UNIVERSITY OF CALIFORNIA

SANTA CRUZ

PHYLOGENETICS IN THE PANDEMIC ERA

A dissertation submitted in partial satisfaction

of the requirements for the degree

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Jakob McBroome

June 2023

The Dissertation of Jakob McBroome is approved:

_____
Professor Russell Corbett-Detig, advisor

_____
Professor Ed Green, chair

_____
Professor Chris Vollmers

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

TABLE OF CONTENTS

List of Figures

List of Tables

Abstract

PHYLOGENETICS IN THE PANDEMIC ERA

Jakob D. McBroome

The COVID-19 pandemic of 2020 was one of the first major global public health crises in the post-genomic era, inspiring truly unprecedented levels of viral genome sequencing. In the realm of phylogenetics, or the reconstruction of ancestral relationships between extant sequences, essentially no software existed capable of handling the full dataset in a timely and effective manner. Phylogenetics is critical for the identification and tracking of major variants, particularly the famous Variants of Concern (VOC), leading to a desperate need for scalable tools. I, along with several collaborators, developed an efficient toolkit for the construction, manipulation, and analysis of massive phylogenetic trees. Our core data structure, the mutation annotated tree (MAT), is capable of storing millions of SARS-CoV-2 genomes in less than a gigabyte of data. My key contribution was the development of matUtils, a C++ library and command line toolkit to manipulate these highly compact data files. I additionally developed BTE, a highly efficient API making our phylogenetics software available in a Python environment. I subsequently developed analytical approaches taking advantage of these new tools with the availability and massive scale of the SARS-CoV-2 data. Among these is scalable phylogeographic inference, through the daily-updated website Cluster-Tracker. Cluster-Tracker uses a simple heuristic I developed to efficiently identify and present local transmission clusters for

public health track-and-trace efforts. I also designed an approach to the identification of novel SARS-CoV-2 strains and integrated it with the popular Pango lineage system. Altogether, this dissertation presents a body of work contributing substantially to effective global public health response to the SARS-CoV-2 pandemic.

Acknowledgements

This work would not have been possible without the contribution and support of many people. First, I would like to thank my fellow developers and network of collaborators. The UShER development group, with whom I worked closely on the work presented here, includes Rusell Corbett-Detig, Yatish Turakhia, Angie Heinrichs, Alex Kramer, Cheng Ye, Lily Karim, and others. I also want to thank our extended collaborators in SARS-CoV-2 genomic epidemiology, including Aine O'Toole, Chris Ruis, Cornelius Roemer, Andrew Rambaut, Nicola Di Maio, and Theo Sanderson, who have provided invaluable feedback, discussion, and integration with their own research and toolkits.

Of course, I am deeply grateful for the everyday support and conversation with members of the Corbett-Detig lab, including Russ himself, Alex Kramer, Lily Karim, Erik Enbody, Chris Condon, Maximillian Genetti, Mara Baylis, Cade Mirchandani, Nicolas Ayala, Adriano de Bernadi Schneider, Gabriel Penuri, and Nick Chan. Their vibrant lab culture and supportive environment eased the isolation and struggle of earning a PhD.

Finally, I would like to thank my parents and my brother. Without their guidance and fostering of my curiosity from a young age, I would never have been set on the path which led to this accomplishment. Their love has been a rock on which I could rely in challenging times, and I would never have made it this far without them.

The text of this dissertation includes reprints of the following previously published materials:

1. McBroome, J., Liang, D. & Corbett-Detig, R. Fine-Scale Position Effects Shape the Distribution of Inversion Breakpoints in Drosophila melanogaster. Genome Biology and Evolution 12, 1378–1391 (2020).
2. McBroome, J., Martin, J., de Bernardi Schneider, A., Turakhia, Y. & Corbett-Detig, R. Identifying SARS-CoV-2 regional introductions and transmission clusters in real time. Virus Evolution 8, veac048 (2022).
3. McBroome, J. et al. A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. Molecular Biology and Evolution (2021) doi:10.1093/molbev/msab264.
4. McBroome, J., Turakhia, Y. & Corbett-Detig, R. BTE: a Python module for pandemic-scale mutation-annotated phylogenetic trees. Journal of Open Source Software 7, 4433 (2022).
5. McBroome, J. et al. Automated Agnostic Designation of Pathogen Lineages. 2023.02.03.527052 Preprint at https://doi.org/10.1101/2023.02.03.527052 (2023).

The co-authors listed in these publications directed and supervised the research which forms the basis for the dissertation.

Introduction

The COVID-19 pandemic was, in many ways, the first major public health crisis of the post-genomic era. Tens of millions of SARS-CoV-2 genomes were sequenced between 2020 and 2023 from all across the world, representing a more complete picture of its genetic diversity than had ever been seen for any other pathogen. While the scale of the data created great opportunities for analysis, it also stressed existing toolkits and software beyond the breaking point. Tools developed for academic, intensive analysis of a relative handful of samples were unable to scale to the incredible surge of sequence data. Fixing the so-called "bioinformatics bottleneck" was one of the greatest challenges presented by the SARS-CoV-2 pandemic, and it is to this challenge that this dissertation is dedicated.

When the pandemic began, I largely abandoned my original research plan to instead address this unprecedented crisis. Here I present a set of software tools and analytical approaches I developed to address the massive scale of SARS-CoV-2 sequencing data. This work underlies much of SARS-CoV-2's global track-and-trace infrastructure, including the popular Pango lineage system, and BigTree, a dashboard used by the California Department of Public Health. The impact of this work on the global response to SARS-CoV-2 cannot be overstated.

In the first chapter, I review the groundwork innovation that supports the rest of the work- the mutation annotated tree (MAT) data structure. The MAT represents an extremely efficient representation of genome sequence data, capable of storing millions of viral genomes in less than a gigabyte of data. While I did not design the

MAT or the tool which constructs them (UShER), I wrote software for the conversion, manipulation, and analysis of these data structures (matUtils). The MAT, as a relatively new innovation, is not commonly supported by other analysis software. The development of an efficient toolkit for manipulating, extracting, and converting MATs to other formats was critical to drive the adoption of the MAT as the gold standard for SARS-CoV-2 genome data sharing. matUtils provides command line access for the use of this file type, enabling a wide variety of other tools and analyses, including accurate Pango lineage assignments. I additionally developed BTE, a Cython-based API which exposed our efficient libraries for MAT analysis to Python. BTE allows for sophisticated analysis of global SARS-CoV-2 genome datasets in Python and serves as a critical tool for prototyping, development, and research with SARS-CoV-2 data. This toolkit for working with MATs is critical to both the following chapters and to the research and analysis of many other groups globally.

After creating the basic toolkit, I developed scalable analytical methods for the MAT datasets. In my second chapter, I present a heuristic method for phylogeographic analysis. Phylogeography, or the analysis of a phylogeny with respect to a geography, has been generally done with complex modeling software such as BEAST. These methods, which are designed to extrapolate highly informed models from relatively little data, require so much time and compute power when applied to large datasets that their use is simply not feasible for most public health groups. Public health groups needed a fast approximate method to continuously update them as to local transmission dynamics and inform any immediate public

health actions. My approach is a heuristic, designed to scale linearly with the size of the phylogeny, that identifies clusters of samples representing a regionally circulating strain for any number of regions. It is capable of identifying thousands of distinct infection clusters from millions of samples across dozens of regions in less than two hours, making it capable of producing nation-wide reports on the latest data on a daily basis. I therefore created a website to do exactly this, called ClusterTracker. ClusterTracker is a simple one-page site that displays a table of clusters identified across each of the fifty United States and an accompanying map colored by interstate transmission dynamics, updated daily. My method and website code was adopted by the California Department of Public Health (CDPH), who constructed a county-level version of the site that produces daily reports of transmission dynamics within California. This is just one way our MAT structure and surrounding code supports the analysis of massive genomic datasets for public health action.

In the third chapter, I turn my attention to the identification and tracking of SARS-CoV-2 lineages. The Pango nomenclature system underlies global understanding of SARS-CoV-2's evolution. Pango lineages, such as B.1.1.7, were commonly reported in the media and were generally the first names applied to emerging Variants of Concern (VOC) before they were formally named. UShER and the MAT format were quickly adopted by the Pango group as a stable alternative to their random forest approach for the assigning of samples to existing Pango labels. The actual designation process of Pango lineages remained unfortunately ad hoc, however. New Pango lineages were largely identified through crowdsourced

proposals, as individual researchers and epidemiologists visualized the data, picked out strains they thought were potentially worth naming, and made arguments in Github issue threads. This process is time-consuming, dependent on individual bias, and not sustainable as researchers and experts move on to the next stages in their careers. I sought to automate Pango lineage designation, at least in part, by automatically identifying lineage candidates for review from the latest data. My approach is based in information theory, attempting to create lineage labels that represent as much genetic information as possible. I included many additional options and weighting schema to control the lineage designation system. For example, the user can strongly emphasize the representation of mutations known to be associated with increased immune evasion, or of samples from underrepresented parts of the world. In collaboration with the Pango team, I integrated my designation pipeline into their designation infrastructure and started producing automated lineage designations for review. This step streamlined SARS-CoV-2 track and trace at one of the most fundamental levels, preparing our infrastructure for a future of lower funding and less researcher hours invested. I additionally generalized my core method for lineage designation, making it an applicable system for any research group dealing with a novel pathogen, or for future pathogens with a profusion of sequencing data to sort through.

This dissertation represents a substantial body of work contributing to the global response to SARS-CoV-2. I developed scalable tools and methods to open the bioinformatics bottleneck, leveraging the massive global sequencing effort to inform

public health action. Over the course of my dissertation I presented to local public health offices, collaborated with a variety of research groups and developers, and was part of a major effort to understand and interpret the ongoing evolution of SARS-CoV-2. My work will serve evolutionary analysis and tracking for SARS-CoV-2, and for other pathogens, for years to come.

Chapter 1

A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees

[This chapter has been adapted from publication, "A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees" (McBroome et al 2021, Molecular Biology and Evolution]

1.1: Background

The COVID-19 pandemic has inspired unprecedented levels of genome sequencing for a single pathogen (Hodcroft et al. 2021). Over a million SARS-CoV-2 genomes have been sequenced worldwide so far, and tens of thousands of new genomes are getting uploaded daily (Maxmen 2021). This data has enabled scientists to closely track the evolution and transmission dynamics of the virus at global and local scales (Deng et al. 2020; Chaillon and Smith 2021; da Silva Filipe et al. 2021). However, the scale of this data is posing serious computational challenges for comprehensive phylogenetic analyses (Hodcroft et al. 2021). Platforms like Nextstrain (Hadfield et al. 2018) have been invaluable in studying viral transmission networks and genomic surveillance efforts, but they only provide subsampled

SARS-CoV-2 trees consisting of a tiny fraction of available data, omitting phylogenetic relationships with most available sequences. A single, comprehensive SARS-CoV-2 reference tree of all available data could not only facilitate detailed and unambiguous phylogenetic analyses at global, country and local levels, but may also help promote consistency of results across different research groups (Turakhia et al. 2020).

The massive volume of SARS-CoV-2 data also poses numerous data sharing challenges with existing file formats, such as Fasta or Variant Call Format (VCF), which are bulky and necessitate network speeds and computational capabilities that are beyond the reach of many research and scientific groups.

1.2: New Approaches

In this work, we simultaneously address the issue of maintaining a comprehensive SARS-CoV-2 reference tree and its associated data processing, sharing and analysis challenges. Specifically, we are maintaining and openly sharing a daily-updated database of mutation-annotated trees (MATs) containing global SARS-CoV-2 sequences from public databases, without any downsampling (other than for quality control, see 1.8), including annotations for Nextstrain clades (Hadfield et al. 2018) and Pango lineages (Rambaut et al. 2020) (Figure A1.1). The MAT is an extremely efficient data format proposed recently (Turakhia et al. 2021) which uses a form of phylogenetic compression (Ané and Sanderson 2005) to facilitate sharing of extremely large genome sequence datasets. An uncompressed

MAT of 834,521 SARS-CoV-2 public sequences requires only 65 MB to store and encodes more information than in a 43 GB VCF containing single-nucleotide variation of all sequences (the MAT format does not handle insertions and deletions (Turakhia et al. 2021)) and a 38 MB Newick file containing the phylogenetic tree topology.

To accompany this database, we present 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.

1.3: A daily-updated mutation-annotated tree database of global SARS-CoV-2 sequences

To aid the scientific community studying the mutational and transmission dynamics of the SARS-CoV-2 virus and its different variants, we are maintaining a daily-updated database of SARS-CoV-2 mutation-annotated trees (MATs) composed of public data. Starting with the final Newick tree release dated November 13, 2020,

of Rob Lanfear's sarscov2phylo (https://github.com/roblanf/sarscov2phylo) that is re-rooted to Wuhan/Hu-1 (GenBank MN908947.3, RefSeq NC_045512.2), we have set up an automated pipeline to aggregate public sequences available through GenBank (Clark et al. 2007), COG-UK (Nicholls et al. 2020), and the China National Center for Bioinformation on a daily basis and incorporate them into our MAT using UShER (see Section 1.8). GISAID data (Shu and McCauley 2017) is not included in our MATs because its usage terms do not allow redistribution. Similar to GISAID, our database is subject to the sampling bias resulting from the vast disparity in the sequencing efforts of various countries (Cyranoski 2021, Figure A1.1B). We also use the matUtils annotate command (Appendix A1.1) to add Nextstrain clade and Pango lineage annotations to individual branches of our MAT. As of June 9, 2021, our MAT consists of 834,521 sequences, includes 14 Nextstrain clade and 895 Pango lineage annotations for all samples, and is only 65 MB, or 14 MB when gzip-compressed (Figure A1.1, Table A1.1). To our knowledge, this is the most comprehensive representation of the SARS-CoV-2 evolutionary history using publicly available sequences as of June 9, 2021. It can be freely used to study evolutionary and transmission dynamics of the virus at global, country and local levels, and can be visualized using the Cov2Tree tool (https://cov2tree.org/) developed by Theo Sanderson.

1.4: matUtils provides a wide range of functions to analyze and manipulate mutation-annotated trees

We have created a high-performance command line utility called matUtils for performing a wide range of operations on MATs for rapid interpretation and analysis in genomic surveillance and contact tracing efforts. matUtils is distributed with the UShER package (Turakhia et al. 2021) and uses the same mutation-annotated tree (MAT) format as UShER. matUtils is organized into five different subcommands: annotate, summary, extract, uncertainty and introduce (Figure 1.1), described briefly below. We provide detailed instructions for the usage of each module on our wiki (https://usher-wiki.readthedocs.io/en/latest/matUtils.html).

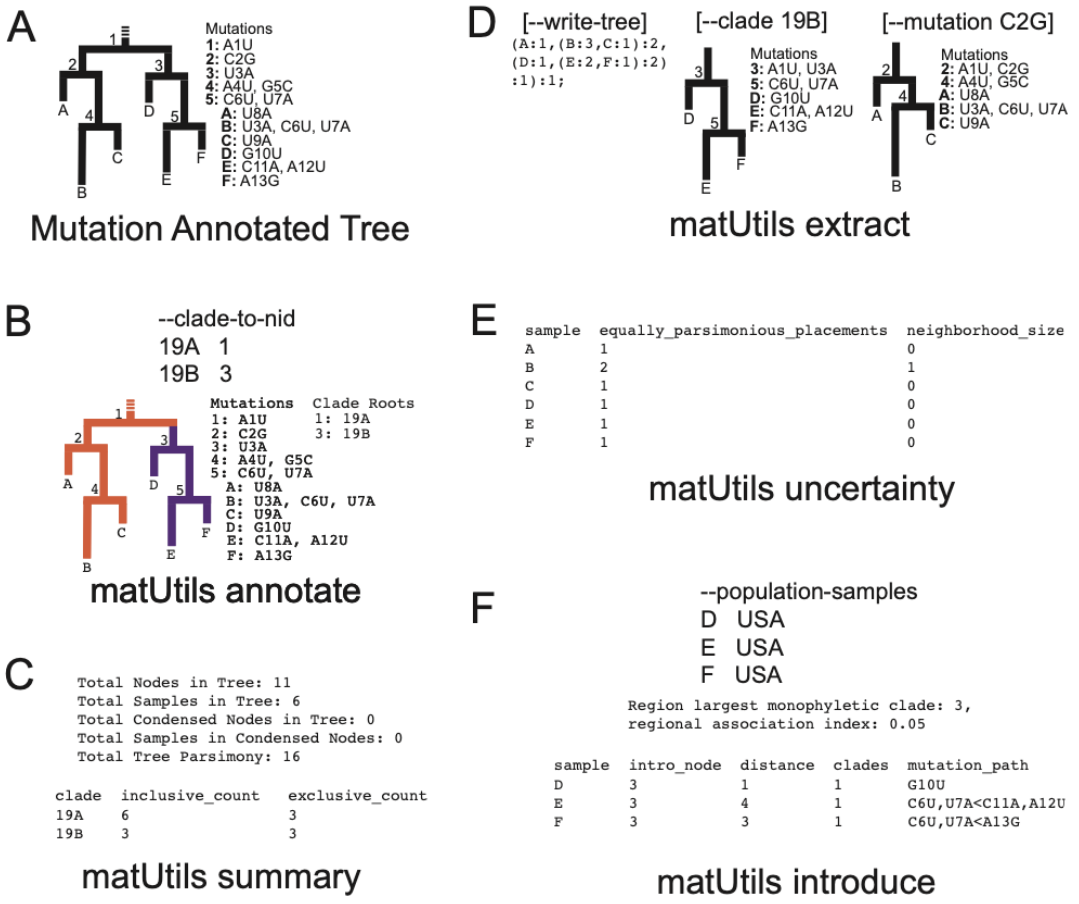

Figure 1.1: **matUtils functions enable fast, user-friendly analysis of mutation-annotated trees (MATs).** (A): An example MAT with tree topology corresponding to the MAT on the left and the mutation annotations on each node shown on the right. (B): matUtils annotate allows the user to annotate internal nodes with clade names. In this example, nodes 1 and 3 are annotated with clade names 19A and 19B, respectively. This MAT serves as an input to commands shown in panels C-F. (C): matUtils summary outputs sample-, clade-, and tree-level statistics for the input MAT. (D): matUtils extract allows users to convert a MAT to Newick format (left), subset the MAT for a specified clade (center) or mutation (right), among other functions. (E): matUtils uncertainty outputs parsimony scores, equally parsimonious placements and neighborhood sizes for each sample of an input MAT. Sample B has two equally parsimonious placements, as it could also be placed as a descendant of node 5 with terminal mutations C2G, A4U, and G5C. (F): matUtils introduce can take a list of samples of interest as input and output the largest monophyletic clade and regional association index associated with the input population , along with their predicted introduction nodes and paths. In all panels, user input commands are shown in large fonts (e.g. "matUtils annotate") and output text from these commands are shown in monospaced fonts.

matUtils Annotate

This function annotates clades in a MAT. One of the central uses of phylogenetics during the pandemic is to trace the emergence and spread of new viral lineages. Nextstrain (Hadfield et al. 2018), Pango (Rambaut et al. 2020) and GISAID (Shu and McCauley 2017) provide different nomenclatures for SARS-CoV-2 variants that have been used widely in genomic surveillance. Our MAT format provides the ability to annotate internal branches of the tree with an array of clade names, one for each clade nomenclature. matUtils annotate provides two methods for annotation: (i) directly providing the mappings of each clade name to its corresponding node or (ii) providing a set of representative sample names for each clade from which the clade roots can be automatically inferred (see Section 1.8). Both methods ensure that the clades remain monophyletic, but we use the second approach to label Nextstrain clades and Pango lineages in our SARS-CoV-2 MAT database since it can be automated using available data (see Section 1.8). matUtils annotate has high congruence with Nextstrain clades and Pango lineage annotations (Table A1.1).

Once clades are annotated on a MAT, the UShER placement tool (Turakhia et al. 2021) can assign each newly placed sequence to its corresponding Pango lineage. This is being used as a feature in Pangolin 3.0 (https://github.com/cov-lineages/pangolin/releases/tag/v3.0) to perform clade assignments in a fully phylogenetic framework.

matUtils Summary

This function provides a brief summary of the available data in the input MAT file and is meant to serve as a typical first step in any MAT-based analysis. It provides a count of the total number of samples in the MAT, the size of each annotated clade, the total parsimony score (i.e. the sum of mutation events on all branches of the MAT), the number of distinct mutations, phylogenetically-informed translation of mutations, and other similar statistics.

matUtils Extract

Many SARS-CoV-2 phylodynamic studies involve restricting the analysis to a smaller tree of interest. While it can be computationally challenging to identify samples most closely related to a given sample or cluster from over a million other sequences, it is straightforward to retrieve subtrees from a comprehensive phylogeny. matUtils extract provides an efficient and robust suite of options for subtree selection from a MAT. A user can use matUtils extract to subsample a MAT to find samples that contain a mutation of interest, are members of a specific clade, have a name matching a specific regular expression pattern (such as the expression "(IND*|India*)" to select samples from India), among other criteria (see Section 1.8). matUtils extract also includes options to identify from a MAT sequences which have descended from long internal branches in the tree, which can sometimes arise from recombination (Jackson et al. 2021; Turkahia et al. 2021), or those with an unusually high parsimony score, which are indicative of low-quality sequences (Mai and

Mirarab 2018). Notably, matUtils extract can produce an output Auspice v2 JSON that is compatible with the Auspice tree visualization tool (Hadfield et al. 2018) (Figure 1.2, see 1.8). matUtils extract can also convert a MAT into other file formats, such as a Newick for its corresponding phylogenetic tree and a VCF for its corresponding genome variation data. matUtils extract also provides an option to resolve all polytomies in a MAT arbitrarily, similar to the muti2di functionality in ape (Paradis and Schliep 2019), for compatibility with phylogenetic tools that do not allow polytomies.

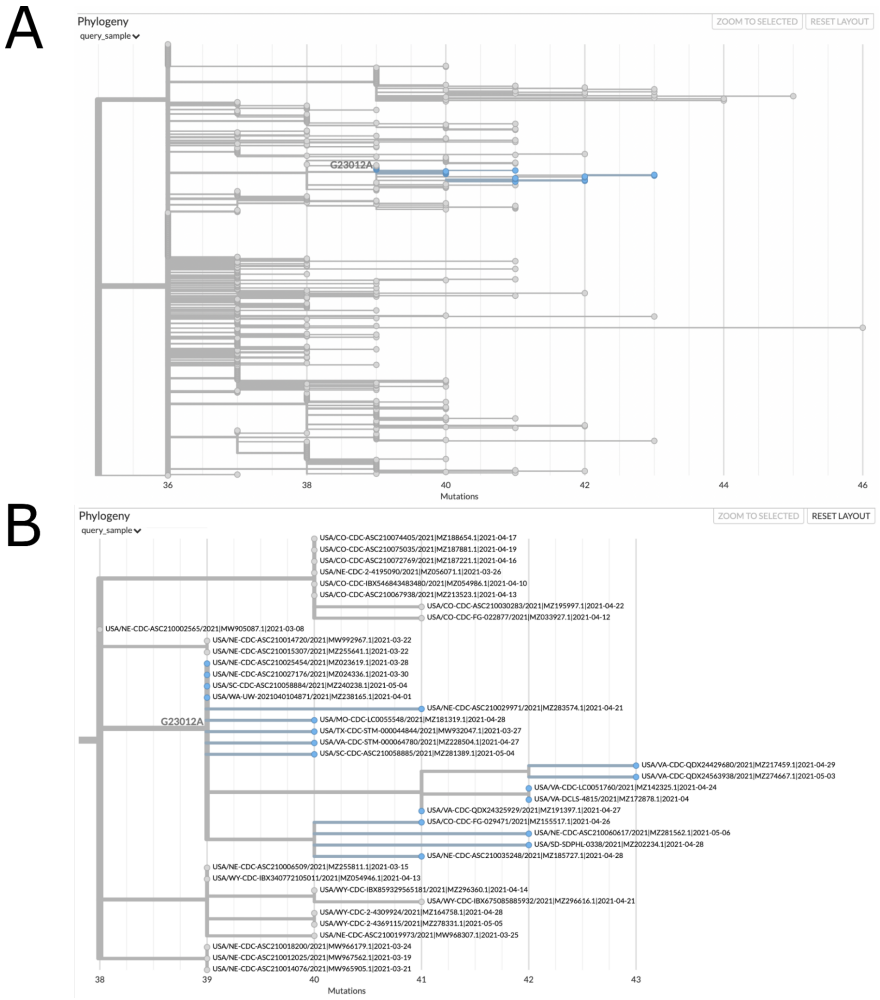

Figure 1.2: matUtils can generate informative visuals with Auspice. The above trees represent a clade of related B.1.1.7 samples from the USA which secondarily acquired the potentially important spike protein mutation E484K, which is caused by the nucleotide mutation G23012A. These trees were obtained by running the command “matUtils extract -i public-2021-06-09.all.masked.nextclade.pangolin.pb.gz -c B.1.1.7 -m G23012A -H "(USA.*)" -N 500 -j clade_trees -d clade_out”, which selects all samples from clade B.1.1.7 which acquired this mutation and are from the USA, then identifies the minimum set of five hundred sample subtrees which contain all of these samples, creating an Auspice v2 format JSON for each subtree (Hadfield et al 2018). This results in thirty-five distinct subtree JSON files of five hundred samples each in the output directory. Panel A represents the entirety of subtree six as viewed with Auspice (Hadfield et al 2018), including blue highlights and a branch label where our mutation of interest occurred. Panel B is zoomed in on this subtree and its sister clade; at this scale we can read individual sample names and observe that this specific strain has been actively spreading in the United States during April 2021.

matUtils Uncertainty

A fundamental concern in SARS-CoV-2 phylogenetics is topological uncertainty (Hodcroft et al. 2021), which may result from contaminated sequences or sample mixtures (Turakhia et al. 2021). The impact of this concern depends on the biological context of the analysis. matUtils uncertainty provides a topological uncertainty statistic which computes the number of equally parsimonious placements that exist for each specified sample in the input MAT. Importantly, matUtils also allows the user to calculate equally parsimonious positions for already placed samples. This is accomplished by pruning the sample from the tree and placing the sample back to the tree using the placement module of UShER (Turakhia et al. 2021) (see Section 1.8). matUtils uncertainty additionally records the number of mutations separating the two most distant equally parsimonious placements, reflecting the distribution of placements across the tree (see Section 1.8). The output file is compatible as "drag-and-drop" metadata with the Auspice platform, which allows for a rapid visualization of potentially problematic placements (Figure 1.3).

matUtils Introduce

Public health officials are often concerned about the number of new introductions of the virus genome in a given country or local area. To aid this analysis, matUtils introduce can calculate the association index (Wang et al. 2001) or the maximum monophyletic clade size statistic (Salemi et al. 2005; Parker et al. 2008) for arbitrary sets of samples, along with simple heuristics for approximating points of introduction into a region (see Section 1.8).

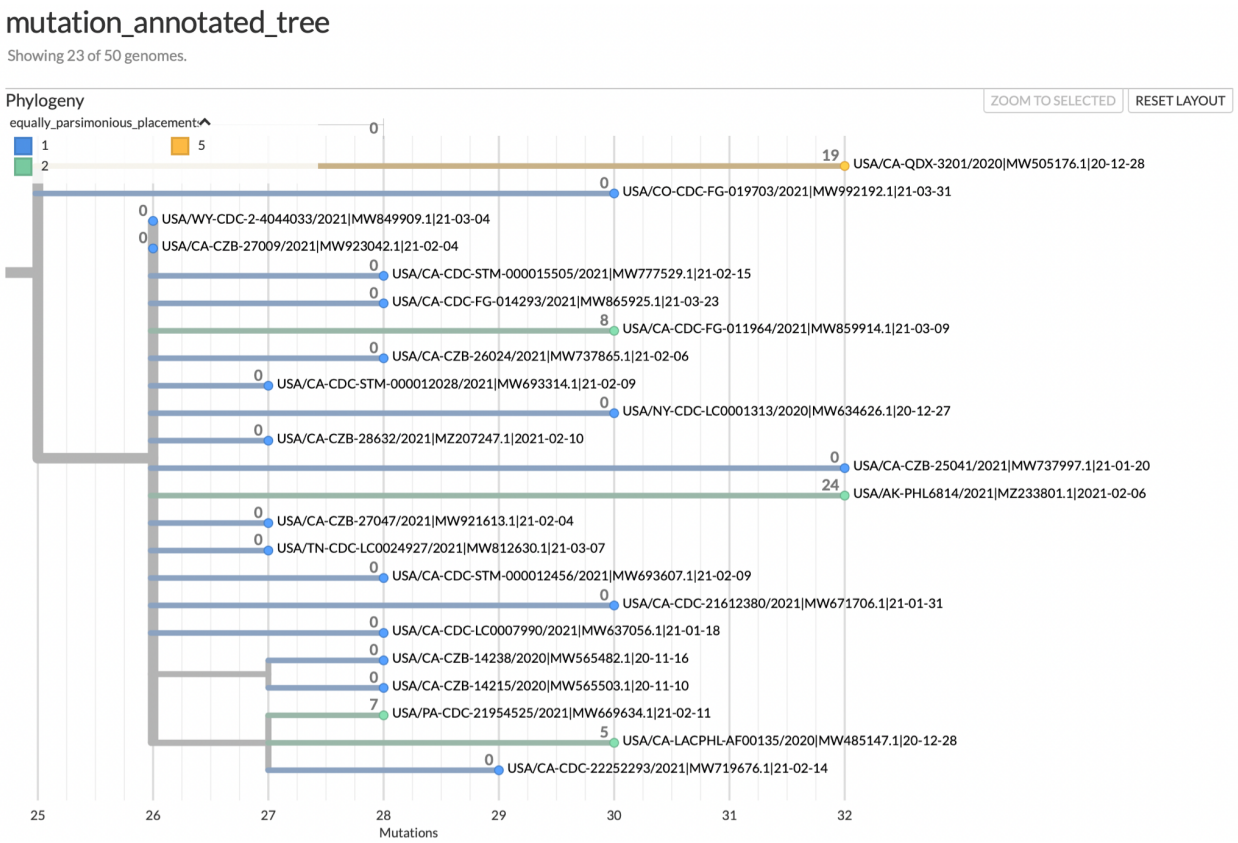

Figure 1.3: matUtils uncertainty statistics reveal low-quality sample placements. This Auspice view of an example subtree is annotated with both equally parsimonious placements (in color) and neighborhood size (branch label integers). 18 of our 23 samples in the subtree have a single placement and a neighborhood size of 0, indicating high placement certainty for those samples. Of the five samples with multiple equally parsimonious placements, one sample has 5 equally parsimonious placements with an NSS value of 19, indicating a high level of placement uncertainty for this sample spanning a relatively large neighborhood.

1.5: matUtils enables rapid analysis of a comprehensive SARS-CoV-2 tree

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 (Tables A1.1.2-A1.1.6). 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 (Table A1.1.2). 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 (Tables A1.1.2-A1.1.6). 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 A1.1.3). 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.

Certain functions of matUtils (such as extracting subtrees of provided sample names or identifiers) have also been ported to UCSC SARS-CoV-2 Genome Browser (Fernandes et al. 2020) and are available from https://genome.ucsc.edu/cgi-bin/hgPhyloPlace. Our database and utility fill a critical need for open, public, rapid analysis of the global SARS-CoV-2 phylogeny by health departments and research groups across the world, with highly-efficient file formats

that do not require high speed internet connectivity or large storage devices, and tools capable of rapidly performing large-scale analyses on laptops.

1.6: Maintaining a daily-updated mutation-annotated tree database of global SARS-CoV-2 sequences

We are maintaining a daily-updated mutation-annotated tree (MAT) database of global SARS-CoV-2 sequences at http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/. Our database is organized into sub-directories sorted by year, month and date. To update the MATs daily, we have set up a CRON job on a server at UCSC which downloads SARS-CoV-2 sequences daily from GenBank (Clark et al. 2007) and COG-UK (Nicholls et al. 2020) (see https://github.com/ucscGenomeBrowser/kent/blob/master/src/hg/utils/otto/sarscov2phylo/updatePublic.sh which calls other scripts in the same directory). We also include 253 sequences downloaded from the China National Center for Bioinformation (https://bigd.big.ac.cn/ncov/release_genome) in October 2020 that are not associated with GenBank IDs.

New sequences are added to the previous day's MAT using the UShER placement tool (Turakhia et al. 2021) with options to place the samples in the order of the fewest ambiguous bases and exclude sequences with 5 or more equally parsimonious placements. Previously excluded sequences are reconsidered for placement during each build. We also use matUtils extract to prune samples with 30

or more private mutations and those internal branches longer than 30 mutations, as these are highly indicative of error-containing sequences (Mai and Mirarab 2018). The trees are rooted to Wuhan/Hu-1 (GenBank MN908947.3, RefSeq NC_045512.2), and nodes with no associated mutations are collapsed (Turakhia et al. 2021). Our first MAT was created by starting with the last Newick tree release (dated November 13, 2020) of Rob Lanfear's sarscov2phylo (Lanfear and Mansfield 2020) containing 82,358 public sequences, adding the later additional public sequences using UShER. Each MAT is then annotated with Nextstrain clade and Pango lineage annotations using matUtils annotate -c with a file containing representative sequences for each clade/lineage. For Nextstrain clades, Nextclade assignments (https://github.com/nextstrain/nextclade) for all sequences are used. For Pango lineages, designated lineage representative sequences from https://github.com/cov-lineages/pango-designation/ are mapped to the corresponding public sequence IDs where possible.

In addition to MATs, we provide in each sub-directory: (i) a Variant Call Format (VCF) file containing the genotypes of public sequences, generated from the corresponding MAT with matUtils extract such that missing or ambiguous bases have been imputed by UShER using maximum parsimony (Turakhia et al. 2021), (ii) a Newick file also generated from the corresponding MAT using matUtils extract (iii) a tab-separated file containing information about each public sequence e.g. collection date, location, Nextstrain clade and Pango lineage, (iv) a tab-separated file with Nextstrain clades assigned to sequences by Nextclade

(https://github.com/nextstrain/nextclade) and (v) a tab-separated file with Pango lineages assigned to sequences by pangolin (https://github.com/cov-lineages/pangolin).

Our script to update the MAT daily is available at https://github.com/ucscGenomeBrowser/kent/blob/master/src/hg/utils/otto/sarscov2phylo/updateCombinedTree.sh.

1.7: matUtils: Design Overview

matUtils is implemented using the C++ programming language and is developed and maintained within the phylogenetic placement package of UShER (Turakhia et al. 2021), since matUtils shares the core mutation-annotated tree (MAT) data structure with UShER, which helps us ensure cross-compatibility of both tools. matUtils complements UShER through its ability to analyze and manipulate the MAT output but can be used as a standalone phylogenetics tool independent of UShER. Installing matUtils requires installing the UShER package that can be done via (i) a Docker container (https://hub.docker.com/repository/docker/yatisht/usher), (ii) the Conda package manager using the bioconda (Grüning et al. 2018) channel (http://bioconda.github.io/recipes/usher/README.html) or (iii) the installation scripts that we provide on our GitHub repository (https://github.com/yatisht/usher) for some recent Linux and MacOS releases. Detailed installation and usage instructions are available on our wiki: https://usher-wiki.readthedocs.io/en/latest/matUtils.html.

