Towards Pandemic-Scale Ancestral Recombination Graphs of SARS-CoV-2

Shing H. Zhan[1], Anastasia Ignatieva[2,3⋆], Yan Wong[1⋆], Katherine Eaton[4], Benjamin Jeffery[1], Duncan S. Palmer[1], Carmen Lia Murall[4], Sarah P. Otto[5], and Jerome Kelleher[1†]

June 8, 2023

Abstract

Recombination is an ongoing and increasingly important feature of circulating lineages of SARS-CoV-2, challenging how we represent the evolutionary history of this virus and giving rise to new variants of potential public health concern by combining transmission and immune evasion properties of different lineages. Detection of new recombinant strains is challenging, with most methods looking for breaks between sets of mutations that characterise distinct lineages. In addition, many basic approaches fundamental to the study of viral evolution assume that recombination is negligible, in that a single phylogenetic tree can represent the genetic ancestry of the circulating strains. Here we present an initial version of `sc2ts`, a method to automatically detect recombinants in real time and to cohesively integrate them into a genealogy in the form of an ancestral recombination graph (ARG), which jointly records mutation, recombination and genetic inheritance. We infer two ARGs under different sampling strategies, and study their properties. One contains 1.27 million sequences sampled up to June 30, 2021, and the second is more sparsely sampled, consisting of 657K sequences sampled up to June 30, 2022. We find that both ARGs are highly consistent with known features of SARS-CoV-2 evolution, recovering the basic backbone phylogeny, mutational spectra, and recapitulating details on the majority of known recombinant lineages. Using the well-established and feature-rich `tskit` library, the ARGs can also be stored concisely and processed efficiently using standard Python tools. For example, the ARG for 1.27 million sequences—encoding the inferred reticulate ancestry, genetic variation, and extensive metadata—requires 58MB of storage, and loads in less than a second. The ability to fully integrate the effects of recombination into downstream analyses, to quickly and automatically detect new recombinants, and to utilise an efficient and convenient platform for computation based on well-engineered technologies makes `sc2ts` a promising approach.

1 Introduction

Recombination via template switching is a common feature of the evolution of coronaviruses (Graham and Baric, 2010; De Klerk et al., 2022), including SARS-CoV-2 (VanInsberghe et al., 2021; Jackson et al., 2021; Ignatieva et al., 2022). By bringing together mutations carried by different lineages, recombination plays an important role in generating genetic diversity, with recombinant lineages associated with adaptation to new host species and with the production of more immune evasive variants (Graham and Baric, 2010; De Klerk et al., 2022; Tamura et al., 2023). Early in the COVID-19 pandemic, the levels of genetic diversity were too low to enable the detection of distinctive recombinant strains. By late 2020, however, the appearance and spread of variants of concern (VoC), designated into classes such as Alpha and Delta which harboured multiple characteristic mutations, created the conditions required to detect recombinant strains and their onward transmission (Jackson et al., 2021). More recently, the high prevalence of Omicron, with multiple co-circulating deeply divergent lineages (BA.1 to BA.5), has accelerated the rate of coinfection and the potential for recombination (Bal et al., 2022). In

[1] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, United Kingdom
[2] Department of Statistics, University of Oxford, United Kingdom
[3] School of Mathematics and Statistics, University of Glasgow, United Kingdom
[4] National Microbiology Laboratory, Public Health Agency of Canada, Canada
[5] Department of Zoology and Biodiversity Research Centre, University of British Columbia, Canada
[*] Joint second author
[†] Correspondence: jerome.kelleher@bdi.ox.ac.uk

early 2023, multiple recombinant lineages have successfully established and spread to high frequency, and accounting for recombinant ancestry is now essential in understanding the ongoing evolution of SARS-CoV-2.

Detecting recombination in SARS-CoV-2 is difficult and identifying new recombinant strains is a time-consuming, manual process (Smith et al., 2023). Most genomic surveys for SARS-CoV-2 recombinants search for mosaic genomes that combine specific subsets of characteristic mutations from different lineages (e.g., VanInsberghe et al., 2021; Jackson et al., 2021; Wertheim et al., 2022; Sekizuka et al., 2022) and as a result can only identify inter-lineage recombination events. Turakhia et al. (2022) presented a phylogeny-based approach ("RIPPLES") to identify putative recombinants among over ten million SARS-CoV-2 genomes, without pre-specifying sets of characteristic mutations. RIPPLES finds candidate recombinants by using an existing phylogeny (built assuming no recombination) and finding potential recombinants by scanning for branches containing many mutations. It then determines if these candidates would be better explained by recombination by exhaustively breaking each sequence into segments and attempting to find more parsimonious placements for each segment on the phylogeny. If such placements are found, the sequence is identified as a putative recombinant. Although it enables rapid searches for genomic evidence of recombinants, RIPPLES relies on a SARS-CoV-2 phylogeny that accounts for only mutations, treats recombinants *post hoc*, and is an incomplete representation of the reticulate evolutionary history of SARS-CoV-2. As noted by the authors, a *post hoc* treatment of recombination is possible when recombinant lineages are rare and leave few descendants. However, the proliferation of recombinant lineages is making this increasingly untenable; for example, more than half of the sequences sampled in February 2023 are from the recombinant strain XBB and its descendants (Chen et al., 2022). This also means that future evolution of SARS-CoV-2 is likely to involve multiple sequential recombination events on top of existing recombinant lineages, creating a highly reticulated genealogy.

It is well known that recombination distorts phylogenies (Schierup and Hein, 2000) and affects the results of downstream analyses, such as inference of selection (Anisimova et al., 2003). Standard phylogenetic methods do not account for recombination (e.g., Ronquist et al., 2012; Minh et al., 2020; Guindon and Gascuel, 2003), and there is no standard method for incorporating the effects of recombination into phylogenetic analyses. Ancestral Recombination Graphs (ARGs) are a means of describing such network-like ancestry (Griffiths, 1981; Gusfield, 2014), but until recently lacked software support and sufficiently scalable inference methods to be of practical use. However, approaches to infer ARGs now exist that can scale to tens of thousands of human genomes and beyond (Speidel et al., 2019; Kelleher et al., 2019; Schaefer et al., 2021; Zhang et al., 2023), dealing with levels of recombination far in excess of those seen in viral phylogenies. The "succinct tree sequence" is an ARG data structure which has led to significant computational advances across a range of applications (Kelleher et al., 2016, 2018, 2019; Ralph et al., 2020; Wohns et al., 2022), and the supporting `tskit` software library is now widely used in population genetics applications. The methods in `tskit` have been developed to support millions of whole human genomes (Kelleher et al., 2019), and so it is particularly well suited to representing large SARS-CoV-2 genealogies, which currently encompasses over 15 million sequences in the GISAID database (Shu and McCauley, 2017). See Section 4.1 for more details on ARGs and the succinct tree sequence data structure.

Here we present a preliminary version of `sc2ts`, a novel method for inferring ARGs for SARS-CoV-2 at pandemic scale, in real time. Building on the open-source `tskit` library, the method explicitly reconstructs genealogies with both mutation and recombination, which can be conveniently and efficiently analysed using standard Python data science tools. As illustrated in Figure 1, inference is based on incrementally adding batches of sequences based on their collection dates and proceeds in three phases. First, possible paths connecting each sample to the current ARG are inferred (allowing for recombination) using the Li and Stephens (LS) model (Figure 1A, B); the LS "copying process" is a Hidden Markov Model (HMM) approximating the effects of mutation and recombination, widely used in large-scale genomics (Section 4.2). Then, since many samples in a batch can share an attachment path, we infer phylogenetic trees for each of these clusters separately using standard methods (Figure 1C; Section 4.3). Finally, we attach the trees for these sample clusters to the current ARG and apply some parsimony-based heuristics to address issues introduced by the inherent greediness of this strategy (Figure 1D, E; Section 4.4). Using the current preliminary version of `sc2ts`, we infer two large ARGs (with 1,265,685 and 657,239 samples, respectively) and study the properties of these ARGs to illustrate the power of the method and to inform subsequent development. We find that

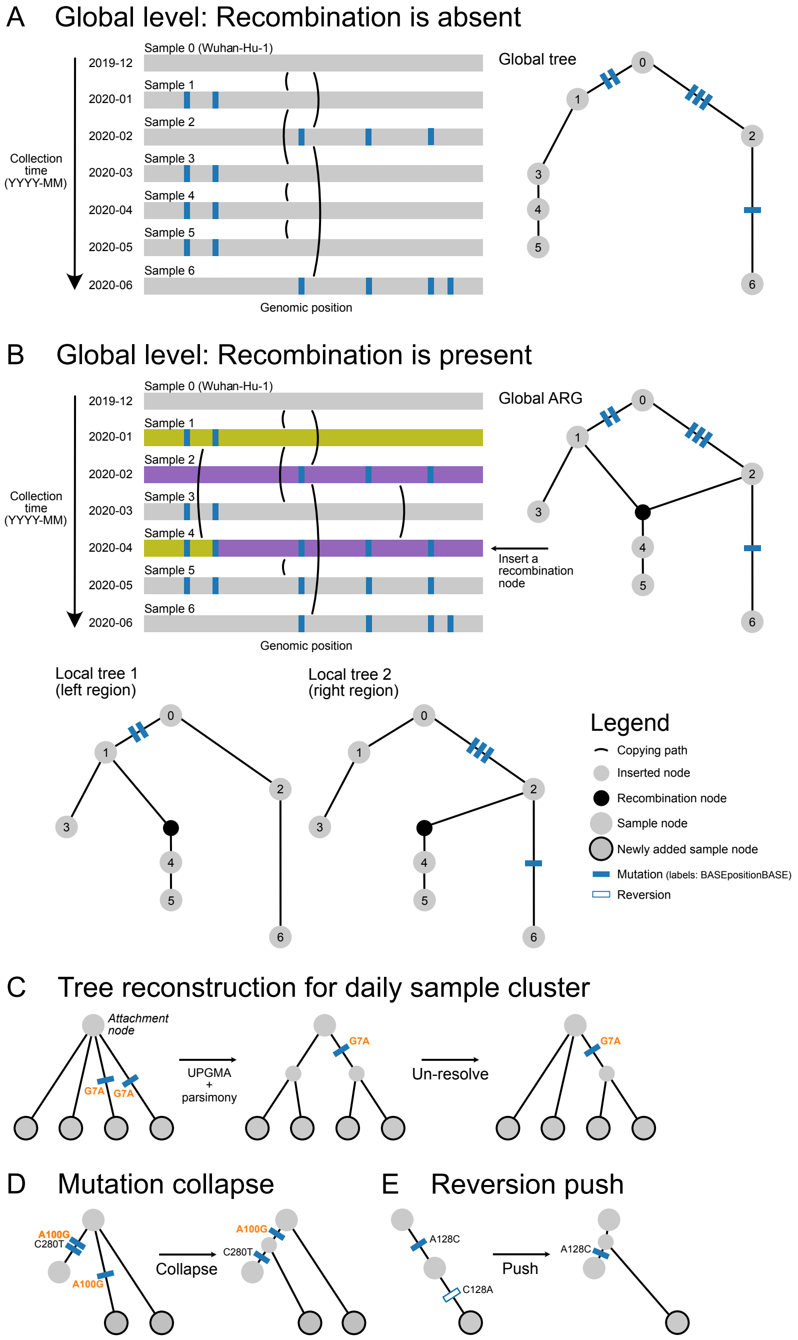


Figure 1: A schematic of the sc2ts method. The genetic relationships among SARS-CoV-2 genomes is reconstructed by using the Li and Stephens model to infer attachment paths for samples to an existing ARG (curved lines). Each daily iteration involves three stages: attachment of new samples to the growing ARG (A, B); reconstruction of trees relating the samples under each attachment node (C); and parsimony-based tree topology adjustments (D, E). In the absence of recombination, sc2ts infers an ARG that is a single tree relating the samples (A). When recombination is detected, sc2ts infers an ARG that concisely encodes a sequence of local trees relating segments of the sample genomes (B). Additionally, mutation-collapsing nodes (D) and reversion-push nodes (E) are inserted to make more parsimonious placements of mutations that should be shared or should not be immediately reverted, respectively.

		Wide ARG		Long ARG	
Sample filtering		collection ≤ 2021-06-30 max-delay=30		collection ≤ 2022-06-30 max-delay=30 max-daily=1000	
Nodes		1,453,347		783,231	
	Node type				
	Sample	1,265,685	(87.09%)	657,239	(83.91%)
	Daily sample cluster tree	102,709	(7.07%)	51,807	(6.61%)
	Reversion push	40,538	(2.79%)	34,358	(4.39%)
	Mutation collapse	40,292	(2.77%)	37,749	(4.82%)
	Recombination	4,123	(0.28%)	2,078	(0.27%)
Mutations		1,231,193		1,062,072	
	Per node per genome	0.83	±1.40	1.36	±1.72
	Per sample per genome	0.77	±1.39	1.38	±1.77
	Per site per ARG	41.23	±108.16	36.10	±80.03
Compressed size (inc metadata)		58MB		37MB	
Bytes/sample (exc metadata)		8.29		10.83	
Load time		0.9s		0.5s	

Table 1: Summary of the inferred ARGs. Nodes are classified as either samples or by the inference process that produced them (see Sections 4.3, 4.4 and 4.5 for details). The mean and standard deviation (±) are reported for the number of mutations per node, sample and site.

these ARGs accurately capture known phylogenetic relationships (Section 2.2) and mutational spectra (Section 2.3), and automatically identify the majority of known recombinant lineages (Sections 2.4 and 2.7) with a high level of precision in the genomic location of recombination breakpoints (Section 2.5) and relationship between parental sequences (Section 2.6). We hope that these benefits of accurate joint estimation of genetic inheritance with mutation and recombination will generate community interest and development of the `sc2ts` method, and more generally in applying the efficient and mature software of the `tskit` ecosystem to pandemic-scale SARS-CoV-2 data.

2 Results

2.1 Inferred ARGs

The goals of this preliminary study are to illustrate the utility of `sc2ts` and to investigate the properties of the inferred ARGs to inform subsequent development. We work with a representative subset of the available data, limited to inferences that can be performed on a single server in a few weeks (see below for further details on timings and computer hardware used). The cut-off dates for sampling are arbitrary. We inferred two ARGs, which we refer to as the "Wide" and "Long" ARGs throughout. The Wide ARG is densely sampled but time-limited and includes 1.27 million sequences collected up to June 30, 2021 which pass some quality-control filters (Section 4.7) and have a maximum delay between sampling and submission dates of 30 days (Section 4.8). For the Long ARG, we randomly sub-sample a maximum of 1,000 genomes per day (again restricting the delay between sampling and submission to 30 days) and include an additional year's worth of samples (to June 30, 2022).

