders of amphibians predicts that the ACE2 proteins of species within these vertebrate classes are not likely to bind SARS-CoV-2 S. Thus, vertebrate classes other than mammals are not likely to be an intermediate host or reservoir for the virus, despite predictions reported in a recent study (Qiu et al., 2020), unless SARS-CoV-2 can use another receptor for infection. With many different non-mammal vertebrates sold in the seafood and wildlife markets of Asia and elsewhere, it is still important to determine if SARS-CoV-2 can be found in non-mammalian vertebrates.

4.15 Relevance to Threatened Species

Among the 103 species that scored *very high*, *high* and *medium* for ACE2 SARS-CoV-2 S RBD binding, 41 (40%) are classified in one of three 'Threatened' categories (*Vulnerable*, *Endangered*, and *Critically Endangered*) on the IUCN Red List of

Threatened Species, five are classified as *Near Threatened*, and two species are classified as *Extinct in the Wild* (IUCN, 2019). This represents only a small fraction of the threatened species potentially susceptible to SARS-CoV-2. For example, all 20 catarrhine primate species in our analysis, representing three families (Cercopithecidae, Hylobatidae, and Hominidae) scored *very high*, suggesting that all 185 species of catarrhine primates, most of which are classified Threatened (Bosch et al., 2005), are potentially susceptible to SARS-CoV-2. Similarly, all three species of deer, representatives of a family of ~92 species (Cervidae), scored as *high* risk, as did species representing Cetacea (baleen and toothed whales), and both groups contain a number of threatened species. Toothed whales have potential for viral outbreaks and have lost function of a gene key to the antiviral response in other mammalian lineages (Braun et al., 2015). If they are susceptible to SARS-CoV-2 , human-to-animal transmission could pose a risk through sewage outfall (Bosch et al., 2005) and contaminated refuse from cities, commercial vessels and cruise liners (Copeland, 2005). In contrast, some threatened species scored *low* or *very low*, such as the giant panda (*low*), potentially positive news for these at risk populations.

Our results have practical implications for populations of threatened species in the wild and those under human care (including those in zoos). Established guidelines for minimizing potential human to animal transmission should be implemented and strictly followed. Guidelines for field researchers working on great apes established by the IUCN have been in place since 2015 in response to previous human disease outbreaks (Gilardi et al., 2015) and have received renewed attention because of SARS-CoV-2

(Estrada et al., 2017; Gilardi et al., 2015; Gillespie and Leendertz, 2020). For zoos, guidelines in response to SARS-CoV-2 have been distributed by several Taxon Advisory Groups of the North American Association of Zoos and Aquariums (AZA), the American Association of Zoo Veterinarians (AAZV), and the European Association of Zoo and Wildlife Veterinarians (EAZWV), and these organizations are actively monitoring and updating knowledge of species in human care considered to be potentially sensitive to infection (A. Lecu, M. Bertelsen, C. Walzer, EAZWV Infectious Diseases Working Group, 2020; J. Johnson, A. Moresco, S. Han, 2020). Although *in silico* studies suggest potential susceptibility of diverse species, verification of infection potential is warranted, using cell cultures, stem cells, organoids, and other methods that do not require direct animal infection studies. Zoos and other facilities that maintain living animal collections are in a position to provide such samples for generating crucial research resources by banking tissues, and cryobanking viable cell cultures in support of these efforts.

4.16 Animal models for COVID-19

A variety of animal models have been developed for studying SARS and MERS coronavirus infections (Sutton and Subbarao, 2015). Presently, there is a tremendous need for animal models for studying SARS-CoV-2 infection and pathogenesis, as the only species currently known to be infected and show similar symptoms of COVID-19 is rhesus macaque. Non-human primate models have proven to be highly valuable for other infectious diseases, but are expensive to maintain and numbers of experimental

animals are limited. Our results provide an extended list of potential species that might be useful as animal models for SARS-CoV-2 infection and pathogenesis, including Chinese hamster and Syrian/Golden hamster (Chan et al., 2020), and large animals maintained for biomedical and agricultural research (e.g., domesticated sheep and cattle).

4.17 Conclusions

We predict that species scored as *very high* and *high* for SARS-CoV-2 S binding to ACE2 will have a high probability of becoming infected by the virus. We also predict that many species having a *medium* score have some risk of infection, and species scored as *very low* and *low* are unlikely to be infected by SARS-CoV-2 via the ACE2 receptor. Importantly, our predictions are based solely on *in silico* analyses and must be confirmed by direct experimental data. Until such time, other than for species in which SARS-CoV-2 infection has been demonstrated to occur using ACE2, we urge caution not to over-interpret the predictions made in the present study. This is especially important with regards to species, endangered or otherwise, in human care. While species ranked *high* or *medium* may be susceptible to infection based on the features of their ACE2 residues, pathological outcomes may be very different among species depending on other mechanisms that could affect virus replication and spread to target cells, tissues, and organs within the host. Furthermore, we cannot exclude the possibility that infection in any species occurs via another cellular receptor, as has been shown for other betacoronaviruses. Nonetheless, our predictions provide a useful

starting point for selection of appropriate animal models for COVID-19 research and for identification of species that may be at risk for human-to-animal or animal-to-animal transmissions by SARS-CoV-2. The approach we used for ACE2 can be extended to other cellular proteins known to be involved in coronavirus infection and immunity to better understand infection, transmission, inflammatory responses and disease progression.

4.18 Methods

Angiotensin I converting enzyme 2 (ACE2) coding and protein sequences

All human ACE2 orthologs for vertebrate species, and their respective coding sequences, were retrieved from NCBI Protein (March 20, 2020) (NCBI Resource Coordinators, 2016). ACE2 coding DNA sequences were extracted from available or recently sequenced unpublished genome assemblies for 123 other mammalian species, with the help of genome alignments and the human or within-family ACE2 orthologs. The protein sequences were predicted using AUGUSTUS v3.3.2 (Mario Stanke, 2005) or CESAR v2.0 (Sharma et al., 2017) and the translated protein sequences were checked against the human ACE2 orthologue. ACE2 gene predictions were inspected and manually curated if necessary. For four bat species (*Micronycteris hirsuta*, *Mormoops blainvillei*, *Tadarida brasiliensis* and *Pteronotus parnellii*) the ACE2 coding region was split into two scaffolds which were merged, and for *Eonycteris spelaea* a putative 1bp frameshift base error was corrected. Eighty ACE2 predictions were

obtained from the Zoonomia project, 19 from the Hiller Lab, 12 from the Koepfli lab, 8 from the Lewin lab and 4 from the Zhou lab. The source, and accession numbers for the genomes or proteins retrieved from NCBI are listed in Dataset S1. The final set of ACE2 sequences comprises 410 vertebrate species. To assure alignment robustness, the full set of coding and protein sequences were aligned independently using Clustal Omega (Sievers and Higgins, 2014), MUSCLE (Tabebordbar et al., 2016) and COBALT (Papadopoulos and Agarwala, 2007) all with default parameters. All resulting protein alignments were identical. Clustal Omega alignments were used in the subsequent analysis. Each amino acid replacement present in our dataset was classified as neutral, semi-conservative and non-conservative as in Clustal Omega.

Identification of ACE2 residues involved in binding to SARS-CoV-2 S protein

We identified 22 ACE2 protein residues that were previously reported to be critical for the effective binding of ACE2 RBD and SARS-CoV-2 S (Lan et al., 2020; Shang et al., 2020). These residues include S19, Q24, T27, F28, D30, K31, H34, E35, E37, D38, Y41, Q42, L45, L79, M82, Y83, N330, K353, G354, D355, R357, and R393. All these residues were identified from the co-crystallization and structural determination of SARS-CoV-2 S and ACE2 RBD (Lan et al., 2020; Shang et al., 2020). The known human ACE2 RBD glycosylation sites N53, N90 and N322 were also included in the analyzed residue set (Sun et al., 2020a).

ACE2 and SARS-CoV-2 binding ability prediction

Based on the known interactions of ACE2 and SARS-CoV-2 residues, we developed a set of rules for predicting the likelihood of the SARS-CoV-2 S binding to ACE2. Each species was classified in one of five categories: *very high*, *high*, *medium*, *low* or *very low* likelihood of binding SARS-CoV-2 S. Species in the *very high* category have at least 23/25 critical residues identical to the human; have K353, K31, E35, M82, N53, N90 and N322; do not have N79; and have only conservative substitutions among the non-identical 2/25 residues. Species in the *high* group have at least 20/25 residues identical to the human; have K353; have only conservative substitutions at K31 and E35; do not have N79; and can only have one non-conservative substitution among the 5/25 non-identical residues. Species scoring *medium* have at least 20/25 residues identical to the human; can only have conservative substitutions at K353, K31, and E35; and can have up to two non-conservative substitutions in the 5/25 non-identical residues. Species in the *low* category have at least 18/25 residues identical to the human; can only have conservative substitutions at K353; can have up to three non-conservative substitutions on the remaining 7/25 non-identical residues. Lastly, species in the *very low* group have less than 18/25 residues identical to the human or have at least four non-conservative substitutions in the non-identical residues.

