UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Evolutionary Genomics of Transfer RNA Genes and SARS-CoV-2

A dissertation submitted in partial satisfaction of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Bryan Thornlow

December 2021

_____

Professor Russell Corbett-Detig, co-chair

_____

Professor Todd Lowe, co-chair

_____

Professor David Haussler

_____

Peter Biehl, Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Evolutionary Genomics of Transfer RNA Genes and SARS-CoV-2

Bryan Thornlow

Abstract

Transfer RNAs (tRNAs) are essential components of translation across all domains of life. The importance of this function is reflected in the strength of their conservation at the genome level, as well as their presence in hundreds of copies across each eukaryotic genome. Their strong conservation and high copy number at the genome level, in conjunction with their extensive post-transcriptional modifications and extreme variation in transcriptional activity by locus, make tRNA genes an enticing but as yet understudied model gene family.

The requirement of tRNA transcripts in exceptionally large quantities causes tRNA loci to experience among the highest rates of transcription in the genome. Consequently, transcription-associated mutagenesis (TAM) and natural selection leave distinct genomic signatures at highly transcribed tRNA loci, such that tRNA genes are strongly conserved despite elevated mutation rates, and their immediate flanking regions are among the most variable sites in the genome. Here, I characterize the relationship between expression, mutation, and selection at tRNA loci in detail by using population genetics, comparative genomics, epigenetics, and transcriptomic data. I then use these findings to engineer a random-forest model to predict tRNA gene transcriptional activity using only DNA data.

In the second half of this dissertation, I use the comparative genomics skills developed in the first part to help develop a novel phylogenetics toolkit. I identify the effects of sequencing errors on large SARS-CoV-2 phylogenies at global and local scales, demonstrate a novel method to quickly add samples to phylogenies, and explore recombination events in SARS-CoV-2 data, finding an excess in the region surrounding the Spike protein.

In this dissertation, I use publicly available DNA, RNA, and epigenetic data to develop novel bioinformatic analysis methods. Together, the conclusions drawn in this dissertation for both tRNA biology and SARS-CoV-2 answer fundamental evolutionary questions.

Acknowledgments

My parents, who nurtured my interest in learning, healthy habits in work and life, and courage to move across the country to pursue a Ph.D. in the first place. My older sister, Dana, has been an ideal role model growing up and I am forever grateful for her advice from her own Ph.D. experience. Most importantly, Dana was the first to give me an opportunity to pursue bioinformatics research when I was an undergraduate, which was the catalyst for my interest in graduate school and is perhaps the biggest turning point in my life.

My friends, especially those who have gone through graduate school along with me: Jon Akutagawa, Kishwar Shafin, Alison Tang, Roger Volden, Colleen Bosworth, Brandon Saint-John, Henry Gong, Akshar Lohith, many others. I am grateful for our trivia nights and board games days, and to have like-minded friends to vent to when things were rough and celebrate with when things were going well.

My labmates in both labs: Paloma Medina, Jakob McBroome, Max Genetti, Rachel Mendelson, Alex Kramer, Jesper Svedberg, Erik Enbody, Iskander Said, Joseph Yull, Evan Pepper, Cade Mirchandani, Erik Hanson, Landen Gozashti; Patricia Chan, Jon Howard, Andrew Holmes, Brian Lin, Alex Bagi, Aidan Manning, Qiuxia Tang, Jesse Leavitt. I could not have asked for a better working environment. I am grateful that we got to spend so much time together and have learned so much from you all. The undergraduates that I was fortunate enough to mentor: Jackie Roger, Skylar Kensinger, Tracey Ramirez, Vanessa Garcia, Trevor Ridgley. I am extremely lucky to have worked with such talented, hard-working people.

My "third lab": Yatish Turakhia, Angie Hinrichs, Nicola De Maio, Rob Lanfear, Cheng Ye, many others. Especially Yatish, whose ability to write an entirely new, vastly improved way of doing phylogenetics has inspired me to prioritize developing my programming skills as much as possible. Although I put more time and therefore effort into my tRNA work, I am perhaps most proud of our body of work on SARS-CoV-2, especially considering that it was

done without ever meeting in person. This was an important silver lining of the pandemic for me and I am extremely grateful to have been able to contribute.

The faculty who supported my first years of rotations, classes, seminars, and my advancement exam: David Haussler, Manny Ares, Grant Pogson, David Bernick, Mark Akeson, Ed Green, Josh Stuart, Angela Brooks, many others. The strength of this program is not only in its research but its mentorship, encouragement of good science, and the willingness of everyone involved to help others turn their weaknesses into strengths. UCSC has been my home away from home for five years now, and I am especially grateful for that.

Finally, my research advisors: Russ Corbett-Detig and Todd Lowe. I was lucky enough to have not only two of the best possible advisors, but advisors who complement each other's strengths perfectly. Russ's energy and drive and work ethic are legitimately unparalleled among anyone I have ever met, but thankfully also contagious, and I cannot imagine having worked as hard and gotten as much done in any other lab. I will never catch up to Russ in terms of accomplishments, but he has graciously taught me so much of what has made him so successful, and I am an infinitely better scientist for it. From Todd, I learned to enjoy and take pride in my work, and make sense of the "stories" that pop out of each discovery. Most importantly, his love of noncoding RNAs -- and extraordinary attention to detail -- has absolutely rubbed off on me, which is as valuable a gift as I could ever receive.

Introduction

Transfer RNAs (tRNAs) are essential to protein synthesis across all of life. Their primary function is in translation of the genetic code into the corresponding amino acid sequences that make up proteins. Thus, tRNA molecules are critical for virtually all cellular processes, and the genes encoding tRNA molecules have been highly conserved over evolutionary time (Tang et al. 2009; Chan and Lowe 2016). Mitochondrial tRNAs have been the subject of many studies, as mutations in these genes lead to a large number of maternally inherited genetic diseases (Suzuki, Nagao, and Suzuki 2011). However, eukaryotic genomes contain ~10- to 20-fold as many tRNA genes encoded in their nuclear chromosomes, which are required for cytosolic protein translation (Chan and Lowe 2016; Schimmel 2018). Despite their importance to the cell, there has been little study of evolutionary conservation or pathogenic mutations in cytosolic tRNA genes (Kutter et al. 2011; Ishimura et al. 2014). tRNAs are required in exceptionally large quantities, and therefore tRNA genes may experience greater levels of transcription than even the most highly transcribed protein-coding genes (Kirchner and Ignatova 2015; Molla-Herman et al. 2015). As the largest, most ubiquitous RNA gene family, cytosolic tRNAs constitute an ideal gene set for studying the interplay between natural selection and elevated mutation rates.

In this dissertation, I primarily explore tRNA conservation patterns at the DNA level, and describe how the extensive needs of protein translation shape the distribution of these essential genes across eukaryotic genomes. tRNA genes demonstrate exceptionally strong conservation, despite their high copy number. Curiously, their immediate flanking regions, which are transcribed but post-transcriptionally removed, are among the most *divergent* sites in the human genome, based on 100-way vertebrate alignments (Hubisz, Pollard, and Siepel 2011). In the first chapter of this dissertation, I combine comparative genomics, transcriptomics, and analyses of allele frequency spectra across multiple model species to

demonstrate that transcription-associated mutagenesis (TAM) subjects tRNAs to exceptionally elevated mutation rates.

The universality of both TAM and tRNAs across multicellular eukaryotes suggests a reverse-engineering of this signal to infer tRNA transcriptional activity using comparative genomics data alone. In the second chapter, I develop a random-forest model that uses only DNA data to predict tRNA gene activity with 94% accuracy. In the process, I explore DNA-based correlates to tRNA gene expression and find that proximity of tRNA genes to CpG islands is very strongly correlated, and most likely causative, of tRNA transcription. My approach demonstrates that both intrinsic -- based on the tRNA gene itself -- and extrinsic -- relating only to the tRNA gene's surroundings -- features are informative of transcriptional activity. I apply my model to 29 placental mammal species and provide predictions for over 10,000 tRNA genes.

It would be remiss not to mention the SARS-CoV-2 pandemic, a public health crisis that has claimed millions of lives and shaken the scientific community. My research priorities drastically shifted in early 2020. In the second half of this dissertation, I detail my analyses on SARS-CoV-2 genomic data. The breadth of whole-genome sequencing data necessitated development of an entirely new phylogenetics toolkit. In the third chapter, I describe a protocol developed by our research team to identify erroneous mutations. I generate simulations to demonstrate the effects of these mutations on inference of phylogenetic trees, as well as identify the effects of masking lab-specific errors on our global tree.

In the fourth chapter, I demonstrate the power of our software package, UShER, to add samples to existing phylogenies via maximum parsimony, without re-inferring the entire phylogeny. I demonstrate that UShER updates 40,000-sample phylogenies at speeds 3,000 times faster than the previous state-of-the-art. As the total number of sequenced SARS-CoV-2 genomes in the GISAID database climbed past 100,000, and more recently past 1,000,000, UShER has become the only viable option for SARS-CoV-2 phylogenetics. I simulate SARS-CoV-2 data with both random and systematic errors and demonstrate that it is

no more strongly affected by erroneous sites than its maximum-likelihood tree inference counterparts. I also briefly demonstrate several features included in matUtils, the sister package for manipulation and analysis of mutation-annotated trees created by UShER. Finally, I use these packages together to tell the story of a short-lived viral lineage identified in the Santa Cruz community in spring 2021.

A growing concern as novel Variants of Concern (VoCs) arise is the prevalence of recombination among SARS-CoV-2 lineages. Recombination requires a coinfection of two viral lineages to the same host and results in a novel lineage that is a mosaic of both. In the absence of recombination, viral lineages rely on mutation and selection alone to adapt to our protective measures, including lockdowns, masks, and medical treatments. Recombination has the capacity to produce VoCs much more quickly. To address these concerns, our group developed RIPPLES, which I describe in chapter 5. I pruned our global tree of likely erroneous samples and used RIPPLES to search for recombination on a 1,607,799-sample phylogeny. I develop and apply extensive post-hoc filtering strategies to remove putative false positives and identify 606 unique recombination events, encompassing 2.7% of samples in our phylogeny.

In the first half of this dissertation, I combine genomics approaches informed by both population genetics and RNA biology to elucidate the forces acting on a fundamental, ubiquitous gene family. In the second half, I use what I have learned about both population genetics and RNA biology to contribute to a more urgent cause. Overall, I use various comparative genomics and phylogenetics techniques to find fundamental insights to the evolution of both tRNA genes and SARS-CoV-2 lineages.

Chapter 1: Transfer RNA genes experience exceptionally elevated mutation rates

1.1: Background

Transcription affects the mutation rates of transcribed genes (Jinks-Robertson and Bhagwat 2014) through the unwinding and separation of cDNA strands (Gnatt et al. 2001). During transcription, a nascent RNA strand forms a hybrid DNA–RNA complex with a template DNA strand. While the complementary tract of nontemplate DNA is temporarily isolated, it is chemically reactive and thus accessible by potential mutagens (Gnatt et al. 2001). Transcription can lead to the formation of noncanonical DNA structures, which can hinder repair pathways and promote errors by the polymerase (Gaillard and Aguilera 2016). The RNA strand can also reanneal to the template DNA strand, prolonging isolation and increasing vulnerability to mutations (N. Kim and Jinks-Robertson 2012; Aguilera and García-Muse 2013). Furthermore, if transcription and DNA replication occur concomitantly at a particular locus, collisions between RNA polymerase and the DNA replication fork may also damage DNA (Gaillard and Aguilera 2016; Jinks-Robertson and Bhagwat 2014; Helmrich, Ballarino, and Tora 2011). In human cancer cells, increased transcription and replication induce torsional stress and collisions (Gaillard and Aguilera 2016).

Several cellular agents have also been shown to cause damage in highly expressed genes (Timakov et al. 2002). Among the most notable sources of mutation associated with transcription is activation-induced cytidine deaminase (AID) (Gómez-González and Aguilera 2007). AID accompanies RNA polymerase II and deaminates cytosine nucleotides. To resolve the resulting base-pair mismatch, the opposing guanine is converted to adenine and uracil to thymine, resulting in excess C→T mutations on the nontemplate strand and excess G→A mutations on the template strand (Jinks-Robertson and Bhagwat 2014; Green et al. 2003). AID is a member of the APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like)

gene family, many of which are involved in double-stranded break repair in transcription (Jinks-Robertson and Bhagwat 2014). Some members of the APOBEC family act strongly on short genes, suggesting increased activity at tRNA loci (Taylor, Wu, and Rada 2014; Saini et al. 2017). For example, APOBEC3B causes 1,000-fold more DNA damage at tRNA loci than at other genomic regions in yeast (Saini et al. 2017). AID also acts on highly transcribed genes in immune B cells, causing transition mutations and double-stranded breaks (Jinks-Robertson and Bhagwat 2014). Due to the strong association of the APOBEC family with transcription, relative excesses of C→T and G→A mutations are a signature of TAM (Jinks-Robertson and Bhagwat 2014).

To conserve mature tRNA sequence identity in the presence of an elevated mutation rate, tRNA genes should experience strong purifying selection. tRNA transcription requires sequence-specific binding of transcription factors to the internal box A and box B promoter elements (White 2011). Once transcribed, precursor tRNAs must fold properly to undergo maturation, which can be disrupted by sequence-altering mutations. The unique structure of tRNAs dictates processing by RNases, addition of modifications, accurate recognition by aminoacyl tRNA synthetases, incorporation into the translating ribosome, and accurate positioning of the anticodon relative to mRNA codons (Zhang and Ferré-D'Amaré 2016). Because of the need to maintain sequence specificity, DNAs encoding the mature portions of tRNAs are well conserved (Zhang and Ferré-D'Amaré 2016). Therefore, we expect that a large proportion of mutations arising in tRNA genes will be deleterious and will quickly be purged by natural selection.

While most human tRNA genes do not have external promoters (White 2011; Zhang and Ferré-D'Amaré 2016), tRNA transcripts include leader and trailer sequences extending roughly two to five nucleotides upstream and 5–15 nucleotides downstream of the annotated mature tRNA gene, based on the position of the genomically encoded poly(T) transcription termination sequence. Aside from the termination sequence, these flanking sequences appear to have limited sequence-specific functionality in most cases (Ziehler et al. 2000;

Hopper 2013; Hasler et al. 2016; Lee et al. 2009). Very early in maturation, all tRNA flanking sequences are removed by RNase P (Hopper 2013; Ziehler et al. 2000) and RNase Z (Richard J. Maraia and Lamichhane 2011). Because these flanking genomic sequences are frequently unwound and therefore vulnerable to TAM, we expect that these regions will experience mutation rates similar to those of tRNAs. Whereas tRNA genes should experience purifying selection, the flanking regions should be neutral or under weak selection. Here we investigate the patterns of conservation, divergence, and within-species variation of cytosolic tRNAs in humans and other model organisms to elucidate the forces shaping the evolution of this essential RNA gene family.

The text of this dissertation includes reprints of the following previously published material: Thornlow, B. P., Hough, J., Roger, J. M., Gong, H., Lowe, T. M., & Corbett-Detig, R. B. (2018). *Transfer RNA genes experience exceptionally elevated mutation rates. Proceedings of the National Academy of Sciences,* 115(36), 8996-9001. The co-authors listed in this publication directed and supervised the research which forms the basis for the dissertation.

1.2: Flanking regions of tRNA genes are highly variable despite strong conservation of mature tRNA sequences.

To estimate evolutionary conservation, we examined phyloP, which measures the conservation of each human genomic position across 100 vertebrate species (Pollard et al. 2010), by position within each tRNA locus (Methods). Positive phyloP scores indicate strong conservation, and negative scores indicate accelerated evolution. To study the effects of evolution on a shorter timescale, we also estimated sequence divergence between *Homo sapiens* and *Macaca mulatta* at each tRNA locus. Mature tRNA sequences are highly conserved across all positions, based on both average phyloP score (Figure 1.1A) (Pollard et al. 2010) and *M. mulatta* alignment (Figure 1B). However, the inner 5′ flanking region (20 bases upstream of the tRNA; see Methods) is roughly four times more divergent than the

untranscribed reference regions. We also find increased rates of divergence in the inner 3′ flanking region, which is roughly three times more divergent than the reference regions (Figure 1B). Both the outer 5′ flank (21–40 bases upstream of the tRNA) and the outer 3′ flank (11–40 bases downstream of the tRNA) are also roughly 1.5 times more divergent than the reference regions. For tRNAs that contain introns (Chan and Lowe 2016), we find that intronic variation correlates with flanking variation. Furthermore, intergenic regions within clusters of active tRNAs show similar patterns in their phyloP scores.

We also studied population-level variation at low-frequency SNPs (minor allele frequency <0.05) for each tRNA locus. Low-frequency SNPs are evolutionarily young and are less affected by selection (Messer 2009). Consistent with our divergence analyses, we find that low-frequency SNPs are more common across both the tRNA gene sequence and flanking regions than in untranscribed reference regions (Figure 1.1C). Although the inner flanking regions are most polymorphic, the mature tRNA sequences have about twice as many low-frequency SNPs as reference regions. Overall, our results are consistent on multiple timescales, indicating that tRNAs and flanking sequences are prone to mutation. Indeed, of the 247 sites in the genome that have phyloP scores of −20 -- the lowest possible score (Pollard et al. 2010; Karolchik et al. 2004) -- 14 are within 10–15 bases upstream of the start of an active tRNA gene, indicating disproportionate enrichment (hypergeometric test, P < 1.65e-48) and that tRNA flanking regions are among the least conserved in the genome. Nonetheless, mature tRNA gene sequences are strongly conserved by purifying selection, which purges mutations.

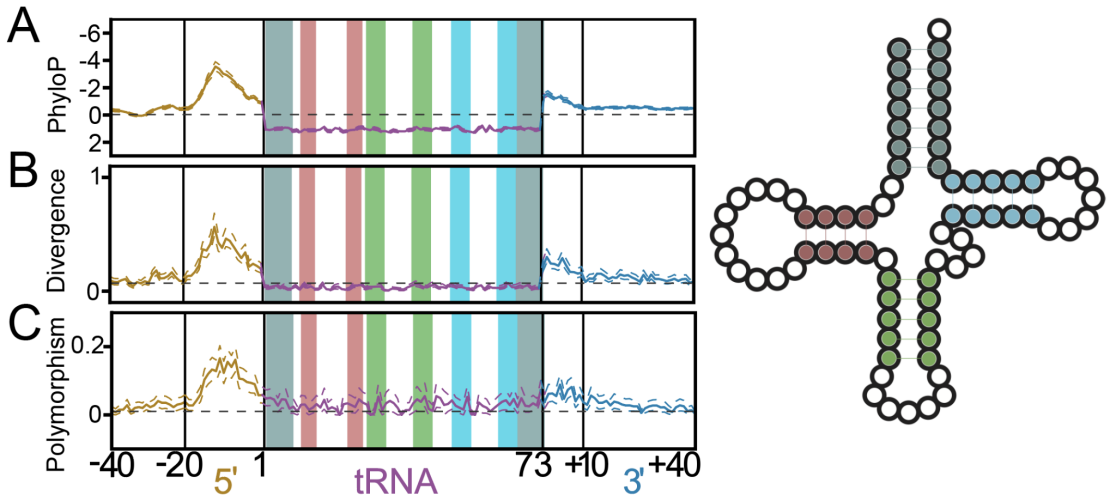


Figure 1.1: tRNA genes are strongly conserved despite extremely divergent immediate flanking regions. *There is a strong pattern of variation in regions flanking human tRNA genes by three measures: relative to vertebrates, by comparison with Rhesus macaque alone, and within the human population. (A) The average phyloP score (comparing humans to 100 vertebrate species) is plotted for each position within the tRNA and flanking region across all human tRNAs. (B) Divergence between the human and M. mulatta tRNA genes and their flanking regions. (C) Frequency of low-frequency SNPs (minor allele frequency ≤ 0.05) across all human tRNAs. The acceptor stem (gray), D-stem (red), anticodon stem (green), and T-stem (blue) are highlighted within the tRNA both in the linear plots and in the 2D structure legend to the right (Chan and Lowe 2016; Sprinzl et al. 1998). Nucleotide numbering below the plots is relative to mature tRNA boundaries, with inner and outer flanks demarcated by a shift in mutation rate (Methods). Dotted lines surrounding plots depict 95% confidence intervals calculated by nonparametric bootstrapping by tRNA loci.*

1.3: Transcription is correlated with variation in tRNA and flanking regions.

We hypothesized that, if transcription-associated mutagenesis drives variation among tRNA loci, highly active tRNA genes would show the greatest mutation rates. Because tRNA transcript abundance measures are often not attributable to individual loci due to identical gene copies and difficulty sequencing full-length tRNAs, we estimated relative transcriptional activity based on chromatin state data from the Epigenomic Roadmap Project (Roadmap Epigenomics Consortium et al. 2015). Based on these data, we classified human tRNA genes as “active” if they are located in expressed regions in several cell lines and otherwise as “inactive” (Methods and Figure 1.2). In some cases, multiple cell lines correspond to a single

tissue or organ, so tissue-specific tRNAs [*e.g.*, the brain-specific arginine tRNA in mouse (Ishimura et al. 2014)] are considered active.

We find that active tRNA genes are significantly more conserved than inactive tRNA loci (Mann–Whitney U test, P < 8.40e-53), and the flanking regions of active tRNAs are significantly more divergent than the flanking regions of inactive tRNAs (P < 7.98e-61). The peak measure of divergence between human and *M. mulatta* tRNA genes in the inner 5′ flanking regions is roughly five times greater in active tRNAs than in inactive tRNAs (Figure 1.2E-F). Active tRNAs in human populations also have significantly more low-frequency SNPs per site than inactive tRNAs across the entire locus, including the tRNA and flanking regions (P < 3.72e-36). Inactive tRNAs are still significantly more conserved (P < 2.02e-12) and polymorphic (P < 0.007) than the untranscribed reference regions, and their flanks are significantly more divergent than the reference regions (P < 1.36e-16).

That the peak in both divergence and polymorphism in all species is consistently 12–15 nucleotides upstream of the mature tRNA sequence is curious. At the most divergent position, 55% of all tRNA loci differ between human and *M. mulatta*, and 15% of human tRNA loci have a low-frequency SNP (Figure 1.1). Furthermore, virtually all active tRNA loci differ at this nucleotide between human and *M. mulatta*, and 25% have a low-frequency SNP at this site. This implies that this region either does not face uniform selective pressures or is not uniformly vulnerable to TAM. While distant flanking sequences can affect tRNA expression in yeast (Bloom-Ackermann et al. 2014), few studies have shown that flanking regions affect expression in higher eukaryotes (Doran, Bingle, and Roy 1988). Transcription initiation is long relative to elongation (Dieci and Sentenac 1996; Graczyk, Cieśla, and Boguta 2018), which may lead to prolonged isolation of the nontemplate DNA strand at the initiation site and increased vulnerability to TAM. A poised initiation complex might also increase the likelihood of collisions between Pol3 and the replication fork (Helmrich, Ballarino, and Tora 2011). Thus, frequent initiation at highly transcribed tRNA loci may contribute to the nonuniform pattern of variation. The crystal structure of a tRNA transcription initiation complex in yeast indicates

that the transcription bubble extends outward to ~12 nt upstream of the start of the tRNA gene (Abascal-Palacios et al. 2018), indicating that torsion may play a role in the elevated mutation rate in this region.

TAM may also explain the increased variation in the outer 3′ flank relative to the outer 5′ flank, as positioning of downstream transcription termination sites varies among tRNA genes (Chan and Lowe 2016; Orioli et al. 2011), whereas transcription start site positions are more consistent. While most tRNAs do not have clear TATA boxes, the TATA-binding protein (TBP) still binds to the DNA duplex ~25 nucleotides upstream of the tRNA (Juo et al. 1996), which coincides with a decrease in variability. Furthermore, while both flanking regions for many other Pol3-transcribed genes are divergent, the 5′ flanking regions are generally more divergent than the 3′ flanking regions, suggesting that the underlying mechanism is not tRNA-specific (Figure 1.3).

Two orthogonal analyses strengthen the observed correlations between gene expression and variation at tRNA loci. First, we find a significant correlation between the TBP intensity peaks (ENCODE Project Consortium 2012, 2011; Kharchenko, Tolstorukov, and Park 2008) and conservation of the mature tRNA sequence (Spearman's rho = 0.64, $P < 2.2e-16$) across all human tRNAs and the opposite relationship in the flanking regions (Spearman's rho = −0.64, $P < 2.2e-16$) (Figure 1.2). TBP ChIP-sequencing (ChIP-seq) data directly reflect transcriptional activity for each locus, as its occupancy is significantly correlated with and required for transcription (White 2011; Roberts et al. 2003; Mason and Struhl 2003; Kuras et al. 2000; Zanton and Pugh 2004; X. Y. Li et al. 1999). Second, mature tRNA sequence read counts are strongly correlated with tRNA conservation (Spearman's rho = 0.18, $P < 0.001$) and flanking region divergence (Spearman's rho = −0.61, $P < 2.2e-16$) (Figure 1.2). These read counts were collected from a single HEK cell line by Zheng et al. (Zheng et al. 2015) using DM-tRNA-seq, a specialized tRNA-sequencing method. These correlations are consistent with the idea that more highly transcribed tRNAs vary more in their flanking regions.

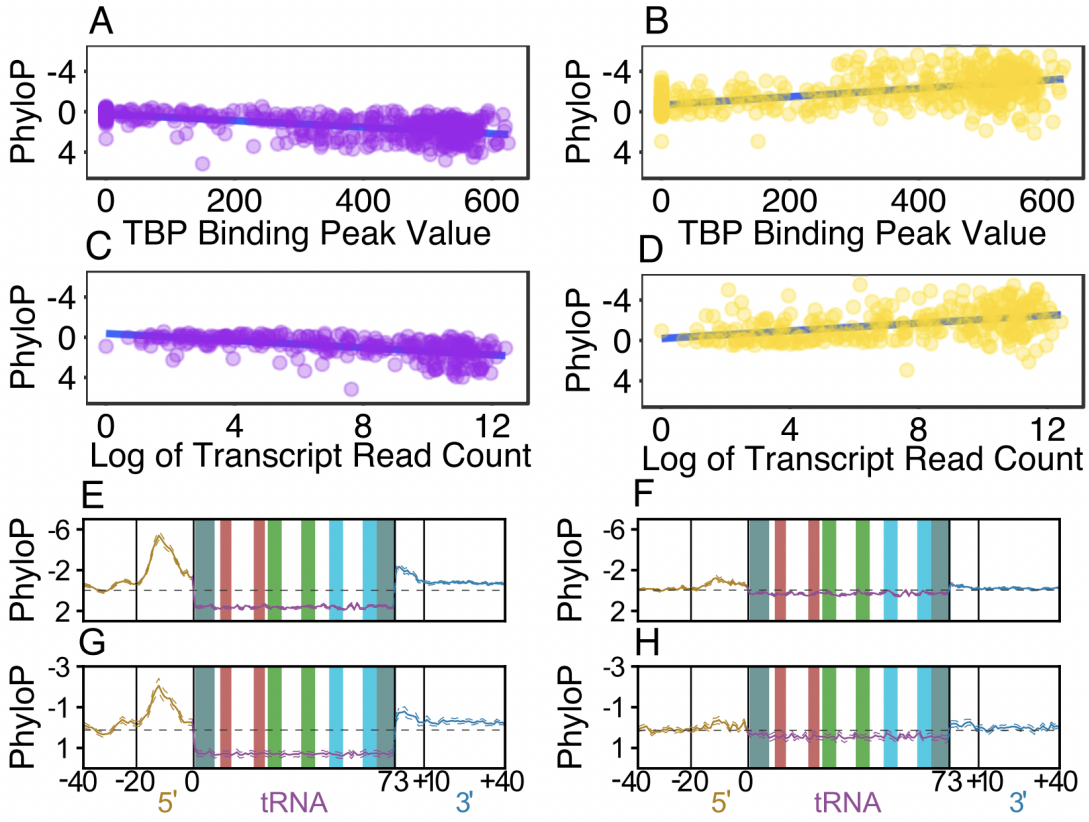


Figure 1.2: tRNA expression is significantly correlated to both tRNA conservation and flanking region divergence. *(A and B): TBP peak value (expression) is plotted versus phyloP score (conservation) for each mature tRNA (A) and adjacent inner 5' flanking region (B). (C and D): Log of the HEK293T cell DM-tRNA-seq read count (expression) (46) is plotted versus phyloP score (conservation) for each gene encoding a unique mature tRNA sequence (C) and the corresponding inner 5' flanking region (D). Both TBP occupancy and transcript abundance are greater for highly conserved mature tRNA loci (A and C) and those with the most divergent flanks (B and D). (E and F) Plotted as in Figure 1, human tRNA loci that are separated into active (E) versus inactive (F) groups show the characteristic differences seen in A–D. (G and H) Mouse tRNA loci split into active (G) versus inactive (H) groups show a pattern strikingly similar to that seen in human (A–F).*

1.4: Variation patterns observed at tRNAs are observed in few other gene families.

Among the histone protein-coding genes less than 1,000 nucleotides in length, the average phyloP score per nucleotide across the coding sequence and flanking regions is 3.449 and −2.052, respectively, comparable to tRNA loci. Histone protein coding genes are transcribed by RNA Polymerase II, but are not polyadenylated, often do not have introns, and

are highly transcribed at specific times in the cell cycle (Rattray and Müller 2012). However, most genes transcribed by RNA Polymerase II do not appear to demonstrate TAM in the same way as tRNAs. For example, ribosomal proteins are very highly transcribed (Thul et al. 2017) and have well-conserved exons, but their introns and flanking regions are not as divergent as tRNA flanking regions (Pollard et al. 2010; Casper et al. 2018). tRNAs are likely ideal for studying TAM because they have predictable transcript start and end sites, internal promoters, and high transcription rates. Other small non-coding RNAs transcribed by RNA Polymerase III demonstrate signals suggestive of TAM, similar to tRNAs (*e.g.* spliceosomal RNAs (snRNAs) and YRNAs, Figure 1.3).

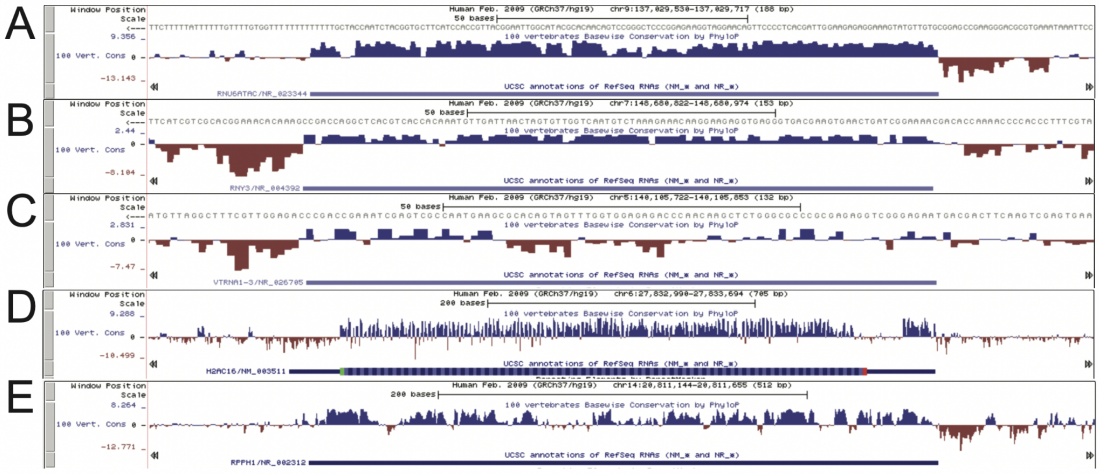


Figure 1.3: Histone protein coding genes and RNAs transcribed by Pol III also demonstrate phylogenetic signals consistent with TAM. *UCSC Genome Browser screenshots for a snRNA (A), Y-RNA (B), vault-RNA (C), histone protein coding gene (D) and RNAse P (E). Note the scale on the left side of each panel, as phyloP scores in immediate flanking regions of each locus reach lower than -5. The phyloP valleys shown here indicate strong divergence, similar to what is observed flanking highly transcribed tRNA genes.*

1.5: Patterns of low-frequency SNPs are consistent with TAM.

In TAM, repair pathways activated in response to deaminations lead to excess conversions between guanine and adenine and between thymine and cytosine on the coding strand (Jinks-Robertson and Bhagwat 2014; Green et al. 2003). Across all tRNA loci, we

found that the most common low-frequency SNPs are C→T and G→A and that these mutations are significantly more common in both tRNA genes and flanking regions than in untranscribed reference regions (Fisher's exact test, $P < 0.05$ for all comparisons) (Figure 1.4). Removal of CpG sites (Schmidt et al. 2008) does not significantly affect these results. The relative excesses of these SNPs are much more pronounced in active tRNA loci than in inactive tRNA loci (Fig 1.5). These results suggest that deamination of the noncoding strand due to TAM and the DNA repair mechanisms acting in response to deamination is especially common at these loci (Jinks-Robertson and Bhagwat 2014; Green et al. 2003; Saini et al. 2017).

It is difficult to discern whether this increased prevalence is due to TAM or selection to preserve the structural integrity of the tRNA. To preserve tRNA secondary structure, we expect transition mutations (e.g., A–U to G–U base pairs, C–G to U–G base pairs) to be more common than transversions, as they should disrupt stem helices less often. However, the mutational skew expected of regions affected by TAM is stronger in regions flanking tRNAs. Transcription initiation is relatively long compared with elongation (Dieci and Sentenac 1996; Graczyk, Cieśla, and Boguta 2018), which might contribute to increased mutagenesis by APOBEC enzymes or more collisions (Helmrich, Ballarino, and Tora 2011) or double-stranded breaks. However, divergence at tRNA flanking regions is correlated with divergence at introns in both human (Spearman's rank, rho = 0.734, $P < 5.58e-6$) and mouse (rho = 0.733, $P < 5.24e-4$), indicating similar mutation rates across tRNA loci. Our results therefore suggest that TAM drives the excess of transitions among low-frequency SNPs across tRNA loci.

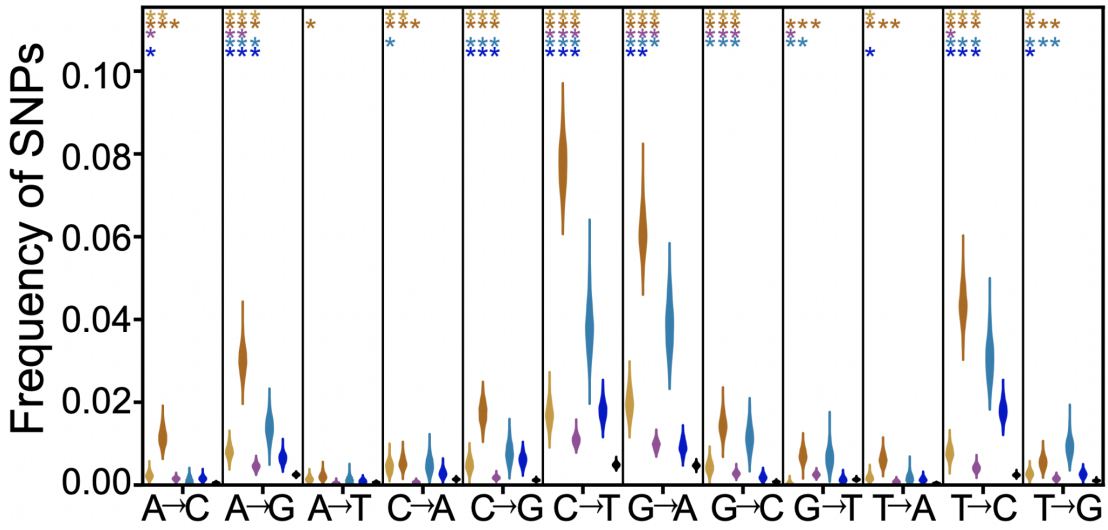


Figure 1.4: The SNP classes most common in regions affected by TAM are also most common at tRNA loci. *The distribution of each class of low-frequency polymorphisms by region across all human tRNAs. Stars indicate the significance levels of Fisher's exact tests comparing the SNP distribution within each region of the tRNA and flank (outer 5' flank in yellow, inner 5' flank in orange, tRNA in purple, inner 3' flank in cyan, outer 3' flank in blue) with that of the untranscribed reference region (black): one star, P ≤ 0.05; two stars, P ≤ 0.005; and three stars, P ≤ 0.0005.*

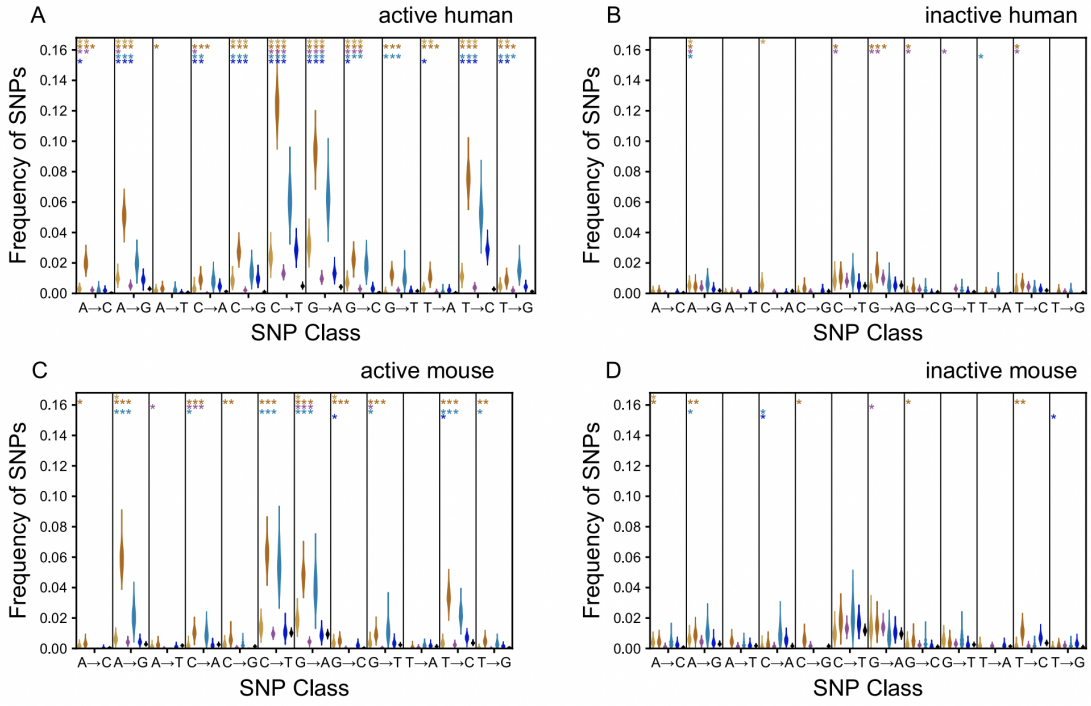


Figure 1.5: Excess SNP classes consistent with TAM are more common in active tRNAs in both human and mouse. *The distribution of each class of low-frequency polymorphisms, defined as a SNP with a minor allele frequency less than or equal to 0.05, is shown by region across active human tRNAs (A), inactive human tRNAs (B), active mouse tRNAs (C) and inactive mouse tRNAs (D). As in Figure 1.4, the significance levels of Fisher's exact tests comparing the SNP distribution within each region of the tRNA and flank (outer 5' flank is yellow, inner 5' flank is orange, tRNA is purple, inner 3' flank is cyan, outer 3' flank is blue) to that of the untranscribed reference region (black), are represented by stars at the top of each panel. One star represents a p value ≤ 0.05, two stars represents a p value ≤ 0.005, and three stars represents a p value ≤ 0.0005.*

1.6: tRNA flanking region variation in other model organisms is consistent with variation observed in humans.

To confirm that our results are not restricted to humans, we also analyzed tRNAs in *Mus musculus, Drosophila melanogaster*, and *Arabidopsis thaliana*. We find similar patterns of sequence conservation of tRNA loci in each when measuring phyloP or divergence to outgroups. The 5′ flanks are consistently more divergent than the 3′ flanks, and the most divergent sites are roughly 10–15 bases upstream of the tRNA in all species. We also used ChIP data across nine mouse tissues to classify mouse tRNAs based on their expression (Bogu et al. 2015). Active mouse tRNAs are more strongly conserved than their inactive counterparts (Mann–Whitney U test, P < 1.81e-19), and their flanks are more divergent (P < 7.04e-22) (Figure 1.2G-H), consistent with our results from the human data (Figure 1.2E-F). Active mouse tRNAs also have more low-frequency SNPs in their flanking regions than inactive mouse tRNAs (P < 2.23e-4). Such consistency suggests that a shared underlying molecular mechanism drives these patterns of sequence variation.

Low-frequency SNPs in the tRNA gene sequences also follow qualitative patterns similar to those in the human data. We observe excess transitions in all species studied, and active mouse tRNAs show a greater excess of low-frequency transitions than do inactive mouse tRNAs (Fig 1.5C-D). However, these patterns vary across species (Figure 1.6B). For example, in mice, tRNA genes have more low-frequency SNPs than the untranscribed reference regions, but the opposite is true in *D. melanogaster.* Low-frequency SNPs are thought not to be strongly affected by selection (Messer 2009), but selection is more efficient

in species with greater effective population sizes (Figure 1.6A). Effective population size (Tenesa et al. 2007; Phifer-Rixey et al. 2012; Cao et al. 2011; Shapiro et al. 2007) and tRNA copy number vary across species, and because the sample sizes and data quality differ among population samples, these differences may be attributable to differences in the impact of selection or in ascertainment of low-frequency variation.