The properties of the inferred ARGs are summarised in Table 1. The majority of the nodes in the ARGs represent sample genomes (Wide ARG: 87%, Long ARG: 84%), with the remainder mostly representing the ancestral sequences inferred from daily sample clusters (Section 4.3) and parsimony heuristics (Section 4.4). Both ARGs contain 29,422 sites (Section 4.7), and a large number of mutations. The average number of mutations per site is high, although some of this may be explained by outlier sites with artefactually high mutation counts (see Figure 4). Despite this, however, the number of mutations per node is small. In the Wide ARG, for example, we have a mean of 0.77 mutations per sampled genome, demonstrating that most added samples fit into the ARG parsimoniously. See

Section 2.3 for more analysis of the patterns of inferred mutations. Recombination plays a relatively minor role, with $< 0.3\%$ of nodes in the ARGs representing inferred recombination events. Of these recombination nodes, the majority are ancestral to only one sample (Wide ARG: 63.1%, Long ARG: 63.3%). We analyse signals of recombination in Sections 2.4, 2.5, 2.6, and 2.7.

Table 1 also summarises some of the computational properties of the inferred ARGs. The ARGs are encoded as a "succinct tree sequence" using the `tskit` library, which provides an extensive suite of operations for constructing and analysing ARGs (Section 4.1). For example, the Wide ARG which contains complete genomes (with imputed missing data) for around 1.2 million samples, along with extensive sample and debugging metadata, requires only 58MB of space (compressed using the `tszip` utility). The majority of this space is used by the metadata, which when discarded results in an encoding that requires an average of only 8.29 bytes per SARS-CoV-2 genome stored. Loading these ARGs takes less than a second, and they can be interactively analysed using Jupyter notebooks (Kluyver et al., 2016) on a standard laptop. The majority of the analyses in this preprint can be carried out in seconds with the `tskit` Python API, using a few gigabytes of RAM.

Inferring these ARGs does require substantial computation. The Wide ARG required 17 days to infer on a server with 128 threads and 512 GB RAM (2x AMD EPYC 7502 @ 2.5GHz). The Long ARG required 23 days on a (much older) machine with 40 threads and 256 GB RAM (2x Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz). The majority of the time is spent on running the LS HMM (Section 4.2) to find the copying path for each sequence, and the process is therefore highly amenable to distributing across multiple machines. We therefore anticipate that further development of the inference methods and scaling out across multiple servers will enable inferences at the full pandemic scale.

2.2 Backbone phylogeny

Compared to organisms like humans that recombine in every generation, recombination is relatively rare in SARS-CoV-2, with recombination nodes accounting for $<0.3\%$ of the inferred ancestry (Table 1). As a result, relationships between strains can often be represented by a single phylogenetic tree, particularly when looking at a subset of strains. We expect ARGs to be particularly treelike early in the pandemic, when co-infection was less likely and divergence between lineages relatively low.

A classic tree-based summary of SARS-CoV-2 ancestry is provided by the Nextstrain project (Hadfield et al., 2018). The trees available from Nextstrain are based on small subsamples of the dataset, and early in the pandemic tend to be restricted to a sample of strains that were not thought to be recombinants. For validation purposes, we compare our ARGs with a downloaded Nextstrain tree, restricted to the time period covered by each ARG. To enable this, we "simplify" (Kelleher et al., 2018) the `sc2ts` ARGs to a backbone containing only those samples present in the Nextstrain tree. This results in a small set of shared samples (Wide ARG: 180, Long ARG: 88), none of which are assigned to Pango recombinant lineages by Nextclade (see Section 2.7).

The `sc2ts` backbone phylogenies for these Nextstrain subsamples contain small amounts of recombination, with 7 recombination nodes in the Wide ARG backbone (8 for the Long ARG backbone). However, the recombination events involve minor, local topological rearrangements where recombination only occurs between close relatives. (The majority of these recombinations are likely false-positives, as discussed in Section 2.8.)

Figure 2 compares the backbone phylogeny of the Wide ARG with a GISAID global "all-time" tree from Nextstrain. We illustrate the backbone phylogeny by visualising a single tree in the middle of the viral genome, although other regions of the genome show almost identical topologies. It is clear that the backbone topology of the `sc2ts` tree shows very close agreement with the Nextstrain tree. The sample genomes cluster by their assigned Pango lineage status, and many variants and their descendants form identical monophyletic clades in both the trees (e.g., the Alpha and Delta VoC clades, labelled). Figure S1 shows the same comparison for the Long ARG, with similar results.

Figure 2 also reveals some notable differences between the trees. Firstly, the `sc2ts` tree is generally less well resolved, particularly in early 2020 when sampling density was much lower than later in the pandemic. Resolution early in the pandemic could be improved by using a tree inferred using classical phylogenetic approaches for the first few months of the pandemic, before the scale of data began to overwhelm these methods. Indeed, this is the approach taken by UShER (Turakhia et al., 2021). Using a pre-existing tree for the early stages of the pandemic would be straightforward in `sc2ts`, and the main reason we did not do this for the initial version under consideration here was to evaluate the

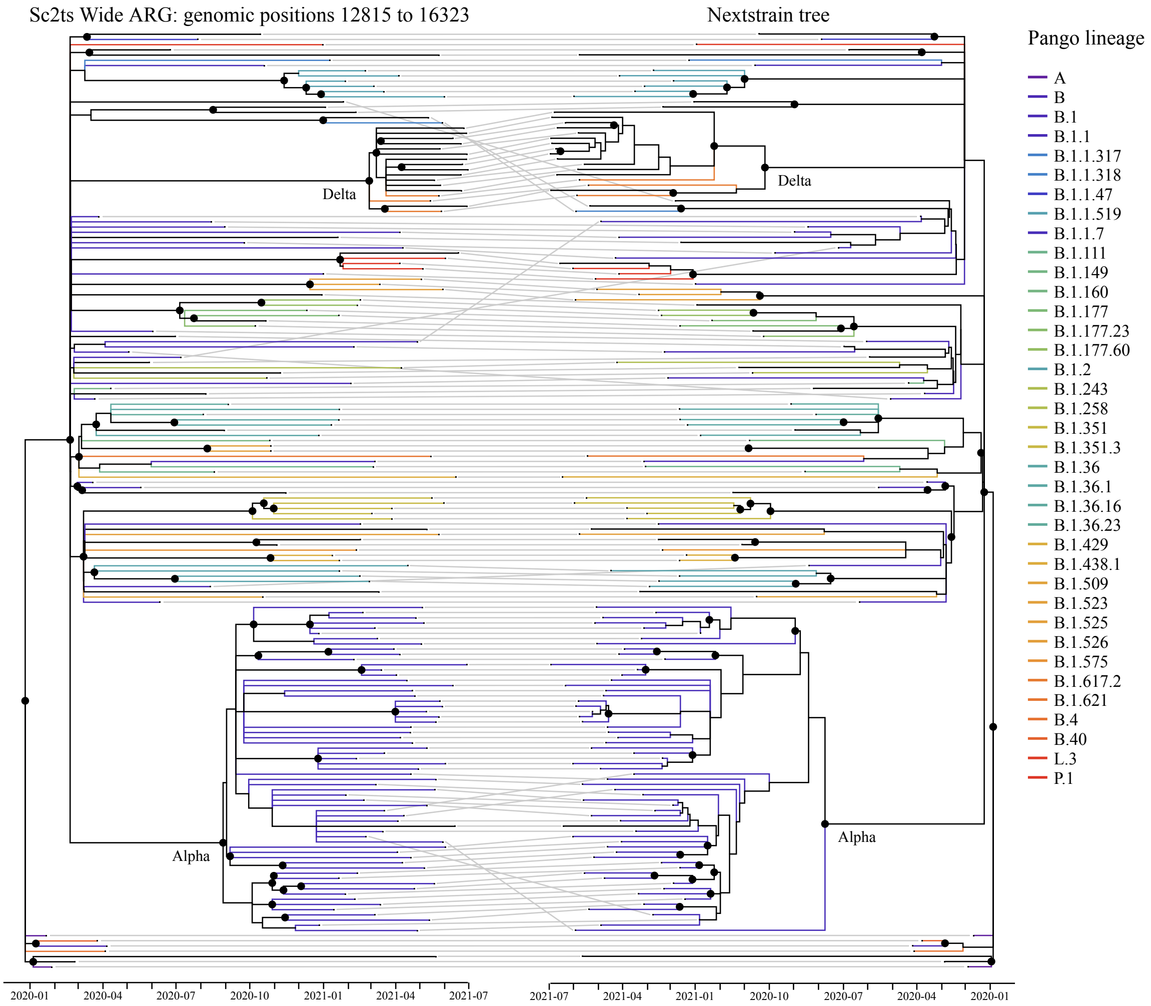
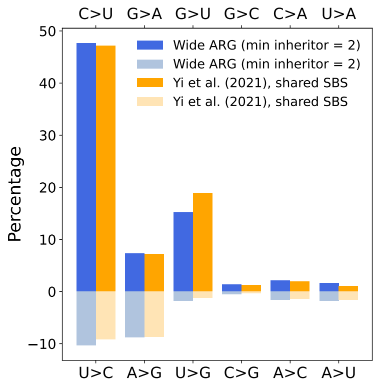


Figure 2: Tanglegram comparing a local tree from the Wide ARG (sampled to mid-2021) and an “all-time” global Nextstrain tree (downloaded on 2023-01-21). Phylogenies are pruned down to those samples present as tips in both datasets, with the horizontal axis representing time (tips end at the sample collection date). Light grey lines match the corresponding samples between the two trees; black circles indicate identical sample partitions between the two trees. Terminal branches are colour-coded according to the Pango lineage status assigned to the tip samples. The tanglegram was generated using the Neighbor-Net algorithm (Scornavacca et al., 2011) implemented in Dendroscope version 3.8.5 (Huson and Scornavacca, 2012). See Figure S1 for the equivalent cophylogeny for the Long ARG.



(A)

	Wide ARG		Long ARG	
Total	1,213,193		1,062,072	
Private	758,903	(62.55%)	767,111	(72.23%)
Transitions	873,487	(72.00%)	783,773	(73.80%)
Transversions	326,053	(26.88%)	270,333	(25.45%)
Insertions	6,191	(0.51%)	2,814	(0.26%)
Deletions	7,462	(0.62%)	5,152	(0.49%)
Recurrent	74,719	(6.16%)	50,099	(4.72%)
Of which reversions	72,617	(5.99%)	48,226	(4.54%)

(B)

Figure 3: (A) Mutational spectrum in the Wide ARG compared to Yi et al. (2021). Mutations are categorised by type (i.e., inherited state > derived state). The percentages of each mutation type from the Wide ARG are represented by blue bars and the percentages from Yi et al. by orange bars, with the darker colours representing one direction (e.g., C>U) and the lighter colours the reverse (e.g., U>C). (B) Summary of mutations in the Long and Wide ARGs. Private mutations occur on terminal branches. Insertions are mutations in which the inherited state is the gap state "-" and the derived state is a nucleotide, and vice versa for deletions. See Section 4.1 for precise definitions of mutations, and the recurrent and reversion classifications.

algorithm's performance with sparse data. Given the simplicity of the algorithm, tree inference for the early pandemic is surprisingly good. The second difference between the sc2ts and Nextstrain trees that we would like to highlight are a few non-identical sample partitions near the tips (e.g. criss-crossing assignments within the Alpha clade). It is unclear what particular differences between the phylogenetic reconstruction algorithms are driving these differences, and more study is required to characterise and address them. Finally, some branch lengths differ substantially between the sc2ts and Nextstrain trees. As discussed in Section 4.6, the dating of nodes other than samples in sc2ts is currently quite crude, but there are likely straightforward expedients that would yield substantial improvements.

2.3 Mutational spectrum

The ARGs inferred by sc2ts and represented using the tskit library (Section 4.1) are a joint estimate of the genealogy with recombination and mutation. Unlike most approaches to phylogenetic analysis, mutations are included in the tskit data model alongside the topological representation of genetic inheritance. This has many advantages, for example allowing us to compute statistics of the observed sequences efficiently (Kelleher et al., 2016; Ralph et al., 2020) and to provide high levels of data compression (Kelleher et al., 2019). The same idea has recently been used to represent SARS-CoV-2 data in UShER's "mutation annotated tree" format (Turakhia et al., 2021).

The properties of the mutations inferred in the Wide and Long ARGs are summarised in Figure 3B. In both cases we have a large number of mutations, and a majority of these (Wide ARG: 62.55%, Long ARG: 72.23%) are private to a single sample, i.e. on terminal branches. Although the average number of mutations per sample is small (Wide ARG: 0.77, Long ARG: 1.38; Table 1), the average number per site is large in both ARGs (Wide ARG: 41.23; Long ARG: 36.10; Table 1). However, a lower median count (Wide ARG: 14; Long ARG: 13), suggests that these high mutation counts are partly driven by some hypermutable sites (e.g., site 28,271 has over 7,000 mutations in the Wide ARG; Figure 4), which may be artefactual.

The current version of sc2ts infers a large number of "reversion" mutations, with 5.99% of all mutations in the Wide ARG (Long ARG: 4.54%) reverting the state change of the immediately ancestral mutation (see Section 4.1 for precise definitions). These are symptomatic of both data quality issues such as "time travellers" (Section 4.8) and problematic sites (e.g., 28,271 as discussed in Section 2.5; see also Figure S7 for an example of multiple reversions at this site), as well as indicating opportunities for improvement in tree building heuristics (Section 4.4). For example, the current "reversion push"

Group	Sequences	Method	Interval	Parent lineages
A (XA)	4	Jackson	21,256–21,615	B.1.177+B.1.1.7
		`sc2ts`	21,256–22,228	B.1.177.18+B.1.1.7
B	2	Jackson	6,529–6,955	B.1.36+B.1.1.7
		`sc2ts`	6,529–6,955	B.1.36+B.1.1.7
C	3	Jackson	25,997–27,443	B.1.1.7+B.1.221
		`sc2ts`	25,997–27,973	B.1.1.7+B.1.221
D	3	Jackson	21,576–23,064	B.1.36.17+B.1.1.7
		`sc2ts`	22,445–23,064	B.1.36.39+B.1.1.7

Table 2: Comparison of recombination breakpoint intervals and parent lineages for Groups A-D reported by Jackson et al. (2021) with the corresponding recombination events in the Wide ARG. The second column gives the number of sequences in the group, limited to the samples considered by Jackson et al. (2021). The 3SEQ (Boni et al., 2007) coordinates reported by Jackson et al. have been altered as follows: we add one to both left and right coordinates to correspond to the `tskit` definition of inheritance on either side of a breakpoint, and add one to the right coordinate to make the intervals right-exclusive. See Section 4.5 for a precise definition of sequence inheritance at recombination events and the corresponding breakpoint intervals. Details for all 16 sequences are given in Table S1.

operation only eliminates reversions in the newly added portion of the tree and not reversions around earlier nodes created by this algorithm.