Protein structure analysis

We applied an orthogonal approach to assess the likelihood of binding of a sampling of species that were predicted to bind SARS-CoV-2 across the categories of *high*, *medium*, *low* or *very low* likelihood of binding. ACE2 amino acid sequences from 28 species were extracted from the multiway alignment and loaded into SWISS-MODEL (Waterhouse et al., 2018) in order to generate homology derived models. The output files were aligned to the crystal structure 6MOJ (Lan et al., 2020) in order to assess the overall similarities to human ACE2. We used two recently solved crystal structures of the complex for ACE2 and SARS-CoV-2 S RBD, 6MOJ (Lan et al., 2020) and 6VW1 (Shang et al., 2020) as ground truth for the human ACE2/SARS-CoV-2 S interaction. In the program CHIMERA (Pettersen et al., 2004), we utilized the rotamer function to model each individual variant that species exhibit separately, and chose the rotamer with the least number of clashes, retaining the most initial hydrogen bonds and containing the highest probability of formation as calculated by CHIMERA from the Dunbrack 2010 backbone-dependent rotamer library (Shapovalov and Dunbrack, 2011). The rotamer was then evaluated in the context of its structural environment and assigned a score based on likelihood of interface disruption. Neutral (N) was assigned if the residue maintained a similar environment as the original residue, and was predicted to maintain or in some cases increase affinity. Weakened (W) was assigned if hydrophobic contacts were lost and contacts that appear disruptive are introduced that are not technically clashes. Unfavorable (U) was assigned if clashes are introduced

and/or a hydrogen bond is broken. Additional structural visualizations were generated in Pymol (PyMOL).

Human variants analysis

All variants at the 25 residues critical for effective SARS-CoV-2-ACE2 binding (Lan et al.; Shang et al., 2020; Sun et al., 2020a) were compiled from from dbSNP (Sherry, 2001), 1KGP (Voight et al., 2015), Topmed (NHLBI), UK10K (UK10K Consortium et al., 2015) and CHINAMAP (28). Specific population frequencies were obtained from gnomAD v.2.1.1 (Karczewski et al., 2020).

Phylogenetic reconstruction of the vertebrate *ACE2* species tree

The multiple sequence alignment of 410 ACE2 orthologous protein sequences from mammals, birds, fishes, reptiles and amphibians was used to generate a gene tree using the maximum likelihood method of reconstruction, as implemented in IQTREE (Minh et al., 2020). The best fit model of sequence evolution was determined using ModelFinder (Kalyaanamoorthy et al., 2017) and used to generate the species phylogeny. A total of 1000 bootstrap replicates were used to determine node support using UFBoot (Hoang et al., 2018).

Identifying sites undergoing positive selection

Signatures of site-specific positive selection in the *ACE2* receptor were explored using CODEML, part of the Phylogenetic Analysis using Maximum Likelihood (PAML, (Yang, 2007)) suite of software. Given CODEML's computational complexity, a smaller subset of mammalian taxa (N=64, Dataset S1), which included species from all prediction categories mentioned above, was used for selection analyses. To calculate likelihood-derived dN/dS rates (ω), CODEML utilises both a species tree and a codon alignment. The species tree for all 64 taxa was calculated using IQTREE (Minh et al., 2020) and the inferred best-fit model of sequence evolution (JTT+F+R4). This gene topology was generally in agreement with the 410 taxa tree, however bats were now sister taxa to Perissodactyla. Therefore all selection analyses were run using both the inferred gene tree, and a modified tree with the position of bats manually modified to reflect the 410 taxa topology. All species trees used were unrooted. A codon alignment of the 64 mammals was generated using pal2nal (Suyama et al., 2006) with protein alignments generated with Clustal Omega (Sievers and Higgins, 2014) and their respective CDS sequences.

Site-models M7 (null model) and M8 (alternative model) were used to identify *ACE2* sites undergoing positive selection in mammals. Both M7 and M8 estimate ω using a beta distribution and 10 rate categories per site with $\omega <= 1$ (neutral or purifying selection), but with an additional 11^{th} category allowing $\omega > 1$ (positive selection) in M8. A likelihood ratio test (LRT) calculated as $2*(\ln L_{alt} - \ln L_{null})$, comparing the fit of both null

and alternative model likelihoods was carried out, with a p-value calculated assuming a chi-squared distribution. Sites showing evidence of positive selection were identified by a significant (>0.95) Bayes Empirical Bayes (BEB) score, and validated by visual inspection of the protein alignment. To explore order-specific instances of positive selection, separate multiple sequence alignments and gene trees for Chiroptera (N=37), Cetartiodactyla (N=45), Carnivora (N=44), Rodentia (N=46) and Primates (N=39) were also generated and explored using M7 vs. M8 in CODEML.

In addition to site-models, branch-site model A1 (null model) and model A (alternative model) were also implemented targeting various mammalian orders, specifically Chiroptera, Cetartiodactyla, Rodentia and Primates, to identify lineage-specific positive selection in the *ACE2* receptor sequence. Branch-site Model A1 constrains both the target foreground branch (Carnivora, Chiroptera, Cetartiodactyla, Rodentia and Primates) and background branches to $\omega<=1$, while the alternative Model A allows positive selection to occur in the foreground branch. Null and alternative models were compared using LRTs as above, with significant BEB sites identified.

We also looked for positively selected sites in the viral spike protein, using SARS-CoV-2 (MN908947.3), Bat coronavirus RaTg13 (MN996532.1), Bat SARS-like coronavirus isolate Rs4231 (KY417146.1), SARS-related coronavirus strain BtKY72 (KY352407.1), SARS coronavirus Urbani (AY278741.1), SARS coronavirus PC4-227 (AY613950.1), Coronavirus BtRs-BetaCoV/YN2018B (MK211376.1) and the more divergent Bat Hp-betacoronavirus/Zhejiang2013 (NC_025217.1) viral strains. Protein and codon

alignments were generated as above, with the viral species tree inferred using full genome alignments of all strains generated with Clustal Omega (Sievers and Higgins, 2014). Site-test models were applied using CODEML, and significant BEB sites identified.

Analysis for departure from neutral evolutionary rate in ACE2 with PHAST

Neutral models were trained on the specified species sets using the REV nucleotide substitution model implemented in phyloFit using an expectation maximization algorithm for parameter optimization. The neutral model fit was based on third codon positions to approximate the neutral evolution rate specific to the *ACE2* gene, using a 410-species phylogenetic tree generated by IQTREE as described above and rooted on fishes. The program phyloP was then used to identify codons undergoing accelerated or conserved evolution relative to the neutral model using --features to specify codons, --method LRT --mode CONACC, and --subtree for lineage-specific tests, with p-values thus assigned per codon based on a likelihood ratio test. P-values were corrected for multiple testing using the Benjamini-Hochberg method (Pollard et al., 2010) and sites with a corrected p-value less than 0.05 were considered significant. PhyloFit and phyloP are both part of the PHAST package v1.4 (Hubisz et al., 2011; Ramani et al., 2019).

5 Investigating the role of lamina associated domains in establishment and maintenance of cell identity

The work below reflects a manuscript currently in preparation for submission. The work is a collaboration with the Jain lab at the University of Pennsylvania, particularly Parisha Shah, who generated all of the unpublished data in the manuscript. Parisha and I are co-leading this project, with her leading the experimental side and me leading the computational analysis.

5.1 Introduction

Adult human bodies are composed of trillions of cells, comprising more than 200 distinct cell types, which are faithfully established and maintained throughout a healthy lifespan. Identifying and understanding molecular mechanisms regulating establishment and maintenance over time of cellular identity are areas of intense interest. In particular, understanding how cell type specific responses are achieved – the ability of a cell to respond to specific stimuli to differentiate or to attenuate a stimulus response to maintain established identity – is fundamental to understanding cellular diversity.

Many decades of study have uncovered unique cell type-specific transcriptional profiles. While these distinct transcriptomes underscore the diversity of cellular function, it remains incompletely understood how such coordinated genome-wide transcriptional regulation is achieved during development. In vitro differentiation models have identified

key signaling molecules and stimuli that function at branch-points of differentiation pathways, revealing delicate and complex processes involving lineage-specific enhancers and resulting in cell type-specific gene patterning (Takahashi et al., 2007). Though important for understanding key developmental cues, these models do not capture the full complexity of genome-wide transcriptional regulation or maintenance of cellular identity over chronological time.

Three-dimensional genome organization has emerged as a potential mechanism to coordinate cell-type specific gene regulation and maintain cell type transcriptional fidelity. In particular, genome organization at the nuclear periphery may provide a key platform for cell type-specific transcription. The nuclear lamina is a filamentous network of lamins A/C, B1, and B2 proteins residing on the inner surface of the nuclear envelope (Burke and Stewart, 2006; Worman and Bonne, 2007). A large proportion of the genome is localized towards the lamina, termed lamina-associated domains (LADs), which range in size from hundreds of kilobases to megabases (Guelen et al., 2008). These loci are generally heterochromatic, and genes within LADs are generally transcriptionally repressed and undergo active silencing, while genes away from the lamina are more often competent for transcriptional activation (Guelen et al., 2008). In some cell types, the chromatin is naturally inverted, and this inversion can be generated by ablation of nuclear lamin genes, indicating an active role for lamins in anchoring LADs to the nuclear periphery (Solovei et al., 2013). The degree of conservation of LADs between cell types and species implies an important role for these loci (Guelen et al., 2008; Meuleman et al., 2013). Our group and others have shown that spatial LAD

positioning regulates organogenesis and show a high degree of transcriptional repression (Peric-Hupkes et al., 2010; Poleshko et al., 2017; Robson et al., 2019). In particular, subsets of LADs are repositioned away from or to the lamina during differentiation in a cell type-specific manner (Peric-Hupkes et al., 2010). During mESC neuronal differentiation model, key neuronal genes lose lamina occupancy in neuronal precursor cells (Meuleman et al., 2013), and a similar phenomenon is observed during mESC cardiac differentiation (Poleshko et al., 2017). Likewise, preventing normal "release" of LAD-bound chromatin impacts normal mESC cardiac differentiation (Poleshko et al., 2017). These studies underscore the biological relevance of nuclear organization and changes therein, but are limited in scope to individual differentiation pathways or cell types.