Several matUtils functions have multi-threaded parallel implementations through Intel's Thread Building Blocks library (https://github.com/oneapi-src/oneTBB).

1.8: matUtils: Implementation details

matUtils annotate

matUtils annotate is designed to annotate clades on the internal branches of the MAT. Our MAT format (specified in https://github.com/yatisht/usher/blob/master/parsimony.proto) provides an ability to annotate internal branches with an array of clade names, one for each clade nomenclature. Each run of matUtils annotate extends the clade name array size in a MAT by one to accommodate a new nomenclature. Only the node corresponding to a clade root is labeled with its clade name, as descendants of that node can be automatically inferred to belong to that clade. Clades can be nested, so that each sequence can be assigned to a clade corresponding to the lowest-level clade root to which it is a descendant. matUtils annotate provides two different ways to annotate clades in a MAT. Both ways, by design, ensure that all clades remain monophyletic. In the first, a user can directly provide the internal node identifiers corresponding to the root of each clade. In the second, a user can provide a list of representative sequences for each clade, such as training data for Pango lineages (https://github.com/cov-lineages/pango-designation), from which the clade root can be inferred in the tree. Not all sequences in the tree need to be designated by a clade. Since the training data is imperfect, and the representative sequences for lineages are

sometimes non-monophyletic in our tree, we have found the simple approach of using the most recent common ancestor (MRCA) does not yield accurate results. The matUtils annotate inference method works instead by first building a "consensus" sequence (where, by default, the consensus sequence requires an allele to be present in at least 80% of representative sequences, with lower frequency alleles marked as ambiguous) for each clade and finding its phylogenetic placement using UShER's placement module to obtain the clade root. When multiple equally parsimonious placements are available for the clade root, the algorithm uses a heuristic formula to compute the "best fit" for the training data, which rewards the placement containing a higher proportion of samples designated by that clade in the training data and penalizes descendants designated by some other clade in the training data. When the same root is found for multiple clades, the clade with fewest equally parsimonious placements, followed by the number of representative sequences in the training data, is prioritized.

matUtils Extract

The extract subcommand acts as a simple prebuilt pipeline with three distinct stages. The first of these, sample selection, collects the set of samples which fulfill each of the conditions indicated by input parameters, then gets the intersection of these sets to identify samples which fulfill all conditions specified on the command. Multiple conditions can be simultaneously specified in a single command for selecting samples, such as clade membership, maximum parsimony score, presence of

a particular mutation, and whether the sample name matches a specific regular expression pattern, among others. The second stage edits the input tree object to generate the indicated subtree, either by pruning excluded samples or by generating a subtree in a parallelized fashion, depending on the size of the chosen sample input. The third stage generates each of the requested output files representing the final tree. These files include Newick for pure tree information, parsimony-resolved VCF for variation information, and Auspice v2 format JSON for both (Hadfield et al. 2018). VCF production is parallelized for efficiency with large sample selections. A sample metadata table in CSV or TSV format can be incorporated into the JSON output. The full list of options can be found at our wiki: https://usher-wiki.readthedocs.io/en/latest/matUtils.html.

matUtils Uncertainty

matUtils uncertainty can calculate two different metrics for characterizing the phylogenetic certainty of a sample placement. The first metric is "equally parsimonious placements" (EPPs), which is the number of places on the tree a sample could be placed without affecting the parsimony score. An EPP score of 1 indicates a high placement certainty of a sample in its local neighborhood in that there is a single most parsimonious placement location for that sample on the entire tree, and a higher EPP score suggests the sample placement is less certain. This metric is calculated by computing the number of most parsimonious placements after remapping the input sample(s) against the same tree (disallowing it from mapping to itself) with UShER's

optimized placement module. About 85% of samples in our SARS-CoV-2 MAT database have an EPP score of 1. The second metric is “neighborhood size score” (NSS), which is the longest distance (in number of edges) between any two equally parsimonious placement locations for a given sample. This metric is complementary to EPPs – when multiple EPPs are possible for a sample, NSS indicates whether the placement uncertainty is restricted to a small neighborhood (small NSS value) or spans a large portion of the tree (large NSS value).

matUtils Introduce

matUtils introduce is aimed to help epidemiologists and public health officials estimate the number of new introductions of the virus in a given area or country. It includes a command which calculates maximum monophyletic clade size and association index statistics for phylogeographic trait association for user-provided input regions. Maximum monophyletic clade size (Parker et al. 2008) is the largest monophyletic clade of samples which are in the region – it is larger for regions which have relatively fewer introductions per sample and correlates with overall sample size. Association index (Wang et al. 2001) is a more complex metric which performs a weighted summation across the tree accounting for the number of child nodes and the frequency of the most common trait, such as membership in a particular geographical region of interest. Association index is smaller for stronger phylogeographic association and increases with the relative number of introductions into a region. For association index, matUtils introduce performs a series of

permutations to establish an expected range of values for the random distribution of samples across the tree. matUtils introduce also implements the regional weight heuristic described in chapter 2 of this work.

1.9: Performance benchmarking of matUtils and other phylogenetics software

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 exact commands used for each comparison can be found in Tables A.1.1.2-A.1.1.6, and the input data used for each comparison can be found at https://github.com/bpt26/matutils_benchmarking/ (DOI: 10.5281/zenodo.7983499). Some additional benchmarking tables can be found at the publishing journal in the supplementary materials (McBroome et al 2021, https://academic.oup.com/mbe/article/38/12/5819/6361626#322927285).

1.10: BTE: a Python module for pandemic-scale phylogenetic trees

[This section has been adapted from publication, “BTE: a Python module for pandemic-scale
mutation-annotated phylogenetic trees” (McBroome et al 2022, JOSS)]

Big Tree Explorer (BTE) is a Python extension of the highly optimized Mutation Annotated Tree (MAT) C++ library, which underlies the popular and highly effective phylogenetics tool UShER. BTE is written in Cython and provides an efficient and intuitive interface for traversing and manipulating mutation-annotated trees in a Python environment. It can load a mutation-annotated tree structure directly from a MAT protocol buffer file, provided by UCSC, or from a Auspice-format JSON. Alternatively, it can automatically infer mutation annotations and create a MAT from a Variant Call Format (VCF) file and a Newick format tree file. BTE provides all forms of standard traversal and methods for tree manipulation, including node and mutation creation, relocation, and deletion. BTE also provides native support for node-level clade and lineage annotations, such as those included in UCSC SARS-CoV-2 MAT protocol buffers. BTE's Cython code also includes functionality not present in the original MAT library, such as nucleotide diversity estimation. Altogether, BTE provides much of the same basic tree-level functionality as competing packages, while also supporting mutation and lineage annotations, allowing any user to take advantage of the powerful MAT data structure.

While there are multiple Python packages for phylogenetics available (Huerta-Cepas et al 2016; Talevich et al 2012), none are designed with mutation-annotated trees in mind and are less optimized for scalability. When attempting to use these packages with mutation-annotated trees, cumbersome file conversions to Newick and separate storage of mutations and tree structures adds substantial overhead to any analysis. BTE is designed explicitly for working with extremely large mutation-annotated parsimony phylogenetic trees. It is both more computationally efficient than competing packages and stores mutations and the tree within a streamlined data structure. BTE makes MATs and pandemic-scale phylogenies in general more accessible and useful to developers worldwide, helping to widen the SARS-CoV-2 bioinformatics bottleneck.

1.11: Comparison with Gold-Standard Alternatives

We compared performance of BTE as compared to two other popular packages for Python phylogenetics, ETE3 and Biopython.Phylo (Huerta-Cepas et al 2016; Talevich et al 2012). Benchmarking was performed by extracting random subtrees of the specified size from one of UCSC's global SARS-CoV-2 MATs, converting the subtree to Newick format, and performing the specified operation with each package. We both tracked total computation time and profiled memory use for each tool. Generally, BTE loads and traverses a tree more quickly than competitors, with a particular advantage at large tree sizes. It also has a substantially improved implementation for identifying the ancestors associated with a given node, with

multiple orders of magnitude improvement. In terms of memory use, it is generally comparable to both ETE3 and Biopython.Phylo, with memory use ranging in the dozens of Kb for most operations, with up to a few hundred Mb for loading and large subtreeing operations. BTE can be found at (https://github.com/jmcbroome/BTE; DOI: 10.5281/zenodo.7983513). All code for benchmarking is available at a dedicated repository (https://github.com/jmcbroome/bte-benchmark; 10.5281/zenodo.7566955).

Chapter 2

Identifying SARS-CoV-2 regional introductions and transmission clusters in real time

[This chapter has been adapted from publication, “Identifying SARS-CoV-2 regional introductions and transmission clusters in real time” (McBroome et al 2022, Virus Evolution)]

2.1: Phylogeography

The massive scale of the SARS-CoV-2 sequencing effort has revealed deep inadequacies in our current methodology for phylogenetic analysis. Tools designed to work on small, sparse, static datasets have adapted poorly to the demands of a pandemic where tens of thousands of new genome sequences are generated and shared daily (Hodcroft et al 2021). Some have made progress by adopting generalized statistical methods built for large data such as random forest regression (O’Toole et al), while others have continued to improve on existing methods (Gill et al 2020, Vöhringer et al 2021), but phylogenetic solutions capable of scaling to millions of samples need to be developed. While our group, among others, has laid the groundwork for pandemic-scale phylogenetics (Dellicour et al 2021, Maio et al 2021, McBroome et al 2021, Schneider et al 2020, Shchur et al 2021, Turakhia et al 2021,

Ye et al 2021) much remains to be done to translate evolutionary inferences to public health understanding and action.

The unprecedented scale of the genomic sequencing effort requires novel approaches to evolutionary, medical, and public health inference. Some groups have developed phylogenetically informed statistics for identifying mutations associated with increased transmissibility and other fitness-related parameters (Richard et al 2021, van Dorp et al 2020). In other cases, simple methods- such as the assaying of groups of identical samples- have been successfully applied to identify superspreader events and similar infection clusters (Bello et al 2022, Gómez-Carballa et al 2020). Unfortunately, many analyses still lack scalable or phylogenetically informed approaches.

The intersection of geography and phylogenetics, phylogeography, has often relied on heavily downsampled and static trees or limiting their analysis to early stages of the pandemic (Dellicour et al 2021, du Plessis et al 2021, Lemey et al 2020, Lemey et al 2021, Lemieux et al 2021, Ragonnet-Cronin et al 2021, Rito et al 2020). Some authors have analyzed several tens of thousands of samples with a divide and conquer approach, subdividing the overall tree by lineage and combining separately inferred results (McCrone et al 2021). Others have had similar success tracking the introduction and spread of a distinct new lineage over the first weeks after its emergence (Kraemer et al 2021). While useful for assessing transmissions between countries and major introductions, downsampling limits our ability to assign specific samples to regional infection clusters or identify clusters of potential interest. Even

creative techniques taking advantage of phylogenetic tree structure to make analysis more tractable will not always be applicable and are limited in their ability to scale to millions of samples across dozens of regions. Additionally, much of these analyses are not readily interpretable for an efficient public health response, lacking intuitive visualization and data exploration tools. There is therefore a significant need for fast, automated, scalable and interpretable phylogeographic approaches for an effective public health response to emerging situations.

To address this need, we present here a phylogenetically-informed summary heuristic (the "regional index"), implementation (matUtils introduce), and data exploration and visualization tool (Cluster Tracker: https://clustertracker.gi.ucsc.edu/) for identifying introduction events and associated clusters of descendants in a given region. Our approach can be used to efficiently identify infection clusters and evaluate transmission dynamics across dozens of regions and millions of samples. Results obtained using this method are congruent with gold-standard Bayesian analyses and are accurate when applied to simulated data. Our visualization platform enables researchers and public health workers to explore new SARS-CoV-2 introductions across the USA, updated daily with all available global public data. This work will empower real time research and public-health applications of genomic epidemiology during the SARS-CoV-2 pandemic and beyond.

2.2: Cluster Concept and Definitions

A cluster, in terms of our analytical approach, is a set of closely-related samples from the same region and descended from a common ancestor with a regional introduction event. Under our definition, the complete set of actively circulating pathogens in a region will be composed of one or more genetically distinct clusters, which resulted from unique introduction events. In the phylogenetic tree, they appear as a set of leaves (samples) from a given geographic region that are descended from a shared common ancestor. A cluster may be monophyletic or paraphyletic, depending on whether some descendants of the cluster common ancestor left the geographic region. We consider location, or region, as a categorical state across the phylogenetic tree. A regional transmission event is where a child node is from a different region than the parent node. These patterns reflect cases of infected travelers moving between regions, followed by local transmission and eventual sampling of a number of descendant infections.

2.3: A Heuristic for Identifying Introductions and Clusters

The core of our heuristic is the “regional index”, which is a weighted summary of the composition of descendants of a node of a phylogenetic tree. Intuitively, if all descendents of an internal node were found in region A, we would assume that the ancestor represented by that internal node was circulating in region A. Similarly, if we sampled a virus from region A which had exactly the inferred genome for this internal node, we would assume the ancestor represented by this node

was in region A. The same logic would apply if no descendents were in region A. Therefore, by computing a heuristic which ranges from 0 to 1 based on the genetic distance to and composition of downstream descendents under a binary model of region membership, we can effectively approximate our intuition that the viral ancestor represented by that node was inside or outside a given region. It is defined as

$$Regional\ Index\ (C) = \frac{1}{1+\frac{\frac{Di}{Li}}{\frac{Do}{Lo}}}$$

where "Li" is the number of downstream leaves that are in a given region, "Di" is the minimum total branch length to a leaf descendent in the focal region, and "Lo" and "Do" are the same for out-of-region leaves (Figure 2.1). On a tree inferred using maximum-parsimony, total branch length is equivalent to the distance in mutations between the query node and the descendant leaf.

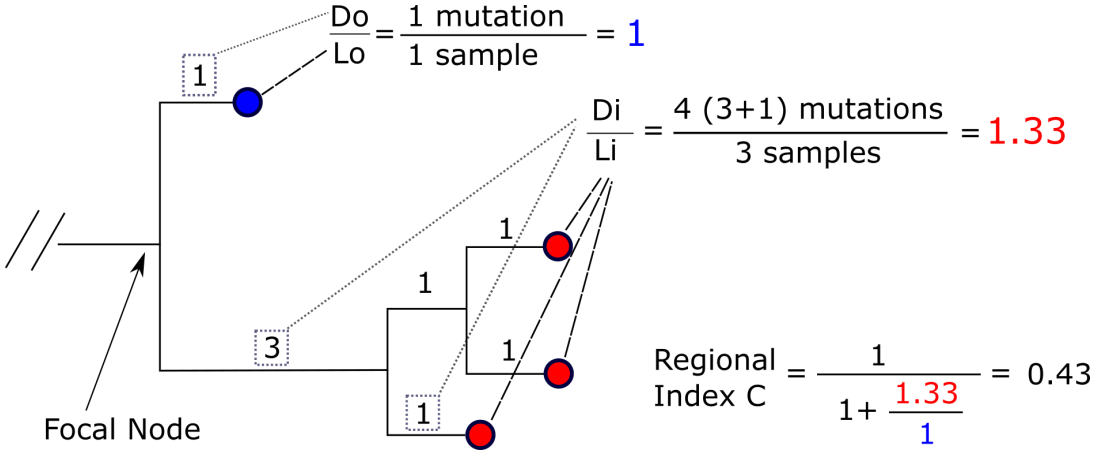

Figure 2.1: Example Index Calculation. This small example tree demonstrates a computation of our index, using blue to indicate "out-of-region" and red to indicate "in-region" leaves. The focal node at the base has an index value below 0.5, suggesting that it is out-of-region by our heuristic. Our introduction point is therefore along the long branch below the root, and the ancestor of the downstream in-region sample cluster would have existed along that branch.

We apply additional rules to handle cases where C is undefined or can't be computed. When a descendent leaf is genetically identical to the internal node and is in-region, C is 1. Similarly, when a genetically identical leaf is out-of-region, we treat C as 0. When such identical children exist both in and out of the region, we treat the node as in-region, as some infection with this genome must have existed in that region. We do not apply this index calculation to leaf nodes, which do not have children, and assume simply that the leaf is either in or out of the region as a given. This requires that each leaf included in the analysis be accompanied by accurate geographic location metadata.

This heuristic has several useful behaviors. For example, a sample identical to a specific internal node will always confer complete confidence about the location of that node, as we have sampled one genome that is identical to the ancestor directly. This can effectively identify nested clusters, where a new group of infections resulting from a regional introduction in turn produce clusters in other regions. It also accounts for the number of leaves downstream in our heuristic, on the assumption that introductions of a strain from one region to another require the lineage to be locally circulating in the origin region, but not necessarily lead to significant local transmission in the target region. This reduces the overall number of introductions we infer. If we account for the number of descendents, internal nodes will generally be assigned to the dominant region if distances are similar, reducing the number of consecutive reciprocal regional transmissions that might be inferred otherwise. Our heuristic strikes a balance between the principles of descendent composition and

genetic distance, allowing us to efficiently analyze a large phylogenetic tree with minimal metadata.

Once indices for a given region have been calculated for each node, the second step is to identify clusters of samples putatively associated with an introduction. This is accomplished on a per-sample basis. The path from the sample to root is traversed and the indices for each ancestor being in the focal region is noted. Generally, the index declines from 1 to 0 along the ancestry path from leaf to root. The introduction point is called where the index for an ancestor being in-region is below 0.5, or the root, whichever is encountered first. 0.5 is our natural cutoff, representing the index value in a scenario where the composition and distance of downstream samples in and out of the region are equivalent, but can be adjusted by the user to modify cluster calling behavior. Once each sample has an ancestor chosen as the introduction node, they are grouped into clusters that share their ancestral introduction node. Generally, a larger threshold value will lead to more, smaller clusters, while a lower threshold value will lead to fewer, larger clusters.

As this heuristic is independent and specific to a region, it can be computed for an arbitrary number of regions across a single tree in parallel. When multiple regions are included, origins of putative clusters can be identified after introduction points are found by examining index scores across all other regions for the origin node and noting the region with the highest index. This metric can be calculated for one region of any size in a single post-order traversal with dynamic programming

(see Section 2.8), which makes it very fast to compute even on extremely large phylogenies with expansive regions.

2.4: Evaluation of Our Heuristic Method

Our implementation is part of the matUtils online phylogenetics package (McBroome et al 2021) and uses the efficient mutation annotated tree protocol buffer format and associated library (Turakhia et al 2021). To test runtime efficiency conditioned on a tree, we applied random subsampling and recorded time to compute our heuristic for a single region. We found that it takes less than forty five seconds on a single thread even for trees of more than two million samples (Table A2.1).

To validate our results, we performed simulations consistent with viral evolutionary dynamics with inter-region dispersal events using the tools VGSim (Shchur et al 2021) and phastSim (Maio et al 2021) (see Section 2.8). We found that our heuristic with default parameters recovered the true geographic location of internal nodes up to 99.8% of the time under realistic conditions for SARS-CoV-2 across an exactly correct bifurcating tree. We further attempted to model our ability to correctly recover clusters on a simulated tree with collapsed branches and realistic mutation rates for SARS-CoV-2. In comparing the clusters we recovered with the true set, we obtained an adjusted Rand index (Rand 1971) of up to 0.999. This suggests that our approach is generally quite accurate, though high migration rates or extremely low mutation rates can be confounding, as these scenarios are associated with minimal geographic and phylogenetic signal respectively (Table A2.2; see

Section 2.8). More practically, this implies that our method will perform best when within-region transmission is substantially more common than between-region transmission (as in e.g., country-level or state-level analyses).

To compare our results to widely used but much slower (days to months) analyses, we used our method to replicate a published phylogeographic analysis for the SARS-CoV-2 pandemic. Alpert et al used Bayesian phylogeography (Lemey et al 2009) to identify 23 distinct introductions of B.1.1.7 into the United States as of March 4th 2020. We obtained their subsampled tree and applied our heuristic using country labels to define the relevant regions (see Section 2.8). With our method, we exactly replicated their identified clusters (Adjusted Rand Index 1.0). Alpert et al additionally predicted "sink" states, or the state to which each of the 23 introductions initially transmitted. We find that for all 23 clusters, samples in the identified sink state are closest or tied for closest in branch length to our inferred introduction point. This suggests that our approach can produce results congruent with more complex statistical models in a fraction of the time.

Another relevant method used in similar situations and that scales well to larger phylogenies is parsimony reconstruction, where region membership is treated as a character trait and inferred across the tree using the standard Fitch-Sankoff algorithm (Sankoff et al 1975, Vöhringer et al 2021, Volz et al 2021). This is more efficient than Bayesian approaches, but is heavily influenced by variation in sampling and low mutation rates relative to sampling and transmission. We performed a simple parsimony reconstruction based on the Fitch algorithm (Fitch 1977) similar to that of

Volz et al on simulated data (Table A2.3). We found that while parsimony performs as well or better than our heuristic on well resolved trees, when the average number of mutations per node is less than one and polytomies are common (as in SARS-CoV-2) our approach has greater accuracy. Our approach is more efficient than the Fitch algorithm because it requires only a single traversal of the phylogeny to compute.

2.5: Global SARS-CoV-2 Transmission Dynamics and Infection Clusters

Using our method, we traced transmission clusters in 102 countries from across the world (Figure 2.2A) using the global parsimony phylogenetic tree, built from 5,563,847 available sequences on GISAID (Shu et al 2017), GenBank (Sayers et al 2021), and COG-UK (Lancet Microbe 2020) on 11-28-2021 (see Section 2.8). Cluster size is highly skewed (Figure 2.2C), with approximately 20% of distinct regional clusters containing 89% of samples. This suggests that the majority of novel introductions do not lead to the establishment of a new locally-circulating strain, consistent with previous findings (du Plessis et al 2021).

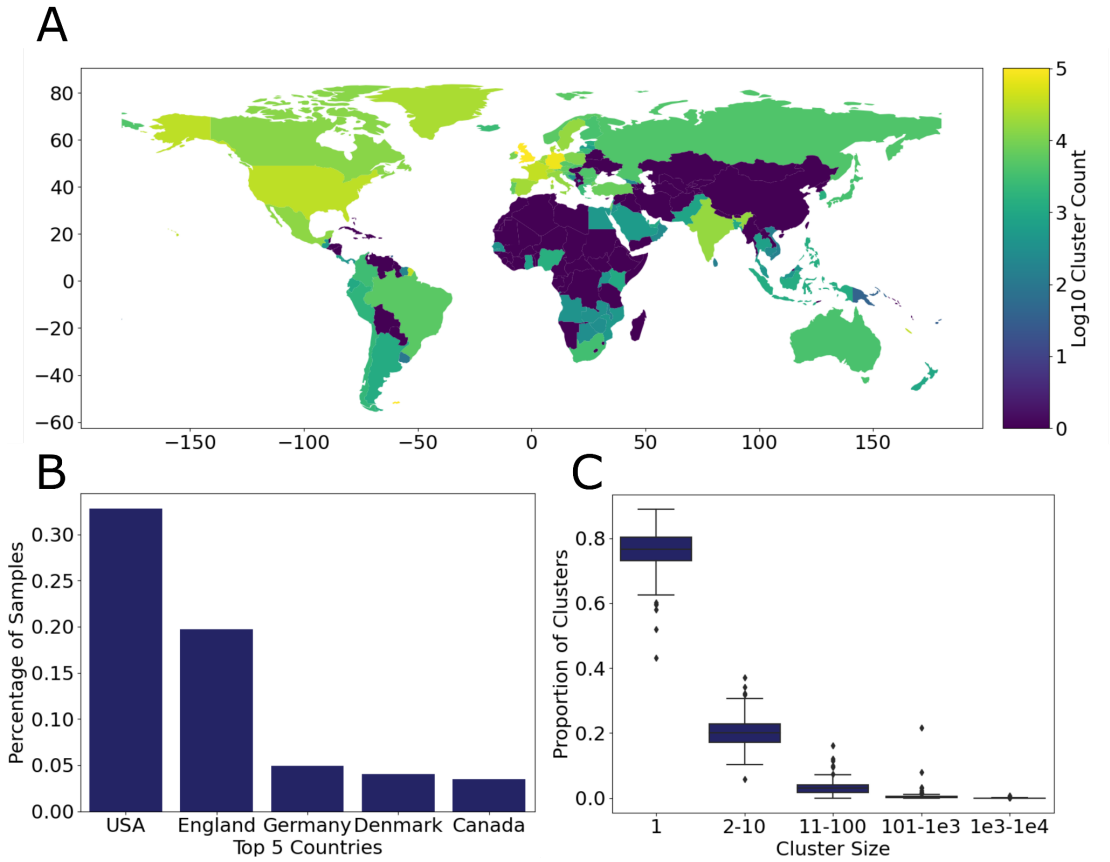

Figure 2.2: Global Distribution of SARS-CoV-2 Transmission Clusters. **A:** The log count of clusters detected across each of the 102 countries surveyed. The number of clusters detected is largely a function of total local sequencing effort. **B:** The five countries with the highest representation in the data. The USA and England together constitute more than half of all available sequences. **C:** Cluster sizes are consistent across countries. Most clusters are small, implying most newly introduced SARS-CoV-2 lineages quickly die out.

Global contributions to sequence repositories are notably biased, with 51% of all samples belonging to either the USA or the United Kingdom (Figure 2.2B). This is a significant restriction on global transmission analysis, especially as the inference of the origin of a cluster is highly dependent on robust sequencing at the origin (see Section 2.8). We therefore narrowed the next step of our analysis to the United States,

which has robust and relatively comprehensive sequencing across each state as well as detailed state-level metadata for the vast majority of available samples.

2.6: SARS-CoV-2 Transmission Into and Across the USA

We identified more than three hundred thousand distinct state-level SAR-CoV-2 infection clusters in the United States over the course of the pandemic, as of November 2021 (Figure 2.3). Approximately 84% of these clusters have an assigned origin using our method (see Section 2.8). Only 7% of our clusters appear to be of international origin, with the majority reflecting transmission within the USA. Mexico and Canada are among the most common international origin regions, in line with expectations given their long land borders (Table A2.4). England is also relatively common, likely because it is very well sampled. This indicates that it is possible that some clusters originate from less sampled intermediate regions and are assigned to the UK or other highly sampled locations. This suggests that relative sequencing effort in a given region is an important bias with respect to accurately identifying the origins of newly identified clusters and results should be interpreted with caution. International introductions rates are correlated with higher total sampling and therefore population size, particularly for California, Texas, New York, Massachusetts, and Florida (Figure 2.3B).

Within the USA, introductions come from a mix of neighboring states and high-population travel centers (Table A2.5). We attempt to mitigate sampling biases-resulting from larger populations, higher case rates, increased sequencing, or other

factors that are not specific to geography- by calculating a log-fold enrichment for rates of introduction from a given source region (see Section 2.8; Figure 2.4). Note that while log-fold enrichment may reveal spatial relationships, it does not reflect the absolute importance of a region as a source or sink of viral transmission.

As with results from international introductions, we also find an enrichment for introductions that originate in geographically adjacent states. Log-fold enrichment is more than five times greater for neighboring states than for non-neighboring states within the USA ($p=1.5e-117$, Mann-Whitney U). Simple counts of inferred introductions are also enriched to a lesser extent between geographically adjacent states ($p=2.2e-16$, Mann-Whitney U). This suggests that SARS-CoV-2 transmission over interstate land borders is a major mechanism for spread within the USA. These results are largely in line with previous results in other viruses (Kozińska et al 2019) and SARS-CoV-2 (Tiwari et al 2021), suggesting that this heuristic is capturing and summarizing true geographic structure within the global SARS-CoV-2 phylogenetic tree.

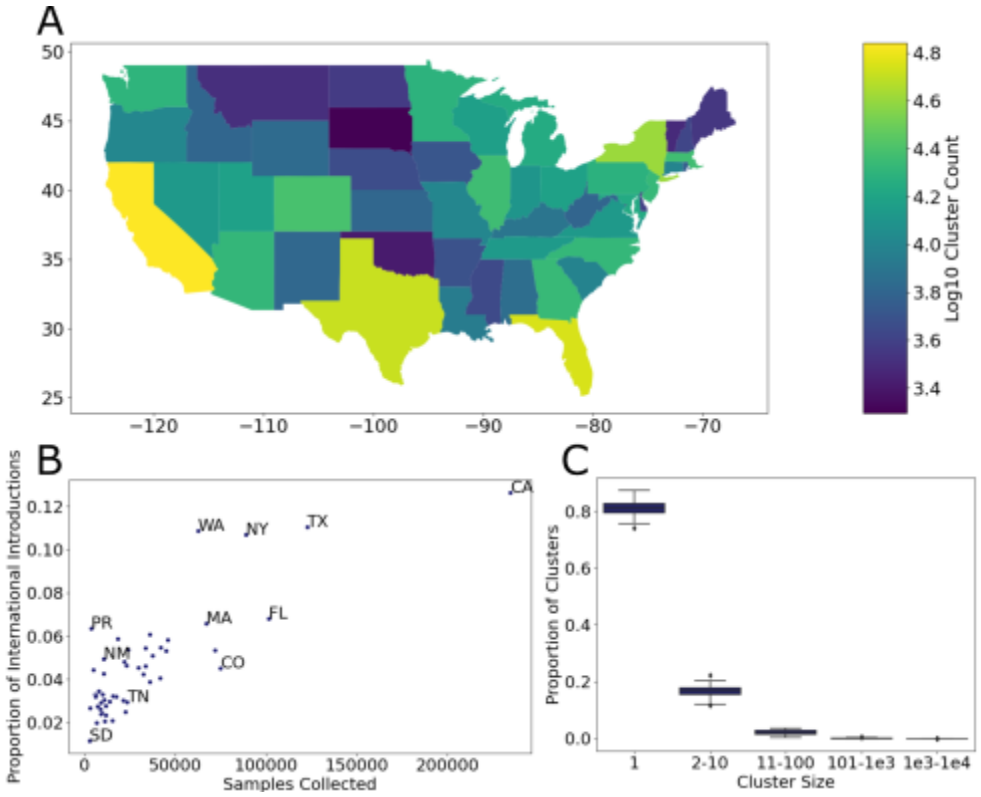

Figure 2.3: International and Interstate Introductions across the USA. A: The log count of clusters identified across the continental USA. California, Texas, Florida, and New York are associated with the greatest number of unique clusters. **B:** The proportion of international introductions in each state plotted against total samples collected in that state. This relationship is largely linear, reflecting the correlation between sampling, population size, and levels of international travel. PR (Puerto Rico) exhibits relatively more international introductions for its sampling than other territories and states of the United States. **C:** The distribution of cluster sizes across states. These are largely consistent with clusters identified at the international level.

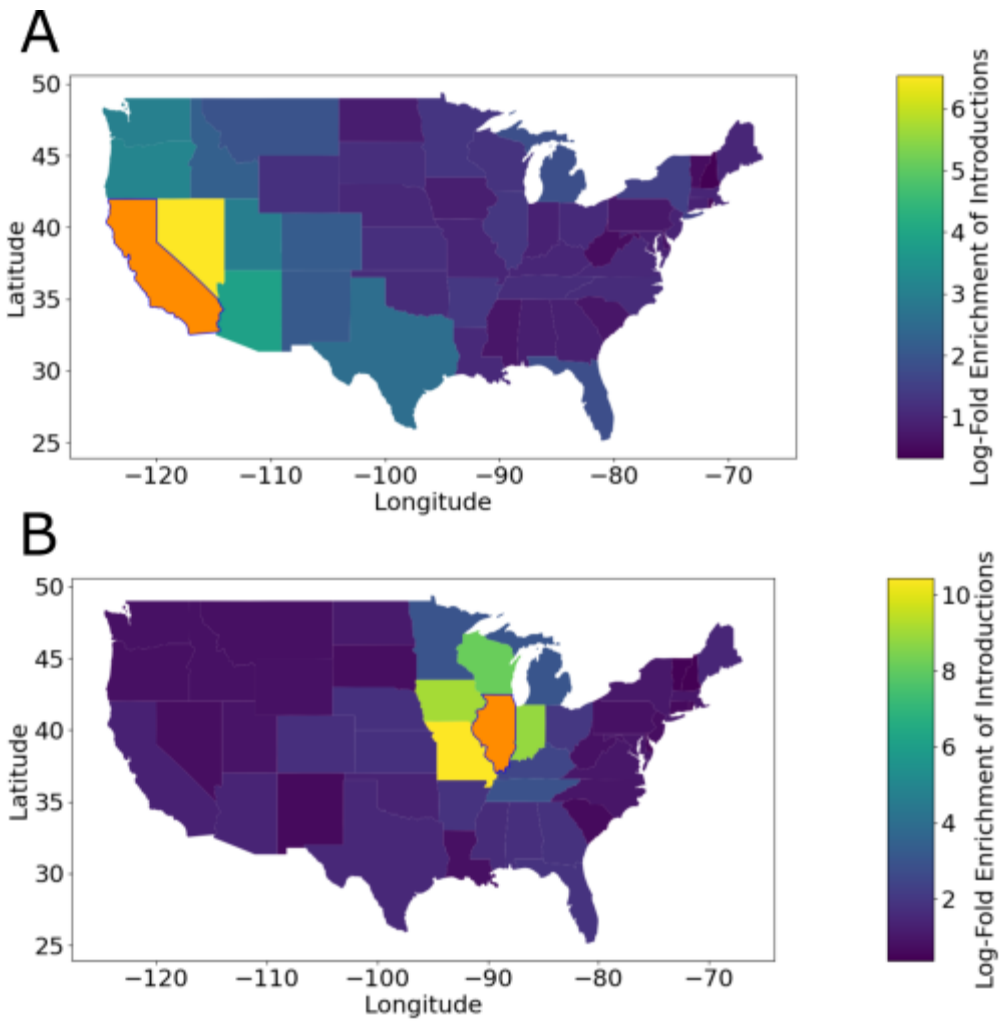

Figure 2.4: Log-Fold Interstate Transmission for the States of California (A) and Illinois (B). A: Interstate introductions of COVID-19 into California are relatively more likely to originate on the West Coast, particularly from Nevada. **B:** Interstate introductions of COVID-19 into Illinois are relatively more likely to come from the immediate surroundings, particularly Iowa and Missouri.

2.7: A Daily-Updated Website To Explore SARS-CoV-2 Clusters in the USA

To make the results of this work broadly useful for the research and public health community, we have developed a visualization and exploration platform. Cluster-Tracker is a publicly-available, daily-updated website displaying the latest results for applying our heuristic to sequences collected from across the United States of America interactively (clustertracker.gi.ucsc.edu; see Methods; Figure 2.5). Cluster-Tracker is open-source with a flexible backend pipeline that allows any user to construct a similar site for any set of regions they have geographic information and sample identification for (https://github.com/jmcbroome/introduction-website; 10.5281/zenodo.7566936).