Despite the presence of some artefactual mutations, the properties of the mutations inferred by `sc2ts` largely follow established results. In Figure 3A we compare the mutational spectrum in the Wide ARG to the results of Yi et al. (2021), who reconstructed a SARS-CoV-2 phylogeny of over 350,000 genomes sampled globally from 2019-12-24 to 2021-01-12 and classified the mutations occurring along the phylogeny. We categorised all single nucleotide mutations in the Wide ARG by type (defined by the inherited and derived states), excluding mutations inherited by only a single sample (which are more likely to be sequencing errors). Similarly, we took the data for single nucleotide mutations from Yi et al. (2021, `https://github.com/ju-lab/SC2_evol_signature`), excluding mutations occurring along terminal branches, and tallied them up by type. Figure 3A shows that the mutational spectrum from the Wide ARG (based on 448,825 mutations) matches that reported by Yi et al. (2021, based on 92,344 mutations). In both spectra, C-to-U mutations and G-to-U mutations occur more frequently than U-to-C and U-to-G, respectively. Similar results are obtained when including the mutations inherited by only a single sample or those occurring on terminal branches (data not shown).

2.4 Early recombinants

RNA viruses are known to recombine at high rates when cells are co-infected (Simon-Loriere and Holmes, 2011), and recombination has been widely documented to be commonplace in animal and human coronaviruses (Su et al., 2016). While recombination in SARS-CoV-2 was shown early on to be frequent *in-vitro* (Gribble et al., 2021), the relatively slow accumulation of genetic diversity early in the pandemic hampered efforts to detect recombinant strains. A number of early studies relying on analysing patterns of linkage disequilibrium and searching for mosaic genomes carrying characteristic mutations of different lineages either failed to detect recombination or posited that this occurred at low rates (e.g., Nie et al., 2020; Tang et al., 2020; VanInsberghe et al., 2021; Varabyou et al., 2021). The first clear evidence of recombinant lineages was presented by Jackson et al. (2021), who performed a careful analysis of sequences circulating in the UK in late 2020 to early 2021 and found evidence of multiple independent recombination events and onward transmission. By searching for samples combining genomic segments from Alpha (B.1.1.7) and from the parental lineage B.1.1 based on a list of 22 Alpha-defining mutations, they found 16 recombinant sequences from 8 putative origins (groups A to D and four singletons). These findings are closely replicated in both the Wide and Long ARGs.

The Wide ARG contains 15 of these 16 recombinant sequences (sample MILK-103C712 was removed during preprocessing; see Section 4.7). Table 2 shows the groups of sequences identified by Jackson et al. as likely independent recombination events with onward transmission. In each case we have

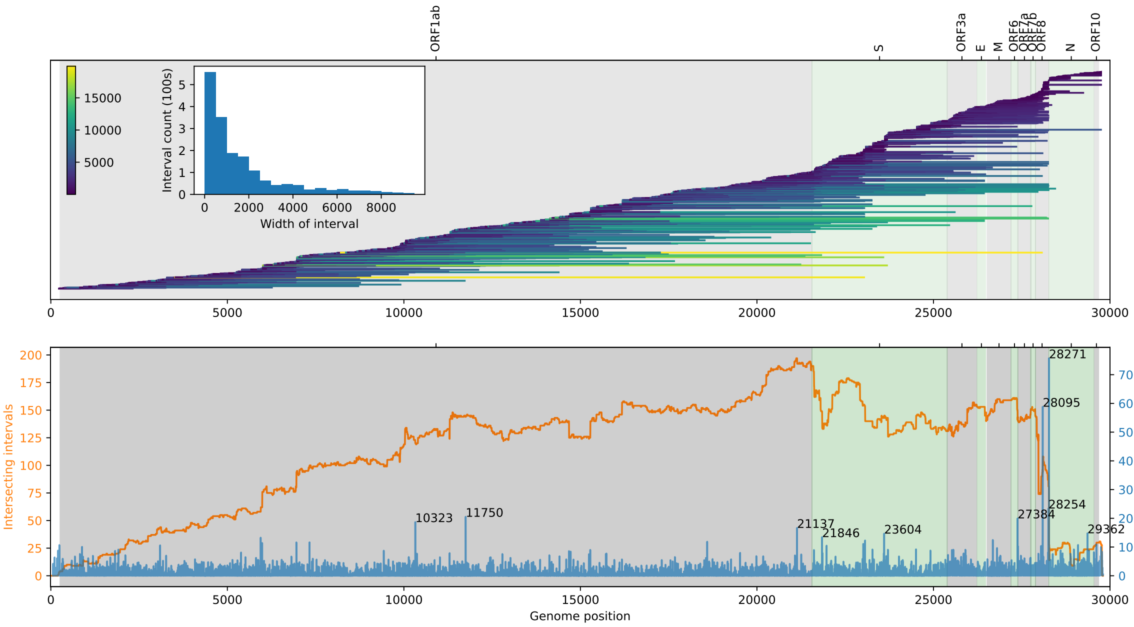


Figure 4: Distribution of recombination breakpoints and mutations along the genome in the Wide ARG. Top panel shows the intervals for 1,769 breakpoints associated with 1,522 recombination nodes with at least two descending samples, plotted along the genome as line segments (coloured by interval width). The inset histogram shows the distribution of these interval widths (truncated at 10kb). The bottom panel shows the number of intervals that span each site along the genome (left axis, orange) and the number of mutations per site (right axis, blue). The top-ten sites by mutation count are annotated. See Figure S2 for the equivalent plot for the Long ARG.

a corresponding recombination node in the Wide ARG, from which all the sequences in the group descend. The parent lineages and breakpoint intervals agree closely (see Section 2.5 for more details on breakpoint intervals). For groups B, C and D, these recombination nodes form clades consisting only of the identified sequences. Group A sequences were subsequently given the Pango XA designation following onward transmission, and there are 44 XA designated samples in the Wide ARG (including the 4 sequences analysed by Jackson et al.). The Group A recombination node forms a monophyletic clade of these 44 samples. Table S1 shows the details for each of the 16 sequences individually and showing generally a strong concordance in mosaic structure and parent lineages (including sample CAMC-CB7AB3, which is inferred to have two breakpoints under both methods).

The Long ARG contains 5 of the sequences: two each from groups A and B and sample QEUH-1067DEF. These cluster under three recombination nodes, as expected, and have identical breakpoint intervals and parental lineage assignments to those of Wide ARG. The recombination nodes for Group B and sample QEUH-1067DEF are ancestral only to the sequences involved. The recombination node for group A forms a monophyletic clade of all 5 XA samples present in the Long ARG (Figure 6A)

2.5 Recombination breakpoint intervals

It is rarely possible to be precise about the position on the genome at which a recombinant sequence switches from inheriting from one parent to another. Even if we observe the recombinant sequence before subsequent divergence occurs (during onward transmission), there is no way to identify the exact breakpoint if the two parent sequences are similar. Here we define the interval within which a particular breakpoint may have occurred as the genome coordinates over which the sequences for the left and right parent nodes are identical. The right-hand extreme of the breakpoint interval is chosen by the LS HMM Viterbi algorithm (Section 4.2), and the left endpoint is then derived by directly comparing the parent sequences. See Section 4.5 for a precise definition of breakpoints and their intervals.

Figure 4 shows the distribution of breakpoint intervals and patterns of recurrent mutation along the genome in the Wide ARG (see Figure S2 for the same information for the Long ARG). We focus on the Wide ARG here as it covers roughly the same time period as the analyses of Turakhia et al. (2022), facilitating comparisons of the results. To reduce the effect of artefactual recombinants, we consider only the breakpoints associated with the 1,522 recombination nodes that are ancestral to more than one sample. The mean length of these intervals is 1,685 bases (median 962), and the length distribution is summarised in the inset histogram in Figure 4.

Although further work is required to filter out spuriously inferred recombination events (see Section 2.8) we can draw some preliminary conclusions from Figure 4. The number of intervals spanning a site (orange curve) is lower at the ends of the genome, as expected due to the lack of information about recombination with few flanking sites. In addition, the number of intervals spanning a site often drops near the beginning of a gene. This is particularly apparent at the ORF8/N gene interface, with the N gene containing fewer potential recombination breakpoints than other genes, in agreement with the results of Turakhia et al. (2022). Noticeable declines in the number of spanning intervals are also seen near the beginning of S, M, ORF7a, and ORF8. These declines are sometimes associated with hypermutated sites (e.g., 27,384 and 28,271 near the beginning of ORF7a and N, respectively), as expected because sites that undergo mutation at high rates are more likely to differ between the parents of a recombinant and so provide information about the location of a breakpoint. This pattern may, however, also be an artefact of sequencing errors causing sites to appear different between the parents when they are not (see the discussion of potential errors at site 28,271 below). In other cases, however, the drop in intervals does not appear to coincide with hypermutability and may reflect shifts in the actual rate of recombination between genes. Indeed, template switching is known to occur at hotspots, which often involve transcription-regulatory sequences preceding genes in SARS-CoV-2 (Yang et al., 2021). The rate of recombination may also depend on the relative abundance of different subgenomic RNA intermediates that span different genes, affecting the availability of templates and influencing the rate of homologous recombination (Kim et al., 2020; Zou et al., 2021).

It is important to note that the precise endpoints of intervals can be somewhat arbitrary, because they are defined by the sequence differences that happen to be present in the recombinant's parents. Thus, breakpoint intervals will tend to be truncated at sites that are hypermutable, either due to increased information about parentage or spurious inferences caused by sequencing errors. It is therefore helpful to compare the number of intersecting intervals with the number of mutations per site (Figure 4). For instance, position 28,271 (located in the ORF8-N intergenic region) has the largest number of mutations and appears as an endpoint of 66 intervals. The 7,572 mutations (including 5,782 insertions and 1,605 deletions) at this site are an indicator that this homopolymeric region may be prone to sequencing errors and potentially should be included in the list of "problematic sites" that are excluded from analysis (Section 4.7). On the other hand, 58 of the breakpoint intervals have endpoints within one base of site 27,972, which has undergone 770 mutation events (including 425 C>T, 314 T>C). The C>T mutation has the effect of truncating ORF8, and it has been posited that the truncation is neutral or advantageous for transmission, and disadvantageous within-host (Jungreis et al., 2021), suggesting that the high rate of recurrent mutation at this site may be due to selection.

2.6 Divergence between recombinant parents

In this section we explore the detailed ancestral relationships between recombinant parents by investigating the patterns of divergence between them. We focus on the Long ARG because it covers time periods where substantial recombination is known to have occurred and contains samples from 33 Pango X lineages (i.e., those inferred to have recombinant ancestry; see Section 2.7 for further analysis). Figure 5 shows the estimated date of the most recent common ancestor (MRCA) of the parent nodes for each recombination breakpoint, plotted against the divergence between these parents (i.e., the total branch length from the parents to their MRCA in the trees to the immediate left and right of the breakpoint). As in Section 2.5, in these plots we exclude breakpoints associated with "singleton" recombination nodes (those ancestral to only one sample). Larger points distinguish those breakpoints which occur in nodes ancestral to 5 or more samples, comprising 316 breakpoints from 291 recombination nodes. The criterion of 5 descendants matches the minimum number required to designate a new Pango lineage (Rambaut et al., 2020). Note that as each plotted point represents a breakpoint, a recombination node with more than one breakpoint (e.g., with 3 or more parents, comprising ∼10% of the recombination nodes in Figure 5) will be represented by several points.

The date of the MRCA of recombinant parents is concentrated in several banded rows in Figure 5.

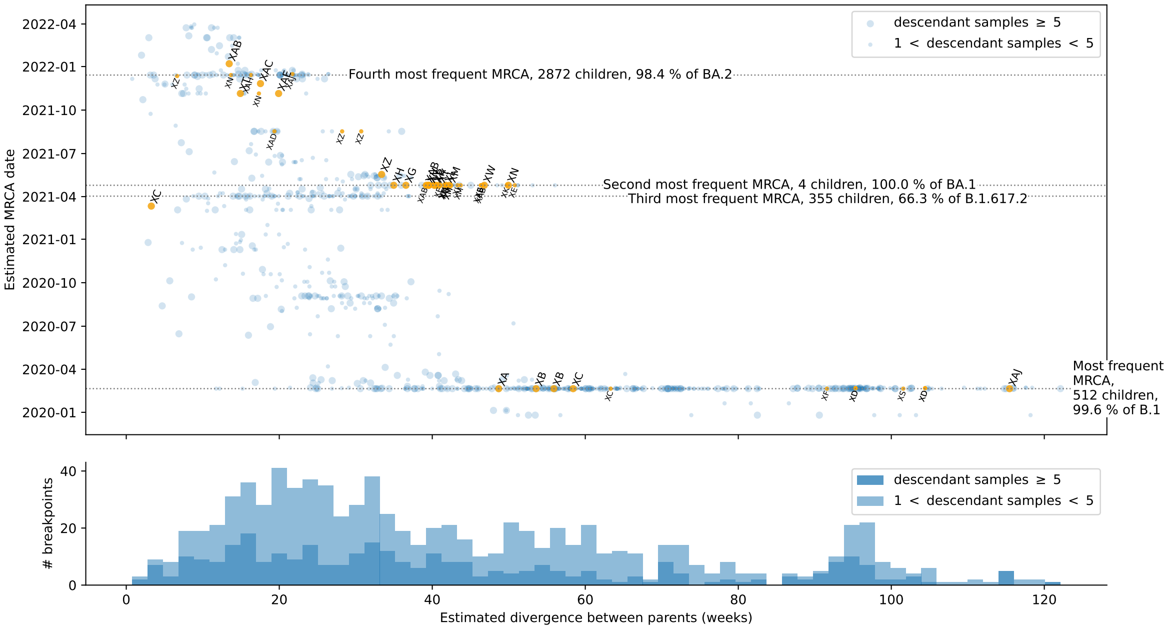


Figure 5: Date of common ancestry between the parents on either side of recombination breakpoints, as a function of the divergence time between the parents. MRCAs of parents associated with Pango designated recombinants (XA, etc) are identified in orange. Larger symbols represent breakpoints in recombination nodes ancestral to five or more samples. Horizontal dotted lines show the four most common MRCA nodes, which tend to be associated with major outbreaks and with many immediate children. The stacked histogram shows the distribution of parental divergence times, ranging from parents that have diverged only a few days ago, to much more divergent parent lineages. See Figure S3 for equivalent plots broken down by parental VoC classification.

These are largely due to a few MRCAs shared by many recombinants (the top four are indicated in the figure). These shared MRCAs lie near the root of large expansions, and the majority are associated with large polytomies, likely indicating a rapid and under-sampled expansion of a clade (e.g. major Delta and Omicron waves).

Figure 5 shows that there is a large range in divergence times among parents of detected recombinants, with a broad spread of times from about 10 to 80 weeks prior to the recombination event and a minor additional peak involving recombination between lineages that diverged ∼95 weeks ago, corresponding to common ancestors which trace back to early 2020. This latter group should contain, for example Delta-Omicron recombinants. More widely, we can further classify the breakpoints by the VoC combinations of their parent lineages. Considering only the Alpha, Delta, and Omicron VoC classes, such a classification reveals that the majority of breakpoints have two Delta or two Omicron parents, and that Omicron and Delta are the variants associated with the most recombination (Figure S3). This may reflect either sampling intensity, the prevalence of cases (which increases the chance of coinfection and recombination), or possible heterogeneity in recombination probabilities among lineages.

The time estimates in these figures should be treated with a degree of caution, because non-sample nodes are crudely dated in the current version of `sc2ts` (see Section 4.6, in particular for discussion of how these dates might be improved using existing methods). Nevertheless, it is clear that `sc2ts` can identify recombination between lineages that are only a few weeks diverged.