An additional question is how peripheral chromatin itself is organized. An intriguing finding from previous studies suggests that a subset of LADs have varying characteristics – reduced lamin occupancy and increased gene density compared to other LADs, indicating that the concept of a LAD as a monolithic block needs to be revised. Work in single cells has shown the frequency at which LADs contact the nuclear lamina varies by locus, and correlates with gene density, implicating a structural role for LADs with higher contact frequency (Kind et al., 2015). Also, only a subset of LADs re-position away from the lamina during differentiation, and individual genomic regions have varying probabilities of becoming re-localized to or from the nuclear lamina (Kind et al., 2013, 2015). Moreover, chromatin at the nuclear periphery is frequently marked by the histone modification H3K9me2, and genomic loci enriched for this signal

have been shown to have a high degree, but not perfect, overlap with LADs (Poleshko et al., 2017). These results raise the intriguing possibility that peripheral heterochromatin may comprise distinct compartments defined by unique sets of features. Defining these various categories of peripheral chromatin domains across multiple cell types, and characterizing distinct or dynamic subtypes of those features, will provide critical knowledge about how peripheral chromatin is organized and how nuclear organization regulates cellular identity.

Here, we have defined the nuclear organization signatures based on LB1 and H3K9me2 occupancy via ChIP-seq across thirteen isogenic human cell types from all four germ layers derived from H9 embryonic stem cells. By linking these with transcriptional data, we identify evidence of cooperative shifts between chromatin structure and gene expression associated with each cell type. Overall, this work provides an atlas of peripheral chromatin and associated features in multiple human cell types across all four germ layers.

5.2 A 3-state Hidden Markov Model approach identifies two categories of LADs that vary by cell type

We generated ChIP-seq datasets for lamin-B1 from thirteen human ES-derived cell types comprising all four germ layers and representative of multiple early differentiation trajectories (Table 2). Visual inspection of these data confirmed the presence of large,

diffuse domains of enrichment of LB1 signal consistent with the presence of LADs in all cell types investigated. Due to the diffuse nature of these LB1-enriched domains and inspired in part by previous work (Meuleman et al., 2013), I implemented a Hidden Markov Model trained on the tier one cell types (Table 2) to identify LADs based on the LB1 ChIP-seq input. A three-state rather than the previously described two-state model was better able to accurately identify LAD-like domains, evidenced by their LB1 binding, genome , region size and concordance with enriched domains evident by visual inspection. Among the three states, one state demonstrated the greatest enrichment for LB1 signal, and so was designated as LADs, with the remaining states were classified as LADlite and nonLAD by descending LB1 signal (Figure 24). LADs in each cell type

Table 2: Cell types with LB1 and H3K9me2 ChIP-seq data

LB1 and H3K9me2 ChIP-seq data were generated from each of the listed cell types.

Cell Type	Germ Layer	Tier
H9-Derived Embryonic Stem Cells	Embryonic	1
Cardiac Myocytes	Mesoderm	1
Early Somite	Mesoderm	1
Paraxial Mesoderm	Mesoderm	1
Epicardium	Mesoderm	2
Day 4 Artery	Mesoderm	2
Sorted Cardiac Myocytes	Mesoderm	2
Mid-Hindgut	Endoderm	1
Liver	Endoderm	1
Endothelial Progenitors	Endoderm	2
Definitive Ectoderm	Ectoderm	1
Day 5 Midbrain	Ectoderm	1
Border Ectoderm	Ectoderm	2

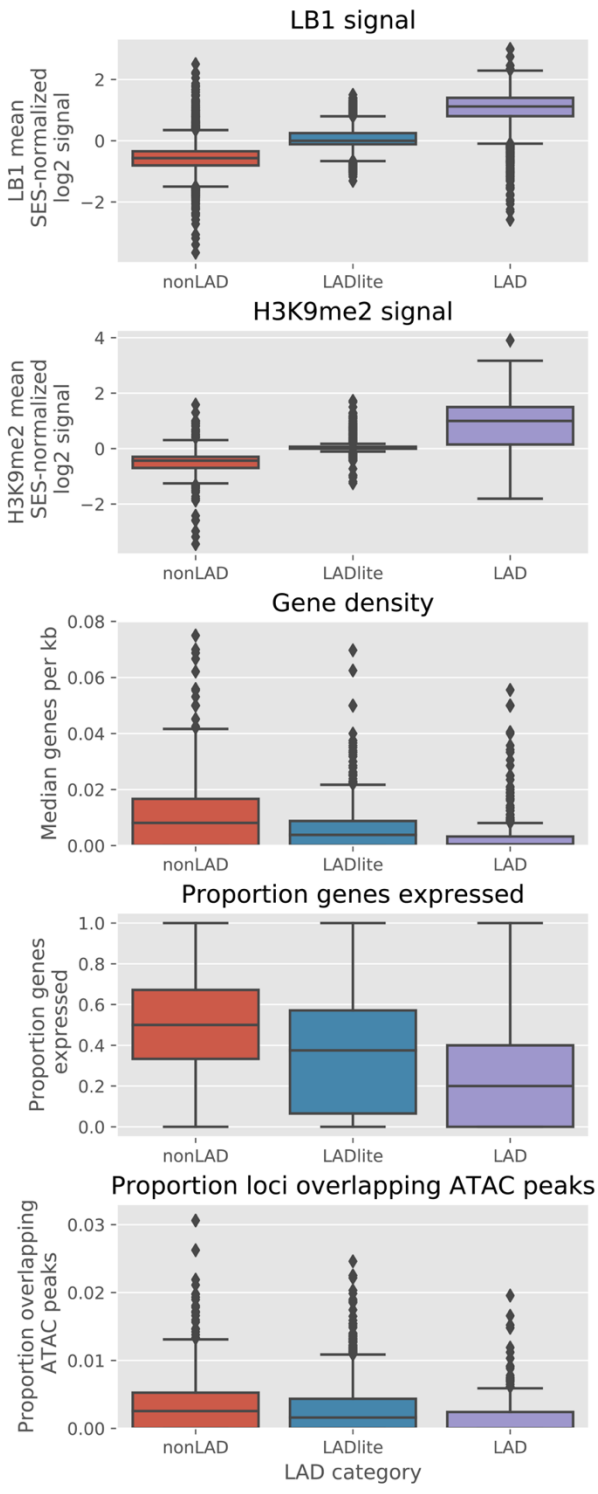

Figure 24: Properties of LADs

LB1, H3K9me2, gene density, gene expression and ATACseq peak overlap in LAD categories in embryonic stem cells.

demonstrated characteristic features, including enrichment for the repressive histone mark H3K9me2, lower gene density, generally lower gene expression, and less overlap with ATACseq peak compared to LADlites and nonLADs. Median LAD sizes for tier one cell types ranged from 160-280kb, covering 21.4%-38.3% of the genome. We found that some genomic loci were classified as LADs in every cell type assessed, consistent with previously described constitutive LADs, while others varied by cell type, consistent with facultative LADs that vary between cell types (Meuleman et al., 2013; Peric-Hupkes et al., 2010). LADs and LADlites are generally more conserved, and tend to have lower GC content compared with nonLADs, and genomic loci that are categorized at LADs in all cell types are depleted for CpG islands. LADs and LADlites tend to have lower amounts of within-domain CTCF

binding, while their boundaries show higher relative binding, particularly on the 5' boundary. LADs are enriched for late-replicating genomic regions, indicating that they are replicated later than nonLAD and LADlites during cell division. While in general LADs tend to be depleted for transposable elements, LADs and LADlites vary in their signal depending on the category of transposable element. For example, LADlite are depleted for LINE elements while LADs are enriched, and LADs are enriched for simple repeats while LADlites are depleted. These findings support the conclusion that our model was successfully able to identify and differentiate two distinct categories of lamin-associated genomic loci based on lamin-B1 ChIP-seq data.