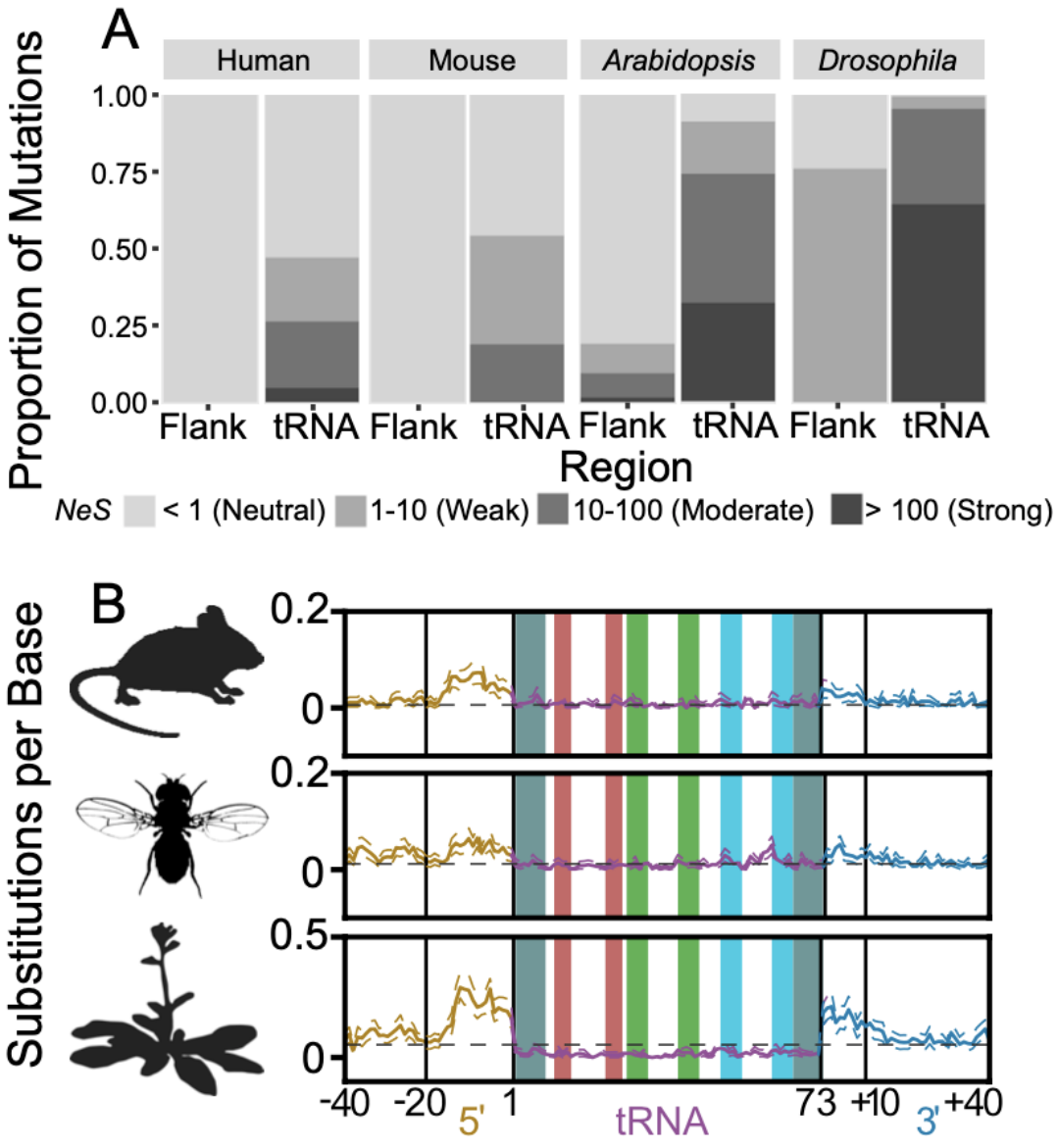


Figure 1.6: tRNA flanking regions are subject to high mutation rates and minimal selection. *The estimated DFE indicates that high proportions of deleterious mutations in tRNAs are under strong selection. (A) Estimated DFE of new deleterious mutations for tRNA genes and inner 3' flanking regions shown in human, mouse, A. thaliana, and D. melanogaster. Proportions of deleterious mutations are shown for each bin of purifying selection strength, estimated on a scale of NeS. Species are arranged by increasing effective population size. (B) Low-frequency SNPs plotted as in Figure 1C for mouse, A. thaliana, and D. melanogaster.*

1.7: Functional tRNA sequences experience strong purifying selection in all species studied.

Our analysis of the distribution of fitness effects (DFE) of deleterious mutations demonstrates that tRNAs evolve under strong purifying selection in all analyzed species. In contrast, regions flanking tRNAs are inferred to be either neutral or subject to weak selection (*NeS* < 10, where *Ne* is the effective population size and *S* is the strength of selection) (Figure 1.6A). Our estimates of the proportions of new mutations falling into each *NeS* range of the DFE for tRNAs indicate far fewer nearly neutral mutations (*NeS* <1) and substantially more strongly deleterious mutations (*NeS* >100) in *D. melanogaster* and *A. thaliana* than in the human or mouse populations (Figure 1.6A). Given that estimates of effective population size in humans (7,000) (Tenesa et al. 2007) and mouse (25,000–120,000) (Phifer-Rixey et al. 2012) are substantially lower than in *A. thaliana* (300,000) (Cao et al. 2011) and *D. melanogaster* (>1,000,000) (Shapiro et al. 2007), this difference in strength of selection may partially reflect differences in effective population size and might explain the differences in low-frequency SNPs in tRNA loci across species (Figure 1.6B). In turn, this might indicate that the strength of purifying selection, independent of effective population size, at tRNA loci is consistent across diverse species.

The strength of selection across species may also reflect the number of unique tRNA gene sequences in each genome. For example, roughly half of all human tRNA genes have unique sequences, but the majority of *D. melanogaster* tRNAs have identical copies (Chan and Lowe 2016). tRNAs with the same anticodon but different sequences may have different functions, and this may affect strength of selection at each locus as well. Indeed, a

significantly greater proportion of sites are invariant (Fisher's exact test, P < 7.50e-5) and fewer sites are divergent (P < 3.85e-8) in active single-copy human tRNA genes than in active multicopy human tRNA genes. We observe the same patterns in the inner 5′ (P < 5.87e-5; P < 0.025) and inner 3′ (P < 8.90e-5; P < 4.04e-4) flanks of active tRNA genes, suggesting increased transcription of active multicopy tRNA genes. However, few SNP data are available for multicopy tRNAs compared with single-copy tRNAs, limiting our ability to identify consistent differences among tRNA subgroups.

1.8: tRNA loci contribute disproportionately to mutational load.

Our discovery of a highly elevated mutation rate at tRNA loci suggests that tRNA genes may contribute disproportionately to mutational load, the reduction in individual fitness due to deleterious mutations (Haldane 1937; Agrawal and Whitlock 2012). To estimate the relative mutation rates at active tRNA loci, we calculated the average ratios of θ for the inner 3′ and 5′ flanking regions of active human tRNA genes to the untranscribed reference regions using the approach of Messer (Methods) (Messer 2009). We estimate θ in the flanking regions instead of the tRNAs because strong selection can cause underestimation of θ (Messer 2009), and our results indicate that active human tRNAs are subject to strong selection while the flanking regions are likely selectively neutral (Figure 1.6A). We therefore estimate that the mutation rate is between 7.24 (inner 3′; 95% CI 7.12–7.33) and 10.36 (inner 5′; 95% CI 10.16–10.41) times greater at tRNA loci than the genome-wide average. Given that there are 25,852 base pairs of active human tRNA sequence, and using 1.45e-8 as the genome-wide mutation rate (Narasimhan et al. 2017), we estimate that U (the genome-wide rate of deleterious mutation per diploid genome) contributed by tRNAs is between 0.0054 and 0.0078. Since active tRNAs make up only 0.0009% of the human genome (Chan and Lowe 2016), this implies that mutations in tRNAs contribute disproportionately to mutational load. Our findings highlight that mutations at tRNA loci are likely an important source of fitness and disease variation in human populations.

1.9: Methods

We used tRNA coordinates from GtRNAdb (Chan and Lowe 2016) for the human, *M. musculus, D. melanogaster,* and *A. thaliana* genomes. For each species, we defined untranscribed reference regions by searching 10 kilobases upstream of each tRNA and selecting a 200-nucleotide tract. If this tract was within a highly transcribed region of the genome [based on genome-wide ChIP data (Bogu et al. 2015)], overlapped a conserved element [defined as a region with a phastCons log odds score greater than 0 (Pollard et al. 2010)], was within 1,000 nucleotides of a known gene (Casper et al. 2018), or overlapped a reference region assigned to another tRNA, we selected a different tract 1,000 bases further upstream and repeated the selection until we found an acceptable region. For the mouse genome, we checked known genes, previously assigned reference regions, and conserved elements. For the *D. melanogaster* and *A. thaliana* genomes, we began our searches only 1,000 bases upstream of each tRNA and searched for 200-nucleotide tracts that were at least 100 nucleotides away from any annotated genetic element (Lack et al. 2015, 2016; Cao et al. 2011) due to the high functional densities of these species' genomes.

For each tRNA in all species, we defined the inner 5′ flank as the 20 bases immediately upstream of the 5′ end of the tRNA gene on the coding strand and the outer 5′ flank as the 20 bases directly upstream of the inner 5′ flank. The inner 3′ flank refers to the 10 bases downstream of the tRNA gene, and the outer 3′ flank refers to the 30 bases downstream of these 10 bases. We made these decisions based on inflection points in our data, as the flanking regions up to 20 bases upstream and 10 bases downstream of tRNA genes have less variation. Transcription usually ends about 10 bases downstream of tRNA genes (Orioli et al. 2011).

The Roadmap Epigenomics Consortium compiled genome-wide epigenomic data across 127 human tissues and cell lines to characterize the chromatin state across the

genome (Roadmap Epigenomics Consortium et al. 2015). We analyzed the regions surrounding each tRNA in each epigenome sample and used clustering to classify each genomic region according to its most common epigenomic state. We classified all human tRNAs based on the epigenomic state annotation in the genome. In the corresponding model, regions in state 1 are likely to be transcribed. The 342 tRNAs in state 1 in at least 4 of the 127 tissues analyzed are active tRNAs, and we consider the remaining 254 tRNAs to be inactive. To classify mouse tRNAs, we used a 15-state Hidden Markov Model based on ChIP data in which states 5 and 7 corresponded to regions near active promoters (Bogu et al. 2015). We considered the 272 tRNAs in genomic regions annotated as state 5 or 7 in at least 3% of tissues as active and the remaining 188 tRNAs as inactive.

We aligned all tRNAs across all species using covariance models (Chan et al., n.d.) and assigned coordinates to each position in each tRNA and flank based on the Sprinzl numbering system (Sprinzl et al. 1998). We averaged the phyloP, divergence, and low-frequency SNP data for all sites assigned to the same Sprinzl coordinate for their respective tRNA loci. Because some tRNAs have variations in structure (Chan and Lowe 2016), this alignment was necessary for position-wise comparisons between tRNAs. We filtered tRNAs with fewer than 50 aligned bases from our analyses. If a conserved element (regions with a phastCons log odds score greater than 0; (Pollard et al. 2010)) was present 4–10 bases up- or downstream of a tRNA, the tRNA was excluded from our analyses, as these regions might contribute to the secondary structure of mature tRNAs and be subject to anomalous levels of selection. We also excluded nuclear-encoded mitochondrial tRNA genes.

We aligned the hg19 human reference genome to the *Macaca mulatta* reference genome (rheMac2; (Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007)), both from the UCSC Genome Browser (Casper et al. 2018). We also compared the mouse (*Mus musculus*, mm10) and rat (*Rattus norvegicus*, rn6) genomes, and the *A. thaliana* (TAIR10) and *A. lyrata* (v.1.0) genomes (Kersey et al. 2016) using the same methods. For *D. melanogaster*, we used an alignment of the dm6 and droYak2 (*D. yakuba*) genomes

(Drosophila 12 Genomes Consortium et al. 2007). Non-gap nucleotide mismatches in the alignments were classified as divergent sites. To account for the possibility that multiple substitutions occurred at a single site, we used a Jukes-Cantor correction (Jukes, Cantor, and Others 1969).

We analyzed human variation data from the African superpopulation of 661 humans from phase 3 of the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015). We acquired *D. melanogaster* variation data for the Siavonga, Zambia populations from the Drosophila Genome Nexus Database (Lack et al. 2015, 2016). We obtained M. musculus and *A. thaliana* data from (Consortium and Mouse Genome Sequencing Consortium 2002) and (Arabidopsis Genome Initiative 2000), respectively. All nonhuman data were aligned and genotypes curated as described in (Corbett-Detig, Hartl, and Sackton 2015).

Within each gene, flank, or reference region, we considered positions with minor allele frequencies between 0 and 0.05 to be low-frequency SNPs. We also determined the frequency each class of mutations (e.g., A→G) within each region of each tRNA locus where the identity of each base is defined according to the coding strand sequence. We found the frequency of divergences and low-frequency SNPs by position across all tRNAs and flanking regions. For conservation studies across multiple species, we used the phyloP track (Pollard et al. 2010) from the University of California, Santa Cruz (UCSC) Genome Browser (48) and calculated the average score for each position within the tRNAs and flanking regions. No phyloP data were available for *A. thaliana* (Pollard et al. 2010). For direct comparisons between the species of interest and an outgroup, we used the Multiz track from the UCSC Table Browser (Karolchik et al. 2004) and the Stitch MAFs tool from Galaxy (Afgan et al. 2016) to create sequence alignments. Details are available in SI Appendix.

The ENCODE Project Consortium used ChIP-seq data to identify binding regions for regulatory factors (ENCODE Project Consortium 2012) including the TBP and Pol3 transcription factors in the human genome (White 2011). These data were taken from the UCSC Genome Browser (Casper et al. 2018). The intensity of a given peak correlates with a

greater frequency of transcription factor binding to that region. For each human tRNA, we found the strongest TBP peak in the 50 base pairs immediately upstream of the tRNA across the GM12878, H1-hESC, HeLa-S3, HepG2, and K562 cell lines. We also calculated the average phyloP score across the flanking regions for each tRNA and used Spearman's rank correlation test on these data.

We used demethylation sequencing data for tRNAs within HEK293T cells from (Zheng et al. 2015). We used Spearman's rank correlation tests to correlate mature tRNA transcript read counts and tRNA and flanking region conservation. Because Zheng et al. sequenced mature tRNAs, which are often encoded by multiple genes, we excluded identical genes to control for the correlation between gene copy number and overall expression (Figure 1.2 C-D). Separately, we summed the average phyloP scores at these loci and correlated the summed scores to total tRNA read counts.

We estimated the DFE for each species using the method of (Keightley and Eyre-Walker 2007) and the DFE-α software. The DFE estimation method is based on site frequency spectra (SFS) obtained from within-species SNP data, and assumes a simple model of recent demographic change to correct the SFS at functional sites for skews caused by demography. We used a two-epoch model of demographic change and estimated the DFEs for tRNAs and inner 3' flanking regions for each species. Each class of sites was assumed to be subject to mutation, selection and drift, with gamma-distributed DFEs and an initial shape parameter (β) of 0.5.

We used the equation $\theta(k) = kG^k$ (Messer 2009) (Equation 22) to estimate the mutation rate at active tRNA loci. We calculated the ratios of θ in active tRNA flanking regions to θ in the reference regions for k = 1,2,3 and bootstrapped by tRNA loci to calculate 95% CIs. We calculate the maximum-likelihood estimator for θ (4Nu) from low-frequency sites for untranscribed reference regions and flanking regions of active tRNA genes. The ratios of θ_{flank} : $\theta_{reference}$ should then provide estimates for the ratio of the mutation rates in these regions. If we then assume the reference regions have a mutation rate equal to 1.45e-8 per site per

generation (Narasimhan et al. 2017), multiplying by this ratio yields an estimate of the per site per generation mutation rate at tRNA loci. To calculate U_{tRNA}, or the contribution of mutations at active tRNA loci to the genome-wide rate of deleterious mutation per diploid genome, we multiply the human genome-wide mutation rate per nucleotide per haploid genome (1.45e-8; (Narasimhan et al. 2017)), times 2 to correct for diploidy, times the number of nucleotides in active human tRNAs (25,852), times the ratios of $\theta_{flank} : \theta_{reference}$ (7.24 for inner 3', 10.36 for inner 5'). Using these ratios, we estimate that UtRNA is between 0.0054 and 0.0078.

Chapter 2: Predicting transfer RNA activity from sequence and genome context

2.1: Background

tRNA molecules are required in large abundance to meet the dynamic metabolic needs of cells, and tRNA genes are believed to be among the most highly transcribed genes in the genome (Palazzo and Lee 2015; Boivin et al. 2018). Despite high cellular demands, numerous individual tRNA genes have no direct evidence for expression (Kutter et al. 2011; Palazzo and Lee 2015; Hummel, Warren, and Drouard 2019; Gogakos et al. 2017). High duplication rates and consequent weakened purifying selection may lead to an abundance of pseudogenes. Additionally, many of these genes may be tRNA-derived short interspersed nuclear elements (SINEs), which often retain strong promoter elements. However, even after removal of apparent pseudogenes and SINEs, more than 60 human tRNA genes and over 100 mouse tRNA genes are in constitutively silenced regions of the genome for all tissues and cell lines, suggesting they are never or rarely transcribed (Roadmap Epigenomics Consortium et al. 2015; Bogu et al. 2015; A. Holmes 2018; Thornlow et al. 2018). Chromatin-immunoprecipitation sequencing (ChIP-seq) data supports this conclusion, as one multi-species study detected occupancy by RNA Polymerase III (Pol III) for only 224 of 417 high-confidence tRNA genes in human liver, with other mammals showing similar patterns (Kutter et al. 2011).

tRNA gene expression may co-evolve with phenotypic differences between species. Data from prior studies suggests that the rate of evolution of protein coding gene expression levels differs by clade (Necsulea and Kaessmann 2014; Brawand et al. 2011; W. H. Li et al. 1996). The rate of evolution of gene expression also varies among non-coding RNA gene families (Meunier et al. 2013; Necsulea and Kaessmann 2014; Necsulea et al. 2014). Due to difficulties in high-throughput, accurate quantification of tRNA abundance, the complexity of

tRNA gene expression across mammals is not well understood. The expanding functional repertoire of tRNA transcripts and tRNA-derived small RNAs (Kirchner and Ignatova 2015; Goodarzi et al. 2015; Mleczko, Celichowski, and Bąkowska-Żywicka 2014; Sun et al. 2018) indicates that changes in tRNA gene expression between species could have profound cellular effects.

Expression of tRNA genes has clear importance for organismal development and contribution to disease, but our understanding of its regulation and evolution is severely lacking for several reasons (Schaffer et al. 2014; Hanada et al. 2013; Yoo et al. 2016). Measuring expression of unique mRNA transcripts has become relatively straightforward. However, tRNA sequencing by the methods originally developed for unmodified small RNAs (e.g. microRNAs) is frequently impeded by numerous RNA modifications at the reverse transcription phase. Only very recently have specialized sequencing library preparation methods been developed to remove or overcome these modifications, enabling effective sequencing (Cozen et al. 2015; Zheng et al. 2015). Furthermore, because fully processed tRNA gene transcripts from different loci are often identical, simple tRNA-seq abundance measurements are often insufficient to determine the true transcriptional activity at each gene locus. Therefore, in order to determine which tRNA genes are potentially constitutively active, highly regulated, or silenced, other methods are needed.

Several genome-wide methods examine tRNA loci in their generally unique genomic contexts, bypassing the problem of identical mature tRNA transcripts. Such assays include chromatin immunoprecipitation (ChIP; (Bogu et al. 2015; Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018), RNA Polymerase III (Pol III) ChIP-seq (Kutter et al. 2011), and ATAC-seq (Foissac et al. 2018), among others. These high-throughput assays remain cost- and resource-intensive, so currently available data are often limited to few species and tissues. Nonetheless, these data show that identical tRNA genes do vary in expression profiles (Pan 2018; Kutter et al. 2011; Schmitt et al. 2014), supporting the need to incorporate extrinsic factors into the prediction of when or if tRNA genes are active. The study of the local

genomic context is therefore essential, and has not been tackled comprehensively by any tRNA gene prediction method.

In this chapter, I begin to resolve these concerns by developing a model to predict whether individual tRNA genes are actively transcribed in at least one tissue, or transcriptionally silent. In the prior chapter, I demonstrated that tRNA gene transcription may be inferred based on DNA variation driven by transcription-associated mutagenesis (Thornlow et al. 2018). I leverage this correlation, further enhanced by other genomic features, to infer expression of tRNA genes with high accuracy. This novel advance in tRNA research uses, but does not require, comparative genomic information, enabling its broad application. I demonstrate tRAP (tRNA Activity Predictor) using 29 placental mammalian genomes, most of which have no tRNA expression data. I also developed a robust mapping of syntenic tRNA genes across all 29 species. By combining tRAP with this comprehensive ortholog set, I analyze and compare expression classifications of over 10,000 tRNA genes, yielding a first look at the rate of tRNA gene regulation evolution in placental mammals, as well as bringing attention to the high frequency of silenced “high-scoring” canonical tRNA genes.

The text of this dissertation includes reprints of the following previously published material: Thornlow, B. P., Armstrong, J., Holmes, A. D., Howard, J. M., Corbett-Detig, R. B., & Lowe, T. M. (2020). Predicting transfer RNA gene activity from sequence and genome context. *Genome Research*, 30(1), 85-94. The co-authors listed in this publication directed and supervised the research which forms the basis for the dissertation.

2.2: Exploring correlates to tRNA activity and developing predictive model

Our goal was to develop a tRNA activity predictive model that could be applied to as many species as possible. To date, the most facile method for inferring tRNA gene function has been the use of tRNAscan-SE covariance model bit scores, which quantify similarity to primary sequence and secondary structure profiles derived from an alignment of reference

tRNAs (Lowe and Eddy 1997; Chan et al., n.d.). However, comparison to RNA Polymerase III ChIP-seq data from multiple mouse tissues (Kutter et al. 2011) suggests that high covariance model bit scores do not always correspond to occupancy by RNA Polymerase III (Pol III) (Supplemental Figure S1). More generally, this is therefore consistent with the idea that tRNAscan-SE bit scores alone are not strongly predictive of gene expression.

To improve prediction of tRNA functional roles and better understand the basis of tRNA gene regulation in mammals, we evaluated many additional sequence features easily obtained from a single reference genome (Figure 2.1). We explored genomic features correlated with activity based on comprehensive epigenomic data across 127 human tissues and cell lines (Roadmap Epigenomics Consortium et al. 2015), and then reduced this set to just those yielding the best predictions for our training data (Table 2.1).

To create our predictive model, we evaluated and incorporated two types of function-predictive statistics: intrinsic features related to tRNA gene sequence, and extrinsic features derived entirely from the genomic context. First, we reasoned that highly expressed tRNA genes should generally encode strong internal promoter sequences, and their transcripts must fold stably into the canonical tRNA structure. Both of these types of information are incorporated into tRNAscan-SE bit scores (Chan et al., n.d.). Furthermore, our previous study found that tRNA gene conservation is highest for actively transcribed tRNA genes, presumably due to stronger purifying selection on required sequence features (Thornlow et al. 2018). Thus, we included tRNA gene conservation in the form of the phyloP score, a nucleotide-level quantitative measure of conservation using multiple alignments (Pollard et al. 2010). We also assessed the correlation of gene activity with the length of each pre-tRNA's 3' tail, measured by the nucleotide distance from the end of the mature tRNA gene to the beginning of the poly(T) transcription termination sequence (Allison and Hall 1985; Koski et al. 1980). Multiple studies on tRNA transcription termination (R. J. Maraia, Kenan, and Keene 1994; Hamada et al. 2000; Orioli et al. 2011; Arimbasseri, Rijal, and Maraia 2013) observed that the RNase Z-trimmed 3' sequences vary in overall length,

composition, and terminator strength (poly(T) length), each potentially affecting tRNA maturation and processing.

We found that tRNAscan-SE bit scores and average phyloP scores across tRNA gene sequences are significantly correlated with tRNA gene activity based on epigenomic data. We also found that the total number of tRNA genes with identical anticodons and the distance to transcription termination sites are significantly anti-correlated with activity, as higher anticodon redundancy in tRNA genes and tRNA genes with more distal transcription termination sites are more frequently inactive (Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018); Spearman's rank correlation, $p < 1\times10^{-4}$ for all comparisons).

Second, because mRNA expression depends heavily on local chromatin context, we explored features of the genomic environment. Protein coding genes in regions rich in CG, or CpG, dinucleotides are known to be more frequently expressed (Gardiner-Garden and Frommer 1987; Krinner et al. 2014). Gardiner-Garden and Frommer define CpG islands scores as the observed frequency of CpG dinucleotides compared to their expected frequency given the G+C content of a region. We found that these scores, when calculated for the 350 bases upstream of each gene, are significantly correlated with active tRNA genes (Spearman's rank correlation, $p < 2.1\times10^{-24}$). Similarly, the frequency of CpG dinucleotides spanning from 350 bases upstream to 350 bases downstream of each tRNA gene is even more significantly correlated with expression (Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018); $p < 1.9\times10^{-27}$).

We also previously found that the putatively neutral regions flanking highly expressed tRNA genes are more divergent, consistent with transcription-associated mutagenesis (Thornlow et al. 2018). We observed that the average phyloP score of the 20-nucleotide 5' flanking regions of tRNAs is significantly anti-correlated with tRNA gene activity, as active tRNA genes more often have highly divergent flanking regions (Roadmap Epigenomics Consortium et al. 2015; Thornlow et al. 2018; $p < 8.9\times10^{-16}$). Finally, based on an expectation for increased chromatin accessibility for tRNA genes near other genes, we found that tRNA

genes are indeed more likely to be in an active chromatin state if near protein-coding genes ($p < 8.9\times10^{-5}$) or other tRNA genes ($p < 9.7\times10^{-7}$).

We hypothesized that some combination of both intrinsic and extrinsic features could enable robust computational inference of potential for tRNA gene activity (Figure 2.1). To develop an integrated model, we tested several common frameworks, including random forest (RF), logistic regression, and support vector machines. The RF classifier was most effective, achieving the greatest area under the Receiver Operating Characteristic curve (AUC, Figure 2.2A-B) based on ten-fold cross-validation of human tRNA gene data and subsequent application, without retraining, to mouse tRNA gene data (Methods).

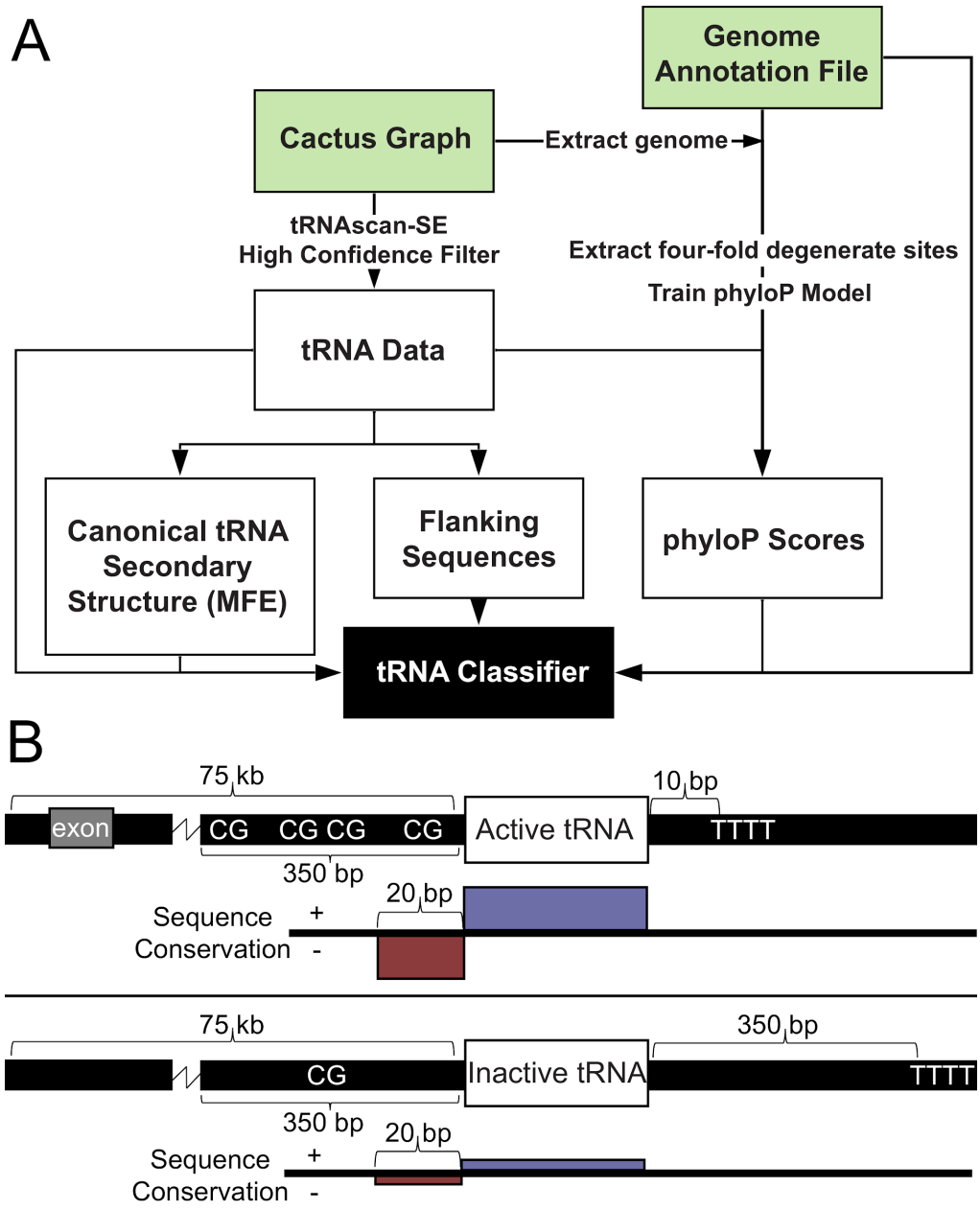


Figure 2.1: Schematic of tRNA activity classifier and key features used in prediction. *(A) Flowchart of analysis pipeline, which extracts tRNA information solely from genomic data and classifies tRNA genes as active or inactive. Green blocks indicate files not created by the pipeline. By default, the method uses a Cactus graph (Armstrong et al. 2019), which is a reference-free whole genome alignment, and a genome annotation file as input. (B) Active tRNA genes generally have more CpG dinucleotides in their 350 base pair upstream flanking regions, more proximal transcription termination sequences ("TTTT"), are within 75 kb of more exons, have more highly conserved gene sequences and more evolutionarily divergent 20-nucleotide 5' flanking regions.*

2.3: Features derived from CpG islands are most informative.

To better understand and improve our classifier, we determined the relative importance of each feature in our random forest model (Table 2.1; (Pedregosa et al. 2011). All features contribute to model accuracy and are significantly correlated with the activity labels (Spearman's rank correlation, $p < 1\times10^{-4}$ for all features). Among high-confidence tRNA genes, as determined by tRNAscan-SE (Chan et al., n.d.), the most informative features predictive of activity (Feature Importance, Table 2.1); (Pedregosa et al. 2011), are derived from CpG content at each tRNA locus. Upon comparing our CpG data to epigenomic data for each mouse tRNA gene (Bogu et al. 2015), we find that both CpG density and CpG islands scores are exceptionally highly correlated with breadth of activity (Spearman's rank, $p < 2.4\times10^{-78}$ for CpG Density, $p < 5.0\times10^{-61}$ for CpG Islands Score). This supports the idea that CpG-derived genomic data are particularly highly informative of tRNA gene activity.

One might expect that the tRNAscan-SE general bit score should be the most informative single feature. However, we start with relatively high quality tRNAs with likely pseudogenes already removed using the tRNAscan-SE bit score-based high-confidence filter (Chan et al., n.d.). Thus, for a starting set of tRNAs already vetted for reasonably strong features, the contribution of the tRNAscan-SE bit score to the model is marginally smaller than other features not previously used to estimate gene function. By incorporating both tRNA gene sequence and genome context, our classifier represents a substantial improvement over using tRNAscan-SE covariance bit scores alone.

		Feature Importance	Active Mean (95% CI)	Inactive Mean (95% CI)	Glu-CTC-1-1	Lys-CTT-11-1	Gln-TTG-3-1	Gln-TTG-4-1
	Activity	**-**	**-**	**-**	**active**	**inactive**	**active**	**inactive**
	Probability	**-**	**-**	**-**	**+0.994**	**-0.975**	**+0.984**	**-0.895**
Intrinsic	Total Number of tRNA Genes with Identical Anticodon	0.089	11.1 (10.4, 11.8)	17.9 (15.3, 20.2)	14	15	6	6
	Minimum Free Energy of Canonical tRNA Secondary Structure	0.074	-27.4 (-27.8, -27.0)	-23.5 (-24.6, -22.5)	-26.4	-16.0	-22.0	-24.1
	tRNAscan-SE General Bit Score	0.070	76.2 (75.3, 77.3)	69.9 (67.3, 72.6)	73.2	56.6	66.9	58.1
	Average PhyloP Score in tRNA Sequence	0.063	0.86 (0.79, 0.92)	0.35 (0.25, 0.46)	1.37	-0.081	0.51	-0.047
	Distance to Nearest TTTT Transcription Termination Sequence	0.040	15.9 (13.1, 19.2)	66.9 (43.5, 94.7)	8	351	7	14
Extrinsic	CpG Density Across tRNA Locus	0.303	0.043 (0.041, 0.045)	0.018 (0.014, 0.023)	0.045	0.014	0.024	0.014
	Observed/Expected CpG Islands Score Upstream of tRNA Gene	0.225	0.67 (0.64, 0.69)	0.27 (0.21, 0.33)	0.68	0.092	0.62	0.11
	Average PhyloP Score in 5' Flanking Region	0.092	-2.61 (-2.74, -2.45)	-1.21 (-1.44, -0.99)	-3.88	0.17	-2.50	0.009
	tRNA Genes Within 10kb	0.023	1.97 (1.75, 2.18)	0.84 (0.54, 1.19)	3	0	5	0
	Exons within 75kb	0.019	35.3 (32.2, 38.7)	21.6 (17.2, 26.7)	88	43	37	0

Table 2.1: Both intrinsic (tRNA-specific) and extrinsic (genome context) features are integral to the model. *All features included in the model with their relative importance values as measured by decrease in node impurity by scikit-learn ((Pedregosa et al. 2011); see Methods). Greater feature importance scores indicate greater contribution to discrimination between active and inactive tRNA genes by the model. The Active Mean and Inactive Mean columns refer to the mean value across all human tRNA genes in our training set that are known to be active and inactive, respectively, with 95% confidence intervals (CI) in parentheses, calculated for each mean using bootstrapping. Minimum free energy of canonical tRNA structure refers to the minimum free energy when constrained to folding into the canonical cloverleaf secondary structure (Lorenz et al. 2011). For calculating CpG-related statistics, we consider the tRNA locus to begin 350 bp upstream and end 350 bp downstream of the gene. To calculate the phyloP score in the 5' flanking region, we considered only the 20 bp immediately upstream of each tRNA gene. As examples, the human tRNA genes predicted most likely active (Glu-CTC-1-1) and most likely inactive (Lys-CTT-11-1), across all human tRNAs, are shown, as well as two examples from the same anticodon family (tRNA-Gln-TTG), one active and one inactive. For the distance to the nearest transcription*

termination sequence, if the motif "TTTT" was not found within 350 nucleotides of a tRNA gene, 351 was used as its value, as is the case for Lys-CTT-11-1.

2.4: tRAP is 94% accurate in classifying mouse tRNA genes based on epigenomic data.

Because our classifier was trained using comprehensive epigenomic data mined from human tRNA gene loci (Roadmap Epigenomics Consortium et al. 2015); Supplemental Table S3), we required an independent data set to validate our predictions. Therefore, we tested the accuracy of our classifier using epigenomic data evaluating histone marks at mouse tRNA genes across 9 tissues from (Bogu et al. 2015). Our mouse tRNA gene set contains 376 genes, with 259 observed as active and 117 believed silent based on epigenomic data. Our classifier predicted that 264 of these genes are active and that 112 are inactive, correctly categorizing 353 tRNA genes and achieving 93.9% accuracy (Figure 2.2B-D). Of the 23 misclassified mouse tRNA genes, 14 are misclassified as active and 9 are misclassified as inactive. We note that these genes are not biased by isotype, nor by genomic location, and are therefore most likely misclassified for a variety of reasons (Supplemental Table S5). To ascertain the performance of our classifier on non-conserved tRNA loci between human and mouse, we also tested the classifier on only the 184 mouse tRNA genes in our test set without syntenic human orthologs. We correctly classify 167 such genes, achieving 90.8% accuracy in this highly biased subset of tRNA genes.

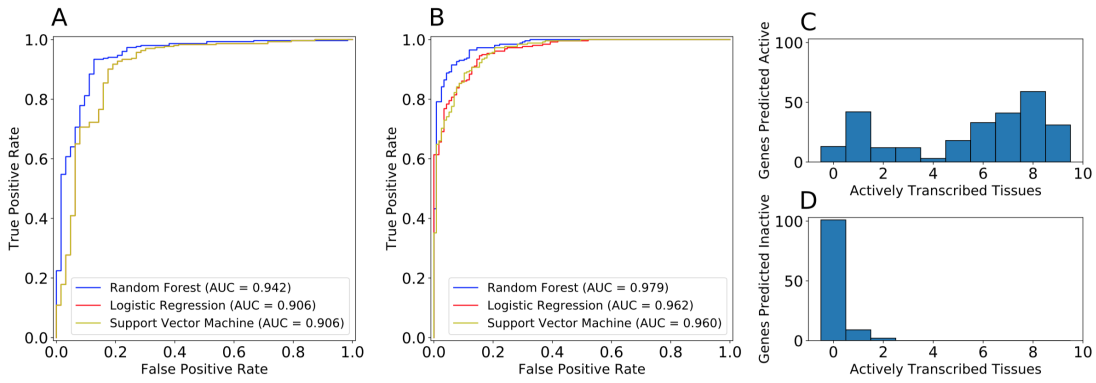


Figure 2.2: Random forest classifier achieves 94% accuracy on mouse tRNA genes. *Receiver operating characteristic curves for random forest (blue), logistic regression (red) and*

support vector machine (yellow) upon application to (A) human training data with ten-fold cross-validation and (B) mouse test data are shown. The number of mouse tRNA genes predicted as (C) active and (D) inactive are compared to the number of tissues in which they are actively transcribed according to (Bogu et al. 2015). We considered a mouse tRNA gene active if it is actively transcribed in at least one tissue.

2.5: Classification without alignment or annotation is similarly accurate.

We developed our method such that it could potentially be applied to any species with a sequenced genome. For best performance, we used a Cactus graph (Armstrong et al. 2019; Paten, Diekhans, et al. 2011; Paten, Earl, et al. 2011; N. Nguyen et al. 2015), which is a reference-free whole genome alignment. Usage of a Cactus graph enhances detection of synteny and facilitates extraction of alignments for specific regions in multiple genomes. The Cactus graph used in this study includes 29 mammalian genomes (Table 2.2).

Nonetheless, we recognize that Cactus graphs are not yet available for all species. To accommodate species for which no alignments or protein-coding gene annotations have been developed, we included an option to omit the requirement of this feature information. Use of this simplified classifier led to decreases in accuracy in both human (AUC = 0.927, 91.8% accuracy compared to AUC = 0.942 and 93.2% accuracy in the full model) and mouse (AUC = 0.974, 92.6% accuracy compared to AUC = 0.979 and 93.9% accuracy in the full model), which may be exacerbated upon application to more phylogenetically distant species.