2.7 Recombinant Pango lineages

In this section we focus on the detailed ancestry of samples that have been previously identified as recombinants, i.e., designated as belonging to a Pango lineage with a name starting with an "X". We focus primarily on the Long ARG which contains many more recombinants (the status of Pango X lineages in the Wide ARG is briefly summarised in Section 2.7.4). Designation of samples to Pango lineages is not a straightforward task, and there can be significant variation between methods (De Bernardi Schneider et al., 2023). Here, we consider two different assignments of Pango lineages to samples in the Long ARG, those provided by Nextclade and by GISAID. Using the Nextclade assignments, the Long ARG contains 749 samples from 33 Pango X lineages (711 samples from 33 lineages when we remove singleton recombinants). In contrast, using the GISAID designations we have 515 samples from 38 X lineages (511 samples from 35 lineages when we filter singleton recombinants). Of the GISAID designations, 28 are shared with Nextclade (26 after filtering singletons). This variation in classifications highlights the uncertainty that exists when assigning Pango X lineages to samples, and is important to keep in mind when interpreting the results here.

We focus here on the 749 Nextclade-designated recombinant samples. There are two samples designated XP which do not descend from a recombination node, and a likely explanation for the absence of a corresponding recombination event in the Long ARG is that the characteristic multibase deletion for XP (`https://github.com/cov-lineages/pango-designation/issues/481`) is masked during our preprocessing (see Section 4.7 for details and potential improvements). Of the remaining X designated samples, 38 are singleton recombinants, descending from recombination nodes that are ancestral only to that sample (10 are labelled XZ; 6 are XE; 3 each from XN and XK; 2 each from XC, XS, XV, XQ and XAB; and 1 sample from each of XB, XM, XJ, XAF, XAH and XAJ). Such samples are likely to be enriched for sequencing errors and lineage designation artefacts, and so we exclude them from further analysis in this section. A further 79 samples (XN: 53, XZ: 16, XAJ: 6, XE:1, XAD: 1, XAH: 1, XAK: 1) trace back to a most recent recombination node that is likely to be a false positive (row C in Table 3, see Section 2.8). For simplicity these samples are likewise excluded from further analyses.

The remaining 630 samples (31 Pango lineages) trace back to 50 different most recent recombination nodes, summarised in Table S2. These fall into three classes: single origin, multiple origin, and multiple nested origins, which we discuss in the following sections.

2.7.1 Single origin

In the absence of genealogical information, a reasonable initial assumption is that all sequences assigned to a given Pango X lineage are descendants of a single recombinant sequence, arising as a result of a mixed infection followed by onward transmission. We would expect our ARGs to reveal evolutionary histories of this nature, where all the samples assigned to a given recombinant lineage trace back to

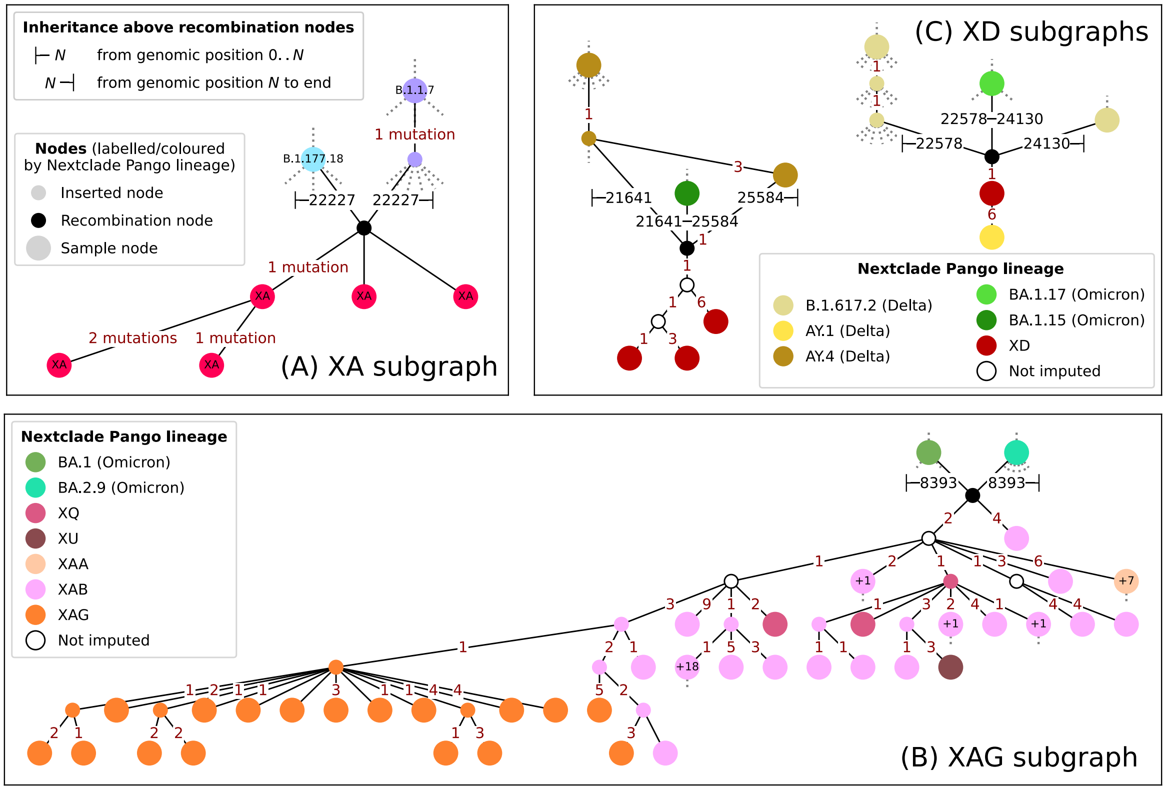


Figure 6: Examples of non-nested Nextclade Pango X lineages. (A) Subgraph for XA in the Long ARG: all five samples designated as XA by Nextclade, together with their ancestral lineages, are shown outwards to the nearest sampled viral genome; dotted lines show ARG continuation. Vertical position of nodes does not show absolute time, but relative rank (parents above children). Nodes are coloured by Nextclade Pango designation; smaller symbols are non-sample nodes inserted by sc2ts, whose Pango status is imputed. Genomic regions inherited by the recombination node are shown; breakpoints correspond to the rightmost breakpoint position inferred by sc2ts. (B) Equivalent subgraph for the 17 XAG samples in the Long ARG, with abbreviated labelling. Where non-XAG samples are ancestral to further unplotted samples, the number of unplotted descendant samples is marked as "+1", "+7", etc. (C) Equivalent subgraphs for both origination events involving the four XD samples in the Long ARG. Details of the mutations and sample node identities for all three plots are provided in supplementary Figures S4, S5, and S6, which also provide alternative GISAID Pango designations.

one recombination node, representing a single originating recombination event. In the Long ARG, 16 of the 31 Pango recombinants lineages identified by Nextclade fall into this category (Table S2)

One of the simplest examples is XA, corresponding to group A of Jackson et al. (2021) as discussed in Section 2.4. Figure 6A shows the exact relationships inferred by sc2ts as a subgraph of the Long ARG. Here, paths have been traced from all Nextclade-identified XA samples (in red) to the closest other sample nodes in the ARG. Sample nodes are plotted as larger circles, but the subgraph also includes intermediate, non-sample nodes (i.e., inserted by sc2ts, see Sections 4.3 and 4.4). Dotted lines show where this subgraph links to the rest of the ARG. Above recombination nodes, only ancestral nodes are shown, meaning that the subgraph is not extended to show additional descendants of recombinant parents.

It is clear from the XA subgraph that all the samples labelled XA by Nextclade trace to a single originating recombination node, whose genome is a composite of a B.1.177.18 lineage on the left of the genome and a B.1.1.7 lineage on the right. In the subgraph we show the rightmost genomic position for the recombination breakpoint, here at position 22,227 (corresponding to a breakpoint strictly less than 22,228, see Table 2)

A more complex single-origin case is XAG, illustrated in Figure 6B. Here, the XAG samples all trace back to the same most recent recombination node (combining BA.1 on the left and BA.2.9 on the right), but we infer this recombination event to also be the originating event for all the recombinant samples designated XAA, and some, but not all, of those identified as XAB, XQ, and XU by Nextclade.

The classification of originating recombination events is dependent on accurate designation of Pango lineages to samples. It is therefore important to note that if the GISAID Pango designations are used, many of the samples marked here as XAB are reclassified as BA.2 and XAG becomes fully monophyletic (although still not an immediate descendant of the originating recombination node, see Supplementary Figure S5). This is an independent confirmation of the uncertainty in designation of these XAG-related samples.

Six of the 16 Nextclade designated lineages (XA, XAC, XAE, XF, XK, and XS) are of the basic (XA) type with no other Pango designations among their descendants. The remaining ten are of the XAG type with multiple Pango designated lineages as additional descendants of the originating recombination event. In some cases, these may, however, be a result of erroneous Pango designations. Table S2 also shows the official Pango designated parent lineages and `sc2ts` inferred parent lineages, which extensively agree (although `sc2ts` provides a more precise parent designation).

2.7.2 Independent multiple origins

The `sc2ts` inference process has no pre-defined knowledge of Pango X lineage assignments, and there is therefore no particular requirement that all the samples assigned to a given lineage must trace back to a single recombination event. Using Nextclade designations, 15 Pango X lineages are inferred to have multiple recombinant origins, such that their samples trace to more than one most recent recombination node in the ARG. Of these, 11 are cases where the recombinants are independent rather than nested (i.e., there is no overlap in the list of descendant samples for each recombination node). Most have a single "main" recombination event from which the majority of the corresponding recombinant samples descend and which agrees with the official Pango designated parent lineages (see Table S2, but note that in XJ and XU there are too few Pango X samples to decide on a "main" clade).

Figure 6C shows a simple multiple-origin example, consisting of the 4 samples labelled XD by Nextclade in the Long ARG. The left hand subgraph (containing three XD samples, all sampled in France) has an earliest sample (strain France/HDF-biopath-7747831001/2022) dated 2022-02-26, while the right hand subgraph has a single XD sample (strain Turkey/HSGM-F12594/2021) dated 2021-12-30. Both involve an Omicron lineage being inserted into the middle of a Delta genome, but the breakpoints in each case are slightly different: the start of the Omicron insertion in the French samples has an estimated rightmost position of 21641bp (and a leftmost of 21619, not shown) with the insertion end occurring at a rightmost position of 25584 (and a leftmost of 25470). By contrast, the Omicron insertion in the Turkish sample is inferred to have occurred from position 21619–22578 to position 23605–24130. The breakpoint difference, the different geographical locations, the time between the samples, and the fact that the two samples differ at 23 nucleotide positions, suggests that these may indeed represent independent Delta–Omicron recombinants. The canonical XD definition is based entirely on samples from northern Europe, particularly France (`https://github.com/cov-lineages/pango-designation/issues/444`) so it seems plausible that the earlier Turkish sample has been mislabelled as XD by Nextclade. Indeed, GISAID does not label any of these samples XD (see Figure S6 which gives exact mutations and sample identifiers). Investigation of other multiple-origin examples reveals somewhat similar patterns, suggesting that most of the simple multiple origin examples are due to incorrect Pango labelling.

2.7.3 Nested recombinant origins

As well as cases where Pango X lineage origins are attributed to independent recombination events, four Pango X lineages in the Long ARG have Nextclade-designated samples whose ancestry involves further recombination events (marked by † in Table S2; the most complex appears to be XAB). Figure 7 plots the earliest example, XB, which is present in both the Wide and Long ARGs. The subgraph shows a recombination between a B.1 sample and B.1.627 sample that leads not only to all the XB-labelled samples but also to a "hairball" of further recombination nodes whose descendants are often not identified as recombinants by Nextclade (plotted on the left, in blue). A similar pattern is seen when examining XB in the Wide ARG (see discussion below).

Note that in the Long ARG, the nested recombination events account for only one XB sample (pink upper left, with 7 mutations above it); moreover, this sample is not identified as XB by GISAID

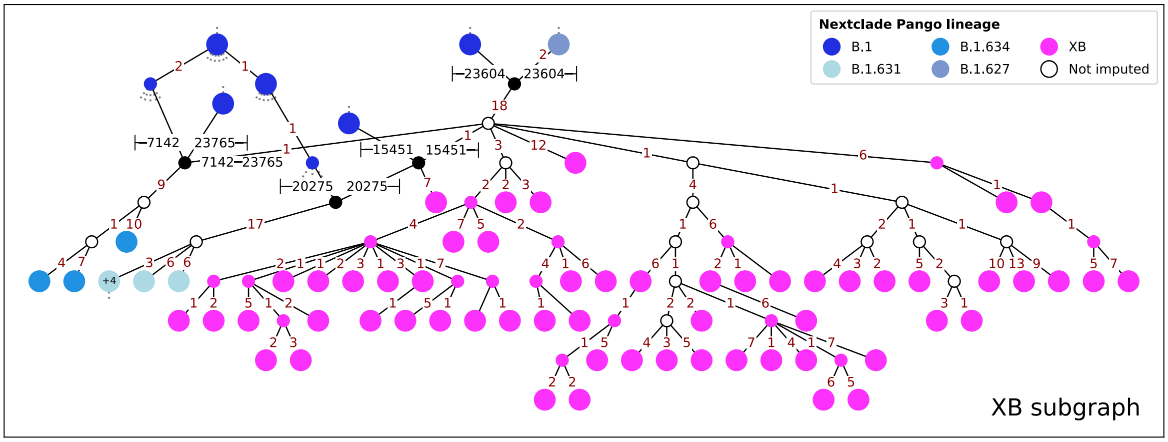


Figure 7: A subgraph of the Long ARG showing nested recombination events involving Pango lineage XB. All XB samples trace to a single recombination node (top centre), but three further recombinations also descend from this node. The samples descending from these nested recombinations include 9 that are assigned by Nextclade to various non-recombinant pre-Alpha lineages (blue).

(Figure S7) indicating some uncertainty in lineage assignment in this part of the ARG. Also note that the number of mutations on the lineages immediately above and below the recombination node (totalling 18+2) is rather large, suggesting that the sampled recombinant which induced the recombination node in the Long ARG is only distantly related to the true originating recombinant. This could account for complex and potentially artefactual relationships around these nodes and is likely to be due to undersampling of the XB outbreak. Investigating examples of nested recombinant origins, and identifying which (if any) of the nested recombination events may be artefactual, is an important area of future research.

2.7.4 Wide ARG

Because the Wide ARG is restricted to data collected prior to mid-2021, it contains samples from only three Pango-designated recombinant lineages: XA, XB, and XC. Both Nextclade and GISAID designate 44 samples as XA, while 237 samples are designated as XB by Nextclade (231 by GISAID), and 6 as XC by Nextclade (none by GISAID). After removing singleton recombinants, XA numbers remain unchanged, but XB is reduced to 235 Nextclade-designated samples (229 GISAID) and XC is reduced to 4 (none in the GISAID designations). We confirm that all samples designated as XA, XB, or XC by any method have one or more recombination nodes in their ancestry.