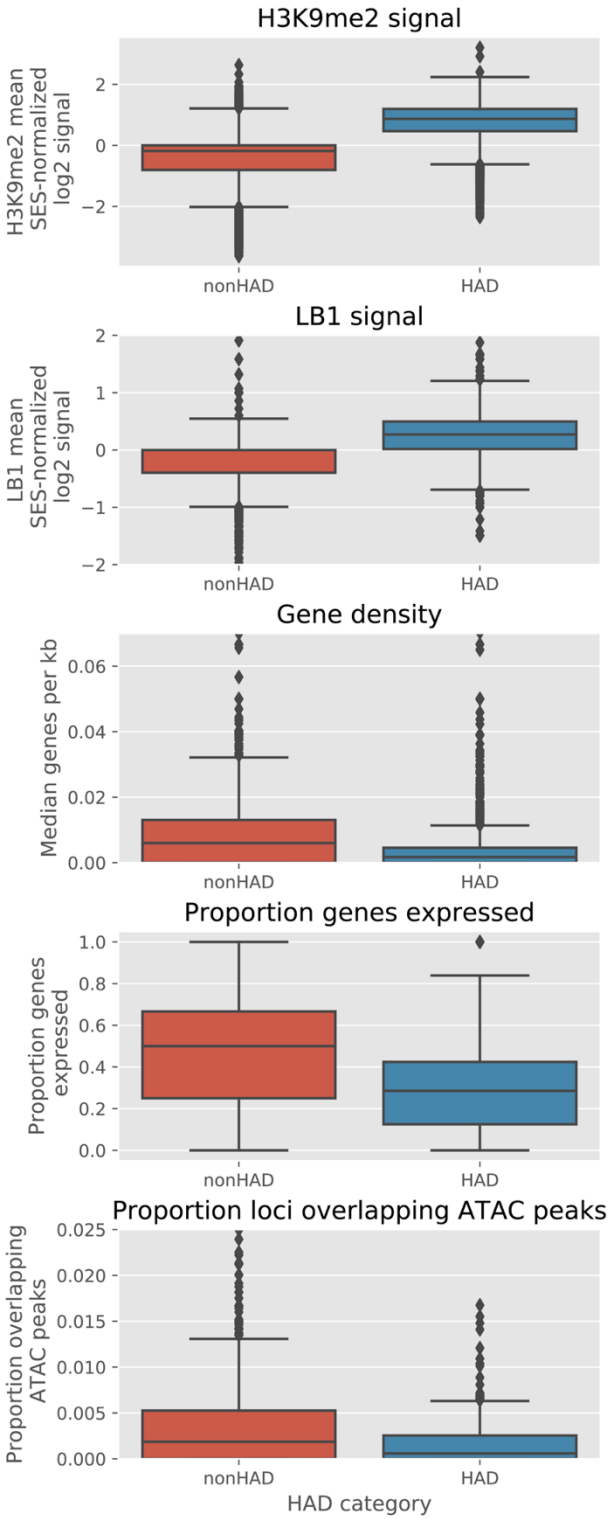

Figure 25: Properties of KDDs

H3K9me2, LB1, gene density, gene expression and ATACseq peak overlap in KDD categories in embryonic stem cells.

5.3 A 2-state Hidden Markov Model identifies H3K9me2 domains

We generated ChIP-seq datasets for H3K9me2 from thirteen human ES-derived cell types comprising all four germ layers and representative of multiple early differentiation trajectories (Table 2). In order to distinguish H3K9me2 domains from LADs described above, I trained another HMM on the H3K9me2 ChIP-seq tier one data, in this case finding a two-state model to be the best fit to the data. From the two states identified by the model, the state with higher H3K9me2 signal was assigned the label of H3K9me2-associated domain (K9-dimethyl domain, “KDD”), and the other state “nonKDD” (Figure 25). KDDs in each cell type demonstrated enrichment for LB1, lower gene density, generally lower gene expression, and less overlap with ATACseq peaks compared to nonKDDs (Figure 25). Median KDD sizes for tier one cell types ranged from 380-2360kb, covering 44.2%-82.8% of the genome. As found with LADs and LADlites, some genomic loci were classified as KDDs in all cell types assessed, while others varied per cell type. KDDs are depleted for CTCF, with a sharp increase in binding at their boundaries. Overall, these data confirm prior findings that H3K9me2-enriched regions share many characteristics with LADs (Kind et al., 2013).

5.4 LADs and KDDs are overlapping but distinct

We found that generally most LADs (>90% for 7 of 8 tier one cell types) are in KDDs, while generally about half of KDDs were in LADs. Exceptions, such as in the case of definitive ectoderm where only 10% of LADs are in KDDs, likely stem from the

additional noise present in the LB1 ChIP-seq data for that cell type, leading to instances where loci that are likely LADs were instead categorized as LADlite. Median region sizes for LADs that overlap KDDs range from 300-820kb, skewing larger in comparison to LADs and falling within the lower range of KDD sizes. Genomic loci categorized as both LADs and KDDs tend to have a higher LB1 ChIP-seq signal, while loci that are LADs but not KDDs have slightly higher LB1 signal compared to loci that are KDDs but not LADs. Genomic loci categorized as both LADs and KDDs also tend to higher H3K9me2 ChIP-seq signal, while loci that are KDDs but not LADs have slightly higher H3K9me2 signal compared to loci that are KDDs but not LADs. Genomic loci that are KDDs but not LADs seem to have higher boundary-associated CTCF signal. Taken together, these data may indicate that genomic loci categorized as both LAD and KDD constitute a particularly robust category of LAD with strong signals of genomic repression and association with the nuclear lamina.

5.5 LADs, KDDs and A and B compartments

LADs share many characteristics with B compartments, large swaths of heterochromatin with many of the same characteristics as LADs, such as low gene expression and low gene density. Therefore we assessed the concordance between LADs, LADlites and KDDs from cardiomyocytes and ESCs with the B compartment calculated from previously published Hi-C data matched cell types (Zhang et al., 2019). The Jaccard index, a measure of similarity, is greater for KDDs (Jaccard=0.62) than for LADs (Jaccard=0.55), indicating greater overlap with B compartments. However, the

Jaccard index for the concatenation of LADs and LADlites (Jaccard=0.75) is the largest, and greater even than the concatenation of LADs, LADlites and KDDs (Figure 26). The same trend was observed in ESCs. This validates the idea that lamina-associated chromatin occupies a similar fraction of chromatin as B compartments, consistent with the similar characteristics of both. However, the overlap of these different categories of chromatin is far from perfect, suggesting that while these domains are similar, they may be functionally distinct.

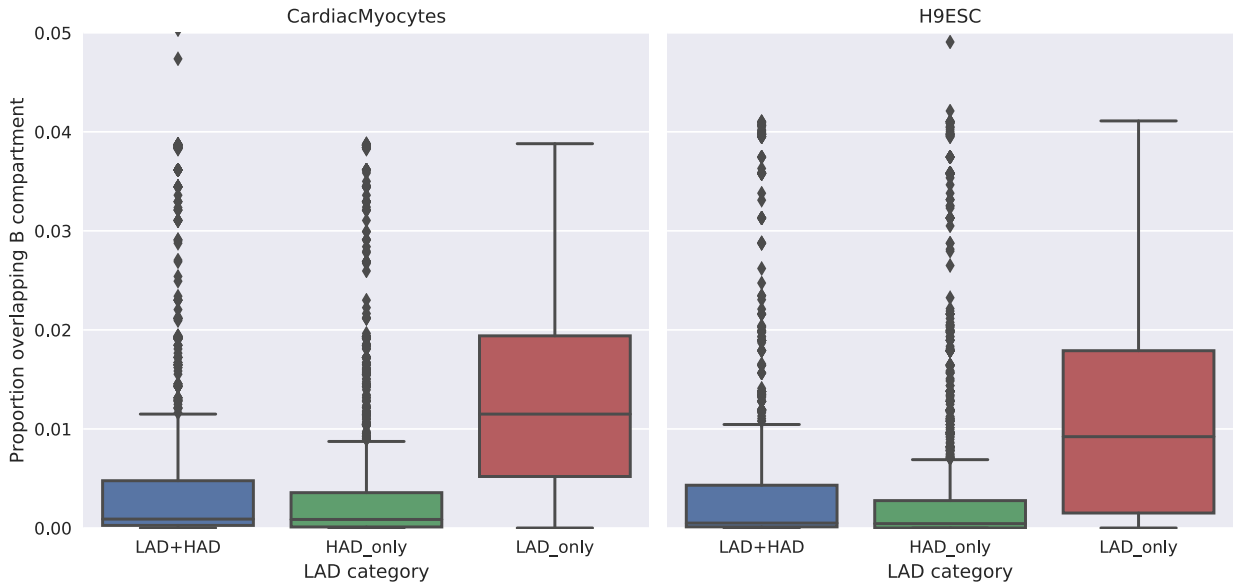

Figure 26: B compartment overlap LADs and HADs

Overlap with B compartment from Hi-C data for matched cell types for LADs overlapping HADs, and LADs and HADs alone.

5.6 Cell type specificity in LADs, LADlites and KDDs

As previously noted, a subset of LADs, LADlites and KDDs vary between different cell types while the remainder are cell-type invariant. A caveat here is that it is likely that some of the regions identified as invariant by this analysis do actually vary in other cell

types not tested in this study. Supporting this, we found immune-related transcription factor binding motifs in loci that are LADs in all cell types, indicating that potentially these loci detach from the lamina in immune cell types. However, some of the loci identified by this study are likely to actually be cell-type invariant. We found that LADs that overlap KDDs are significantly more likely to overlap a cell-type invariant LAD in most cell types (Fisher's exact test p-value < 0.05). Therefore, the combination of LAD and KDD may represent a genomic region that is more stably linked to the nuclear lamina both within and between cell types.

I investigated whether cell type variable loci contained cell-type-specific features. Gene ontology (GO) analysis on genes that fall in genomic loci that are LADs or LADlites in all cell types assayed resulted in processes that might be expected to be shared, such as chromosome segregation. Genomic loci that were categorized cell-type-specifically as LADs or LADlites tended to include more cell-type-relevant terms, such as embryo development and gastrulation for H9ESCs and regulation of neurogenesis for Day5-Midbrain.