Common Name	Species Name	Assembly Name	Accession	Scaffold N50	Contig N50	Family	Order
dog	*Canis lupus familiaris*	CanFam3.1	GCF_00000228	45876610	267478	Canidae	Carnivora
killer whale	*Orcinus orca*	Oorc_1.1	GCF_00033195	12735091	70300	Delphinidae	Cetacea
black flying fox	*Pteropus alecto*	ASM32557v1	GCF_00032557	15954802	31841	Pteropodidae	Chiroptera
big brown bat	*Eptesicus fuscus*	EptFus1.0	GCF_00030815	13454942	21392	Vespertilionidae	Chiroptera
nine-banded armadillo	*Dasypus novemcinctus*	Dasnov3.0	GCF_00020865	1687935	26277	Dasypodidae	Cingulata
western European hedgehog	*Erinaceus europaeus*	EriEur2.0	GCF_00029675	3264618	21359	Erinaceidae	Insectivora
rabbit	*Oryctolagus cuniculus*	OryCun2.0	GCF_00000362	35972871	64648	Leporidae	Lagomorpha
goat	*Capra hircus*	ASM170441v1	GCF_00170441	87277232	26244591	Bovidae	None
pig	*Sus scrofa*	Sscrofa11.1	GCA_00000302	88231837	48231277	Suidae	None
horse	*Equus caballus*	EquCab2.0	GCF_00000230	46749900	112381	Equidae	Perissodactyla
western gorilla	*Gorilla gorilla*	GSMRT3	GCA_90000665	10016017	10016017	Hominidae	Primates
human	*Homo sapiens*	GRCh38.p11	GCA_00000140	59364414	56413054	Hominidae	Primates
Rhesus monkey	*Macaca mulatta*	Mmul_8.0.1	GCF_00077287	4193270	107156	Cercopithecidae	Primates
gray mouse lemur	*Microcebus murinus*	Mmur_3.0	GCA_00016544	108171978	210702	Cheirogaleidae	Primates
Ma's night monkey	*Aotus nancymaae*	Anan_2.0	GCA_00095205	8268663	126456	Aotidae	Primates
house mouse	*Mus musculus*	GRCm38.p5	GCF_00000163	52589046	32273079	Muridae	Rodentia
domestic guinea pig	*Cavia porcellus*	Cavpor3.0	GCF_00015173	27942054	80583	Caviidae	Rodentia
European marmot	*Marmota marmota*	marMar2.1	GCF_00145813	31340621	66492	Sciuridae	Rodentia
lesser Egyptian jerboa	*Jaculus jaculus*	JacJac1.0	GCF_00028070	22080993	15675	Dipodidae	Rodentia
degu	*Octodon degus*	OctDeg1.0	GCF_00026025	12091372	19847	Octodontidae	Rodentia
Ferret	*Mustela putorius furo*	MusPutFur1.0	GCA_00021562	9335154	44823	Mustelidae	Carnivora
Praire vole	*Microtus ochrogaster*	MicOch1.0	GCA_00031737	17270019	21250	Cricetidae	Rodentia
Cape golden mole	*Chrysochloris asiatica*	ChrAsi1.0	GCA_00029673	13470186	19631	Chrysochloridae	Afrosoricida
Naked mole rat	*Heterocephalus glaber*	HetGla_female_1.0	GCA_00024769	20532749	47778	Heterocephalidae	Rodentia
Chinchilla	*Chinchilla lanigera*	ChiLan1.0	GCA_00027666	21893125	61105	Chinchillidae	Rodentia
Norway rat	*Rattus norvegicus*	Rnor_6.0	GCF_00000189	14986627	100461	Muridae	Rodentia
cattle	*Bos taurus*	Bos_taurus_UMD_3.1.1	GCF_00000305	6380747	96955	Bovidae	None
chimpanzee	*Pan troglodytes*	Pan_tro 3.0	GCF_00000151	26972556	384816	Hominidae	Primates
orangutan	*Pongo abelii*	Susie_PABv2*	GCA_00288077	98475126	11074009	Hominidae	Primates

Table 2.2: Genome, species and assembly information for all species found in the Cactus graph used in our study. *For the orangutan genome, an assembly prior to submission to GenBank was used. Therefore, our coordinates and annotation do not precisely match up to those found in Susie_PABv2.*

2.6: ChIP-seq, DM-tRNA-seq and ATAC-seq data independently validate our classifications in additional species.

To further validate our model, which was trained on human chromatin data, we compared our predictions to RNA Polymerase III (Pol III) ChIP-seq data previously collected from the livers of four species (*Mus musculus*, *Macaca mulatta*, *Rattus norvegicus* and *Canis lupus familiaris*; (Kutter et al. 2011). While Pol III ChIP-seq measures Pol III occupancy rather than transcription, it is a requirement for transcription, and our usage of ChIP-seq data instead of transcription data ameliorates the common problem of mature tRNA transcripts mapping ambiguously to multiple tRNA loci. We found roughly expected agreement between our classifications and the Pol III ChIP-seq read counts from a single tissue (Figure 2.3A-D). Our predictions are similarly accurate when compared to mouse muscle and testes ChIP-seq data.

We predicted many tRNA genes as active despite a lack of Pol III-binding at these loci in liver, muscle, and testes. This is a consequence of our methodology, as our model does not predict activity in specific tissues, but is instead trained to predict tRNA genes as active if epigenomic data indicates active transcription in at least one of many tissues (Roadmap Epigenomics Consortium et al. 2015). For example, in mouse, 259 total tRNA genes are active in at least one tissue based on the epigenomic data, but 90 of these (35%) are not expected to be active in the liver based on the same data. Based on human and mouse epigenomic data, a large proportion of tRNA genes are expressed exclusively in stem cells and cell lines (A. Holmes 2018). This may explain many of the discrepancies we observe in predicting tRNA genes as active that do not have any evidence for Pol III occupancy in one or a small number of differentiated tissues. We predict the brain-specific mouse tRNA gene, tRNA-Arg-TCT-4-1 (Ishimura et al. 2014), which has no ChIP-seq reads in mouse liver, muscle or testes (Kutter et al. 2011), as active with 0.664 probability. This is consistent with our goal to predict any tRNA gene with known activity in any tissue as active.

Our model predicted 120 (macaque), 67 (rat) and 142 (dog) tRNA genes as active despite Pol III ChIP-seq read counts of zero in the liver (Kutter et al. 2011). Although ChIP-seq has not been performed on macaque, rat and dog tRNA loci for any other tissues, we find that virtually all tRNA genes with measured Pol III binding are predicted to be active by our classifier. Among tRNA genes with Pol III ChIP-seq read-counts greater than zero, we predicted that 95.3% are active in mouse, 97.3% in macaque, 98.3% in rat and 98.5% in dog. This consistency in tRNA distributions and classifier behavior across species suggests that the classifier is similarly accurate in mouse, macaque, rat and dog (Figure 2.3).

As additional validation, we have compared our predictions to new tRNA transcript abundance data for mouse brain and liver, collected by our lab using DM-tRNA-seq (Zheng et al. 2015). Compared to other assays, DM-tRNA-seq is a more direct measure of transcript abundance. However, because this sequencing method captures mostly mature tRNA transcripts, we were limited to only the 153 single-copy mouse tRNA loci in our data set, as it

is impossible to determine the source loci for transcripts produced by multi-copy tRNA genes. We conducted DM-tRNA-seq in mouse liver and brain in three replicates each, and compared the average normalized read counts across each tissue to our activity predictions for each single-copy tRNA gene (Figure 2.3E-F, see Methods). Our DM-tRNA-seq data supports the tissue specificity of tRNA-Arg-TCT-4-1, as we see moderate expression across all of our brain replicates, and detect no expression in any of our liver replicates. We also found that our DM-tRNA-seq data from mouse liver is significantly correlated with the Pol III ChIP-seq data from mouse liver (Spearman's rank, $p < 3.6\times10^{-29}$). When we consider tRNA genes with an average of at least 20 reads in either tissue to be active and all others to be inactive, we achieve 84% accuracy, which may be a low estimate based on the small number of tissues tested.

To validate our predictions in more species, we used ATAC-seq data captured in liver, CD4 and CD8 cells for the cow, pig and goat genomes (Foissac et al. 2018). We compared our predictions to the ATAC-seq peaks across these tissues for the regions spanning 250 base pairs upstream and downstream of each tRNA gene (Supplemental Figure S6-S7). Again, due to the inclusion of only a small subset of tissues in this data, many tRNA loci that do not show activity in these tissues but were predicted as active by our model may be active in other tissues. Among tRNA genes with ATAC-seq peaks, we predicted 90.4%, 95.1%, and 90.8% as active in cow, goat and pig, respectively. These results are comparable to measurements obtained from ChIP-seq data in mouse, macaque, rat and dog.

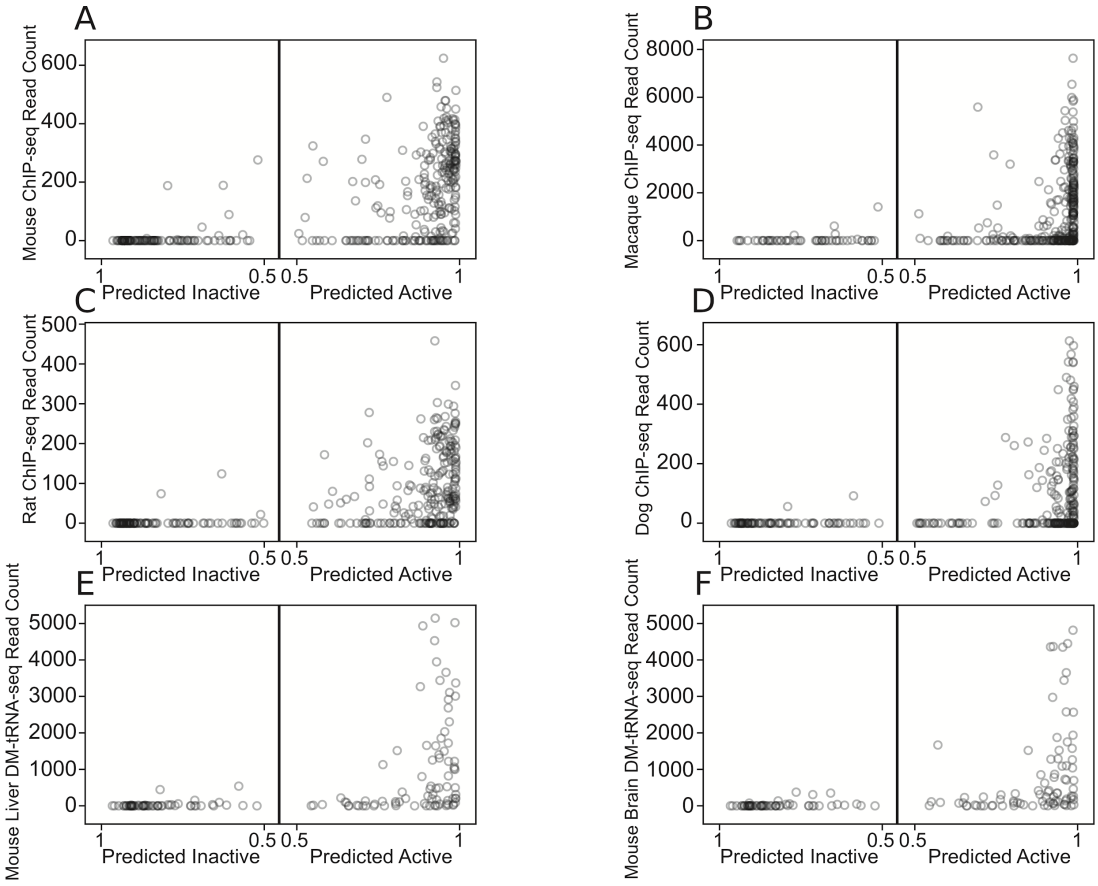


Figure 2.3: Classification of gene activity based on genomic data achieves similar results to Pol III ChIP-seq analysis in four species and DM-tRNA-seq in two tissues. *Probability scores output by the classifier for (A) mouse, (B) macaque, (C) rat and (D) dog are shown on the x-axis where tRNA genes further left are predicted inactive with greater probability, and tRNA genes further right are predicted active with greater probability. The y-axis shows Pol III ChIP-seq read counts from the liver of each species for each tRNA gene, from (Kutter et al. 2011). Similar patterns are observed for predicted active versus inactive mouse tRNA genes with uniquely mapping DM-tRNA-seq data, comparing to the average normalized read count across three replicates in (E) mouse liver and (F) mouse brain.*

2.7: tRNA gene classifications follow similar distributions across the eutherian phylogeny.

We applied our model to 29 mammalian species (Figure 2.4, Table 2.2) to glean new insights into the evolution of tRNA complements. We determined the distributions of active and inactive tRNA genes by anticodon across these species, finding that most species have approximately 250-350 predicted active genes, comprising roughly 75% of their tRNA gene sets. We observe similar distributions by clade, with a few exceptions. *Bos taurus, Capra*

hircus and *Orcinus orca* (cow, goat and orca, respectively) have more than 300 tRNA genes predicted inactive while no other species has more than 154. This most likely reflects decreased ability of tRNAscan-SE to discriminate tRNA-derived SINEs (short interspersed nuclear elements) from tRNA genes in these species (Chan et al. 2019). Furthermore, we verified that all species had at least one tRNA gene predicted as active for each expected anticodon (Grosjean et al. 2010), with only three exceptions that likely represent genome assembly errors.

2.8: Establishing mammalian ortholog sets enables further evolutionary analysis of tRNA gene regulation.

In order to investigate the relationship between evolutionary conservation and transcriptional activity, we developed a complete set of placental mammal tRNA gene orthologs using a Cactus graph (Armstrong et al. 2019). Cactus graphs are state-of-the-art alignments that allow greater detection of synteny across many species. Of the 11,724 tRNA genes in our 29-species alignment, 3,554 genes in total, or about 123 per species on average, appear to be species-specific, although this may be an overestimate due to our limited ability to definitively call orthologs. The rest were grouped into 1,097 ortholog sets. Out of these, 750 ortholog groups contain only tRNA genes predicted to be active, approximately mirroring the distribution of active to inactive tRNA genes predicted at the species level (Figure 2.4B). On average, each of our 1,097 ortholog sets spans 7.4 species, indicating that tRNA genes are generally either fairly deeply conserved or recently evolved. In aggregate, this is consistent with prior studies in *Drosophila* showing that tRNA genes can be “core” or “peripheral” (Rogers, Bergman, and Griffiths-Jones 2010).

We identified a “core” set of 97 primate tRNA genes for which all seven primate species (human, chimpanzee, gorilla, orangutan, macaque, *Microcebus murinus* (gray mouse lemur) and *Aotus nancymaae* (Nancy Ma’s night monkey)) have a syntenic ortholog, which are of interest for future experimentation (Supplemental Figure S10). These represent tRNA

genes likely present in the primate common ancestor that have not been lost in any lineage leading to the sampled genomes. These genes encode 19 amino acids. A single standard amino acid isotype is not represented: cysteine. tRNA-Cys genes are often present in high numbers, and every species in the primate phylogeny has at least 19 of these genes. However, these genes are prone to accumulating nucleotide substitutions, as the human genome contains 23 unique high-confidence tRNA-Cys-GCA gene sequences, the most of any isotype. Therefore, the lack of a “core” eutherian tRNA-Cys gene may be due to relatively rapid evolution of this gene family, or perhaps difficulty in alignment due to their high variation in sequence.

In 15 of these 97 “core” ortholog sets, we predicted at least one member of the ortholog group to be active and at least one inactive among the different primate species. Across all 97 “core” ortholog sets, we predict 98% of all member tRNA genes as active, consistent with the previously mentioned correlation between tRNA gene conservation and transcriptional activity. Additionally, upon comparison to measurements of Pol III-specific transcription factors (Canella et al. 2010), we find that all 97 “core” human tRNA genes have peaks greater than zero.

A

Initial \ Final	Absent	Inactive	Active
Absent	0.994	0.004	0.002
Inactive	0.005	0.994	**0.001**
Active	0.015	**0.002**	0.983

B

Species	Active	Inactive
Homo sapiens	314	52
Pan troglodytes	336	53
Gorilla gorilla gorilla	315	48
Pongo abelii	344	51
Macaca mulatta	334	57
Aotus nancymaae	285	43
Microcebus murinus	282	40
Octodon degus	241	57
Chinchilla lanigera	250	51
Cavia porcellus	266	56
Heterocephalus glaber	238	56
Mus musculus	264	112
Rattus norvegicus	243	102
Microtus ochrogaster	340	128
Jaculus jaculus	244	74
Marmota marmota	249	154
Oryctolagus cuniculus	371	73
Bos taurus	391	373
Capra hircus	207	317
Orcinus orca	321	334
Sus scrofa	287	105
Canis lupus familiaris	271	89
Mustela putorius furo	290	86
Equus caballus	335	63
Pteropus alecto	222	123
Eptesicus fuscus	199	15
Erinaceus europaeus	341	104
Dasypus novemcinctus	371	150
Chrysochloris asiatica	275	97

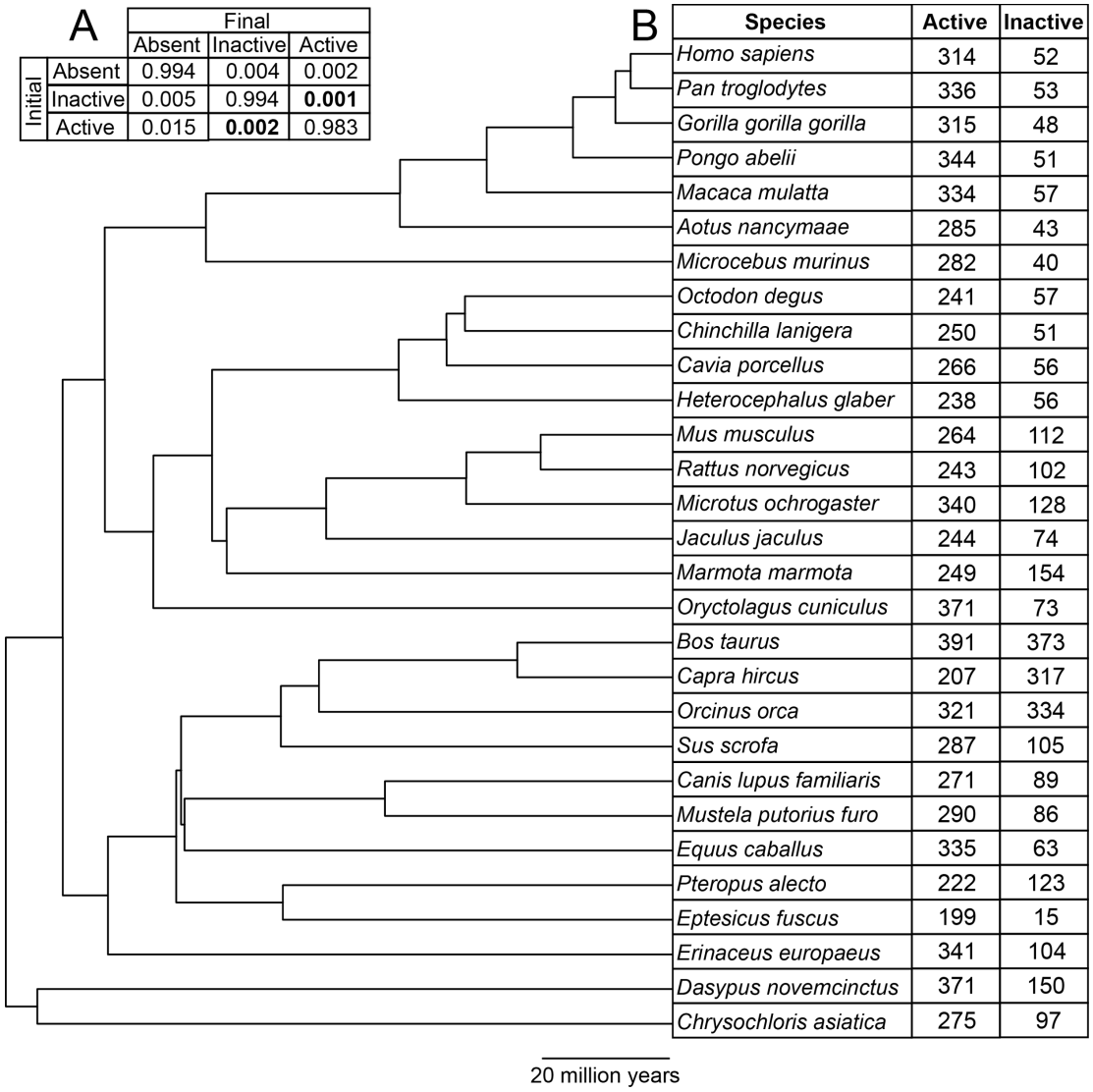


Figure 2.4: Placental mammals demonstrate consistent distributions of predicted active and inactive tRNA genes. *(A) Estimated transition probabilities between each predicted activity state over a branch length of 1 million years using RevBayes. The probabilities of transition from inactive to active (0.001) and from active to inactive (0.002) are in bold. (B) The number of tRNA genes in each predicted activity class are shown for each species in our phylogeny (Hedges, Dudley, and Kumar 2006), after removal of tRNA genes in segmental duplications. For human and mouse, tRNA genes with no epigenomic data are excluded from this table as well (see Methods).*

2.9: Transitions between active and inactive are rare.

We fit all of our ortholog sets and the predicted activity states of their constituent genes to a Markov model of evolution of discrete characters using RevBayes (Höhna et al.

2016) (Figure 2.4A, see Methods). By fitting our data to the model, we estimated transition probabilities to and from three states: active, inactive, and absent (no detected ortholog). We held the phylogeny constant and solved only for the transition rate parameters. Our model finds that the probability of observing a tRNA gene transition from active to inactive for a given tRNA gene over 1 million years is only 0.002 (Figure 2.4A), suggesting that activity state transitions are rare.

Our classifier does well to detect these rare transition events. There are 183 human/mouse ortholog pairs spanning our training and test data sets, and in 171 (93%) of them, human and mouse have the same activity state based on epigenomic data. However, we correctly classified 180 human (98%) and 177 mouse (97%) tRNA genes within this set, indicating that we detected activity state changes between these species, including 11 human tRNA genes and 8 mouse tRNA genes whose activity states differ from their orthologous counterparts. Assuming that the activity state of orthologous tRNA genes remains constant across closely related species would yield largely accurate activity state predictions for annotating tRNAs in additional new species. However, our classifier represents an improvement over this assumption, and is particularly applicable to species-specific tRNA genes, which are especially common and have no ortholog data.

Inactive tRNA genes that are conserved most often remain inactive (Figure 2.4A), hinting at undiscovered biological roles for conserved, apparently silent tRNA genes. We also observed some variation in the relative transition probabilities within clades. Primate tRNA genes are less likely to remain in their initial predicted activity state than rodent tRNA genes. This is consistent with prior studies on the rate of evolutionary change of protein coding gene expression between clades (Brawand et al. 2011; Necsulea and Kaessmann 2014) but likely also reflect differences in sample size between clades. Based on our results, turnover in tRNA gene expression class generally appears to be slow, similar to the expression of protein coding genes (Brawand et al. 2011).

2.10: Discussion

Greater understanding of tRNA regulation is a difficult and unmet challenge. There are many obstacles preventing direct measurement of expression at the gene level, including extensive post-transcriptional modifications impeding sequencing, and multiple genomic loci encoding identical transcripts. Nonetheless, we show that accounting for the genomic context allows for improved tRNA gene annotation, and that in order to determine the transcriptional potential of tRNA genes, direct measurement across many tissues is not necessarily required if the gene sequence and genomic context is known. We leverage features intrinsic to tRNA genes, which relate directly to tRNA function and processing, as well as those extrinsic, which relate to regulation of the chromosomal region.

There are numerous challenges to validating any method for predicting tRNA transcriptional potential. Comprehensive epigenomic data is available for only a few species. Similarly, ChIP- and ATAC-seq data are generally conducted only on a few tissues for a few species of interest. The prevalence of identical tRNA genes in most placental mammal genomes also prevents the identification of source loci for tRNA transcript sequencing data and further limits our ability to support our predictions. However, this relative scarcity of available data motivates the creation of our classifier. Our estimates are comparable to experimental results, but with much greater ease of use and cost-effectiveness. Epigenomic data (Bogu et al. 2015; A. Holmes 2018) indicate that only 8% of mouse tRNA genes in our test set are active in all nine tissues, and 13% are expressed in only one tissue. This suggests that tissue-specificity of tRNA expression is common. This is an area of great interest (Ishimura et al. 2014), but few examples have been characterized. Because our classifier infers expression in at least one tissue, our methods will be useful in guiding experiments to find more examples of tightly regulated tRNA genes.

The genome-based nature of this method allows for expansion to incorporate much more data in the future. For example, variation within populations may be useful for predicting relative transcript expression within gene families. We previously determined that actively

transcribed tRNA genes accumulate more rare single nucleotide polymorphisms (SNPs) in both their flanking regions and gene sequences (Thornlow et al. 2018). Therefore, we expect that when population variation data is available for more species, we may infer expression differences at narrower timescales. The model may also be expanded to accommodate non-binary classification of expression levels in different tissue types, and capture the nuance of tRNA gene expression regulation. This approach might also be adapted for the study of other large gene families, as we have previously shown that histone protein coding genes exhibit similar genetic variation to tRNA genes (Thornlow et al. 2018).

In conclusion, we demonstrate reliable classification of tRNA genes using an algorithm that requires little input data and can easily be expanded in the future. Annotations created by our method will be useful in prioritizing tRNA characterization experiments, as well as interpreting the biological effects of mutations in and surrounding tRNA genes. This work informs the broader question of tRNA gene function evolution, illustrating that tRNA gene expression regulation is dependent on the tRNA gene sequence as well as the varied genomic environment.

2.11: Methods

For the training data, we used coordinates from human genome assembly GRCh38 for tRNA genes not removed by the tRNAscan-SE high confidence filter (Chan and Lowe 2016; Chan et al., n.d.). For all species, including human and mouse, we extracted the genomes from our Cactus alignment (Armstrong et al. 2019); Supplemental Table S6), ran tRNAscan-SE 2.0 and applied the EukHighConfidenceFilter to exclude tRNA pseudogenes and tRNA-derived SINEs (Chan et al., n.d.). We used custom Python scripts to find tRNA loci that were identical from 80 nucleotides upstream to 40 nucleotides downstream of the gene start and end. We considered these segmental duplications and excluded them from classification. If any of these loci also did not align to any tRNA loci in any other species, they were also removed from our ortholog calls, as they most likely represent assembly errors. For

genome assemblies in which at least 85% of nucleotides were found on chromosomes, we excluded all tRNA genes not found on chromosomes. For the human tRNA gene set, because our epigenomic data is based on GRCh37 assembly gene annotations, we removed any tRNA genes that were not included in the older assembly (determined by performing liftOver (Casper et al. 2018) conversion from GRCh38 to GRCh37), as well as genes in segmental duplications in either assembly.

We used the PHAST (Hubisz, Pollard, and Siepel 2011) and HAL (Hickey et al. 2013) toolkits to generate phyloP data, and RNAfold (Lorenz et al. 2011) to estimate minimum free energy, using the constraints on secondary structure output by tRNAscan-SE 2.0. We used custom Python scripts in conjunction with tRNAscan-SE 2.0 output and genome annotation files (accession numbers listed in Table 2.2) to obtain data for all other features. When phyloP data were unobtainable due to lack of alignment, we replaced feature values for each tRNA gene with the mean value for that feature across all tRNA genes in that species, using the SimpleImputer() module in scikit-learn (Pedregosa et al. 2011). We used scikit-learn to train the model and classify each gene (Pedregosa et al. 2011). Our pipeline and corresponding data are available in the Supplemental Material, as well as at https://github.com/bpt26/tRAP/. We used Spearman's rank correlation test to ensure that no features were perfectly correlated (Guyon and Elisseeff 2003). We used CfsSubsetEval (Hall et al. 2009) to remove uninformative features and scikit-learn to determine feature importance (Pedregosa et al. 2011). To determine the threshold distances for the "Exons Within 75 Kilobases" and "tRNA Genes Within 10 Kilobases" features, we conducted the Mann-Whitney *U* test for several threshold distances and selected the distance that yielded the smallest p-value.

To train and test our model, we used epigenomic data from the NIH Roadmap Epigenomics Program (Roadmap Epigenomics Consortium et al. 2015) and the chromatin state-associated gene study in mice (Bogu et al. 2015) for human and mouse tRNA gene activity states, respectively. These studies used histone marks to identify regions of active transcription across 127 human tissues and 9 mouse tissues, respectively. In both species,

we excluded tRNA genes for which epigenomic data was not available, and tRNA genes contained within large segmental duplications. Our training set includes 366 human tRNA genes, 303 active and 63 inactive. For both species, we considered tRNA loci as active if they had an open chromatin state in at least one tissue. We considered all others to be inactive. To determine performance on the human data, we used ten-fold cross-validation, as is commonly used for gene classification studies (McLachlan, Do, and Ambroise 2005; Chen et al. 2018; Sethi et al. 2018). We also tested three-fold cross-validation, but observed very little difference in the model. To validate our model, we compared our classifications to ChIP-seq read counts taken directly from Kutter et al 2011 and ATAC-seq peaks taken directly from Foissac et al 2018, using liftOver (Casper et al. 2018) conversion to accommodate differences in genome assembly.

For information on library preparation methods for DM-tRNA-seq assays, see Supplemental Methods. Following sequencing, we used a specialized tRNA sequencing data analysis pipeline available at https://github.com/UCSC-LoweLab/tRAX, which aligns reads to the tRNA transcripts and reference genome, and compute normalized read counts for each transcript. Because some tRNAs have multiple identical copies in the genome, those sequencing reads were aligned to all corresponding gene loci. To avoid ambiguities, we analyzed only the single-copy tRNA gene loci in this study, as well as only the reads corresponding to whole tRNA molecules.

We used hal2maf to create 29-way alignments for all tRNA loci of interest for the species in our phylogeny (Hickey et al. 2013). For each tRNA locus, we considered the best aligning tRNA locus from all other species as orthologous, allowing only one ortholog per species per locus. We allowed tRNA genes in segmental duplications to be included, but only if they had an ortholog in at least one other species, as species-specific segmental duplications may be the result of assembly errors. We augmented our ortholog sets with syntenic human/mouse, human/dog and human/macaque tRNA gene ortholog pairs from (A. Holmes 2018). For all instances in which each tRNA gene in a Holmes 2018 ortholog pair

aligned to mutually exclusive sets of species in our Cactus graph, we combined them into one ortholog set. We found that 29 Holmes 2018 human-mouse ortholog pairs align to each other in the Cactus graph, 152 align to mutually exclusive sets of species in our Cactus graph, and 17 align to overlapping sets of species in our Cactus graph. Therefore, we combined the 152 human-mouse tRNA gene pairs with the corresponding ortholog sets defined by our Cactus graph into larger ortholog sets.

We used a phylogeny from TimeTree (Kumar et al. 2017) and fit our data to a Markov model using RevBayes (Höhna et al. 2016). We held the phylogeny constant and allowed RevBayes to optimize only the Q matrix using our tRNA data. We then determined transition probabilities over 10 million years using the RevBayes function getTransitionProbabilities() across all species (Figure 2.4A) and by clade.

Chapter 3: Stability of SARS-CoV-2 Phylogenies

3.1: Background

The SARS-CoV-2 pandemic has led to unprecedented, nearly real-time genetic tracing due to the rapid community sequencing response. This global sequencing effort is inherently decentralized and must rely on data collected by many labs using a wide variety of molecular and bioinformatic techniques. Therefore, a strong possibility exists that systematic errors associated with lab-specific practices affect a non-negligible number of analyzed sequences. These errors may be created or accentuated when samples that contain unidentified sequencing errors are incorporated into the phylogenetic tree. Defining stable and easily referenced major clades of the virus is essential for epidemiological studies of viral population dynamics (Rambaut et al., n.d.; Mavian et al., n.d.). An understanding of how errors might be affecting the trees that are being published is essential to achieving that goal.

It can be difficult to distinguish sequencing errors of different types from genuine transmitted and non-transmitted mutations in genome sequences. Taking a conservative approach, many researchers remove mutations that are observed only once during the evolution of the virus when constructing a phylogenetic tree, as these may be more likely to be errors (Rayko and Komissarov, n.d.; Akther et al. 2020), or non-transmitted mutations. However, systematic errors, where the same error from a common source is introduced many times in otherwise distinct viral genome sequences, are not removed by that approach (Freeman et al. 2020; NicolaDeMaio et al. 2020). These are more problematic, as they can appear as if they are genuine transmitted mutations. This might result from recurring errors in data generation or processing, or due to contamination among samples. Each case induces an apparent mutation that may be challenging to rectify with the real structure of the viral tree. Consequently, systematic errors can produce support for erroneous relationships between viral isolates and destabilize tree-building efforts. One possible approach is to mask out

specific sites in the genome sequence where recurring errors are suspected, as suggested previously (NicolaDeMaio et al. 2020). However, genuine recurrent mutations that may contain important information about properties of viral evolution (van Dorp et al. 2020, n.d.; B. Korber et al., n.d.; Yi 2020; Lythgoe et al., n.d.) are sometimes hard to distinguish from recurrent systematic errors, and this could obscure important biology.

Another basic problem in current investigations of viral evolution is widespread phylogenetic uncertainty. Many groups have inferred phylogenetic trees with widely varying goals. Consequently, resulting topologies vary dramatically in structure, owing to differences in analysis choices and to phylogenetic uncertainty stemming from limited genetic diversity in the expanding viral populations. Consistent approaches for identifying commonalities and rectifying differences among trees are therefore foundational to the efforts to characterize viral evolution and epidemiology. A maximally stable topology is essential for consistent nomenclature and facilitating conversations between research groups (Dellicour, Durkin, Hong, Vanmechelen, Martí-Carreras, Gill, Meex, Bontems, André, Gilbert, Walker, De Maio, Hadfield, et al., n.d.; Rambaut et al., n.d.).

In this chapter, I examine both systematic errors and phylogenetic uncertainty. First, I demonstrate that hundreds of samples in the current SARS-CoV-2 sequencing datasets are affected by lab-associated mutations, which are potentially erroneous (see also (NicolaDeMaio et al. 2020)). These mutations distort phylogenetic inferences at scales most relevant to local lineage tracing and impact inferred patterns of mutational recurrence and recombination. However, many can be identified and removed by cross-referencing patterns of recurrence against the source sequencing lab, and we provide automated methods for detecting suspicious and highly recurrent mutations. Second, to facilitate communication and comparison across different SARS-CoV-2 phylogenies, we develop approaches for efficiently comparing and visualizing differences among trees. The tools outlined here have since been widely adopted by the SARS-CoV-2 research community.

The text of this dissertation includes reprints of the following previously published material: Turakhia, Y.*, De Maio, N.*, Thornlow, B.*, Gozashti, L., Lanfear, R., Walker, C. R., ... & Corbett-Detig, R. (2020). Stability of SARS-CoV-2 phylogenies. *PLoS Genetics*, 16(11), e1009175. The co-authors listed in this publication directed and supervised the research which forms the basis for the dissertation.

3.2: Systematic error could be mistaken for recurrent mutation or recombination.

Non-random errors can present a fundamental challenge for phylogenetic inference and to the interpretation of viral evolutionary dynamics. There are at least four possible sources of (real or apparent) mutations that recur within independent lineages in a tree, and each makes distinct predictions about the source of recurrent mutations (Table 4.1). In particular, recent work has shown a strong bias towards C>U mutation in the SARS-CoV-2 genome (Rayko and Komissarov, n.d.; Dellicour, Durkin, Hong, Vanmechelen, Martí-Carreras, Gill, Meex, Bontems, André, Gilbert, Walker, De Maio, Hadfield, et al., n.d.; Rice et al., n.d.; Xia 2020). Systematic errors, which usually result from consistent errors in molecular biology techniques or bioinformatic data data processing, need not reflect this bias and are not subject to natural selection. We therefore anticipate that many systematic errors will affect many mutation types, modify protein sequences, and strongly correlate with genome sequences generated in particular labs (NicolaDeMaio et al. 2020).

Source	Heritable	Typical Allele Frequency	C>U Biased	Lab Correlation	Extremal*
Recurrent Mutation	Y	Low-Moderate	Y**	N	Y
Recombination	Y	High	Possible	N	N
Systematic Error	N	Low	Possible	Y	Y
Contamination Error	N	High	N	Possible	N

Table 3.1: Expectations for various sources of apparent recurrent mutation.
**As defined in the main text (below). **Owing to ours and previous works, we expect that most recurrent mutations will usually demonstrate a C>U bias; however, this may not be uniformly the case for example in mutation hotspots.*

3.3: Many apparently recurrent mutations found in SARS-CoV-2 genome sequences

To examine patterns of recurrent mutation we employ a simple statistic, the parsimony score, which is the count of the minimum number of unique mutation events consistent with a tree and sample genotypes ((Fitch 1971; Sankoff 1975), computed using our software from https://github.com/yatisht/strain_phylogenetics, Methods). More sophisticated statistics could be employed, but this simple one is effective, is readily interpretable, and can be computed rapidly. We restrict most analysis to bi-allelic sites, i.e. sites that contain one the allele in the reference genome from the root of the tree (here and in Nextstrain this is, Wuhan-Hu-1, obtained in December 2019 in the city of Wuhan) and a single alternate allele. Across the 4/19/2020 Nextstrain tree, we found 2533, 395, 94, 40, and 44 bi-allelic sites with parsimony score one, two, three, four, and five or more, respectively (Figure 4.1). In particular, there is a strong “on diagonal” component of the data that is defined by a linear relationship between the log of the alternate allele count and parsimony score (dashed line in Figure 4.1A, log2-based slope = 3.188). These mutations reoccur across the phylogeny at exceptional rates relative to their allele frequencies. Hereafter, we refer to the set of variants in this on-diagonal group as extremal sites (blue, red, and orange in Figure 4.1A). This relationship suggests that the extreme accumulation of independent clades for the alternate allele is logarithmically related to the number of instances of the alternate allele in the phylogeny (Figure 4.1B). This suggests that even the most mutable or error prone sites in the genome will sometimes have alternate alleles grouped into clades during phylogenetic inference thereby appearing to be inherited.

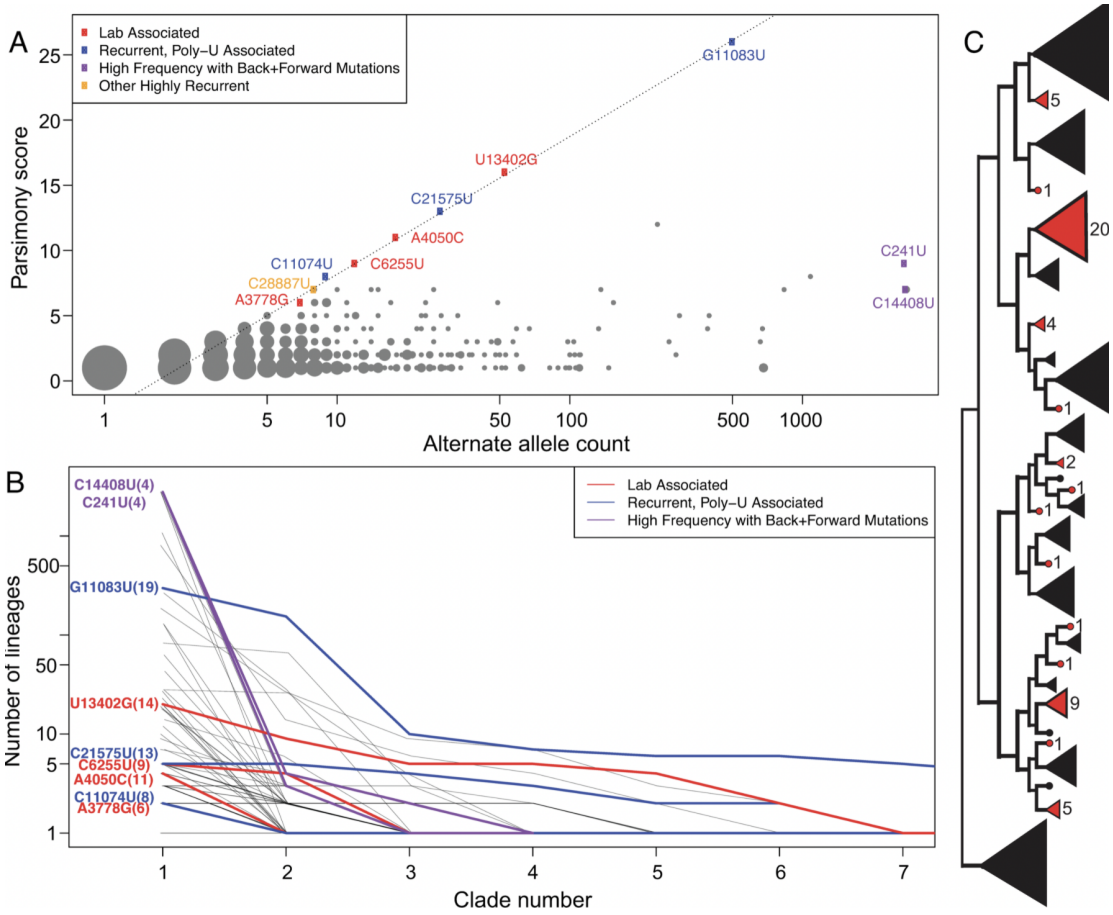


Figure 3.1: The relationship between alternate allele count and parsimony score. (A) *The relationship between alternate allele count and parsimony score. Point radius indicates how many sites share a single parsimony score and alternate allele count. Several noteworthy recurrent mutations are labelled. Note that the X-axis is log-scaled.* ***(B)*** *The sizes of independent clades for the same alternate allele arranged in descending order. The number of lineages per clade is shown on logarithmic scale facilitating comparison with Panel (A). These data indicate that when alternate allele clade sizes for a given site are sorted in decreasing order, their sizes are reduced going from left to right by a multiplicative factor at each step, consistent with the log-linear relationship displayed in Panel (A). Mutations with remarkably high recurrence are shown with color reflecting their properties: lab-associated (red), recurrent and associated with a poly-U stretch (blue), and high frequency with many forward and backward mutations (purple). Grey lines in the background are the same values but for all other mutations with parsimony score 4 or greater. The values in parentheses in the mutation names indicate the number of unique clades associated with the alternate allele. Note that in some cases, this extends beyond the limit of the X-axis and that the Y-axis is log-scaled for visibility.* ***(C)*** *An example of the observed patterns of evolution at one highly recurrent site with reference allele U and alternate allele G, site 13402 and parsimony score 14, where 14 alternate allele clades (in red) each represent an apparently independent incidence of the mutation substituting the alternate allele.*

3.4: SARS-CoV-2 data contains many lab-associated errors.