Cluster Tracker is composed of two primary sections and some descriptive text (Figure 2.5). The first section is an interactive map of the United States. In the default view, this map is colored by the number of clusters detected across each state throughout the course of the pandemic. The true number of introductions into a given region is likely to be substantially larger

because many small clusters will not be sampled by ongoing viral surveillance efforts, but major local transmission clusters should be represented. By clicking on a state, the site changes to a view specific to that state. In the default view, the map is colored by the log-fold enrichment of introductions from each other state to that state. Optionally, the user can switch the color to raw counts of detections with the toggle in the upper right.

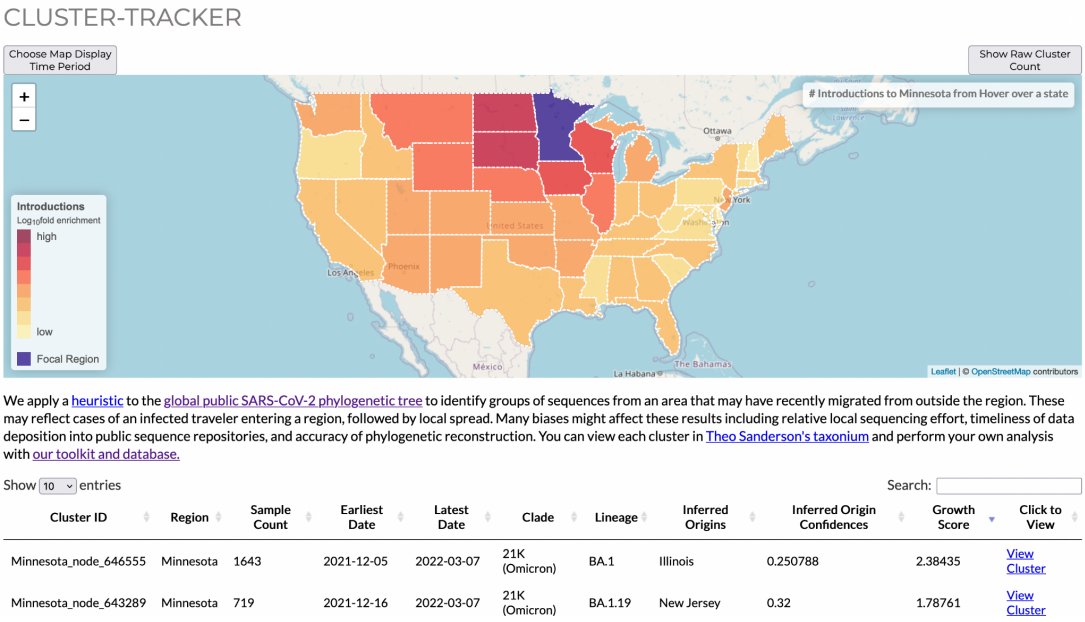

Figure 2.5: The Cluster-Tracker Site. The Cluster-Tracker tool is updated daily at clustertracker.gi.ucsc.edu. Users can interactively explore the latest results of our heuristic applied to each of the continental United States, by sorting the interactive table, selecting states to focus on in the map, and using the Taxonium tree-viewing platform to examine clusters of interest in detail.

The second section is a sortable, searchable table display of the highest priority clusters. In the default view, these are the top 100 clusters overall as sorted by "growth score". We define "growth score" as the square root of the number of samples divided by the number of weeks since the introduction occurred. The goal of this metric is to weight clusters by relative size and how recently they entered a given area, so that clusters of interest to public health appear first. When a state is selected, this table changes to the top 100 clusters obtained from that particular state. Basic information including clade, lineage, the earliest and latest dates of detection, and inferred origins are displayed for each cluster. The "inferred origin confidences" column is the highest or tied for highest regional index among all other regions for

the parent node to the cluster origin, with a floor of 0.05 below which the cluster is simply marked “indeterminate”. The “inferred origins” column is the regions which match these scores, and generally represents our best guess at the origin of this cluster. The last column of the table contains links to the Taxonium viewer (https://github.com/theosanderson/taxonium) which will automatically render the full tree and zoom to the cluster of interest when opened (Figure 2.6). Full results and the taxonium protocol buffer file, which encodes the tree and all cluster IDs, are available to be downloaded at the bottom of the page.

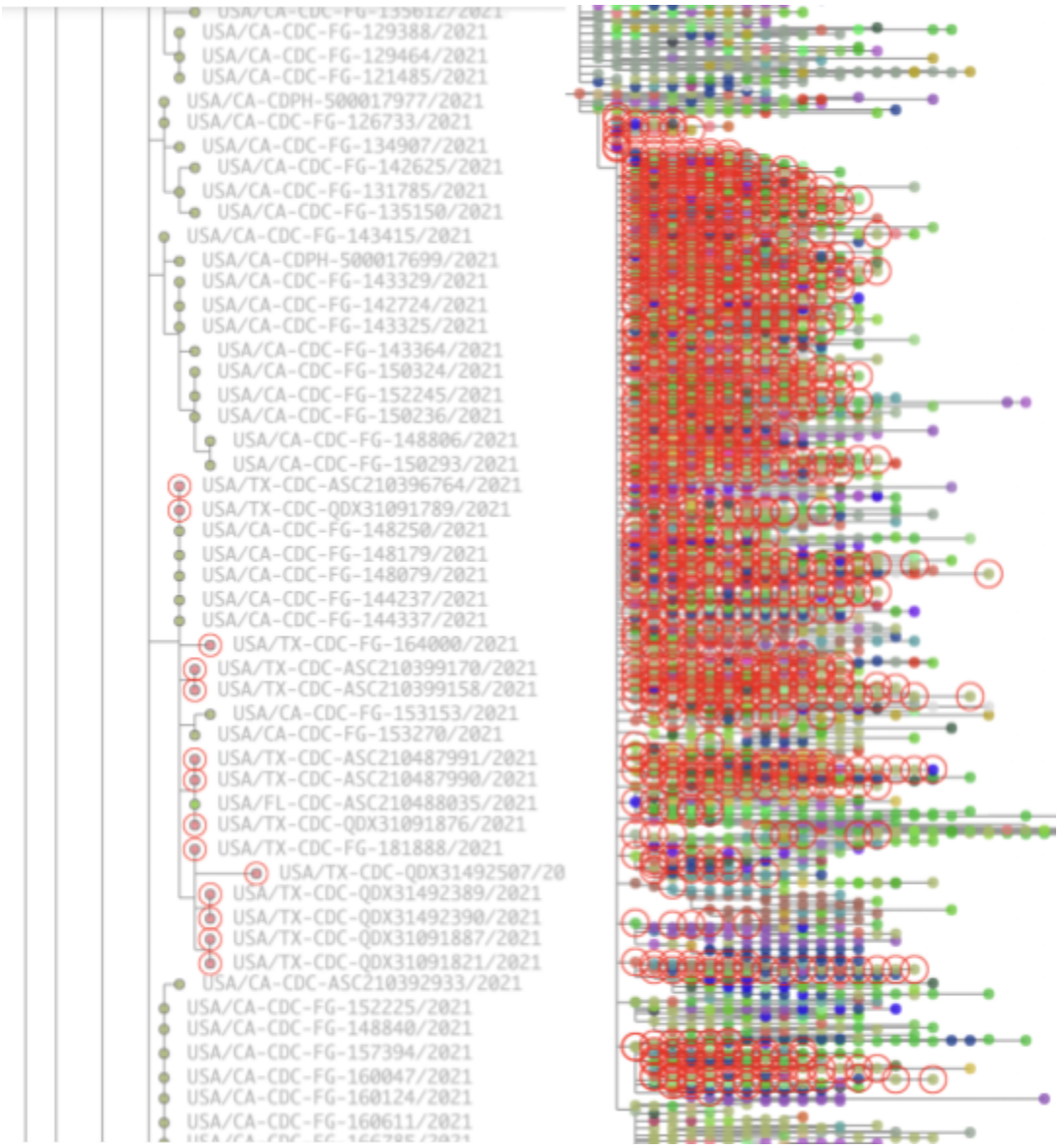

Figure 2.6: Example Clusters in the Taxonium Phylogenetic Tree Viewer. (A) An example cluster in Texas (member samples circled in red) that is inferred to have originated from California (Regional Index = 0.94). There are many samples from California closely related to the cluster's common ancestor, supporting California as the most likely origin. (B) A different, much larger, 9,533 leaf cluster in California. This represents a lineage of SARS-CoV-2 commonly circulating in California, descended from one of the original introductions of the Delta variant into California in mid June 2021. Descendants from this cluster have transmitted to other regions many times, but members of this cluster have been found in California as recently as December 7th, 2021.

2.8: matUtils Implementation Details

We implemented a calculation of this heuristic as a part of our online phylogenetics package, matUtils, under the command “matUtils introduce” (McBroome et al 2021) (https://github.com/yatisht/usher). Our implementation uses dynamic programming based on a post-order traversal to compute the regional index for each node in the tree in a single pass for each region. This is because the four parameters which define regional index- distance to the nearest descendent and total descendents for in-region and out-of-region- can be computed from these same metrics for each child of a node plus the branch length to each child. The total number of leaves descended from a query parent node is the sum of all leaves descended from each of their children, and the shortest distance traversed to a leaf is the minimum of each child’s minimum distance traversed plus the branch length between that child and the query parent. Therefore, by computing it first for nodes with only leaf children, then progressively deeper internal nodes, we only have to reference the children of each internal node and check their stored values instead of having to traverse from each node. This step is optionally parallelized across distinct regions, if multiple regions are passed.

The secondary step is an ancestry traversal for each sample in the tree, identifying the most recent ancestor which has a regional index below the set threshold, which is inferred to be the introduction point for this lineage. Once introduction points have been inferred for each sample, samples are grouped by

shared introduction points into clusters, basic statistics and information are computed, and results are reported.

Ultimately, our implementation can compute this heuristic, identify clusters, and report all results in less than two minutes for a tree containing more than two and a half million samples (Table A2.1). The speed of calculation is a major attraction of this heuristic approach over more complex Bayesian models. Calculating in minutes on minimal computing resources makes this method accessible and applicable to update results daily, identifying clusters and introductions as they occur and new data is uploaded globally. Accordingly, this implementation underlies our website Cluster-Tracker, which is updated with all new uploaded data each day and a recalculation of our heuristic.

Handling Nested Clusters and Unstructured Regions

We implemented a few additional parameters that can be used to control behavior at the secondary cluster identification step. Once that is useful is setting a short-range maximum index requirement- that is, looking ahead at some additional number of ancestors and ensuring that each of those have a lower regional index than the intended ancestor node. Setting this parameter causes small nested clusters to be merged into larger overarching clusters. Another useful parameter is a minimum required branch length between the ancestor inferred to be in-region to its parent; if the branch length is less than the minimum, then the parent instead of the in-region node is inferred as the introduction point. Setting this parameter allows sibling

clusters to be merged if both of their branch lengths are below minimum; this also resolves unstructured parts of the tree where large polytomies of identical samples with branch length 0 both in and out of a region are included.

Prioritization and Bias Handling

Another significant point of consideration is cluster prioritization. This cluster identification method is based solely on the phylogenetic tree and simple sample-region association, and while this makes it lightweight and flexible, identifying clusters which died out locally months ago is not of use to public health offices doing real-time transmission cluster tracking. We therefore in our implementation sort the output by a "growth score", defined as the square root of the number of samples associated with the cluster divided by the time in weeks from the oldest sample in the cluster to the current date plus one. This means that large, recent clusters will appear at the top of any output tables, and makes the method more easily accessible when thousands of clusters are being inferred simultaneously.

When using this method to examine inter-region transmission dynamics, we rely on comparable and significant levels of sequencing in order to identify introduction origins. Intuitively, the less sequencing is performed in a region, the less likely we are to recognize sequences from that region when they appear in another region. We can compensate for this bias to an extent by calculating log-fold enrichment of introductions between regions. This is computed as

$$Log\ Fold\ Enrichment\ (LFE)\ = log_{10}(\frac{Iab \times Ixx}{Iax \times Ixb})$$

Where I_{ab} is introductions from region A to region B, I_{xx} is introductions from any region to any region, I_{ax} is introductions from region A to any other region, and I_{xb} is introductions from anywhere to region B. This computation can remove biases in rates of detected introduction which would apply to any pair of regions, but requires many regions to be computed as points of comparison. This score is used to color the map on Cluster-Tracker when a state is selected and has a very strong correlation with geographic distance.

Simulation for Validation

To assay the performance of our heuristic, we fully simulated a pandemic phylogeny with VGsim (Shchur et al 2021) and phastSim (Maio et al 2021). From the resulting mutation-annotated tree, we calculated true node region states based on VGsim's migration event output and applied our heuristic with matUtils (McBroome et al 2021). We then computed accuracy as the proportion of internal nodes which have a heuristic value above 0.5 for the true state. Leaves are excluded from this calculation as they are taken as an input in our heuristic and will always be 100% accurate.

For our specific results, we simulated a one-million-leaf SARS-CoV-2 tree under a simple model with two equivalently-sized regions with an even rate of migration between them, no strain or site selection and complete immunity for

recovered individuals (Table A2.2). We included a lockdown parameter starting at 5% infected and ending at 1% infected, with a 10-fold reduction in transmissivity under lockdown, and a sampling multiplier of 0.2 in order to deepen the tree by effectively extending the time for one million samples to be collected.

ARI (Adjusted Rand Index) and IAC (Internal Assignments Correct) are our quality metrics. ARI represents how well our method correctly groups samples into true clusters descended from a single introduction event. ARI performs best when migration is low, leading to large and clean clusters which are easily separated heuristically, and performs somewhat better when scale is increased. IAC is the proportion of internal nodes which are assigned to the true region by our heuristic across the bifurcating tree. It is computed on the correct bifurcating tree because collapsing true nodes from different regions leads to nodes that are naturally indeterminate. IAC is generally robust, only performing slightly worse with an increased migration rate, likely as deeply set internal nodes tend towards indeterminacy with high distances to many leaves across different regions. This suggests that the primarily limitation of our heuristic is simply the number of mutations available to distinguish samples from across varying regions rather than any structural or fundamental issues.

All code for this simulation is available as a modular and reproducible Snakemake pipeline at github.com/jmcbroome/pandemic-simulator (10.5281/zenodo.7566940).

Global Phylogenetic Tree Construction

At UCSC we maintain a large phylogeny of all GISAID (Shu et al 2017), GenBank (Sayers et al 2021), and COG-UK (Lancet Microbe 2020) sequences using the script https://github.com/ucscGenomeBrowser/kent/blob/master/src/hg/utils/otto/sarscov2phylo/updatePublic.sh and the UShER online phylogenetics suite (McBroome et al 2021, Turakhia et al 2021). Updates are performed daily by obtaining all newly uploaded sequences from each database and placing them on the previous day's global phylogenetic tree with UShER (see McBroome et al).Starting with our phylogeny updated on 11-28-2021, we pruned all samples with long branch lengths and path lengths using the matUtils parameters --max-branch-length 45 and --max-path-length 100 and performed a round of optimization with an SPR radius of 8. The resulting phylogeny contained 5,563,847 samples with a total tree parsimony of 4,847,954.

Computing USA state transmission

We obtained the latest mutation-annotated phylogenetic tree representing the entirety of all public samples and all samples available on GISAID on 11-28-2021. As the standard format for publicly uploaded SARS-CoV-2 sequence identifiers is "Country/(Area)-CollectingAgencyInfo/Year|Date", we extracted sample labels for samples in the USA by identifying samples with names beginning with "USA/" and then extracting the two-letter state code, if it matches with a two-state letter code.

This resulted in 1,764,019 labeled samples belonging to the USA. Samples from outside the USA were labeled by country; countries and ambiguous labels with less than 500 samples in GISAID and public data were excluded and their samples removed. Samples from "mink" were additionally excluded as they may not be from human sources. The resulting tree contained 5,237,796 of the total of 5,563,847 samples available, reflecting more than 94% of all SARS-CoV-2 genomic data collected and incorporated to date.

We applied matUtils introduce with default parameters to this tree and sample set and produced the full by-sample output. After computing basic statistics, we calculated log-fold enrichment of introductions between all pairs of states, and a selection of other countries to and from the USA. All code for this paper is provided at (https://github.com/jmcbroome/cluster-heuristic; 10.5281/zenodo.7566933).

Cluster-Tracker Website Development

All relevant javascript and some example data files are provided at (https://github.com/jmcbroome/introduction-website; 10.5281/zenodo.7566936). This github includes a brief description of how to set up a local test site and run the backend pipeline for generating new results to display for your regions of interest. It is based on Leaflet (https://leafletjs.com/) and DataTables (https://datatables.net/) for the primary view, and includes links to the Taxonium tree viewer (https://taxonium.org/) for detailed cluster exploration.

We include Python scripts to create the backend data for the website display, contained in the "data" directory. This includes two versions of the primary pipeline. One is specific to the United States, which fills in many default parameters and uses data included in the repository. The second is a more flexible and configurable pipeline which given a tree, sample labels, and a geojson, can create a Cluster-Tracker equivalent website for any set of regions.

Comparison with Published Studies

To compare our approach to that of Alpert et al 2021, we retrieved the Auspice JSON they used to generate Figure 2.3 from (https://github.com/grubaughlab/CT-SARS-CoV-2) and obtained table S3 from their supplementary data online, which contains cluster labelings for samples from the tree represented by the JSON. We converted the Auspice JSON to the UShER MAT protocol buffer format using python. We labeled all samples in the resulting tree by their country of origin and ran matUtils introduce with default parameters. The resulting labels were compared to the cluster labels presented in table S3 and the Adjusted Rand Index was computed across all labeled samples with scikit-learn (Pedregosa et al 2011). We performed this analysis twice- once including all samples in their tree from any region and once excluding samples from the USA in their tree that were excluded from their clusters. The first method resulted in an ARI of 0.9 and the second a perfect 1.0; this discrepancy results from a single difference where a pair of large clusters, sibling to one another, are merged by our results when samples

excluded from their clusters are included in our analysis. This is because a sample identical to the parent node of these two sibling clusters from the USA is excluded from Alpert et al's clusters. In any case, the clusters we identify are highly concordant with Alpert et al's results. All code for this analysis is available on https://github.com/jmcbroome/cluster-heuristic (10.5281/zenodo.7566933).

2.9: Discussion

The goal of this resource is to make cluster identification, exploration, and prioritization more accessible and digestible for public health offices and policy makers. A significant roadblock for public health action is the sheer quantity of daily new data and the speed with which we can draw inferences from these data. Cluster-Tracker can assist exploration and prioritization of the latest genome sequences, quickly identifying the clusters most likely to be of interest for public health action for a given region. Our construction pipeline is flexible and can be applied for any set of regions (e.g., county-level), allowing groups anywhere to construct web interfaces for intuitive SARS-CoV-2 phylogenetic data exploration.

While simple and efficient, our heuristic does exhibit some weaknesses. It is not a model; while simulations have demonstrated its efficacy in describing simple patterns of transmission, it can fail to correctly infer more complex scenarios and requires substantial and dense input data. Simulations indicate that it performs best with larger and more homogenous regions with low rates of migration, such as countries. If the user attempts to infer introductions with very small regions with high

rates of inter-regional transmission, it may fail to properly recapitulate transmission patterns. Additionally, regionally-biased differences in sequencing effort (Brito et al 2021, Colson et al 2021) can lead to significant biases in raw counts and our ability to correctly identify introductions, making individual cluster origins difficult to interpret in many cases. In terms of functional limitations, the heuristic is based on a binary regional labeling model, and does not have the ability to directly interpret lat-long coordinates or unique location values for samples like some Bayesian phylogeographic methods. Overall, it remains a useful tool for quickly assaying viral diversity and inter-regional transmission patterns on a global scale.

The pandemic has made the need for rapid and powerful tools to unlock the potential of pandemic-scale genomic epidemiology. The method we developed and the efficient software package we provide will empower researchers worldwide to make fast inferences from vast sequence datasets. Our results have revealed geographic structure at scales below the level of pango-lineage (O'Toole et al 2021) within the global SARS-CoV-2 phylogeny. We have provided tools and resources with which to explore this geographic structure and draw useful inferences for specific areas. Additionally, to empower public health officers and the public to explore the spread of SARS-CoV-2 across the USA, we developed an accessible open-source interactive interface for our results, which can automatically compute and display introductions and clusters with each update to the global phylogenetic tree. Our work can support public health groups across the world to quickly understand and apply insights obtained from the latest genomic data.

Chapter 3

Automated Agnostic Designation of Pathogen Lineages

[This chapter has been adapted from publication, “Automated Agnostic Designation of Pathogen Lineages” (McBroome et al 2023, bioarxiv)]

3.1: Pathogen Nomenclature

Pathogen lineage nomenclature, or the designation of epidemiologically distinct groups below the level of species, are important for facilitating effective research, treatment, and communication about diseases. Despite the universal importance and long history of nomenclature systems for pathogens, there remains a plurality of approaches to apply to new emerging pathogens. These lineage systems are generally based on some combination of three elements: phenotype, genotype, and geography. Phenotype-based systems are often predicated on vulnerability to antibiotics (Collins et al. 1982) or serology (Lancefield 1933); pathogens with serology based nomenclature systems include Salmonella spp. (Brenner et al. 2000), dengue viruses (Cuypers et al. 2018; Simmonds et al. 2017), and Streptococcus spp. (Facklam 2022; Lancefield 1933). Geography-based classification systems may be appropriate for pathogens where the primary reservoir is in non-human species, such

as Chikungunya virus (CHIKV) (de Bernadi Schneider et al. 2019) and the Zaire Ebola viruses (Kuhn et al. 2014). Finally, genotype-based nomenclature divides a species-wide phylogeny into statistically well-supported, mutually exclusive taxa generally referred to as “lineages” or “clades”. These groups can be defined as clusters of samples below a genetic diversity threshold or as the descendents of an inferred common ancestor on a single phylogeny. Genotype-based classification has become increasingly common in application to viruses such as RSV (Ramaekers et al. 2020), dengue (Cuypers et al. 2018) and influenza viruses (Anderson et al. 2016).

The COVID-19 pandemic presented a unique challenge to these nomenclature system approaches. Other diseases often have lineages inferred to have originated several years to decades in the past, with well-defined characterizing genetic changes. These stable nomenclatures rarely need active updates, being defined with respect to a single phylogeny that remains largely unchanged. However, in SARS-CoV-2, a single mutation may be all that defines a new epidemiologically distinct lineage (O’Toole et al. 2021). Additionally, the SARS-CoV-2 genomic data is orders of magnitude greater in volume than that for extant pathogens, as well as constantly growing as new data is collected (Hodcroft et al. 2021). The expansion of the dataset means that the SARS-COV-2 phylogeny is regularly updated (McBroome et al 2021), necessitating further review and updates to any genotype-based lineage system.

The current solution to these challenges is the popular Pango lineage system. Pango is a genotype-based dynamic lineage nomenclature for SARS-CoV-2 characterized by the manual designation of new lineages from a global phylogenetic

tree (Rambaut et al. 2020). Pango lineages are hierarchical and comprehensive, including hundreds of nested designations for any subgroup of viruses that may be of concern. When compared to traditional nomenclature, these often initially contain fewer samples, are less genetically distinct, and are regularly updated as new genetic data is collected. These small, dynamic lineages serve a critical function in organizing genetic data for public health tracking efforts. The Pango system has provided initial names used for all Variants of Concern (VOCs), including B.1.1.7 (Alpha) and B.1.1.529 (Omicron), and defined the serial replacement of Omicron lineages through time (BA.1, BA.2, BA.5). Pango has accordingly played a central role in facilitating effective tracking of and communication about emerging SARS-CoV-2 strains over the course of the pandemic.

Currently, Pango relies on manual curation and designation, including the crowdsourcing of lineage proposals on a public forum (https://github.com/cov-lineages/pango-designation). More than 2500 SARS-CoV-2 variants have been named under the Pango system as of January 2023. The trained human eye is excellent at distinguishing new lineages of interest from groups of low-quality or contaminated isolates, but the Pango group's resources have become strained as the volume of data has increased and public investment has decreased. Furthermore, crowdsourced proposals are vulnerable to delays as well as regional and personal bias, as individual researchers have differing opinions on the importance of various mutations and are more or less likely to search for clades from specific parts

of the world. A more objective metric to evaluate candidates for lineage designation could help to reduce this bias and streamline the lineage proposal and review process.

We propose a simple heuristic approach for the definition and expansion of genotype-based dynamic nomenclature systems. Our method is rooted in information theory, optimizing for the representation of sample-level haplotype information. It requires only a phylogeny with branch lengths scaled to genetic distance, with additional forms of information emphasizing specific mutations or samples being optional. It is efficient in application to extremely large phylogenies and produces a comprehensive hierarchy of genetically distinct lineages. Our lineage system is flexible and can be effectively weighted in any number of ways, allowing epidemiologists and researchers to prioritize critical elements for lineage definition and tracking efforts. Importantly, it can expand a preexisting lineage system, making adoption of this approach for the maintenance and expansion of existing nomenclature straightforward. We, in collaboration with the Pango designation team, have implemented this system as a new input for the existing Pango lineage designation infrastructure (https://github.com/jmcbroome/autolin; DOI: 10.5281/zenodo.7566921). Additionally, as sequencing technology becomes more widely applied, both novel and extant pathogens will develop similarly dense and expanding genomic datasets. This approach will provide a scalable solution to creating and managing these dynamic lineage systems for any pathogen.

3.2: The Genotype Representation Index (GRI)

A nomenclature system can be likened to a language, where additional words, analogous to lineages, are defined for common, unique concepts to reduce the average number of words per sentence. Along these lines, an effective nomenclature summarizes a complex phylogeny into useful, distinct categories to facilitate effective analysis and communication. The lineage hierarchy is generally defined with respect to a specific rooted phylogeny, where a number of specific ancestral nodes are designated as lineage roots. Higher-level lineages are divided hierarchically into finer sublineages. Individual samples, represented as tips of the tree, are members of every lineage that is rooted in its inferred ancestry. To automate the definition of this hierarchy, we need some objective measure of distinctiveness or importance that can be computed for lineages. One approach is to compute a distinctiveness value for every node on the tree, as individual lineages in these systems are generally defined as the descendents of a single, specific ancestor. Once we have a node-level measure of lineage efficacy, we can iteratively construct a nomenclature by selecting high-value nodes and designating them as new lineage roots. These lineages can then be presented to an end user, or directly incorporated into an expanding nomenclature.

To this end, we define the following index, hereafter referred to as the "genotype representation index" (GRI) (Figure 3.1).

$$Genotype\ Representation\ Index = \frac{N \cdot D}{\frac{S}{N} + D}$$

The GRI takes values with respect to a specific node on the tree, hereafter referred to as the "focal node". Here, N is the number of descendent tips from the focal node, D is the total branch length from the focal node to the root of the tree or previously designed parent lineage, and S is the sum of branch lengths from the focal node to each descendent tip. In natural language, the GRI is the mean branch length position of the focal node along the ancestry paths of all its descendants, multiplied by the total number of descendants. The GRI increases both with an increasing number of descendents (N) and with being closely related on average to those descendents (lower S). Nodes with an overall high GRI will be closely related to many descendents, representing a group of consistently genetically distinct samples- a good choice for lineage labeling. For a mutation-annotated tree (Turakhia et al 2021), such as those used for SARS-CoV-2, the branch lengths (D and S) are in units of total mutations across the genome. However, the GRI can be computed on any rooted tree topology, as long as branch lengths are scaled by genetic distance. The GRI is high for focal nodes where descendent samples are genetically similar to one another and the focal node itself is genetically distinct from the rest of the phylogeny, desirable qualities for lineage designation (Rambaut et al. 2020). The motivation behind this formulation is presented in the Methods section.

Autolin defines a lineage system based on the GRI by applying a simple greedy maximization algorithm. Initially, the GRI is computed for each node on the tree and the node with the highest value is chosen as a new lineage root. Additional mutually exclusive lineages are defined by disregarding all samples covered by an

existing lineage label and recomputing the GRI for all remaining samples and their ancestors. To prevent the retroactive definition of lineage parents that might interfere with an existing hierarchy, we additionally disregard nodes that are directly ancestral to existing or newly added lineages. Additional hierarchical lineages are defined similarly by only considering samples within a specific existing "parent" lineage. This process is repeated until a desired number of lineage labels have been defined or all available nodes fail to pass thresholds for designation. This iterative approach is not guaranteed to find the highest overall GRI lineage configuration among many possible combinations of lineages, but it scales well to millions of samples and a rapid pace of lineage updates.

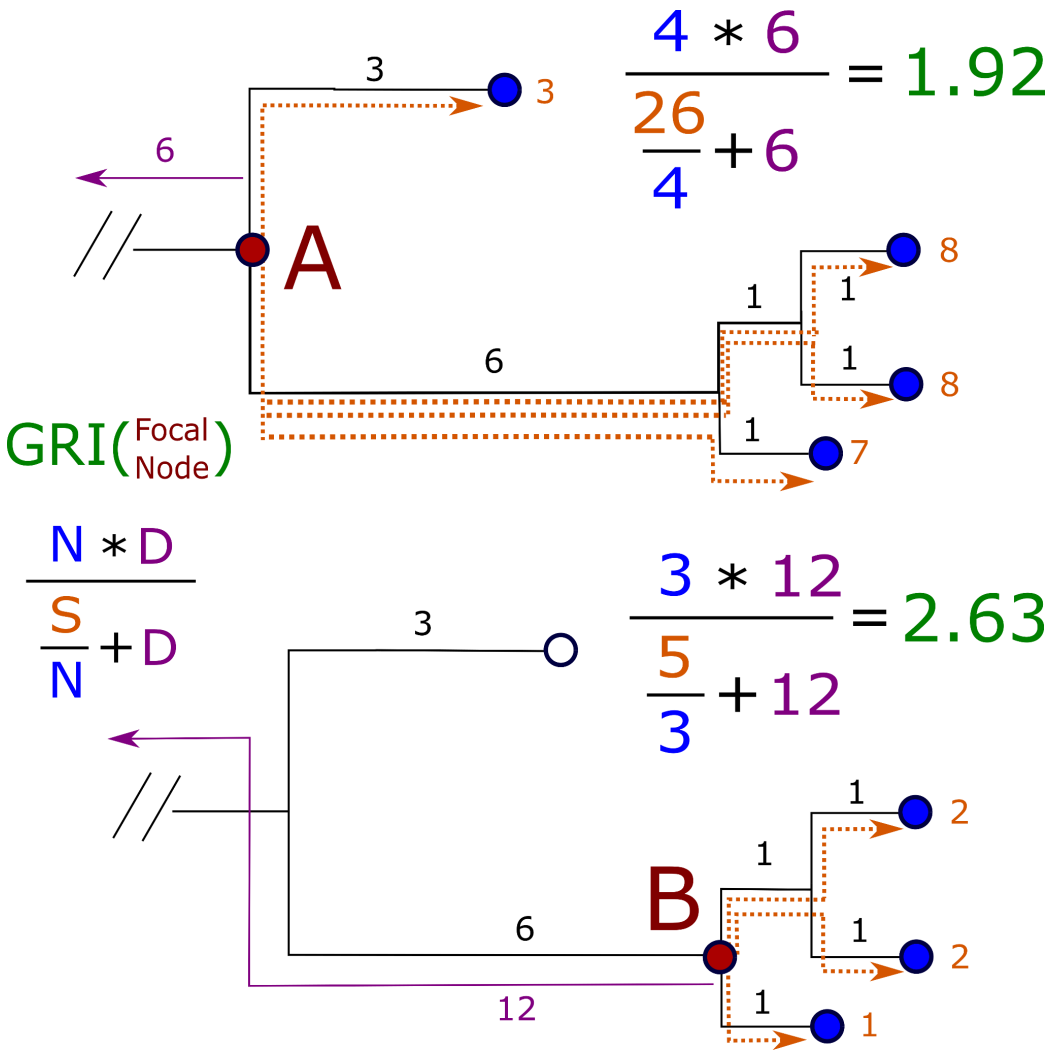

Figure 1: Computation of GRI. This figure depicts the computation of GRI values for two nodes on a small example subtree. The base of this subtree is a total distance of 6 from the last lineage root, represented in purple. The node at the base of this subtree (A) has a total path length to descendents (S) of 26, 4 total descendents (N), and is a total distance of 6 from the last root (D), leading to a GRI of 1.92. The lower child node (B) only has 3 descendents (N), but has a much lower path length (S) and a longer distance to the last root (D), meaning that it scores much higher at 2.63. In this case, we would choose to assign a lineage label to the lower child node (B).

3.3: Adjustments to the Genotype Representation Index

In practice, pathogen lineage nomenclature systems are generally designed for purposes beyond summarization of a phylogenetic tree. Lineages often carry connotations of distinct phenotypic behavior, such as serological types, immune evasion, transmissibility, and other metrics. Some parts of the genome may contribute more than others to these phenotypes. For example, some spike protein changes are known to alter immune escape in SARS-CoV-2 (Greaney et al. 2022). Information about parts of the genome associated with important phenotypes is inherently more valuable and more worth representing in our lineage system. Accordingly, we may want to weight our GRI calculation by giving additional value to these mutations when computing distances. Conversely, we may want to disregard parts of the genotype that are not informative for phenotypic behavior or that are not readily interpretable, such as repetitive noncoding sequences or sites prone to recurrent errors.

The GRI, while based on genotype representation, can be flexibly altered to focus on the representation of important elements. The original Pango rules for the definition of SARS-CoV-2 lineages have requirements around evidence for international transmission and changes to proteins to designate lineages that are more likely to be epidemiologically important (Rambaut et al. 2020). By using these weighting schemes for GRI, we can automatically propose lineage designations of high epidemiological import. This allows researchers to develop fully-informed and highly applicable nomenclature systems.

3.4: Sorting and Prioritizing Novel Lineages

In some cases, curators may prefer to designate a smaller number of lineages that are of higher apparent epidemiological impact, to improve the average impact and simplicity of the lineage system. In this case, our approach can be applied to identify a large number of individual lineage candidates, which can then be filtered and prioritized according to lineage-level statistics. While many simple filters we support, such as the number of countries a lineage has been detected in, are simply applied to the tabular report, we do also provide a more informed sorting schema based on lineage growth.

To sort putative lineages for manual inspection after the initial designation procedure, we fit a geographically stratified exponential growth model to each proposed lineage using Markov Chain Monte Carlo (MCMC). Bayesian methods of this type are appropriate for inference with small, noisy datasets, as the uncertainty in the model is directly quantified. To summarize, we construct a posterior distribution of exponential growth coefficient scores filtered through a binomial sequencing model. Lineage proposals which have a high, low-variance posterior distribution of growth are more likely to be rapidly expanding and are of high priority for labeling. Additional information can be found in the Methods section. This model is extremely simple compared to standard epidemiological models due to the constraints of available data and necessary speed. Accordingly, it does not directly inform the initial designation of lineages, but instead serves as an optional out-of-the-box solution for

users to identify putative lineages of immediate and critical public health importance without significantly adding to overall compute time for the pipeline.