I assessed differential enrichment transcription factor binding motifs (TFBMs) analysis across various sets to determine whether this would reflect cell-type-specific signals. Relative to the union of LADs from all cell types, cardiomyocyte LADs were enriched for motifs for genes involved in differentiation and processes of cell types other than cardiomyocytes, such as *CDX2*, implicated in the intestinal epithelium, *NKX2.2*, implicated in immune- and neuronal-related gene regulation, neuronal gene *SOX6*,

epithelial gene *HOXB13*, and insulin metabolism gene *FOXO1*. This supports the idea that non-cardiac genes are silenced and therefore located in LADs in cardiomyocytes. Some of the transcription factors listed are known to have a preference to bind methylated DNA and/or have repressive functionality, and therefore may contribute to the silencing of their target genes and potentially the generation of heterochromatin and/or sequestration to the nuclear lamina in those loci. Similarly, Day 5 Midbrain LADs relative to all LADs were enriched for TFBMs for non-neuronal genes, for example some relevant to heart including *SOX17, NR2F2* and *DLX2*, and some involved in pluripotency, including *OCT4* and *SOX1*. ESC LADs relative to all LADs were enriched for all of the transcription factors listed above for cardiomyocytes, as well as many additional related to differentiation and functionality in differentiated cell types, in line with the repression of these genes in pluripotent cells. Overall these findings are consistent with a model in which repression of expression of genes regulated by cell-type-relevant transcription factors is accomplished through sequestration of TFBMs and transcription start sites in LADs of other cell types in order to protect cell identity and differentiation fidelity.

I next sought to determine whether there was evidence for difference in enrichment of cell-type-relevant TFBMs in different LAD categories and KDDs, which may lend insight into the different functions of these domains. I found enrichment of TFBMs for many cardiac-defining transcription factors in LADlites compared to LADs in cardiomyocytes, including *GATA2, PLAGL1, HAND2,* and *TBX5*. Furthermore, we found enrichment of multiple pluripotency-maintenance genes in LADlites relative to LADs in

cardiomyocytes, including *OCT4, SOX1, SOX2, SOX3, SOX15, KLF4, KLF1, MYC*, and *NANOG*. The same trend was seen in H9ESCs. Overall, this supports the hypothesis that LADlites may represent a distinct chromatin region relative to LADs, important for cell type differentiation and maintenance. KDDs were enriched for multiple heart- and pluripotency-important genes relative to LADs in cardiomyocytes, which may reflect the overlap of KDDs with LADlites. It also may reflect the higher density of genes in KDDs relative to LADs. Overall it seems that LADlites and KDDs are more dynamic compared to LADs and important for cell type differentiation and maintenance.

5.7 Discussion

Maintenance of chromatin structure via association with the nuclear periphery is among the many factors contributing to successful cell type specification and identity (Poleshko et al., 2017). LADs have previously been shown to demonstrate heterogeneity among different cell types and even between single cells. In this work we have captured some of that heterogeneity using bulk measurements of LB1 and H3K9me2 DNA binding in various cell types, validating previous findings of LAD cell type specificity in a wide range of cell types and germ layers and further differentiating two distinct LAD categories within each cell type. The driving forces of these distinct LAD categories are as of yet unknown, but likely possibilities include the categories reflecting different frequencies of lamina-attachment between single cells as previously demonstrated (Kind et al., 2015). Furthermore, we were able to define an independent domain, the KDD, based on H3K9me2. Various characteristics of KDDs relative to LADs and

LADlites, such as gene expression, gene density, and ATACseq accessibility, support the idea that KDDs represent a specific kind of functional domain. Genomic loci that are both LADs and KDDs appear to represent the most stable peripheral heterochromatin both within and between cell types.

Overall, this work takes domains that were previously viewed as a single monolithic entity, LADs, and characterizes a cell-type-specific set of overlapping but distinct domains. This follows the general trend of genomics, in which no locus can be simply defined, and most often has multiple functions depending on its context (Halfon, 2019). These findings lay the groundwork for future studies aimed at defining the driving cause behind the difference between LADlites and LADs – are these truly more dynamic regions, or do they vary cell-to-cell, leading to the distinct bulk LB1 measurements? Is the enrichment of cell-type-specific TFBMs in LADs, LADlites and KDDs a driving force in, or a consequence of, differentiation and cell identity?

5.8 Methods

Methods are provided here for the computational parts of this project only as that was my contribution to this work.

ChIP-sequencing data processing for Lamin-B1 and H3K9me2

Adapters were trimmed using Trimmomatic [v0.39] (Bolger et al., 2014). Sequencing reads were aligned to human reference hg38 using BWA-MEM [v0.7.17] (Li and Durbin, 2010). The FASTA file for hg38 was downloaded from the UCSC Genome Browser. Aligned reads were converted to BAM and sorted using Samtools [v0.1.19] (Heintzman et al., 2009), with quality filter ("-F") set to 1804. Duplicates were removed using Picard [v2.18.7] MarkDuplicates. Sequencing reads from the ENCODE blacklist were removed using Bedtools [v2.29.0] (Quinlan and Hall, 2010). Each replicate had at least 1 million mapped sequencing reads. Data for both LB1 and H3K9me2 ChIP-seq were divided into higher quality ("tier one") and lower quality ("tier two") as assessed by replicate correlation values and visual assessment in order to generate stringent sets to train the LAD- and KDD-calling models (Table 2). Spearman correlations between biological replicates was greater than 0.7 for all tier one cell types, and greater than 0.6 for tier two cell types assessed by comparing bigwig files for SES-normalized signal over controls generated with bamCompare using multiBigWigSummary and plotCorrelation from deepTools [v3.3.2] (Ramírez et al., 2014).

Identification of LADs

LB1 ChIP-seq signal were calculated and converted into BedGraph files using deepTools bamCompare [v3.3.2] (Ramírez et al., 2014) with 20kb bins, using the signal extraction scaling method (Diaz et al., 2012) for sample scaling. A 3-state HMM was

implemented using pomegranate [v0.11.1] (Schreiber, 2017). The HMM was initialized using a normal distribution and k-means to initialize the distribution with a uniform transition matrix. The Baum-Welch algorithm was then used to train the model, with tier one cell types (Table 2) used together in the model training. The model was then applied to predict LAD state genome-wide per 20kb bins for each cell type from both tier one and tier two individually, filtering regions from the ENCODE blacklist from consideration. States were labeled as LAD, LADlite or nonLAD based on median LB1 signal for the bins with that state label, with the highest median LB1 signal being assigned LAD, second highest LADlite, and lowest nonLAD.

Identification of KDDs

H3K9me2 ChIP-seq signal were calculated and converted into BedGraph files using deepTools bamCompare [v3.3.2] (Ramírez et al., 2014) with 20kb bins, using the signal extraction scaling method (Diaz et al., 2012) for sample scaling. A 2-state HMM was implemented using pomegranate [v0.11.1] (Schreiber, 2017). The HMM was initialized using a normal distribution and k-means to initialize the distribution with a uniform transition matrix. The Baum-Welch algorithm was then used to train the model, with tier one cell types (Table 1) used together in the model training. The model was then applied to predict KDD state genome-wide per 20kb bins for each cell type from both tier one and tier two individually, filtering regions from the ENCODE blacklist from consideration. States were labeled as KDD or nonKDD based on median H3K9me2

signal for the bins with that state label, with the highest median H3K9me2 signal being assigned KDD and lowest nonKDD.

RNA-sequencing analysis

Transcriptome data were quantified using Kallisto [v0.44.0] quant with fragment length determined by BioAnalyzer, standard deviation of 10, and 30 bootstraps, assigning reads using the Ensembl [v96] genome annotation (Bray et al., 2016). TPM values were quantile-normalized between cell types. Differentially expressed transcripts (q<=0.01) between cell types were identified using Sleuth [0.30.0] (Pimentel et al., 2017). RNA-seq data for cardiomyocytes, embryonic stem cells and day-15 endothelial cells were generated in the Jain lab, early somite and paraxial mesoderm were from (Koh et al., 2016), mid-hindgut from (Loh et al., 2014) and neural ectoderm from (Tchieu et al., 2017). All RNA-seq data were reanalyzed as described above.

Transcription factor binding motif analysis

Differential transcription factor binding was analyzed using Homer [v4.11.1] (Heinz et al., 2010).

A and B compartment analysis

Hi-C data for cardiomyocytes and embryonic stem cells were downloaded as Cooler files from the 4D Nucleome Data Portal (Zhang et al., 2019). A and B compartments were called using cooltools [v0.3.0] (Abdennur and Mirny, 2020).

Supporting analyses

Plotting, statistical analyses and supporting analyses were conducted in Python [v3.6] with packages Jupyter, matplotlib (Hunter, 2007), seaborn (Waskom et al., 2018), upsetplot (Lex et al., 2014), scikit-learn (Pedregosa et al., 2011), numpy (Walt et al., 2011) and pybedtools (Dale et al., 2011; Quinlan and Hall, 2010).