To search for systematic errors associated with a particular lab, we extracted the set of sites with parsimony score 4 or more. We then flagged sites as lab-associated mutations if more than 80% of the samples containing the alternate allele were generated by a single group. Using this heuristic approach, we found 16 such sites. We note that this set of sites contains two mutations previously identified as lab-associated mutations (NicolaDeMaio et al. 2020), some others identified as highly homoplastic (Dellicour, Durkin, Hong, Vanmechelen, Martí-Carreras, Gill, Meex, Bontems, André, Gilbert, Walker, De Maio, Hadfield, et al., n.d.; van Dorp et al. 2020; NicolaDeMaio et al. 2020; van Dorp et al., n.d.), as well as several identified as evidence for recombination (B. Korber et al., n.d.). These mutations in lab-associated sites display a range of base compositions and only one is a C>U transition (C6255U). This rate of C>U mutation is much less than the genome-wide average rate of C>U mutation for non-singleton sites (49%, P = 0.0004914, Fisher's exact test), and differs significantly from the rate of C>U mutation among our set of highly recurrent mutations that are not strongly associated with a single sequencing lab (P = 1.005e-07, Fisher's exact test). Furthermore, our set of lab-associated mutations is weakly enriched for protein altering mutations relative to other highly recurrent mutations (P = 0.09372). Collectively, our results suggest that some recurrent mutations among these 16 could be lab-associated systematic errors.

The potential causes of lab-associated mutations are numerous. A non-exhaustive list follows. First, primers for reverse transcription or PCR might introduce systematic errors either via errant priming, because they "overwrite" true variation, or because of errors during bioinformatic processing. For example, the commonly used ARTIC primer sets amplify the viral genome from metatranscriptomic cDNA by tiling the viral genome with PCR amplicons (https://artic.network/). Second, if a portion (perhaps a single amplicon) from a contaminating sample were present in many sequencing reactions from a single lab, this could propagate

variants across all genome sequences from a single group. Third, contamination from the human transcriptome itself might be inadvertently included in assembled viral genomes.

Two labs contributed a disproportionate number of lab-associated mutations in our dataset, suggesting a consistent source of these alternate alleles. One lab group is strongly associated with two adjacent high parsimony score mutations A24389C and G24390C. These occur in a 10bp sequence that otherwise closely resembles an Oxford Nanopore sequencing adapter, CAGCA**CC**TT, and is adjacent to an ARTIC primer binding site. Here, the differences between the genome sequence and adapter are bolded (NicolaDeMaio et al. 2020). Additionally, A4050C, U8022G, U13402G, and A13947U (Figure 3.2) are associated with this same lab and either overlap or are within 10bp of ARTIC primer binding sites (14_left_alt4, 26_right, 44_right, and 47_left, respectively), suggesting that a consistent bioinformatics data processing error may be responsible. Sequences submitted by another lab group are strongly associated with four additional high parsimony score mutations, G2198A, G3145U, A3778G, and C6255U (Figure 4.1). Here again, each of these intersects one of the ARTIC primer binding sites (8_left, 11_left, 13_left and 20_right respectively, Figure 3). In aggregate, our set of lab-associated mutations are significantly closer to ARTIC primer binding sites than would be expected by chance (P = 0.0283, permutation test, Figure 3). Another lab-associated mutation, C22802G, also overlaps an ARTIC primer (76_left), but the ultimate source is unrelated. In this case, the cause appears to be misalignment of a human ribosomal RNA sequence that was incorporated into the consensuses for a subset of genomes produced by this group (Dr. Darrin Lemmer, *Pers. Comm.*). Our results highlight the broad range of possible causes of lab-associated mutations.

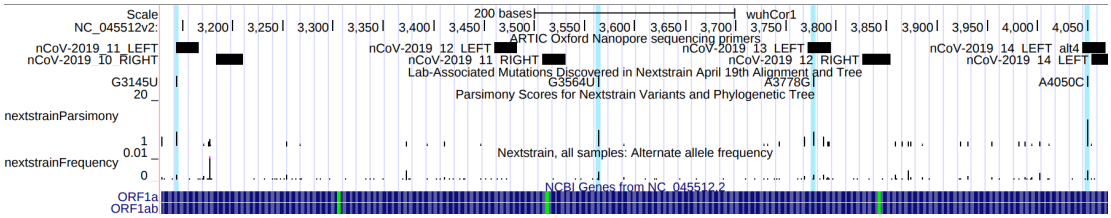


Figure 3.2. UCSC Genome Browser display of lab-associated mutations and ARTIC primers. *Bases 3130 to 4070 of the SARS-CoV-2 genome are displayed, containing four lab-associated mutations highlighted in light blue. G3145U, A3778G and A4050C overlap ARTIC primer bind sites. An interactive view of this figure is available from: http://genome.ucsc.edu/s/SARS_CoV2/labAssocMuts.*

3.5: Lab-associated mutations are consistent with simulated systematic error.

To study how systematic errors affect phylogenetic inference, we introduced simulated errors in replicate experiments. We found that the parsimony score displays a roughly linear relationship with the log of the alternate allele count, as it does for extremal sites in Nextstrain built on different days in April, but with varying slope (Figure 3.3). This is expected because errors will sometimes occur in sample genomes whose positions are close on the real phylogeny and in sister lineages. Tree-building methods could then group these samples into a single clade. Importantly, the effect of drawing samples together can cause systematic error, or hypermutable sites for that matter, to appear heritable.

Additionally, we find that viral genetic background and mutation type is an important contributor to this relationship. When errors are placed randomly across Australian samples (Figure 3.3A), we see much higher parsimony scores than when errors are placed only in samples from France collected between March 1 and March 17 (Figure 3.3B). The difference likely reflects the fact that the samples from France are more closely related. Because many of the lab-associated mutations that we identified are derived from a similarly restricted time and geographic region as our samples from France, parsimony scores at those sites closely resemble these sets of simulated error (Figure 3.3B). This suggests that the identification of lab-associated mutations becomes increasingly straightforward as the viral populations accumulate genetic diversity. We also observe that mutations that truly occur less often

during SARS-CoV-2 evolution (*e.g.,* C to G) have slightly lower parsimony scores. This is likely due to modelling nucleotide-specific mutation rates during tree-building where mutations consistent with viral mutational processes are less likely to be erroneously grouped. Our results suggest that a simple heuristic based on each site's parsimony score and recurrence is sufficient to identify most lab-associated mutations above very low frequencies. However, extremely infrequent lab-associated error could be challenging to distinguish from more conventional sequencing error.

Because systematic errors also affect the inferred tree, they can impact inferred patterns of mutational recurrence at other positions in the genome. In 50 out of 54 total experiments where we introduced a single recurrent error, we found that the parsimony score increased at other sites (range 2 to 44). This emphasizes the importance of identifying and excluding such mutations prior to inferring the final tree and downstream analyses.

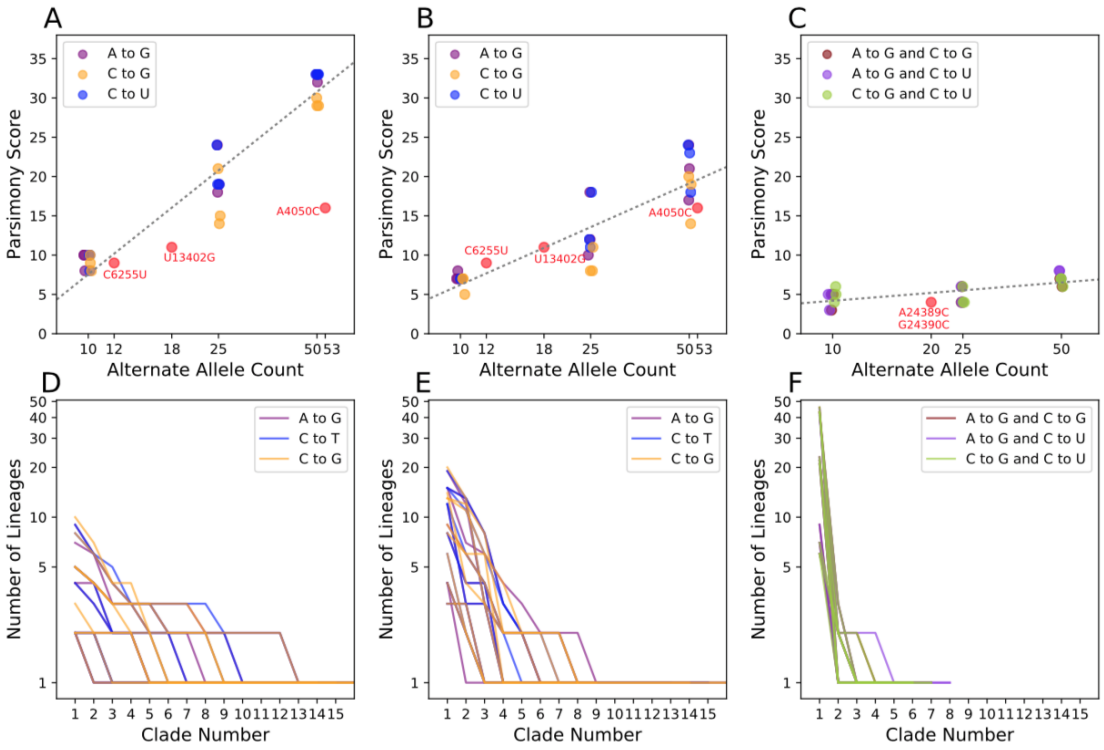


Figure 3.3: Parsimony scores at sites with introduced systematic errors. *We added artificial errors to 10, 25, and 50 Australian (A) and early-March French (B) samples at the*

sites A11991G (purple), C22214G (blue), and C10029U (orange) in three replicates, then produced phylogenies and computed the parsimony score at each site. (C) We also introduced errors to the early-March French samples two at a time per sequence rather than individually. For comparison, we also show the values for three lab-associated mutations (C6255U, U13402G, A4050C; A, B) and for pair of linked lab-associated mutations (A24389C and G24390C; C). Each panel (A-C) contains a best-fit line (as in Figure 3.1A), for the relationship between log2 alternate allele count and parsimony in simulated error data (slopes = 10.0, 5.55, and 1.0). (D-F) Corresponding clade sizes arranged in descending order for error simulations in (A-C, respectively, as in Figure 3.1B).

3.6: Correlated lab-associated mutations have large impacts on phylogenetic inference.

If infrequent but highly correlated errors were introduced at different sites in many samples, this could cause more samples to be grouped into a clade. We might not easily detect these errors based on recurrence. Two lab-associated mutations, A24389C and G24390C, are not just on adjacent genomic locations but are nearly perfectly correlated across samples. These sites have low parsimony scores when compared to other lab-associated mutations (4 and 5, respectively, Figure 3.4C). When we introduced similar correlated errors, we found that the parsimony scores were lower than in single error introduction experiments. Nonetheless, in only two error introduction experiments (out of 9) with 10 affected samples did we see a parsimony score as low as 3. Although low frequency and highly correlated error could be challenging to identify in general, we believe this is infrequent in our dataset. Therefore we have not included tests for correlated errors in our suggested methods for finding lab-associated mutations, but adjacent correlated sites should be carefully scrutinized.

3.7: Lab-associated mutations affect phylogenetic inferences on scales relevant to local lineage tracing

To investigate the impacts of lab-specific mutations on phylogenetic inference, we removed (“masked”) each of the 16 sites with a lab-associated mutation. Importantly, removing lab-associated mutations sometimes impacted phylogenetic patterns at other sites.

For example, after removing all lab-associated mutations, the evidence for back-mutations at C14408U is eliminated, while many forward-mutations remain (*e.g.*, Figure 3.4). In fact, the parsimony score changed for 107 sites and decreased for 53 sites on the tree that we inferred after removing all of the lab-associated mutations relative to the tree inferred including all sites. Additionally, we find that many samples containing lab-associated mutations have been repositioned on local topologies (*e.g.,* Figure 3.4). Furthermore, in some cases the placement of closely related lineages that are unaffected by lab-associated mutations is also affected. These mutations therefore affect phylogenetic inferences at scales relevant to local lineage tracing, which may obscure dynamics of local transmission.

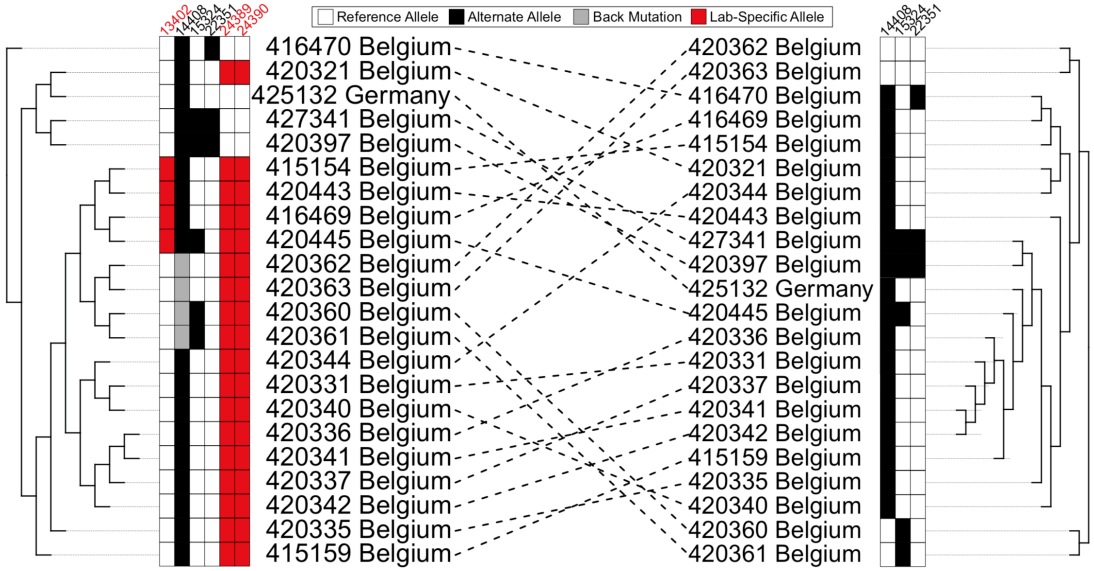


Figure 3.4. Lab-associated mutations impact phylogenetic inferences. *Part of the tree we inferred from the 4/19/2020 Nextstrain dataset (left) compared to the corresponding part of tree after removal of sites with lab-associated mutation (right). Lab-associated mutations (red) can affect the inferred phylogeny and are associated with apparent back-mutation to the ancestral allele (grey in column 14408, left) at other sites (white). When lab-associated mutations are removed, the resulting tree (right) shows no evidence for back-mutation at those sites (now white in column 14408), though several independent forward mutations remain evident.*

To examine the effect of each lab-associated mutation and the other extremal sites in isolation from one another, we individually masked each site and inferred a phylogeny. As a

comparison, we also masked a set of sites that have similar alternate allele frequencies as the lab-associated mutations, but each has a parsimony score of one. The distributions of entropy-weighted total distance (a measure of distance between trees, described below) are remarkably similar when masking individual lab-associated sites, other extremal sites, and our control sites (Figure 3.5). Most exceed the distance we observed when we independently inferred two trees from the same input alignment (dashed black line). Our results therefore suggest that the lab-associated and extremal sites can impact tree-building approaches on par with real mutations, although the effects are typically small on the scale of whole topologies, as is expected given their typically low allele frequencies (Figure 3.5).

Phylogenies made after removing two mutations, one control and one lab-associated are outliers for entropy-weighted total distance (Figure 3.5) and other tree distance statistics. In each case, however, the likelihood of the tree produced from the full dataset is actually higher, suggesting that our tree-building method discovered a different locally optimal but less favorable topology rather than a dramatic impact of each site individually. These results suggest higher level uncertainty in the tree topology largely independent of the effects of lab-associated mutations.

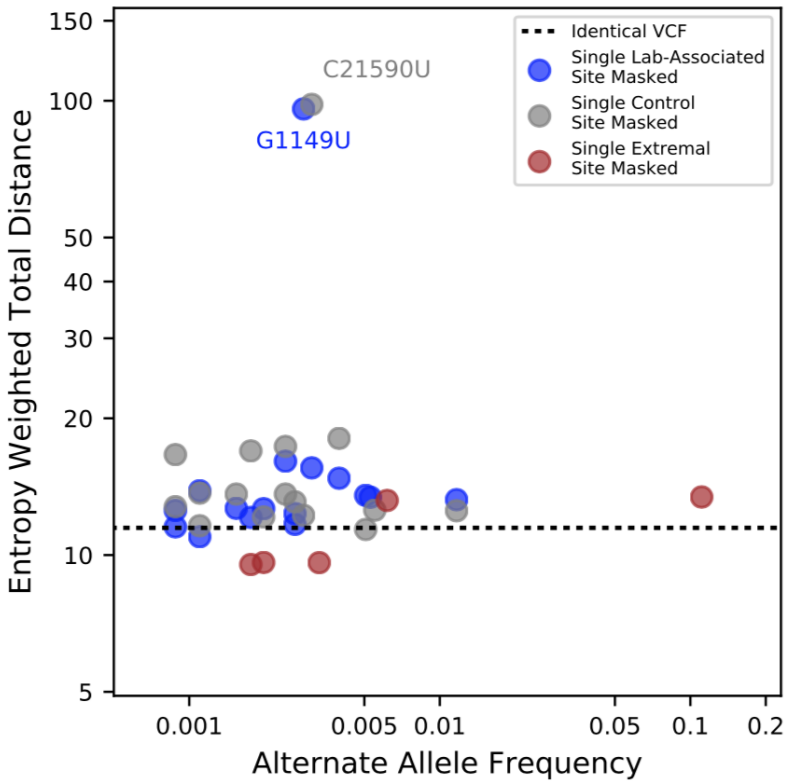


Figure 3.5: The relationship between alternate allele frequencies of lab-associated mutations and effect of masking on inferred tree topology. *Entropy-weighted total distances relative to the reference maximum likelihood phylogeny are shown for phylogenies constructed after masking individual sites. Blue points correspond to sites with lab-specific alternate alleles, grey points correspond to control sites with parsimony scores of 1 and similar alternate allele frequencies to the sites with lab-specific alternate alleles, and brown points correspond to non-lab-specific extremal sites. The black horizontal line indicates the entropy-weighted total distance value for a maximum likelihood phylogeny constructed from an alignment identical to that of the reference phylogeny. Two outliers, C21590U (control) and G1149U (lab-associated) have outsize effects on inferred tree topology.*

3.8: Nextstrain phylogenies vary significantly over time.

As expected for a relatively slowly-evolving and rapidly expanding viral population (Sanjuán et al. 2010), there is substantial uncertainty in the SARS-CoV-2 phylogeny. This extends well beyond the typically localized impacts of lab-associated and highly recurrent mutations, and instead derives from the fact that most branches in the SARS-CoV-2 phylogeny are supported by few mutations. Undoubtedly, thousands of unique phylogenies

will be produced by groups studying this viral outbreak and these may sometimes support conflicting evolutionary relationships. Our research team therefore sought to provide tools to facilitate interpretation of commonalities and differences among such large phylogenies.

We explored differences among trees made by the same group from slightly different sample sets with the goal of understanding phylogenetic stability as new samples are incorporated. For the purposes of comparison, we restricted 31 Nextstrain trees produced between March 23, 2020 and April 30, 2020 to just the 468 samples they all have in common. Comparing topologies, we found that a number of these 468 samples moved back and forth between different clade designations during the month , including samples in the specific clades (A1a, A2, A2a, A6, A7, B, B1, B2, B4) named and analyzed by the Nexstrain consortium during this period. Note that the Nextstrain clade ID system was updated while we were finalizing this work (Hodcroft EB, Hadfield J, Neher RA, Bedford T 2020). We then measured all pairwise tree distances between restricted trees and found that they varied widely (normalized entropy-weighted total distances ranged from 0.089 to 0.352, Figure 3.6). There is therefore substantial variation in Nextstrain phylogenies over time.

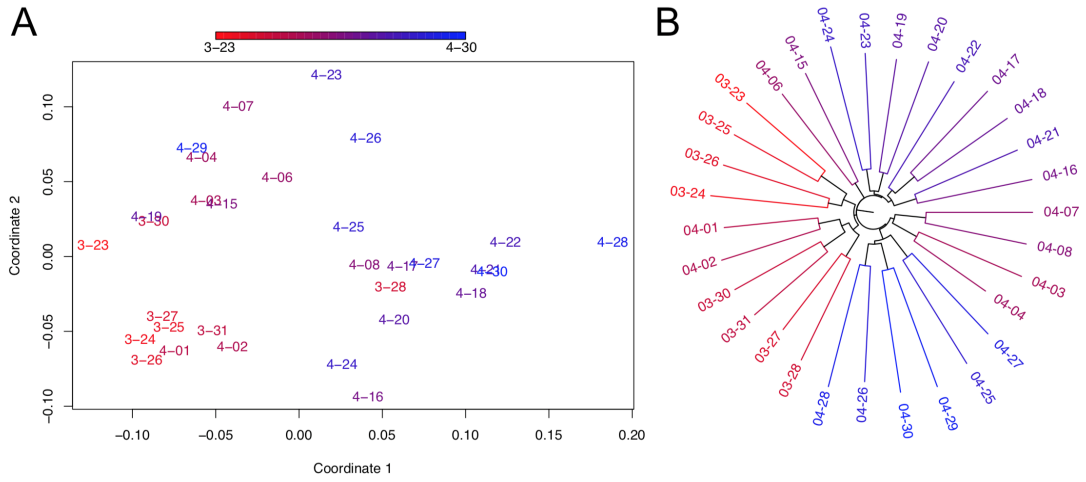


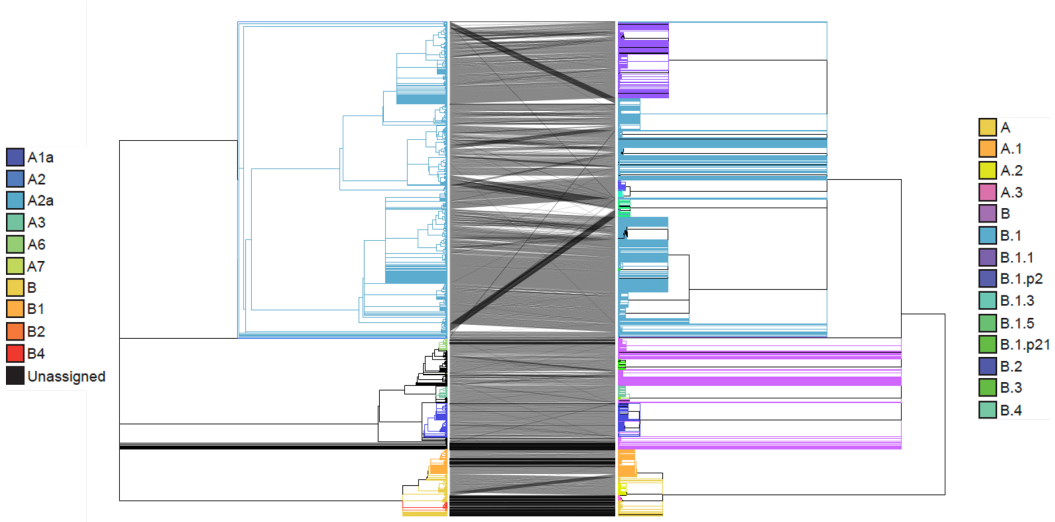
Figure 3.6. Comparisons of Nextstrain trees over time. *(A) Multidimensional scaling of normalized entropy-weighted total distances among phylogenetic trees produced by Nextstrain from March and April. Each topology is labelled with its date and dates are depicted in a color gradient from 3/23 (red) to 4/30 (blue). Coordinates 1 and 2 are plotted*

here and each contributes 34% and 15% of the total variance explained, respectively. (B) Relationships between Nextstrain phylogenies are shown in a tree-of-trees, "meta-tree" (Nye 2008) we constructed, which displays the distances among topologies of the constitutive trees. .

3.9: Higher-level branches are remarkably consistent across analyses.

Even if it was possible to obtain error-free data and multiple alignments as well as have all groups use that same data, different tree inference approaches can produce different topologies. Furthermore, there is substantial uncertainty inherent to SARS-CoV-2 evolution because there are few mutations that uniquely mark each branch. Nonetheless, it is essential that epidemiologists studying the pandemic be able to communicate phylogenetically informed observations (Rambaut et al., n.d.; Dellicour, Durkin, Hong, Vanmechelen, Martí-Carreras, Gill, Meex, Bontems, André, Gilbert, Walker, De Maio, Hadfield, et al., n.d.). As discussed above, the clade placements of individual samples, even when inferred by the same group, can vary as different datasets are incorporated into the tree construction process (*e.g.* Figure 3.6). Differences between groups are expected to be even more pronounced. Without a 1-1 correspondence between the topologically defined clades in the Nextrain and COG-UK phylogenetic trees, it is difficult to translate nomenclature in order to conduct precise scientific discourse pertaining to the evolutionary conclusions reached by these groups. As clade based comparisons are an essential part of consistent scientific discourse, tools are needed to ameliorate these difficulties.

A



B

NS	COG-UK	J
A1a	**B.2**	**0.948**
A2	**B.1.1**	**0.993**
A2a	**B.1.1**	**0.997**
A3	B.4	1.0
A6	B	0.027
A7	B	0.001
B	**A**	**1.0**
B1	**A.1**	**0.916**
B2	A	0.038
B4	A	0.088

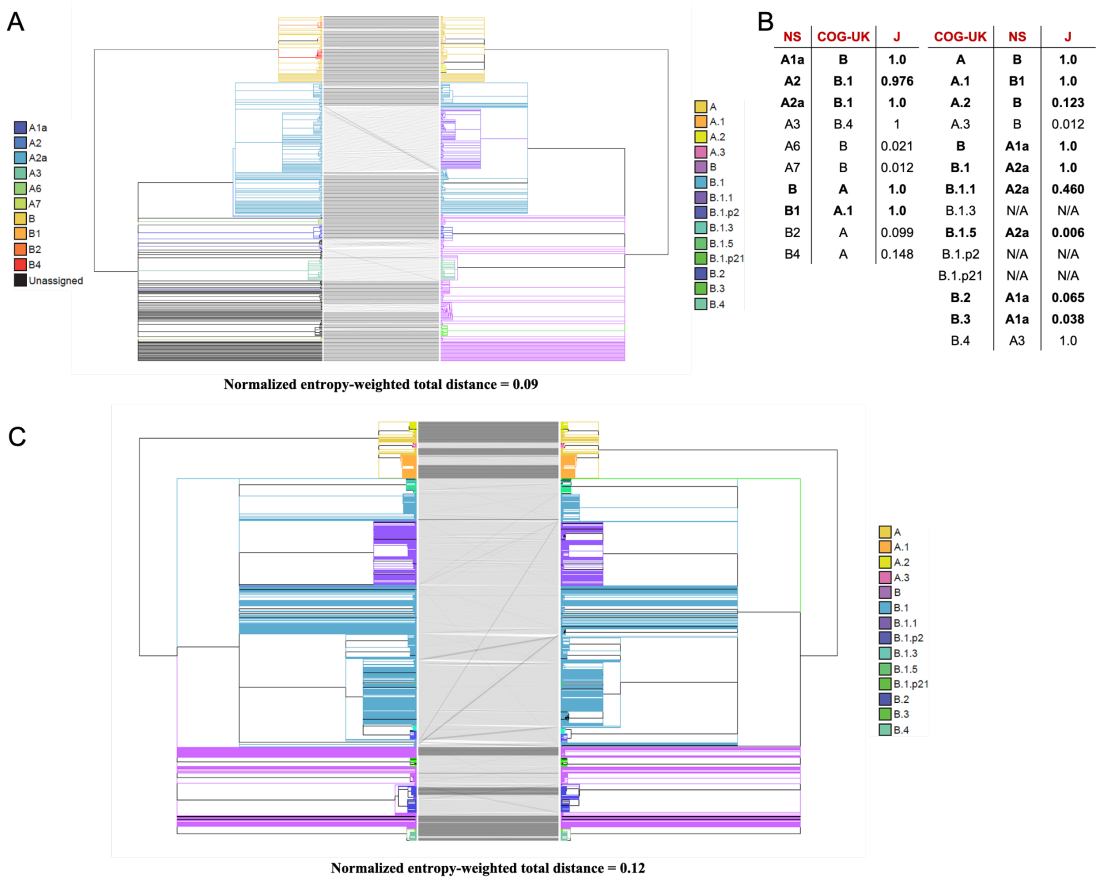
COG-UK	NS	J
A	**B**	**1.0**
A.1	**B1**	**0.916**
A.2	B	0.182
A.3	B	0.073
B	**A2**	**0.741**
B.1	**A2**	**0.741**
B.1.1	**A2a**	**0.997**
B.1.3	A2a	0.019
B.1.5	A2a	0.057
B.1.p2	A2a	0.033
B.1.p21	A2a	0.008
B.2	**A1a**	**0.948**
B.3	**A2**	**0.741**
B.4	A3	1.0

Figure 3.7: Comparison of Nextstrain and COG-UK trees. (A) *A tanglegram of the Nextstrain tree from 4/19 (left) with the COG-UK tree from 4/24 (right). Each tree has 4167 samples.* ***(B)*** *The COG-UK clades (which they term "lineages") having the highest Jaccard similarity coefficient (J) with each Nextstrain (NS) named clade and vice versa, where the Jaccard similarity coefficient is computed using the set of samples from the root of that clade. Clades with more than 200 samples are shown in bold font and considered "big", the others "small". For each big Nexstrain clade there is a closely corresponding COG-UK clade, and vice-versa.*

To explore the differences among available phylogenies and to provide guidelines for clade-based comparisons across possible evolutionary histories, we used our approach to

identify the correspondence between the Nextstrain phylogeny produced on April 19, 2020 and the COG-UK phylogeny produced on April 24, 2020 (Figure 3.7A). We observe agreement between the big Nextstrain-named clades and their corresponding best matching named clades in the COG-UK tree and vice versa (*e.g.,* “A2a” clade in Nextstrain, “B.1” clade in COG-UK, etc, Figure 3.7B), suggesting that these clades are reasonably stable across different analyses. However, in small named subclades within those big clades, there are many noteworthy differences between the two topologies, and the overall congruence is significantly reduced (Figure 3.7A). In addition to differences in methodology, this reflects a difference in the time when clades were originally named and the intents of each nomenclature system. Nextstrain named clades much earlier and many did not increase in size subsequently, others have since emerged and were named by COG-UK later. Additionally, the COG-UK system is intentionally dynamic and clades that have become inactive are removed. As a consequence, some clades do not have an obvious named analog in the two systems resulting in low similarities (Figure 3.7B).

Perhaps the most obvious difference between the topologies is that the COG-UK tree has many more large polytomies (Figure 3.7A). This reflects the decisions motivating their analysis (Rambaut et al., n.d.; “An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network” 2020), where the authors’ goal is to provide a well-supported and stable topology to facilitate lucid communication about viral lineages for evolutionary as well as epidemiological studies. This contrasts with the Nextstrain consortium’s primary goal of up-to-date transmission tracing. As is typical in phylogenetics, topological stability comes as a tradeoff against the cost of articulation in the branches. Because of the many different motivations for constructing phylogenetic trees, it is a certainty that many independent trees will be used to study the evolution of SARS-CoV-2. Comparisons using our approaches can enable communication about evolving viral lineages across disparate analyses by facilitating the identification and visualization of the most closely matching clades.



B

NS	COG-UK	J	COG-UK	NS	J
A1a	**B**	**1.0**	**A**	**B**	**1.0**
A2	**B.1**	**0.976**	**A.1**	**B1**	**1.0**
A2a	**B.1**	**1.0**	**A.2**	**B**	**0.123**
A3	B.4	1	A.3	B	0.012
A6	B	0.021	**B**	**A1a**	**1.0**
A7	B	0.012	**B.1**	**A2a**	**1.0**
B	**A**	**1.0**	**B.1.1**	**A2a**	**0.460**
B1	**A.1**	**1.0**	B.1.3	N/A	N/A
B2	A	0.099	**B.1.5**	**A2a**	**0.006**
B4	A	0.148	B.1.p2	N/A	N/A
			B.1.p21	N/A	N/A
			B.2	**A1a**	**0.065**
			B.3	**A1a**	**0.038**
			B.4	A3	1.0

Figure 3.8: Comparison of our consensus tree to the COG-UK trees. (A) *A tanglegram of our Nextstrain consensus tree (left) and COG-UK tree from 4/24 (right). Each tree has 422 samples.* ***(B)*** *The COG-UK lineages having the highest Jaccard similarity coefficient (J) with each Nextstrain consensus (NS) named clade and vice versa. "Big clades" are in bold. Lineages in 'N/A' (B.1.3, B.1.p2 and B.1.p21) were pruned out as a result of restricting the trees to common samples.* ***(C)*** *A tanglegram of our tree produced after masking all lab-associated and extremal mutations except 11083 (left) and COG-UK tree from 4/24 (right). Each tree has 4172 samples and the samples (branches) have been colored based on COG-UK lineage labels.*

3.10: Higher branches in our tree closely mirror a Nextstrain "consensus" tree and the COG-UK tree

To identify stable nodes across analyses we compared a Nextstrain "consensus tree" and the COG-UK tree. To do this, we produced a majority rule clade consensus tree (Margush and McMorris 1981) for the 422 common samples in 31 Nextstrain releases

between 3/23 to 4/30, and restricted the COG-UK tree to these same samples. We find exceptionally good congruence between our Nextstrain consensus and the COG-UK phylogenies (Figure 3.8A), even though the inference methods differed substantially. Specifically, the COG-UK tree is built using a more typical bootstrapping approach (Hoang et al. 2018) whereas our approach for building a Nextstrain “consensus” from trees produced on subsequent days would resemble a kind of “bootstrapping by samples” approach. This congruence reaffirms the idea that the COG-UK tree provides a stable “backbone” to enable direct conversations in epidemiology. Nonetheless, we still observe several small rearrangements between the two topologies, suggesting that both will likely be subject to clade refinements in the future.

We also observed congruence between the tree that we produced after removing lab-associated and extremal mutations (except 11083, see above) and the COG-UK tree (Figure 3.8C). Here, the sample size is much larger, 4172, allowing for a much more quantitative comparison. The correspondence between the two trees is very high with normalized entropy weighted total distance of just 0.12. Because lab-associated and extremal mutations were used in the COG-UK tree but not in our tree, this consistency among topologies supports our assertion that the effect of lab-associated and extremal mutations will typically not result in large-scale reorganizations of large clades across the phylogeny.

3.11: Methods

To detect mutations that reoccur many times through viral evolution, we computed the parsimony score (Fitch 1971; Sankoff 1975) for each polymorphic site (our program is available from https://github.com/yatisht/strain_phylogenetics). Briefly, conditional on a tree, we compute the minimum number of branches that have experienced a mutation at a single site to accommodate the phylogenetic distribution of the mutant and reference allele. These are candidate highly recurrent mutations, but we note that these mutations, or others elsewhere on the chromosome, might also impact the process of tree building itself, and the

score should be interpreted with caution if counting the specific rate of occurrence at a given site is of interest.

We systematically flagged possible variants resulting from lab-specific biases based on the proportion of lab-specific alternate allele calls and respective alternate allele frequency (https://github.com/lgozasht/COVID-19-Lab-Specific-Bias-Filter). To do this, we first filtered variants with parsimony score greater than 4 using concurrent Nextstrain tree and vcf files from 4/19/2020. Next, we obtained metadata for all COVID-19 genomes on GISAID (accessed 4/28/2020) and computed the proportion of alternate allele calls contributed by each “originating lab” and “submitting lab” for each filtered variant. We then employed a Fisher’s exact test associating the number of major and alternate alleles attributed to each specific “originating” and “submitting” lab and the respective global major and alternate allele counts. We flagged variants for which one lab accounts for more than 80% of the total alternate allele calls and for which a Fisher’s Exact Test suggests a strong correlation (at the $p < 0.01$ level) between that lab and samples containing the alternate allele. We note that these cutoffs are somewhat arbitrary, and may require modification in the future, but the subdivision of the data is consistent with our expectations as described in Results. Because samples are not independent and identically distributed, p-values may not reflect error but rather relatedness among samples sequenced at a single facility. For example, if a single lab sampled a transmission chain, many mutations could be strongly associated with that facility. These should be interpreted cautiously, however, there is no obvious reason why unrelated samples sequenced at the same facility should share an excess of homoplasious mutations.

We obtained the phylogenetic tree hosted by Nextstrain (accessed 4/19/2020) and used this in our comparisons of clades among trees and as our primary data object for examining apparently recurrent mutation on the tree. We did separately confirm that most apparently recurrent mutations are recovered on the trees produced on different days by Nextstrain.

For comparison of clades among different tree-building approaches, we obtained variant datasets, and phylogenies from Nextstrain (https://nextstrain.org/ncov accessed 4/19/2020-4/26/2020), and from COG-UK (https://cog-uk.s3.climb.ac.uk/20200424/cog_2020-04-24_tree.newick, accessed 4/24/2020)

From the 04/19 Nextstrain release, we created a “reference phylogeny“ using IQ-TREE-2 (L.-T. Nguyen et al. 2015; Minh et al. 2020) to build phylogenies from each of these alignments using the GTR+G nucleotide substitution model. For all other phylogenies, we altered the input by removing or “masking” individual sites, then produced phylogenies from these altered alignments using the same IQ-TREE-2 parameters.

The likelihood of a tree given the alignment from which it was constructed was automatically calculated by the IQ-TREE command used above (*iqtree -s <alignment.phy> -m GTR+G*). However, to compute the likelihood of a particular alignment given a different tree, we used the command *iqtree -s <alignment.phy> -te <phylogeny.nh> -m GTR+G*.

To generate our final tree having masked lab-associated and extremal mutations, we used the same command but also included the ultrafast bootstrapping option “-bb 1000” to assist with quantifying uncertainty in our final phylogeny (Hoang et al. 2018). We used the same command but included the full multiple alignment to compare the tree obtained to one obtained from the full dataset using identical methodology. Finally we collapsed all branches that were not supported by at least one mutation using parsimony to identify nodes that experienced a mutation.

To investigate the effects of lab-specific alleles on phylogenetic topology, we also introduced artificial errors at control sites. We chose three sites at which to introduce these errors: A11991G, C22214G, and C10029U. To introduce an error, we manually changed a reference allele to an alternate allele for a given sample at a given site. For each of these sites, we chose 10, 25, and 50 samples for which we introduced errors. To mimic the effects of a lab-specific allele, we ensured that each set of samples with artificial errors must come from the same country. We chose Australia due to its high representation in the Nextstrain

data, as 372 samples in the 04/19 Nextstrain release came from Australia. To further mimic lab-specific behavior, we separately introduced errors at the same sites for 10, 25, and 50 randomly selected French samples collected between March 1 and March 17. After introducing these errors, we constructed phylogenies from the modified alignments using IQ-TREE 2 (L.-T. Nguyen et al. 2015; Minh et al. 2020), as described above. In total, we produced 54 phylogenies in this experiment, introducing errors at three sets of random samples for each of the three sites, at 10, 25, and 50 samples each, for Australian and French samples.

We also repeated this experiment, but introducing errors at pairs of sites simultaneously rather than at individual sites (i.e. A11991G and C22214G, A11991G and C10029U, and C10029U and C22214G). We used the same randomly chosen sets of French samples for this aspect of the experiment, and produced phylogenies by the same methods. In total, we produced 27 phylogenies in this experiment, introducing errors at three sets of randomly chosen samples, at 10, 25 and 50 samples each, for each of the three pairs of sites.

To understand commonalities in tree structure over time, we used multidimensional scaling of a distance matrix of normalized entropy-weighted total distances among Nextstrain releases (pruned to 468 shared samples) spanning from March 23 to April 30. To do this, we used the cmdscale() function in base R and retained the first six coordinates because they accounted for the vast majority of the total variance explained (approximately 80%). We computed the correlation between our distance matrix and the proportion of samples shared among topologies produced each day using a Mantel test implemented within the ade4 package in R. We also used the meta-tree package (Nye 2008) to relate Nextstrain trees to each other.

To produce a Nextstrain consensus tree we first pruned all Nextstrain trees to a common set of samples included in each tree. We then used the sumtrees script within the dendropy package (Sukumaran and Holder 2010) to produce a majority rule consensus tree

out of each tree requiring at least 50% of trees support a clade for inclusion in the final consensus tree. Specifically, we used the sumtrees function to perform this task. In our cases, that is equivalent to requiring at least 16 of 31 trees contain a given clade to include it.