3.5: Systematic Application to SARS-CoV-2 and Example Designations

As a basic demonstration of our method, we applied the pipeline to the complete SARS-CoV-2 global public phylogenetic tree, as of 2022-12-11 from http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/ (McBroome et al. 2021). In the absence of an extant lineage system and considering all samples, the GRI based approach assigns more than 170,000 lineages to this phylogeny. These lineages are divided into twelve levels, representing recursive levels of child lineages, with the first level being trivially defined by the root of the phylogeny itself. The majority of these lineages are small, with only 10% of designations being larger than 100 samples. In general, users may wish to restrict autolin to output only lineages above some minimum size (as in e.g., Rambaut et al. 2021). Of the approximately 2000 Pango lineages included in this phylogeny (Figure A3.1), more than 1175 are closely matched with a GRI equivalent lineage, including the major Delta and Omicron lineages. Another 586 Pango have a corresponding GRI identified lineage with a Jaccard similarity of overlapping samples greater than 0.5 (Figure A3.2). The remaining unmatched 217 lineages are mostly extremely small, with more than 95% of them including <10 samples in this phylogeny, and therefore would not pass the default filters for Autolin (Figure A3.3). Overall, the systems are concordant, especially with regards to major variants.

To evaluate the utility of our method for maintaining and expanding dynamic lineage nomenclature specifically, we applied Autolin to the same phylogeny as above, but built on the extant Pango lineages. We generated 187 new lineage designations using the default configuration parameters, which only considers samples collected in the preceding 8 weeks. 24 of these lineages were actively sampled in December 2022 as of 2022-12-11. These active designations were highly dispersed in size, with a mean size of 82 samples and a median of 45 samples. The full report for the active designations is available in Table A3.1.

We fit an exponential growth model to each active lineage (Table A3.1, Figure 3.2) and obtained a 95% confidence interval estimate of the rate of exponential growth. The average confidence interval for the exponential growth interval was relatively large (0.07, 0.49), due primarily to the effects of limited sample sizes. 16 of the 24 lineages had a positive lower interval bound, which is evidence for active spread in the countries they are present in. The width of the interval is naturally dependent on the data available; while the average estimate for our lineages is +/- 0.2, estimates for lineages with at least 50 total collected samples had a much narrower average value of +/- 0.07. All model confidence intervals are reported in Table A3.1. All code for fitting and reproducing these results is available at https://github.com/jmcbroome/lineage-manuscript (10.5281/zenodo.7566938).

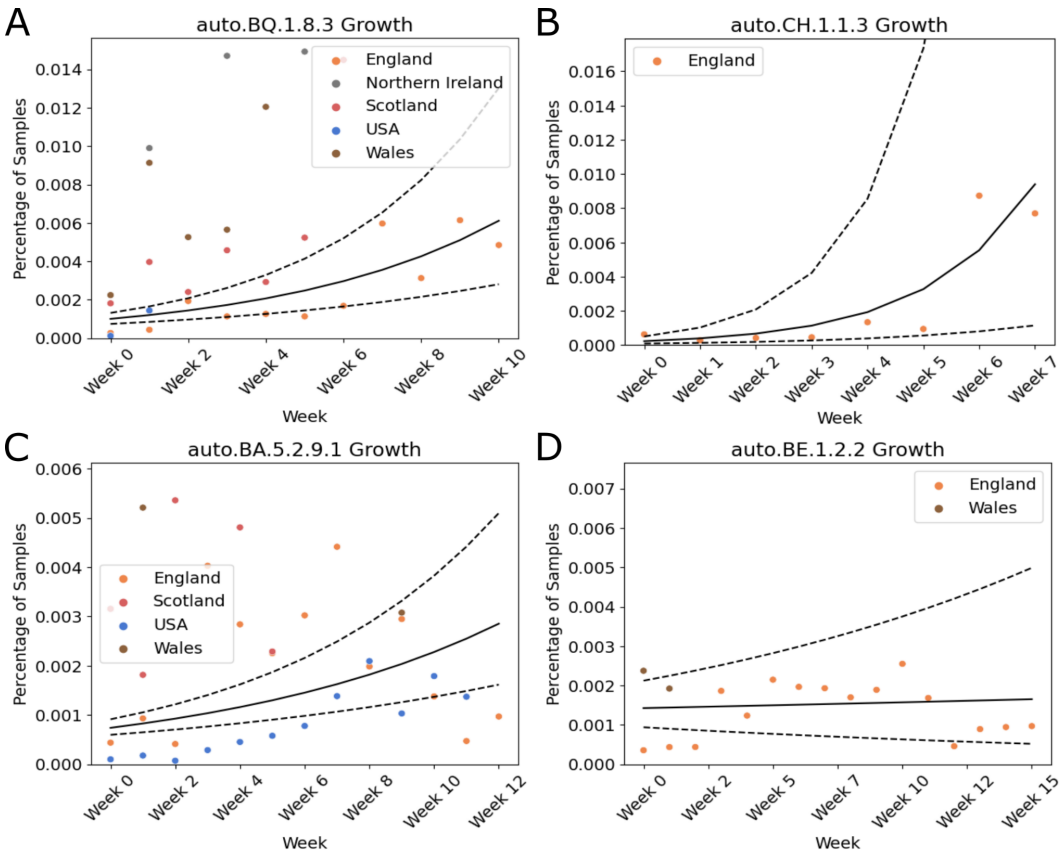

Figure 3.2: Exponential Growth Modeling. The above four plots describe some of the lineage annotations produced by our method based on the public SARS-CoV-2 data. The black line is the median estimated growth trajectory, while the dotted lines represent the trajectories that would result from the lower and upper bounds of the 95% credible interval of the growth rate. The x-axis is represented in weeks since first detection among each country.

This procedure can serve to organize and prioritize lineage designations, despite suffering from high uncertainty. Figure 3.2 displays a small example selection of lineages and model fits in further detail. The naming schema matches the Pango naming schema, with the addition of an “auto” prefix denoting that the lineage in question was created by our approach and not manually designated by the Pango team. ”auto.CH.1.1.3”, while exclusive to England, exhibits a very rapid expansion in

latter weeks that drive a very high, if wide, estimate of growth. “auto.BQ.1.8.3” and “auto.BA.5.2.9.1” are more international, but less consistent; the latter appears to grow consistently in the United States, but fluctuates to a much greater degree in England. Finally, “auto.BE.1.2.2.” is an example of a low-priority designation, with no strong evidence of positive growth. Altogether, our models are capable of capturing a diverse set of lineage trajectories and rapidly and effectively identifying lineages undergoing exponential expansion.

We have collaborated with the Pango team to incorporate our approach into the existing SARS-CoV-2 lineage designation infrastructure. Statistics such as lineage size, associated mutations, and geographic localization are computed and reported as a part of a pull request to the curated Pango repository. Our update includes links to external data exploration sources such as cov-spectrum (Chen et al 2022) and taxonium (Sanderson 2022; Kramer et al 2023), as well as programmatic generation of all files requisite for the incorporation of the new designations. All code for this procedure can be found at https://github.com/jmcbroome/autolin (10.5281/zenodo.7566921).

3.6: Application to Other Pathogens

The GRI approach can be used to generate lineage proposals for any pathogen, with or without an existing base nomenclature. We compared our approach to a recent Zika (ZIKV) nomenclature proposal (Seabra et al 2022), applying Autolin directly to their likelihood phylogeny (see Appendix 3). We find high level

concordance between the automated system and the formal nomenclature (ARI 0.47, p<0.001) (Figure 3.3). The formal Zika nomenclature proposal is the result of the application of Bayesian clustering directly on aligned sample haplotypes (Seabra et al 2022), so while this system is genotype-based, it does not directly depend on the phylogeny. This may explain some of the inconsistencies between these systems, particularly as regards basal groups like ZA. However, we do see high level concordance between these groups, particularly in the widespread ZB.2 variants.

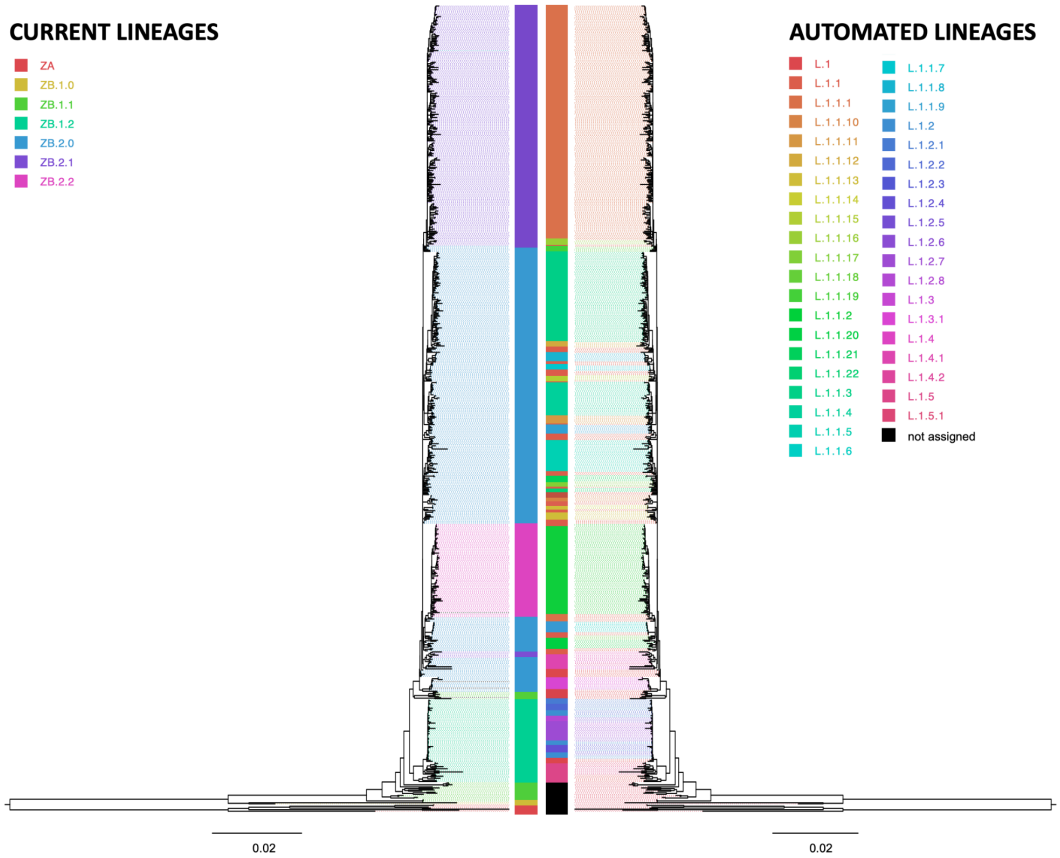

Figure 3.3: Comparison of Zika lineage designation. Comparison of a published proposed lineage system for Zika virus (left tree) based on phylogenetic analyses, clustering techniques, within- and between-group pairwise genetic distances, and evolutionary analyses to define genetic groups (Seabra et al 2022) with automated lineage designation (right tree) visualized on FigTree v.1.4.4.

We analyzed two additional pathogens, Chikungunya virus (CHIKV) and Venezuelan equine encephalitis virus complex (VEE). These phylogenies are provided as Nextstrain Auspice JSON (Hadfield et al 2018), so we used an alternative implementation of Autolin found at https://github.com/jmcbroome/automated-lineage-json (DOI: 10.5281/zenodo.7566925) designed to work with arbitrary Auspice JSON formatted phylogenies. It is provided as both a command line interface tool and as an online Streamlit app, accessible at https://jmcbroome-automated-lineage-json-streamlit-app-3adskh.streamlit.app/. Specifically, we used the currently available nextstrain builds (CHIKV Nextstrain build 5.1 (https://nextstrain.org/groups/ViennaRNA/CHIKVnext) and VEE Nextstrain build 2.1 (https://nextstrain.org/groups/ViennaRNA/VEEnext) to generate our novel lineages.

The rationale for choosing VEE and CHIKV as additional examples stems from their respective lineage systems. The VEE lineage system relies solely on serology, disregarding phylogenetic relationships and displaying paraphyletic groups. Conversely, the CHIKV lineage system is geographically-driven and, although most often presenting monophyletic groups, relies on arbitrary thresholds to define lineages based on location. Overall, the CHIKV geographic nomenclature aligns with the automated lineage designations at its base level (ARI=0.69, p=0.018), with further breaking down of the tree in certain regions such as the Indian Ocean Lineage (Figure A3.4). VEE's serology based nomenclature, by comparison, is paraphyletic and does

not represent phylogenetic lineages or clades (Forrester et al. 2017, de Bernardi Schneider & Wolfinger 2023). We elected to present two levels of annotation, reflecting the distinction between VEE viruses generally and the Venezuelan Equine Encephalitis Virus (VEEV) and its subtypes. VEEV itself is successfully identified from the VEE complex by our lineage approach at the first level of annotation (ARI=0.9, p=0.0003). However, our method was unable to reliably recapitulate VEEV serotypes at the second level of annotation (ARI=0.28, p=0.25, Figure A3.5), due largely to the paraphyletic nature of VEEV's serotype-based nomenclature.

Altogether, these examples show how this method can generate de novo lineage classification of pathogens, independent of context and consistent with human intuition. Moreover, the significance of these examples lies in showcasing the potential advantages of our methodology in mitigating user interference in lineage classification by updating biased nomenclature systems. This, in turn, enhances the possibility of epidemiological discoveries that might otherwise be overlooked. This and similar implementations of the GRI method will be able to support dynamic lineage systems for any future pathogen.

3.7: Mathematical Underpinnings

A lineage system can be formulated as a sender/receiver information scenario. The sender possesses the full phylogenetic tree and a lineage system L, while the receiver possesses only the lineage system L and the associated mutation paths that define each lineage. S may or may not be a member of any lineage within system L.

If it is, the receiver already has all ancestry information associated with that specific lineage L for the sample S. In this scenario, we can compute how much additional information is required to specify the full ancestry of sample S.

A single site's state can be represented in a finite number of bits; 2 bits to represent the state and 15 bits to represent the location, for SARS-CoV-2. Therefore, the full ancestry path of a given branch can be represented in a finite number of bits, proportional to the number of mutations separating it from the root.

$$I(N) = \sum_{1}^{m} 2 + 15$$

Therefore, the additional information required to specify the ancestry of sample S, given a lineage system with a label at branch B, is

$$A(S,B) = \begin{cases} I(S) - I(B) & if \ \ S \in D(B) \\ I(S) & otherwise \end{cases}$$

Where D(B) is the set of samples descended from a labeled branch B.

We further refined this concept to represent instead the average proportion of information about sample S conveyed by a lineage B. This normalization procedure ensures that all samples are treated equally and that the lineage itself is an effective representation of the member samples. By normalizing to total distance, a cluster of samples near the reference will be treated the same as a similar group of samples positioned further from the root, given that the groups are similarly distinct from their last respective lineage labels. We therefore compute the following:

$$P(S,B) = \begin{cases} \frac{I(S)-I(B)}{I(S)} & if\, S \in D(B) \\ 1 & otherwise \end{cases}$$

We extend this to compute the total amount of information for a system of multiple lineage branches B. These may be hierarchically arranged, where a single sample is descended from multiple, nested lineage labels B; in this case, the minimum value is taken.

$$Y = \{B_1, \ldots, B_n\}$$

$$O(Y) \;=\; \sum_{S \in T} \min(\{P(S,B) : B \in Y\})$$

When adding a new branch B to this system, we can compute the difference in overall information represented by this addition. Adding a lineage B will always either reduce O(Y) or leave it the same, as any altered values summed to O(Y) are replaced by a smaller value.

$$Y' = \{B_1, \ldots, B_{n+1}\}$$

$$O(Y') \leq O(Y)$$

The difference between O(Y') and O(Y) can be computed as the sum of differences in P(S,B) for all samples where B_{n+1} is the terminal lineage of that sample; that is, where $P(S,B_{n+1})$ is the minimum value of P(S,B) for all B. For all other values, P(S,B) is identical, and therefore can be disregarded.

$$O(Y) - O(Y') = \sum_{S \in T} (\min(\{P(S,B) : B \in Y\}) \; - P(S, B_{n+1}))$$

Our goal is to choose the value of B_{n+1} that maximizes the overall difference. Therefore, we can disregard the first term, since it remains unchanged by the choice

of B_{n+1}. The difference between the systems is negatively proportional to the remaining term.

$$O(Y) - O(Y') \propto -\sum_{S \in T} P(S, B_{n+1})$$

Samples S where B_{n+1} is not the terminal lineage will be valued the same regardless of the choice of terminal lineage, so this can be further reduced for the purposes of comparison.

$$O(Y) - O(Y') \propto -\left(\sum_{S \in D_t(B_{n+1})} P(S, B_{n+1}) \right)$$

Where Dt is the set of samples for which B is the terminal lineage. As all samples in Dt(B) are necessarily members of D(B), this is equivalent to the following:

$$O(Y) - O(Y') \propto -\left(\sum_{S \in D_t(B_{n+1})} \frac{I(S) - I(B_{n+1})}{I(S)} \right)$$

The I(S)/I(S) term simplifies to 1, which can be disregarded when comparing these values between different choices of Bn because each contains the same scalar. Simplification leaves a term to which the difference between the systems is directly proportional.

$$O(Y) - O(Y') \propto -\left(\sum_{S \in D_t(B_{n+1})} 1 - \frac{I(B_{n+1})}{I(S)} \right)$$

$$O(Y) - O(Y') \propto \sum_{S \in D_t(B_{n+1})} \frac{I(B_{n+1})}{I(S)} - 1$$

$$O(Y) - O(Y') \propto \sum_{S \in D_t(B_{n+1})} \frac{I(B_{n+1})}{I(S)}$$

In practice, we often track the information about the branch I(B) and the distances to the descendent samples S from that branch B as explicit quantities.

$$F(S, B) = I(S) - I(B)$$

$$O(Y) - O(Y') \propto \sum_{S \in D_t(B_{n+1})} \frac{I(B_{n+1})}{F(S, B_{n+1}) + I(B_{n+1})}$$

This equation is the basis of the Autolin heuristic, which is a computationally practical representation of lineage information content.

3.8: GRI and the Autolin Algorithm

We want to avoid computing the set of samples D_t for each node B on the tree explicitly, as this requires either repetitive traversal or storing large arrays of values. The only dependent term on this set of samples S is $F(S,B_{n+1})$. We therefore replace this term by dynamically computing the mean $F(S,B_{n+1})$ for all samples S and multiplying the entire equation by the number of descendents, meaning we only have to compute this overall equation once. While this is not exactly equivalent to the sum, except under special conditions, it is strongly correlated with it and can reduce the effect of outlier samples on the overall computation.

$$O(Y) - O(Y') \propto |D(B_{n+1})| \cdot \frac{I(B_{n+1})}{\frac{\sum_{S \in D(B_{n+1})} F(S,B_{n+1})}{|D(B_{n+1})|} + I(B_{n+1})}$$

This allows us to only track three values for each node- the sum of distances F(S,B), the number of descendents |D(B)|, and the information of the branch I(B), and only perform a single computation. The sum of F(S,B) and the number of descendents |D(B)| can both be dynamically computed by a single reverse postorder traversal of the tree and stored as single float values.

$$SUM(B) = \begin{cases} branch\,length\,of\,B & if\,children(B) = \emptyset \\ \sum_{C \in children(B)} SUM(C) & otherwise \end{cases}$$

$$COUNT(B) = \begin{cases} 1 & if\,children(B) = \emptyset \\ \sum_{C \in children(B)} COUNT(C) & otherwise \end{cases}$$

I(B) can be dynamically computed by a single forward traversal, as the branch length I(B) is equal to the branch length of B plus the information of its parent. We perform one pass to compute the sum and count values, and we track I(B) on the forward pass where candidate nodes are evaluated. With these values for each node, we can compute the following:

$$GRI(B) = \frac{COUNT(B) \cdot I(B)}{\frac{SUM(B)}{COUNT(B)} + I(B)}$$

Notationally, we use single letters to refer to the values of these functions for a branch B.

$S = SUM(B)$

$N = COUNT(B)$

$D = I(B)$

$$GRI = \frac{N \cdot D}{\frac{S}{N} + D}$$

This final equation is the GRI heuristic we use to select our lineages. It does not require identifying the explicit set of descendent samples D(B), which for large phylogenies either requires storing large vectors in memory or repeated tree traversal, instead using single values for the sum and count. It also has useful properties; it can never have a higher value than N, limiting the effect of extremely long branches, and approaches 0 as S becomes large, where the lineage proposal would be a poor representative of its descendents.

$$\lim_{D \to \infty} \frac{N \cdot D}{\frac{S}{N} + D} = N$$

$$\lim_{S \to \infty} \frac{N \cdot D}{\frac{S}{N} + D} = 0$$

In the simplest case, the construction of a lineage system will involve the stepwise addition of lineage labels. Finding the overall system which maximizes the relative gain for multiple simultaneous lineage definitions is excessively complex and unscalable for systems of more than a handful of lineages, due to the extremely high number of possible combinations of lineage labels to evaluate. However, a system of arbitrary size can be constructed efficiently through a simple greedy stepwise algorithm, where the best choice for each step is taken without regard for the impact on potential future choices. Therefore, our implementation computes this metric for every node on the tree, assigns a new lineage at the highest value node, and then

repeats this process until no candidates pass minimum thresholds set by the user.

Serial" or non-overlapping lineages, where

$$D(L_1) \cap D(L_2) = \varnothing$$

Can be assigned by repeating the minimization procedure while disregarding all samples that are a member of existing lineages. This can be repeated until some minimum percentage of samples are contained within some set D(L).

"Hierarchical" or nested lineages, where

$$D(L_2) \subseteq D(L_1)$$

Can be assigned by treating L_1 as the root of the tree, with ancestry information conveyed with respect to it. There are no other types of lineage relationship, as a rooted phylogenetic tree is a directed acyclic graph and lineages are always defined as a monophyletic clade. It is not possible for two clades to partially overlap when they are defined by internal nodes on a fixed phylogenetic tree.

There is one obvious failure case with this model; if the number of lineage labels B is not limited or penalized, every node in the tree can be given individual labels, reproducing the original phylogeny and all accompanying information exactly in the lineage system. However, this degenerate case is not desirable, as the goal of lineage systems is generally to compress phylogenetic information to a more manageable set of groups while keeping key elements. Two simple restrictions are a minimum lineage size and a minimum distinction from the parental lineage or root.

To require a minimum number of samples to be represented by a putative lineage label, we define a minimum *m* and we subtract the weighted mean

information represented by a theoretical set of m samples with the same path length distribution from the true information distribution for the node. If the net information represented is negative, then we reject this node as a candidate for a new lineage definition. We define the following inequality:

$$\frac{(N-m)\cdot D}{\frac{S}{N}+D}>0$$

Essentially, we require that $N > m$, where m is a user selected parameter, in order to define a new lineage. Setting this to a positive value will produce only proposed lineages that convey some information about at least that many leaves.

Similarly, we can set a minimum distinguishing distance from the subtree root/parent lineage. Often lineage designation systems require some number of unique distinguishing mutations for a new sublineage. We therefore define

$$\frac{N\cdot(D-p)}{\frac{S}{N}+D}>0$$

When $p < D$, this value is negative and we reject this candidate node. Setting this to two, for example, will produce only lineages that convey at least two unique mutations distinct from the parent lineage or tree root. Combining both of these filters, we reject nodes where either or both of these inequalities are not passed. Together, this allows automatic proposals to fulfill standard conditions required by lineage nomenclature review groups.

Our pipeline implementation includes a substantial set of configurable parameters. These include minimum lineage size and minimum distinction, as

outlined above. We also can simply threshold on the GRI itself, ignoring marginal designations that contain relatively little additional information.

Notably, we can additionally incorporate arbitrary sample-level weighting. This allows our lineage system to prioritize effective representation of high-interest samples. R(S), below, is a function representing the "importance" of sample S. This might be high for a sample S from an undersequenced region, or lower for a sample S from a heavily sequenced time or place.

$$W = \sum_{S \in D(B)} R(S)$$

$$\frac{W \cdot D}{\frac{S}{N} + D}$$

Samples from regions that contribute a small percentage of all samples will have substantially higher weights than ones from regions that contribute a large percentage of sequences, though all samples will have a weight greater than 1 under this schema. This is just one potential weighting schema for handling geographic sequencing bias, and the user can define any schema and set weights on a per-sample basis.

Similar concepts can apply to computing path lengths- we may consider only part of the haplotype, or assign additional weight to specific mutations of interest that we want our lineage system to prioritize representing. We provide options for the user to select genes of interest for representation, as well as the ability to ignore mutations

that do not change amino acid content of proteins and represent coding haplotypes only.

We also provide arbitrary weighting schema for mutations of interest, similar to samples.. As an example, we provide a parameter that heavily weights mutations that are predicted to increase vaccine escape (Greaney et al 2022). This parameter multiplies the escape weight value estimated by the Bloom lab calculator by the user's parameter and adding 1. In this schema, mutations that are not predicted to contribute to immune escape have a weight of 1, while mutations that do contribute have a weight greater than 1 that is proportional to the strength of escape conferred. The resulting lineage system is more likely to include designations that have a change in immune escape. This is just one possible schema and the user can define weights on a per-mutation basis in our implementation.

All parameters and configuration information used in the production of these results can be found at 10.5281/zenodo.7566938.

3.9: Bayesian Growth Modeling

Our simplified Bayesian growth model is a geographically stratified estimate of a fundamental rate of exponential growth over a weekly time series. For lineage L in country C, we model the true percentage P as increasing in an approximately exponential fashion. This is appropriate for newly emerging lineages that consist of a small percentage of total cases in any country where they are found but are successfully spreading. Each data point consists of the total number of samples from

lineage L found in a specific country during a specific week. We assume that the inherent exponential growth coefficient for L is shared across all countries in which it is found and combine all data points across countries and times for each lineage. The first week that any sample from lineage L was found in country C is treated as the initial timepoint (t=0) for data from that country.

We do not directly observe the true percentage of cases P that are of lineage L. Instead, some number N of all cases are sequenced, and we observe some number X of these samples to be lineage L. As the number of cases is much larger than the number of samples, we can model this process as a binomial sampling procedure with N trials and a probability of success being the true percentage P.

Our Bayesian model combines both this sampling procedure and the exponential growth model to yield a posterior distribution of growth values which can explain the behavior of lineage L. Often these distributions are wide, due to sparse sampling and noise over few datapoints. Additionally, some lineages may not fit an exponential growth model at all, due to being outcompeted by newly introduced lineages or simple epidemiological noise, leading to highly variable estimates of growth. Accordingly, we compute the 0.025 and 0.975 quantiles (95% CI) for this distribution for each lineage L and sort the output by the lower quantile. Lineages with a large positive value for the lower quantile will reliably resemble a high exponential growth model and are more likely to be of epidemiological concern.

All code for our modeling and reporting process can be found at https://github.com/jmcbroome/lineage-manuscript (DOI: 10.5281/zenodo.7983421) and https://github.com/jmcbroome/autolin (DOI: 10.5281/zenodo.7566921).

3.10: Discussion

We have presented a new index-based method, capable of both expanding existing dynamic lineage systems and generating novel lineage designations for understudied or emerging pathogens. Originally designed for the demands of the SARS-CoV-2 pandemic, this approach can be easily applied to any rooted tree with branch lengths scaled by genetic distance. Our implementation is efficient and includes several parameters to adjust the behavior of the metric, including prioritizing the labeling of specific mutations or specific samples and only considering mutations with effects on specific proteins.

Nonetheless, our approach does exhibit a few potential issues, shared with many lineage nomenclatures. First, it is defined with respect to a specific phylogeny. This can be problematic when attempting to maintain lineages over time, as new data is collected and the phylogeny is updated. Phylogenetic inference is naturally uncertain, and optimization of an existing phylogeny may alter lineage relationships or invalidate identified lineages. In rare cases, lineages may need to be retracted or redefined, as is the case for current Pango lineages when new data suggest alternative relationships than the one originally used for lineage designation. While these lineages are generally stable (see Appendix 3), spuriously duplicated samples due to

redundancy between data sources can lead to inflated lineage counts or spurious lineage definitions. Appropriate filters, such as removing low-quality or duplicate samples from the input tree, will be necessary to ensure the stability and viability of these lineage systems.

Second, SARS-CoV-2 recombines at low rates (Jackson et al. 2021; Turakhia et al. 2022). The apparently long branches which occur on the phylogeny as a result of recombination between genetically distinct lineages will often be picked up by this method as a new lineage annotation, but the ancestry of that lineage annotation cannot be accurately represented by a single tree topology. In this scenario, a recombinant lineage may have to be retracted or renamed. Alternatively, lineages identified as recombinants can receive special designation names. We have previously developed methods for comprehensively identifying recombinant lineages within SARS-CoV-2 phylogenies (Turakhia et al. 2022) that may facilitate this effort in the future.

One fundamental challenge for SARS-CoV-2 genomic analysis is variation in sequencing among different parts of the world. The United States of America and the United Kingdom contribute a massive quantity of data to public repositories compared to many other countries, and so strains of the virus specifically circulating in these countries may appear more important if this bias is not corrected for. We provide options for correction of these biases at multiple steps of our pipeline. First, we provide methods for users to indicate the base weight that should be assigned to individual samples, allowing users to emphasize samples from particular regions or with particular attributes for representation. We additionally provide a built-in

category-frequency-based weighting scheme for these samples, where samples labeled as from being from a rare group are given proportionally higher weight; we have applied this with country as the category, leading to the designation of lineages specifically representing strains from less well surveilled countries. Second, we explicitly build regional sequencing bias into our growth modeling, by stratifying the growth curve by country and normalizing to the total sequence contribution from that country for that time period. While this may introduce some noise into the model, it means that rapid growth in two countries with uneven sequencing will be treated equally. In any case, user discretion in sample weighting parameters used will be important to mitigate regional sequencing bias.

SARS-CoV-2 is likely to become an endemic pathogen, similar to the influenza virus (Otto et al. 2021). Accordingly, there is likely to be a long-term pattern of replacement of existing strains, demanding ongoing designation of new lineages for effective monitoring of pathogen diversity (Rambaut et al. 2020). Investing into infrastructure to reduce manual curation will lead to long-term consistency and effectiveness of designation. Additionally, it is likely that automated approaches will be faster than many human-based systems, thereby promoting stability of public and scientific discourse by labeling potentially important lineages before they are widespread and contributing to major epidemiological patterns worldwide. The results we present here may serve for consistent, immediate SARS-CoV-2 lineage designation for years to come.

Overall, this approach for lineage designation is generic, flexible, and applicable to future datasets with unclear nomenclature or expansive phylogenies. With global pathogen sequencing on the rise, and generalized toolkits for the creation and maintenance of dynamic lineage systems will be critical for future public health challenges.

Conclusion

In the first chapter, I outlined the mutation annotated tree (MAT) and described the toolkit I designed for the manipulation of these structures. matUtils, along with the rest of the Online Phylogenetics Toolkit, was quickly adopted by the SARS-CoV-2 research community, with more than 250,000 downloads as of January 2023. BTE further expanded the capabilities of our codebase by making it accessible in Python, one of the most popular scientific programming languages.

In the second chapter, I presented a heuristic approach I developed for the identification of geographically localized SARS-CoV-2 transmission clusters. Standard tools were designed for small, sparse datasets with no demand for immediate output, and failed to produce rapid results to inform public health action. My method and the accompanying website, ClusterTracker, were adopted as well by the SARS-CoV-2 public health community, and used directly by the California Department of Public Health.

In the third chapter, I described an information theory informed approach to the identification of novel pathogen lineages. While the major implementation of this

method was focused on identifying and labeling new SARS-CoV-2 strains, the core approach itself is easily generalized to arbitrary pathogens. I applied it to generate novel nomenclature for two understudied pathogens, VEE and CHIKV, demonstrating its capacity to serve as a general framework for the definition of subspecies pathogen nomenclature. This method was adopted by the Pango team and has become part of the core workflow that underlies almost all SARS-CoV-2 track and trace procedures.

The COVID-19 pandemic has been an incredibly challenging time for many people around the world. It has been incredibly limiting in some respects; compared to most graduate students, I have had little opportunity to attend conferences and network with other academic researchers. I was largely unable to pursue my original research program and day to day collaboration with my fellow graduate students was rare. For me, however, it was also a great opportunity. The COVID-19 bioinformatics bottleneck opened opportunities for high impact publications and contributions. Most graduate students can only expect their theses to contribute in some small way to a niche area of interest; my work, by contrast, has had direct, worldwide impact. I have been given opportunities to work with public health groups and international researchers afforded to few. This work will form the foundation of public health genomics for years to come.

Appendix 1

Chapter 1 Supplementary Material

Group	Number of clades/lineages	Training Size	Test Size	Mean Training Accuracy	Minimum Training Accuracy	Mean Test Accuracy	Minimum Test Accuracy
Nextstrain	14	651446	184877	0.973	0.972	0.971	0.97
Pangolin	895	651446	184877	0.881	0.881	0.881	0.88

Table A1.1: matUtils annotate can quickly and effectively assign clade lineage roots. This table was generated by taking training data associated with Nextstrain clades and Pango lineages from our public repository (lineageToPublicName.gz and cladeToPublicName.gz), splitting the data 80/20 into training and test sets, and assigning roots based on the 80% selected training data with matUtils annotate on the 06-09-2021 public MAT tree. Accuracy was scored as the percentage of the training or test set which matches Nextclade or Pangolin assignments. This process was repeated 9 times and mean and minimum accuracy values were collected.

Program	Command	Time (M:S)	Memory (kB)
matUtils	matUtils summary -i public-2021-06-09.all.masked.nextclade.pangolin.pb	0:05.65	987616
matUtils	matUtils extract -i public-2021-06-09.all.masked.nextclade.pangolin.pb -S sample-paths.txt	0:13.90	977792
matUtils	matUtils summary -i public-2021-06-09.all.masked.nextclade.pangolin.pb -A	0:15.87	984360
newick_utils	nw_stats public-2021-06-09.all.masked.nextclade.pangolin.nwk	0:01.14	218920

Table A1.2: Time and memory usage to summarize the tree.

Program	Command	Time (M:S)	Memory (kB)
matUtils	matUtils extract -i public-2021-06-09.all.masked.nextclade.pangolin.pb -R -t public-2021-06-09.all.masked.nextclade.pangolin.resolved.nwk	0:09.43	1224652
ape	t1=read.tree("public-2021-06-09.all.masked.nextclade.pangolin.nwk") t1b<-multi2di(t1) write.tree(t1b,file="public-2021-06-09.all.masked.nextclade.pangolin.resolved.nwk")	37:30.56	735688

Table A1.3: Time and memory usage to resolve all polytomies in the tree.

Program	Samples	Command	Time (M:S)	Memory (kB)
matUtils	70	matUtils introduce -i public-2021-06-09.all.masked.nextclade.pangolin.pb -s spanish_samples.txt -o spanish_introductions.txt	0:25.93	976540
matUtils	13302	matUtils introduce -i public-2021-06-09.all.masked.nextclade.pangolin.pb -s aus_samples.txt -o aus_introductions.txt	0:31.23	982104
matUtils	39678	matUtils introduce -i public-2021-06-09.all.masked.nextclade.pangolin.pb -s welsh_samples.txt -o wales_introductions.txt	0:33.67	1003856
matUtils	296257	matUtils introduce -i public-2021-06-09.all.masked.nextclade.pangolin.pb -s usa_samples.txt -o usa_introductions.txt	0:37.12	1271304
matUtils	398396	matUtils introduce -i public-2021-06-09.all.masked.nextclade.pangolin.pb -s british_samples.txt -o british_introductions.txt	0:33.63	1338316

Table A1.4: Time and memory usage to calculate introduction statistics for subsets of samples within the tree.