6 Conclusion

The work I have presented here in this dissertation investigates and emphasizes the role of genomic variation, in multiple dimensions and timescales, in health, development and evolution. This work provides a tool to use sequence variation to design therapeutics for genomic disease, investigates the role of structural and sequence-based genomic variation in evolution, uses sequence genomic variation across species to predict susceptibility to a viral pathogen and investigates the role of structural genomic variation in cell type specification. This work is a reflection of and contribution towards our growing understanding of the complexity of each aspect of our genomes. In

chapter 2, I describe a software tool I built, AlleleAnalyzer, which incorporates genomic sequence variation into gRNA design. This tool enables allele-specific gRNA design, which may help to enable a cure for diseases driven by a single allele, such as dominant negative or imprinting diseases. AlleleAnalyzer also may enable experiments that depend on targeting a specific allele, or enable more accurate editing of a region with genetic variation relative to a reference genome. In chapter 3, I investigate how genomic variation makes us human, and how structural genomic variation may contribute to accelerated sequence evolution. By implementing the HAR discovery pipeline using Nextflow, it is now more reproducible and easier to change parameters, and ascertain the impact of various components of the pipeline. Via this process, I discovered that alignment and assembly quality are important for HAR discovery, and defined a new set of HARs. My finding that HARs are enriched in TADs with human-specific structural variants suggests the possibility of enhancer hijacking as a driving factor in the accelerated evolution of HARs, a line of investigation which will need to be followed up by further analyses and experimental work. As I was writing this thesis, the world was upended by COVID-19, a disease caused by the virus SARS-CoV2. Therefore, I joined a team of scientists from all over the world, applying my skills in comparative genomics developed in work pertaining mainly to chapter 3 towards this virus. In chapter 4, I assess selection in a receptor known to mediate infectivity by SARS-CoV2 which contributed toward a method to predict risk of infection to other species. This will inform future risk predictions, potentially choice of model animals for therapeutic and vaccine development, and conservation efforts for endangered species. In chapter 5, I investigated how changes in genomic loci that are associated with

binding of LB1 or H3K9me2 impact cell type specificity. Through this work I discovered that LADs are not a single monolithic entity, but instead a set of distinct but overlapping domains. The differences between these domains appear to be related to cell type differentiation and maintenance, and may better inform how the 3D genome participates in and potentially influences cell type specification. Overall, my dissertation work improves our understanding of how genomic variation, in all its forms, is important to evolution, development, and disease.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

A. Lecu, M. Bertelsen, C. Walzer, EAZWV Infectious Diseases Working Group (2020). Science-based facts & knowledge about wild animals, zoos, and SARS-CoV-2 Virus. Eur. Assoc. Zoo Wildl. Vet. - Transm. Dis. Handb.

Abdennur, N., and Mirny, L.A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics *36*, 311–316.

Alekseev, K.P., Vlasova, A.N., Jung, K., Hasoksuz, M., Zhang, X., Halpin, R., Wang, S., Ghedin, E., Spiro, D., and Saif, L.J. (2008). Bovine-like coronaviruses isolated from four species of captive wild ruminants are homologous to bovine coronaviruses, based on complete genomic sequences. J Virol *82*, 12422–12431.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nat Med *26*, 450–452.

Anthony, S.J., Johnson, C.K., Greig, D.J., Kramer, S., Che, X., Wells, H., Hicks, A.L., Joly, D.O., Wolfe, N.D., Daszak, P., et al. (2017). Global patterns in coronavirus diversity. Virus Evol *3*, vex012.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From fastQ

data to high-confidence variant calls: The genome analysis toolkit best practices pipeline.

Bakondi, B., Lv, W., Lu, B., Jones, M.K., Tsai, Y., Kim, K.J., Levy, R., Akhtar, A.A., Breunig, J.J., Svendsen, C.N., et al. (2015). In Vivo CRISPR/Cas9 Gene Editing Corrects Retinal Dystrophy in the S334ter-3 Rat Model of Autosomal Dominant Retinitis Pigmentosa. Mol. Ther. *24*, 556–563.

Banerjee, A., Baker, M.L., Kulcsar, K., Misra, V., Plowright, R., and Mossman, K. (2020). Novel Insights Into Immune Systems of Bats. Front Immunol *11*, 26.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bosch, A., Xavier Abad, F., and Pintó, R.M. (2005). Human Pathogenic Viruses in the Marine Environment. In Oceans and Health: Pathogens in the Marine Environment, (Springer, Boston, MA), pp. 109–131.

Braun, B.A., Marcovitz, A., Camp, J.G., Jia, R., and Bejerano, G. (2015). Mx1 and Mx2 key antiviral proteins are surprisingly lost in toothed whales. Proc Natl Acad Sci U A *112*, 8036–8040.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome

Institute at Washington University, Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet *81*, 1084–1097.

Burke, B., and Stewart, C.L. (2006). The Laminopathies: The Functional Architecture of the Nucleus and Its Contribution to Disease. Annu. Rev. Genomics Hum. Genet. *7*, 369–405.

Cao, Y., Li, L., Feng, Z., Wan, S., Huang, P., Sun, X., Wen, F., Huang, X., Ning, G., and Wang, W. (2020). Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov *6*, 11.

Capra, J.A., Erwin, G.D., McKinsey, G., Rubenstein, J.L.R., and Pollard, K.S. (2013). Many human accelerated regions are developmental enhancers. Philos. Trans. R. Soc. B Biol. Sci. *368*, 20130025.

Chan, J.F.-W., Zhang, A.J., Yuan, S., Poon, V.K.-M., Chan, C.C.-S., Lee, A.C.-Y., Chan, W.-M., Fan, Z., Tsoi, H.-W., Wen, L., et al. (2020). Simulation of the clinical and pathological manifestations of Coronavirus Disease 2019 (COVID-19) in golden Syrian hamster model: implications for disease pathogenesis and transmissibility. Clin Infect Dis.

Chatterjee, S., and Ahituv, N. (2017). Gene Regulatory Elements, Major Drivers of Human Disease. Annu. Rev. Genomics Hum. Genet. *18*, 45–63.

Chen, W., Yan, M., Yang, L., Ding, B., He, B., Wang, Y., Liu, X., Liu, C., Zhu, H., You, B., et al. (2005). SARS-associated Coronavirus Transmitted from Human to Pig. Emerg Infect Dis *11*, 446.

Chen, X., Xu, F., Zhu, C., Ji, J., Zhou, X., Feng, X., and Guang, S. (2014). Dual sgRNA-directed gene knockout using CRISPR/Cas9 technology in Caenorhabditis elegans. Sci Rep *4*, 7581.

Christie, K.A., Courtney, D.G., DeDionisio, L.A., Shern, C.C., De Majumdar, S., Mairs, L.C., Nesbit, M.A., and Moore, C.B.T. (2017). Towards personalised allele-specific CRISPR gene editing to treat autosomal dominant disorders. Sci. Rep. *7*, 1–11.

Clarkson, K.L. (1993). Algorithms for polytope covering and approximation. In Lecture Notes in Computer Science, pp. 246–252.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

Copeland, C. (2005). Cruise Ship Pollution: Background, Laws and Regulations, and Key Issues. Congr. Res. Serv. Libr. Congr. *RL32450*.

Dale, R.K., Pedersen, B.S., and Quinlan, A.R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics *27*, 3423–3424.

Damas, J., Hughes, G.M., Keough, K.C., Painter, C.A., Persky, N.S., Corbo, M., Hiller, M., Koepfli, K.-P., Pfenning, A.R., Zhao, H., et al. (2020). Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. BioRxiv 2020.04.16.045302.

Danecek, P., Schiffels, S., and Durbin, R. (2014). Multiallelic calling model in bcftools (-m). 10–11.

Delmas, B., Gelfi, J., L'Haridon, R., Vogel, Sjöström, H., Norén, and Laude, H. (1992). Aminopeptidase N is a major receptor for the enteropathogenic coronavirus TGEV. Nature *357*, 417–420.

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319.

Diaz, A., Park, K., Lim, D.A., and Song, J.S. (2012). Normalization, bias correction, and peak calling for ChIP-seq. Stat. Appl. Genet. Mol. Biol. *11*.

Dixon, D.L., Trankle, C., Buckley, L., Parod, E., Carbone, S., Van Tassell, B.W., and Abbate, A. (2016). A review of PCSK9 inhibition and its effects beyond LDL receptors. J. Clin. Lipidol. *10*, 1073–1080.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat. Biotechnol. *32*, 1262–1267.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 1–12.

Drubin, D.G., and Hyman, A.A. (2017). Stem cells: the new “model organism.” Mol. Biol. Cell *28*, 1409–1411.

Eres, I.E., Luo, K., Hsiao, C.J., Blake, L.E., and Gilad, Y. (2019). Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. PLOS Genet. *15*, e1008278.

Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492.

Estrada, A., Garber, P.A., Rylands, A.B., Roos, C., Fernandez-Duque, E., Di Fiore, A., Nekaris, K.A.-I., Nijman, V., Heymann, E.W., Lambert, J.E., et al. (2017). Impending extinction crisis of the world’s primates: Why primates matter. Sci Adv *3*, e1600946.

Feng, Y., Yue, X., Xia, H., Bindom, S.M., Hickman, P.J., Filipeanu, C.M., Wu, G., and Lazartigues, E. (2008). Angiotensin-converting enzyme 2 overexpression in the subfornical organ prevents the angiotensin II–mediated pressor and drinking responses and is associated with angiotensin II type 1 receptor downregulation. Circ Res *102*, 729–736.