Chapter 4: Tools for Sample Placement on Existing Trees and Manipulating Mutation-Annotated Trees

4.1: Background

In late 2019, the SARS-CoV-2 virus emerged from a presumably zoonotic source to spread across human populations worldwide (T. T.-Y. Lam et al. 2020; Andersen et al. 2020; Zhou et al. 2020). By midsummer 2020, over 2,000 groups worldwide had generated 97,733 high coverage whole-genome SARS-CoV-2 sequences, made available by GISAID (Y. Shu and McCauley 2017). These vast datasets and rapid sequencing turnaround times are enabling a type of “genomic contact tracing” where genetic similarities (or dissimilarity) between viral genomes isolated for different hosts carries important information about the transmission dynamics of the virus. For example, these data can be used to infer the number of unique introductions of the viral genome in a given area (Stefanelli et al. 2020; Surleac et al. 2020; Deng et al. 2020; Pattabiraman et al. 2020; Maurano et al. 2020; Gámbaro et al. 2020; Thielen et al. 2020) and to identify “transmission chains” among otherwise seemingly unrelated infections (Rockett et al. 2020; Dellicour, Durkin, Hong, Vanmechelen, Martí-Carreras, Gill, Meex, Bontems, André, Gilbert, Walker, De Maio, Faria, et al., n.d.; Fauver et al. 2020; Lu et al. 2020; Bedford et al. 2020).

Despite great potential, this unprecedented and ongoing accumulation of sequencing data has overwhelmed previously existing systems for analysis and interpretation of viral transmission and evolutionary dynamics. In part, this is because typical phylogenetics applications accumulate all of the relevant sequence data before beginning phylogenetic inference. For genomic contact tracing to work effectively, each new viral genome sequence must be contextualized within the entire evolutionary history of the virus rapidly and accurately as it is collected. This could be accomplished by re-inferring the full phylogeny, but with current SARS-CoV-2 datasets, this takes more than a day even using powerful

computational resources. Alternatively, new genome sequences could be contextualized by placing samples onto an existing “reference phylogeny” and several methods have been developed for this purpose (Minh et al. 2020; Barbera et al. 2019; Löytynoja, Vilella, and Goldman 2012; Ruan et al. 2008). These methods have been used to place new samples onto a phylogeny created from a small subset of available SARS-CoV-2 isolates (Singer et al., n.d.), and to provide regular updates to a global phylogeny of SARS-CoV-2 (Lanfear, Robert 2020). Nonetheless, existing algorithms for placing sequences onto reference phylogenies are far too slow to enable real-time genomic contact tracing.

Quantification of uncertainty is a fundamental aspect of interpreting phylogenetic inferences (Simon 2020) and sample placements onto a reference phylogeny. Non-parametric bootstrapping (Felsenstein 1985) has been a cornerstone of phylogenetic inference for decades, but this is impractical for the extremely large sample sizes and the limited phylogenetic information in SARS-CoV-2 genome isolates. More recently developed methods such as Ultrafast Bootstrapping (Hoang et al. 2018; Minh, Nguyen, and von Haeseler 2013) are fast, but not applicable to the problem of placing individual samples onto a reference phylogeny. An alternative to these approaches is the approximate likelihood ratio test (Anisimova and Gascuel 2006), but its computation is prohibitively slow and interpretation challenging. Quantifying uncertainty in sample placement on reference phylogenies is therefore an important unsolved problem and particularly relevant during this pandemic.

In this chapter, I describe UShER, a novel method developed by our research group to place samples on existing phylogenies by maximum parsimony without re-inferring the entire phylogeny. I demonstrate its vast improvement over previous methods in both speed and memory usage, as well as its nearly identical performance relative to competitors in accuracy and robustness to errors and missing data. I also detail UShER's companion package, matUtils, which enables further analysis of mutation annotated tree (MAT) objects created by UShER. Finally, I describe a case in which our team used these tools together to categorize 335 whole genome sequences from Santa Cruz County, several of which shared

mutations with known Variants of Concern (VoCs). This community analysis, completed and submitted to bioRxiv one week after receiving the 335 whole-genome sequences, is a living example that phylogenetics in real time is now possible and is invaluable for ameliorating public health crises.

The text of this dissertation includes reprints of the following previously published material: Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., ... & Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, *53*(6), 809-816. McBroome, J.*, Thornlow, B.*, Hinrichs, A. S., Kramer, A., De Maio, N., Goldman, N., ... & Turakhia, Y. (2021). A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Molecular Biology and Evolution*. Thornlow, B.*, Hinrichs*, A. S., Jain, M., Dhillon, N., La, S., Kapp, J. D., ... & Corbett-Detig, R. (2021). A new SARS-CoV-2 lineage that shares mutations with known Variants of Concern is rejected by automated sequence repository quality control. *bioRxiv*. The co-authors listed in these publications directed and supervised the research which forms the basis for the dissertation.

4.2: Prior sample placement tools are inadequate for pandemic-scale phylogenetics.

Genomic contact tracing during this global pandemic necessitates algorithms that efficiently place samples onto the vast global tree. With this requirement in mind, we evaluated the performance of several existing approaches (Minh et al. 2020; Barbera et al. 2019; Löytynoja, Vilella, and Goldman 2012; Ruan et al. 2008) and compared their runtime and memory usage by adding just one additional sequence to a SARS-CoV-2 global phylogeny containing 38,342 leaves, our “reference phylogeny”, which comes from the 11/7/2020 release of (Lanfear, Robert 2020). We found that the time required to place a single sample is unacceptably large. For example, EPA-NG (Barbera et al. 2019) takes approximately 28 CPU minutes to place one sample and requires 791 GB of memory (Table 5.1).

To address the challenge of real-time sample placement, we developed a new tool called Ultrafast Sample placement on Existing tRees (UShER). UShER can place a SARS-CoV-2 sample onto our reference phylogeny in just 0.5 seconds – several orders of magnitude improvement over the next fastest tool. A part of the increased efficiency of UShER stems from its heavily optimized encoding of mutations compared to a multiple sequence alignment (MSA) and from its pre-computed data object storing the inferred histories of mutation events on the tree before placing samples during each execution.

Program	Average time to place one sample	Time range over 20 replicates	Average peak memory used (GB)	Memory range (GB) over 20 replicates
IQ-TREE 2.1.0	46m 31s	29m 56s - 68m 52s	12.85	12.82 - 12.89
EPA-NG	27m 38s	25m 19s - 31m 13s	791.82	781.80 - 800.85
PAGAN2	120m 32s	102m 5s - 156m 15s	470.74	468.10 - 473.84
TreeBeST	48+ hours	N/A	N/A	N/A
UShER (with pre-processed mutation-annotated tree)	**0.5s**	**0.40s - 0.65s**	**0.28**	**0.17 - 0.32**
UShER (without pre-processed mutation-annotated tree)	**1m 43s**	**1m 40s - 1m 46s**	**1.02**	**0.99 - 1.04**

Table 4.1: Average time and time range required to place one sample and peak memory usage across 20 replicate runs of each placement algorithm. *A typical use case for placing SARS-CoV-2 samples onto the global phylogeny will often require placing 10–100 sequences. We did not evaluate that in this study because we found that several other algorithms could not be run on larger sample sets due to exceptionally high memory usage and runtimes. Note that while the other tools use an MSA as input, UShER accepts a VCF for new samples, which can be generated very quickly (compared to adding sequences to an existing MSA) using pairwise alignments (in, for example, minimap2 (H. Li 2018)) and whose overhead we ignore. We also note that TreeBeST was not developed explicitly for this purpose; we include it in this table because it has tree placement capabilities. UShER's time and memory usage are highlighted in bold. N/A, not applicable.*

4.3: UShER stores its data in an efficient mutation-annotated tree (MAT)

Existing approaches to sample placement use a Multiple Sequence Alignment (MSA) of genomes that requires storing a whole-genome sequence for each sample (Figure 4.1, Methods). UShER's primary data structure is substantially more efficient. It starts with a list of variants with respect to a reference sequence for each sample and represents genotype data based on the inferred phylogeny of the viral population itself. UShER uses the Fitch-Sankoff algorithm to infer the placement of mutations on a given tree and on the variant list (Fitch 1971; Sankoff 1975). Besides the phylogeny itself, UShER records only the nodes for which mutations are inferred to have occurred on the branches leading to them in a representation that we call mutation-annotated tree (Figure 4.1). This representation is particularly favorable for the SARS-CoV-2 phylogeny in which the mutations are relatively rare and often shared across several samples. This approach has parallels to efficient tree-based representations used recently in population genetics (Ralph, Thornton, and Kelleher 2020; Kelleher et al. 2018). For our SARS-CoV-2 reference phylogeny, UShER's mutation-annotated tree uses only 3.4MB of memory (that fits easily in a last-level cache (Hennessy and Patterson 2017)) to encode virtually the same information as the full MSA which requires 1.14GB (>300x improvement).

UShER can generate a mutation-annotated tree for our reference tree with 38,342 leaves and 15,129 variant positions in just 2 minutes 24 seconds using four threads. This data structure is then stored as a pre-processed protocol buffer (https://developers.google.com/protocol-buffers), which is a customizable binary file format that can be rapidly loaded (~150 milliseconds) during sample placement and data visualization (Figure 4.1) and obviates the need to recompute the assignments for each execution.

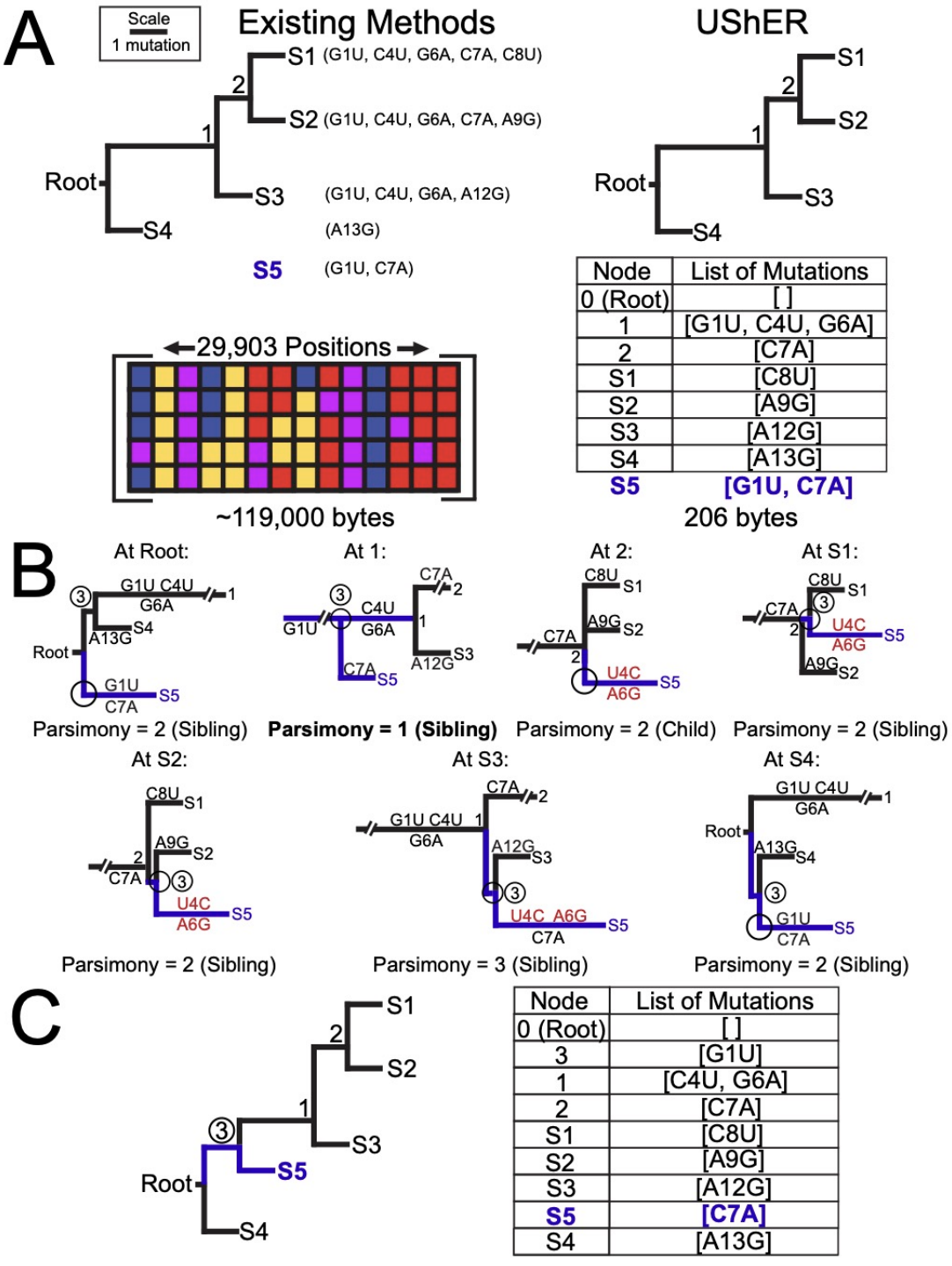


Figure 4.1: Overview of UShER's placement algorithm and data object. *(A) Prior methods rely on a full multiple sequence alignment (MSA) to inform phylogenetic structure (left), while UShER uses a mutation-annotated tree (right). The MSA shown is color-coded to match the mutations present in the tree above (A is in red, C in yellow, G in purple, U in blue). (B) UShER evaluations of the parsimony score for placing the sample S5 (blue) at each*

possible position (see Methods) of our example phylogeny (shown in panel A). We consider the branch leading to a given node to be the parent branch. The branches that need to be modified or added to the phylogeny to accommodate S5 are shown in blue; back mutations (if present) are colored in red and new nodes are circled. For example, if S5 is placed at S1, new node 3 has children S1 and S5, and two back-mutations (U4C and A6G) occur at the branch leading to S5, giving this placement a parsimony score of 2. Placing S5 at node 1 is optimal by parsimony. (C) The final tree with S5 added, in which an additional internal node 3 is added to support S5 (left), and the mutation annotations for the final tree with S5 colored in blue (right). Note that the memory efficiency of the mutation-annotated tree can vary depending on the dataset. In all panels, the length of each branch is proportional to the number of mutations that occurred on that branch. Zero-length branches, which are not associated with any mutations (e.g. those leading to node 3 in "At Root", "At S1", "At S2", "At S3", and "At S4" in panel B) are shown as very short branches for visibility.

4.4: UShER quickly and accurately places samples by maximum parsimony.

UShER uses this mutation-annotated tree to rapidly place newly acquired samples onto the tree of SARS-CoV-2 variation. More specifically, UShER uses a maximum-parsimony approach where it searches the entire reference tree (Figure 4.1, Methods) for a placement that requires the fewest additional mutations to accommodate the added sample (*i.e.*, the maximum-parsimony placement of a sample). UShER breaks ties based on the number of descendant leaves at the placement nodes when multiple placements are parsimony-optimal (Methods). When a pre-processed mutation-annotated tree is already available, this procedure takes approximately 0.5 seconds to place a single sample onto the SARS-CoV-2 reference tree (Table 5.1) and is even more efficient when placing larger sets of samples since the time to load the mutation-annotated tree gets amortized. For example, it only takes ~18 seconds to place 1000 samples onto our reference tree using 16 threads. This means that our implementation is fast enough to facilitate real-time placement of SARS-CoV-2 sequences and sufficiently memory efficient (Table 5.1) that everything we present could be run on a basic laptop, which should facilitate widespread adoption of this approach.

To evaluate the accuracy of UShER's maximum parsimony-based placement algorithm when the viral evolutionary history is known, we generated a SARS-CoV-2 simulated dataset using a fixed tree that we supplied (see Methods). UShER places samples

with the correct sister node in 97.2% of cases. For samples with just one parsimony-optimal placement, UShER achieves 98.5% accurate local placements. When incorrect, UShER's placements tend to be quite close to the correct node on the SARS-CoV-2 global phylogeny – separated by just 1.1 edges from the correct position on the tree, on average (Figure 4.2, Methods). We therefore conclude that UShER is capable of accurately placing new samples onto a fixed SARS-CoV-2 global reference phylogeny in practice and could indeed facilitate the ongoing genomic contact tracing efforts. Although UShER works well for SARS-CoV-2, it will not be as accurate for phylogenetic analyses in which maximum parsimony algorithms are known to perform poorly (*e.g.,* cases of long branch attraction (Felsenstein 1978)).

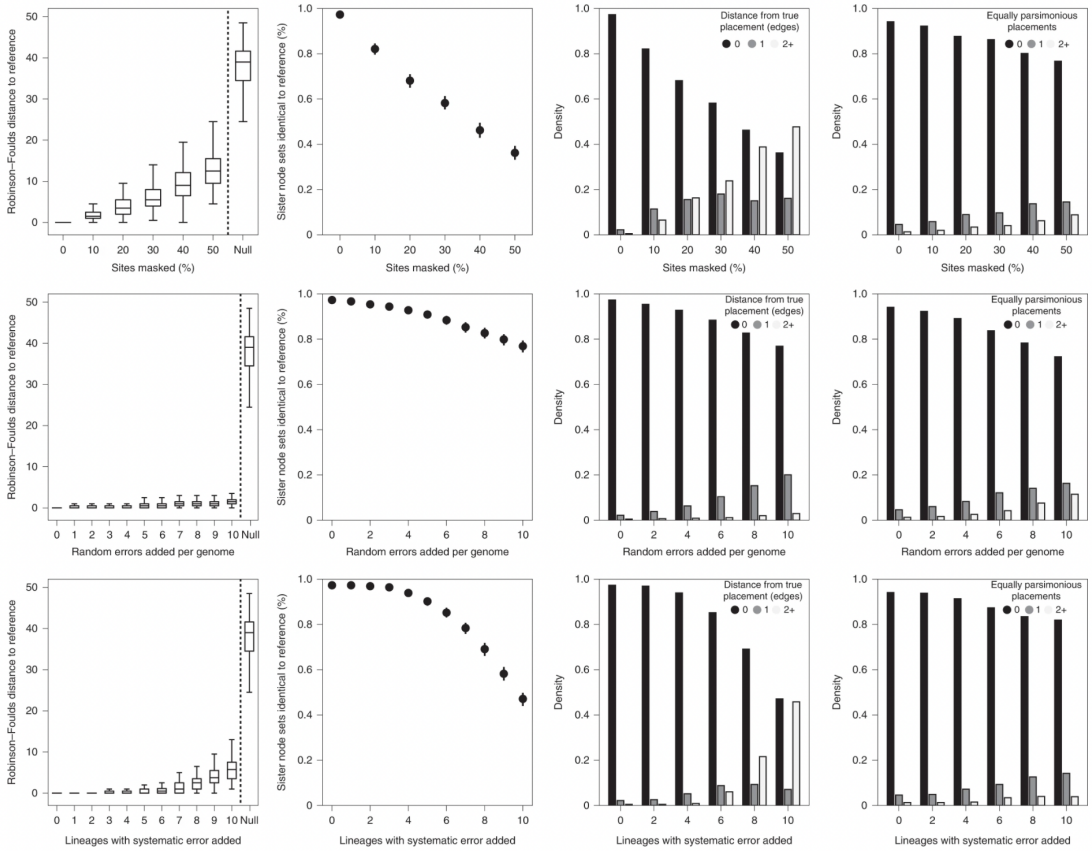


Figure 4.2: The maximum parsimony algorithm used by UShER is robust to moderate rates of missing data and simulated errors in SARS-CoV-2 genomes. *Top: We independently masked sites at 10, 20...50 percent of sites for each of 10 simulated genomes to be added to the phylogeny and computed the Robinson-Foulds distance (Robinson and Foulds 1981), the average number of lineages added that had identical sister node sets to those in the simulated reference tree, the distance from true placement for each lineage added (Methods) and the number of equally parsimonious sites per placement for each lineage added. Middle: We introduced random nucleotide substitutions to the genomes of the 10 lineages added to the tree by UShER at a rate of 1, 2…10 sites per genome, drawn independently, and computed the same measures of coherence to the reference tree, with the error bars representing the 95% confidence intervals. Bottom: We introduced one systematic error to 1, 2…10 of the genomes added to the tree by UShER and computed the same metrics as above. For each experiment, the distance from true placement was strongly correlated with the amount of missing data ($P < 3.34 \times 10^{-112}$ for all experiments). For each panel depicting Robinson-Foulds scores, the distribution of scores across 100 replicates where 10 lineages were added randomly to the phylogeny is shown to the far right for a null model comparison and is labeled 'Null'. In the error bar panels (second from the left), the data points are centered on the mean of the data and extend to the bounds of the 95% confidence interval, calculated by 1,000 iterations of bootstrapping.*

4.5: UShER is similarly robust to erroneous and missing data compared to maximum-likelihood counterparts.

Given the low mutation rate and therefore low phylogenetic signal in SARS-CoV-2 viral genomes, missing data has a large impact on phylogenetic placement, as expected (Figure 4.2). When we randomly masked between 0 and 50% of positions in samples to be placed by UShER, all measures of placement accuracy were negatively impacted. With 50% of all sites masked, we find that only 41.9% of samples are assigned identical sister nodes as their true placement on the reference tree. However, the mean distance between UShER and correct placements on the tree remained relatively small – just 1.61 edges– and 81.0% of lineages have sister node sets in the UShER tree that are a subset of the sister nodes in the reference tree, or vice versa (see Methods).

High rates of missing data have a slightly larger effect on the precision of UShER's placements than for maximum-likelihood tree inference methods when constructing a complete subtree *de novo* (Figure 4.3). When using Robinson-Foulds distance to measure congruence with the correct tree, we find that when no sites were masked, the average distance values from the correct tree for the trees obtained from the three methods are within

12.7% of each other and 12.9-13.3 times lower than a null model obtained from random tree construction (Figure 4.3). With no missing data, UShER produces the most congruent tree (*i.e.* having the lowest RF distance) to the correct tree, on average. The distance values increased by up to 11.1% with only 2.5% missing data and up to 76% with 10% missing data, with UShER being slightly more adversely affected by missing data than the other methods. Based on these observations, we recommend that the reference tree should ideally be maintained using only genomes with nearly complete sequences regardless of the tree inference method (*e.g.,* by filtering data obtained from the GISAID database using “complete” and “high coverage” tags).

Two types of errors in SARS-CoV-2 consensus sequences also affect the accuracy of sample placements. First, stochastic errors are likely present in many available SARS-CoV-2 sequences (Morel et al. 2020). When we simulated independent errors, we found the effects on UShER’s accuracy are modest (Figure 4.2). With 10 errors on average, the placement is approximately 20% less likely to select the correct sister node, and other distance metrics are similarly impacted (Figure 4.2). Our results indicate that especially low quality samples should be rigorously identified and excluded from analyses using UShER. Additionally, poor quality samples can be easily flagged because they will tend to appear as unrealistically long terminal branches in UShER’s placements. UShER reports all newly added samples with a parsimony score greater than 3 along with a list of parsimony-increasing sites.

Second, systematic error, where the same apparent variant is introduced into many sequences, are present in some SARS-CoV-2 sequences and have the potential to affect phylogenetic inference because they appear as inherited mutations (Turakhia, Thornlow, Gozashti, et al. 2020; N. De Maio et al. 2020). Whereas UShER appears to be robust to a single systematic error present in fewer than five samples (Figure 4.2), a single systematic error present in all 10 samples had a similar overall effect on placement accuracy as 50% missing data in error-free sequences. Consistent with our previous work (Turakhia, Thornlow, Gozashti, et al. 2020; N. De Maio et al. 2020), addition of two perfectly correlated systematic

errors can drastically affect UShER performance. Systematic errors should be rigorously identified and removed before sample placements are performed. We refer readers to methods that we developed previously to detect and eliminate such errors (Turakhia, Thornlow, Gozashti, et al. 2020; N. De Maio et al. 2020) and the UShER package includes a tool to remove known problematic positions when preparing input data.

We emphasize that sequencing errors pose similar challenges for other phylogenetic inference tools (Figure 4.3) and our analysis is meant to serve as a guideline to the user rather than highlight the limitations of UShER.

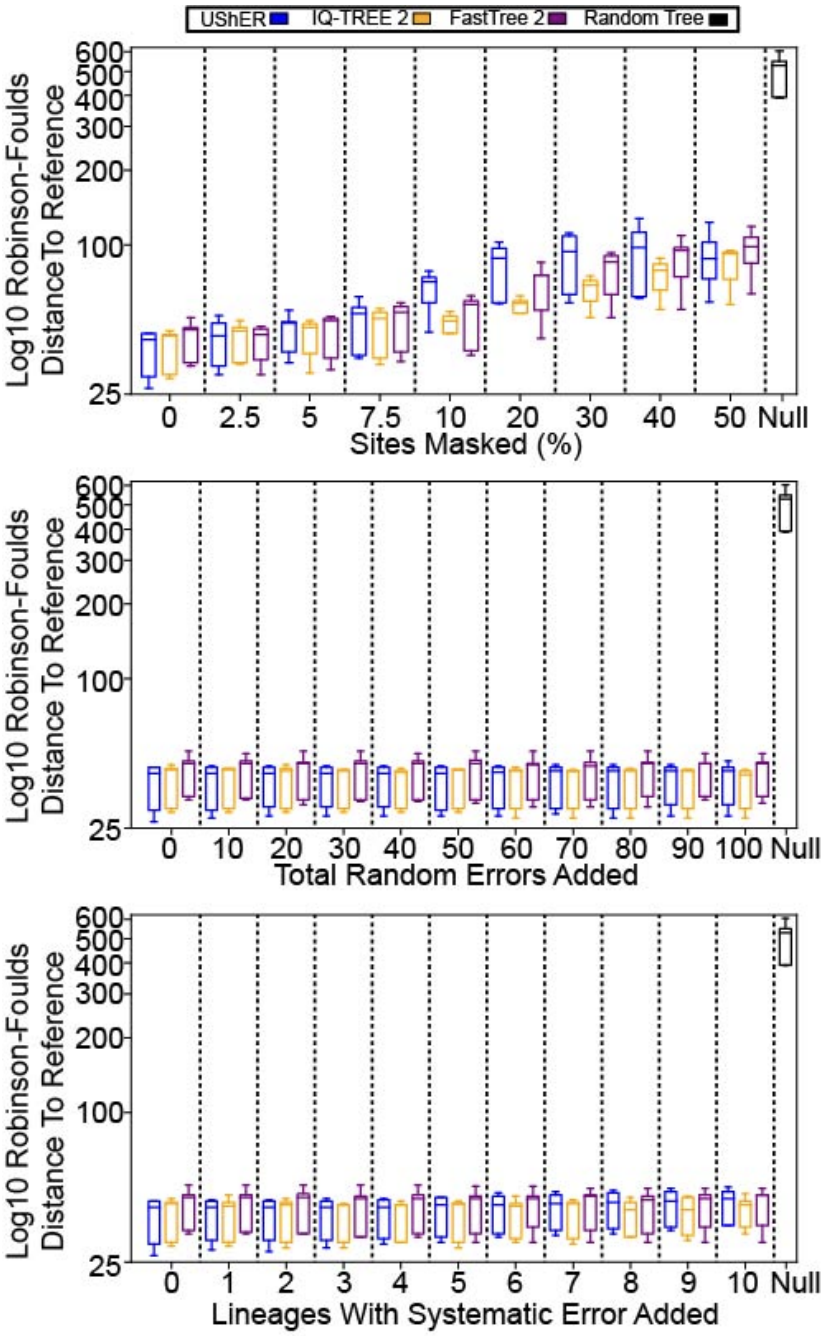


Figure 4.3: UShER is similarly robust to maksed sites and nucleotide errors as IQTREE-2 and FastTree2. *We pruned 5 independent clades of roughly 1,000 lineages each and applied the same methods as in Figure 4.2, masking 2.5, 5, 7.5, 10, 20...50 percent of sites (top, note that the X-axis does not use a linear scale), adding 10, 20...100 independently drawn random nucleotide substitutions across the lineages to be placed (center), and adding one error to 1, 2...10 of the genomes of interest (bottom). We then used UShER (blue), IQ-TREE 2 (Minh et al. 2020) (orange), and FastTree 2 (Price, Dehal, and Arkin 2010)*

(purple) to reconstruct these clades. We determined the Robinson-Foulds distance of each to the original clade using TreeCmp (Bogdanowicz, Giaro, and Wróbel 2012), as well as the distance of randomly constructed trees to the far right (black, labeled 'Null') as a null model comparison.

4.6: UShER is congruous with standard methods for SARS-CoV-2 data

To evaluate the performance of our approach under realistic conditions with genuine SARS-CoV-2 data, we used UShER to place real samples onto a global reference phylogeny. Because the phylogeny was necessarily inferred from real data (see Methods), this approach measures the consistency of placement between more typical tree-building approaches and UShER placement algorithm rather than placement accuracy *per se*. To evaluate the consistency, we randomly pruned and replaced 100 sets of 10 samples each using the reference tree (see Methods). We found that UShER placed each with an identical sister node as in the reference tree in 90.0% of cases (Figure 4.4). Additionally, the placements tend to be quite close to correct and the mean number of edges between the reference position and UShER's placement is just 0.159 and the mean Robinson-Foulds distance for trees with 10 samples added is 1.27 (Figure 4.4A). When we mimicked a plausible use case by removing larger sets of related sequences, we found that UShER is also able to accurately reconstruct larger subtrees for the added samples (Figure 4.4D-G). Collectively, our metrics are not far from those we obtained when analyzing the simulated datasets, and indicate that missing data, errors and other features of real sequences occasionally impact UShER's placements.

We found that samples causing inconsistent placements between the reference tree and UShER were mostly challenging cases. In particular, six of the 1000 sequences that we attempted to place using UShER have large numbers of equally parsimonious placements (5-65) and were placed inconsistently relative to the reference tree. Each of these consensus sequences has a large number of ambiguous nucleotide positions (8-15) that overlap many phylogenetically informative sites in the reference tree. This may suggest a mixture of two genetically divergent samples–either a true mixed infection or laboratory induced. Regardless

of the source, we believe future versions of the reference tree should rigorously filter sequences containing ambiguities at phylogenetically informative positions.

Additionally increasing genetic distance and sequencing errors are expected to affect placement accuracy. We found that samples are more likely to be placed inconsistently when the parsimony score is higher ($P = 2.98\times10^{-5}$, one-tailed Mann-Whitney U Test). Incorrectly placed samples also had significantly more equally parsimonious placements ($P = 1.3\times10^{-21}$, one-tailed Mann-Whitney U Test). In fact, 15% of real samples have more than one equally parsimonious placement on the reference phylogeny and many distinct nodes are identical in the reference tree. However, if we restrict the analysis to samples with only a single most parsimonious placement, we find that 97% of UShER's placements are consistent with the maximum-likelihood reference tree. We suggest that the placements of samples that are unusually genetically distant or that have many equally parsimonious placements on a reference tree should be regarded with caution. Both statistics are reported by UShER.

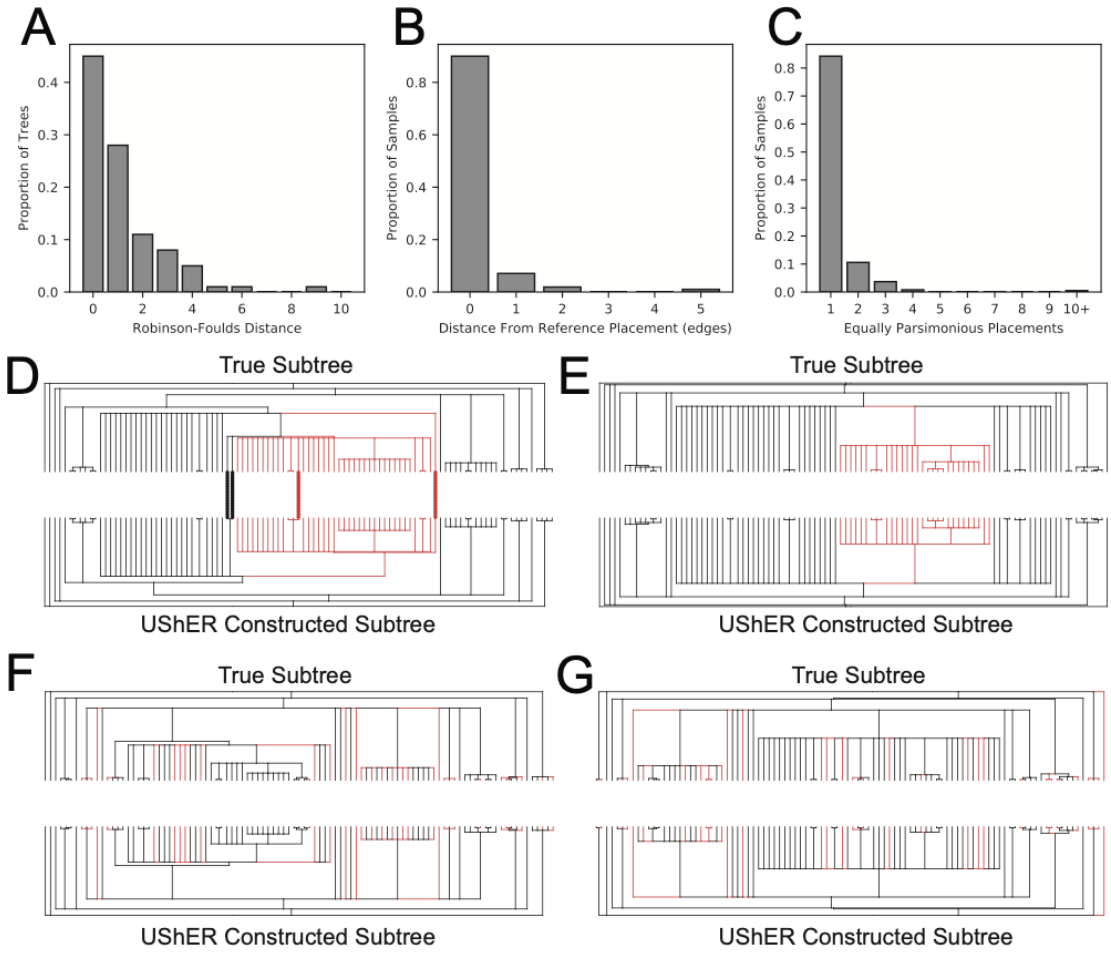


Figure 4.4: UShER is accurate using real data: *The Robinson-Foulds distance between 100 reference and UShER-generated trees produced by removing and re-adding 10 samples in each (A), the distance from reference placement for each of 1,000 placed samples (B), and the number of equally parsimonious placements for each of the 1,000 placed samples (C). Comparisons of subsets of the global phylogeny released on 11/7 with reconstruction of this phylogeny using UShER (D-G). In each case, we pruned lineages colored in red from the phylogeny and added them back using UShER. UShER accurately places subtrees lineages collected in the Western United States in March/April (D) and in Europe in March (E), as well as more distantly related lineages whose times and places of collection differed more widely (F, G). Differences in tree topology (D) are highlighted in bold. Other topologies (E-G) are identical. All trees in this figure are ultrametric and branch lengths are arbitrary.*

4.7: UShER can maintain a global phylogeny.

We propose that UShER could form the basis for real-time phylogenetics platforms in periodically updating the reference tree itself or be used in conjunction with maximum-likelihood updates. To investigate this, we used UShER to add all of the 9437

additional sequences in the 31/7/2020 release of the global tree to our 11/7/2020 reference tree. We also extensively optimized both trees using a maximum likelihood approach in FastTree 2 (Price, Dehal, and Arkin 2010). The Robinson-Foulds distance between all trees is similar, suggesting that the UShER updated topologies are close to *de novo* phylogenies. Additionally, the optimized version of the phylogeny produced by UShER resulted in a substantially increased likelihood over the 31/7/2020 tree inferred *de novo* with similarly extensive optimization. We obtained the highest likelihood topology from a heavily optimized 11/7/2020 tree, sample addition with UShER, and then another round of tree optimization (Table S4). This indicates that UShER, combined with additional rounds of optimization, does not result in unrecoverable local-minima but rather may help avoid them. In combination with periodic maximum-likelihood updates to the global phylogeny, UShER can offer an appealing combination of real-time phylogenetic methods and model-based practices. This combination can be used to maintain an updated phylogeny for the SARS-CoV-2 pandemic, which is not remotely feasible by any other existing software package.

The SARS-CoV-2 pandemic has been accompanied by unprecedented levels of pathogen genomic sequencing which has truly empowered near real-time monitoring of viral transmission and evolution. This seemingly endless flood of genome sequence data has also pushed phylogenetic analysis frameworks over the edge of their capabilities, requiring new approaches to rapidly incorporate and contextualize newly sequenced viral genomes. UShER is an extremely efficient software package inspired by the ongoing evolution of the virus itself, that provides a method to immediately incorporate viral genome isolates into a global phylogenetic tree. Compared to its closest counterpart, UShER is over 3,000x faster and orders of magnitude more memory efficient. It is currently the only tool with actual real-time capabilities. Although several challenges still remain for routinely deploying pathogen surveillance methodologies, UShER removes a key barrier by significantly decreasing the turnaround time from sample to analysis and empowering real-time genomic contact tracing efforts during the SARS-CoV-2 pandemic and beyond.

4.8: Methods

The code used for the statistical analyses and to produce the figures is available at https://github.com/bpt26/USHER_ANALYSES. We measured UShER's accuracy in placing samples onto a reference phylogeny using simulated (described above) and real data. For simulated data, both reference phylogeny and sequences were simulated; for real data, we used the global phylogeny dated 11 July 2020 (https://github.com/roblanf/sarscov2phylo) as reference and its corresponding sequences were obtained from GISAID (Yuelong Shu and McCauley 2017). In each case, we first randomly pruned out ten samples from the global phylogeny, which was then used as the input phylogeny while adding back the pruned samples using UShER. UShER's accuracy in placing back the samples was computed using the average values of three different statistics (described below) over 100 such replicates.

We initially used TreeCmp (Bogdanowicz, Giaro, and Wróbel 2012) to compute the Robinson-Foulds distance between the reference phylogeny and the tree constructed by samples using UShER. Separately, we recorded whether the sister clade for each placed sample was identical to the true sister clade (that is, the sister clade in the reference phylogeny). Finally, we computed the distance between the UShER placement and the correct placement in terms of the minimum number of edges separating them as described below.

Ordinarily, the distance between two nodes in a tree can be computed using their lowest common ancestor44 by taking the sum of the number of edges to each node from the lowest common ancestor. To determine the distance between the node placement in two trees (reference phylogeny and the one resulting from UShER placement), we developed a utility that reports all descendant lineages from an n-th generation ancestor of any given node in a tree, with n provided as input (that is, when n = 1, it reports unpruned lineages in the sister clade). For each pruned lineage, we found the descendants varying the number of generations, N1 and N2, in global and UShER phylogenies, respectively and reported the

distance between the UShER placement and the correct placement as the smallest (N1 + N2 − 2), which resulted in the same set of descendant lineages in both phylogenies. Note that the second statistic records cases for which the sister clades in the two trees are identical, which would always have 0 distance in our third statistic (N1 = N2 = 1).

We also measured UShER's accuracy in a more realistic scenario of placing closely related samples that form their own subtree. In this case, during pruning, we required that the pruned samples together formed a subtree (that is, not a trivial polytomy) in the reference phylogeny.

To evaluate the accuracy and robustness to error of UShER compared to IQ-TREE 2 and FastTree 2, we identified 5 clades of approximately 1,000 lineages and reconstructed each from scratch using each of the 3 methods, and repeated these experiments after randomly masking between 2.5 and 50% of sites to 'N', adding 10, 20 … 100 independently drawn random single-nucleotide errors across the lineages to be placed and adding identical single-nucleotide errors to 1–10 of the genomes to be placed. We measured the accuracy of these placements by calculating the Robinson-Foulds distance using TreeCmp.

We compared UShER to four other lineage placement algorithms: IQ-TREE multicore v.2.1.1; EPA-ng v.0.3.8; PAGAN2 v.1.54; and TreeBEsT v.1.9.2 (Minh et al. 2020; Barbera et al. 2019; Löytynoja, Vilella, and Goldman 2012; Ponting 2007). We initially attempted to add 1,000 lineages to our simulated phylogeny; however, except for UShER, which required 18 seconds to finish using 64 threads, none of the placement programs finished within 24 h. Due to time and memory constraints, we instead added only 1 lineage to the tree in 20 replicates, recording the average and time range and peak memory usage across these 20 replicates in Table 5.1. We installed and ran each program on a server with 160 processors (Intel Xeon CPU E7-8870 v.4, 2.10 gigahertz), each with 20 CPU cores.