Program	Command	Time (M:S)	Memory (kB)
matUtils	matUtils extract -i public-2021-06-09.all.masked.nextclade.pangolin.pb -t public-2021-06-09.all.masked.nextclade.pangolin.nwk	0:06.38	976628
matUtils	matUtils extract -i public-2021-06-09.all.masked.nextclade.pangolin.pb -v public-2021-06-09.all.masked.nextclade.pangolin.vcf	25:38.85	23406404

Table A1.5: Time and memory usage to convert into Newick and VCF formats.

		Replicate 1		Replicate 2		Replicate 3		Average	
Program	Samples Calculated	Time (M:S)	Memory (kB)	Time (M:S)	Memory (kB)	Time (M:S)	Memory (kB)	Time (M:S)	Memory (kB)
matUtils	100	1:23.65	2015864	1:18.13	1905420	1:17.99	1798000	01:19.9	1906428
matUtils	500	6:21.72	2166900	6:23.78	2159228	6:21.05	2406620	06:22.2	2244249
matUtils	1000	12:45.96	2395516	12:41.89	2201184	12:30.68	3130644	12:39.5	2575781

Table A1.6: Time and memory usage to determine equally parsimonious placements for subsets of samples in the tree.

The above tables, along with some additional benchmarking information too lengthy for inclusion here, can be found at 10.5281/zenodo.7566973.

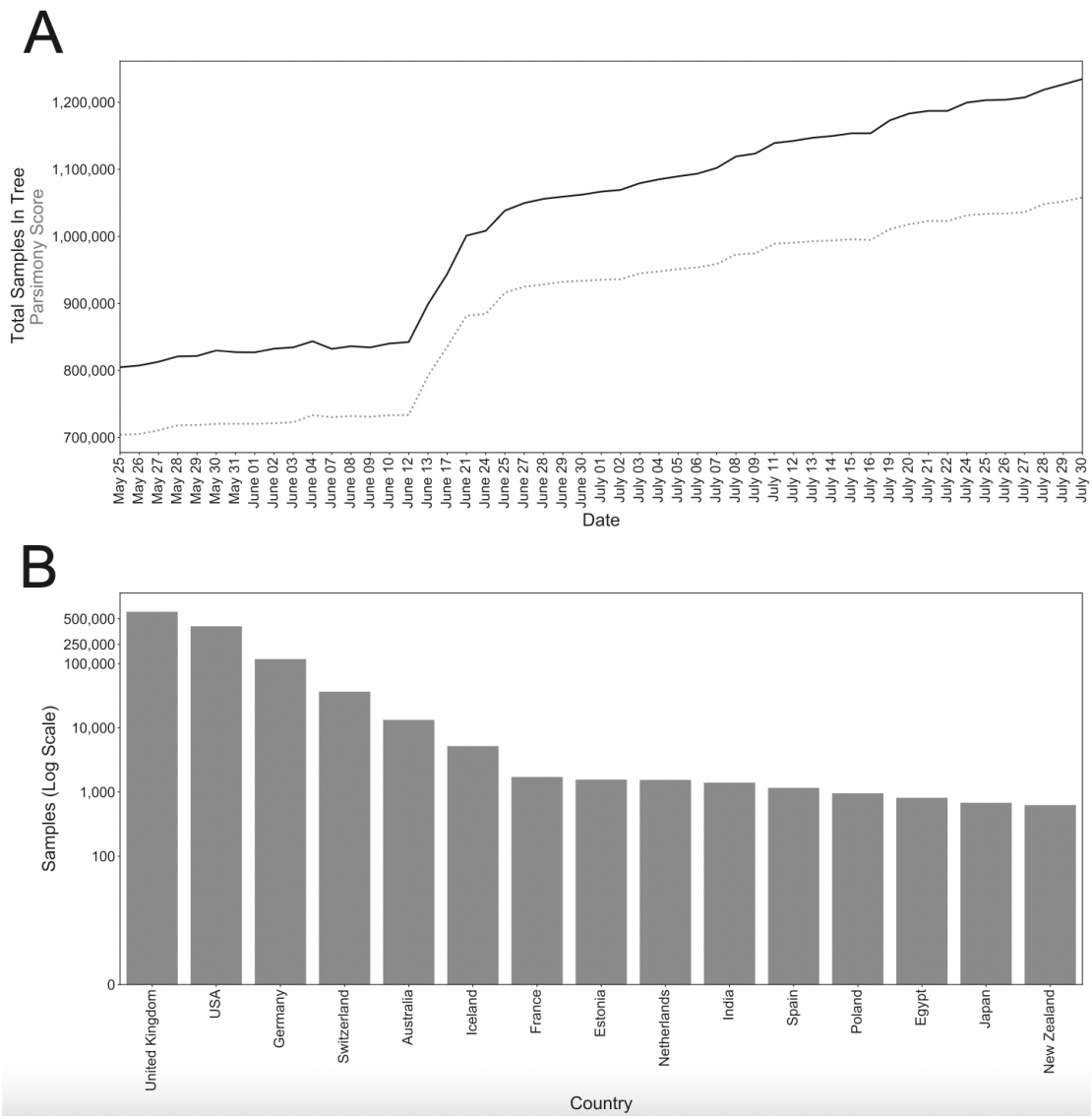

Figure A1.1: Our global phylogeny contains 1,234,612 samples as of July 30, 2021. (A): Our database, containing all high-quality publically available SARS-CoV-2 whole-genome sequences and their clade assignments, is updated daily. As of July 30, our phylogeny contains 1,234,612 sequences with a total parsimony score of 1,058,062. Sequences that have 5 or more equally parsimonious placements on the tree are removed at each build (Supplementary Methods), so the total samples sometimes drop during successive builds. (B): Distribution of samples from July 30, 2021, phylogeny based on their country of origin. A few countries, such as the United Kingdom and the United States, have contributed a disproportionately large fraction of sequences to the public SARS-CoV-2 databases.

Appendix 2

Chapter 2 Supplementary Material

Sample Count	Samples In-Region	Time (seconds)
100	25	0.02106654644
1000	250	0.03384798765
10000	2500	0.2106143832
50000	12500	0.9356530309
100000	25000	1.890872002
500000	125000	9.770167351
1000000	250000	19.75107902
2500000	625000	37.4885782

Table A2.1: Basic benchmarking information for our method. For this benchmark, we took the public tree obtained on 11-08-21 and randomly generated subtrees containing a set number of samples. We further selected at random 25% of these samples to be considered in-region, under a single region model. We find that runtime is approximately linear with the number of samples in the tree (which, in turn, is correlated with the number of nodes in the tree). Even for a tree of two and a half million samples and a region with 625,000 samples, a single region on a single thread doesn't take more than one minute to compute our heuristic for.

Scale	Migration Rate	Nodes Collapsed	Mutations Per Node	Parsimony	ARI	Tree Depth	IAC
0.001	0.001	569615	1.24	2476860	0.94	19.85	0.998
0.001	0.005	589133	1.16	2316475	0.93	13.8	0.991
0.001	0.01	580482	1.2	2393013	0.89	14.16	0.984
0.001	0.05	583369	1.2	2399857	0.15	17.65	0.943
0.005	0.001	209173	6.08	12166774	0.99	26.86	0.998
0.005	0.005	212422	6.0	11970552	0.95	28.76	0.993
0.005	0.01	220101	5.754	11494897	0.91	25.35	0.987
0.005	0.05	204296	6.47	12746185	0.47	28.995	0.950
0.01	0.001	116730	12.21	24416981	0.998	28.54	0.998
0.01	0.005	123025	11.38	22754698	0.95	27.53	0.993
0.01	0.01	111625	12.78	25561431	0.43	30.55	0.988
0.01	0.05	118963	12.1	24166836	0.43	28.47	0.951

Table A2.2: Results from a set of simulations generated via PhastSim and VGsim (see Methods). For reference, the real tree that we considered in this work had an overall parsimony of 4,847,954 and a mean tree depth of 35. “Scale” is the parameter passed to phastSim --scale, representing a scalar applied to the branch lengths to rescale. Smaller values of scale imply fewer mutations per site per branch. “Migration Rate” is a reciprocal value representing the rate of migration events between two equally-sized “regions” under simulation. “Nodes Collapsed” is the number of nodes which have no mutations on their branch after phastSim, resulting in their collapse with their parent. “Parsimony” is the total tree parsimony score, or the count of all mutations across all branches, and also reflects mutations per node and scale. “Tree Depth” is the mean distance in mutations between the root of the tree and a leaf. ARI, or adjusted rand index, is the computed adjusted rand index for sample cluster labels on the final collapsed tree versus the true clusters. True clusters here are defined as the set of samples which share a single true migration event into their region at their common ancestor. IAC stands for “internal assignments correct”, or the proportion of internal nodes which have their true regional states correctly assigned by the heuristic on the uncollapsed, bifurcating tree.

Scale	Migration Rate	Nodes Collapsed	Mutations Per Node	Mutation Parsimony	IAC	Parsimony IAC	ARI	Parsimony ARI
0.0001	0.01	1114193	0.11	223737	0.99	0.94	0.0031	0.00091
0.0001	0.05	1137090	0.12	230852	0.95	0.83	0.0033	0.173
0.0001	0.1	1203907	0.11	214417	0.92	0.77	0.0087	0.0096
0.0001	0.25	1287683	0.12	227881	0.88	0.76	0.00067	0.018
0.0005	0.01	940046	0.56	1112307	0.994	0.96	0.77	0.0069
0.0005	0.05	976594	0.56	1116632	0.96	0.82	0.43	0.017
0.0005	0.1	1084199	0.48	968699	0.94	0.83	0.04	0.0239
0.0005	0.25	1188100	0.49	971192	0.895	0.81	0.006	0.028
0.001	0.01	766682	1.19	2377529	0.99	0.90	0.85	0.0015
0.001	0.05	910117	0.99	1987227	0.98	0.85	0.51	0.01
0.001	0.1	909412	1.09	2183873	0.95	0.84	0.23	0.018
0.001	0.25	1149881	0.81	1631128	0.90	0.84	0.021	0.021
0.005	0.01	494106	5.71	11433065	0.99	0.99	0.94	0.115
0.005	0.05	565153	5.61	11216328	0.97	0.95	0.53	0.51
0.005	0.1	525005	6.68	13350037	0.96	0.95	0.33	0.43
0.005	0.25	960374	3.79	7576739	0.913	0.9	0.0997	0.13

Table A2.3: Efficacy on Simulated Data. This table is similar to Table A2.2 with several columns in common, including “scale”, “migration rate”, “nodes collapsed”, and “mutations per node”. For clarity here tree parsimony is referred to as “mutation parsimony”. “IAC” again refers to “internal assignments correct” on the full, correct, bifurcating tree structure, from our method and from parsimony reconstruction respectively. Our method is better than parsimony across most conditions at correctly recovering internal node states under the simulated parameters. The Adjusted Rand Index (ARI) is a measure representing how effectively true descendent clusters were recaptured, with scores closer to 1 being more accurate. Again, our method outperforms parsimony when there are few mutations per node; once the tree is mostly resolved with several mutations per node, their efficacy becomes comparable. Overall, our method is more robust than parsimony approaches when applied to phylogenetic trees with few mutations relative to total sampling, as in the case of SARS-CoV-2.

Tables A2.4, A2.5 and Data A2.1: Inferred introduction counts to each of the fifty United States from international sources, inferred introduction counts between each of the fifty United States, and full sample accreditation, respectively. These large data files can be found at 10.5281/zenodo.7566973.

Appendix 3

Chapter 3 Supplementary Material

A3.1: Lineage Stability Analysis

To identify any potential issues with respect to the stability of lineages defined by Autolin with respect to the SARS-CoV-2 tree, we identified new lineages via Autolin and tracked the phylogenetic placements of the associated samples over the following month. We began with the global public phylogeny as of 2023-04-01, and computed a set of 65 new lineage annotations with Autolin. The new designations covered 8,951 samples, with a median size of 50 samples per lineage and the majority being sub-variants of the XBB lineage. We then transferred these designations to each successive daily tree through 2023-04-30 using matUtils annotate and tracked lineage membership of the initial 8,951 samples. By the end of the month, the set of samples covered by all 65 lineages grew to 11,877.

Over the course of the month, only 28 samples changed lineages (0.23% of samples), affecting 4 of the 65 designations. auto.XBB.1.5.40 was the primary affected lineage, with 9 samples being added to it and 7 samples being removed, but these 16 samples only constitute 0.7% of the 2119 samples in this large lineage. A full table of samples which change lineages can be found in Table A3.2. Overall, we

observe high stability in our lineage designations with regards to samples that remain consistently present on the tree.

However, some of these apparent lineage changes and related stability issues result from duplicated, dropped, or renamed sample data. Several hundred samples were dropped from the global phylogeny on 2023-04-13 and 2023-04-14, affecting 38 of the 65 lineages. In some cases, this resulted in an apparent decline in lineage membership; the lineage designation “auto.XBB.1.29.1” began with 12 samples on 2023-04-01, but declined to only 7 by 2023-04-30. It is likely that the majority of these were replaced with a new name; for example, the sample OX451312.1, a member of auto.XBB.1.29.1, was dropped from the tree on 2023-04-13. Present throughout this period, also within auto.XBB.1.29.1, was the sample Scotland/SCOT-26390/2023|OX451312.1|2023-02-08, with the same tag and date of collection. It’s likely that the inclusion of OX451312.1 represents a spurious duplication of this sample within the global phylogeny, perhaps because it was uploaded separately to different public databases. Of the 12 original samples of auto.XBB.1.29.1, 5 are dropped on the 13th while matching fuller names with the same collection date and tag information in auto.XBB.1.29.1. Therefore, the original 12 sample set represents an inflated, spurious group and this marginal lineage designation must be dropped or revised. It is worth noting, however, that this lineage remained present throughout the period and the deduplicated samples remained stable members of this group.

SARS-CoV-2 is a densely sampled pathogen and most branches are well supported by many samples worth of data. Additionally, samples incorporated into the SARS-CoV-2 global public phylogenetic trees used for lineage designation are rigorously quality filtered to remove low quality consensus sequences. Generally, lineage designations of more than a few dozen samples are largely stable, with <1% of samples overall changing designations over the course of weeks. Even in the case of recombination, where the base of a lineage cannot be properly represented in a tree, the group itself is generally stable and cohesive. All code for this analysis can be found at https://github.com/jmcbroome/lineage-manuscript (DOI: 10.5281/zenodo.7983421).

A3.2: Methods for Application to Other Pathogens

To validate that this method can be applied to pathogens other than SARS-CoV-2, we selected two nextstrain instances for Chikungunya virus and the Venezuelan Equine Encephalitis complex viruses, which are currently classified based on their geography and serology, respectively. We applied our generalized implementation (https://github.com/jmcbroome/automated-lineage-json (DOI: 10.5281/zenodo.7566925)) under default settings for the Auspice JSON files of each virus (CHIKV Nextstrain build 5.1 available at https://nextstrain.org/groups/ViennaRNA/CHIKVnext (doi:10.5281/zenodo.7514289) and VEE Nextstrain build 2.1 available at https://nextstrain.org/groups/ViennaRNA/VEEnext (doi:10.5281/zenodo.7524848)) to

obtain lineage assignments. These Nextstrain JSON were generated by the Augur pipeline (nextstrain-augur v19.1.0, treetime v 0.9.4, iqtree v2.2.0). We then downloaded the nexus file with annotations from the new JSON file from Nextstrain and visualized and compared the annotations using FigTree v.1.4.4. Tree figure comparisons were made by extracting them in pdf format as displayed in FigTree, mirrored and aligned on a photo editing software. Taxon labels were colored according to the lineage assignment and were replaced with bars representing the color of the lineage for best visualization.

There was no available Auspice build for the ZIKV nomenclature (Seabra et al 2022). We therefore had to construct a MAT to make a file compatible with Autolin. We obtained the phylogeny directly from the authors and sample names and lineage assignments from their Supplementary Table 3. We downloaded sample sequences using the Entrez API and aligned them to the same Zika reference (KJ776791) used by Seabra et al with minimap2 (Li et al 2018) to produce a VCF. We then combined this VCF and their likelihood phylogeny into a MAT with likelihood branch lengths using UShER (Turakhia et al 2021). We applied Autolin to this MAT with a minimum lineage size of 3 and a minimum distinction (distance in total branch length from the last annotated lineage) of 0. Finally, we extracted the new lineage annotations for each sample using matUtils (McBroome et al 2021). Strictly, the mutations inferred did not affect this process, as the GRI is dependent on the branch lengths, but constructing the MAT was necessary to make the data compatible with the Autolin implementation of the GRI. All code to reproduce this process can be

found at https://github.com/jmcbroome/lineage-manuscript/zika-lineages (DOI: 10.5281/zenodo.7983421). Figure 3.3 was produced as described above with FigTree.

We compared the automated lineage assignments with the previous nomenclature using the Adjusted Rand Index (ARI). We randomly selected nodes in the amount of the number of categories found for each annotation to create a distribution of random ARI's to evaluate the robustness of the method. By selecting random nodes within the tree and taking their descendants to construct our null comparisons, we account for natural correlation from the tree structure, while the Adjusted Rand Index itself accounts for variations in group sizes. We then compute the percentile of the true Adjusted Rand Index of our lineage proposals against the existing nomenclature from the permuted null distribution, yielding the reported p-values. All code for this can be found at https://github.com/jmcbroome/lineage-manuscript (DOI: 10.5281/zenodo.7983421).

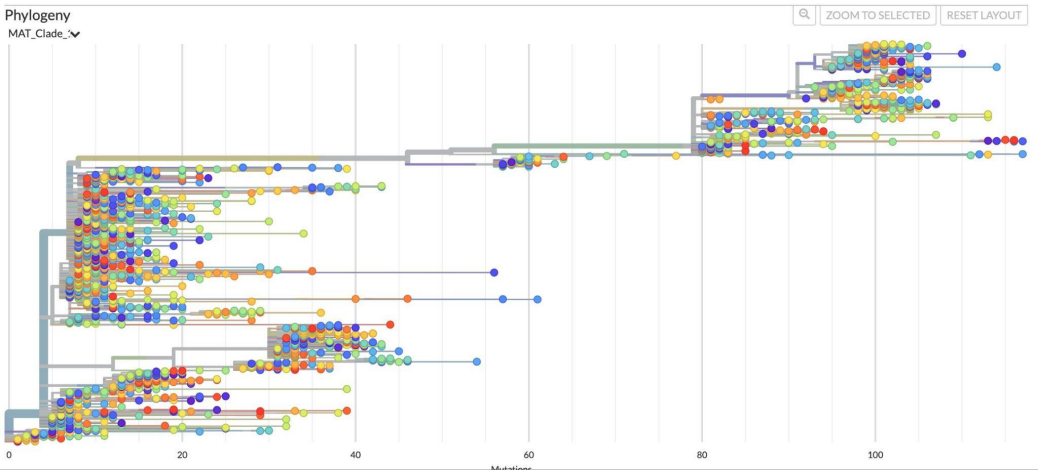

Figure A3.1: Pango Lineage Hierarchy. This figure displays the hierarchical and serial relationships among the defined Pango lineages as of 2022-12-11. Each individually colored tip (dot) on this tree represents a specific Pango lineage.

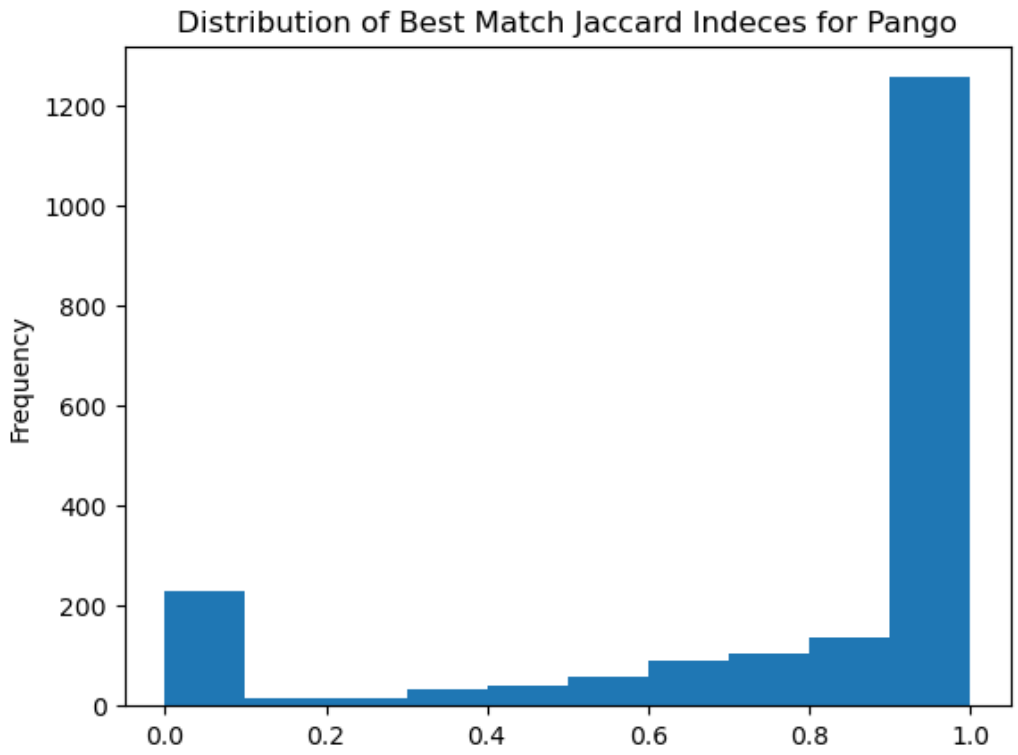

Figure A3.2: Jaccard Index Distribution for Pango Lineages. The distribution is highly bimodal, with a plurality of lineages being perfectly or partially matched by an automatically identified lineage, but with a substantial body of Pango lineages with no strong matching label.

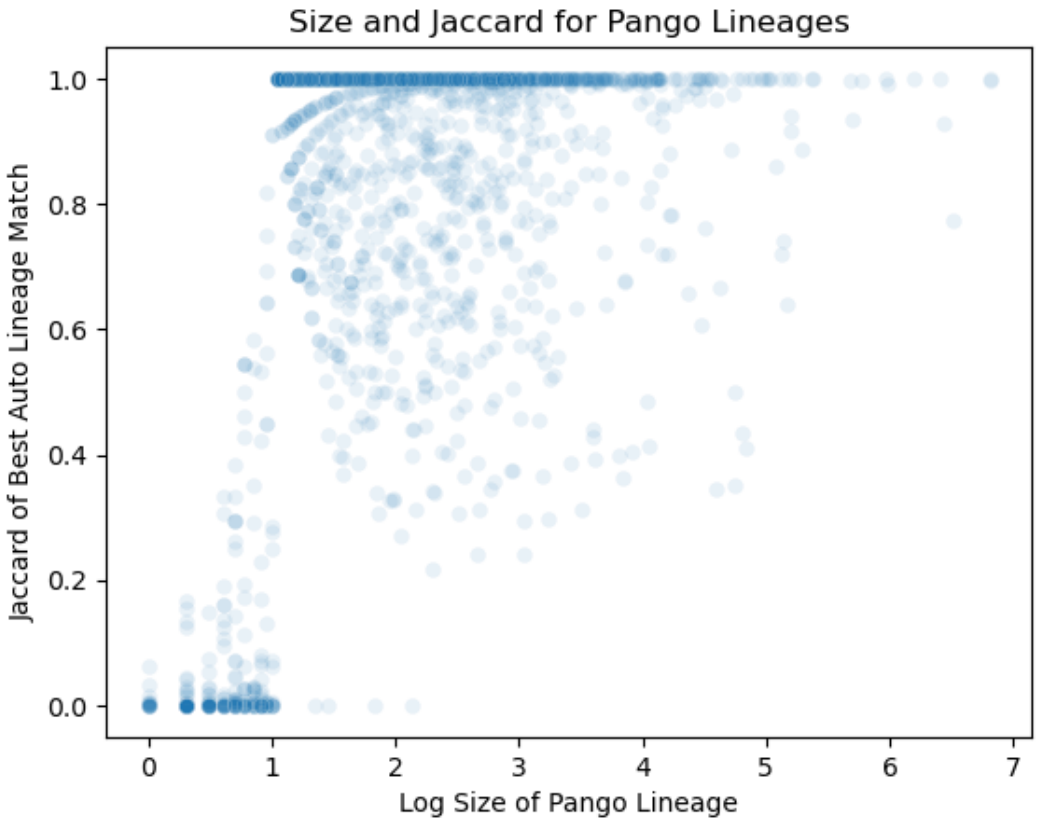

Figure A3.3: Pango Jaccard Indices by Size. We generally find that larger lineages are recaptured well by Autolin. Pango lineages below size 10 (1 on the log10 scale) are poorly recaptured because Autolin filters lineage proposals of less than 10 samples by default.

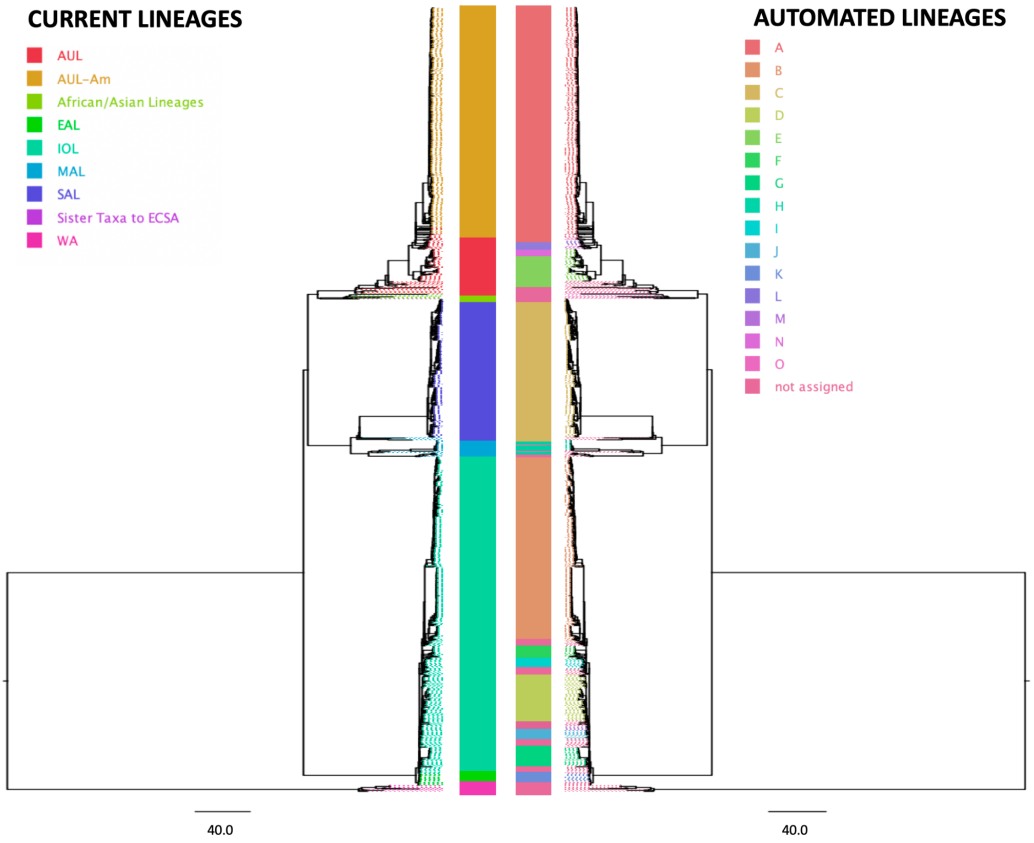

Figure A3.4: Comparison of CHIKV lineage annotations. Comparison of the geography lineage designation (left tree) with automated lineage designation (right tree) of Chikungunya virus, based on a tree previously generated by the Augur pipeline (Huddleston et al 2021) and visualized on FigTree v.1.4.4.

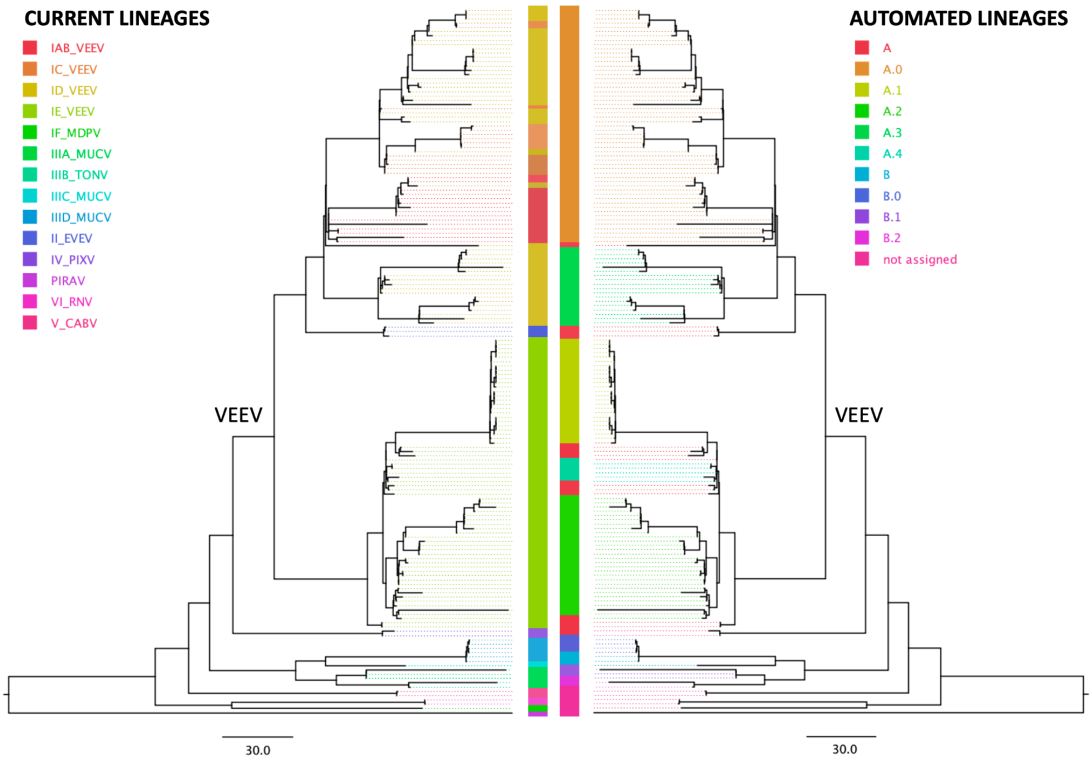

Figure A3.5: Comparison of VEE lineage annotations. Comparison of the serology subtype designation (left tree) with automated lineage designation (right tree) of the Venezuelan equine encephalitis virus complex (VEE), based on a tree previously generated by the Augur pipeline (Huddleston et al 2021) and visualized on FigTree v.1.4.4. According to the current nomenclature, VEE encompasses Everglades virus (EVEV), Mucambo virus (MUCV), Tonate virus (TONV), Pixuna virus (PIXV), Cabassou virus (CABV), Rio Negro virus (RNV), Mosso das Pedras virus (MDPV), Pirahy virus (PIRAV) and the Venezuelan equine encephalitis virus (VEEV) . The VEEV clade is labeled in the tree.

Lineage Name	Parent Lineage	Size	Exponential Growth Coefficient CI	Earliest Appearance	Latest Appearance	Regions	Nucleotide Changes	Amino Acid Changes
auto.BQ.1.1.8.1	BQ.1.1.8	23	[0.54 1.48]	2022-10-25	2022-12-02	England	G26526T,C29353T,A13581G	M:A2S
auto.CH.1.1.3	CH.1.1	32	[0.33 0.75]	2022-10-17	2022-12-05	England	C28093T,C21811T,G13441A	ORF8:S67F
auto.BQ.1.1.24.1	BQ.1.1.24	102	[0.33 0.53]	2022-11-01	2022-12-03	England	C5407T,G25459T	ORF3a:A23S
auto.BN.1.4.1	BN.1.4	183	[0.22 0.33]	2022-10-16	2022-12-03	England, USA, and Scotland	C14318T,G10364A	ORF1ab:V3367I,ORF1ab:T4685I
auto.BQ.1.1.13.1	BQ.1.1.13	65	[0.22 0.5]	2022-10-21	2022-12-04	Scotland and England	G13822A	ORF1ab:V4520I
auto.BQ.1.1.12	BQ.1.1	176	[0.210.35]	2022-10-16	2022-12-03	USA, Scotland, and England	C19547T,C25821T	ORF1ab:S6428L
auto.BQ.1.10.2	BQ.1.10	55	[0.2 0.47]	2022-10-17	2022-12-04	USA and England	A1320C,C3040T,C9286T	ORF1ab:E352A
auto.BQ.1.5.1	BQ.1.5	71	[0.19 0.36]	2022-10-16	2022-12-05	USA and England	G25855T,C823T,T8200C	ORF3a:D155Y
auto.CH.1.1.2.1	CH.1.1.2	16	[0.13 1.12]	2022-10-28	2022-12-02	England	A3569G	ORF1ab:S1102G
auto.BQ.1.8.3	BQ.1.8	109	[0.12 0.23]	2022-10-17	2022-12-03	England and Northern Ireland	A21137G	ORF1ab:K6958R
auto.BQ.1.2.1	BQ.1.2	145	[0.11 0.24]	2022-10-16	2022-12-02	England and USA	G19677T,C26147T,C3318T,T24163C	ORF3a:S252L,ORF1ab:P1018L,ORF1ab:Q6471H
auto.BA.5.1.13	BA.5.1	299	[0.11 0.14]	2022-10-16	2022-12-05	England and USA	G25352T,G15451A,A4595G,C28567T	ORF1ab:T1444A,ORF1ab:G5063S,S:V1264L

auto.BA.4.6.3.1	BA.4.6.3	36	[0.10 0.40]	2022-10-17	2022-12-04	England	T11377C,T4839C,G7675T	ORF1ab:I1525T
auto.BA.5.2.13.1	BA.5.2.13	138	[0.08 0.16]	2022-10-16	2022-12-03	England	G7743T	ORF1ab:S2493I
auto.BN.1.2.2	BN.1.2	31	[0.08 0.51]	2022-10-19	2022-12-01	England	C2574T,C1915T	ORF1ab:T770I
auto.BA.5.2.9.1	BA.5.2.9	255	[0.08 0.15]	2022-10-16	2022-12-04	USA and England	A27250G,C14697T,A12061C,C10369T,C18570T,C5806T,A22600T	ORF1ab:E3932D,S:R346S,ORF6:I17V
auto.BQ.1.18.1	BQ.1.18	22	[0.03 0.53]	2022-10-16	2022-12-06	England	C15017T	ORF1ab:A4918V,E:S71P
auto.BQ.1.10.1.1	BQ.1.10.1	22	[-0.0026 0.73]	2022-11-06	2022-12-05	England	G23401T,A3481T	S:Q613H
auto.BE.1.2.2	BE.1.2	79	[-0.047 0.061]	2022-10-16	2022-12-02	England	C5183T,C27393T,C9979T	ORF1ab:P1640S
auto.BQ.1.1.22.1	BQ.1.1.22	25	[-0.052 0.72]	2022-11-08	2022-12-01	England	A2977G,C27213T,C19185T,C16726T	ORF1ab:H5488Y
auto.BN.1.3.1.1	BN.1.3.1	28	[-0.107 0.213]	2022-10-16	2022-12-01	England	G20578T,C3787T,T7456C	ORF1ab:V6772L
auto.XBB.3.2	XBB.3	30	[-0.28 0.27]	2022-10-18	2022-12-02	USA	C29614T,T22092C,C28054T,C8146T	ORF8:S54L,S:M177T
auto.BA.5.3.5.1	BA.5.3.5	27	[-0.31 0.60]	2022-11-08	2022-12-03	England	C22993T,C7858T,T1592C	ORF1ab:S443P
auto.CR.1.2	CR.1	10	[-0.56 0.97]	2022-10-31	2022-12-01	England	G15957T,T22308C	S:L249S

Table A3.1: Output Report for 24 New Lineage Designations. This table includes basic statistics for 24 new lineage designations actively sampled in December 2022.