As in the Long ARG, all XA samples in the Wide ARG trace back to a unique originating recombination node, consistent with Figure 6A. This is the product of a recombination between a B.1.177.18 sample (specifically the strain Wales/ALDP-115BF41/2021) which contributed the majority of the genome from the start to a rightmost position of 22227, and an unknown (inserted) node with imputed Pango lineage B.1.1.7, which contributed the remaining right hand portion. The recombination node has five immediate children: four sample leaves (strains Wales/ALDP-11CF93B/2021, Wales/ALDP-125C4D7/2021, Wales/LIVE-DFCFFE/2021, and Wales/ALDP-130BB95/2021) and a inserted node which is the ancestor of all other XA samples in the dataset. The geographical clustering inferred by sc2ts for these samples matches the findings of Jackson et al. (2021).

For XB, all samples trace back to an originating node which is the product of a recombination between a B.1 sample (specifically the strain England/CAMB-7B47D/2020, which contributed the majority of the genome from the start up to a rightmost position of 23604bp), and two UPGMA nodes with imputed Pango lineages B.1.627 (up to a rightmost position of 27389bp) and B.1.36.8 (the remaining fragment of the genome). Figure 7 shows that in the Long ARG the equivalent recombination node has only 2 parents, with no involvement of B.1.36.8; it is possible that the third parent in the Wide ARG is artefactual. As in the Long ARG, additional non-X-designated samples such as B.1.634 also descend from this recombination, and there are also a small number of nested recombination nodes. However, all but one of these nested recombinations are unimportant, being ancestral to

	Strain	Descendants	Lineage	Rank in lineage	HMM Cost
A	Germany/HH-RKI-I-061284/2021	178,405	B.1.617.2	3 / 12829	17
B	India/ILSGS00961/2021	177,649	B.1.617.2	8 / 12829	25
C	Denmark/DCGC-281594/2021	127,227	BA.2.9	1 / 10897	32
D	USA/NJ-GBW-EWR000001/2021	127,230	BA.3	1 / 15	25

Table 3: False positive recombination events. Details show are for the top four recombination nodes in the Long ARG ordered by number of descending samples. See the text for details on the remaining columns. Rows are labelled A, B, C and D for ease of reference.

negligible fractions of the Nextclade-designated XB samples. The one exception accounts for about 17% of the designated XB nodes and involves a recombination between descendants of the originating XB recombination. More specifically, strain USA/TX-HMH-MCoV-43092/2021 is inferred to be a recombinant between an XB grandparent and its XB grandchild, involving an intermediate UPGMA node. It seems likely to be an artefactual recombination event caused by undersampling, but it could also reflect true recombination between closely related lineages.

For the few XC-labelled samples, the Wide ARG identifies more than one originating recombination node. However, as none of the samples designated as XC by Nextclade are designated as XC by GISAID, these patterns could be due to mislabelling, and a greater number of XC samples would be needed to draw reasonable conclusions.

2.8 False positive recombinants

The Long and Wide ARGs both contain several recombination nodes that are ancestral to a large number of samples. These inferences of recombination may well be artefacts caused by the appearance of new variants of concern carrying more than the expected number of mutations (Otto et al., 2021). Table 3 shows details of four recombinants in the Long ARG that are likely false positives and have come to be the ancestors of a large number of samples. These are the top-four recombinants from the Long ARG in terms of numbers of descendant samples (the fifth-largest has substantially fewer, at 14,869 descendant samples), labelled A–D. For each row, we show the sample's Pango lineage designation and the temporal rank of that sample out of all samples with that designation. Thus, rows C and D are the earliest samples seen in the Long ARG from the BA.2.9 and BA.3 lineages, and A and B are respectively the third and eighth earliest samples among 12829 samples assigned by Nextclade to the B.1.617.2 lineage (Delta VoC). Also shown is the overall "cost" of the Viterbi solution computed by the LS HMM (Section 4.2), which is 3× number of recombinations + number of additional mutations (for a mismatch ratio of $k = 3$). This column shows that each of these samples was a large "distance" from the current ARG when it was added. The mean HMM cost over all 2,078 recombinants in the Long ARG is 8.39 (median: 7), and the mean cost for the 763 non-singleton recombinants is 8.5 (median: 7). Thus, the recombinants in Table 3 are in the tail of this distribution.

Occasional evolutionary leaps, in which a large number of mutations are acquired in sudden jumps, is a signature feature of SARS-CoV-2 (Corey et al., 2021; Otto et al., 2021; Nielsen et al., 2023). Such "saltations" naturally present challenges to `sc2ts` and the current HMM parameterization of three mutations per recombination (mismatch ratio). The first few sequences from these new lineages will be a poor match to the existing ARG, and the HMM will therefore search for ways to reduce the number of mutations required by recombining segments. This appears to be the case for the Delta VoC, corresponding to rows A and B of Table 3, which emerged and quickly rose to high prevalence in early-mid 2021, carrying 30 characteristic mutations compared to the reference sequence (McCrone et al., 2022), including nine mutations in the S gene not seen in earlier VoCs. The origins of the Delta VoC are are illustrated in Figure S8, which shows the large numbers of mutations involved, and the multiple closely related recombinations inferred before tree-like behaviour is resumed at the ancestor of 90% of Delta samples (node `tsk261771`). Similar patterns around the emergence of Delta are seen in the Wide ARG (not shown).

Given these considerations, it is important to note that the number of ultimate descendants is not an entirely reliable indicator of the quality of inferred recombinants. Further study is required to systematically identify such false positive recombinants, and to update the topology with a more parsimonious explanation of the data.

3 Discussion

Although the COVID-19 pandemic is no longer considered a global emergency by the WHO, the prevalence of SARS-CoV-2 continues to be high worldwide. This fuels proliferation of many variants, with more than 600 Pango-designated lineages circulating globally in the last three months (January to March, 2023; `https://gisaid.org/`; accessed on 2023-03-27). High prevalence increases the risk of coinfection, providing opportunities for new phenotypically distinct recombinants to emerge and spread. Phylogenetic approaches have been central to responses to the pandemic thus far (Attwood et al., 2022; Bloom and Neher, 2023; Abbas et al., 2022; McLaughlin et al., 2022). However, with the rise to high frequency of recombinant lineages (e.g. XBB; Tamura et al., 2023), it is imperative that these methods are updated to incorporate the effects of recombination, so that future public health interventions are not based on incomplete and potentially biased evolutionary models. Here we have introduced the first method to infer an evolutionary history that jointly estimates genealogies with both mutation and recombination at pandemic scale and illustrated how this single structure accurately captures results derived by many different means.

Nonetheless, `sc2ts` is currently "alpha" quality software, and we caution against over interpreting current results. As we have sought to illustrate throughout, there are some clear areas for improvement. The pipeline used to identify and mask erroneous sites in the input alignments is simplistic, and, among other issues, results in multi-base indels being marked as missing data (Section 4.7). A more sophisticated approach (e.g., Aksamentov et al., 2021) would likely yield significant improvements and reduce the effect of sites with artefactually high levels of mutation (e.g., site 28271; Section 2.5). Using a pre-existing tree built using state-of-the-art phylogenetic methods for the early stages of the pandemic (Section 2.2), and minor adaptations to standard node-dating methods (Section 4.6) should help resolve the most notable issues with the inferred backbone phylogeny (Figure 2). Trees constructed from daily sample clusters have a surprisingly large influence on the overall ARG topology (Section 4.3), and so using a more sophisticated tree building approach should yield clear improvements. The unrealistically large number of reversion mutations (Section 2.3) may be reduced by improvements to the current parsimony heuristics (Section 4.4). A major source of errors are the "time-traveller" samples, whose recorded collection dates are months (or years) too early (Section 4.8). While it is unclear how we might solve this problem in general, some simple solutions such as filtering out sequences that exceed a given cost in the LS HMM (i.e., number of mutations and recombination switches) may work well in practice. Such an approach would also reduce the impact of sequences with high levels of sequencing error (which currently contribute a large number of mutations). Taken together, these and other relatively minor improvements should enable inference over much larger subsets of the dataset and give a clearer picture of the combined processes of recombination and mutation over the pandemic so far.

An attractive feature of `sc2ts` is that the most difficult part of the inference problem—finding likely recombinant paths through the existing ARG for new samples—is solved exactly under a well-defined statistical model, using established HMM methodology (Section 4.2). The implementation currently uses a single, arbitrarily chosen, maximum likelihood path via the Viterbi algorithm, but there are numerous ways in which the HMM machinery could be extended in order to explore the set of possible paths, or to quantify the uncertainty around it. Similarly, the current parameterization of the HMM with a single mismatch ratio is simplistic, and it would be straightforward to condition on per-site mutation rates (and nucleotide-dependent state transitions, with some additional development). Recombination breakpoints for the ARG are currently inserted at the right-most extent of the possible interval (Section 2.5). More likely locations for the breakpoint could be chosen within the interval, for example based on sequence motifs (Gallaher, 2020; Yang et al., 2020). It is likely that the basic machinery of finding matches and quantifying the uncertainty around them under a well-defined statistical model in large ARGs would have many applications besides those explored here.

The vast volume of whole genome sequence data generated during the pandemic has presented classical phylogenetic methods and software with major difficulties (Hodcroft et al., 2021). Standard interchange formats such as FASTA, Newick and VCF were simply not designed to deal with millions of samples, and their limitations have come sharply into focus (Turakhia et al., 2021; De Maio et al., 2023). Replacements that can scale to millions of genomes have had to be developed at speed, usually focusing on compiled programming languages to maximise performance. Here, however, we have developed a new method based on an existing data structure and library infrastructure, designed from the beginning to scale to millions of samples (Kelleher et al., 2016, 2019). The sc2ts package is written entirely in Python and by reusing existing high-performance components can infer recombinant viral ancestry at unprecedented scale. Similarly, all of the analyses shown here are written in Python, using the tskit API, mostly running in a few seconds on standard laptop computers (see the Data Availability section for details of the corresponding Jupyter notebooks). Retooling methods to scale up rapidly with expanding data and to encode recombination promises to improve tracking during this and any future pandemic.

4 Methods

Sc2ts (pronounced "scoots", optionally) is a method for inferring Ancestral Recombination Graphs (ARGs; see Section 4.1) from densely sampled pandemic-scale data in real time, in which recombination occurs at a low but significant rate. The basic idea is to incrementally update an ARG each day with the sequences collected on that day (Figure 1). The first step is to find likely "copying paths" under the Li and Stephens model for each sequence in the daily batch to the current ARG (Section 4.2). These copying paths will mostly consist of a new sample sequence copying from a node in the ARG with a small number of mutations, and often many samples from a daily batch will copy from the same ARG node. The second step is then to "resolve" these implied polytomies by using standard tree-building techniques (Figure 1C; Section 4.3). This greedy update strategy inevitably leads to unparsimonious topologies, and the third step is to then increase the overall parsimony of the inferred ARG by making some simple topological updates (Figure 1C,D; Section 4.4). Recombination is inferred as an integral and ongoing part of this process, requiring only a few additional steps to facilitate later analysis (Section 4.5). The result of inference for a given day is then a genealogy recording genetic inheritance as well as mutation and recombination events for all of the sequences inserted into the ARG up to that day, which can be conveniently and efficiently analysed using the mature and feature-rich tskit library (Kelleher et al., 2018; Ralph et al., 2020; Tskit developers, 2023).

4.1 Ancestral Recombination Graphs

The term "Ancestral Recombination Graph" was introduced by Griffiths and colleagues (Griffiths, 1991; Griffiths and Marjoram, 1997) and originally defined as an alternative formulation of the coalescent with recombination stochastic process (Hudson, 1983). Subsequently, the term ARG came to be used in a more general way to describe not just realisations of this model, but to any recombinant genetic ancestry (Minichiello and Durbin, 2006; Zhang et al., 2023). While there is some subtlety in the details (Wong et al., 2023), we can think of an ARG as being any graph that encodes the reticulate genetic ancestry of a sample of colinear sequences under the influence of recombination. This definition encompasses various types of graphs often described using the broader term of phylogenetic networks.

The "succinct tree sequence" is an ARG data structure that is both general (in terms of the types of ancestry that can be described) and computationally efficient (Wong et al., 2023). Originally developed to facilitate large-scale coalescent simulations (Kelleher et al., 2016), the methods have been extended and applied to forward-time simulations (Kelleher et al., 2018; Haller et al., 2019), calculation of population genetics statistics (Ralph et al., 2020) and ARG inference (Kelleher et al., 2019; Wohns et al., 2022). The succinct tree sequence is based on a simple tabular representation, which defines a set of nodes, edges, sites and mutations. A *node* represents a particular genome, which may be an observed sample or an inferred genetic ancestor. The genetic inheritance between a pair of nodes along a segment of genome is defined by the *edge* (ℓ, r, p, c), which states that child node c inherited its genome from parent node p from left coordinate ℓ to right coordinate r. A *site* defines a position on the genome and the ancestral state (allele) at that site. A *mutation* records the site and node IDs where a mutation occurs and the derived state (allele). In addition to these basic elements of the data model, the tskit library supports additional tables and fields, the ability to associate arbitrary metadata with table rows, and facilities to record provenance information (Tskit developers, 2023).

Tskit supports arbitrarily complex patterns of mutation at a particular site, and it is useful to define some terminology to classify them. A mutation's "parent" is the first mutation encountered (at that site) on the path to root from the mutation's node in the local tree corresponding to the site's position. If no other mutation is encountered on the path to root, the mutation's parent is null. The "derived state" of a mutation is the allelic state inherited by nodes in the subtree rooted at the mutation's node (in the local tree), assuming there are no subsequent descendant mutations. The "inherited state" of a mutation is the derived state of the parent mutation, if it exists, or the site's ancestral state otherwise. A "recurrent mutation" is a mutation with a non-null parent, and a "reversion" is a recurrent mutation that reverses the state change of its parent. For example, if we have two mutations a and b, such that a is the parent of b, with state transitions (inherited state → derived state) $a : A \rightarrow T$ and $b : T \rightarrow A$, we define b as a reversion.

Given the node, edge, site and mutation tables we can efficiently construct the local genealogical trees along the genome (arising from recombination) and perform a range of calculations efficiently by reasoning about the differences between these local trees (Kelleher et al., 2016; Ralph et al., 2020). These algorithms have led to performance increases of several orders of magnitude over the state-of-the-art in a range of applications (Kelleher et al., 2016, 2018, 2019; Ralph et al., 2020; Baumdicker et al., 2022). The succinct tree sequence encoding is also very concise, allowing, for example, for millions of complete human genomes to potentially be stored in a few gigabytes of space (Kelleher et al., 2019).