For the 11 and 31 July 2020 reference trees, we created optimized versions of each using FastTree 2 (ref. 38) using ten iterations of the command '*fasttree -nt -nni 0 -spr 1 -sprlength 1000 -nosupport -intree <initial tree> global.fa > <new tree>*', replacing the initial

tree with the new tree from the previous iteration each time, followed by the command '*fasttree -nt -nni 0 -spr 1 -sprlength 1000 -nosupport -gamma -intree <initial tree> global.fa > <new tree>*'. In these commands, -nt indicates that the input is a nucleotide alignment, -nni 0 indicates that no minimum-evolution nearest neighbor interchanges are done, -spr 1 -sprlength 1000 indicates 1 round of SPR with a maximum distance of 1,000, meaning that a single SPR move could move a subtree to any new branch of the global tree. The -nosupport flag indicates that support values are not output and the -gamma flag indicates that the lengths are rescaled to optimize the Gamma likelihood. Because FastTree 2 requires binary trees, we randomly resolved all polytomies before optimization. We also generated two other trees using UShER, by taking the original and optimized 11 July trees, pruning out all lineages in the 11 July 2020 tree not present in the 31 July 2020 tree and using UShER to add in all lineages present in the 31 July 2020 that were not present in the 11 July tree. We then optimized these two new trees using ten iterations of FastTree 2, followed by another round of optimization using the -gamma flag as described above.

4.9: matUtils: tools for analyzing comprehensive mutation-annotated trees.

Perhaps the greatest strength of UShER is its efficiency, in both time and data usage. By far the most important component of this efficiency is the MAT data object, stored as a protobuf (https://developers.google.com/protocol-buffers). However, if UShER is to be used extensively, a full companion phylogenetics package is required to handle these data objects. To accomplish this, we developed matUtils – a toolkit for rapidly querying, interpreting and manipulating the MATs included in our database or constructed with UShER (Turakhia et al. 2021). Using matUtils, common operations in genomic surveillance and contact tracing efforts, including annotating a MAT with new clades, extracting specific subtrees, or converting the MAT to standard Newick or VCF format, can be performed in a matter of seconds to minutes even on a laptop. We also provide a web interface for matUtils through the UCSC SARS-CoV-2 Genome Browser (Fernandes et al. 2020). Together, our

SARS-CoV-2 database and matUtils toolkit can simultaneously democratize and accelerate pandemic-related research.

The matUtils toolkit is designed to scale efficiently to SARS-CoV-2 phylogenies containing millions of samples. Using matUtils, common pandemic-relevant operations described in the earlier section can be performed in the order of seconds to minutes with the current scale of SARS-CoV-2 data (Table 4.2). For example, it takes only 5 seconds to summarize the information contained in our June 9, 2021 SARS-CoV-2 MAT of 834,521 samples and only 15 seconds to extract the mutation paths from the root to every sample in the MAT. Since matUtils is primarily designed to work with the newly-proposed and information-rich MAT format, it does not have direct counterparts in other bioinformatic software packages currently, but its efficiency is similar or better than state-of-the-art tools that offer comparable functionality (Table 4.2). For example, matUtils is able to resolve polytomies in a 834,521 sample tree in 9 seconds, a task which takes over 37 minutes using ape (Paradis and Schliep 2019) (Table 4.2). matUtils is also very memory-efficient, requiring less than 1.4 GB of main memory for most tasks, making it possible to run even on laptop devices.

matUtils is effectively a superset of phylogenetics functions, and has expanded well beyond what is possible in packages such as *tree_doctor* (Hubisz, Pollard, and Siepel 2011) or *ape* (Paradis and Schliep 2019). Therefore, we did not benchmark several of its functions, as there were no viable comparisons. Overall, matUtils performs admirably in both time and memory usage, especially when considering that matUtils reads in trees with mutations added -- effectively both a .vcf and a .nwk combined, and that matUtils is capable of reading and editing trees larger than one million samples with any number of polytomies.

Function	Software Package	Average Time (M:S)	Average Memory (kB)
Prune 50,000 Samples	matUtils	00:14.3	1,433,618
	tree_doctor	07:30.3	667,670
	bcftools	14:31.9	314,965
	newick_utils	01:17.3	837,756
	ape	00:51.2	706,961
Annotate Clade	matUtils	00:33.7	1,447,677
	tree_doctor	01:56.2	677,764
Extract Clade	matUtils	00:10.0	1,271,104
	bcftools	08:06.9	281,844
Resolve Polytomies	matUtils	00:09.4	1,224,652
	ape	37:30.6	735,688

Table 4.2: Benchmarking matUtils and other phylogenetics software packages. *The average time and memory usage across three replicates for listed functions on our annotated public tree dated June 9 2021, which contains 834,521 samples. The clades annotated and extracted had roughly 250,000 samples. For a more extensive comparison, see Tables S2-S9 in (McBroome et al. 2021).*

All performance benchmarking experiments were carried out on a Google Cloud Platform (GCP) instance n2d-standard-16 with 16 vCPUs (Intel Xeon CPU E7-8870 v.4, 2.10 GHz) with 64 GB of memory using our public SARS-CoV-2 MAT dated June 9, 2021. matUtils does not have direct counterparts for its ability to work with the mutation-annotated tree (MAT) format, but we compared the performance of matUtils with state-of-the-art tools that offer some comparable functionality on Newick or VCF formats. Specifically, we compared the most recent version of matUtils (version 0.3.1) to newick_utils version 1.6 (Junier and Zdobnov 2010), tree_doctor (from version 1.5 of the phast package; (Hubisz et al. 2011), ape version 5.5 (Paradis and Schliep 2019), and bcftools version 1.7 (Danecek et al. 2011). The input data used for each comparison can be found at https://github.com/bpt26/matutils_benchmarking/.

4.10: Applying phylogenetics tools to SARS-CoV-2 samples from the Santa Cruz community

4.10.1: Background

Rapid SARS-CoV-2 genome sequencing enables researchers to trace the virus' evolution as it spreads and adapts within human populations. Several important mutations have arisen within the SARS-CoV-2 population that are thought to increase the transmissibility of the virus (Bette Korber et al. 2020; Volz, Hill, et al. 2021; Volz, Mishra, et al. 2021) or to improve the virus' ability to evade host immune defenses (Hoffmann et al. 2021; Planas et al. 2021). The early detection of new variant lineages of SARS-CoV-2 that might have similar traits is essential to prioritizing public health responses, developing variant-specific diagnostics and vaccines, and beginning research into the possible immunological and general health implications of newly discovered variants.

Parallel evolution occurs when the same mutations arise in distantly related lineages at rates and may suggest positive selection recurrently favors the same alleles. Mutations undergoing parallel evolution that do not confer fitness advantages to the virus may be due to hypermutation, genetic drift, or founder effects (Chiara et al. 2020). However, several previously described recurrent mutations may confer viral fitness advantages, via *e.g.* increased transmissibility (Volz, Hill, et al. 2021; Davies et al. 2021) or ability to evade host immune response (McCarthy et al. 2021; Starr et al. 2021; Thomson et al. 2021). Such mutations undergoing parallel evolution are most likely to be adaptive (Hodcroft, Domman, et al. 2021; Wu 2020). Increased genomic surveillance and detection of such parallel evolution events may therefore be useful for identifying previously undescribed Variants of Interest (Martin et al. 2021; Hodcroft, Domman, et al. 2021).

In this section, I describe a novel variant lineage, B.1.623, discovered by our research team, that shares several mutations with known SARS-CoV-2 Variants of Concern, including Spike protein mutations S494P, N501Y and P681H (Thornlow et al. 2021). Notably,

each of these mutations within B.1.623 has apparently occurred independently to any previously described variant of concern, including B.1.1.7. While no longer in circulation in Santa Cruz County, B.1.623 is one of likely hundreds of short-lived variants from neighborhoods across the country that serves to highlight an example of widespread parallel evolution among SARS-CoV-2 lineages, and to demonstrate the need for increased real-time monitoring for the public health and safety of communities everywhere.

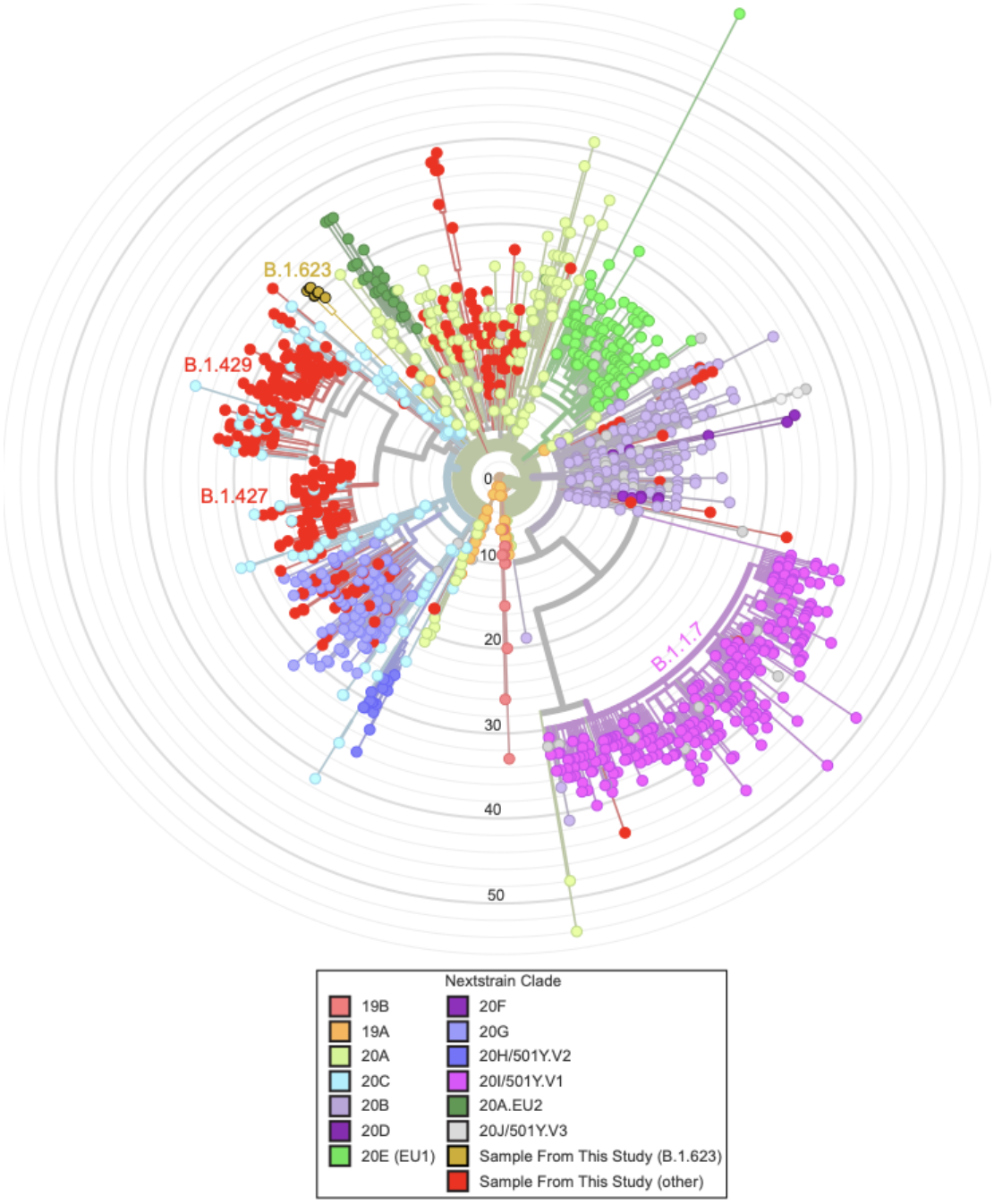


Figure 4.5: The phylogenetic distribution of 339 samples obtained from SARS-CoV-2 sequencing in Santa Cruz County plus 1000 samples from elsewhere. *The tree is produced via hgPhyloPace (https://genome.ucsc.edu/cgi-bin/hgPhyloPlace). To produce it, we added the 339 genomes from the Santa Cruz County samples to a global phylogeny of more than 1 million SARS CoV-2 genomes and then pruned back to retain only the Santa Cruz genomes plus 1,000 others selected at random. We visualized the tree using the Auspice.us platform. The 339 samples from Santa Cruz County are colored in red, with the eight samples representing B.1.623 highlighted in gold, and the remaining 1,000 samples*

colored by Nextstrain clade. Note that clade sizes reflect both prevalence and local sampling effort and we have not attempted to correct for the effect of either.

4.10.2: Lineage B.1.623 shares mutations with Variants of Concern.

We obtained a consensus sequence from 339 samples from Santa Cruz County. The majority of these samples are high quality, and 88% have a single maximally parsimonious placement on the global SARS-CoV-2 phylogeny that we maintain (Turakhia, Thornlow, Hinrichs, et al. 2020). This suggests phylogenetic inference using these samples is reliable (Turakhia, Thornlow, Hinrichs, et al. 2020). Approximately 58% of these 339 sequences are associated with lineages B.1.427 and B.1.429, first identified in California (Deng et al. 2021). We also detected two samples in our dataset of lineage B.1.1.7, first detected in the UK (Volz, Mishra, et al. 2021) and increasing in southern California, USA (Washington et al. 2021). We did not detect any other CDC-designated Variants of Concern or Variants of Interest (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html). The distribution of lineages in Santa Cruz County is similar to reports from the San Francisco Bay Area approximately 100km to the north conducted at approximately the same time (Peng et al. 2021) (Figure 4.5).

Eight samples collected between mid February and early March represent a novel lineage, named B.1.623 by Pangolin (Rambaut et al. 2020), and discovered by our research team. These samples contain several mutations shared with B.1.1.7 and other VOCs. The eight samples share six non-synonymous mutations within the S or Spike protein (S494P, N501Y, D614G, P681H, K854N, and E1111K) relative to the Wuhan-Hu-1 SARS-CoV-2 reference sequence (RefSeq NC_045512.2). Among these, N501Y is thought to be important for viral replication because it enables the virus to bind ACE2 and enter host cells more efficiently (Gu et al. 2020; Starr et al. 2020). S494P is also located within the ACE2 receptor binding domain and experimental evidence suggests that mutations at this position decrease antibody binding affinity (Starr et al. 2020). Similarly, P681H is located within the Spike protein furin cleavage site which is thought to be a hotspot of viral adaptive evolution (*e.g.,*

(Peng et al. 2021; Zuckerman et al. 2021)). D614G became globally dominant in 2020 possibly due to higher viral loads (Volz, Hill, et al. 2021). The effect of the other two Spike mutations is unknown.

In addition to the Spike mutations, B.1.623 includes N:M234I (G28975A), which also appears in Variants of Interest B.1.526 and P.2 (G28975T). The three nucleotide mutations that cause N:M234I (G28975A, G28975C, G28975T) have all been observed at the roots of several Pango lineages and the frequency of N:M234I in 480,704 samples available from GISAID as of 2 April 2021 with collection dates 2021-01-01 to 2021-03-31 is 7.0%. N:M234I has been predicted to be stabilizing for the protein structure (Jacob et al. 2020). More generally, because each of these mutations appear to have occurred independently from other VOCs in B.1.623, these substitutions reveal significant evolutionary parallelism between B.1.623 and known VOCs.

The parallel mutations that we observed within B.1.623 and known VOCs likely reflect strong adaptation rather than hypermutability of specific positions (Figure 4.6). Although there is evidence that supports highly variable mutation rates across sites within the SARS-CoV-2 genome (Goswami et al. 2021; Pachetti et al. 2020; Badua, Baldo, and Medina 2021), none of the amino acid substitutions that we identified here are particularly recurrent in SARS-CoV-2 evolution (Nicola De Maio et al. 2021). Given the strong experimental evidence for immunity evasion associated with some of these specific positions, positive selection appears to be a more likely explanation for the strong parallelism with previously described VOCs, themselves thought to be evolving in parallel due to positive selection (Martin et al. 2021).

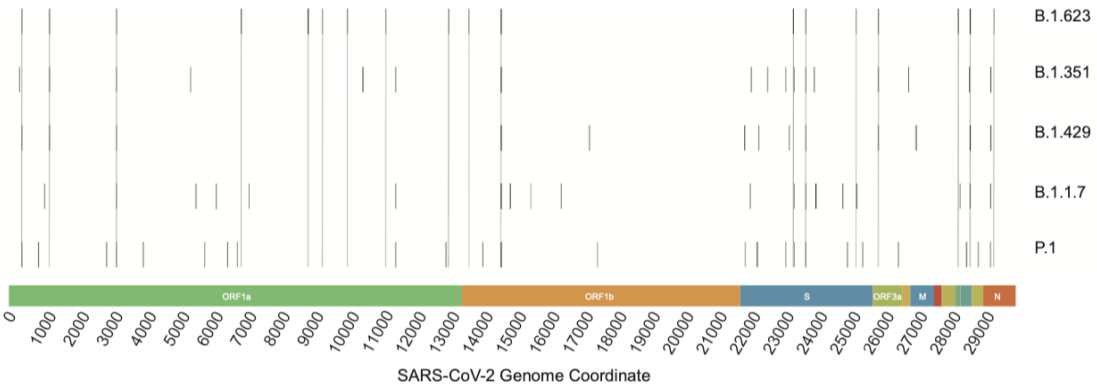


Figure 4.6: Sequence variation across each VOC and lineage B.1.623. *Dashed lines indicate mutations relative to the reference sequence that are shared by B.1.623 and at least one VOC. No stretch of variation appears to clearly match any single VOC, as we would expect if recent recombination generated the shared mutations. In particular, the spike region is quite distinct except for the shared mutations with known VOCs such as N501Y and P681H as noted in the main text.*

4.10.3: Lineage B.1.623 has a novel 35-nt deletion in ORF8

The B.1.623 lineage contains a 35-nt deletion that induces a frameshift and early stop codon in ORF8 (bases 27922-27956 of RefSeq NC_045512.2) that is reminiscent of B.1.1.7, which also contains a nonsense mutation that causes a premature stop codon in ORF8. However, the functional significance of this mutation is not known. This parallel evolution with the B.1.1.7 lineage suggests that inactivation of ORF8 may be favorable to the virus, possibly in combination with shared amino acid substitutions within the Spike protein (Zinzula 2021).

Variable genome consensus generation methods and difficulties in accurately genotyping large indels make it challenging to establish the evolutionary pattern of this deletion with certainty. There appears to be dramatic variation in the presence/absence of this deletion across closely related sequences (Figure 4.7). However, many related consensus sequences have this section of the genome replaced with Ns, and others contain nucleotide sequences that suggest that the deletion may have been replaced with reference alleles (see below). We believe that it is unlikely that this large, specific deletion evolved more than once in this lineage. Conversely, a reversion to the reference, via reinsertion of identical 35bp

nucleotide sequence is unlikely. Instead, we suggest that this highly unusual evolutionary pattern primarily reflects apparent differences in consensus genome assembly and submission methods where some tools may insert reference bases and others call deletions as missing data. Of course, confident phylogenetic inference with SARS-CoV-2 is difficult (Morel et al. 2020; Turakhia, Thornlow, Gozashti, et al. 2020), so it is possible that errors in the inferred phylogeny contribute to this apparent improbable evolutionary pattern at this deletion. Comparison to raw sequence data will be necessary to confirm or refute this. However the pattern suggests that the deletion is a shared feature among most genomes in this lineage and that its apparent sporadic absence in some genomes is an artefact of the way virus genomes are currently assembled and submitted.


Nucleotides at 27921-27956
Deletion at 27921-27956
Ns at 27921-27956
≥ 40 Ns in ORF8
Samples Collected in Santa Cruz
E1111K
N:M234I
K854N
P681H
S494P
N501Y
E484K

Figure 4.7: Phylogenetic distribution of 322 sequences in the B.1.623 lineage and 15 nearest neighbors. *The phylogeny was inferred based on whole-genome sequences and colored based on ORF8. Our samples and others containing the deletion are shown in blue and purple respectively. Samples with missing data (Ns) throughout the deletion are shown in red, and samples with nucleotide sequence are shown in black. To accommodate slight variation in consensus sequences, we refer to all samples with at least 30 "-" characters in the multiple sequence alignment as deletion and similarly all sequences with at least 30 "N" characters as missing data. Samples with missing data in ORF8 extending beyond the 35 nt locus are shown in gray. The eight samples collected in Santa Cruz County are shown in gold. Branch lengths are proportional to the number of mutations inferred to occur on that branch. All spike amino-acid altering mutations in B.1.x are shown except D614G which occurred in an ancestor that is not shown in this phylogeny. Note that two samples isolated in the UK also contain E484K as indicated.*

4.11: Conclusion

Pathogen genomic surveillance has the potential to identify novel lineages with important immunological and epidemiological consequences. After genome sequencing and assembly, UShER enables placement of samples on large phylogenies in less than a second per sample. matUtils enables extraction of just the nearest neighbors of the samples of interest, and exportation into files viewable with FigTree (Rambaut 2012) or the auspice.us platform (Huddleston et al. 2021). Our discovery of the B.1.623 lineage and full analysis within a week of receiving the sequencing data is an example of real-time phylogenetics at work. Genomic surveillance can now be done anywhere with these publicly available tools.

Since our analysis, B.1.623 has died out in Santa Cruz County. This is expected for the vast majority of novel variants, especially in areas where the vaccine has been effectively rolled out to the public. Nonetheless, given its genomic parallels to known VOCs, its characterization was briefly an important local health matter. Our work demonstrates that communities everywhere are now equipped with all of the bioinformatics tools necessary to characterize any persistent novel variants that might arise in the future, and are therefore more able to quell additional threats to public health.

Chapter 5: Pandemic-Scale Phylogenomics Reveals Elevated Recombination Rates in the SARS-CoV-2 Spike Region

5.1: Background

Recombination is a primary contributor of novel genetic variation in many prevalent pathogens, including *betacoronaviruses* (Forni, Cagliani, and Sironi 2020), the clade that includes SARS-CoV-2. By mixing genetic material from diverse genomes, recombination can produce novel combinations of mutations that have potentially important phenotypic effects (Didelot and Maiden 2010). For example, recombination is thought to have played an important role in the recent evolutionary histories of MERS (Dudas and Rambaut 2016) and SARS-CoV (Lau et al. 2015; E. C. Holmes and Rambaut 2004). Furthermore, a recombination event that transferred a portion of the Spike protein coding region into the ancestor of SARS-CoV-2 may have contributed to the emergence of the COVID-19 pandemic in human populations (X. Li et al., n.d.). Recombination is thought to have the potential to generate viruses with zoonotic potential in the future (X. Li et al., n.d.). Therefore, accurate and timely characterization of recombination is foundational for understanding the evolutionary biology and infectious potential of established and emerging pathogens in human, agricultural, and natural populations.

Now that substantial genetic diversity is present across SARS-CoV-2 populations (Nicola De Maio et al. 2021) and co-infection with different SARS-CoV-2 variants has been known to sometimes occur (Taghizadeh et al. 2021), recombination is expected to be an important source of new genetic variation during the pandemic. Whether or not there is a detectable signal for recombination events in the SARS-CoV-2 genomes has been fiercely debated since the early days of the pandemic (X. Li et al., n.d.). Nonetheless, several apparently genuine recombinant lineages have been identified using *ad hoc* approaches (Jackson et al. 2021) and semi-automated methods that cope with vast SARS-CoV-2

datasets by reducing the search space for possible pairs of recombinant ancestors (Jackson et al. 2021; VanInsberghe et al. 2021). Because of the importance of timely and accurate surveillance of viral genetic variation during the ongoing SARS-CoV-2 pandemic, new approaches for detecting and characterizing recombinant haplotypes are needed to evaluate new variant genome sequences as quickly as they become available. Such rapid turnaround is essential for driving an informed and coordinated public health response to novel SARS-CoV-2 variants.

The text of this dissertation includes reprints of the following previously published material: Turakhia, Y., Thornlow, B., Hinrichs, A. S., Mcbroome, J., Ayala, N., Ye, C., ... & Corbett-Detig, R. (2021). Pandemic-Scale phylogenomics reveals elevated recombination rates in the SARS-CoV-2 spike region. *bioRxiv*. The co-authors listed in this publication directed and supervised the research which forms the basis for the dissertation.

5.2: How RIPPLES works

We developed a novel method for detecting recombination in pandemic-scale phylogenies, Recombination Inference using Phylogenetic PLacEmentS (RIPPLES, Figure 5.1). Because recombination violates the central assumption of many phylogenetic methods, *i.e.,* that a single evolutionary history is shared across the genome, recombinant lineages arising from diverse genomes will often be found on “long branches” which result from accommodating the divergent evolutionary histories of the two parental haplotypes (Figure 5.1). RIPPLES exploits that signal by first identifying long branches on a comprehensive SARS-CoV-2 mutation-annotated tree (Turakhia et al. 2021; McBroome et al. 2021). RIPPLES then exhaustively breaks the potential recombinant sequence into distinct segments that are differentiated by mutations on the recombinant sequence and separated by up to two breakpoints. For each set of breakpoints, RIPPLES places each of its corresponding segments using maximum parsimony to find the two parental nodes – hereafter termed donor and acceptor – that result in the highest parsimony score

improvement relative to the original placement on the global phylogeny (Text S1). Our approach therefore leverages phylogenetic signals for each parental lineage as well as the spatial correlation or markers along the genome. We establish significance for the parsimony score improvement through a null model conditioned on the inferred site-specific rates of *de novo* mutation (Text S2-S3).

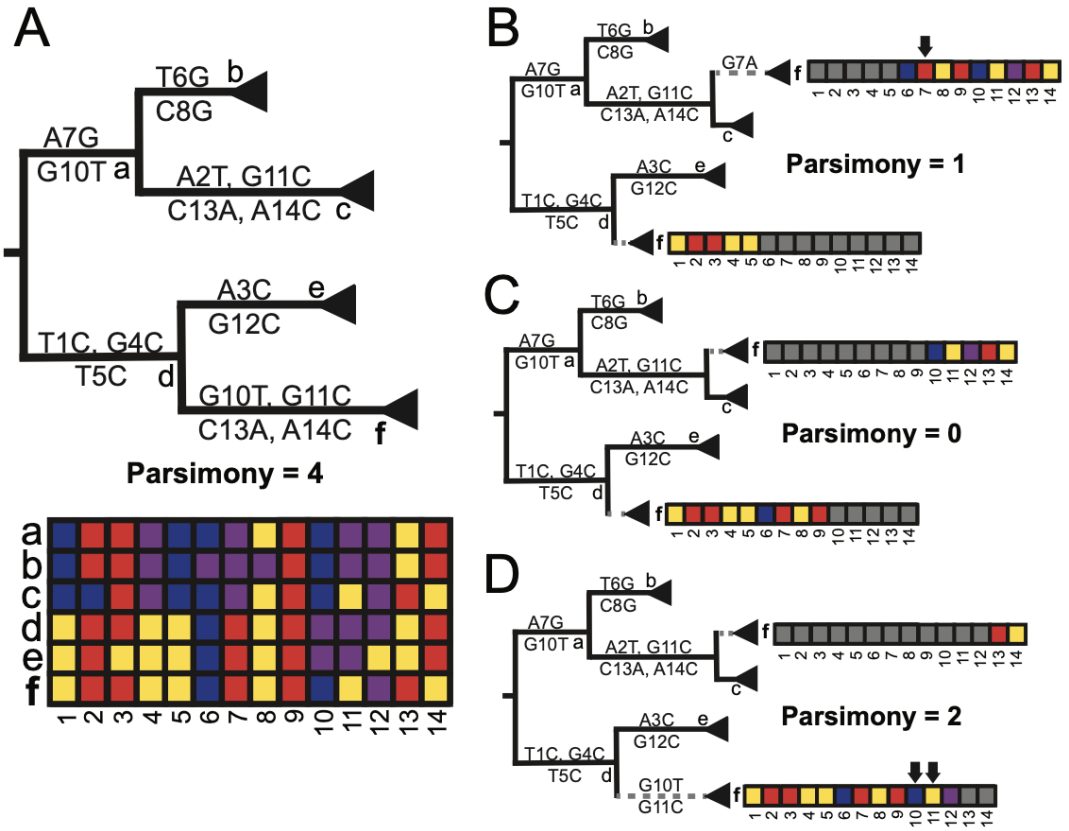


Figure 5.1: RIPPLES exhaustively searches for optimal parsimony improvements using partial interval placements. (A)*: A phylogeny with 6 internal nodes (labeled a-f), in which node f is the one being currently investigated as a putative recombinant. The initial parsimony score of node f is 4, according to the multiple sequence alignment below the phylogeny, which displays the variation among samples and internal nodes. Note that internal nodes may not have corresponding sequences in reality, but test for recombination using reconstructed ancestral genomes.* ***(B-D)****: Three partial placements given breakpoints are shown with their resulting parsimony scores. Arrows mark sites that increase the sum parsimony of the two partial placements of f. The optimal partial placement and breakpoint prediction for node f is in the center (C), with one breakpoint after site 9 and with partial placements both as a sibling of node c and as a descendant of node d.*

5.3: Applying RIPPLES to a 1.607M-sample phylogeny

Substantial testing via simulation indicates that RIPPLES is sensitive and can confidently identify recombinant lineages. On our tree containing over 1.6 million SARS-CoV-2 sequences, RIPPLES takes just 6.25 minutes of wall time using 4 CPU threads and 1.94 GB of RAM per tested node, on average (Text S4-S5). Nonetheless, recombination breakpoints close to the edges of the SARS-CoV-2 genome are challenging to identify with certainty (Figure 5.2), which makes RIPPLES weakly biased against identifying recombination events near the edges of the viral genome. As expected, when recombination occurs between genetically similar sequences, it is harder to detect it using RIPPLES (Figure 5.3). The low identifiability of recombination events among closely related lineages is a well-known challenge in population genetics (Stephens 1986). Nonetheless, RIPPLES detects simulated recombinants with 93% sensitivity (Table S1), and is able to detect each of the highly confident recombinant samples identified by Jackson et al. (Jackson et al. 2021) (Text S6). In contrast to previous methods for investigating recombination in the vast SARS-CoV-2 genomic datasets (Varabyou et al. 2021; Jackson et al. 2021; VanInsberghe et al. 2021), RIPPLES can search for recombination on the inferred internal nodes of the phylogeny and does not require that phylogenetically informative sites or the set of parental lineages be selected *a priori.* This allows RIPPLES to achieve high sensitivity and be able to identify recombinant lineages without retraining the underlying model.

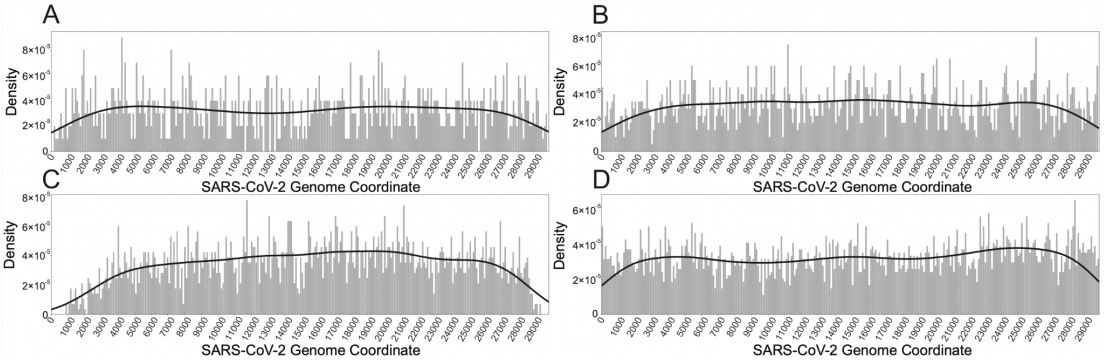


Figure 5.2: RIPPLES is highly sensitive and able to detect 93% of simulated breakpoints. *We simulated 8,000 recombinant samples: 4,000 with one breakpoint and 4,000 with two breakpoints. The distributions of the true breakpoints for the simulated one- and two-breakpoint samples are shown in A and B, and were generated by choosing random positions across the genome, except that two breakpoints in a sample, if present, must be 1,000 nucleotides apart. The distribution of breakpoints detected for each simulated sample is shown by breakpoint position, with one-breakpoint samples in C, and two-breakpoint samples in D.*

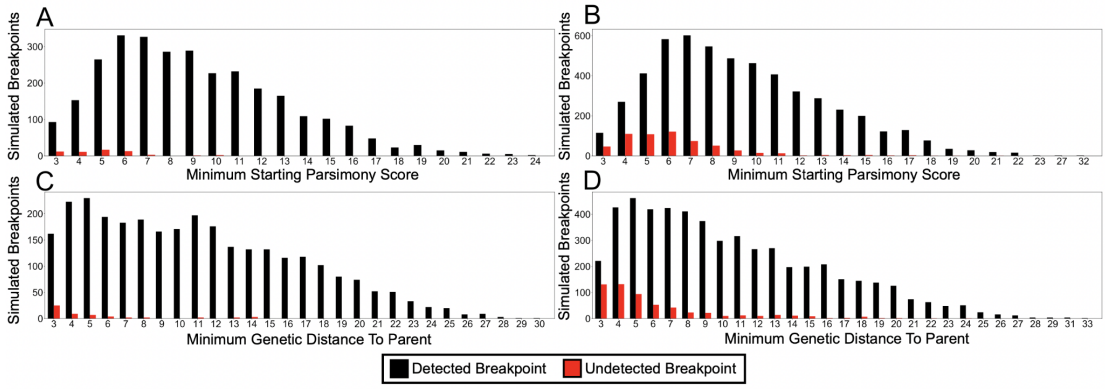


Figure 5.3: RIPPLES more easily detects breakpoints causing large changes in parsimony score. *The distribution of simulated breakpoints detected for each simulated sample is shown by minimum starting parsimony score for the simulated one-breakpoint (A) and two-breakpoint (B) sample, and minimum genetic distance from simulated one-breakpoint (C) and two-breakpoint (D) sample to parent. Minimum starting parsimony (A, B) is dependent upon the initial placement of the recombinant node in the tree and refers to the genetic distance in mutations between the recombinant node and its direct parent in the phylogeny. Minimum genetic distance from sample to parent (C, D) refers to the number of mutations relevant to recombination that separate the recombinant node from either the donor or the acceptor, and is not dependent on the initial phylogeny. Detected breakpoints are shown in black and undetected breakpoints are shown in red. We condition on locating the true breakpoints and observing a significant parsimony score according to our phylogenetic null model. Therefore, we exclude recombination events with minimum starting parsimony scores and genetic distances of less than 3, as these are not significant under our null model.*

Recombination analysis using RIPPLES on a global phylogeny of approximately 1.6 million SARS-CoV-2 genomes reveals that a significant fraction of the sequenced SARS-CoV-2 genomes belong to detectable recombinant lineages. To mitigate the impacts of sequencing and assembly errors, we exclude all nodes with only a single descendant, we applied conservative filters to remove potentially spurious samples from the recombinant sets flagged by RIPPLES, and we manually confirmed mutations in a subset of putatively recombinant samples using raw sequence read data (Figure 5.4). After this, we retained 606

unique recombination events, which have a combined total of 43,163 descendant samples (Table 6.1). This means that approximately 2.7% of total sampled SARS-CoV-2 genomes are inferred to belong to detectable recombinant lineages. *Post hoc* statistical analysis yields an empirical false discovery rate estimate of 11.6% for our statistical thresholds (Text S9, Table S4). Additionally, excess similarity of geographic location and date metadata among the descendants of donor and acceptor nodes strongly supports the notion that the ancestors of recombinant genomes co-circulated within human populations – a prerequisite for recombination. Because recombination events that occur between genetically similar viral lineages are challenging to detect (Figure 5.3), ours is expected to be a potentially large underestimate of the overall frequency of recombination. As a result, the RIPPLES estimate is likely conservative with respect to the global frequency of recombination in the SARS-CoV-2 population.

Figure 5.4: Examples of detected trios filtered out due to sequence quality concerns. ***A:*** *Partial alignment of consensus sequences from a filtered recombinant trio of nodes 77695, 169585, and 77690, centered on site 28225, has consensus sequences of mostly 'N' spanning several sites meant to be informative of a recombination event. This can occur when many descendant samples have missing data. Mismatches between the three consensus sequences immediately flanking this region may be the result of poor sequencing quality as well.* ***B:*** *Partial alignment of consensus sequences from a filtered recombinant trio of nodes 173213, 173209, and 173274, centered on site 16846, has 7 recombination-informative mutations in an 8-nucleotide window that are unlikely to be true mutation events, but rather an alignment artifact or a complex indel event.* ***C:*** *Partial alignment of consensus sequences from a filtered recombinant trio of nodes 293461, 293460, and 211841, centered on site 29769, has 3 mismatches in a 5-nucleotide window, immediately flanked by a large gap in the alignment and are unlikely to be true mutations.*

5.4: The region surrounding the Spike protein is enriched for recombination events.

RIPPLES uncovered a strikingly non-uniform distribution of recombination breakpoint positions across the SARS-CoV-2 genome, consistent with previous analyses in

betacoronaviruses (Patiño-Galindo, Filip, and Rabadan 2021; Müller, Kistler, and Bedford 2021). In particular, there is an excess of recombination breakpoints towards the 3' end of the SARS-CoV-2 genome relative to expectations based on random breakpoint positions ($p < 1\times10^{-7}$; permutation test; Text S11). Importantly, no such bias is apparent when we simulate recombination breakpoints following a uniform distribution (Figure 5.2). Change-point analysis identifies an increase in the frequency of recombination breakpoints immediately 5' of the Spike protein region (20,787 bp; Text S12). The rate of recombination breakpoints is approximately three times higher towards the 3' of the change-point than the 5' interval (Figure 5.5) – which is similar to the relative recombination rates in the genomes of other human coronaviruses (Müller, Kistler, and Bedford 2021).

Several lines of evidence suggest that the skewed distribution of recombination breakpoint positions results primarily from a neutral mechanistic bias rather than being a consequence of positive selection. First, many of these recombinant clades have existed for a relatively short period of time, and might already be extinct. The mean timespan between the earliest and latest dates of observed descendants of detected recombinant nodes is just 37 days. Second, of the subset of recombination events that we inferred to occur between Variants of Concern (VOC; lineages B.1.1.7, B.1.351, B.1.617.2, and P.1 (Rambaut et al. 2020)) and other lineages, VOCs contribute slightly fewer Spike protein mutations than non-VOC lineages on average (58 out of 123 VOC/non-VOC recombinants, P = 0.765, sign test). Third, recombinant clade size does not greatly differ from the remaining clade sizes, which would be expected if recombinant lineages experienced strong selection (P = 0.8401, permutation test). Therefore, although natural selection on recombinant lineages could also impact the observed distribution of recombinant breakpoint positions (Müller, Kistler, and Bedford 2021), our data indicates that an important mechanistic bias shapes the distribution of recombination events across the SARS-CoV-2 genome.

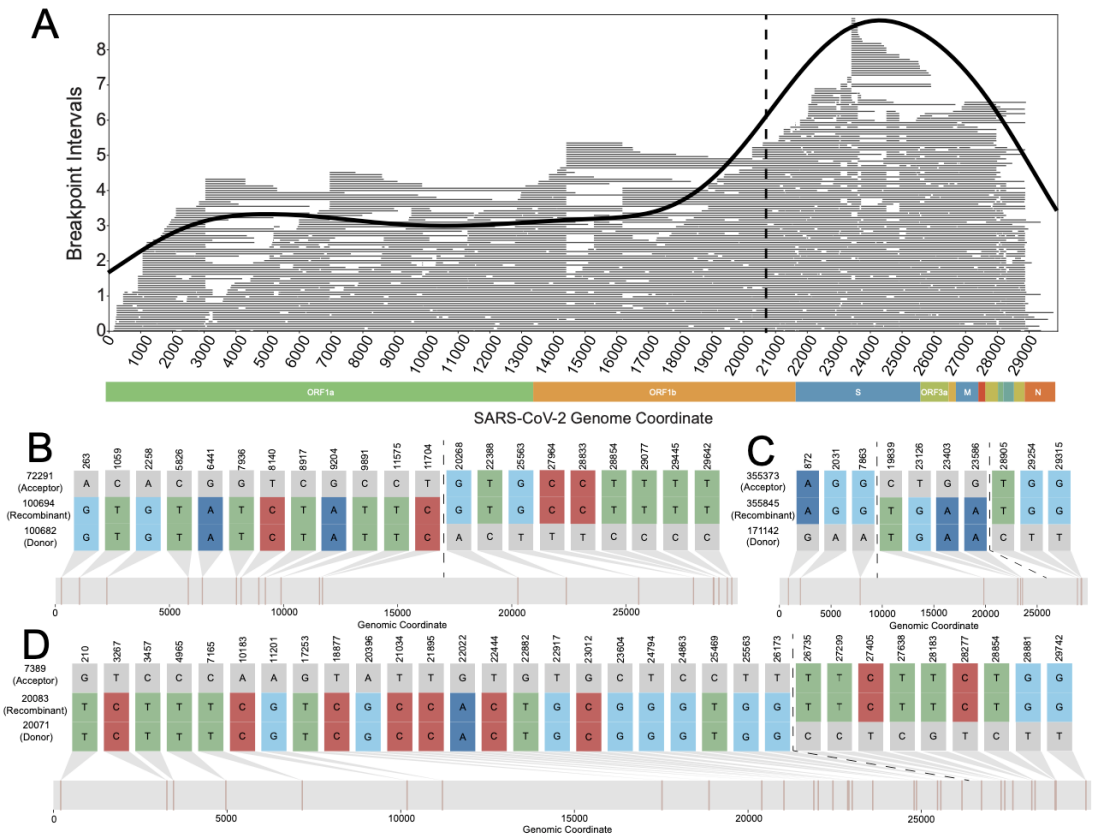


Figure 5.5. RIPPLES detects an excess of recombination in the Spike protein region. ***(A)****: The distribution of midpoints of each breakpoint's prediction interval are shown as a density plot, with the underlying recombination prediction intervals plotted as individual lines in gray. We used the midpoint of the breakpoint prediction interval because recombination events can only be localized to prediction intervals which are the regions between two recombination informative SNPs. A dashed vertical line at position 20,787 delimits recombination rate regions identified by change-point analysis. The apparent lack of recombination towards the chromosome edges likely reflects a detection bias we describe above (Figure 5.2)* ***(B-D)****: Recombination-informative sites (i.e. positions where the recombinant node matches either but not both parent nodes) for three example recombinant trios detected by RIPPLES. The numbers to the left of each sequence correspond to the node identifiers from our MAT. B and D are examples of a recombinant with a single breakpoint (shown in dotted lines), C is an example of a recombinant with two breakpoints. B-D were generated using the SNIPIT package (https://github.com/aineniamh/snipit).*

5.5: Pango lineage B.1.355 is likely the result of a recombination event

Although not yet widespread among circulating SARS-CoV-2 genomes, recombination has measurably contributed to the genetic diversity within SARS-CoV-2 lineages. The ratio of variable positions contributed by recombination versus those resulting

from *de novo* mutation, R/M, is commonly used to summarize the relative impacts of these two sources of variation (Patiño-Galindo, Filip, and Rabadan 2021). Using our dataset of putative recombination events, we estimate that R/M = 0.00264 in SARS-CoV-2 (Text S13). This is low for a coronavirus population (*e.g.* for MERS, R/M is estimated to be 0.25-0.31, (Patiño-Galindo, Filip, and Rabadan 2021)), which presumably reflects the extremely low genetic distances among possible recombinant ancestors during the earliest phases of the pandemic and the conservative nature of our approach. As SARS-CoV-2 populations accumulate genetic diversity and co-infect hosts with other species of viruses, recombination will play an increasingly large role in generating functional genetic diversity and this ratio could increase (D. Kim et al. 2020). RIPPLES is therefore poised to play a primary role in detecting novel recombinant lineages and quantifying their impacts on viral genomic diversity as the pandemic progresses.