Strain	Moved	Lineage	On Date
IND/2947/2023|OQ701301.1|2023-03-10	To	auto.XBB.1.16.1.1	2023-04-02
Scotland/SCOT-24049/2022|OX435521.1|2022-12-31	From	auto.B.1.1.529.6	2023-04-02
Scotland/SCOT-26462/2023|2023-02-21	To	auto.XBB.1.5.8.1	2023-04-02
Scotland/SCOT-26462/2023|OX451476.1|2023-02-21	To	auto.XBB.1.5.8.1	2023-04-15
Scotland/SCOT-26819/2023|2023-02-21	To	auto.XBB.1.5.8.1	2023-04-02
Scotland/SCOT-26819/2023|OX453590.1|2023-02-21	To	auto.XBB.1.5.8.1	2023-04-15
Scotland/SCOT-26821/2023|2023-02-21	To	auto.XBB.1.5.8.1	2023-04-02
Scotland/SCOT-26821/2023|OX453648.1|2023-02-21	To	auto.XBB.1.5.8.1	2023-04-15
Scotland/SCOT-27434/2023|OX457315.1|2023-03-02	To	auto.B.1.1.529.6	2023-04-02
Scotland/SCOT-28483/2023|2023-03-12	To	auto.XBB.1.5.8.1	2023-04-06
Scotland/SCOT-28483/2023|OX462874.1|2023-03-12	To	auto.XBB.1.5.8.1	2023-04-21
USA/CA-CDC-QDX48192978/2023|OQ728437.1|2023-03-20	To	auto.XBB.1.5.40	2023-04-05
USA/CA-CDC-QDX48192979/2023|OQ728624.1|2023-03-20	To	auto.XBB.1.5.40	2023-04-05
USA/FL-CDC-QDX47418754/2023|OQ644056.1|2023-03-01	From	auto.XBB.1.5.40	2023-04-02
USA/FL-CDC-QDX48193288/2023|OQ782385.1|2023-03-21	From	auto.XBB.1.5.40	2023-04-17
USA/FL-CDC-QDX48587517/2023|OQ827304.1|2023-03-31	To	auto.XBB.1.5.40	2023-04-22
USA/LA-BIE-LSUH003690/2023|OQ736767.1|2023-03-20	To	auto.XBB.1.5.40	2023-04-06
USA/LA-BIE-LSUH003692/2023|OQ736769.1|2023-03-22	To	auto.XBB.1.5.40	2023-04-06
USA/M00621433-F1-1/2023|OQ820697.1|2023-03-29	From	auto.XBB.1.5.40	2023-04-20
USA/NJ-CDC-QDX48685195/2023|OQ827907.1|2023-04-03	To	auto.XBB.1.5.40	2023-04-22
USA/NV-CDC-LC1035085/2023|OQ808432.1|2023-03-30	To	auto.XBB.1.5.40	2023-04-19
USA/NY-CDC-QDX47311927/2023|OQ610887.1|2023-02-27	From	auto.XBB.1.5.40	2023-04-02
USA/PA-CDC-QDX46827201/2023|OQ577387.1|2023-02-14	From	auto.XBB.1.5.40	2023-04-02
USA/PA-CDC-QDX47973735/2023|OQ713000.1|2023-03-15	From	auto.XBB.1.5.40	2023-04-02
USA/TX-CDC-QDX47418188/2023|OQ643781.1|2023-02-28	To	auto.XBB.1.5.40	2023-04-02
USA/VA-CDC-LC1031818/2023|OQ734369.1|2023-03-25	From	auto.XBB.1.5.40	2023-04-06
USA/WA-CDC-UW23032239573/2023|OQ833295.1|2023-03-22	To	auto.XBB.1.5.40	2023-04-22
USA/WI-CDC-VSX-A065209/2023|OQ748644.1|2023-03-22	To	auto.XBB.1.16.1.1	2023-04-08

Table A3.2: 28 Samples that Change Lineages between 2023-04-01 and 2023-04-30. It's likely that some of these represent redundant or mislabeled samples, such as Scotland/SCOT-26462/2023|2023-02-21 and Scotland/SCOT-26462/2023|OX451476.1|2023-02-21. auto.XBB.1.5.40 is the most affected lineage, but these samples are <1% of its constituency.

Appendix 4

Fine-Scale Position Effects Shape the Distribution of Inversion Breakpoints in Drosophila melanogaster

[This appendix has been adapted from publication, “Fine-Scale Position Effects Shape the Distribution of Inversion Breakpoints in Drosophila melanogaster” (McBroome et al 2020, Genome Biology and Evolution)]

A4.1: Context

The following section is work I completed over the course of 2019 and published in early 2020, before I pivoted my research program towards addressing the pandemic bioinformatics bottleneck. Accordingly, it has little to no relation with the main body of this dissertation. As it represents published scientific work I completed during my time at UCSC, I chose to include it in this appendix. It represents an analysis examining the effects of inversion events on local regulatory landscapes and how that might inform genome structural evolution in *Drosophila melanogaster*.

A4.2: Chromosomal Inversions

Chromosomal inversions, which are large genomic regions that are generated by double-strand breakage and repair in reverse orientation, are widespread in many natural populations. These rearrangements have a long history of study in *Drosophila* species (Dobzhansky 1962; Sturtevant 1917). The primary theories explaining the prevalence of inversions in natural populations are that suppressed recombination over the inverted region is favored by natural selection (Corbett‐Detig & Hartl 2012; Fuller et al 2019; Kapun et al 2016a; Kirkpatrick and Barton 2006; Langley et al 2012; Mukai 1971; Sturtevant and Beadle 1936). Alleles contained in inversions can interact epistatically or additively to maintain a complex polygenic phenotype such as body size, stress resistance, fecundity, and lifespan (Hoffmann et al 2004; Hoffmann and Rieseberg 2008; Kirkpatrick 2010). Inversions that suppress recombination between alleles that contribute to a beneficial phenotype can be selected for. Biogeographic data supports this hypothesis; natural populations of *Drosophila melanogaster* maintain inversion frequency clines strongly correlated with climatic clines (Kapun et al 2016a; Kapun et al 2016b; Knibb 1982; Mettler et al 1977; Rane et al 2015; Simões and Pascual 2018). Furthermore, an ever expanding set of taxa appear to contain polymorphic inversions that are associated with adaptive phenotypes (Butlin et al 1982; Huynh et al 2011; Oneal et al 2014). It is increasingly accepted that a major source of positive selection on chromosomal inversions is the maintenance of linkage among alleles that are favorable in similar contexts.

Whereas the potential fitness benefits of maintaining linkage among synergistic alleles are well established, the impacts of inversion breakpoints on the individuals that carry them are not well understood. Nonetheless, these impacts are likely to play an important role in shaping evolutionary outcomes for new arrangements. An inversion breakpoint that disrupts a key gene sequence could result in the death or sterility of the individual that carries it, preventing the inversion from reaching polymorphic frequencies in natural populations. Accumulated evidence is consistent with the idea that an inversion's breakpoint positions might have large impacts on its fitness. The distribution of polymorphic inversion breakpoints along the genome is not random (Calvete et al 2012; Gonzalez et al 2007; Orengo et al 2015; Pevzner and Tesler 2003; Puerma et al 2014; Puerma et al 2016b; Tonzetich et al 1988). In fact, many apparently independently formed inversions seem to precisely share breakpoint locations (Gonzalez et al 2007; Puerma et al 2014; Pevzner and Tesler 2003; Corbett-Detig et al 2019). Even when inversion breakpoints are not precisely reused at the molecular level, their broad-scale distributions across the genome are non-uniform (Pevzner and Tesler 2003; Ranz et al 2007). Though this pattern is well-established, the factors underlying breakpoint localization and the fitness of new arrangements are poorly understood.

There are two mechanisms that shape the fine-scale distribution of inversion breakpoints. First, mutational biases are factors that affect the probability that an inversion breakpoint occurs at a specific genomic location (Calvete et al 2012; Guillen and Ruiz 2012; Pevzner and Tesler 2003; Tonzetich et al 1988). In many

species inversions occur through ectopic recombination between repetitive sequences, an example of a mutational bias (Guillen and Ruiz 2012), though this is relatively rare in the *melanogaster* subgroup (Ranz et al. 2007; Corbett-Detig and Hartl 2012). Additionally, some evidence indicates that physical instability due to unstable secondary structure or local chromatin environment may also bias breakpoint localization (Falk et al 2010). Second, specific breakpoint positions can affect the fitness of a new arrangement. These "position effects" have been identified in a variety of organisms (Castermans et al 2007; Frischer et al 1986; Hough et al 1998; Lakich et al 1993; Puig et al 2004). Deleterious effects associated with breakpoint positions could hypothetically be as large as positive impacts from the maintenance of allele complexes contributing to polygenic traits. Deleterious position effects are therefore expected to limit the number of individual inversions that could evolve and maintain polygenic phenotypes in the population.

There are several specific factors that could influence the fitness of inversion breakpoints. First, disruption of gene sequence and enhancer-promoter interactions can cause mRNA truncation, chimeric transcripts, or misregulation of genes overlapping and near to breakpoints (Castermans et al 2007; Frischer et al 1986; Lupianez et al 2016; Ren and Dixon 2015). Previous work has found that common inversions in *D. melanogaster* and fixed inversions in *D. pseudoobscura* are less likely to disrupt gene coding sequences that would be expected under a random breakpoint model (Corbett-Detig and Hartl 2012; Fuller et al. 2017), possibly indicating that natural selection acts against inversions which disrupt gene sequences.

However, a mutational bias that preferentially creates breakpoints in intergenic regions is also consistent with these findings.. Inversions in the *D. melanogaster* species group tend to create inverted duplications of sequence at their breakpoints in the repair process, which can preserve copies of disrupted sequence (Puerma et al 2016b; Ranz et al 2007). Duplication size may therefore also influence the fitness of an inversion breakpoint because large duplications can avoid disrupting individual genes. A study in *Anopheles gambiae* has shown that the inversion $2L+^{a}$ is likely viable because it preserves functional copies of disrupted genes through this mechanism (Sharakhov et al 2006). Location in respect to gene sequence and duplication size should both contribute to the fitness of a new inversion arrangement.

Factors related to gene regulation may also impact the fitness of newly formed arrangements. These include Topologically Associated Domains (TADs), chromatin state, and the locations of insulator elements. TADs are genomic features that appear in HiC proximity ligation mappings (Lieberman et al 2009) which reflect the physical folding and arrangement of the genome (Lupianez et al 2016; Jost et al 2014; Sexton et al 2012). Disruption of these domains may alter local gene expression (Lupianez et al 2015). Chromatin marks often determine local expression and repressive chromatin is capable of suppressing nearby gene activity when translocated (Cryderman et al 1998). Boundaries between domains are often associated with insulator elements in *D. melanogaster* (Sexton et al 2012). Insulators limit the influence of repressive chromatin marks and block ectopic enhancer activity, and could therefore act as a compensatory mechanism to maintain native regulatory environments (Bushey et al

2008; Gaszner and Felsenfeld 2006; Sigrist and Pirrotta 1997; Yang and Corces 2012). We hypothesize that high-fitness inversions disrupt local gene regulation less than would be expected by chance, by avoiding disrupting crucial domains or by colocalization with insulator elements.

Comparisons among fixed, high frequency and low frequency inversions can reveal the impact of natural selection on chromosomal inversion breakpoints (Caceres et al. 1997; Corbett-Detig 2016). Because they have persisted and spread within natural populations, we expect both high population frequency and ancestrally fixed chromosomal inversion breakpoints to show a biased distribution of features consistent with higher fitness. Conversely, low-frequency inversions, often identified in only a single individual within a population, are most likely recently arisen arrangements. The low-frequency inversions' breakpoint distribution should therefore primarily reflect mutational biases. By examining the distributions of fixed, high frequency, and low frequency inversion breakpoints, we can identify the factors that shape the fitness of newly-arisen arrangements.

We leverage population resequencing datasets from more than 1,000 *D. melanogaster* isolates to detect and *de novo* assemble both breakpoints of 18 rare naturally occurring inversions. We compare these "rare" inversion breakpoints to known high frequency inversion breakpoints in *D. melanogaster* (Corbett-Detig and Hartl 2012) as well as a set of fixed inversion breakpoints between species in the *Melanogaster* subgroup (Ranz et al. 2007). By comparing rare, common, and fixed inversion breakpoints, we find evidence supporting the idea that both mutational

biases and natural selection play important roles in shaping the fine-scale distribution of inversion breakpoints in natural populations.

A4.3: Methods

Defining Inversion Categories

In our analysis, we define three classes of inversion population frequency. Previous work in *D. melanogaster* has typically referred to four categories of inversion, “common cosmopolitan”, “rare cosmopolitan”, “recurrent endemic”, and “unique endemic” (Mettler et al 1977; Krimbas and Powell 1992). The latter half of each of these terms refers to the geographic distribution of the inversion. As long as an inversion reached high frequency in any population, it has not been strongly impacted by negative selection. We label these high-frequency inversions “common” inversions. We use “rare” to refer to inversions which were found in only single samples (with the exception of *In(2R)Mal*, which is present in three samples studied here). The distribution of rare inversions, while possibly containing high-fitness inversions that could eventually spread to high frequencies, are likely to primarily reflect mutational biases in their overall breakpoint distribution. To summarize, “common cosmopolitan”, “rare cosmopolitan”, and “recurrent endemic” will all fall under our label “common” while we refer to “unique endemic” as “rare” inversions, similarly to the analysis in (Corbett-Detig 2016).

The third class in our framework, “fixed” inversions, are inversions that have gone to fixation within one lineage during divergence of the *Drosophila melanogaster*

subgroup (Ranz et al 2007). Originally all fixed inversions occurred as unique events in a *Drosophila* ancestor. They subsequently spread until they reached fixation in populations ancestral to contemporary species in the *melanogaster* subgroup. These fixed inversions were discovered by comparing the locations of homologous sequences in the genomes of between *D. melanogaster* and its relatives (Lemeunier and Ashburner 1976) and have been molecularly characterized previously (Ranz et al 2007). It is important to note that the vast majority of these fixed inversions occurred on the *Drosophila yakuba* branch and not in a direct *D. melanogaster* ancestor (Krimbas and Powell 1992; Ranz et al 2007). The reference genome of *D. melanogaster* should therefore generally reflect the ancestral state and the genetic background on which these inversions originated rather than a derived state evolved after fixation. Common and rare inversions annotated here occurred in contemporary *D. melanogaster* populations and thus in the absence of additional changes unrelated to genome structure, on a similar genetic background to that on which the *D. yakuba* inversions were fixed. The functional annotations used here are also based on the *D. melanogaster* standard arrangement, meaning these annotations should represent the genetic background of all three inversion frequency categories.

Short Read Alignment

We obtained short read data as fastq files from the Sequence Read Archive. All short read data is described in Lack et al 2016 and was originally produced in (Pool et al 2012; Lack et al 2015; Mackay et al 2012; Kao et al 2015; Grenier et al.

2015). We aligned the short read data using bwa v0.7.15 using the "mem" function and default parameters (Li et al 2013). All post-processing (sorting, conversion to BAM format, and filtering) was performed in SAMtools v1.3.1 (Li et al 2009). We filtered these BAM files to include only those alignments with a minimum mapping quality of 20 or more.

Rare Breakpoint Identification

As in previous works that characterized structural variation using short-insert paired-end Illumina libraries (*e.g.* (Corbett-Detig et al 2012; Cridland et al 2010; Rogers et al 2014), we first identified aberrantly-mapped read "clusters". Briefly, here, a cluster is defined as 3 or more read pairs that align in the same orientation (for inversions, this is either both forward-mapping or both reverse-mapping) and for which all reads at one edge of the cluster map to within 1 Kb of all other reads in the cluster. We considered only aberrant clusters where both ends mapped to the same chromosome arm as the vast majority of inversions in *Drosophila* are paracentric (Krimbas and Powell 1992). We required that all read pairs included in a cluster map a minimum of 500 Kb apart. We then retained only those potential inversions for which we recovered both forward and reverse mapping clusters there were within 100 Kb of one another. The choice of a maximum distance between possible breakpoint coordinates was included to reduce the possible rates of false positives and because none of the known inversions whose breakpoints have previously been characterized included a duplicated region of 100 Kb or more (Corbett-Detig and Hartl 2012; Ranz

et al. 2007). When breakpoint assemblies existed in very close proximity or appeared to delete short sequences, we set the duplication size to 1 base. We further filtered all breakpoint assemblies that overlapped annotation transposable elements as these are the primary source of aberrantly-mapping read clusters in previous works (*e.g.*, Corbett-Detig and Hartl 2012).

As an additional check for the accuracy of our newly discovered breakpoints, we compared our distribution of rare breakpoints to the known cytogenetic distribution and found no chromosomal or by-region differences ($p=0.7$, chi-square test; cytogenetic data from Corbett-Detig 2016 who summarized Krimbas and Powell 1992). The short insert size from previous sequencing experiments ranged from ~200bp to ~600bp, which may have led to a non-trivial false negative rate of breakpoint discovery particularly if the breakpoints contain repetitive elements or other large DNA insertions. However, we do not expect that these potential false negatives will bias our downstream analyses, and all previously characterized inversion breakpoints in the *Melanogaster* species complex occurred in unique sequences (Ranz et al. 2007; Corbett-Detig and Hartl 2012). All software used to perform these analyses is available from the github repositories associated with this project. Specifically, scripts used for breakpoint detection and assembly are in https://github.com/dliang5/breakpoint-assembly (DOI: 10.5281/zenodo.7983441).

De novo Rare Breakpoint Assembly

For each putative inversion, we then extracted all reads for which either pair mapped to within 5 Kb of the predicted breakpoint position. We converted all fastq read files to fasta and qual files as is required by Phrap, and we assembled each using otherwise default parameters but including the "-vector_bound 0 -forcelevel 10" command line options (Corbett Detig and Hartl 2012; Rogers et al 2014). We then used blast to align the resulting de novo assembled contigs to the *D. melanogaster* reference genome to identify the contig that overlapped the predicted breakpoint using the flybase blast tool (www.flybase.org/blast). We retained only inversions for which we could *de novo* assemble contigs overlapping both breakpoints, and we further discarded any contigs where the sequence intervening two distant genomic regions contained sequence with homology to known transposable elements. All of the assembled breakpoint sequences are available in File S1. Assembly scripts are available from https://github.com/dliang5/breakpoint-assembly (DOI: 10.5281/zenodo.7983441).

Overlapping Inversions and In(2R)Mal

We also attempted to find sets of overlapping inversions. Briefly, for overlapping inversions, where one inversion arises on a background that contains another inversion with one breakpoint inside and one outside of the inverted region, the breakpoint-spanning read clusters should be largely the same as inversions that arose on a standard arrangement chromosome. However, the key difference is that

rather than pairs of forward and reverse mapping read clusters, we expect to observe two distantly mapping read clusters in the reverse-forward and forward-reverse arrangements. We applied this approach for the 17 rare inversions that we initially discovered as well as to all samples that contained common inversions that are known from previous work (Corbett-Detig and Hartl 2012; Lack et al. 2015). We found only one such overlapping rare inversion, which is consistent with the known segregation distorter associated chromosomal inversion *In(2R)Mal*, which is composed of two overlapping inversions (Presgraves et al 2009). In our analysis here, we treat these overlapping inversions as independent, but our results are qualitatively unaffected if we simply exclude the second inversion.

Genome Version, Insulator, and Gene Annotations

All our analyses are based on alignments to *D. melanogaster* genome version 6.26 (Hoskins et al 2015)**.** We obtained genome annotation data including gene locations from flybase. We treated long non-coding RNAs as genes for our purposes, as they perform essential functions and can be disrupted in the same way as protein-coding genes. We obtained insulator binding site positions from (Negre et al 2010, accession GSE16245). As necessary, we converted the coordinates of genomic features from genome version 5 to 6 using the flybase coordinate batch conversion tool (https://flybase.org/convert/coordinates).

Selection of Public Datasets for Topological Domains and Chromatin Marks

We obtained topologically associated domain (TAD) data including annotations of chromatin state from Sexton et al. (2012). This dataset is composed of domains detected by genome-wide chromosome conformation capture sequencing, HiC, on early-stage embryos, and annotated with an epigenetic state using a clustering method applied to another source of linear epigenomic data (Sexton et al 2012). Their annotations include four categories: “active”, “null”, “PcG” (polycomb), and “HP1” (centromeric heterochromatin). For the sake of consistency we refer to Sexton et al.’s “null” domains as “inactive”. Early stage embryos are likely to be the environment in which any regulatory disruption induced by inversions is most deleterious given the sensitive nature of development, which makes this a promising source of context for our analysis of inversion frequency. This dataset also allows us to separately analyze breakpoint occurrence within topologically associated domains and chromatin states in tandem, since they are derived from the same source. It should be noted, however, that the annotations of these TADs are relatively coarse and may not reflect the more local environment of an inversion breakpoint.

We therefore performed a second analysis on finer scales using the dataset of Kharchenko et al (2011, accession GSE25321). This dataset in its raw form consists of short spans marked with one of a set of chromatin markers, in both a nine-state model and a thirty-state model. As we desired a representation of the local chromatin environment around inversion breakpoints, we chose to bin the nine-state

representation into total counts of bases assigned to a state of the given type over windows of 10kb. 10kb was selected based on the average heterogeneity of the windows; we wanted our window size to be as small as possible but for most windows to contain at least one region with an annotated chromatin state. This yielded a distribution of values for each window which represented the overall enrichment of each state in each 10kb span. As we lacked statistical power to evaluate these mark types individually with our relatively small inversion breakpoint datasets, we further assigned each 10kb window an activity state based on the majority of present marks. Windows in which the vast majority of sites were assigned states one through five, annotated by Kharchenko et al. as being various components of genes including promoters, exons, and introns, were designated “Active”. Windows where states six through nine, which include PcG, HP1, and other heterochromatic marks, were most prominent, were designated “Inactive”. Windows in which both groups each constituted at least five percent of all marks were designated “Mixed”. This yields an alternative representation of chromatin environments surrounding inversion breakpoints that is much finer-grained than the annotations of Sexton et al.

We compared this representation to Sexton et al.’s annotated chromatin states as an additional check for the validity of our approach. We found that 10kb windows located within each annotated TAD generally aligned with the annotation of that TAD, but that substantial heterogeneity of chromatin marks exists within each TAD span (Figure A4.3). For example, approximately 19% of windows within TADs annotated as “active” are enriched for chromatin state 9, which is associated with

extended silenced regions, and conversely 26% of windows within TADs annotated as inactive are enriched for chromatin state 2, which is associated with the active transcription. This indicates that one cannot be treated as a direct substitute for the other.

As a final check on the validity of the domains obtained from Sexton et al, we obtained polytene domain data from Eagen et al (2015), repeated our analysis, and found them to be generally consistent with our conclusions. These results may be found in section A4.12.

Permutations and Statistical Tests

To compare inversion breakpoint positions to a randomized distribution, permutations for all categories of inversions (rare, common, and fixed) were performed with 1000 iterations of a group of randomly located breakpoints, holding the inversion number, duplication lengths, and chromosome arms constant. Specifically, for each inversion breakpoint, one thousand starting positions were chosen from a uniform distribution between the start of that chromosome arm and the end minus the length of the duplication- that is, from the entire set of possible points for that size of breakpoint. Random breakpoints were located independently for most tests, as most values were calculated for each breakpoint individually rather than the inversion as a whole. The exception is the chromatin-blending test, in which we additionally controlled for inversion lengths to account for the role of inversion length in biasing pairs of chromatin environments. Features of the genome at each of these

breakpoints were recorded as our expected value for the random distribution of breakpoints.

Tests were divided by the nature of the factor. For factors that are a discrete numerical value for each break, such as distance to an element or length of a duplication, p-values were calculated as percentiles of real values within a large set of random distributions. Tests between categories of the distance-based factors and the duplication length test were performed distribution to distribution with pairwise Mann-Whitney rank-sum tests.

For categorical values, such as disrupting a gene span or not, rates of category occurrence were calculated for one thousand permutations. We define disruptions of genes and other elements as both forward and reverse single-strand breaks occurring within a single annotated functional element (Figure A4.1). It is important to note that our method of defining disruption is likely to overestimate the proportion of fixed inversion breakpoints that truly disrupt genic sequences. Ranz et al (2007)'s method to identify sequences duplicated by the original break relies on sequence homology, and in fixed inversions divergence of noncoding sequences can interfere with the precise identification of breakpoint regions. For example, if the original duplicated region includes a gene coding span and some non-coding bases, a complete gene copy will be produced along with a partial duplication (Figure A4.1). Over time, the non-coding region will tend to accumulate more mutations than the intact gene copy. In this case, coordinates obtained from BLAST alignments may not detect the homology between the non-coding regions and instead only yield apparent homology

from duplication within the conserved gene span. This would be counted as a gene disruption event by our analysis. This bias will tend to make our analysis conservative with respect to identifying the impacts of natural selection, because breakpoints are more likely to be identified within coding regions and because we should tend to underestimate the sizes of breakpoint adjacent duplicated regions after sequence homology has decreased. All scripts used to produce the results of the permutation tests described above are available from the github repository associated with this project https://github.com/jmcbroome/breakpoint_analysis (DOI: 10.5281/zenodo.7983441).

Lethal and Sterile Phenotype Analysis

Additionally, we obtained phenotype data from Flybase using the query builder (https://flybase.org/cgi-bin/qb.pl) to get the IDs of all genes which have lethal phenotypes and sterile phenotypes. This data was incorporated into the gene disruption analysis and we sought evidence of difference in disruption rates between genes annotated with these phenotypes and the overall set of annotated genes. Table A4.2 contains the set of inversion breakpoints which appear to disrupt these genes.

A4.4: Common and Fixed Inversion Breakpoints

Common and fixed inversion breakpoints have been characterized extensively in *D. melanogaster* and in the *Melanogaster* species complex in previous works. We obtained the breakpoint locations for nine common inversions from (Corbett-Detig

and Hartl 2012; Lack et al 2016). We note that although population frequencies and geographic ranges vary among common inversions (Corbett-Detig and Hartl 2012; Krimbas and Powell 1992; Lack et al 2016), each has reached frequencies of at least 10% within local subpopulations and all have been observed in several geographically-widespread populations, suggesting that their breakpoints do not cause strong deleterious fitness consequences. From Ranz et al (2007) we obtained the breakpoint positions of 26 inversions that have fixed in a lineage since the common ancestor of the *Melanogaster* species complex. To confirm that the breakpoint-adjacent regions have not been modified or updated in the more recent genome assemblies for either *D. melanogaster* or *D. yakuba*, we extracted each surrounding 100Kb region from the genome that contains the ancestral arrangement and used BLAST to align these to the genome containing the derived rearrangement. We recorded the most breakpoint proximal high quality, *i.e.* BLAST score greater than 50, sequence alignment as the putative location of the inversion breakpoint.

A4.5: Rare Inversion Breakpoints Discovered

We realigned all sequence data from over 1,000 *D. melanogaster* natural isolates that have been sequenced previously using paired-end sequencing methods (Grenier 2015; Langley 2012; Pool et al 2012; Mackay et al 2012; Lack et al 2015; Kao et al 2015; summarized in detail in Lack et al 2016). We identified 5,318 short read clusters that corresponded to possible inversion breakpoints that are a minimum of one Megabase from each other and for which we found both forward and reverse

mapping read clusters (Figure A4.4). That is, for a given inversion relative to the reference genome, we expect to find a cluster of read pairs where both maps in the "forward" orientation and another cluster where each pair of reads both map in the "reverse" orientation (Corbett-Detig and Hartl 2012; see Methods). We also searched for overlapping inversions using a slight modification of this approach (Methods). To be as conservative as possible with our analysis, we retained only the set for which we recovered and successfully *de novo* assembled both breakpoints for a given inversion. Additionally, we removed any putative breakpoint-spanning contig that mapped with high confidence to multiple locations in the *D. melanogaster* reference genome. We ultimately retained 18 rare inversions. Three of our candidate rare inversions are corroborated by previous cytological evidence (Huang et al 2014; Presgraves et al 2009). Similarly, previous molecular evidence (Grenier et al 2015) supports the identified breakpoints of another chromosomal inversion. The breakpoints of our putative rare inversions do not show unusual genetic distances from other samples isolated from the same populations, suggesting that these are relatively recent events and not older inversions that have recently gone to lower frequencies (Section A4.13, Table S1).

The genomic and population distributions of candidate rare inversions are largely consistent with our expectations based on extensive cytological work. First, our estimated rate of occurrence of rare inversions, 1.6% per genome, is within the range of estimates from cytological data across diverse populations 0.47%-2.71% (Aulard et al 2002; Krimbas and Powell 1992). Furthermore, we found no rare

inversions on the X chromosome, which contains very few chromosomal inversions in natural populations of this species (Aulard et al 2002; Krimbas and Powell 1992). However, because we conservatively required that both breakpoints are detected from discordant short-read alignments and completely assembled *de novo*, and because we excluded any breakpoints that contained homology to annotated transposable elements, it is possible that our approach has underestimated the prevalence of rare inversions in these datasets. It is also possible that a portion of the rare inversions may be false positives owing to the challenges of short-read based de novo assembly and interpretation. Nonetheless, as an additional check to ensure the robustness of our results, we repeated all of our analyses on the subset of rare inversions which have been cytologically or molecularly characterized or are very simple in their breakpoint structures and found no major differences between datasets (Section A4.14).

A4.6: Inversion Breakpoints Could Truncate Coding Sequences

Inversions can strongly disrupt sequences at their breakpoints (Figure A4.1). This has multiple classes of potential negative consequences, including the truncation of gene spans and the creation or alteration of enhancer-gene interactions (Castermans et al 2007; Frischer et al 1986; Lupianez et al 2016; Ren and Dixon 2015). We investigated interactions with gene spans with the hypothesis that higher-frequency inversions are more likely to exhibit features which reduce large-scale disruptions of local functional elements. For each category, we calculated the percentile of the count of disrupting breakpoints against the permuted distribution, where low percentiles

correspond to less disruption than expected. All three inversion frequency categories disrupt annotated gene spans less often than the random expectation (rare $p=0.0415$, common $p=0.0055$, fixed $p<0.001$, permutation test). The proportion of gene-disrupting inversions is inversely correlated to population frequency category (44% of rare inversion breakpoints, 28% of common inversion breakpoints, 24% of fixed inversion breakpoints).

Our results are consistent with gene disruption being negatively selected after inversion formation. We note here that the baseline rate of disruption is still relatively high even in the most conservative category, at 24% of fixed inversion breakpoints. Nonetheless, for reasons described above (methods), this should be considered a conservative upper-bound on the rate of gene disruption in the fixed inversion class. In all cases of putative disruption, the *D. yakuba* genome contains an intact ortholog; this indicates that if breakpoints occurred within an annotated gene, they rarely completely disrupt the coding sequence or that secondary sequence evolution can suppress the deleterious effects. All putatively disrupting breakpoints within the fixed inversion class lie within 1000 bases of the start or the end of the disrupted gene (Table S3). The trend across categories indicates that there is a negative association between population frequency and the occurrence of inversion breakpoints within gene sequences in our data.

We also note that rare inversions appear to disrupt genes less often than expected by chance. This could be explained by the critical nature of many genes to survival. At a minimum, each inversion must not be lethal for us to discover it. The

preservation of gene spans by rare inversions may also be explained by a mutational bias of chromatin state or basepair composition favoring intergenic regions, reducing gene disruption rates below random expectations. As a final possible explanation we note that because many of the samples used in this work were inbred, either intentionally or passively as isofemale lines, inversions that induce recessive strongly deleterious fitness effects might still be exposed to selection and purged from the line prior to sequencing.

We further investigated the possible fitness impacts of disrupted gene sequences by examining the subset of disrupted genes that are annotated as having lethal or sterile alleles. We expect to observe a reduction in the rates that inversion breakpoints interrupt genes with lethal alleles and sterile alleles owing to the importance for organism survival and reproduction. In applying a similar permutation test as above, but instead asking if inversion breakpoints are less likely than expected by chance to disrupt essential genes specifically, we do not find a significant decrease in the rate of essential gene disruptions compared to genes overall (Table S4). We note that only one gene with an annotated sterile phenotype was disrupted among all inversion breakpoints considered here. However, we still failed to reject the null model possibly due to a general paucity of known sterility-inducing genes compared to unannotated genes.

Furthermore, it is possible that a significant portion of genes remain functional despite the presence of both breaks within the annotated span. For example, the common inversion *In(X)A* disrupts a gene with annotated lethal alleles. The disrupted

gene, NFAT, encodes an important transcription factor (Keyser et al 2007). The inversion breakpoint is very near the 5' start of the gene, where some annotated transposable element insertions have produced viable alleles (Bellen et al 2011). It is possible that the breakpoint does not actually render the gene nonfunctional and is therefore not lethal. Further functional work will be needed to understand the specific effects of localized gene disruption on individual phenotypes.

A4.7: Larger Inverted Duplications May Prevent Gene Disruption

Duplications that occur during inversion formation may maintain functional elements and suppress the local gene-interrupting effects of inversion breakpoints. Paired staggered double-strand breakage is the major mechanism by which inversion events occur in the *D. melanogaster* subgroup (Puerma et al 2016b; Ranz et al 2007). These breaks leave an overhang of sequence at each end of the putative inversion. After repair in inverted orientation, the result is inverted duplicated regions on either side of a new inversion with length equal to the overhang left over after the double strand break (Figure A4.1). To guarantee disruption of a given functional element at the sequence level without creating a complete duplicate, both sides of a double-strand break must fall into that same functional element. Longer duplications are thus less likely to disrupt individual elements. Therefore we hypothesized that selection will favor longer duplicated regions that minimize impacts on local sequence functions.

To test this, we first verified that tandem duplications of inversion breakpoints which do not disrupt genes are longer than those that do ($p=0.0096$, Mann-Whitney U test). Dividing the data by frequency category, we found that common polymorphic inversions have significantly longer duplications than rare inversions (Figure A4.1, $p=0.0095$, Mann-Whitney U test). We did not include fixed inversion breakpoints, as secondary sequence evolution and gaps between synteny blocks made determination of exact original duplication length inaccurate and likely an underestimate. These results are consistent with the idea that long duplications act as a compensatory mechanism for otherwise negative position effects by preserving intact functional elements or by maintaining proximity among functional elements within duplicated regions.