The `tskit` library (Tskit developers, 2023) is a liberally licensed open source toolkit that provides a comprehensive suite of tools for working with ARGs. Based on core functionality written in C, it provides interfaces in C, Python and Rust. The Python interface is based on NumPy (Harris et al., 2020) and provides a convenient platform for interactive analysis of large-scale data using, for example, Jupyter notebooks (Kluyver et al., 2016) and taking advantage of the analysis tools in the burgeoning PyData ecosystem. (It is possible to access the toolkit from R via the `reticulate` package, and the `slendr` library (Petr et al., 2022) also provides some native R support. A full R interface would be a valuable addition to the ecosystem.) Tskit is mature software, widely used in population genetics, and has been incorporated into several downstream applications (e.g., Haller and Messer, 2019; Speidel et al., 2019; Terasaki Hart et al., 2021; Fan et al., 2022; Korfmann et al., 2022; Mahmoudi et al., 2022; Petr et al., 2022; Rasmussen and Guo, 2022; Zhang et al., 2023). It is important to note that this ecosystem for storing and manipulating ARGs can generally be used to efficiently record and analyse SARS-CoV-2 genealogies reconstructed using other methods, not only the `sc2ts` approach that we describe here. Note also that there is no requirement that recombination be present, and the methods are also very efficient when working with a single tree.

4.2 The Li and Stephens model

The Li and Stephens (LS) model (Li and Stephens, 2003) is an approximation of the coalescent with recombination (Hudson, 1983) which captures many of the important features of the joint processes of mutation and recombination. It is a Hidden Markov Model (HMM) in which a focal genome is modelled as a sequence of nucleotides that are probabilistically emitted as an imperfect mosaic of a set of reference genomes (Figure 8). The LS model is used in a wide variety of applications in genomics, including modern approaches to statistical genotype phasing and imputation (Delaneau et al., 2019; Browning et al., 2021, 2018; Rubinacci et al., 2020), and estimation of parameters such as recombination rates (e.g., Hinch et al., 2011) and intensity of selection within and across hosts in viral sequence data (e.g., Palmer et al., 2019). See McVean and Kelleher (2019) for further review and discussion of the LS model.

The generative process of the LS model is summarised in Figure 8. Here, a transition matrix, Q, governs the process of switching (recombining) between members of the reference panel (the hidden states). An emission matrix, E, allows for differences between the focal sequence and the hidden state from which it is copied (due to mutation as well as sequencing error). Both E and Q may be a function of the reference panel members, but transitions are generally assumed to be independent of the hidden states (Figure 8, pink panel). This assumption dramatically increases performance as the state space drops to two states (i.e., switching or not switching). Emissions may also be a function of the nucleotide states, but in our RNA virus case we assume that mutations occur from all possible alleles (A, C, G, U and a gap in the alignment, $-$) to any other with equal probability $\mu_\ell/4$. This is reasonable for rapidly evolving pathogens, but we note that setting the number of alleles at site ℓ (a_ℓ) to the set of observed alleles across all analysed samples ($\mu_\ell/(a_\ell - 1)$) will often be more appropriate (Figure 8, blue panel). We use the Viterbi algorithm (Viterbi, 1967) to find the most likely copying path, given Q, E, and a set of reference sequences. Throughout, we refer to the probabilities of mismatching to a member of

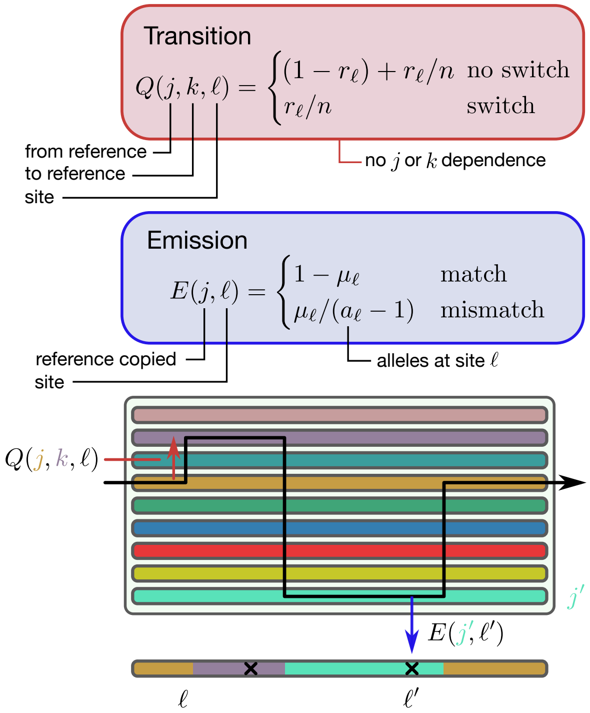

Transition
$Q(j,k,\ell) = \begin{cases} (1-r_\ell)+r_\ell/n & \text{no switch} \\ r_\ell/n & \text{switch} \end{cases}$
from reference
to reference
site
no j or k dependence
Emission
$E(j,\ell) = \begin{cases} 1-\mu_\ell & \text{match} \\ \mu_\ell/(a_\ell-1) & \text{mismatch} \end{cases}$
reference copied
site
alleles at site ℓ
$Q(j,k,\ell)$
j'
$E(j',\ell')$
ℓ
ℓ'

Figure 8: A schematic of the Li and Stephens (LS) model, in which a focal sequence (bottom) is described as an imperfect mosaic of the sequences in a reference panel. Black crosses along the focal sequence show sequencing errors or mutations. In the standard formulation, at site ℓ, the recombination probability is r_ℓ, the mutation probability is μ_ℓ and n denotes the size of the reference panel. The Viterbi algorithm can be used to find a "copying path" through the reference panel for a given focal sequence that maximises the likelihood under these parameters. Unseen states in the reference panel are shown as coloured lines enclosed by the grey box. The black arrow describes the true path through the data which leads to the emitted focal sequence below. Examples of transition and emission probabilities along this trajectory are shown by the red and blue arrows, respectively.

the reference panel at site ℓ as μ_ℓ, and the probability of recombining between two members of the reference panel between site $\ell - 1$ and site ℓ as r_ℓ. For convenience, μ_ℓ is commonly referred to as the 'mutation probability', but we note that this probability of mismatch also encompasses various other error modes that result in a mismatch (such as sequencing and alignment errors). Note also that it is these probabilities, not the rates of mutation and recombination, that are required to fully define the HMM; see Donnelly and Leslie (2010) for a discussion of how these parameters relate to the coalescent process.

The probabilities of these competing processes of mismatch and recombination are usually controlled by the site-specific parameters μ_ℓ and r_ℓ, respectively. For this work we used a slightly different formulation, which uses one parameter, the mismatch ratio (MMR), to control the relative importance of mutation and recombination in the HMM. Specifically, an MMR value of k will prefer k mismatches (mutations) to a single recombination that results in copying from a template with no mismatches. To map between recombination probability and mutation probability for a particular mismatch ratio, we simply consider the two paths that we wish to be equally likely and rearrange for the mutation or recombination probability. Consider the simple case where we assume $\mu_l = \mu$ and $r_l = r$. Without loss of generality, consider a region of length m. Up until this region, two paths are equally likely, so we can re-scale by the likelihood of observing the focal sequence up to the site before the region starts, c. For an MMR of k, we only need to consider two of the n members of the reference panel. One, $\mathcal{P}_1$, for which there are no mismatches, but we need to recombine in, and a second, $\mathcal{P}_2$, for which there are k (randomly chosen) mismatches in the region. The probability of tracking along each of those paths is given by:

$$\mathbb{P}[\mathcal{P}_1] = \frac{cr}{n}\alpha^{m-1}\left(1-\mu\right)^{m} \qquad \text{recombine to a template with no mismatches,}$$

$$\mathbb{P}[\mathcal{P}_2] = c\alpha^{m}\left(1-\mu\right)^{m-k}\mu^{k} \qquad \text{stay in a template with } k \text{ mismatches,}$$

where $\alpha = (1 - r) + r/n$ represents the probability of a sequence not recombining at a given position with any of the other $n-1$ members of the reference panel (Figure 8, red panel) and c is the likelihood of the path up to this point (assumed to be equal by construction). We then set these path probabilities to be equal, and rearrange to relate r and μ to one another:

$$r = \frac{n\mu^k}{\left(1-\mu\right)^k + \left(n-1\right)\mu^k}, \qquad \mu = \frac{1}{\sqrt[k]{\frac{n}{r} - \left(n-1\right)} + 1}.$$

Thus, for lower MMR values (k here), recombination is increasingly favoured over multiple mutations in a specific ancestral genome. We use an MMR value of $k = 3$ in this work, because of the relatively high rate of recombination relative to mutation typical of coronaviruses (Amoutzias et al., 2022). Exploring different mismatch ratios and more sophisticated parameterizations of the HMM are important avenues for future work.

Genetic data for SARS-CoV-2 contains substantial amounts of missingness, (nucleotides coded as 'unknown' or masked out as described in Section 4.7) and it is important to account for this missingness in a systematic way. In `sc2ts` we automatically impute missing data as samples are added into the ARG, using the LS model. To do this, we assume that sites with missing nucleotides are uninformative to the path probability, by setting the emission probability from any state $(A, C, G, U, -)$ to 'unknown' equal to 1 (though we could have chosen any constant). As a result, emissions of unknown nucleotides will not contribute to differences in path probabilities. Once the most likely copying path is determined, we then attach the sample to the ARG (see Figure 1 and following subsections). For each newly attached sample we encode its nucleotide sequence by recording mutations where it differs from the nucleotide sequence of its parental node, but (importantly) ignoring sites with missing data in the new sample. Thus, once the newly added sequence has been attached to the ARG any missing data is imputed from its parental node. There is therefore no missing data in the ARG; all missing bases are "hard called" at attachment time. Note that this approach is equivalent to using state-of-the-art imputation methods (e.g. Browning et al., 2018; Delaneau et al., 2019) with a reference panel consisting of all sequences in the ARG, since these methods are also based on the LS model. Evaluating the accuracy of missing data imputation using `sc2ts` is an important facet of future work.

In `sc2ts`, we use the efficient ARG-based implementation of the LS Viterbi algorithm from `tsinfer` (Kelleher et al., 2019) to find the most likely copying path for each sample sequence among all sequences (sampled and inferred) in the current ARG. In the majority of cases, with non-recombinant sample sequences, the most likely solution is to copy from one of the nodes in the ARG that minimises the number of mutations required to insert the focal sequence. Importantly, because the reference panel here consists of *every* node in the ARG, we can match to both older sample sequences or internal nodes representing an inferred ancestral sequence (see subsequent subsections for details about how these are added). Thus, when no recombination is present, the LS Viterbi algorithm is implementing a version of parsimony, in which we are guaranteed to find a sequence that minimises the number of additional mutations required to incorporate a newly-added sample into the ARG. Recombination is then inferred when the most likely solution to the LS HMM is to copy from more than one ARG node along the genome, for a given sample sequence.

The Viterbi algorithm enables us to find a path through the reference panel from among the n^m paths that is provably at the optimum, under the LS model. We can solve this massive optimisation problem exactly because the ARG-based implementation of the LS HMM used in `tsinfer` (Kelleher et al., 2019) scales approximately logarithmically with reference panel size (as opposed to linearly, for standard matrix-based approaches). This efficient HMM algorithm is the main reason for `tsinfer`'s scalability, and here allows us to find closely matching sequences and recombination paths among millions of SARS-CoV-2 genomes exactly under a well-defined statistical model.

It is important to note that the Viterbi algorithm only returns *one of* the copying paths that maximise the likelihood under the given mutation and recombination parameters. There may be many such paths, from which we choose one arbitrarily. Also, the present choice of using a single mismatch-ratio parameter to control the likelihood of recombination vs mutation may lead to relatively flat likelihood spaces where many different paths have equal likelihood. There are many possibilities in using established HMM methodology to reason about and explore the space of possible matches, which may be a fruitful avenue for future work. Examples include stochastic traceback (e.g., Rasmussen et al., 2014) through the collection of paths at the global optimum to glean further information about the likelihood surface, and determine whether there are downstream implications for our conclusions. Here, we have considered the Viterbi algorithm to make statements about the most likely paths through the data. The machinery used here can be modified to run the forwards and backwards algorithms that determine the probability of observing a focal sequence, integrating over all possible paths through the data, under the LS model (Palmer et al., 2023). This presents an opportunity to estimate parameters of interest under the LS model at pandemic scale.

4.3 Tree inference from HMM daily sample clusters

With tens of thousands of samples being added to the ARG per day, there are often clusters of hundreds of sequences attaching to the same node (or more generally, recombinant path; see Section 4.5). While some of these samples will require no extra mutations (because they are identical to the attachment node), in general there will be complex patterns of shared mutations among the samples reflecting their evolutionary relationships. A natural way to infer these within-cluster evolutionary relationships is to use standard tree-building algorithms. We can infer a likely tree relating the samples in a cluster independently of the other samples in a daily batch and then attach the tree (and mutations) to the ARG at the node identified by the HMM.

We currently use the UPGMA algorithm (Michener and Sokal, 1957) as implemented in SciPy (Virtanen et al., 2020) to build trees from sample clusters, and then map mutations back to this topology using maximum parsimony. We chose this approach mainly for simplicity, and because of the speed and reliability of the SciPy implementation. An issue with the UPGMA algorithm is that it generates a strictly binary tree, creating internal nodes supported by no informative site (i.e., having no mutation immediately ancestral to them). We avoid such false precision by post-processing to remove unsupported internal nodes, representing the relationship between k identical descendants of a node as a polytomy of size k.

There are well-known issues with using such a simple algorithm for inferring evolutionary relationships (Felsenstein, 2004). Table 1 shows that this within-cluster tree building has a significant influence on the overall ARG topology, and therefore applying more sophisticated tree building methods that keep track of the required mutations (rather then inferring post-hoc by parsimony) is a likely avenue for improvements in overall inference quality.

4.4 Parsimony-increasing heuristics

Attaching trees built from the clusters of samples that copy from a particular node (or path of nodes for recombinants, see Section 4.5) under the HMM is an inherently greedy strategy and can produce inferences that are clearly unparsimonious. The final step in adding a daily batch of samples to the ARG is therefore to perform some local updates that target specific types of parsimony violations in the just-updated regions of the ARG. There are currently two parsimony-increasing operations applied, which we refer to as "mutation collapsing" and "reversion pushing" (Figure 1D, E).

Given a newly attached node, mutation collapsing inspects its siblings from previous sample days to check if any of them share (a subset of) the mutations that it carries. If so, we increase the overall parsimony of the inference by creating a new node representing the ancestor that carried those shared mutations and make that new node the parent of the siblings carrying those shared mutations (Figure 1D). The patterns of shared mutations between siblings can be complex, and the current implementation uses a simple greedy strategy for choosing the particular mutations to collapse.

The reversion push operation inspects a newly added node to see if any of its mutations are "immediate reversions"; that is, are reversions of a mutation that occurred on the new node's immediate parent. We increase the overall parsimony of the inference by "pushing in" a new node which descends from the original parent, and carries all its mutations except those causing the reversions on the newly added node (Figure 1E).

Table 1 shows that nodes generated by these operations constitute roughly the same fraction of the total in both the Long and Wide ARGs, and contribute significantly to the overall topology. These nodes are also being chosen by the LS HMM as likely choices of parent (data not shown) demonstrating that the heuristics are successfully capturing features of real sequences. However, they are both simple greedy operations, just examining the local parts of the ARG topology affected by newly added samples. Because the inferred ARGs still contain a large number of reversion mutations which are likely to be mostly artefactual (Section 2.3), it is clear that there is room for improvement and that further parsimony-increasing heuristics (e.g., resolving reversions beyond those on immediately adjacent edges) would likely be of benefit.