Forrest, M.P., Zhang, H., Moy, W., McGowan, H., Leites, C., Dionisio, L.E., Xu, Z., Shi, J., Sanders, A.R., Greenleaf, W.J., et al. (2017). Open Chromatin Profiling in hiPSC-Derived Neurons Prioritizes Functional Noncoding Psychiatric Risk Variants and Highlights Neurodevelopmental Loci. Cell Stem Cell *21*, 305-318.e8.

Francesc Alted, I.V. and others PyTables: Hierarchical Datasets in Python.

Franchini, L.F., and Pollard, K.S. (2017). Human evolution: the non-coding revolution. BMC Biol. *15*, 89.

Fudenberg, G., Kelley, D.R., and Pollard, K.S. (2019). Predicting 3D genome folding from DNA sequence (Genomics).

Gao, X., Tao, Y., Lamas, V., Huang, M., Yeh, W.-H., Pan, B., Hu, Y.-J., Hu, J.H., Thompson, D.B., Shu, Y., et al. (2018). Treatment of autosomal dominant hearing loss by in vivo delivery of genome editing agents. Nature *553*, 217–221.

Gilardi, K.V.K., Gillespie, T.R., Leendertz, F.H., Macfie, E.J., Travis, D.A., Whittier, C.A., and Williamson, E.A. (2015). Best practice guidelines for health monitoring and disease control in great ape populations.

Gillespie, T.R., and Leendertz, F.H. (2020). COVID-19: protect great apes during human pandemics. Nature *579*, 497.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., et al. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol *17*, 148.

Halfon, M.S. (2019). Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. Trends Genet. *35*, 93–103.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol *35*, 518–522.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell *181*, 271-280.e8.

Horlbeck, M.A., Gilbert, L.A., and Villalta, J.E. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *9*, 1–20.

Hubisz, M.J., and Pollard, K.S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. Curr. Opin. Genet. Dev. *29*, 15–21.

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

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Hussain, M., Jabeen, N., Raza, F., Shabbir, S., Baig, A.A., Amanullah, A., and Aziz, B. (2020). Structural Variations in Human ACE2 may Influence its Binding with SARS-CoV-2 Spike Protein. J Med Virol.

IUCN (2019). The IUCN Red List of Threatened Species. Version 2019-2. Http://Www.Iucnredlist.Org.

J. Johnson, A. Moresco, S. Han (2020). SARS-COV-2 Considerations and Precautions. AZA Small Carniv. Taxon Advis. Group.

Jebb, D., Huang, Z., Pippel, M., Hughes, G.M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L.S., Skirmuntt, E.C., Katzourakis, A., et al. (2019). Six new reference-quality bat genomes illuminate the molecular basis and evolution of bat adaptations.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods *14*, 587–589.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493-6.

Keough, K.C., Lyalina, S., Olvera, M.P., Whalen, S., Conklin, B.R., and Pollard, K.S. (2019). AlleleAnalyzer: a tool for personalized and allele-specific sgRNA design. Genome Biol. *20*, 167.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. Cell *153*, 178–192.

Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M., et al. (2015). Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. Cell *163*, 134–147.

King, M., and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116.

Koh, P.W., Sinha, R., Barkal, A.A., Morganti, R.M., Chen, A., Weissman, I.L., Ang, L.T., Kundaje, A., and Loh, K.M. (2016). An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. Sci. Data *3*, 1–15.

Kostka, D., Hubisz, M.J., Siepel, A., and Pollard, K.S. (2012). The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome. Mol. Biol. Evol. *29*, 1047–1057.

Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P., Dougherty, M.L., et al. (2018). High-resolution comparative analysis of great ape genomes. Science *360*, eaar6343.

Lam, T.T.-Y., Shum, M.H.-H., Zhu, H.-C., Tong, Y.-G., Ni, X.-B., Liao, Y.-S., Wei, W., Cheung, W.Y.-M., Li, W.-J., Li, L.-F., et al. (2020). Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature 1–6.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., et al. Crystal structure of the 2019-nCoV spike receptor-binding domain bound with the ACE2 receptor.

Laude, H., Van Reeth, K., and Pensaert, M. (1993). Porcine respiratory coronavirus: molecular features and virus-host interactions. Vet Res *24*, 125–150.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lessard, S., Francioli, L., Alfoldi, J., Tardif, J.-C., Ellinor, P.T., MacArthur, D.G., Lettre, G., Orkin, S.H., and Canver, M.C. (2017). Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. Proc. Natl. Acad. Sci. U. S. A. *114*, E11257–E11266.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. *20*, 1983–1992.

Li, F. (2013). Receptor recognition and cross-species infections of SARS coronavirus. Antivir. Res *100*, 246–254.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589–595.

Li, F., Li, W., Farzan, M., and Harrison, S.C. (2005). Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science *309*, 1864–1868.

Li, W., Moore, M.J., Vasilieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., Luzuriaga, K., Greenough, T.C., et al. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature *426*, 450–454.

Liu, Z., Xiao, X., Wei, X., Li, J., Yang, J., Tan, H., Zhu, J., Zhang, Q., Wu, J., and Liu, L. (2020). Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. J Med Virol.

Loh, K.M., Ang, L.T., Zhang, J., Kumar, V., Ang, J., Auyeong, J.Q., Lee, K.L., Choo, S.H., Lim, C.Y.Y., Nichane, M., et al. (2014). Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. Cell Stem Cell *14*, 237–252.

Lu, G., Wang, Q., and Gao, G.F. (2015). Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol. *23*, 468–478.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. *32*, 225–237.

Mario Stanke, B.M. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res *33*, W465.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, S. van der Walt, and J. Millman, eds. pp. 51–56.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and Steensel, B. van (2013). Constitutive nuclear lamina–genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Mihindukulasuriya, K.A., Wu, G., St. Leger, J., Nordhausen, R.W., and Wang, D. (2008). Identification of a Novel Coronavirus from a Beluga Whale by Using a Panviral Microarray. J Virol *82*, 5084–5088.

Miltenberger-Miltenyi, G., Janecke, A.R., Wanschitz, J.V., Timmerman, V., Windpassinger, C., Auer-Grumbach, M., and Löscher, W.N. (2007). Clinical and electrophysiological features in Charcot-Marie-Tooth disease with mutations in the NEFL gene. Arch. Neurol. *64*, 966–970.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol.

Mitchell, S., OSullivan, M., and others (2011). PuLP: a linear programming toolkit for python. Univ. Auckl.

Morgens, D.W., Wainberg, M., Boyle, E.A., Ursu, O., Araya, C.L., Tsui, C.K., Haney, M.S., Hess, G.T., Han, K., Jeng, E.E., et al. (2017). Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. Nat. Commun. *8*, 15178.

Munster, V.J., Feldmann, F., Williamson, B.N., van Doremalen, N., Pérez-Pérez, L., Schulz, J., Meade-White, K., Okumura, A., Callison, J., Brumbaugh, B., et al. (2020). Respiratory disease and virus shedding in rhesus macaques inoculated with SARS-CoV-2.

National Institutes of Health (2007). NIH Curriculum Supplement Series.

NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *44*, D7.

NHLBI Trans-Omics for Precision Medicine WGS-About TOPMed.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., Shih, D.J.H., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature *511*, 428–434.

Othman, H., Bouslama, Z., Brandenburg, J.-T., da Rocha, J., Hamdi, Y., Ghedira, K., Abid, N.-S., and Hazelhurst, S. (2020). Interaction of the spike protein RBD from SARS-CoV-2 with ACE2: similarity with SARS-CoV, hot-spot analysis and effect of the receptor polymorphism.

Oudit, G.Y., Crackower, M.A., Backx, P.H., and Penninger, J.M. (2003). The role of ACE2 in cardiovascular physiology. Trends Cardiovasc Med *13*, 93–101.

Papadopoulos, J.S., and Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics *23*, 1073–1079.

Park, R.J., Wang, T., Koundakjian, D., Hultquist, J.F., Lamothe-Molina, P., Monel, B., Schumann, K., Yu, H., Krupzcak, K.M., Garcia-Beltran, W., et al. (2016). A genome-

wide CRISPR screen identifies a restricted set of HIV host dependency factors. Nat. Genet. *49*, 193.

Patel, V.B., Zhong, J.-C., Grant, M.B., and Oudit, G.Y. (2016). Role of the ACE2/Angiotensin 1-7 Axis of the Renin-Angiotensin System in Heart Failure. Circ Res *118*, 1313–1326.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Mol. Cell *38*, 603–613.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605–1612.

Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods *14*, 687–690.

Poleshko, A., Shah, P.P., Gupta, M., Babu, A., Morley, M.P., Manderfield, L.J., Ifkovits, J.L., Calderon, D., Aghajanian, H., Sierra-Pagán, J.E., et al. (2017). Genome-Nuclear

Lamina Interactions Regulate Cardiac Stem Cell Lineage Restriction. Cell *171*, 573-587.e14.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006a). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., et al. (2006b). Forces Shaping the Fastest Evolving Regions in the Human Genome. PLoS Genet. *2*, 13.

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

Pouget, J.G., Gonçalves, V.F., Spain, S.L., Finucane, H.K., Raychaudhuri, S., Kennedy, J.L., and Knight, J. (2016). Genome-Wide Association Studies Suggest Limited Immune Gene Enrichment in Schizophrenia Compared to 5 Autoimmune Diseases. Schizophr. Bull. *42*, 1176–1184.

PyMOL The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC. Https://Pymol.Org/2/.