Formally, our analysis is consistent with higher relative fitness of inversions with longer inverted repeats, but does not necessarily require deleterious effects at single breakpoints. It is also possible that inversions are positively selected when they contain larger breakpoint-adjacent duplications because of positive effects associated with gene duplications or chimeric gene products (Puerma et al 2016). However, given that microsynteny is largely maintained over evolution and given that the *Drosophila* genome contains a high density of functional elements, we favor our hypothesis that larger repeats can be favored by natural selection because they can avoid disrupting functional elements.

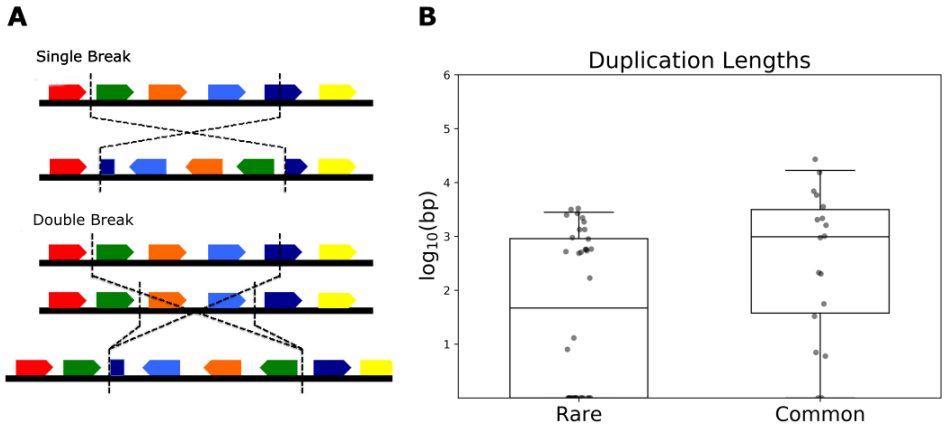

Figure A4.1: A) Staggered breakpoints generate duplications. These might suppress the impacts of sequence disruption**.** In the single break, the dark blue gene span is divided in half in the inverted line with no functional copies remaining. In the double break, the dark blue gene span is duplicated and a functional copy remains to the right of the break region. **B) Common inversion breakpoints exhibit longer duplications.** The boxplots represent duplication lengths of each inversion class. Note that the Y-axis is in logarithmic scale, and all short deletions were set to length 1.

A4.8: Inversions Could Alter Local Regulatory Environments

Impacts on gene regulation in the regions surrounding inversion breakpoints is also likely to be an important determinant of inversion fitness. By translocating large sections of the genome, inversions can reshape local regulatory environments and interfere with nuclear structures. They can separate enhancers from their gene targets,

bring chromatin marks of varying kinds into close proximity, and alter the content and size of local regulatory domains. Translocations of repressive chromatin marks can lead to the silencing of nearby genes, such as in the phenomenon of position-effect variegation, which is variable silencing of a gene near a translocated section of heterochromatin (Cryderman et al 1998; Eissenberg et al 1992; Puig et al 2004; Shatskikh et al 2018; Vogel et al 2008). Chromatin environments also guide the activity of different double-strand break repair mechanisms including non-homologous end joining, which may serve as a mutational bias in the occurrence of inversions (Lemaître and Soutoglou 2014; Marnef et al 2017). We investigated the occurrence of inversions in different chromatin domains, hypothesizing that both mutational biases and selective pressures may influence breakpoints within these domains.

We examined patterns related to chromatin states and marks at two resolutions. The coarser resolution is the level of TADs. TADs are often highly conserved and associated with coordinated gene regulatory blocks (Cavalli and Misteli 2013). In *D. melanogaster*, TADs have been identified through high-resolution chromatin conformation capture, or HiC, sequencing and found to contain distinct chromatin states (Sexton et al 2012). While any inversion whose breakpoints occurs within these domains can and does alter relative TAD boundary positions, inversions with breakpoints that capture boundary elements within associated duplicated regions might form entirely new boundaries and TADs by duplicating those boundary elements. We hypothesized that inversion breakpoints

would be less likely to duplicate boundary elements at higher population frequencies, as the formation or division of TADs may be more deleterious than resizing them.

Only two polymorphic inversion breakpoints, one rare and one from the common inversion In(X)A, could have duplicated a boundary element annotated by Sexton et al (2012). This occurs less often than we would expect by chance for both categories (rare $p=0.02275$, rare and common combined $p=0.001$, permutation test). As common inversions have a modest sample size ($n=9$), no level of boundary duplication for them alone is statistically significant. The low rates of boundary duplication are relatively invariant across frequency categories, so we speculate that a mutational bias may protect boundary regions from breakage. This could occur through a concentration of bound proteins in boundary regions (Sexton et al 2012). Alternatively, it may be extremely deleterious to duplicate boundary regions, purging these inversions from our rare inversion dataset as well as from inversions at higher frequencies.

We also discovered an enrichment of inversion breakpoints within TADs marked with active chromatin by Sexton et al (2012) (rare $p=0.003$, common $p=0.055$, fixed $p<0.001$, all categories $p<0.001$, permutation test). As part of our hypothesis that mixing chromatin states is deleterious, we investigated correlations between domain annotations at either end of an inversion- that is, whether the identity of the domain at one inversion breakpoint is correlated with the domain type at the other inversion breakpoint. There does not appear to be any enrichment for particular combinations of chromatin environments around each breakpoint in our inversions

(p=0.36, chi-square test), suggesting that the biased distribution of inversion breakpoint chromatin domains is driven by a marginal increase of breakpoints within active regions rather than a pairwise effect of breakpoint adjacent chromatin domains. The increased rate of breakpoints in active regions is consistent with a mutational bias, as it is relatively invariant across frequency categories. This bias could be due to a difference in the rate of occurrence of double-strand breaks in open chromatin, or it could be due to a difference in the accuracy and efficiency of double-stranded break repair in these active environments (Marnef et al 2017).

Because this coarse representation may not fully represent the role of chromatin environment on inversion occurrence, we additionally examined enrichment of chromatin marks at a finer scale. Kharchenko et al (2011) created a genome-wide dataset representing local chromatin mark enrichment, represented as a computationally-derived nine-state model with high resolution. We binned this data into windows of 10kb and examined the regions immediately surrounding each inversion breakpoint for the presence of Kharchenko's chromatin states, assigning each window to a general state of "active", "inactive", or "mixed". We found that the enrichment of inversion breakpoints in active regions was not replicated in this finer-scale dataset (p=0.59, permutation test; Figure A4.2). It is likely that this is because on these scales the majority of windows designated "active" contain genes and breakpoints are less likely to occur within genes than in intergenic regions (see Section A4.6).

We discovered an enrichment of mixed chromatin activity states (i.e., both active and inactive states) in windows around fixed inversion breakpoints ($p<0.001$, permutation test) and a decreased occurrence of fixed inversion breakpoints within windows containing only inactive states ($p=0.0075$, permutation test; Figure A4.2). Common inversions show a similar pattern, though with no significant association between inversion breakpoints and windows containing only active chromatin states, but with a significant depletion of breakpoints in windows designated inactive ($p=0.026$, permutation test). Rare inversion breakpoints appear randomly distributed with regards to their fine-scale chromatin environments (Figure A4.2). It is possible that the common inversions would reflect this same pattern with a larger sample size, but we are limited by the scarcity of high frequency inversions. We additionally explored whether there is a correlation between local chromatin windows between two breakpoints of each inversion, but found no statistical enrichment ($p=0.25$, chi-square test). Overall these results suggest that high-frequency breakpoints tend to occur on epigenetic boundaries within regions of the genome that contain some active chromatin.

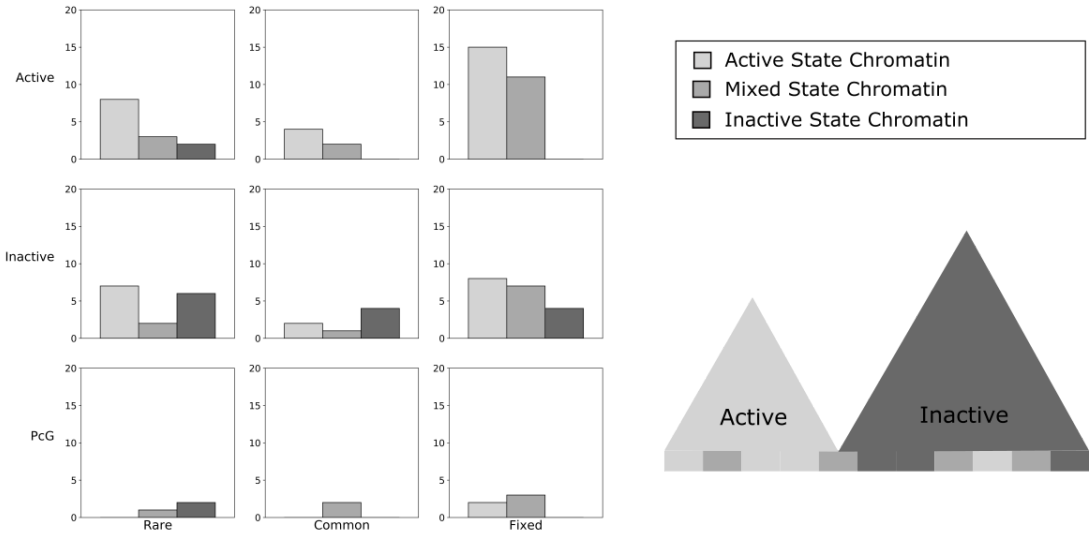

Figure A4.2: Chromatin around Inversion Breakpoints is Active and Heterogenous. Inversion breakpoints are grouped by frequency into columns, then by domain annotation from Sexton et al into rows. Each group is further subdivided into three bars, representing the overall states assigned to the 10kb chromatin windows around each break in that combination of categories (*i.e.,* using the Kharchenko et al. dataset for chromatin state assignment, Methods). The Y-axis of each plot is the count of inversion breakpoints which exist in that combination of states and frequencies. The plot below the legend displays the intuitive relationship between these two levels of annotation; that is, the larger domain annotations (triangles) contain a heterogenous but biased set of local chromatin states (bars).

A4.9: Insulators Maintain Boundaries of Local Regulatory Environments

We discovered an enrichment for inversions within active domains, including inversions which occur between pairs of active and inactive domains. This led us to ask whether we can identify a candidate compensatory mechanism that might suppress the disruption of local regulatory environments by the translocation of chromatin-defining elements. Insulator elements are key to the structure and function of genome regulatory networks, serving as structural anchor points, physical blockers

of enhancer interactions, and boundary elements between TADs or chromatin compartments (Chung et al 1993; Negre 2010; Roseman et al 1993; Sexton et al 2012). Insulators may reduce or prevent the effects of repressive chromatin on local gene activity after translocation (Bushey et al 2008; Gaszner and Felsenfeld 2006; Sigrist and Pirrotta 1997; Yang and Corces 2012). In fact, these elements have been previously shown to be associated with fixed structural rearrangement breakpoints that alter local synteny, which includes inversion breakpoints (Negre et al 2010).

Strong association with these elements may act as a compensatory mechanism that preserves local chromatin state and thereby allows inversions to occur between heterochromatic and euchromatic regions while minimizing negative consequences (Figure A4.3A). Insulator element binding sites are strongly associated with local active and mixed chromatin window states in our data (active p=4.7e-24, mixed p=2.7e-8, Fisher's exact test) and correspondingly rare in windows annotated as inactive (p=2.7e-14, Fisher's exact test). This further supports that these insulator elements represent epigenetic boundaries, and a strong association with these insulator elements may explain the enrichment of mixed chromatin window states around higher frequency inversions in our data.

Association with insulator elements may also prevent ectopic enhancer activity, or the activation of genes other than the target gene by an enhancer (Figure A4.3A). Previous studies have shown developmental disorders can occur in mammals from inversion rearrangements with no additional mutation via ectopic enhancer activity (Ren and Dixon 2015; Lupianez et al 2016). Notably, recent data suggests

inversion breakpoints are rarely associated with local perturbed expression in *Drosophila* (Fuller et al 2016; Ghavi-Helm et al 2019; Said et al. 2018; Lavington and Kern 2017). However, there still appears to be some cases of ectopic enhancer/promoter interactions as in *subdued* and *Dscam4* (Ghavi-Helm et al 2019). This suggests that new ectopic interaction may be promoted by inversions but that these interactions rarely impact overall expression levels.

Our data show that inversion breakpoints are significantly closer to insulator elements than would be expected for randomly distributed breakpoints (rare $p=0.0446$, common $p<0.001$, fixed $p<0.001$, permutation test, Figure A4.3). Common inversion breakpoints are significantly closer to insulators than are rare inversions breakpoints ($p = 0.0373$, Mann-Whitney U test), fixed and common inversion breakpoints are not statistically different ($p=0.312$, Mann-Whitney U test) and fixed inversion breakpoints are significantly closer to insulator elements than are rare inversion breakpoints ($p=0.00242$, Mann-Whitney U test). We asked whether insulators are found within or outside the duplicated regions and found no evidence for any directionality to the association (Figure A4.5). We additionally note that insulator binding sites are much more common in active topologically associated domains (Figure A4.6), and that correspondingly inversions in active regions are somewhat more closely associated with insulators across frequency categories ($p=0.075$, Mann-Whitney U test).

Because inversion breakpoints tend to occur in active domains, the enrichment for proximity to insulator elements might result from the increased density of

insulator elements within those regions and not for selection for proximity to insulators *per se*. We saw a bias towards breakpoints situated in active domains, particularly among fixed inversions. We therefore tested for an association between fixed inversion breakpoints that occur within inactive domains and proximity to insulator elements. After controlling for the domain type, the close proximity of insulators persists, suggesting the observation is partially independent of the correlation between active domains and insulators ($p=0.03$, Mann-Whitney U test). We thus conclude that this is not likely to be a strong confounding factor for our insulator association results. Conversely, if proximity to insulators is beneficial for any reason, this may explain the enrichment of mixed chromatin states around high-frequency inversions.

While these results are consistent with the idea that insulators are a compensation mechanism preventing misexpression, our speculation is not directly proven here. A recent work discovered a lack of gene expression differences over and around inversion breakpoint regions in balancer chromosomes (Ghavi-Helm et al 2019). Because they are maintained as heterozygotes, balancer chromosomes are shielded from selection in homozygotes in a similar fashion to rare inversions. We investigated proximity to insulator elements as a potential cause for this lack of misexpression associated with the balancer chromosome inversion breakpoints but found no association ($p=0.57$, Mann-Whitney). The result challenges our specific hypothesis because Ghavi-Helm et al. found minimal differential gene expression around balancer inversion breakpoints despite the lack of insulator association. The

association between proximity to insulators and population frequencies is robust in our data, but Ghavi-Helm et al's results suggest it may not be necessary to prevent disruption of gene expression. A potential explanation may be that insulators affect more subtle gene expression phenotypes. For example, if the presence of insulator elements decreases the variance in expression across cells or facilitates coordinated timing of expression during development, this might not be reflected in mean expression values obtained via bulk RNA-seq. Regardless of the mechanism, our results suggest that inversions, and possibly synteny changes in general, are more fit when associated with insulator elements in the *D. melanogaster* genome.

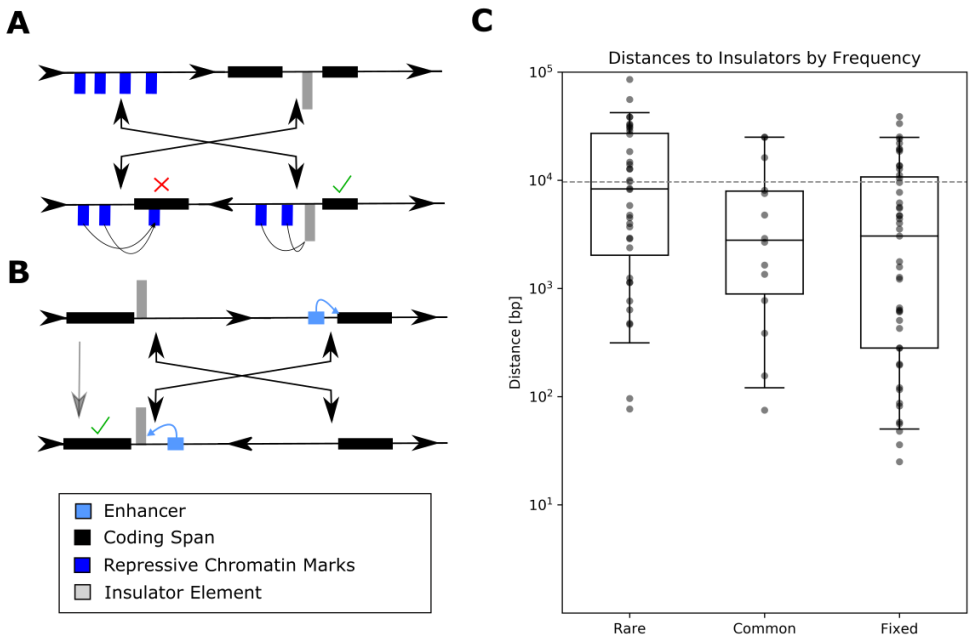

Figure A4.3: A) Insulators could prevent chromatin repression across inversion breakpoints. This pair of hypothetical inversions have four breakpoints that do not disrupt gene spans. The left inversion breakpoint is in a repressive gene-free chromatin landscape indicated by blue chromatin marks, while the other is between two genes separated by an insulator element. After inversion, repressive chromatin marks are relocated to be adjacent to one of the genes, but their repressive effect is blocked by the insulator. The check mark indicates normal gene expression, while the cross indicates disrupted gene expression. **3B) Insulators block enhancer activity.** This simple example displays how an inversion may translocate an enhancer into the proximity of a new gene. In this example, the enhancer's activity is blocked by the presence of an insulator, in grey. **3C) Higher population frequency correlates with insulator proximity.** The distributions of distances are from breakpoints to the nearest insulator element. Note that the y-axis is a logarithmic scale. The dashed grey line is the expected median distance of a random-breakpoint model.

A4.10: Cross Feature Analysis

Each of the features that we examined in this work do not exist in isolation, and it is possible that interactions among features also impact the fitness of new arrangements. Our primary hypothesis for insulator association is that it compensates

for the negative effects of other features; for example, mixing chromatin may be permissible when insulator elements near the inversion breakpoint suppress expression modifying effects (see Section A4.9). Therefore we performed several cross analyses between the features explored here both without condition and conditioning on population frequencies of inversions studied in this work. The results of the cross-feature analysis are described in Section A4.15. We discovered a few obvious associations between features, such as gene disruption and active chromatin, which is expected given that active chromatin is typically associated with genes. We discovered no feature correlations based on population frequency, though this may be due to the modest sample sizes, which likely limited our power to detect interaction effects among features considered here.

Despite our lack of statistical power, we do observe individual cases where a combination of features may mitigate negative fitness effects. For example, in the common polymorphic inversion *In(2L)t* the first breakpoint is in a region containing active chromatin states while the second is in a region that contains inactive states. The active region breakpoint is less than 1Kb from an insulator binding motif; it may be that insulator binding at this site limits the influence of repressive chromatin on the other side of the breakpoint. The common polymorphic inversions *In(2R)NS*, *In(3L)P*, and *In(3R)K* are similarly arranged, with breakpoints occurring in regions of active and inactive chromatin marks and with an insulator binding site very near to the active region breakpoint. Consistent with this idea, Said et al (2018) recently showed that breakpoint adjacent genes in *In(2L)t* and *In(3R)K* are expressed at similar levels

between standard and inverted arrangements. These breakpoint arrangements of common inversions are therefore qualitatively consistent with the idea that insulator elements could suppress deleterious consequences of mixing chromatin states.

A4.11: Conclusion

In this work, we present evidence that mutational biases and natural selection have played a role in shaping the fine-scale distribution of chromosomal inversion breakpoints in the *D. melanogaster* subgroup. Natural selection likely plays a role in maintaining gene sequences, as we found that high-frequency inversions are less likely to disrupt gene coding spans and that they produce correspondingly longer tandem duplications. We also identified two levels of association between chromatin activity and inversion frequency- there appears to be a mutational bias towards occurrence in active regions of the genome, but inversions also tend to occur in areas with locally mixed chromatin states. We found evidence explaining this second pattern consistent with natural selection for the association of inversion breakpoints with insulator elements, which are in turn strongly associated with these mixed chromatin states. Our analyses therefore clarify the mutational context and fitness impacts of novel chromosomal inversions in natural populations and guide future research into specific fitness and gene regulatory impacts of chromosomal inversions.

A stronger understanding of the factors underlying the distribution of polymorphic inversions is requisite for the study of their evolutionary impact. Although it is possible that breakpoint effects sometimes increase the fitness of new

arrangements, such as by creating new expression patterns of transposed loci or chimeric transcripts (Puerma et al 2016), our data are consistent with the conclusion that selection acts against breakpoints that disrupt functional elements within breakpoint regions, consistent with studies in other *Drosophila* species (Fuller et al 2017). Factors that mitigate fitness costs determine what parts of the genome can tolerate polymorphic inversions that maintain complex phenotypes. Variations in these factors may also explain differences in the formation and role of inversions across species. Before an inversion can be selected for recombination suppression or other features it must form in a genomic context where the presence of inversion breakpoints is not immediately and strongly detrimental. Given two additional features of chromosomal inversions, (1) the *de novo* mutation rate is likely very low (Krimbas and Powell 1992), and (2) the conditions for a new arrangement to be favored by natural selection are sometimes restrictive (Charlesworth et al 2018; Kirkpatrick and Barton 2006; Hoffmann et al 2008), the impacts of fine-scale inversion breakpoint positions on the fitness of new arrangements suggest that the availability of suitable, high fitness arrangements may often be rate-limiting for adaptive evolution when suppressed recombination is favorable.

A4.12: Supplementary- Breakage Within Polytene Domains

Polytene chromosomes are bundles of highly replicated chromosomes created in some salivary gland cells in *Drosophila melanogaster* (Urata et al 1995). These highly duplicated homologous chromosomes are aligned along their entire length and

overall compartmentation shared between homologs is visible (Urata et al 1995). These domains were found to represent large-scale TADs (Eagen et al 2015). We performed a separate analysis examining HiC-derived TADs from polytene chromosomes, sourced from Eagen et al (2015). Eagen et al included in their work a direct comparison with the dataset of Sexton et al, which showed substantial concordancy, making these results comparable. It is notable that the dataset created by Eagen et al (2015) has TADs that are substantially larger with a mean length of 200k, versus the mean 100k of Sexton et al (2012)'s domains ($p=5.5x10^{-59}$, Mann-Whitney U test). Of additional importance is that Sexton et al (2012)'s domains are continuous along a chromosome- that is, there were no regions in between TADs not assigned to a TAD. This is a confounding issue for examining the occurrence of inversion breakpoints within non-TAD regions, as only the telomeres are not assigned to some TAD. Eagen et al (2015), by contrast, has substantial gaps between their TADs, with a median of 105kb of space between each annotated TAD. These gaps allow us to analyze the frequencies of inversion breakpoints within versus between annotated TADs.

Fixed inversion breakpoints were less likely to be found within polytene domains ($p=0.01$, permutation test). Common and rare polymorphic inversion breakpoints do not exhibit significant reduced occurrence within polytene TAD domains (rare $p=0.1$; common $p=0.09$; permutation test). As the fixed inversions primarily occurred on lineages other than *Drosophila melanogaster*, the co-evolution of polytene domains as a consequence of previously fixed inversions is unlikely to be

the explanation for this pattern. We therefore conclude that there may be a weak pattern of a reduction of inversion breakpoint formation within TADs versus inter-TAD regions, at least among TADs which are reflected as bands in salivary polytene chromosomes. Our conclusions are further complicated by the fact that polytene chromosomes only occur in specific cell types, such as salivary glands, and so their domains may not represent functional units applicable to most cell types. Further exploration of the spatial conformation of the genome directly around these breakpoint regions is needed before stronger conclusions about this relationship can be made.

A4.13: Supplementary- Breakpoint Sequence Divergence

We verified the likely ages of our discovered rare inversions by comparing overall strain divergence between the inverted strains and all other strains. If an inversion had persisted for a long period of time at low frequencies, its breakpoint regions should display excess divergence compared to other strains within the same populations. Divergence values were first calculated as "# of mismatches / (# of matches + # of mismatches)" between each pair of strains with an assembly, where a match was when two bases at a point were concordant and mismatch discordant, discounting any points where either assembly in the pairwise comparison was ambiguous. Each strain then had its mean divergence value from the set of comparisons to all other strains calculated and the distribution of those values was

used to calculate the percentile of average divergence. The results of this analysis are located in table SA4.1.

A4.14: Supplementary- Robustness to Variation in Breakpoint Structures

Limitations of short read sequencing technologies prevent us from verifying the chromosome order has actually changed when no single read pair spans an entire breakpoint region. Some of our candidate rare inversion breakpoints could also be consistent with duplication in one genomic region and insertion in reverse orientation without actual inversion of the intervening chromosome. If these occur, it would manifest as an apparent inversion with a duplication for at least one breakpoint (i.e., the transposed region). We note that this mechanism certainly does not explain all candidate rare inversions with breakpoint duplication structure as one cytologically-verified inversion, *In(2R)Mal*, includes one inversion with only a single breakpoint duplication. Other, more complex structural variants could also generate breakpoints consistent with inverted structures, including potentially two duplications plus transposition. It is not known how common these phenomena are relative to inversion, but note that most common and fixed inversion breakpoints do show evidence of duplication.

We therefore identified and reanalyzed a set of highest-confidence rare inversions which have been cytologically or molecularly characterized previously or have very simple and obvious breakpoint structures that are not consistent with any other simple rearrangement (i.e. “cut-and-paste” breakpoints with no associated

duplications). This "high confidence" set includes *In(2L)DL3, In(2R)Y1a, In(2L)MAL_1, In(2L)MAL_2, In(3L)DL10*, and *In(3R)Gb*. We find that gene disruption by rare inversions is more common in this set (75% versus 44% of breakpoints interrupt genes) and not significantly reduced relative to random expectations (p=0.886, permutation test). Furthermore, high confidence rare inversion breakpoints are more likely to interrupt genes than are common inversion breakpoints (p=0.0236, Fisher's exact test). Consistent with our results from the full set we find no statistical evidence for further reduced disruption of lethal and sterile genes. Additionally, the high-confidence rare inversion breakpoints have shorter inverted duplications than common inversion breakpoints (p=0.0002, Mann-Whitney U test). We note that the apparent increases in the rates of gene disruption and decreases in the sizes of breakpoint-associated duplications in this subset of breakpoints are likely due to the fact that duplication length is one element used to determine inversion confidence. Including the full set of candidate inversion breakpoints is therefore conservative for testing these hypotheses.

Additionally, we find that high-confidence rare inversion breakpoints occur at a greater rate than expected in domains annotated as active by Sexton et al (p=0.001, permutation test) as with the full candidate rare inversion set. They do not occur at significantly higher rates in active chromatin fine-scale windows, however (p=0.44, permutation test). There does not appear to be any relationship between the types of domain each inversion's pair of breakpoints occur in at any scale (domains p=0.1, windows p=0.3, chi-square tests). The high-confidence rare inversion breakpoints are

significantly associated with insulators (p=0.024, percentile) but are further from insulators than common inversions (p=0.044, Mann-Whitney U test) and marginally further than fixed inversions (p=0.087, Mann-Whitney U test). As we state in the main text, colocalization with insulator elements may in part reflect the tendency for breakpoints to occur in active chromatin windows. The high-confidence rare inversion breakpoints appear to behave similarly to the entire set in regards to both chromatin and insulators despite the reduced statistical power.

Overall, our results are not strongly affected by the exclusion of the lower-confidence rare inversion breakpoints, where there are slight differences they make our reported results more conservative, and we do not believe that the inclusion of rare inversions with breakpoint proximal duplications is confounding for our findings in this work.

A4.15: Supplementary- Cross feature correlation analyses: chromatin windows, gene disruption, and insulator proximity

None of the features we explore are completely independent and it is possible that associations with one factor could bias our analysis of another. We therefore performed several cross-correlation analyses where we evaluated associations between features in our dataset and looked for patterns of association among them. Generally, comparisons between a categorical feature such as gene disruption and a continuous feature such as duplication size were performed by conditioning on the categorical feature and performing Mann-Whitney rank-sum tests comparing the

conditional distributions of the continuous feature. Comparisons between continuous features were performed with linear regression and a Wald test. Comparisons between categorical features were performed with a Fisher's Exact test of the proportions of categories conditioned on the other category type. Many of the tests were not possible as conditioning on multiple features and frequency category often left sample sizes too small to retain sufficient statistical power to identify important differences.

Tests involving specific categories of domain breakage were all nonsignificant due to low sample sizes, with no obvious biases. We instead focused on local chromatin window activity as a possible influence on other factors by subsetting the dataset to breakpoints with these window features. Inversion breakpoints in inactive windows were not tested against mixed or active windows as the sample size within inactive windows was too small for statistical significance. Local chromatin window activity being active versus mixed did not exert a direct influence on breakpoint proximity to insulator elements ($p=0.54$, Mann-Whitney U test).

Some tests related to gene disruption did yield significant results. Duplication size among breakpoints which do not disrupt genes are larger than those that do ($p=0.0096$, Mann-Whitney U test) which is expected given our definition of gene disruption. Inversion breakpoints which disrupt genes are also less associated with insulator elements ($p=0.015$, Mann-Whitney U test), likely because insulator binding sites are only rarely found within coding regions. Inversions which disrupt genes largely occur in active chromatin regions ($p=0.01$, Fisher's Exact test), likely because active chromatin marks exist across gene coding spans.

Chromatin window state and gene disruption were also associated with one another; breakpoints in active regions were more likely to disrupt genes than those in inactive regions ($p=0.01$, Fisher's Exact test). This is likely due to the nature of active chromatin regions and their association with gene transcription. This effect is most robust among rare inversion breakpoints ($p=0.051$, Fisher's Exact test), though the statistical strength declines due to low sample size. The effect disappears when looking at common inversion breakpoints ($p=0.19$, Fisher's Exact test) and fixed inversion breakpoints ($p=0.55$, Fisher's Exact test), likely because these categories interrupt genes less frequently than rare inversions in general regardless of chromatin window state.

Finally, insulator proximity and duplication size were uncorrelated ($p=0.98$, Wald test). Overall, we conclude that there are some connections between features such as chromatin window size and gene disruption, but that these effects are not strongly related to the frequencies of inversions examined here and therefore unlikely to confound our frequency category comparison.

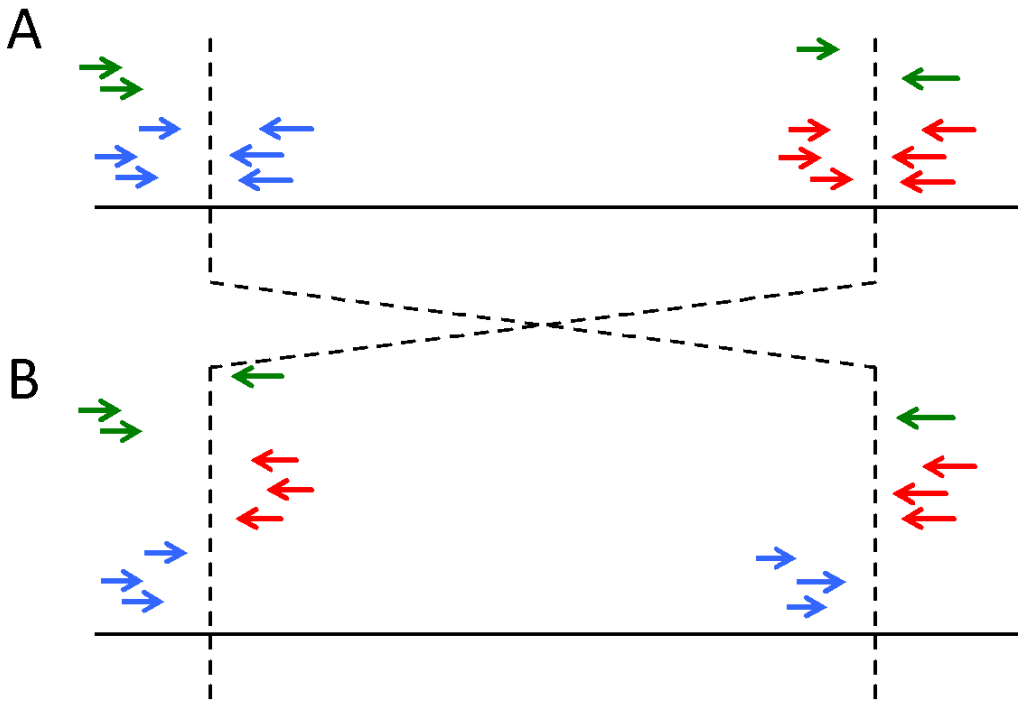

Figure A4.4: Mapping positions along the genome representing inversion breakpoints. (A) Mapping positions along the genome from which short reads that support the presence of inversion breakpoints are derived. (B) Mapping positions along the genome where those reads will map on the standard arrangement reference genome that does not have this inversion. Read pairs that map in forward-forward orientation on the reference genome are shown in blue, read pairs that map in reverse-reverse orientation are shown in red. Finally, reads whose pair does not map, presumably because it overlaps the breakpoint, are shown in green. All reads show, as well as their unmapped pairs and additional adjacent standard mapping read pairs, were used to *de novo* assemble breakpoint adjacent regions.

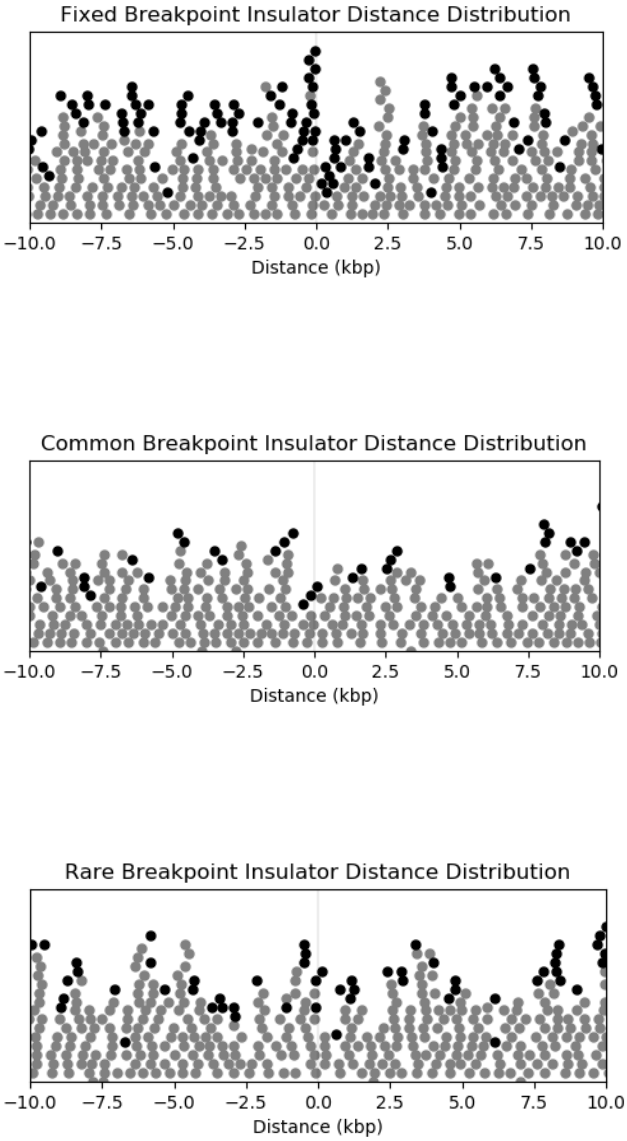

Figure A4.5: Insulator Distance Distribution: Vertical swarmplot of insulator distance values in a 10kbp window on either side of the set of breakpoints. The x-axis is the distance in bases for each permuted insulator. Grey dots are permuted expectations; black is real data. No directional enrichment is evident.