Our extensively optimized implementation of RIPPLES allows it to search the entire phylogenetic tree and detect recombination both within and between SARS-CoV-2 lineages without *a priori* defining a set of lineages or clade-defining mutations. This is a key advantage of our approach relative to other methods that cope with the scale of SARS-CoV-2 datasets by reducing the search space for possible recombination events (*e.g., (Jackson et al. 2021; VanInsberghe et al. 2021; Varabyou et al. 2021)*). RIPPLES discovers 239 recombination events within branches of the same Pango lineages. Our results also include 367 inter-lineage recombination events (Table 6.1). Additionally, we find evidence that recombination has influenced the Pangolin SARS-CoV-2 nomenclature system (Rambaut et al. 2020). Specifically, we discover that the root of B.1.355 lineage might have resulted from a recombination event between nodes belonging to the B.1.595 and B.1.371 lineages (Fig 5.6). These diverse recombination events highlight the versatility and strengths of the approach taken in RIPPLES.

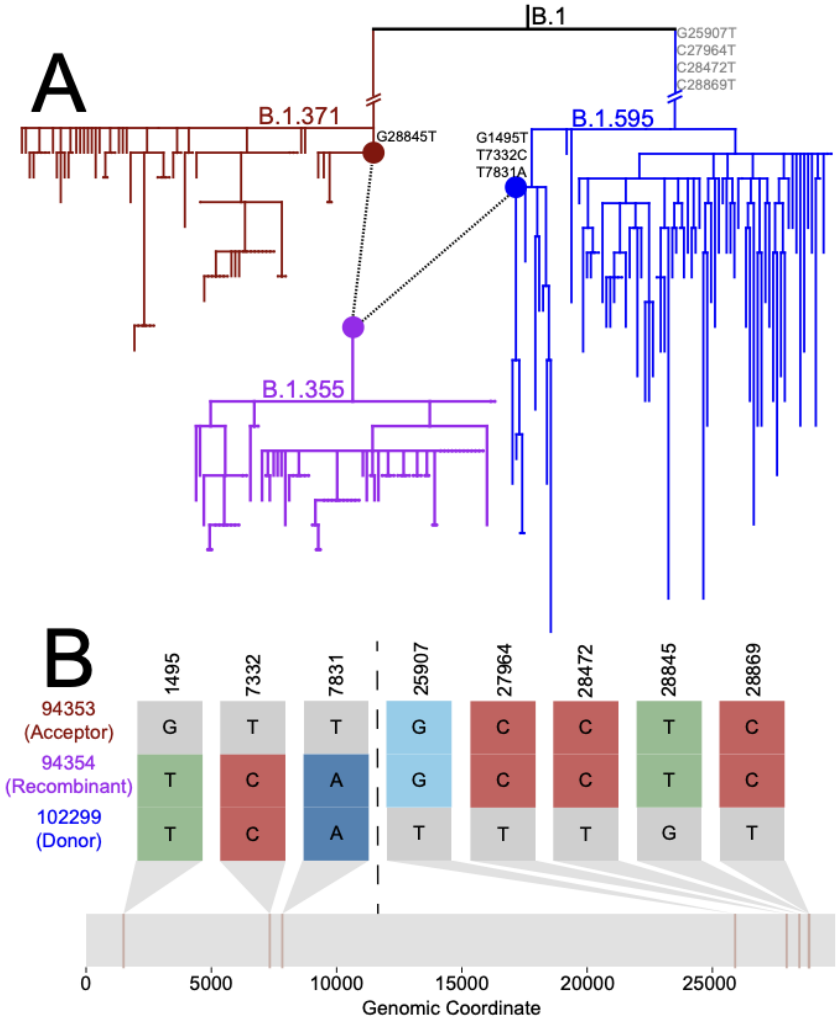


Figure 5.6. RIPPLES uncovered evidence that the B.1.355 lineage might have resulted from a recombination event between lineages of B.1.595 and B.1.371. (A)*: Sub-phylogeny consisting of all 78 B.1.355 samples (purple) and the most closely related 78 samples to nodes 94353 and 102299 from lineages B.1.371 and B.1.595, respectively, using the "k nearest samples" function in matUtils (McBroome et al. 2021). Nodes 94353 (red) and 102299 (blue) are connected by dotted lines to node 94354 (purple), the root of lineage B.1.355. Recombination-informative mutations are marked where they occur in the phylogeny, with those occurring in a parent but not shared by the recombinant sequence shown in gray.* ***(B)****: Recombination-informative sites (i.e. sites where the recombinant node matches either but not both parent nodes) are shown following the same format as Figure 5.5B-D. B was generated using the SNIPIT package (https://github.com/aineniamh/snipit).*

5.6: RIPPLES highlights the need for increased genomic surveillance

The detection of increased recombination rates around the SARS-CoV-2 Spike protein highlight the utility of ongoing surveillance. The Spike protein is a primary location of

adaptive novelty for viral lineages as they adapt to transmission within and among human hosts. Our discovery of the excess of recombination events specifically around the Spike protein, as well as and the relatively high levels of recombinants currently in circulation, underline the importance of monitoring the evolution of new viral lineages that arise through mutation or recombination through real-time analyses of viral genomes. Our work also emphasizes the impact that explicitly considering phylogenetic networks will have for accurate interpretation of SARS-CoV-2 sequences (Müller, Kistler, and Bedford 2021).

Beyond SARS-CoV-2, recombination is a major evolutionary force driving viral and microbial adaptation. It can drive the spread of antibiotic resistance (Didelot and Maiden 2010), drug resistance (Moutouh, Corbeil, and Richman 1996), and immunity and vaccine escape (Golubchik et al. 2012). Identification of recombination is an essential component of pathogen evolutionary analyses pipelines, since recombination can affect the quality of phylogenetic, transmission and phylodynamic inference (Schierup and Hein 2000). For these reasons, computational tools to detect microbial recombination have become very popular and important in recent years (Didelot and Wilson 2015). The SARS-CoV-2 pandemic has driven an unprecedented surge of pathogen genome sequencing and data sharing, which has in turn highlighted some of the limitations of current software in investigating large genomic datasets (Hodcroft, De Maio, et al. 2021). RIPPLES was built for pandemic-scale datasets and is sufficiently optimized to exhaustively search for recombination in one of the largest phylogenies ever inferred in less than four days on the n2d-highcpu-224 Google Cloud Platform (GCP) instance containing 224 vCPUs. To facilitate real-time analysis of recombination among tens of thousands of new SARS-CoV-2 sequences being generated by diverse research groups worldwide each day (Yuelong Shu and McCauley 2017; Sayers et al. 2021), RIPPLES provides an option to evaluate evidence for recombination ancestry in any user-supplied samples within minutes. RIPPLES therefore opens the door for rapid analysis of recombination in heavily sampled and rapidly evolving pathogen populations, as well as providing a tool for real-time investigation of recombinants during a pandemic.

5.7: Methods

5.7.1: Constructing a null model

It is necessary to define a null model in order to determine whether we observe more recombination events than would be expected as false positives. Here, as an alternative to recombination, we define a null model wherein the additional mutations on a branch that we will test for recombination result instead from the underlying observed mutation process. To do this, we selected nodes at random and added k additional mutations, where k is an input parameter. Here, each mutation was drawn proportionally to the parsimony score of that mutation in the global phylogeny. This should make the extended branches we consider here consistent with the underlying null model. Importantly, our correction for *de novo* mutations should be more appropriate than alternative null models that assume that the mutation rate is equal across all sites (*e.g., (VanInsberghe et al. 2021)*). Furthermore, to whatever extent recombination contributes to apparently recurrent mutations, this model will be conservative for establishing significance under the null (below).

After generating sequences with additional mutations as described above, we placed those samples onto the phylogeny using UShER (Turakhia et al. 2021). Then, we searched for all possible partial placements using RIPPLES. We record the resulting improvement in parsimony score in the best partial placement that we found relative to the initial placement. The distribution of parsimony score improvement for each initial parsimony score provides a null model for the amount of improvement that might be expected under a model where mutation generates the long branches we search for and conditional on the phylogeny and the initial parsimony score.

5.7.2: Establishing significance under a null model based on observed mutation rates.

For each putative recombinant, we use the null distribution based on mutation on a single phylogeny without recombination to establish significance. For each node with a given initial parsimony score, we obtain the p-value as the proportion of simulated null distribution samples with the same initial parsimony score where the recombinant parsimony score improved by an equal number or more mutations than in the putative recombinant sample. Because the parsimony score improvement distribution is discrete and relatively small in value, the p-values obtained will typically be conservative. Furthermore, our test statistic is defined as the best possible parsimony score improvement for a given set of partial placements for a single node. The number of tests performed should therefore be linear with respect to the number of potential recombinant nodes evaluated. This property will typically be appealing when applying a false discovery rate correction because many tests will be highly correlated among possible parent nodes due to the nodes' proximity within the phylogenetic tree. This can be a problem with methods that are not phylogenetic, *e.g.,* those that examine all possible trios for donor-acceptor-recombinant relationships (e.g., (H. M. Lam, Ratmann, and Boni 2018)). With such methods, in some cases, if two nodes are distinguished by a SNP that is not contained within a recombinant segment, two or more ancestral nodes can yield identical results. More generally, closely related trios will yield highly correlated results which can impose important challenges for multiple testing corrections.

5.7.3: Tree pruning and sample filtration.

In order to test our method and detect as many SARS-CoV-2 recombination events as possible, we required a large phylogeny encompassing the genetic diversity of the virus. At UCSC, we have been maintaining a daily-updated SARS-CoV-2 phylogeny of all GISAID (Yuelong Shu and McCauley 2017), GenBank (Sayers et al. 2021) and COG-UK (COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk 2020) sequences using the script

https://github.com/ucscGenomeBrowser/kent/blob/master/src/hg/utils/otto/sarscov2phylo/updatePublic.sh and the method described in (Turakhia et al. 2021; McBroome et al. 2021). We started with our phylogeny dated 28/05/2021 containing a total of 1,807,630 sequences with a parsimony score of 1,772,324. We then used the corresponding VCF file and masked all known problematic sites (Turakhia, Thornlow, Gozashti, et al. 2020) and pruned out samples with fewer than 28,000 non-N nucleotides at positions where the SARS-CoV-2 reference genome had a non-N nucleotide. We also pruned out all samples with 2 or more ambiguous (non-[ACGTN-]) nucleotides, and then iteratively removed all samples on branches with length greater than 30 using the *-b 30* flag in matUtils. After this, we ran matOptimize twice using an SPR radius of 10 and 40 in subsequent rounds, and using the masked VCF as an input. Following this, we again iteratively pruned out all samples on branches with length greater than 30. The final tree contains 1,607,799 samples, 1,967,136 nodes, and has a total parsimony score of 1,522,210. We then optimized the tree using matOptimize (https://github.com/yatisht/usher).

5.7.4: Establishing sensitivity

To test RIPPLES' sensitivity, we simulated 8,000 recombinant samples by choosing 2 random internal nodes from our phylogeny with at least 10 descendants and choosing breakpoints at random across the genome. We generated 1,000 simulations each for one and two breakpoints with 0, 1, 2, and 3 additional mutations added to the sequence after the recombination event. We ensured that any two breakpoints were at least 1,000 nucleotides apart. The distribution of breakpoints selected for this experiment is approximately uniform, with slight bias against the ends of the chromosomes caused by this 1,000-nucleotide condition.

We then measured the ability of RIPPLES to detect breakpoints as a function of the position of the breakpoint and the minimum genetic distance from the recombinant node to either parent. Overall, we detect 93% of all breakpoints across all of our simulations. Scripts

used to generate simulated recombinants are available at https://github.com/bpt26/recombination/.

In addition to simulations, we evaluated the sensitivity of RIPPLES by asking if it could detect each of the high-confidence recombinant SARS-CoV-2 clusters of Jackson et al. (Jackson et al. 2021). Briefly, this work used the unique and highly divergent B.1.1.7 haplotype to detect putative recombination events. To do this, we ran RIPPLES while relaxing the requirement that each detected recombinant have a minimum of two descendants. We did this because several of the clusters identified in that work have only a single extant descendant. We found that all putative recombination events identified in that work are also discovered by RIPPLES.

5.7.5: Filtering possible false positives.

We applied several *post hoc* filters to remove putative recombinant nodes that may be false positives resulting from several possible sources of error. For each internal node from each trio (putative recombinant, donor, and acceptor nodes) that comprised a recombinant event, we downloaded the consensus genome sequence for the nearest descendants of each node, from COG-UK, GenBank, GISAID, and the China National Center for Bioinformatics. We then aligned the sequences of all descendants for each trio using MAFFT (Katoh and Standley 2013), focusing specifically on recombination-informative sites, i.e. where the allele of the recombinant node matched one parent node but not the other. From each set of descendants, we created a consensus sequence for the recombinant, donor, and acceptor nodes. We then compared these consensus sequences to determine whether the informative sites for recombination were likely to be true mutations, or alignment artifacts not captured by our initial VCF file.

If an insertion or a deletion (indel) in the alignment or a set of missing bases (Ns) spanned at least one recombination-informative site in at least one of the consensus sequences, or if an informative site was within 5 nucleotides of an indel or set of missing

bases at least 5 nucleotides long, or if more than 5 informative sites were within 2 nucleotides of an indel or set of Ns of any length, we discarded the trio. From careful inspection of individual trios, variation fitting these criteria might be influenced by sequencing quality. We also discarded trios containing more than 5 recombination-informative mutations in a 20-nucleotide span. While multi-nucleotide mutation events do occur, we found upon inspection of the raw sequences that cases of more than 5 mutations in such a small window most often occurred very near to either end of the sequence for that sample. We then discarded trios where 3 or more recombination-informative mutations in a 20-nucleotide span were found within 5 nucleotides of an indel or set of Ns at least 10 nucleotides in length. Finally, we removed trios for which the entire set of recombination-informative mutations in the donor or acceptor sequence occurred in a 20-nucleotide span. We have aimed to be conservative with our filtering and excluding these trios may eliminate some true variation from our dataset, but this conservative approach should limit false positives.

To further remove low-quality recombination events, we removed cases whose p-value in 3seq (H. M. Lam, Ratmann, and Boni 2018) was greater than 0.2. 3seq conducts non-parametric tests for clustering in sequences of binary values. We generated binary sequences using the informative sites for each trio ("A" if the recombinant matched only the donor, "B" if the recombinant matched only the acceptor). Our choice for a p-value of 0.2 is based on visual inspection of binary sequences. For example, a sequence of "AAAABBB" is assigned a p-value of 0.143, and "AAABBB" is assigned a p-value of 0.2. Our intention with this filter is to remove obviously erroneous recombination events, but a recombination event between nodes with few total informative sites could certainly result in such a sequence. However, the sequences "BAAABBABBBBBBBA" and "ABBBAABAAAAAAAB" result in p-values of 0.275. Clustering in these sequences do not resemble what we expect from simple recombination events and might be the result of contamination or mixed infections.

After controlling for sequence quality, we compared each parsimony improvement to the phylogenetically informed null model described above. We retained only trios whose

p-value was less than 0.05, where the p-value represents the proportion of null samples, with parsimony score improvements of at least that observed for the sample of interest, given the same initial parsimony score. We then needed to remove redundant trios from this set of statistically significant predicted recombinants. Several recombinant nodes had predicted recombination events with different sets of parents, and/or different predicted breakpoint intervals, but because multiple recombination events are extremely unlikely to have occurred at one node, we retained only one recombination event for each node. To break ties, we favored recombination events for which we predicted only one breakpoint. Then, we favored trios with fewer informative sites. These represent cases where the donor and acceptor have more similar sequences, and we expect that strains with more similar sequences would be more likely to be in the same place at the same time, as is required for recombination to take place. After this, we resolved the remaining ties by favoring the trio with the smaller 3seq p-value, larger predicted breakpoint interval, and greater sum of descendants of the donor and acceptor nodes. Finally, we found a few cases where two predicted recombinant nodes were the acceptor or donor of each other, and retained only one event for these cases. To accomplish this, we applied the same set of sequential tiebreakers described above. After applying these filters, we retained 606 unique putative recombinant nodes, which are parents to 43,163 unique descendant samples. Scripts used for filtering results as described here are available at https://github.com/bpt26/recombination.

5.7.6: Empirical false discovery rate estimation

To estimate the false discovery rate associated with our specific approach and statistical threshold selected, we computed a *post hoc* empirical false discovery rate. To do this, we obtained the number of internal nodes that we tested and which were associated with a given parsimony score. Then, for each initial parsimony score and parsimony score improvement, we obtained the expected number of internal nodes that would display that parsimony score improvement under the null model, i.e. as a consequence of mutational

processes and in the absence of recombination. We estimate the false discovery rate as the ratio of expected nodes for a given initial and final parsimony score to the number of detected recombinant nodes with the same initial and final parsimony score. As would be expected, more modest parsimony score improvements are associated with a higher estimated false discovery rate (Table 5.1).

Starting Parsimony	Improvement	Nodes in Tree	P-value	Expected False Discoveries	Actual Discoveries	FDR
6	3	2654	0.001143510577	3.034877073	35	0.0867107735
6	4	2654	0.000571755288	1.517438536	14	0.1083884669
6	5	2654	0.000571755288	1.517438536	9	0.1686042818
6	6	2654	0.000571755288	1.517438536	3	0.5058128454
7	3	1456	0.002528445006	3.681415929	23	0.1600615621
7	4	1456	0.000632111251	0.9203539823	7	0.1314791403
7	5	1456	0.000632111251	0.9203539823	4	0.2300884956
7	6	1456	0.000632111251	0.9203539823	3	0.3067846608
7	7	1456	0.000632111251	0.9203539823	2	0.4601769912
8	3	796	0.003248862898	2.586094867	14	0.1847210619
8	4	796	0.000649772579	0.5172189734	6	0.08620316223
8	5	796	0.000649772579	0.5172189734	4	0.1293047433
8	6	796	0.000649772579	0.5172189734	3	0.1724063245
8	7	796	0.000649772579	0.5172189734	1	0.5172189734
9	3	455	0.002652519894	1.206896552	7	0.1724137931
9	4	455	0.000663129973	0.3017241379	5	0.06034482759
9	5	455	0.000663129973	0.3017241379	1	0.3017241379
9	6	455	0.000663129973	0.3017241379	3	0.1005747126
9	7	455	0.000663129973	0.3017241379	3	0.1005747126
9	8	455	0.000663129973	0.3017241379	1	0.3017241379
9	9	455	0.000663129973	0.3017241379	1	0.3017241379
10	3	267	0.005365526492	1.432595573	6	0.2387659289
10	4	267	0.000670690811	0.1790744467	4	0.04476861167
10	5	267	0.000670690811	0.1790744467	1	0.1790744467
10	6	267	0.000670690811	0.1790744467	1	0.1790744467
10	9	267	0.000670690811	0.1790744467	2	0.08953722334
11	10	167	0.000736377025	0.1229749632	1	0.1229749632
11	3	167	0.00736377025	1.229749632	2	0.6148748159
11	6	167	0.000736377025	0.1229749632	1	0.1229749632
12	3	108	0.01051709027	1.135845749	1	1.135845749
12	4	108	0.000876424189	0.09465381245	1	0.09465381245
12	5	108	0.000876424189	0.09465381245	1	0.09465381245
12	8	108	0.000876424189	0.09465381245	2	0.04732690622
13	3	84	0.01445086705	1.213872832	4	0.3034682081
13	5	84	0.001445086705	0.1213872832	1	0.1213872832
13	8	84	0.001445086705	0.1213872832	1	0.1213872832
15	3	32	0.01978417266	0.6330935252	1	0.6330935252
15	5	32	0.001798561151	0.05755395683	1	0.05755395683
15	9	32	0.001798561151	0.05755395683	1	0.05755395683
16	3	23	0.01254480287	0.2885304659	1	0.2885304659
16	8	23	0.001792114695	0.04121863799	1	0.04121863799
17	3	24	0.02312138728	0.5549132948	1	0.5549132948
17	6	24	0.001926782274	0.04624277457	1	0.04624277457
18	4	9	0.002375296912	0.02137767221	1	0.02137767221
18	5	9	0.002375296912	0.02137767221	1	0.02137767221
21	12	6	0.003745318352	0.02247191011	1	0.02247191011
Total FDR				70.22869144	606	0.1158889298

Table 5.1. False discovery rate estimation for each parsimony score improvement observed in our dataset.

5.7.7: Permutation test to evaluate the apparent excess of 3' recombination.

We next sought to determine if identified recombination breakpoints are shifted towards the 3’ end of the genome. To do this, we performed a permutation test comparing the

difference of the mean of the distribution of detected breakpoints when recombination breakpoints are simulated uniformly at random with the mean of the breakpoint position distribution in the true set. Briefly, this is accomplished by randomizing the set of breakpoint positions between two vectors of equivalent lengths to the simulated and real sets. The reported p-value is the proportion of such permutations where the difference between the mean position of the true and simulated vectors was greater than or equal to the observed difference in the true data. Importantly, because both distributions reflect subsets of recombination events that can be detected conditional on the landscape of genetic diversity and phylogeny of SARS-CoV-2, this is an improved null comparison than assuming a distribution, *e.g.,* a uniform distribution.

5.7.8: Estimating R/M.

A central focus of much of microbial evolutionary analysis is distinguishing the relative contributions of recombination and mutation to patterns of variation. To estimate this ratio for SARS-CoV-2, we conservatively assume that RIPPLES successfully detects all recombination events that are present on the phylogeny. Then, the decrease in parsimony score associated with each detected recombination event is an estimate of the total variation that results from recombination. The contribution to the total mutations present in the viral population is then the parsimony score decrease multiplied by the number of descendant lineages of that recombinant node. This is the total number of observed mutations whose genealogies contain a recombination event. For lineages descendant of multiple recombinant nodes, we multiplied by the recombinant node with greater parsimony score improvement. If we subtract this value from the total number of mutations observed across the entire datasets, we obtain an estimate of the number of mutations whose histories are attributed in whole to mutational processes. The ratio of these two numbers is an estimate of R/M averaged across all samples that are included in our tree.

Conclusion

Transfer RNAs are a ubiquitous RNA gene family whose function is conserved in and essential to all domains of life. Their high transcription, mutation rate, and copy number make them an ideal model gene family for various population genetic studies. In the first chapter, I discovered that they are heavily influenced by transcription-associated mutagenesis, and that the most highly transcribed tRNA genes have mutation rates ~10x greater than the genome-wide average. In the second chapter, I leveraged the correlation between mutation rate and expression level to infer tRNA gene activity for over 10,000 tRNA genes, using only DNA data. In the third chapter, I explored the possible effects of somatic mutation rate in relation to tRNA gene copy number, as well as developed a model relating somatic mutations to dominance.

Although I had planned further tRNA projects, I was able to use my knowledge of both population genetics and RNA biology to make important contributions to SARS-CoV-2 phylogenetics. In the fourth chapter, I demonstrated how lab-specific errors can affect phylogenetic inference, and highlighted the instability of large phylogenies consisting of closely related samples. In the fifth chapter, I demonstrated the power of UShER to revolutionize phylogenetics through the use of simulations, comparisons to other widely used methods, and an application to genomic surveillance using real data from my own community. In the final chapter, I demonstrate RIPPLES on our global phylogeny and its discovery of extensive recombination throughout the pandemic, concentrated in the region of the clinically relevant Spike protein.

While on the surface, tRNA biology and SARS-CoV-2 phylogenetics appear very disparate subjects, the methods used in each chapter have significant overlap. In this dissertation, I have used various comparative genomics and phylogenetics techniques to find fundamental insights to the evolution of both tRNA genes and SARS-CoV-2 lineages.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Abascal-Palacios, Guillermo, Ewan Phillip Ramsay, Fabienne Beuron, Edward Morris, and Alessandro Vannini. 2018. “Structural Basis of RNA Polymerase III Transcription Initiation.” *Nature* 553 (7688): 301–6.

Afgan, Enis, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, et al. 2016. “The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2016 Update.” *Nucleic Acids Research* 44 (W1): W3–10.

Agrawal, Aneil F., and Michael C. Whitlock. 2012. “Mutation Load: The Fitness of Individuals in Populations Where Deleterious Alleles Are Abundant,” November. https://doi.org/10.1146/annurev-ecolsys-110411-160257.

Aguilera, Andrés, and Tatiana García-Muse. 2013. “Causes of Genome Instability.” *Annual Review of Genetics* 47 (July): 1–32.

Akther, S., E. Bezrucenkovas, B. Sulkow, and C. Panlasigui. 2020. “CoV Genome Tracker: Tracing Genomic Footprints of Covid-19 Pandemic.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.04.10.036343v1.abstract.

Allison, D. S., and B. D. Hall. 1985. “Effects of Alterations in the 3’ Flanking Sequence on in Vivo and in Vitro Expression of the Yeast SUP4-O tRNATyr Gene.” *The EMBO Journal* 4 (10): 2657–64.

Andersen, Kristian G., Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, and Robert F. Garry. 2020. “The Proximal Origin of SARS-CoV-2.” *Nature Medicine* 26 (4): 450–52.

“An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network.” 2020. *The Lancet Microbe*. https://doi.org/10.1016/s2666-5247(20)30054-9.

Anisimova, Maria, and Olivier Gascuel. 2006. “Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative.” *Systematic Biology* 55 (4): 539–52.

Arabidopsis Genome Initiative. 2000. “Analysis of the Genome Sequence of the Flowering Plant Arabidopsis Thaliana.” *Nature* 408 (6814): 796–815.

Arimbasseri, Aneeshkumar G., Keshab Rijal, and Richard J. Maraia. 2013. “Transcription Termination by the Eukaryotic RNA Polymerase III.” *Biochimica et Biophysica Acta* 1829 (3-4): 318–30.

Armstrong, Joel, Glenn Hickey, Mark Diekhans, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, et al. 2019. “Progressive Alignment with Cactus: A Multiple-Genome Aligner for the Thousand-Genome Era.” https://doi.org/10.1101/730531.

Badua, Christian Luke D. C., Karol Ann T. Baldo, and Paul Mark B. Medina. 2021. “Genomic and Proteomic Mutation Landscapes of SARS-CoV-2.” *Journal of Medical Virology* 93 (3): 1702–21.

Barbera, Pierre, Alexey M. Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. 2019. “EPA-Ng: Massively Parallel Evolutionary Placement of Genetic Sequences.” *Systematic Biology* 68 (2): 365–69.

Bedford, Trevor, Alexander L. Greninger, Pavitra Roychoudhury, Lea M. Starita, Michael Famulare, Meei-Li Huang, Arun Nalla, et al. 2020. “Cryptic Transmission of SARS-CoV-2 in Washington State.” *Science* eabc0523 (September). https://doi.org/10.1126/science.abc0523.

Bloom-Ackermann, Zohar, Sivan Navon, Hila Gingold, Ruth Towers, Yitzhak Pilpel, and Orna Dahan. 2014. “A Comprehensive tRNA Deletion Library Unravels the Genetic Architecture of the tRNA Pool.” *PLoS Genetics* 10 (1): e1004084.

Bogdanowicz, Damian, Krzysztof Giaro, and Borys Wróbel. 2012. “TreeCmp: Comparison of Trees in Polynomial Time.” *Evolutionary Bioinformatics Online* 8 (January): EBO.S9657.

Bogu, Gireesh K., Pedro Vizán, Lawrence W. Stanton, Miguel Beato, Luciano Di Croce, and

Marc A. Marti-Renom. 2015. “Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse.” *Molecular and Cellular Biology* 36 (5): 809–19.

Boivin, Vincent, Gabrielle Deschamps-Francoeur, Sonia Couture, Ryan M. Nottingham, Philia Bouchard-Bourelle, Alan M. Lambowitz, Michelle S. Scott, and Sherif Abou-Elela. 2018. “Simultaneous Sequencing of Coding and Noncoding RNA Reveals a Human Transcriptome Dominated by a Small Number of Highly Expressed Noncoding Genes.” *RNA* 24 (7): 950–65.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, et al. 2011. “The Evolution of Gene Expression Levels in Mammalian Organs.” *Nature* 478 (7369): 343–48.

Canella, Donatella, Viviane Praz, Jaime H. Reina, Pascal Cousin, and Nouria Hernandez. 2010. “Defining the RNA Polymerase III Transcriptome: Genome-Wide Localization of the RNA Polymerase III Transcription Machinery in Human Cells.” *Genome Research* 20 (6): 710–21.

Cao, Jun, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, et al. 2011. “Whole-Genome Sequencing of Multiple Arabidopsis Thaliana Populations.” *Nature Genetics* 43 (10): 956–63.

Casper, Jonathan, Ann S. Zweig, Chris Villarreal, Cath Tyner, Matthew L. Speir, Kate R. Rosenbloom, Brian J. Raney, et al. 2018. “The UCSC Genome Browser Database: 2018 Update.” *Nucleic Acids Research* 46 (D1): D762–69.

Chan, Patricia P., Brian Y. Lin, Allysia J. Mak, and Todd M. Lowe. n.d. “tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes.” https://doi.org/10.1101/614032.

Chan, Patricia P., and Todd M. Lowe. 2016. “GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes.” *Nucleic Acids Research* 44 (D1): D184–89.

Chen, Lei, Jiarui Li, Yu-Hang Zhang, Kaiyan Feng, Shaopeng Wang, Yunhua Zhang, Tao

Huang, Xiangyin Kong, and Yu-Dong Cai. 2018. “Identification of Gene Expression Signatures across Different Types of Neural Stem Cells with the Monte-Carlo Feature Selection Method.” *Journal of Cellular Biochemistry* 119 (4): 3394–3403.

Chiara, Matteo, David S. Horner, Carmela Gissi, and Graziano Pesole. 2020. “Comparative Genomics Suggests Limited Variability and Similar Evolutionary Patterns between Major Clades of SARS-CoV-2.” *bioRxiv*. https://doi.org/10.1101/2020.03.30.016790.

Consortium, Mouse Genome Sequencing, and Mouse Genome Sequencing Consortium. 2002. “Initial Sequencing and Comparative Analysis of the Mouse Genome.” *Nature*. https://doi.org/10.1038/nature01262.

Corbett-Detig, Russell B., Daniel L. Hartl, and Timothy B. Sackton. 2015. “Natural Selection Constrains Neutral Diversity across a Wide Range of Species.” *PLoS Biology* 13 (4): e1002112.

COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. 2020. “An Integrated National Scale SARS-CoV-2 Genomic Surveillance Network.” *The Lancet. Microbe* 1 (3): e99–100.

Cozen, Aaron E., Erin Quartley, Andrew D. Holmes, Eva Hrabeta-Robinson, Eric M. Phizicky, and Todd M. Lowe. 2015. “ARM-Seq: AlkB-Facilitated RNA Methylation Sequencing Reveals a Complex Landscape of Modified tRNA Fragments.” *Nature Methods* 12 (9): 879–84.

Davies, Nicholas G., Sam Abbott, Rosanna C. Barnard, Christopher I. Jarvis, Adam J. Kucharski, James Munday, Carl A. B. Pearson, et al. 2021. “Estimated Transmissibility and Severity of Novel SARS-CoV-2 Variant of Concern 202012/01 in England.” *MedRxiv*, 2020–2012.

Dellicour, Simon, Keith Durkin, Samuel L. Hong, Bert Vanmechelen, Joan Martí-Carreras, Mandev S. Gill, Cécile Meex, Sébastien Bontems, Emmanuel André, Marius Gilbert, Conor Walker, Nicola De Maio, James Hadfield, et al. n.d. “A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2

Lineages.” https://doi.org/10.1101/2020.05.05.078758.

Dellicour, Simon, Keith Durkin, Samuel L. Hong, Bert Vanmechelen, Joan Martí-Carreras, Mandev S. Gill, Cécile Meex, Sébastien Bontems, Emmanuel André, Marius Gilbert, Conor Walker, Nicola De Maio, Nuno R. Faria, et al. n.d. “A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages.” https://doi.org/10.1101/2020.05.05.078758.

De Maio, Nicola, Conor R. Walker, Yatish Turakhia, Robert Lanfear, Russell Corbett-Detig, and Nick Goldman. 2021. “Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2.” *Genome Biology & Evolution*, April. https://doi.org/10.1101/2021.01.14.426705.

De Maio, N., C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, and N. Goldman. 2020. “Issues with SARS-CoV-2 Sequencing Data [Internet]. 2020 [cited 2020 Jun 16].” Virological.org. May 15, 2020. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Deng, Xianding, Miguel A. Garcia-Knight, Mir M. Khalid, Venice Servellita, Candace Wang, Mary Kate Morris, Alicia Sotomayor-González, et al. 2021. “Transmission, Infectivity, and Antibody Neutralization of an Emerging SARS-CoV-2 Variant in California Carrying a L452R Spike Protein Mutation.” *medRxiv : The Preprint Server for Health Sciences*, March. https://doi.org/10.1101/2021.03.07.21252647.

Deng, Xianding, Wei Gu, Scot Federman, Louis du Plessis, Oliver G. Pybus, Nuno R. Faria, Candace Wang, et al. 2020. “Genomic Surveillance Reveals Multiple Introductions of SARS-CoV-2 into Northern California.” *Science* 369 (6503): 582–87.

Didelot, Xavier, and Martin C. J. Maiden. 2010. “Impact of Recombination on Bacterial Evolution.” *Trends in Microbiology* 18 (7): 315–22.

Didelot, Xavier, and Daniel J. Wilson. 2015. “ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes.” *PLoS Computational Biology* 11 (2): e1004041.

Dieci, G., and A. Sentenac. 1996. “Facilitated Recycling Pathway for RNA Polymerase III.” *Cell* 84 (2): 245–52.

Doran, James L., Wade H. Bingle, and Kenneth L. Roy. 1988. “Two Human Genes Encoding tRNAGlyGCC.” *Gene* 65 (2): 329–36.

Dorp, Lucy van, Mislav Acman, Damien Richard, Liam P. Shaw, Charlotte E. Ford, Louise Ormond, Christopher J. Owen, et al. 2020. “Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2.” *Infection, Genetics and Evolution*. https://doi.org/10.1016/j.meegid.2020.104351.

Dorp, Lucy van, Damien Richard, Cedric C. S. Tan, Liam P. Shaw, Mislav Acman, and Francois Balloux. n.d. “No Evidence for Increased Transmissibility from Recurrent Mutations in SARS-CoV-2.” https://doi.org/10.1101/2020.05.21.108506.

Drosophila 12 Genomes Consortium, Andrew G. Clark, Michael B. Eisen, Douglas R. Smith, Casey M. Bergman, Brian Oliver, Therese A. Markow, et al. 2007. “Evolution of Genes and Genomes on the Drosophila Phylogeny.” *Nature* 450 (7167): 203–18.

Dudas, Gytis, and Andrew Rambaut. 2016. “MERS-CoV Recombination: Implications about the Reservoir and Potential for Adaptation.” *Virus Evolution* 2 (1): vev023.

ENCODE Project Consortium. 2011. “A User’s Guide to the Encyclopedia of DNA Elements (ENCODE).” *PLoS Biology* 9 (4): e1001046.

———. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Fauver, Joseph R., Mary E. Petrone, Emma B. Hodcroft, Kayoko Shioda, Hanna Y. Ehrlich, Alexander G. Watts, Chantal B. F. Vogels, et al. 2020. “Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States.” *Cell* 181 (5): 990–96.e5.

Felsenstein, Joseph. 1978. “Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.” *Systematic Zoology*. https://doi.org/10.2307/2412923.

———. 1985. “Confidence Limits on Phylogenies: An Approach Using the Bootstrap.” *Evolution; International Journal of Organic Evolution* 39 (4): 783–91.

Fitch, Walter M. 1971. “Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology.” *Systematic Zoology*. https://doi.org/10.2307/2412116.

Foissac, Sylvain, Sarah Djebali, Kylie Munyard, Nathalie Villa-Vialaneix, Andrea Rau, Kevin Muret, Diane Esquerre, et al. 2018. “Livestock Genome Annotation: Transcriptome and Chromatin Structure Profiling in Cattle, Goat, Chicken and Pig.” *bioRxiv*. https://doi.org/10.1101/316091.

Forni, Diego, Rachele Cagliani, and Manuela Sironi. 2020. “Recombination and Positive Selection Differentially Shaped the Diversity of Betacoronavirus Subgenera.” *Viruses* 12 (11). https://doi.org/10.3390/v12111313.

Freeman, Timothy M., Genomics England Research Consortium, Dennis Wang, and Jason Harris. 2020. “Genomic Loci Susceptible to Systematic Sequencing Bias in Clinical Whole Genomes.” *Genome Research* 30 (3): 415–26.

Gaillard, Hélène, and Andrés Aguilera. 2016. “Transcription as a Threat to Genome Integrity.” *Annual Review of Biochemistry* 85 (June): 291–317.

Gámbaro, Fabiana, Sylvie Behillil, Artem Baidaliuk, Flora Donati, Mélanie Albert, Andreea Alexandru, Maud Vanpeene, et al. 2020. “Introductions and Early Spread of SARS-CoV-2 in France, 24 January to 23 March 2020.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 25 (26). https://doi.org/10.2807/1560-7917.ES.2020.25.26.2001200.

Gardiner-Garden, M., and M. Frommer. 1987. “CpG Islands in Vertebrate Genomes.” *Journal of Molecular Biology* 196 (2): 261–82.

Gnatt, A. L., P. Cramer, J. Fu, D. A. Bushnell, and R. D. Kornberg. 2001. “Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 A Resolution.” *Science* 292 (5523): 1876–82.

Gogakos, Tasos, Miguel Brown, Aitor Garzia, Cindy Meyer, Markus Hafner, and Thomas Tuschl. 2017. “Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP.” *Cell Reports* 20 (6): 1463–75.

Golubchik, Tanya, Angela B. Brueggemann, Teresa Street, Robert E. Gertz Jr, Chris C. A. Spencer, Thien Ho, Eleni Giannoulatou, et al. 2012. “Pneumococcal Genome Sequencing Tracks a Vaccine Escape Variant Formed through a Multi-Fragment Recombination Event.” *Nature Genetics* 44 (3): 352–55.

Gómez-González, Belén, and Andrés Aguilera. 2007. “Activation-Induced Cytidine Deaminase Action Is Strongly Stimulated by Mutations of the THO Complex.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (20): 8409–14.

Goodarzi, Hani, Xuhang Liu, Hoang C. B. Nguyen, Steven Zhang, Lisa Fish, and Sohail F. Tavazoie. 2015. “Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement.” *Cell* 161 (4): 790–802.

Goswami, Pratik, Martin Bartas, Matej Lexa, Natália Bohálová, Adriana Volná, Jiří Červeň, Veronika Červeňová, et al. 2021. “SARS-CoV-2 Hot-Spot Mutations Are Significantly Enriched within Inverted Repeats and CpG Island Loci.” *Briefings in Bioinformatics* 22 (2): 1338–45.

Graczyk, Damian, Małgorzata Cieśla, and Magdalena Boguta. 2018. “Regulation of tRNA Synthesis by the General Transcription Factors of RNA Polymerase III - TFIIIB and TFIIIC, and by the MAF1 Protein.” *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms* 1861 (4): 320–29.