Qian, Z., Travanty, E.A., Oko, L., Edeen, K., Berglund, A., Wang, J., Ito, Y., Holmes, K.V., and Mason, R.J. (2013). Innate Immune Response of Human Alveolar Type II Cells Infected with Severe Acute Respiratory Syndrome–Coronavirus. Am J Respir Cell Mol Biol *48*, 742–748.

Qiu, Y., Zhao, Y.-B., Wang, Q., Li, J.-Y., Zhou, Z.-J., Liao, C.-H., and Ge, X.-Y. (2020). Predicting the angiotensin converting enzyme 2 (ACE2) utilizing capability as the receptor of SARS-CoV-2. Microbes Infect.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raj, V.S., Mou, H., Smits, S.L., Dekkers, D.H.W., Müller, M.A., Dijkman, R., Muth, D., Demmers, J.A.A., Zaki, A., Fouchier, R.A.M., et al. (2013). Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature *495*, 251–254.

Ramani, R., Krumholz, K., Huang, Y.-F., and Siepel, A. (2019). PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. Bioinformatics *35*, 2320–2322.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Renieri, A., Benetti, E., Tita, R., Spiga, O., Ciolfi, A., Birolo, G., Bruselles, A., Doddato, G., Giliberti, A., Marconi, C., et al. (2020). ACE2 variants underlie interindividual variability and susceptibility to COVID-19 in Italian population. MedRxiv.

Robson, M.I., Ringel, A.R., and Mundlos, S. (2019). Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. Mol. Cell *74*, 1110–1122.

Ryu, H., Inoue, F., Whalen, S., Williams, A., Kircher, M., Martin, B., Alvarado, B., Samee, M.A.H., Keough, K., Thomas, S., et al. (2018). Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. BioRxiv 256313.

Saif, L.J. (2010). Bovine respiratory coronavirus. Vet Clin North Am Food Anim Pr. *26*, 349–364.

Schreiber, J. (2017). Pomegranate: fast and flexible probabilistic modeling in python. J. Mach. Learn. Res. *18*, 5992–5997.

Schütze, H. (2016). Coronaviruses in Aquatic Organisms. In Aquaculture Virology, (Academic Press), pp. 327–335.

Scott, D.A., and Zhang, F. (2017). Implications of human genetic variation in CRISPR-based therapeutic genome editing. Nat. Med.

Shan, C., Shi, Z.-L., Yuan, Z.-M., Yao, Y.-F., Yang, X.-L., Zhou, Y.-W., Wu, J., Gao, G., Peng, Y., Yang, L., et al. Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the Rhesus Macaques.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., and Li, F. (2020). Structural basis of receptor recognition by SARS-CoV-2. Nature.

Shapovalov, M.V., and Dunbrack, R.L., Jr (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure *19*, 844–858.

Sharma, V., Schwede, P., and Hiller, M. (2017). CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. Bioinformatics *33*, 3985–3987.

Sherry, S.T. (2001). dbSNP: the NCBI database of genetic variation.

Shi, J., Wen, Z., Zhong, G., Yang, H., Wang, C., Huang, B., Liu, R., He, X., Shuai, L., Sun, Z., et al. (2020). Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS-coronavirus 2. Science.

Shin, J.W., Kim, K.-H., Chao, M.J., Atwal, R.S., Gillis, T., MacDonald, M.E., Gusella, J.F., and Lee, J.-M. (2016a). Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. Hum Mol Genet *25*, 4566–4576.

Shin, J.W., Kim, K.-H., Chao, M.J., Atwal, R.S., Gillis, T., MacDonald, M.E., Gusella, J.F., and Lee, J.-M. (2016b). Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. Hum. Mol. Genet. *0*, ddw286.

Shirley, M.D., Ma, Z., Pedersen, B.S., and Wheelan, S.J. (2015). Efficient "pythonic" access to FASTA files using pyfaidx (PeerJ Inc.).

Sievers, F., and Higgins, D.G. (2014). Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. Methods Mol Biol *1079*.

Solovei, I., Wang, A.S., Thanisch, K., Schmidt, C.S., Krebs, S., Zwerger, M., Cohen, T.V., Devys, D., Foisner, R., Peichl, L., et al. (2013). LBR and Lamin A/C Sequentially Tether Peripheral Heterochromatin and Inversely Regulate Differentiation. Cell *152*, 584–598.

Stawiski, E.W., Diwanji, D., Suryamohan, K., Gupta, R., Fellouse, F.A., Sathirapongsasuti, F., Liu, J., Jiang, Y.-P., Ratan, A., Mis, M., et al. (2020). Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility. BioRxiv 2020.04.07.024752.

Stéfan van der Walt, S.C.C. and G.V. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Comput. Sci. Eng. *13*, 22–30.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sun, J., He, W.-T., Wang, L., Lai, A., Ji, X., Zhai, X., Li, G., Suchard, M.A., Tian, J., Zhou, J., et al. (2020a). COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. Trends Mol Med.

Sun, K., Gu, L., Ma, L., and Duan, Y. (2020b). Atlas of ACE2 gene expression in mammals reveals novel insights in transmisson of SARS-Cov-2. BioRxiv 2020.03.30.015644.

Sutton, T.C., and Subbarao, K. (2015). Development of animal models against emerging coronaviruses: From SARS to MERS coronavirus. Virology *479–480*, 247–258.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res *34*, W609–12.

Tabebordbar, M., Zhu, K., Cheng, J.K.W., Chew, W.L., Widrick, J.J., Yan, W.X., Maesner, C., Wu, E.Y., Xiao, R., Ran, F.A., et al. (2016). In vivo gene editing in dystrophic mouse muscle and muscle stem cells. Science *351*, 407–411.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell *131*, 861–872.

Tchieu, J., Zimmer, B., Fattahi, F., Amin, S., Zeltner, N., Chen, S., and Studer, L. (2017). A Modular Platform for Differentiation of Human PSCs into All Major Ectodermal Lineages. Cell Stem Cell *21*, 399-410.e7.

Temmam, S., Barbarino, A., Maso, D., Behillil, S., Enouf, V., and others (2020). Absence of SARS-CoV-2 infection in cats and dogs in close contact with a cluster of COVID-19 patients in a veterinary campus. BioRxiv.

de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell *172*, 289-304.e18.

UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R.B., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

United States Department of Agriculture Animal and Plant Health Inspection Service USDA APHIS | USDA Statement on the Confirmation of COVID-19 in a Tiger in New York.

Vermunt, M.W., Tan, S.C., Castelijns, B., Geeven, G., Reinink, P., de Bruijn, E., Kondova, I., Persengiev, S., Bontrop, R., Cuppen, E., et al. (2016). Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. Nat. Neurosci. *19*, 494–503.

Vijgen, L., Keyaerts, E., Moës, E., Thoelen, I., Wollants, E., Lemey, P., Vandamme, A.-M., and Van Ranst, M. (2005). Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. J Virol *79*, 1595–1604.

Voight, B.F., Trynka, G., B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, GR. Abecasis, Xue, Y., H. Jung, T. Bleazard, J. Lee, D. Hong, YB. Simons, MC. Turchin, JK. Pritchard, G. Sella, Do, R., D H Alexander, J Novembre, I. Mathieson, G.M., H. Li, R.D., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Voisey, J., Mehta, D., McLeay, R., Morris, C.P., Wockner, L.F., Noble, E.P., Lawford, B.R., and Young, R.M. (2017). Clinically proven drug targets differentially expressed in the prefrontal cortex of schizophrenia patients. Brain. Behav. Immun. *61*, 259–265.

Walt, S. van der, Colbert, S.C., and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Comput. Sci. Eng. *13*, 22–30.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D.C., Augspurger, T., Halchenko, Y., Cole, J.B., et al. (2018). mwaskom/seaborn: v0.9.0 (July 2018).

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res *46*, W296–W303.

Woo, P.C.Y., Lau, S.K.P., Lam, C.S.F., Tsang, A.K.L., Hui, S.-W., Fan, R.Y.Y., Martelli, P., and Yuen, K.-Y. (2014). Discovery of a Novel Bottlenose Dolphin Coronavirus Reveals a Distinct Species of Marine Mammal Coronavirus in Gammacoronavirus. J Virol *88*, 1318–1331.

Worman, H.J., and Bonne, G. (2007). "Laminopathies": A wide spectrum of human diseases. Exp. Cell Res. *313*, 2121–2133.

Xie, X., Chen, J., Wang, X., Zhang, F., and Liu, Y. (2006). Age- and gender-related difference of ACE2 expression in rat lung. Life Sci *78*, 2166–2171.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586–1591.

Yang, L., Grishin, D., Wang, G., Aach, J., Zhang, C.-Z., Chari, R., Homsy, J., Cai, X., Zhao, Y., Fan, J.-B., et al. (2014). Targeted and genome-wide sequencing reveal single nucleotide variations impacting specificity of Cas9 in human stem cells. Nat. Commun. *5*, 5507.

Yang, T., Justus, S., Li, Y., and Tsang, S.H. (2015). BEST1: the Best Target for Gene and Cell Therapies. Mol. Ther. J. Am. Soc. Gene Ther. *23*, 1805–1809.

Zhang, T., Wu, Q., and Zhang, Z. (2020). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol *30*, 1346–1351.e2.

Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R., Lee, A.Y., et al. (2019). Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat. Genet. *51*, 1380–1388.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Kathleen Call Keough
B4A4936EE81B4E5...
Author Signature

6/6/2020
Date