4.5 Treatment of recombinants

A sample sequence is designated as a recombinant if the most likely path inferred by the LS HMM for that sample contains at least one switch between parents. Recombinant sequences are mostly treated identically to non-recombinants, as we simply need to reason about a path of parent nodes along genome intervals rather than a single parent over the whole genome, which is naturally handled by the succinct tree sequence data structure and `tskit` library (Section 4.1). To facilitate analysis and to help understand the robustness of recombinants we perform some additional steps in `sc2ts`.

The LS HMM may infer identical paths and patterns of mutations for multiple samples in a daily batch, and so we create a "recombination" node (marked with a specific "flags" value) for each distinct recombinant. (This node is not strictly necessary but makes it convenient to find recombinants for subsequent analysis.) Variation within a cluster of recombinant sequences is handled in the same way as non-recombinants (see Section 4.3 above for details). When the Viterbi algorithm implementation used by the LS HMM infers recombinant ancestry for a given genome, the point at which inheritance switches from one parent node to another is the last possible position. The left-most extent of the breakpoint interval is derived by sequence comparison between the parents, as described in Section 2.5.

For a particular recombination node in an ARG, a breakpoint is defined as the location at which inheritance switches from one parent to another. In the `tskit` encoding (see Section 4.1) inheritance between nodes is defined by edges (ℓ, r, p, c), which state that child node c inherits from parent node p over the half-closed genome interval $[\ell, r)$. For simplicity, suppose that a recombination node u inherits from two parents p_1 and p_2 with a breakpoint of x. In the ARG, this is defined by two edges $(0, x, p_1, u)$ and (x, L, p_2, u) where L is the length of the genome. Since inheritance intervals are half-closed, u inherits all positions up to x (exclusive) from parent p_1 and all positions from x (inclusive) to the end of the genome from parent p_2. We then define the breakpoint *interval* $[b_\ell, b_r)$ as the half-closed interval defining the range of possible values for x, such that $b_\ell \leq x < b_r$.

The LS HMM machinery, and the interpretation of inferred recombinant paths and breakpoint intervals is a central part of `sc2ts`, and there are many ways to extend and improve. For example, the current parameterization of using a single "mismatch ratio" is very simplistic (Section 4.2) and likely results in a flat likelihood space where many recombinant paths have equal probability of being chosen. Post-processing the match results to produce more parsimonious breakpoints may also be a worthwhile avenue for development. In particular, we may choose to insert breakpoints for the ARG that are chosen from within the possible interval, rather than the current approach of taking the rightmost value. Cases where we have more than two parents may either be the "stacking" of multiple recombination events or instances where the HMM has chosen to switch to a third sequence rather than back to the original parent (where this is equally parsimonious). Many putative recombination events, however, will represent poor quality data, where a recombinant copying path happens to be a more likely explanation of a highly divergent sequence. A thorough analysis of the behaviour of the LS model in the context of a pandemic-scale ARG may lead to significant improvements in our ability to identify recombinants and to filter poor quality data.

4.6 Node dating

The approach to assigning a date to nodes in `sc2ts` is currently ad-hoc, and the inferred timing of events from the ARGs reported here should be treated with caution (e.g., Figure 5). Sample nodes (those corresponding to observed sample sequences) are the most accurately dated, as we use the reported collection date for these nodes. These are not entirely accurate, but our data filtering criteria should remove the most egregious errors (see Section 4.8). Other nodes in the ARG are dated by splitting the time between the attached samples and the chosen parent nodes equally (in the case of trees inferred from daily sample clusters, Section 4.3) or by adding arbitrary small values when creating new nodes using parsimony rules (Section 4.4). Because the ARGs for SARS-CoV-2 are very treelike, with recombination nodes constituting a tiny fraction of the overall topology (Table 1), existing methods (e.g., To et al., 2016) could likely be adapted to accurately date the vast majority of the nodes.

4.7 Data preprocessing

The findings of this study are based on sequences and metadata available on GISAID (`https://gisaid.org/`) up to 2022-08-22 and accessible at `https://doi.org/10.55876/gis8.230329cd`. We removed sequences if they had ambiguous collection dates, were collected before 2020-01-01 or were isolated from a non-human host. We aligned sequences to the Wuhan-Hu-1/2019 reference sequence (GenBank: MN908947.3) using Nextclade v2.3.0 (Aksamentov et al., 2021) (dataset tag 2022-07-26T12:00:00Z). We also excluded sequences if they had a "bad" quality control status in any of the four Nextclade columns ("qc.missingData.status", "qc.mixedSites.status", "qc.frameShifts.status" and 'qc.stopCodons.status").

We encode ambiguous nucleotide letters (i.e., not A, C, G, T, or a gap) in the pairwise genome alignments as missing data (N). Problematic bases in the alignments, which had two or more Ns or gaps within a distance of seven bases, are masked as missing data following the approach used in the "faToVCF" tool used by UShER (Turakhia et al., 2021). Sites that are masked by this process are treated as missing data by the LS HMM (Section 4.2). In addition, we exclude 481 problematic sites flagged as prone to sequencing errors or as highly homoplasic entirely (`https://github.com/W-L/ProblematicSites_SARS-CoV2/`, accessed 2022-09-22).

Although the current masking strategy is simple and robust, there are significant disadvantages because it excludes, for example, any deletions of length greater than one base. Exploring more sophisticated masking strategies is an important route for future improvements.

4.8 Filtering "time travellers"

A major source of error in early versions of `sc2ts` was the existence of "time traveller" sequences: those with erroneously early collection dates. For example, an Alpha sample purportedly collected in 2020 from the United States, before Alpha appeared in the United Kingdom (USA/MN-Mayo-1563/2020), produced significant topological distortions in inferred ARGs. Hence, to exclude such potential "time travellers" we employ two filters.

The first filter is a simple threshold on the time delay between the collection date and submission date. After some preliminary analysis we settled on a maximum submission delay of 30 days when building the ARGs described here. The second filter is to remove any sequence with a collection date

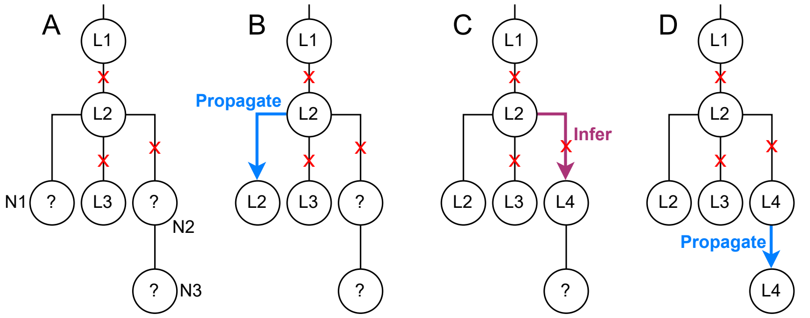


Figure 9: A schematic of the iterative procedure to impute Pango lineage for inserted, non-sample nodes. Here, three nodes (question marks) have unknown Pango lineage (A). The lineage for node N1 can be directly copied from its parent (L2), which has an identical sequence. The lineage for node N2 must be inferred from that of the parent (L2) plus the lineage-defining mutation (red X) on the connecting edge. The lineage for node N3 can then be copied from that of node N2.

that pre-dates the time to the most recent common ancestor (tMRCA) of its corresponding clade. We obtained the tMRCA for each clade from a Nextstrain GISAID reference tree (downloaded on 2022-08-22), and we used the lower bound of the 95% confidence interval of each clade as the minimum date cut-off. This excludes a further 618 samples not covered by the maximum submission delay filter.

As the results of `sc2ts` are sensitive to the existence of time-travellers, an important aspect of future work is to find better ways to identify them. One possibility is to use the LS HMM itself to flag overly divergent sequences and to exclude them from attachment to the ARG. We might also estimate collection dates by adding these potential time travellers back to the ARG, allowing an automated assessment of collection date discrepancies.

4.9 Imputation of Pango lineage for non-sample nodes

While the Pango lineages of sample nodes are imported directly from GISAID metadata, the lineage status of internal nodes inserted into the `sc2ts` ARG must be imputed. We do this using the list of lineage-defining mutations (based on 90% consensus of the sequences analysed) from the COVID-CG website (Chen et al., 2021, `https://covidcg.org/`; accessed on 2022-11-04).

For a given non-sample node u, if the Pango lineage of its parent or one of its children is already known, and there are no lineage-defining mutations on the connecting edge, then u copies this Pango lineage exactly. Otherwise, the lineage for u is inferred by matching its full set of mutations against the COVID-CG list. We apply these two steps to the internal nodes of the ARG iteratively, as illustrated in Figure 9, until all internal nodes are assigned a lineage (where possible—note that sometimes a lineage cannot be assigned to the children of a recombination node).

This method is fast, as the lineages for most of the internal nodes can be imputed by copying from the surrounding nodes (Wide ARG: 80% of nodes, Long ARG: 66%), and significantly more efficient than extracting the haplotypes for each internal node and using existing Pangolin assignment tools (O'Toole et al., 2021). The accuracy depends on the quality of the list of lineage-defining mutations, as well as the source of lineage designation for the sample nodes: we obtain slightly different results when using those recorded on GISAID (which uses pangoLEARN), and those assigned by Nextclade. To gauge the accuracy of imputation, we have used our method to re-impute the lineage designations of each sample node using the surrounding information; this results in 99% of nodes being assigned the same lineage as per the source metadata in the Wide ARG, and 98% in the Long ARG.

5 Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. SHZ is supported by the Janssen-Oxford Translational Genomics Fellowship. JK, YW, and BJ are supported by the Robertson Foundation. AI is supported by the Wellcome Trust.

Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

6 Data Availability

The source code for `sc2ts` and notebooks and code used to produce the results described here available on GitHub:

- `https://github.com/jeromekelleher/sc2ts/`
- `https://github.com/jeromekelleher/sc2ts-paper/`

Details of the GISAID data used are available at `https://doi.org/10.55876/gis8.230329cd` and included in the Supplemental Table (GISAID EPI SET PDF).

The inferred ARGs described here are available on request to those with the appropriate GISAID data access.

The mapping of `tskit` IDs to strain and EPI_ISL identifiers for the subgraph plots in Supplementary Figures S4, S5, S6 and S7 are at `https://github.com/jeromekelleher/sc2ts-paper/blob/main/data/Subgraph_sample_mapping.txt`.

References

Abbas, M., Cori, A., Cordey, S., Laubscher, F., Robalo Nunes, T., Myall, A., Salamun, J., Huber, P., Zekry, D., Prendki, V., et al. Reconstruction of transmission chains of SARS-CoV-2 amidst multiple outbreaks in a geriatric acute-care hospital: A combined retrospective epidemiological and genomic study. *eLife*, **11**: e76854, 2022.

Aksamentov, I., Roemer, C., Hodcroft, E. B., and Neher, R. A. Nextclade: Clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, **6**(67): 3773, 2021.

Amoutzias, G. D., Nikolaidis, M., Tryfonopoulou, E., Chlichlia, K., Markoulatos, P., and Oliver, S. G. The remarkable evolutionary plasticity of coronaviruses by mutation and recombination: insights for the covid-19 pandemic and the future evolutionary paths of sars-cov-2. *Viruses*, **14**(1): 78, 2022.

Anisimova, M., Nielsen, R., and Yang, Z. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, **164**(3): 1229–1236, 2003.

Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R., and Pybus, O. G. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics*, **23**(9): 547–562, 2022.

Bal, A., Simon, B., Destras, G., Chalvignac, R., Semanas, Q., Oblette, A., Quéromès, G., Fanget, R., Regue, H., Morfin, F., et al. Detection and prevalence of SARS-CoV-2 co-infections during the Omicron variant circulation in France. *Nature Communications*, **13**(1): 6316, 2022.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, **220**(3), 2022.

Bloom, J. D. and Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *bioRxiv*, 2023. URL `https://doi.org/10.1101/2023.01.30.526314`

Boni, M. F., Posada, D., and Feldman, M. W. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, **176**(2): 1035–1047, 2007.

Browning, B. L., Tian, X., Zhou, Y., and Browning, S. R. Fast two-stage phasing of large-scale sequence data. *American Journal of Human Genetics*, **108**(10): 1880–1890, 2021.

Browning, B. L., Zhou, Y., and Browning, S. R. A One-Penny imputed genome from Next-Generation reference panels. *American Journal of Human Genetics*, **103**(3): 338–348, 2018.

Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A., and Deverman, B. E. COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest. *eLife*, **10**: e63409, 2021.

Chen, C., Nadeau, S., Yared, M., Voinov, P., Xie, N., Roemer, C., and Stadler, T. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*, **38**(6): 1735–1737, 2022.

Corey, L., Beyrer, C., Cohen, M. S., Michael, N. L., Bedford, T., and Rolland, M. SARS-CoV-2 variants in patients with immunosuppression. 2021.

De Bernardi Schneider, A., Su, M., Hinrichs, A. S., Wang, J., Amin, H., Bell, J., Wadford, D. A., O'Toole, Á., Scher, E., Perry, M. D., et al. SARS-CoV-2 lineage assignments using phylogenetic placement/UShER are superior to pangoLEARN machine learning method. *bioRxiv*, 2023. URL `https://doi.org/10.1101/2023.05.26.542489`

De Klerk, A., Swanepoel, P., Lourens, R., Zondo, M., Abodunran, I., Lytras, S., MacLean, O. A., Robertson, D., Kosakovsky Pond, S. L., Zehr, J. D., et al. Conserved recombination patterns across coronavirus subgenera. *Virus Evolution*, **8**(2): veac054, 2022.

De Maio, N., Kalaghatgi, P., Turakhia, Y., Corbett-Detig, R., Minh, B. Q., and Goldman, N. Maximum likelihood pandemic-scale phylogenetics. *Nature Genetics*, **55**: 746–752, 2023.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., and Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nature Communications*, **10**(1): 5436, 2019.

Donnelly, P. and Leslie, S. The coalescent and its descendants. *arXiv preprint arXiv:1006.1514*, 2010.

Fan, C., Mancuso, N., and Chiang, C. W. A genealogical estimate of genetic relationships. *The American Journal of Human Genetics*, **109**(5): 812–824, 2022.

Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

Gallaher, W. R. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-CoV-2. *Archives of Virology*, **165**(10): 2341–2348, 2020.

Graham, R. L. and Baric, R. S. Recombination, reservoirs, and the modular spike: Mechanisms of coronavirus cross-species transmission. *Journal of Virology*, **84**(7): 3134–3146, 2010.

Gribble, J., Stevens, L. J., Agostini, M. L., Anderson-Daniels, J., Chappell, J. D., Lu, X., Pruijssers, A. J., Routh, A. L., and Denison, M. R. The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathogens*, **17**(1): e1009226, 2021.

Griffiths, R. Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, **19**(2): 169–186, 1981.