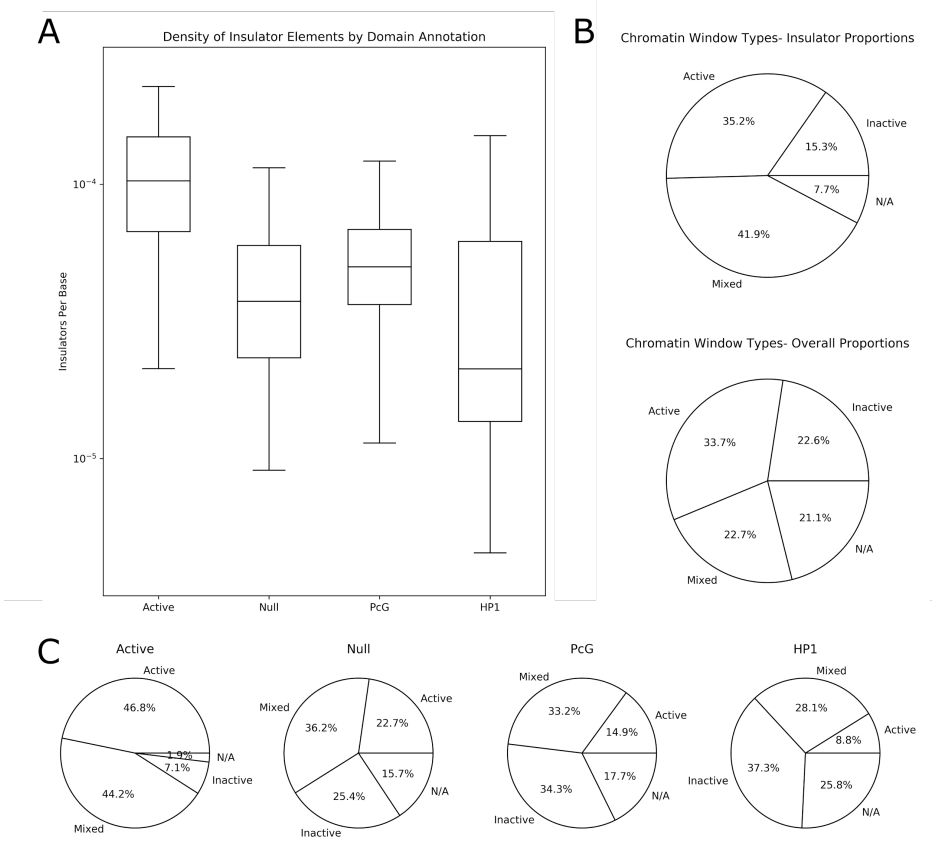

Figure A4.6: Chromatin Marker State Windows- Insulators and Domains

The above figure displays the pairwise relationships between the domain chromatin annotations, the 10kb chromatin windows, and the distribution of insulator binding sites. A the density of insulator elements per base between domains with varying annotations. We can see that insulators are much denser in active regions. **B** The 10kb chromatin windows surrounding insulator elements (upper) and across the entire genome (lower). We can see that the windows which contain insulator elements are strongly enriched for active and mixed chromatin states. **C** is the third pairwise relationship, that between the 10kb chromatin windows and the domain chromatin annotations. We see that domains annotated “Active” are strongly enriched for both active and mixed chromatin states, “Null” domains are relatively evenly split between states, “PcG” domains are primarily inactive and mixed, and “HP1” domains have very little active windows. Note that a chromatin window labeled “N/A” means there was no marker information in that window, which likely indicates highly repetitive regions or a lack of any kind of distinct marker enrichment.

Name	Percentile	Inversion Divergence	Most Diverged Value	Least Diverged Value	Count of Missing Assemblies
In(2L)DL1_1	78.17258883	0.013337749	0.015094053	0.010209212	0
In(2L)DL1_2	1.52284264	0.00772233	0.010368947	0.007557807	0
In(2L)DL2_1	41.66666667	0.007548929	0.009924965	0.006788092	2
In(2L)DL2_2	6.25	0.003030571	0.004950918	0.00295072	2
In(2R)DL4_1	52	0.00493158	0.009634861	0.004113951	10
In(2R)DL4_2	90	0.006984834	0.008292514	0.003785763	15
In(2R)MAL_1_1	72.08121827	0.008800419	0.010477536	0.005842443	0
In(2R)MAL_1_2	80.7106599	0.016620079	0.018559695	0.012634134	0
In(2R)MAL_2_1	99.49238579	0.005584235	0.005657727	0.003371302	0
In(2R)MAL_2_2	9.137055838	0.009280233	0.013281644	0.008329574	0
In(3L)DL5_1	97.87234043	0.009548998	0.009715444	0.00578075	1
In(3L)DL5_2	12.76595745	0.005099891	0.008850537	0.004851756	1
In(3L)DL6_1	81.21827411	0.012675769	0.015284631	0.010456063	0
In(3L)DL6_2	84.26395939	0.013024658	0.014283207	0.009468615	0
In(3L)DL7_1	32.98969072	0.00750953	0.010768587	0.005973036	11
In(3L)DL7_2	80.44692737	0.009858011	0.012344136	0.006002733	26
In(3L)DL9_1	18.59296482	0.004877904	0.008827557	0.004398443	6
In(3L)DL9_2	54.87179487	0.008071974	0.011122911	0.00655541	10
In(3R)G_1	93.33333333	0.003646416	0.004202843	0.00207387	40
In(3R)G_2	80	0.004841514	0.005379021	0.003319676	35
In(3R)DL12_1	15.52795031	0.000987473	0.00185542	0.00079961	44
In(3R)DL12_2	30.81395349	0.002945007	0.006007908	0.002326722	33
In(3R)DL14_1	85.40145985	0.012131614	0.01330209	0.009463809	0
In(3R)DL14_2	72.99270073	0.004730172	0.005748614	0.002839291	0

Table A4.1: Breakpoint Divergence. This table displays for each breakpoint the name and sequence divergence of the breakpoint as well as the highest and lowest divergences observed, and the number of stains missing assemblies over the breakpoint region. This table only includes versions for which an assembly of the chromosome arm for that strain was available. No annotated inversion-bearing strain

was either the most or least divergent strain in its population in a window of 10 kilobases around the breakpoint, indicating that these inversions are likely not exceptionally old.

Label	Chromosome	Frequency	Gene Interrupted	Allele Phenotype
In(2R)MAL_1	2R	Rare	FBgn0034389	Lethal
In(2R)MAL_2	2R	Rare	FBgn0024189	Lethal
In(3L)DL10	3L	Rare	FBgn0036447	Lethal
In(3R)Gb	3R	Rare	FBgn0015589	Lethal
In(3R)DL14	3R	Rare	FBgn0038211	Lethal
In(3R)DL14	3R	Rare	FBgn0250823	Lethal
In(3R)Dl14	3R	Rare	FBgn0029881	Sterile
In(X)A	X	Common	FBgn0030505	Lethal
In(2L)t	2L	Common	FBgn0031403	Lethal
X(1)	X	Fixed	FBgn0026089	Lethal
2L(1)	2L	Fixed	FBgn0031646	Lethal
2L(1)	2L	Fixed	FBgn0031968	Lethal
2L(7)	2L	Fixed	FBgn0051678	Lethal

Table A4.2: Inversion Allele Phenotypes. This table is a display of all lethal phenotype genes disrupted by our breakpoint dataset.

Fixed Inversion	Ortholog Present?	Distance in bp to nearest edge of disrupted gene span
X(1)	Yes	101
X(2)	Yes	278
X(5)	Yes	126
X(6)	Yes	928
2L(1)	Yes	33
2L(1)	Yes	278
2L(2)	Yes	125
2L(6)	Yes	323
2L(7)	Yes	202
2L(7)	Yes	1
2L(8)	Yes	71
3L(1)	Yes	68
3R(7-8)	Yes	626

Table A4.3: Breakpoints Disrupting Genes. This table contains all fixed breakpoints that appear to disrupt a gene regardless of whether the gene is annotated with lethal or sterile phenotypes. We note that in all cases there is a *D. yakuba* ortholog to the gene present and that the closest distance between a forward or reverse strand break and a span edge is less than 1000 bases. This suggests that all of these disruption events are false positives due to bias and low accuracy of breakpoint estimation in the rare category (see Section A4.5).

Frequency	Allele Phenotype	p-value of reduction
Rare	Lethal	0.57
Rare	Sterile	0.53
Common	Lethal	0.54
Common	Sterile	0.35
Fixed	Lethal	0.36
Fixed	Sterile	0.21

Table A4.4: Deleterious Phenotype p-values

These are the non-significant p-values for all categories of strongly deleterious phenotype annotated for genes possibly disrupted. The lack of significance may be due to a combination of low sample size and the nature of the drop in gene disruption rates. As gene disruption levels drop among higher-frequency inversions, and as only a small proportion of genes are annotated with either phenotype, we lack power to detect differences in the rates of disruption for these genes.

Data A4.1

This file contains the sequences of the breakpoint-spanning contigs for the candidate rare inversions in fasta format. It can be downloaded at 10.5281/zenodo.7566973.

Bibliography

Alpert, T. et al. Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. Cell 184, 2595-2604.e13 (2021).

Anderson, T. K. et al. A Phylogeny-Based Global Nomenclature System and Automated Annotation Tool for H1 Hemagglutinin Genes from Swine Influenza A Viruses. mSphere 1,
e00275-16 (2016).

Ané C, Sanderson MJ. 2005. Missing the Forest for the Trees: Phylogenetic Compression and Its Implications for Inferring Complex Evolutionary Histories. Systematic Biology 54:146–157.

Aulard, S., David, J., and Lemeunier, F. (2002). Chromosomal inversion polymorphism in Afrotropical populations of Drosophila melanogaster. Genetical Research 79, 49–63.

Bellen, H.J., Levis, R.W., He, Y., Carlson, J.W., Evans-Holm, M., Bae, E., Kim, J., Metaxakis, A., Savakis, C., Schulze, K.L., et al. (2011). The Drosophila Gene Disruption Project: Progress Using Transposons With Distinctive Site Specificities. Genetics 188, 731–743.

Bello, X. et al. CovidPhy: A tool for phylogeographic analysis of SARS-CoV-2 variation. Environmental Research 204, 111909 (2022).

Brenner, F. W., Villar, R. G., Angulo, F. J., Tauxe, R. & Swaminathan, B. Salmonella Nomenclature. Journal of Clinical Microbiology 38, 2465–2467 (2000).

Brito, A. F. et al. Global disparities in SARS-CoV-2 genomic surveillance. medRxiv 2021.08.21.21262393 (2021) doi:10.1101/2021.08.21.21262393.

Bushey, A.M., Dorman, E.R., and Corces, V.G. (2008). Chromatin Insulators: Regulatory Mechanisms and Epigenetic Inheritance. Molecular Cell 32, 1–9.

Butlin, R.K., Read, I.L., and Day, T.H. (1982). The effects of a chromosomal inversion on adult size and male mating success in the seaweed fly, Coelopa frigida. Heredity 49, 51.

Cáceres, M., Barbadilla, A., and Ruiz, A. (1997). Inversion Length and Breakpoint Distribution in the Drosophila Buzzatii Species Complex: Is Inversion Length a Selected Trait? Evolution 51, 1149–1155.

Calvete, O., González, J., Betrán, E., and Ruiz, A. (2012). Segmental Duplication, Microinversion, and Gene Loss Associated with a Complex Inversion Breakpoint Region in Drosophila. Mol Biol Evol 29, 1875–1889.

Castermans, D., Vermeesch, J.R., Fryns, J.-P., Steyaert, J.G., Ven, W.J.M.V. de, Creemers, J.W.M., and Devriendt, K. (2007). Identification and characterization of the TRIP8 and REEP3 genes on chromosome 10q21.3 as novel candidate genes for autism. European Journal of Human Genetics 15, 422.

Cavalli, G., and Misteli, T. (2013). Functional implications of genome topology. Nature Structural & Molecular Biology 20, 290.

Chaillon A, Smith DM. 2021. Phylogenetic analyses of SARS-CoV-2 B.1.1.7 lineage suggest a single origin followed by multiple exportation events versus convergent evolution. Clinical Infectious Diseases [Internet]. Available from: https://doi.org/10.1093/cid/ciab265

Charlesworth, B., and Barton, N.H. (2018). The Spread of an Inversion with Migration and Selection. Genetics 208, 377–382.

Chen, C. et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics 38, 1735–1737 (2022).

Chung, J.H., Whiteley, M., and Felsenfeld, G. (1993). A 5’ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. Cell 74, 505–514.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218.

Collins, C. H., Yates, M. D. & Grange, J. M. Subdivision of Mycobacterium tuberculosis into five variants for epidemiological purposes: methods and nomenclature. Epidemiology & Infection 89, 235–242 (1982).

Colson, P. & Raoult, D. Global Discrepancies between Numbers of Available SARS-CoV-2 Genomes and Human Development Indexes at Country Scales. Viruses 13, 775 (2021).

Corbett-Detig, R.B., and Hartl, D.L. (2012). Population Genomics of Inversion Polymorphisms in Drosophila melanogaster. PLOS Genetics 8, e1003056.

Corbett-Detig, R.B., Said, I., Calzetta, M., Genetti, M., McBroome, J., Maurer, N.W., Petrarca, V., della Torre, A., and Besansky, N.J. (2019). Fine-Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the Anopheles gambiae Species Complex Using Proximity-Ligation Sequencing. Genetics 213, 1495.

COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe 1, e99–e100 (2020).

Cridland, J.M., and Thornton, K.R. (2010). Validation of Rearrangement Break Points Identified by Paired-End Sequencing in Natural Populations of Drosophila melanogaster. Genome Biol Evol 2, 83–101.

Cryderman, D.E., Cuaycong, M.H., Elgin, S.C.R., and Wallrath, L.L. (1998). Characterization of sequences associated with position-effect variegation at pericentric sites in Drosophila heterochromatin. Chromosoma 107, 277–285.

Cuypers, L. et al. Time to Harmonize Dengue Nomenclature and Classification. Viruses 10, 569 (2018).

Cyranoski D. 2021. Alarming COVID variants show vital role of genomic surveillance. Nature 589:337–338.

da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, Ashraf S, Balcazar C, Brunker K, Campbell A, et al. 2021. Genomic

epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. Nature Microbiology 6:112–122.

de Bernardi Schneider, A. and Wolfinger, M. T. Molecular epidemiology of Venezuelan equine encephalitis complex viruses. bioRxiv (2023).

de Bernardi Schneider, A. et al. Updated Phylogeny of Chikungunya Virus Suggests Lineage-Specific RNA Architecture. Viruses 11, 798 (2019).

Dellicour, S. et al. A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages. Molecular Biology and Evolution 38, 1608–1613 (2021).

Deng X, Gu W, Federman S, Plessis L du, Pybus OG, Faria NR, Wang C, Yu G, Bushnell B, Pan C-Y, et al. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science 369:582–587.

Dobzhansky, T. (1962). Rigid vs. Flexible Chromosomal Polymorphisms in Drosophila. The American Naturalist 96, 321–328.

du Plessis, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science 371, 708–712 (2021).

Eissenberg, J.C., Morris, G.D., Reuter, G., and Hartnett, T. (1992). The heterochromatin-associated protein HP-1 is an essential protein in Drosophila with dosage-dependent effects on position-effect variegation. Genetics 131, 345–352.

Facklam, R. What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. Clinical Microbiology Reviews 15, 613–630 (2002).

Falk, M., Lukasova, E., and Kozubek, S. (2010). Higher-order chromatin structure in DSB induction, repair and misrepair. Mutation Research/Reviews in Mutation Research 704, 88–100.

Fernandes JD, Hinrichs AS, Clawson H, Gonzalez JN, Lee BT, Nassar LR, Raney BJ, Rosenbloom KR, Nerli S, Rao AA, et al. 2020. The UCSC SARS-CoV-2 Genome Browser. Nature Genetics 52:991–998.

Fitch, W. M. On the Problem of Discovering the Most Parsimonious Tree. The American Naturalist 111, 223–257 (1977).

Forrester, Naomi L., et al. "Evolution and spread of Venezuelan equine encephalitis complex alphavirus in the Americas." PLoS neglected tropical diseases 11.8 (2017): e0005693.

Frischer, L.E., Hagen, F.S., and Garber, R.L. (1986). An inversion that disrupts the Antennapedia gene causes abnormal structure and localization of RNAs. Cell 47, 1017–1023.

Fuller, Z.L., Haynes, G.D., Richards, S., and Schaeffer, S.W. (2016). Genomics of Natural Populations: How Differentially Expressed Genes Shape the Evolution of Chromosomal Inversions in Drosophila pseudoobscura. Genetics 204, 287–301.

Fuller, Z.L., Koury, S.A., Phadnis, N., and Schaeffer, S.W. (2019). How chromosomal rearrangements shape adaptation and speciation: Case studies in Drosophila pseudoobscura and its sibling species Drosophila persimilis. Mol Ecol 28, 1283–1301.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nature Reviews Genetics 7, 703.

Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korbel, J.O., and Furlong, E.E.M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nature Genetics 1.

Gill, M. S., Lemey, P., Suchard, M. A., Rambaut, A. & Baele, G. Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. Molecular Biology and Evolution 37, 1832–1842 (2020).

Gómez-Carballa, A., Bello, X., Pardo-Seco, J., Martinón-Torres, F. & Salas, A. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. Genome Res. 30, 1434–1448 (2020).

González, J., Casals, F., and Ruiz, A. (2007). Testing Chromosomal Phylogenies and Inversion Breakpoint Reuse in Drosophila. Genetics 175, 167–177.

Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. Virus Evolution 8, veac021 (2022).

Grenier, J.K., Arguello, J.R., Moreira, M.C., Gottipati, S., Mohammed, J., Hackett, S.R.,
Boughton, R., Greenberg, A.J., and Clark, A.G. (2015). Global Diversity Lines–A Five-Continent Reference Panel of Sequenced Drosophila melanogaster Strains. G3: Genes, Genomes, Genetics 5, 593–603.

Guillén, Y., and Ruiz, A. (2012). Gene alterations at Drosophila inversion breakpoints provide prima facie evidence for natural selection as an explanation for rapid chromosomal evolution. BMC Genomics 13, 53.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

Hodcroft EB, Maio ND, Lanfear R, MacCannell DR, Minh BQ, Schmidt HA, Stamatakis A, Goldman N, Dessimoz C. 2021. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature 591:30–33.

Hodcroft, E. B. et al. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature 591, 30–33 (2021).

Hoffmann, A.A., and Rieseberg, L.H. (2008). Revisiting the Impact of Inversions in Evolution:
From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? Annual Review of Ecology, Evolution, and Systematics 39, 21–42.

Hoffmann, A.A., Sgrò, C.M., and Weeks, A.R. (2004). Chromosomal inversion polymorphisms and adaptation. Trends in Ecology & Evolution 19, 482–488.

Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., Svirskas, R., et al. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 25, 445–458.

Hough, R.B., Lengeling, A., Bedian, V., Lo, C., and Bućan, M. (1998). Rump white inversion in the mouse disrupts dipeptidyl aminopeptidase-like protein 6 and causes dysregulation of Kit expression. Proc. Natl. Acad. Sci. U.S.A. 95, 13800–13805.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 24, 1193–1208.

Huynh, L.Y., Maney, D.L., and Thomas, J.W. (2011). Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (Zonotrichia albicollis). Heredity 106, 537.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. 10.1093/bioinformatics/bty191.

Jackson, B. et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell 184, 5179-5188.e8 (2021).

Jost, D., Carrivain, P., Cavalli, G., and Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res 42, 9553–9561.

Kao, J.Y., Zubair, A., Salomon, M.P., Nuzhdin, S.V., and Campo, D. (2015). Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south-eastern United States and Caribbean Islands. Molecular Ecology 24, 1499–1509.

Kapun, M., Schmidt, C., Durmaz, E., Schmidt, P.S., and Flatt, T. (2016a). Parallel effects of the inversion In(3R)Payne on body size across the North American and Australian clines in Drosophila melanogaster. Journal of Evolutionary Biology 29, 1059–1072.

Kapun, M., Fabian, D.K., Goudet, J., and Flatt, T. (2016b). Genomic Evidence for Adaptive Inversion Clines in Drosophila melanogaster. Mol Biol Evol 33, 1317–1336.

Kauffmann, F. The bacteriology of Enterobacteriaceae. Collected studies of the author and his co-workers. (1966).

Keyser, P., Borge-Renberg, K., and Hultmark, D. (2007). The Drosophila NFAT homolog is involved in salt stress tolerance. Insect Biochemistry and Molecular Biology 37, 356–362.

Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 471, 480–485.

Kirkpatrick, M. (2010). How and Why Chromosome Inversions Evolve. PLOS Biology 8, e1000501.

Kirkpatrick, M., and Barton, N. (2006). Chromosome Inversions, Local Adaptation and Speciation. Genetics 173, 419–434.

Knibb, W.R. (1982). Chromosome inversion polymorphisms in Drosophila melanogaster II. Geographic clines and climatic associations in Australasia, North America and Asia. Genetica 58, 213–221.

Kozińska, M., Zientek, J., Augustynowicz-Kopeć, E., Zwolska, Z. & Kozielski, J. Transmission of tuberculosis among people living in the border areas of Poland, the Czech Republic, and Slovakia. Polish Archives of Internal Medicine 126, 32–40 (2019).

Kraemer, M. U. G. et al. Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. Science 373, 889–895 (2021).

Kramer, A. M., Sanderson, T. & Corbett-Detig, R. Treenome Browser: co-visualization of enormous phylogenies and millions of genomes. Bioinformatics 39, btac772 (2023).

Krimbas, C.B., and Powell, J.R. (1992). Drosophila Inversion Polymorphism (CRC Press).

Kuhn, J. H. et al. Nomenclature- and Database-Compatible Names for the Two Ebola Virus Variants that Emerged in Guinea and the Democratic Republic of the Congo in 2014. Viruses 6, 4760–4799 (2014).

Lack, J.B., Cardeno, C.M., Crepeau, M.W., Taylor, W., Corbett-Detig, R.B., Stevens, K.A., Langley, C.H., and Pool, J.E. (2015). The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics 199, 1229–1241.

Lack, J.B., Lange, J.D., Tang, A.D., Corbett-Detig, R.B., and Pool, J.E. (2016). A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. Mol Biol Evol 33, 3308–3313.

Lakich, D., Kazazian, H.H., Antonarakis, S.E., and Gitschier, J. (1993). Inversions disrupting the
factor VIII gene are a common cause of severe haemophilia A. Nat. Genet. 5, 236–241.

Lancefield, R. C. A Serological Differentiation of Human and Other Groups of Hemolytic Streptococci. J Exp Med 57, 571–595 (1933).

Langley, C.H., Stevens, K., Cardeno, C., Lee, Y.C.G., Schrider, D.R., Pool, J.E., Langley, S.A., Suarez, C., Corbett-Detig, R.B., Kolaczkowski, B., et al. (2012). Genomic Variation in Natural Populations of Drosophila melanogaster. Genetics 192, 533–598.

Lavington, E., and Kern, A.D. (2017). The Effect of Common Inversion Polymorphisms In(2L)t and In(3R)Mo on Patterns of Transcriptional Variation in Drosophila melanogaster. G3: Genes, Genomes, Genetics 7, 3659–3668.

Lemaître, C., and Soutoglou, E. (2014). Double strand break (DSB) repair in heterochromatin and heterochromatin proteins in DSB repair. DNA Repair 19, 163–168.

Lemeunier, F., and Ashburner, M.A. (1976). Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc. R. Soc. Lond., B, Biol. Sci. 193, 275–294.

Lemey, P. et al. Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. Nat Commun 11, 5110 (2020).

Lemey, P. et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. Nature 595, 713–717 (2021).

Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian Phylogeography Finds Its Roots. PLOS Computational Biology 5, e1000520 (2009).

Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, eabe3261 (2021).

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293.

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

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature 482, 173.

Mai U, Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19:272.

Maio, N. D. et al. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. 2021.03.15.435416 https://www.biorxiv.org/content/10.1101/2021.03.15.435416v2 (2021) doi:10.1101/2021.03.15.435416.

Marnef, A., Cohen, S., and Legube, G. (2017). Transcription-Coupled DNA Double-Strand Break Repair: Active Genes Need Special Care. Journal of Molecular Biology 429, 1277–1288.

Maxmen A. 2021. One million coronavirus sequences: popular genome site hits mega milestone. Nature 593:21–21.

McBroome, J. et al. A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees. Molecular Biology and Evolution (2021) doi:10.1093/molbev/msab264.

McCrone, J. T. et al. Context-specific emergence and growth of the SARS-CoV-2 Delta variant. 2021.12.14.21267606 (2021) doi:10.1101/2021.12.14.21267606.
Mettler, L.E., Voelker, R.A., and Mukai, T. (1977). Inversion Clines in Populations of Drosophila melanogaster. Genetics 87, 169–176.

Mukai, T., Mettler, L.E., and Chigusa, S.I. (1971). Linkage Disequilibrium in a Local Population of Drosophila melanogaster. PNAS 68, 1065–1069.

Nègre, N., Brown, C.D., Shah, P.K., Kheradpour, P., Morrison, C.A., Henikoff, J.G., Feng, X., Ahmad, K., Russell, S., White, R.A.H., et al. (2010). A Comprehensive Map of Insulator Elements for the Drosophila Genome. PLOS Genetics 6, e1000814.

Nicholls SM, Poplawski R, Bull MJ, Underwood A, Chapman M, Abu-Dahab K, Taylor B, Jackson B, Rey S, Amato R, et al. 2020. MAJORA: Continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. bioRxiv:2020.10.06.328328.

O'Toole, Á. et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol 7, (2021).

Oneal, E., Lowry, D.B., Wright, K.M., Zhu, Z., and Willis, J.H. (2014). Divergent population structure and climate associations of a chromosomal inversion polymorphism across the Mimulus guttatus species complex.

Orengo, D.J., Puerma, E., Papaceit, M., Segarra, C., and Aguadé, M. (2015). A molecular perspective on a complex polymorphic inversion system with cytological evidence of multiply reused breakpoints. Heredity 114, 610–618.

Otto, S. P. et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Current Biology 31, R918–R929 (2021).

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526–528.

Parker J, Rambaut A, Pybus OG. 2008. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infect Genet Evol 8:239–246.

Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).

Pevzner, P., and Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. PNAS 100, 7672–7677.

Pool, J.E., Corbett-Detig, R.B., Sugino, R.P., Stevens, K.A., Cardeno, C.M., Crepeau, M.W., Duchen, P., Emerson, J.J., Saelao, P., Begun, D.J., et al. (2012). Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLOS Genetics 8, e1003080.

Presgraves, D.C., Gérard, P.R., Cherukuri, A., and Lyttle, T.W. (2009). Large-Scale Selective Sweep among Segregation Distorter Chromosomes in African Populations of Drosophila melanogaster. PLOS Genetics 5, e1000463.

Puerma, E., Orengo, D.J., and Aguadé, M. (2016a). The origin of chromosomal inversions as a source of segmental duplications in the Sophophora subgenus of Drosophila. Scientific Reports 6, 30715.

Puerma, E., Orengo, D.J., and Aguadé, M. (2016b). Multiple and diverse structural changes affect the breakpoint regions of polymorphic inversions across the Drosophila genus. Scientific Reports 6, 36248.

Puerma, E., Orengo, D.J., Salguero, D., Papaceit, M., Segarra, C., and Aguadé, M. (2014). Characterization of the Breakpoints of a Polymorphic Inversion Complex Detects Strict and Broad Breakpoint Reuse at the Molecular Level. Mol Biol Evol 31, 2331–2341.

Puig, M., Cáceres, M., and Ruiz, A. (2004). Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA. Proc Natl Acad Sci U S A 101, 9013–9018.

Ragonnet-Cronin, M. et al. Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions. Nat Commun 12, 2188 (2021).

Ramaekers, K. et al. Towards a unified classification for human respiratory syncytial virus genotypes. Virus Evol 6, veaa052 (2020).

Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5, 1403–1407 (2020).

Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66, 846–850 (1971).

Rane, R.V., Rako, L., Kapun, M., Lee, S.F., and Hoffmann, A.A. (2015). Genomic evidence for role of inversion 3RP of Drosophila melanogaster in facilitating climate change adaptation. Molecular Ecology 24, 2423–2432.

Ranz, J.M., Maurin, D., Chan, Y.S., Grotthuss, M. von, Hillier, L.W., Roote, J., Ashburner, M., and Bergman, C.M. (2007). Principles of Genome Evolution in the Drosophila melanogaster Species Group. PLOS Biology 5, e152.

Ren, B., and Dixon, J.R. (2015). A CRISPR Connection between Chromatin Topology and Genetic Disorders. Cell 161, 955–957.

Richard, D. et al. A phylogeny-based metric for estimating changes in transmissibility from recurrent mutations in SARS-CoV-2. 2021.05.06.442903 https://www.biorxiv.org/content/10.1101/2021.05.06.442903v2 (2021) doi:10.1101/2021.05.06.442903.

Rito, T., Richards, M. B., Pala, M., Correia-Neves, M. & Soares, P. A. Phylogeography of 27,000 SARS-CoV-2 Genomes: Europe as the Major Source of the COVID-19 Pandemic. Microorganisms 8, 1678 (2020).

Rogers, R.L., Cridland, J.M., Shao, L., Hu, T.T., Andolfatto, P., and Thornton, K.R. (2014). Landscape of Standing Variation for Tandem Duplications in Drosophila yakuba and Drosophila simulans. Mol Biol Evol 31, 1750–1766.

Roseman, R.R., Pirrotta, V., and Geyer, P.K. (1993). The su(Hw) protein insulates expression of the Drosophila melanogaster white gene from chromosomal position-effects. EMBO J. 12, 435–442.

Said, I., Byrne, A., Serrano, V., Cardeno, C., Vollmers, C., and Corbett-Detig, R. (2018). Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. PNAS 115, 5492–5497.

Salemi M, Lamers SL, Yu S, de Oliveira T, Fitch WM, McGrath MS. 2005. Phylodynamic Analysis of Human Immunodeficiency Virus Type 1 in Distinct Brain Compartments Provides a Model for the Neuropathogenesis of AIDS. J Virol 79:11343–11352.

Sanderson, T. Taxonium, a web-based tool for exploring large phylogenetic trees. eLife 11, e82392 (2022).

Sankoff, D. Minimal Mutation Trees of Sequences. SIAM J. Appl. Math. 28, 35–42 (1975).

Sayers, E. W. et al. GenBank. Nucleic Acids Research 49, D92–D96 (2021).

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell 148, 458–472.

Sharakhov, I.V., White, B.J., Sharakhova, M.V., Kayondo, J., Lobo, N.F., Santolamazza, F., Torre, A. della, Simard, F., Collins, F.H., and Besansky, N.J. (2006). Breakpoint structure reveals
the unique origin of an interspecific chromosomal inversion (2La) in the Anopheles gambiae complex. PNAS 103, 6258–6262.

Shatskikh, A.S., Olenkina, O.M., Solodovnikov, A.A., and Lavrov, S.A. (2018). Regulated Gene Expression as a Tool for Analysis of Heterochromatin Position Effect in Drosophila. Biochemistry Moscow 83, 542–551.

Shchur, V. et al. VGsim: scalable viral genealogy simulator for global pandemic. medRxiv 2021.04.21.21255891 (2021) doi:10.1101/2021.04.21.21255891.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 22:30494.

Sigrist, C.J.A., and Pirrotta, V. (1997). Chromatin Insulator Elements Block the Silencing of a Target Gene by the Drosophila Polycomb Response Element (PRE) but Allow trans Interactions
Between PREs on Different Chromosomes. Genetics 147, 209–221.

Simmonds, P. et al. ICTV Virus Taxonomy Profile: Flaviviridae. J Gen Virol 98, 2–3 (2017).

Simões, P., and Pascual, M. (2018). Patterns of geographic variation of thermal adapted candidate genes in Drosophila subobscura sex chromosome arrangements. BMC Evolutionary Biology 18, 60.

Sturtevant, A.H. (1917). Genetic Factors Affecting the Strength of Linkage in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 3, 555.

Sturtevant, A.H., and Beadle, G.W. (1936). The Relations of Inversions in the X Chromosome of
Drosophila Melanogaster to Crossing over and Disjunction. Genetics 21, 554.

Tiwari, A., So, M. K. P., Chong, A. C. Y., Chan, J. N. L. & Chu, A. M. Y. Pandemic risk of COVID-19 outbreak in the United States: An analysis of network connectedness with air travel data. International Journal of Infectious Diseases 103, 97–101 (2021).

Tonzetich, J., Lyttle, T.W., and Carson, H.L. (1988). Induced and natural break sites in the chromosomes of Hawaiian Drosophila. PNAS 85, 1717–1721.

Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, Hinrichs AS, Fernandes JD, Borges R, Slodkowicz G, et al. 2020. Stability of SARS-CoV-2 phylogenies.Barsh GS, editor. PLoS Genet 16:e1009175.

Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. 2021. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nature Genetics:1–8.

Turakhia, Y. et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature 609, 994–997 (2022).

Turakhia, Y. et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet 53, 809–816 (2021).

van Dorp, L. et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun 11, 5986 (2020).

Vogel, M.J., Pagie, L., Talhout, W., Nieuwland, M., Kerkhoven, R.M., and van Steensel, B. (2009). High-resolution mapping of heterochromatin redistribution in a Drosophila position-effect variegation model. Epigenetics & Chromatin 2, 1.

Vöhringer, H. S. et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. Nature 600, 506–511 (2021).

Volz, E. et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell 184, 64-75.e11 (2021).

Wang TH, Donaldson YK, Brettle RP, Bell JE, Simmonds P. 2001. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. J Virol 75:11686–11699.

Yang, J., and Corces, V.G. (2012). Insulators, long-range interactions, and genome function. Current Opinion in Genetics & Development 22, 86–92.

Ye, C. et al. Pandemic-scale phylogenetics. 2021.12.03.470766 https://www.biorxiv.org/content/10.1101/2021.12.03.470766v1 (2021) doi:10.1101/2021.12.03.470766.