Green, Phil, NISC Comparative Sequencing Program, Brent Ewing, Webb Miller, Pamela J. Thomas, and Eric D. Green. 2003. “Transcription-Associated Mutational Asymmetry in Mammalian Evolution.” *Nature Genetics*. https://doi.org/10.1038/ng1103.

Gu, Hongjing, Qi Chen, Guan Yang, Lei He, Hang Fan, Yong-Qiang Deng, Yanxiao Wang, et al. 2020. “Adaptation of SARS-CoV-2 in BALB/c Mice for Testing Vaccine Efficacy.” *Science* 369 (6511): 1603–7.

Guyon, Isabelle, and André Elisseeff. 2003. “An Introduction to Variable and Feature Selection.” *Journal of Machine Learning Research: JMLR* 3 (Mar): 1157–82.

Haldane, J. B. S. 1937. “The Effect of Variation of Fitness.” *The American Naturalist* 71 (735): 337–49.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. “The WEKA Data Mining Software: An Update.” *ACM SIGKDD Explorations Newsletter* 11 (1): 10–18.

Hamada, M., A. L. Sakulich, S. B. Koduru, and R. J. Maraia. 2000. “Transcription Termination by RNA Polymerase III in Fission Yeast. A Genetic and Biochemically Tractable Model System.” *The Journal of Biological Chemistry* 275 (37): 29076–81.

Hanada, Toshikatsu, Stefan Weitzer, Barbara Mair, Christian Bernreuther, Brian J. Wainger, Justin Ichida, Reiko Hanada, et al. 2013. “CLP1 Links tRNA Metabolism to Progressive Motor-Neuron Loss.” *Nature* 495 (7442): 474–80.

Hasler, Daniele, Gerhard Lehmann, Yasuhiro Murakawa, Filippos Klironomos, Leonhard Jakob, Friedrich A. Grässer, Nikolaus Rajewsky, Markus Landthaler, and Gunter Meister. 2016. “The Lupus Autoantigen La Prevents Mis-Channeling of tRNA Fragments into the Human MicroRNA Pathway.” *Molecular Cell* 63 (1): 110–24.

Hedges, S. Blair, Joel Dudley, and Sudhir Kumar. 2006. “TimeTree: A Public Knowledge-Base of Divergence Times among Organisms.” *Bioinformatics* 22 (23): 2971–72.

Helmrich, Anne, Monica Ballarino, and Laszlo Tora. 2011. “Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes.” *Molecular Cell* 44 (6): 966–77.

Hennessy, John L., and David A. Patterson. 2017. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann.

Hickey, Glenn, Benedict Paten, Dent Earl, Daniel Zerbino, and David Haussler. 2013. “HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments.” *Bioinformatics* 29 (10): 1341–42.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. “UFBoot2: Improving the Ultrafast Bootstrap Approximation.” *Molecular Biology*

and Evolution 35 (2): 518–22.

Hodcroft EB, Hadfield J, Neher RA, Bedford T. 2020. “Year-Letter Genetic Clade Naming for SARS-CoV-2 on Nextstain.org.” Virological. June 2, 2020. https://virological.org/t/year-letter-genetic-clade-naming-for-sars-cov-2-on-nextstain-org/498.

Hodcroft, Emma B., Nicola De Maio, Rob Lanfear, Duncan R. MacCannell, Bui Quang Minh, Heiko A. Schmidt, Alexandros Stamatakis, Nick Goldman, and Christophe Dessimoz. 2021. “Want to Track Pandemic Variants Faster? Fix the Bioinformatics Bottleneck.” *Nature*. https://doi.org/10.1038/d41586-021-00525-x.

Hodcroft, Emma B., Daryl B. Domman, Daniel J. Snyder, Kasopefoluwa Oguntuyo, Maarten Van Diest, Kenneth H. Densmore, Kurt C. Schwalm, et al. 2021. “Emergence in Late 2020 of Multiple Lineages of SARS-CoV-2 Spike Protein Variants Affecting Amino Acid Position 677.” *medRxiv : The Preprint Server for Health Sciences*, February. https://doi.org/10.1101/2021.02.12.21251658.

Hoffmann, Markus, Prerna Arora, Rüdiger Groß, Alina Seidel, Bojan F. Hörnich, Alexander S. Hahn, Nadine Krüger, et al. 2021. “SARS-CoV-2 Variants B.1.351 and P.1 Escape from Neutralizing Antibodies.” *Cell*, March. https://doi.org/10.1016/j.cell.2021.03.036.

Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Fredrik Ronquist. 2016. “RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.” *Systematic Biology* 65 (4): 726–36.

Holmes, Andrew. 2018. “Analyzing Regulation of tRNAs, tRNA Fragments, and mRNAs in Whole Genomes.” UC Santa Cruz. https://escholarship.org/uc/item/4n09j1gw.

Holmes, Edward C., and Andrew Rambaut. 2004. “Viral Evolution and the Emergence of SARS Coronavirus.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 359 (1447): 1059–65.

Hopper, Anita K. 2013. “Transfer RNA Post-Transcriptional Processing, Turnover, and

Subcellular Dynamics in the Yeast Saccharomyces Cerevisiae." *Genetics* 194 (1): 43–67.

Hubisz, Melissa J., Katherine S. Pollard, and Adam Siepel. 2011. "PHAST and RPHAST: Phylogenetic Analysis with Space/time Models." *Briefings in Bioinformatics* 12 (1): 41–51.

Huddleston, John, James Hadfield, Thomas R. Sibley, Jover Lee, Kairsten Fay, Misja Ilcisin, Elias Harkins, Trevor Bedford, Richard A. Neher, and Emma B. Hodcroft. 2021. "Augur: A Bioinformatics Toolkit for Phylogenetic Analyses of Human Pathogens." *Journal of Open Source Software* 6 (57): 2906.

Hummel, Guillaume, Jessica Warren, and Laurence Drouard. 2019. "The Multi-Faceted Regulation of Nuclear tRNA Gene Transcription." *IUBMB Life*. https://doi.org/10.1002/iub.2097.

Ishimura, Ryuta, Gabor Nagy, Ivan Dotu, Huihao Zhou, Xiang-Lei Yang, Paul Schimmel, Satoru Senju, Yasuharu Nishimura, Jeffrey H. Chuang, and Susan L. Ackerman. 2014. "RNA Function. Ribosome Stalling Induced by Mutation of a CNS-Specific tRNA Causes Neurodegeneration." *Science* 345 (6195): 455–59.

Jackson, Ben, Maciej F. Boni, Matthew J. Bull, Amy Colleran, Rachel M. Colquhoun, Alistair Darby, Sam Haldenby, et al. 2021. "Generation and Transmission of Inter-Lineage Recombinants in the SARS-CoV-2 Pandemic." *medRxiv*. https://www.medrxiv.org/content/10.1101/2021.06.18.21258689v1.abstract.

Jacob, Jobin John, Karthick Vasudevan, Agila Kumari Pragasam, Karthik Gunasekaran, Gagandeep Kang, Balaji Veeraraghavan, and Ankur Mutreja. 2020. "Evolutionary Tracking of SARS-CoV-2 Genetic Variants Highlights Intricate Balance of Stabilizing and Destabilizing Mutations." https://doi.org/10.1101/2020.12.22.423920.

Jinks-Robertson, Sue, and Ashok S. Bhagwat. 2014. "Transcription-Associated Mutagenesis." *Annual Review of Genetics* 48 (September): 341–59.

Jukes, Thomas H., Charles R. Cantor, and Others. 1969. "Evolution of Protein Molecules." *Mammalian Protein Metabolism* 3: 21–132.

Juo, Z. S., T. K. Chiu, P. M. Leiberman, I. Baikalov, A. J. Berk, and R. E. Dickerson. 1996. "How Proteins Recognize the TATA Box." *Journal of Molecular Biology* 261 (2): 239–54.

Karolchik, Donna, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. 2004. "The UCSC Table Browser Data Retrieval Tool." *Nucleic Acids Research* 32 (Database issue): D493–96.

Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.

Keightley, Peter D., and Adam Eyre-Walker. 2007. "Joint Inference of the Distribution of Fitness Effects of Deleterious Mutations and Population Demography Based on Nucleotide Polymorphism Frequencies." *Genetics* 177 (4): 2251–61.

Kelleher, Jerome, Kevin R. Thornton, Jaime Ashander, and Peter L. Ralph. 2018. "Efficient Pedigree Recording for Fast Population Genetics Simulation." *PLoS Computational Biology* 14 (11): e1006581.

Kersey, Paul Julian, James E. Allen, Irina Armean, Sanjay Boddu, Bruce J. Bolt, Denise Carvalho-Silva, Mikkel Christensen, et al. 2016. "Ensembl Genomes 2016: More Genomes, More Complexity." *Nucleic Acids Research* 44 (D1): D574–80.

Kharchenko, Peter V., Michael Y. Tolstorukov, and Peter J. Park. 2008. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins." *Nature Biotechnology* 26 (12): 1351–59.

Kim, David, James Quinn, Benjamin Pinsky, Nigam H. Shah, and Ian Brown. 2020. "Rates of Co-Infection Between SARS-CoV-2 and Other Respiratory Pathogens." *JAMA: The Journal of the American Medical Association* 323 (20): 2085–86.

Kim, Nayun, and Sue Jinks-Robertson. 2012. "Transcription as a Source of Genome Instability." *Nature Reviews. Genetics* 13 (3): 204–14.

Kirchner, Sebastian, and Zoya Ignatova. 2015. "Emerging Roles of tRNA in Adaptive Translation, Signalling Dynamics and Disease." *Nature Reviews. Genetics* 16 (2):

98–112.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. “Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus.” *Cell* 182 (4): 812–27.e19.

Korber, B., W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, B. Foley, et al. n.d. “Spike Mutation Pipeline Reveals the Emergence of a More Transmissible Form of SARS-CoV-2.” https://doi.org/10.1101/2020.04.29.069054.

Koski, R. A., S. G. Clarkson, J. Kurjan, B. D. Hall, and M. Smith. 1980. “Mutations of the Yeast SUP4 tRNATyr Locus: Transcription of the Mutant Genes in Vitro.” *Cell* 22 (2 Pt 2): 415–25.

Krinner, Simone, Asli P. Heitzer, Sarah D. Diermeier, Ingrid Obermeier, Gernot Längst, and Ralf Wagner. 2014. “CpG Domains Downstream of TSSs Promote High Levels of Gene Expression.” *Nucleic Acids Research* 42 (6): 3551–64.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. “TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.” *Molecular Biology and Evolution* 34 (7): 1812–19.

Kuras, L., P. Kosa, M. Mencia, and K. Struhl. 2000. “TAF-Containing and TAF-Independent Forms of Transcriptionally Active TBP in Vivo.” *Science* 288 (5469): 1244–48.

Kutter, Claudia, Gordon D. Brown, Angela Gonçalves, Michael D. Wilson, Stephen Watt, Alvis Brazma, Robert J. White, and Duncan T. Odom. 2011. “Pol III Binding in Six Mammals Shows Conservation among Amino Acid Isotypes despite Divergence among tRNA Genes.” *Nature Genetics* 43 (10): 948–55.

Lack, Justin B., Charis M. Cardeno, Marc W. Crepeau, William Taylor, Russell B. Corbett-Detig, Kristian A. Stevens, Charles H. Langley, and John E. Pool. 2015. “The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila Melanogaster Genomes, Including 197 from a Single Ancestral Range Population.”

Genetics 199 (4): 1229–41.

Lack, Justin B., Jeremy D. Lange, Alison D. Tang, Russell B. Corbett-Detig, and John E. Pool. 2016. “A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus.” *Molecular Biology and Evolution* 33 (12): 3308–13.

Lam, Ha Minh, Oliver Ratmann, and Maciej F. Boni. 2018. “Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm.” *Molecular Biology and Evolution* 35 (1): 247–51.

Lam, Tommy Tsan-Yuk, Na Jia, Ya-Wei Zhang, Marcus Ho-Hin Shum, Jia-Fu Jiang, Hua-Chen Zhu, Yi-Gang Tong, et al. 2020. “Identifying SARS-CoV-2-Related Coronaviruses in Malayan Pangolins.” *Nature* 583 (7815): 282–85.

Lanfear, Robert. 2020. *A Global Phylogeny of SARS-CoV-2 Sequences from GISAID*. https://doi.org/10.5281/zenodo.3958883.

Lau, Susanna K. P., Yun Feng, Honglin Chen, Hayes K. H. Luk, Wei-Hong Yang, Kenneth S. M. Li, Yu-Zhen Zhang, et al. 2015. “Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination.” *Journal of Virology*. https://doi.org/10.1128/jvi.01048-15.

Lee, Yong Sun, Yoshiyuki Shibata, Ankit Malhotra, and Anindya Dutta. 2009. “A Novel Class of Small RNAs: tRNA-Derived RNA Fragments (tRFs).” *Genes & Development* 23 (22): 2639–49.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, W. H., D. L. Ellsworth, J. Krushkal, B. H. Chang, and D. Hewett-Emmett. 1996. “Rates of Nucleotide Substitution in Primates and Rodents and the Generation-Time Effect Hypothesis.” *Molecular Phylogenetics and Evolution* 5 (1): 182–87.

Li, Xiaojun, Elena E. Giorgi, Manukumar Honnayakanahalli Marichann, Brian Foley, Chuan Xiao, Xiang-Peng Kong, Yue Chen, Bette Korber, and Feng Gao. n.d. “Emergence of

SARS-CoV-2 through Recombination and Strong Purifying Selection." https://doi.org/10.1126/sciadv.abb9153.

Li, X. Y., A. Virbasius, X. Zhu, and M. R. Green. 1999. "Enhancement of TBP Binding by Activators and General Transcription Factors." *Nature* 399 (6736): 605–9.

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. 2011. "ViennaRNA Package 2.0." *Algorithms for Molecular Biology: AMB* 6 (November): 26.

Lowe, T. M., and S. R. Eddy. 1997. "tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence." *Nucleic Acids Research* 25 (5): 955–64.

Löytynoja, Ari, Albert J. Vilella, and Nick Goldman. 2012. "Accurate Extension of Multiple Sequence Alignments Using a Phylogeny-Aware Graph Algorithm." *Bioinformatics* 28 (13): 1684–91.

Lu, Jing, Louis du Plessis, Zhe Liu, Verity Hill, Min Kang, Huifang Lin, Jiufeng Sun, et al. 2020. "Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China." *Cell* 181 (5): 997–1003.e9.

Lythgoe, Katrina A., Matthew David Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-Cockett, Amy Trebes, Monique Andersson, et al. n.d. "Shared SARS-CoV-2 Diversity Suggests Localised Transmission of Minority Variants." https://doi.org/10.1101/2020.05.28.118992.

Maraia, Richard J., and Tek N. Lamichhane. 2011. "3' Processing of Eukaryotic Precursor tRNAs." *Wiley Interdisciplinary Reviews. RNA* 2 (3): 362–75.

Maraia, R. J., D. J. Kenan, and J. D. Keene. 1994. "Eukaryotic Transcription Termination Factor La Mediates Transcript Release and Facilitates Reinitiation by RNA Polymerase III." *Molecular and Cellular Biology* 14 (3): 2147–58.

Margush, T., and F. R. McMorris. 1981. "Consensusn-Trees." *Bulletin of Mathematical Biology*. https://doi.org/10.1007/bf02459446.

Martin, Darren P., Steven Weaver, Houryiah Tegally, Emmanuel James San, Stephen D.

Shank, Eduan Wilkinson, Jennifer Giandhari, et al. 2021. “The Emergence and Ongoing Convergent Evolution of the N501Y Lineages Coincides with a Major Global Shift in the SARS-CoV-2 Selective Landscape.” *medRxiv : The Preprint Server for Health Sciences*, March. https://doi.org/10.1101/2021.02.23.21252268.

Mason, Paul B., and Kevin Struhl. 2003. “The FACT Complex Travels with Elongating RNA Polymerase II and Is Important for the Fidelity of Transcriptional Initiation In Vivo.” *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.23.22.8323-8333.2003.

Maurano, Matthew T., Sitharam Ramaswami, Gael Westby, Paul Zappile, Dacia Dimartino, Guomiao Shen, Xiaojun Feng, et al. 2020. “Sequencing Identifies Multiple, Early Introductions of SARS-CoV2 to the New York City Region.” *medRxiv : The Preprint Server for Health Sciences*, April. https://doi.org/10.1101/2020.04.15.20064931.

Mavian, Carla, Simone Marini, Mattia Prosperi, and Marco Salemi. n.d. “A Snapshot of SARS-CoV-2 Genome Availability up to 30th March, 2020 and Its Implications.” https://doi.org/10.1101/2020.04.01.020594.

McBroome, Jakob, Bryan Thornlow, Angie S. Hinrichs, Nicola De Maio, Nick Goldman, David Haussler, Russell Corbett-Detig, and Yatish Turakhia. 2021. “A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees.” *bioRxiv*, 2021–2004.

McCarthy, Kevin R., Linda J. Rennick, Sham Nambulli, Lindsey R. Robinson-McCarthy, William G. Bain, Ghady Haidar, and W. Paul Duprex. 2021. “Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape.” *Science* 371 (6534): 1139–42.

McLachlan, Geoffrey, Kim-Anh Do, and Christophe Ambroise. 2005. *Analyzing Microarray Gene Expression Data*. John Wiley & Sons.

Messer, Philipp W. 2009. “Measuring the Rates of Spontaneous Mutation from Deep and Large-Scale Polymorphism Data.” *Genetics* 182 (4): 1219–32.

Meunier, Julien, Frédéric Lemoine, Magali Soumillon, Angélica Liechti, Manuela Weier, Katerina Guschanski, Haiyang Hu, Philipp Khaitovich, and Henrik Kaessmann. 2013.

"Birth and Expression Evolution of Mammalian microRNA Genes." *Genome Research* 23 (1): 34–45.

Minh, Bui Quang, Minh Anh Thi Nguyen, and Arndt von Haeseler. 2013. "Ultrafast Approximation for Phylogenetic Bootstrap." *Molecular Biology and Evolution* 30 (5): 1188–95.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." *Molecular Biology and Evolution* 37 (5): 1530–34.

Mleczko, Anna M., Piotr Celichowski, and Kamilla Bąkowska-Żywicka. 2014. "Ex-Translational Function of tRNAs and Their Fragments in Cancer." *Acta Biochimica Polonica* 61 (2): 211–16.

Molla-Herman, Anahi, Ana Maria Vallés, Carine Ganem-Elbaz, Christophe Antoniewski, and Jean-René Huynh. 2015. "tRNA Processing Defects Induce Replication Stress and Chk2-Dependent Disruption of piRNA Transcription." *The EMBO Journal* 34 (24): 3009–27.

Morel, Benoit, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, et al. 2020. "Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult." *bioRxiv*. https://doi.org/10.1101/2020.08.05.239046.

Moutouh, L., J. Corbeil, and D. D. Richman. 1996. "Recombination Leads to the Rapid Emergence of HIV-1 Dually Resistant Mutants under Selective Drug Pressure." *Proceedings of the National Academy of Sciences of the United States of America* 93 (12): 6106–11.

Müller, Nicola F., Kathryn E. Kistler, and Trevor Bedford. 2021. "Recombination Patterns in Coronaviruses." *bioRxiv : The Preprint Server for Biology*, April. https://doi.org/10.1101/2021.04.28.441806.

Narasimhan, Vagheesh M., Raheleh Rahbari, Aylwyn Scally, Arthur Wuster, Dan Mason, Yali

Xue, John Wright, et al. 2017. “Estimating the Human Mutation Rate from Autozygous Segments Reveals Population Differences in Human Mutational Processes.” *Nature Communications*. https://doi.org/10.1038/s41467-017-00323-y.

Necsulea, Anamaria, and Henrik Kaessmann. 2014. “Evolutionary Dynamics of Coding and Non-Coding Transcriptomes.” *Nature Reviews. Genetics* 15 (11): 734–48.

Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner, and Henrik Kaessmann. 2014. “The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods.” *Nature* 505 (7485): 635–40.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Nguyen, Ngan, Glenn Hickey, Daniel R. Zerbino, Brian Raney, Dent Earl, Joel Armstrong, W. James Kent, David Haussler, and Benedict Paten. 2015. “Building a Pan-Genome Reference for a Population.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 22 (5): 387–401.

NicolaDeMaio, Sergei Pond, Oscar Maclean, Matthew Parker, and Liam Shaw. 2020. “Issues with SARS-CoV-2 Sequencing Data.” Virological. May 5, 2020. http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Nye, Tom M. W. 2008. “Trees of Trees: An Approach to Comparing Multiple Alternative Phylogenies.” *Systematic Biology* 57 (5): 785–94.

Orioli, Andrea, Chiara Pascali, Jade Quartararo, Kevin W. Diebel, Viviane Praz, David Romascano, Riccardo Percudani, et al. 2011. “Widespread Occurrence of Non-Canonical Transcription Termination by Human RNA Polymerase III.” *Nucleic Acids Research* 39 (13): 5499–5512.

Pachetti, Maria, Bruna Marini, Francesca Benedetti, Fabiola Giudici, Elisabetta Mauro, Paola Storici, Claudio Masciovecchio, et al. 2020. “Emerging SARS-CoV-2 Mutation Hot Spots

Include a Novel RNA-Dependent-RNA Polymerase Variant." *Journal of Translational Medicine* 18 (1): 179.

Palazzo, Alexander F., and Eliza S. Lee. 2015. "Non-Coding RNA: What Is Functional and What Is Junk?" *Frontiers in Genetics* 6 (January): 2.

Pan, Tao. 2018. "Modifications and Functional Genomics of Human Transfer RNA." *Cell Research* 28 (4): 395–404.

Paradis, E., and K. Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics* . https://academic.oup.com/bioinformatics/article-abstract/35/3/526/5055127.

Paten, Benedict, Mark Diekhans, Dent Earl, John St John, Jian Ma, Bernard Suh, and David Haussler. 2011. "Cactus Graphs for Genome Comparisons." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 18 (3): 469–81.

Paten, Benedict, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. 2011. "Cactus: Algorithms for Genome Multiple Sequence Alignment." *Genome Research* 21 (9): 1512–28.

Patiño-Galindo, Juan Ángel, Ioan Filip, and Raul Rabadan. 2021. "Global Patterns of Recombination across Human Viruses." *Molecular Biology and Evolution* 38 (6): 2520–31.

Pattabiraman, Chitra, Farhat Habib, P. K. Harsha, Risha Rasheed, Vijayalakshmi Reddy, Prameela Dinesh, Tina Damodar, et al. 2020. "Genomic Epidemiology Reveals Multiple Introductions and Spread of SARS-CoV-2 in the Indian State of Karnataka." *medRxiv*. https://doi.org/10.1101/2020.07.10.20150045.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research: JMLR* 12 (Oct): 2825–30.

Peng, James, Sabrina A. Mann, Anthea M. Mitchell, Jamin Liu, Matthew T. Laurie, Sara Sunshine, Genay Pilarowski, et al. 2021. "Estimation of Secondary Household Attack

Rates for Emergent SARS-CoV-2 Variants Detected by Genomic Surveillance at a Community-Based Testing Site in San Francisco." https://doi.org/10.1101/2021.03.01.21252705.

Phifer-Rixey, Megan, François Bonhomme, Pierre Boursot, Gary A. Churchill, Jaroslav Piálek, Priscilla K. Tucker, and Michael W. Nachman. 2012. "Adaptive Evolution and Effective Population Size in Wild House Mice." *Molecular Biology and Evolution* 29 (10): 2949–55.

Planas, Delphine, Timothée Bruel, Ludivine Grzelak, Florence Guivel-Benhassine, Isabelle Staropoli, Françoise Porrot, Cyril Planchais, et al. 2021. "Sensitivity of Infectious SARS-CoV-2 B.1.1.7 and B.1.351 Variants to Neutralizing Antibodies." *Nature Medicine*, March. https://doi.org/10.1038/s41591-021-01318-5.

Pollard, Katherine S., Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. 2010. "Detection of Nonneutral Substitution Rates on Mammalian Phylogenies." *Genome Research* 20 (1): 110–21.

Ponting, C. 2007. "TreeBeST: Tree Building Guided by Species Tree."

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PLoS ONE* 5 (3): e9490.

Ralph, Peter, Kevin Thornton, and Jerome Kelleher. 2020. "Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes." *Genetics* 215 (3): 779–97.

Rambaut, Andrew. 2012. "FigTree v1. 4."

Rambaut, Andrew, Edward C. Holmes, Verity Hill, Áine O'Toole, J. T. McCrone, Chris Ruis, Louis du Plessis, and Oliver G. Pybus. n.d. "A Dynamic Nomenclature Proposal for SARS-CoV-2 to Assist Genomic Epidemiology." https://doi.org/10.1101/2020.04.17.046086.

Rambaut, Andrew, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, and Oliver G. Pybus. 2020. "A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology."

Nature Microbiology 5 (11): 1403–7.

Rattray, Alexander M. J., and Berndt Müller. 2012. “The Control of Histone Gene Expression.” *Biochemical Society Transactions* 40 (4): 880–85.

Rayko, Mikhail, and Aleksey Komissarov. n.d. “Quality Control of Low-Frequency Variants in SARS-CoV-2 Genomes.” https://doi.org/10.1101/2020.04.26.062422.

Rhesus Macaque Genome Sequencing and Analysis Consortium, Richard A. Gibbs, Jeffrey Rogers, Michael G. Katze, Roger Bumgarner, George M. Weinstock, Elaine R. Mardis, et al. 2007. “Evolutionary and Biomedical Insights from the Rhesus Macaque Genome.” *Science* 316 (5822): 222–34.

Rice, Alan M., Atahualpa Castillo Morales, Alexander T. Ho, Christine Mordstein, Stefanie Mühlhausen, Samir Watson, Laura Cano, Bethan Young, Grzegorz Kudla, and Laurence D. Hurst. n.d. “Evidence for Strong Mutation Bias Towards, and Selection Against, T/U Content in SARS-CoV2: Implications for Attenuated Vaccine Design.” https://doi.org/10.1101/2020.05.11.088112.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518 (7539): 317–30.

Roberts, Douglas N., Allen J. Stewart, Jason T. Huff, and Bradley R. Cairns. 2003. “The RNA Polymerase III Transcriptome Revealed by Genome-Wide Localization and Activity–occupancy Relationships.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (25): 14695–700.

Robinson, D. F., and L. R. Foulds. 1981. “Comparison of Phylogenetic Trees.” *Mathematical Biosciences* 53 (1): 131–47.

Rockett, Rebecca J., Alicia Arnott, Connie Lam, Rosemarie Sadsad, Verlaine Timms, Karen-Ann Gray, John-Sebastian Eden, et al. 2020. “Revealing COVID-19 Transmission in Australia by SARS-CoV-2 Genome Sequencing and Agent-Based Modeling.” *Nature Medicine*, July. https://doi.org/10.1038/s41591-020-1000-7.

Rogers, Hubert H., Casey M. Bergman, and Sam Griffiths-Jones. 2010. “The Evolution of tRNA Genes in Drosophila.” *Genome Biology and Evolution* 2 (July): 467–77.

Ruan, Jue, Heng Li, Zhongzhong Chen, Avril Coghlan, Lachlan James M. Coin, Yiran Guo, Jean-Karim Hériché, et al. 2008. “TreeFam: 2008 Update.” *Nucleic Acids Research* 36 (Database issue): D735–40.

Saini, Natalie, Steven A. Roberts, Joan F. Sterling, Ewa P. Malc, Piotr A. Mieczkowski, and Dmitry A. Gordenin. 2017. “APOBEC3B Cytidine Deaminase Targets the Non-Transcribed Strand of tRNA Genes in Yeast.” *DNA Repair* 53 (May): 4–14.

Sanjuán, Rafael, Miguel R. Nebot, Nicola Chirico, Louis M. Mansky, and Robert Belshaw. 2010. “Viral Mutation Rates.” *Journal of Virology*. https://doi.org/10.1128/jvi.00694-10.

Sankoff, David. 1975. “Minimal Mutation Trees of Sequences.” *SIAM Journal on Applied Mathematics*. https://doi.org/10.1137/0128004.

Sayers, Eric W., Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Conrad L. Schoch, Stephen T. Sherry, and Ilene Karsch-Mizrachi. 2021. “GenBank.” *Nucleic Acids Research* 49 (D1): D92–96.

Schaffer, Ashleigh E., Veerle R. C. Eggens, Ahmet Okay Caglayan, Miriam S. Reuter, Eric Scott, Nicole G. Coufal, Jennifer L. Silhavy, et al. 2014. “CLP1 Founder Mutation Links tRNA Splicing and Maturation to Cerebellar Development and Neurodegeneration.” *Cell* 157 (3): 651–63.

Schierup, M. H., and J. Hein. 2000. “Consequences of Recombination on Traditional Phylogenetic Analysis.” *Genetics* 156 (2): 879–91.

Schimmel, Paul. 2018. “The Emerging Complexity of the tRNA World: Mammalian tRNAs beyond Protein Synthesis.” *Nature Reviews. Molecular Cell Biology* 19 (1): 45–58.

Schmidt, Steffen, Anna Gerasimova, Fyodor A. Kondrashov, Ivan A. Adzhubei, Ivan A. Adzuhbei, Alexey S. Kondrashov, and Shamil Sunyaev. 2008. “Hypermutable Non-Synonymous Sites Are under Stronger Negative Selection.” *PLoS Genetics* 4 (11): e1000281.

Schmitt, Bianca M., Konrad L. M. Rudolph, Panagiota Karagianni, Nuno A. Fonseca, Robert J. White, Iannis Talianidis, Duncan T. Odom, John C. Marioni, and Claudia Kutter. 2014. “High-Resolution Mapping of Transcriptional Dynamics across Tissue Development Reveals a Stable mRNA–tRNA Interface.” *Genome Research* 24 (11): 1797–1807.

Sethi, Anurag, Mengting Gu, Emrah Gumusgoz, Landon Chan, Koon-Kiu Yan, Joel Rozowsky, Iros Barozzi, et al. 2018. “A Cross-Organism Framework for Supervised Enhancer Prediction with Epigenetic Pattern Recognition and Targeted Validation.” *bioRxiv*. https://doi.org/10.1101/385237.

Shapiro, Joshua A., Wei Huang, Chenhui Zhang, Melissa J. Hubisz, Jian Lu, David A. Turissini, Shu Fang, et al. 2007. “Adaptive Genic Evolution in the Drosophila Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (7): 2271–76.

Shu, Y., and J. McCauley. 2017. “GISAID: Global Initiative on Sharing All Influenza Data - from Vision to Reality.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 22 (13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Shu, Yuelong, and John McCauley. 2017. “GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality.” *Eurosurveillance*. https://doi.org/10.2807/1560-7917.es.2017.22.13.30494.

Simon, Chris. 2020. “An Evolving View of Phylogenetic Support.” *Systematic Biology*, September. https://doi.org/10.1093/sysbio/syaa068.

Singer, Joshua, Robert Gifford, Matthew Cotten, and David Robertson. n.d. “CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation.” https://doi.org/10.20944/preprints202006.0225.v1.

Sprinzl, M., C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. 1998. “Compilation of tRNA Sequences and Sequences of tRNA Genes.” *Nucleic Acids Research* 26 (1): 148–53.

Starr, Tyler N., Allison J. Greaney, Amin Addetia, William W. Hannon, Manish C. Choudhary,

Adam S. Dingens, Jonathan Z. Li, and Jesse D. Bloom. 2021. “Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat COVID-19.” *Science* 371 (6531): 850–54.

Starr, Tyler N., Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, et al. 2020. “Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.” *Cell* 182 (5): 1295–1310.e20.

Stefanelli, Paola, Giovanni Faggioni, Alessandra Lo Presti, Stefano Fiore, Antonella Marchi, Eleonora Benedetti, Concetta Fabiani, et al. 2020. “Whole Genome and Phylogenetic Analysis of Two SARS-CoV-2 Strains Isolated in Italy in January and February 2020: Additional Clues on Multiple Introductions and Further Circulation in Europe.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 25 (13). https://doi.org/10.2807/1560-7917.ES.2020.25.13.2000305.

Stephens, J. C. 1986. “On the Frequency of Undetectable Recombination Events.” *Genetics* 112 (4): 923–26.

Sukumaran, Jeet, and Mark T. Holder. 2010. “DendroPy: A Python Library for Phylogenetic Computing.” *Bioinformatics* 26 (12): 1569–71.

Sun, Chunxiao, Ziyi Fu, Siwei Wang, Jun Li, Yongfei Li, Yanhong Zhang, Fan Yang, et al. 2018. “Roles of tRNA-Derived Fragments in Human Cancers.” *Cancer Letters* 414 (February): 16–25.

Surleac, Marius, Leontina Banica, Corina Casangiu, Marius Cotic, Dragos Florea, Oana Sandulescu, Petre Milu, et al. 2020. “Molecular Epidemiology Analysis of SARS-CoV-2 Strains Circulating in Romania during the First Months of the Pandemic.” *Life* 10 (8). https://doi.org/10.3390/life10080152.

Suzuki, Tsutomu, Asuteka Nagao, and Takeo Suzuki. 2011. “Human Mitochondrial tRNAs: Biogenesis, Function, Structural Aspects, and Diseases.” *Annual Review of Genetics* 45

(September): 299–329.

Taghizadeh, Peyman, Sadegh Salehi, Ali Heshmati, Seyed Massoud Houshmand, Kolsoum InanlooRahatloo, Forouzandeh Mahjoubi, Mohammad Hossein Sanati, et al. 2021. “Study on SARS-CoV-2 Strains in Iran Reveals Potential Contribution of Co-Infection with and Recombination between Different Strains to the Emergence of New Strains.” *Virology* 562 (October): 63–73.

Tang, Dave T. P., Evgeny A. Glazov, Sean M. McWilliam, Wesley C. Barris, and Brian P. Dalrymple. 2009. “Analysis of the Complement and Molecular Evolution of tRNA Genes in Cow.” *BMC Genomics* 10 (April): 188.

Taylor, Benjamin J. M., Yee Ling Wu, and Cristina Rada. 2014. “Active RNAP Pre-Initiation Sites Are Highly Mutated by Cytidine Deaminases in Yeast, with AID Targeting Small RNA Genes.” *eLife* 3 (September): e03553.

Tenesa, Albert, Pau Navarro, Ben J. Hayes, David L. Duffy, Geraldine M. Clarke, Mike E. Goddard, and Peter M. Visscher. 2007. “Recent Human Effective Population Size Estimated from Linkage Disequilibrium.” *Genome Research* 17 (4): 520–26.

Thielen, Peter M., Shirlee Wohl, Thomas Mehoke, Srividya Ramakrishnan, Melanie Kirsche, Oluwaseun Falade-Nwulia, Nidia S. Trovao, et al. 2020. “Genomic Diversity of SARS-CoV-2 During Early Introduction into the United States National Capital Region.” *medRxiv : The Preprint Server for Health Sciences*, August. https://doi.org/10.1101/2020.08.13.20174136.

Thomson, Emma C., Laura E. Rosen, James G. Shepherd, Roberto Spreafico, Ana da Silva Filipe, Jason A. Wojcechowskyj, Chris Davis, et al. 2021. “Circulating SARS-CoV-2 Spike N439K Variants Maintain Fitness While Evading Antibody-Mediated Immunity.” *Cell* 184 (5): 1171–87.e20.

Thornlow, Bryan, Angie S. Hinrichs, Miten Jain, Namrita Dhillon, Scott La, Joshua D. Kapp, Ikenna Anigbogu, et al. 2021. “A New SARS-CoV-2 Lineage That Shares Mutations with Known Variants of Concern Is Rejected by Automated Sequence Repository Quality

Control.” *bioRxiv.org: The Preprint Server for Biology*, April. https://doi.org/10.1101/2021.04.05.438352.

Thornlow, Bryan, Josh Hough, Jackie Roger, Henry Gong, Todd Lowe, and Russell Corbett-Detig. 2018. “Transfer RNA Genes Experience Exceptionally Elevated Mutation Rates.” *Proceedings of the National Academy of Sciences* 115 (36): 8996–9001.

Thul, Peter J., Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, et al. 2017. “A Subcellular Map of the Human Proteome.” *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.

Timakov, Benjamin, Xiaoru Liu, Ismail Turgut, and Ping Zhang. 2002. “Timing and Targeting of P-Element Local Transposition in the Male Germline Cells of Drosophila Melanogaster.” *Genetics* 160 (3): 1011–22.

Turakhia, Yatish, Bryan Thornlow, Landen Gozashti, Angie S. Hinrichs, Jason D. Fernandes, David Haussler, and Russell Corbett-Detig. 2020. “Stability of SARS-CoV-2 Phylogenies.” *PLoS Genetics*, November. https://doi.org/10.1371/journal.pgen.1009175.

Turakhia, Yatish, Bryan Thornlow, Angie S. Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-Detig. 2020. “Ultrafast Sample Placement on Existing Trees (UShER) Empowers Real-Time Phylogenetics for the SARS-CoV-2 Pandemic.” *bioRxiv : The Preprint Server for Biology*, September. https://doi.org/10.1101/2020.09.26.314971.

VanInsberghe, David, Andrew S. Neish, Anice C. Lowen, and Katia Koelle. 2021. “Recombinant SARS-CoV-2 Genomes Are Currently Circulating at Low Levels.” *bioRxiv : The Preprint Server for Biology*, March. https://doi.org/10.1101/2020.08.05.238386.

Varabyou, Ales, Christopher Pockrandt, Steven L. Salzberg, and Mihaela Pertea. 2021. “Rapid Detection of Inter-Clade Recombination in SARS-CoV-2 with Bolotie.” *Genetics*, May. https://doi.org/10.1093/genetics/iyab074.

Volz, Erik, Verity Hill, John T. McCrone, Anna Price, David Jorgensen, Áine O’Toole, Joel Southgate, et al. 2021. “Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on

Transmissibility and Pathogenicity." *Cell* 184 (1): 64–75.e11.

Volz, Erik, Swapnil Mishra, Meera Chand, Jeffrey C. Barrett, Robert Johnson, Lily Geidelberg, Wes R. Hinsley, et al. 2021. "Assessing Transmissibility of SARS-CoV-2 Lineage B.1.1.7 in England." *Nature*, March, 1–17.

Washington, Nicole L., Karthik Gangavarapu, Mark Zeller, Alexandre Bolze, Elizabeth T. Cirulli, Kelly M. Schiabor Barrett, Brendan B. Larsen, et al. 2021. "Genomic Epidemiology Identifies Emergence and Rapid Transmission of SARS-CoV-2 B.1.1.7 in the United States." *medRxiv : The Preprint Server for Health Sciences*, February. https://doi.org/10.1101/2021.02.06.21251159.

White, Robert J. 2011. "Transcription by RNA Polymerase III: More Complex than We Thought." *Nature Reviews. Genetics* 12 (7): 459–63.

Wu, Y. 2020. "Strong Evolutionary Convergence of Receptor-Binding Protein Spike between COVID-19 and SARS-Related Coronaviruses." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.03.04.975995v1.abstract.

Xia, Xuhua. 2020. "Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense." *Molecular Biology and Evolution*, April. https://doi.org/10.1093/molbev/msaa094.

Yi, Huiguang. 2020. "2019 Novel Coronavirus Is Undergoing Active Recombination." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, March. https://doi.org/10.1093/cid/ciaa219.

Yoo, Hyunjin, Dabin Son, Young-Joo Jang, and Kwonho Hong. 2016. "Indispensable Role for Mouse ELP3 in Embryonic Stem Cell Maintenance and Early Development." *Biochemical and Biophysical Research Communications* 478 (2): 631–36.

Zanton, Sara J., and B. Franklin Pugh. 2004. "Changes in Genomewide Occupancy of Core Transcriptional Regulators during Heat Stress." *Proceedings of the National Academy of Sciences of the United States of America* 101 (48): 16843–48.

Zhang, Jinwei, and Adrian R. Ferré-D'Amaré. 2016. "The tRNA Elbow in Structure,

Recognition and Evolution.” *Life* 6 (1). https://doi.org/10.3390/life6010003.

Zheng, Guanqun, Yidan Qin, Wesley C. Clark, Qing Dai, Chengqi Yi, Chuan He, Alan M. Lambowitz, and Tao Pan. 2015. “Efficient and Quantitative High-Throughput tRNA Sequencing.” *Nature Methods* 12 (9): 835–37.

Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. “A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin.” *Nature* 579 (7798): 270–73.

Ziehler, William A., Jeremy J. Day, Carol A. Fierke, and David R. Engelke. 2000. “Effects of 5‘ Leader and 3‘ Trailer Structures on Pre-tRNA Processing by Nuclear RNase P.” *Biochemistry* 39 (32): 9909–16.

Zinzula, Luca. 2021. “Lost in Deletion: The Enigmatic ORF8 Protein of SARS-CoV-2.” *Biochemical and Biophysical Research Communications* 538 (January): 116–24.

Zuckerman, Neta S., Shay Fleishon, Efrat Bucris, Dana Bar-Ilan, Michal Linial, Itay Bar-Or, Victoria Indenbaum, et al. 2021. “A Unique SARS-CoV-2 Spike Protein P681H Strain Detected in Israel.” *medRxiv*. https://www.medrxiv.org/content/10.1101/2021.03.25.21253908v1.abstract.