Griffiths, R. C. The two-locus ancestral graph. In I. V. Basawa and R. L. Taylor, eds., *Selected Proceedings of the Sheffield Symposium on Applied Probability. IMS Lecture Notes-Monograph Series*, vol. 18, 100–117. IMS, Hayward, CA, 1991.

Griffiths, R. C. and Marjoram, P. An ancestral recombination graph. In P. Donnelly and S. Tavaré, eds., *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, vol. 87, 257–270. Springer-Verlag, Berlin, 1997.

Guindon, S. and Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5): 696–704, 2003.

Gusfield, D. *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, Cambridge, MA, 2014.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, **34**(23): 4121–4123, 2018.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., and Ralph, P. L. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, **19**(2): 552–566, 2019.

Haller, B. C. and Messer, P. W. SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, **36**(3): 632–637, 2019.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with NumPy. *Nature*, **585**(7825): 357–362, 2020.

Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., Wang, K., Buxbaum, S. G., Akylbekova, E. L., et al. The landscape of recombination in African Americans. *Nature*, **476**(7359): 170–175, 2011.

Hodcroft, E. B., De Maio, N., Lanfear, R., MacCannell, D. R., Minh, B. Q., Schmidt, H. A., Stamatakis, A., Goldman, N., and Dessimoz, C. Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature*, **591**(7848): 30–33, 2021.

Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**: 183–201, 1983.

Huson, D. H. and Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, **61**(6): 1061–1067, 2012.

Ignatieva, A., Hein, J., and Jenkins, P. A. Ongoing recombination in SARS-CoV-2 revealed through genealogical reconstruction. *Molecular Biology and Evolution*, **39**(2), 2022.

Jackson, B., Boni, M. F., Bull, M. J., Colleran, A., Colquhoun, R. M., Darby, A. C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J. T., et al. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, **184**(20): 5179–5188, 2021.

Jungreis, I., Sealfon, R., and Kellis, M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 sarbecovirus genomes. *Nature Communications*, **12**(1): 1–20, 2021.

Kelleher, J., Etheridge, A. M., and McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, **12**(5): e1004842, 2016.

Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. Efficient pedigree recording for fast population genetics simulation. *PLoS Computational Biology*, **14**(11): e1006581, 2018.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. Inferring whole-genome histories in large population datasets. *Nature Genetics*, **51**(9): 1330–1338, 2019.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J. W., Kim, V. N., and Chang, H. The architecture of SARS-CoV-2 transcriptome. *Cell*, **181**(4): 914–921, 2020.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, eds., *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90. IOS Press, Amsterdam, 2016.

Korfmann, K., Awad, D. A., and Tellier, A. Weak seed banks influence the signature and detectability of selective sweeps. *bioRxiv*, 2022.
URL `https://doi.org/10.1101/2022.04.26.489499`

Li, N. and Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**(4): 2213–2233, 2003.

Mahmoudi, A., Koskela, J., Kelleher, J., Chan, Y.-b., and Balding, D. Bayesian inference of ancestral recombination graphs. *PLOS Computational Biology*, **18**(3): e1009960, 2022.

McCrone, J. T., Hill, V., Bajaj, S., Pena, R. E., Lambert, B. C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., et al. Context-specific emergence and growth of the sars-cov-2 delta variant. *Nature*, **610**: 154–160, 2022.

McLaughlin, A., Montoya, V., Miller, R. L., Mordecai, G. J., COVID, C., Worobey, M., Poon, A. F., Joy, J. B., et al. Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada. *eLife*, **11**: e73896, 2022.

McVean, G. and Kelleher, J. Linkage disequilibrium, recombination and haplotype structure. In D. Balding, I. Moltke, and J. Marioni, eds., *Handbook of Statistical Genomics*, 51–86. Wiley, Hoboken, NJ, 2019.

Michener, C. D. and Sokal, R. R. A quantitative approach to a problem in classification. *Evolution*, **11**(2): 130–162, 1957.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, **37**(5): 1530–1534, 2020.

Minichiello, M. J. and Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *The American Journal of Human Genetics*, **79**(5): 910–922, 2006.

Nie, Q., Li, X., Chen, W., Liu, D., Chen, Y., Li, H., Li, D., Tian, M., Tan, W., and Zai, J. Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Research*, **287**: 198098, 2020.

Nielsen, B. F., Saad-Roy, C. M., Li, Y., Sneppen, K., Simonsen, L., Viboud, C., Levin, S. A., and Grenfell, B. T. Host heterogeneity and epistasis explain punctuated evolution of SARS-CoV-2. *PLoS computational biology*, **19**(2): e1010896, 2023.

Otto, S. P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn, D. J., et al. The origins and potential future of sars-cov-2 variants of concern in the evolving covid-19 pandemic. *Current Biology*, **31**(14): R918–R929, 2021.

O'Toole, Á., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, **7**(2): veab064, 2021.

Palmer, D. S., Turner, I., Fidler, S., Frater, J., Goedhals, D., Goulder, P., Huang, K.-H. G., Oxenius, A., Phillips, R., Shapiro, R., et al. Mapping the drivers of within-host pathogen evolution using massive data sets. *Nature Communications*, **10**(1): 3017, 2019.

Palmer, D. S., Wong, Y., and Kelleher, J. Efficient Li and Stephens on ancestral recombination graphs. 2023. In preparation.

Petr, M., Haller, B. C., Ralph, P. L., and Racimo, F. Slendr: A framework for spatio-temporal population genomic simulations on geographic landscapes. *bioRxiv*, 2022.
URL `https://doi.org/10.1101/2022.03.20.485041`

Ralph, P., Thornton, K., and Kelleher, J. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics*, **215**(3): 779–797, 2020.

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., and Pybus, O. G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, **5**(11): 1403–1407, 2020.

Rasmussen, D. A. and Guo, F. Espalier: Efficient tree reconciliation and ARG reconstruction using maximum agreement forests. *bioRxiv*, 2022.
URL `https://doi.org/10.1101/2022.01.17.476639`

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, **10**(5): e1004342, 2014.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**(3): 539–542, 2012.

Rubinacci, S., Delaneau, O., and Marchini, J. Genotype imputation using the positional Burrows Wheeler transform. *PLoS Genetics*, **16**(11): e1009049, 2020.

Schaefer, N. K., Shapiro, B., and Green, R. E. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Science Advances*, **7**(29), 2021.

Schierup, M. H. and Hein, J. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**(2): 879–891, 2000.

Scornavacca, C., Zickmann, F., and Huson, D. H. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics*, **27**(13): i248–i256, 2011.

Sekizuka, T., Itokawa, K., Saito, M., Shimatani, M., Matsuyama, S., Hasegawa, H., Saito, T., and Kuroda, M. Genome recombination between the Delta and Alpha variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Japanese Journal of Infectious Diseases*, **75**(4): 415–418, 2022.

Shu, Y. and McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, **22**(13), 2017.

Simon-Loriere, E. and Holmes, E. C. Why do RNA viruses recombine? *Nature Reviews Microbiology*, **9**(8): 617–626, 2011.

Smith, E., Wright, S., and Libuit, K. Identifying SARS-CoV-2 recombinants. 2023. Accessed: 2023-06-02.
URL `https://pha4ge.org/resource/identifying-sars-cov-2-recombinants`

Speidel, L., Forest, M., Shi, S., and Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, **51**(9): 1321–1329, 2019.

Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C., Zhou, J., Liu, W., Bi, Y., and Gao, G. F. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends in Microbiology*, **24**(6): 490–502, 2016.

Tamura, T., Ito, J., Uriu, K., Zahradnik, J., Kida, I., Anraku, Y., Nasser, H., Shofa, M., Oda, Y., Lytras, S., et al. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nature Communications*, **14**(1): 2800, 2023.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, **7**(6): 1012–1023, 2020.

Terasaki Hart, D. E., Bishop, A. P., and Wang, I. J. Geonomics: Forward-time, spatially explicit, and arbitrarily complex landscape genomic simulations. *Molecular Biology and Evolution*, **38**(10): 4634–4646, 2021.

To, T.-H., Jung, M., Lycett, S., and Gascuel, O. Fast dating using least-squares criteria and algorithms. *Systematic Biology*, **65**(1): 82–97, 2016.

Tskit developers. Tskit: A portable library for population scale genealogical analysis. 2023. In preparation.

Turakhia, Y., Thornlow, B., Hinrichs, A., McBroome, J., Ayala, N., Ye, C., Smith, K., De Maio, N., Haussler, D., Lanfear, R., et al. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature*, **609**(7929): 994–997, 2022.

Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., and Corbett-Detig, R. Ultrafast sample placement on existing trees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, **53**(6): 809–816, 2021.

VanInsberghe, D., Neish, A. S., Lowen, A. C., and Koelle, K. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. *Virus Evolution*, **7**(2): veab059, 2021.

Varabyou, A., Pockrandt, C., Salzberg, S. L., and Pertea, M. Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *Genetics*, **218**(3), 2021.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**: 261–272, 2020.

Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2): 260–269, 1967.

Wertheim, J. O., Wang, J. C., Leelawong, M., Martin, D. P., Havens, J. L., Chowdhury, M. A., Pekar, J. E., Amin, H., Arroyo, A., Awandare, G. A., et al. Detection of SARS-CoV-2 intra-host recombination during superinfection with Alpha and Epsilon variants in New York City. *Nature Communications*, **13**(1): 3645, 2022.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. A unified genealogy of modern and ancient genomes. *Science*, **375**(6583): eabi8264, 2022.

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., and Kelleher, J. A general and efficient representation of ancestral recombination graphs. 2023. In preparation.

Yang, Y., Yan, W., Hall, A. B., and Jiang, X. Characterizing Transcriptional Regulatory Sequences in Coronaviruses and Their Role in Recombination. *Molecular Biology and Evolution*, **38**(4): 1241–1248, 2020.

Yang, Y., Yan, W., Hall, A. B., and Jiang, X. Characterizing transcriptional regulatory sequences in coronaviruses and their role in recombination. *Molecular Biology and Evolution*, **38**(4): 1241–1248, 2021.

Yi, K., Kim, S. Y., Bleazard, T., Kim, T., Youk, J., and Ju, Y. S. Mutational spectrum of SARS-CoV-2 during the global pandemic. *Experimental & Molecular Medicine*, **53**(8): 1229–1237, 2021.

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F., and Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, **55**: 768–776, 2023.

Zou, W., Xiong, M., Hao, S., Zhang, E. Y., Baumlin, N., Kim, M. D., Salathe, M., Yan, Z., and Qiu, J. The SARS-CoV-2 transcriptome and the dynamics of the S gene furin cleavage site in primary human airway epithelia. *MBio*, **12**(3): e01006–21, 2021.

Supplementary Material

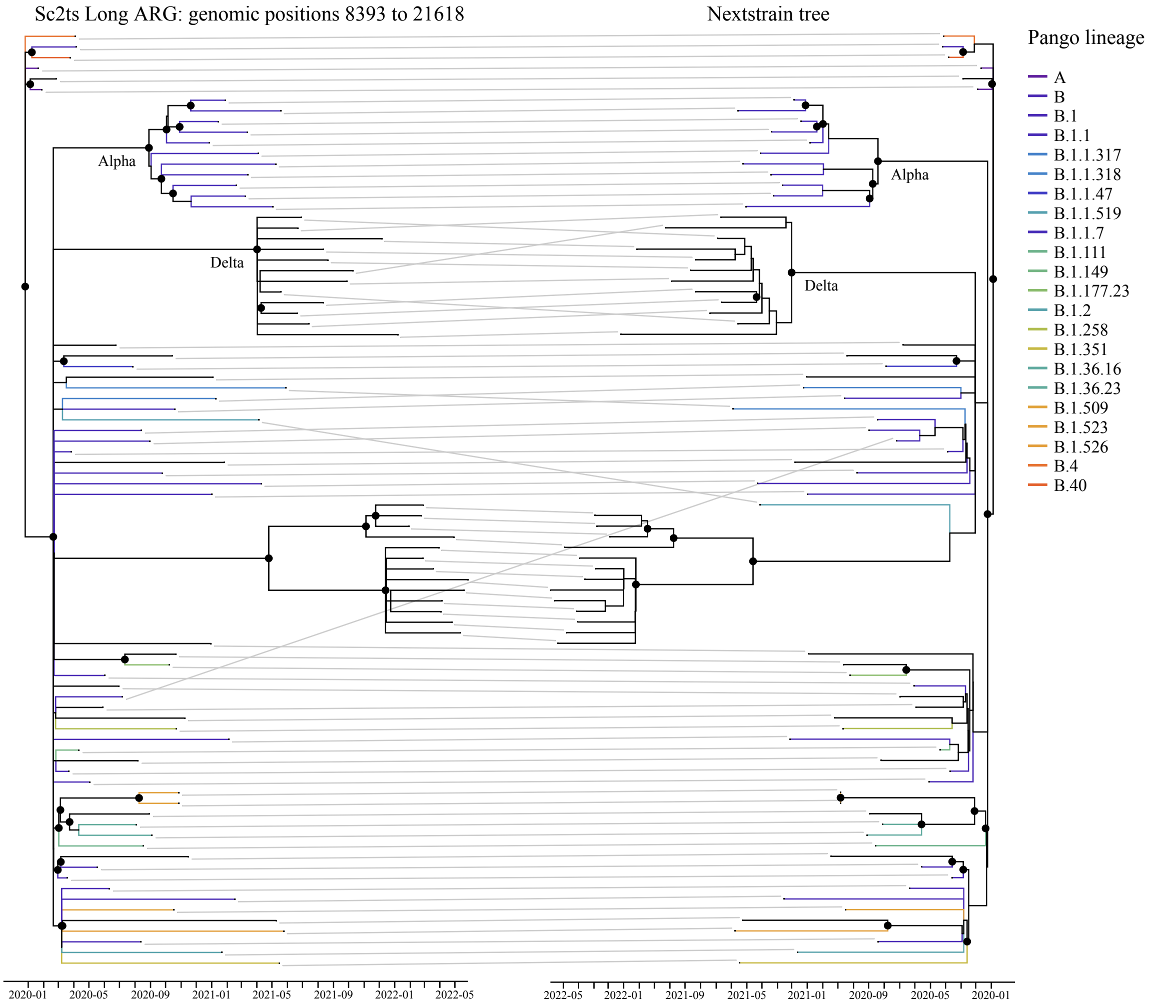


Figure S1: Tanglegram equivalent to that in Figure 2, but for the Long ARG (i.e., subsampled to mid-2022).

		Jackson et al. (2021)		**sc2ts (Wide ARG)**	
Sample	**Group**	**Parents**	Breakpoint interval(s)	**Parents**	Breakpoint interval(s)
ALDP-11CF93B	A	B.1.177 Alpha	21,256–21,615	B.1.177.18 Alpha	21,256–22,228
ALDP-125C4D7	A	B.1.177 Alpha	21,256–21,615	B.1.177.18 Alpha	21,256–22,228
ALDP-130BB95	A	B.1.177 Alpha	21,256–21,615	B.1.177.18 Alpha	21,256–22,228
LIVE-DFCFFE	A	B.1.177 Alpha	18,999–20,296	B.1.177.18 Alpha	21,256–22,228
QEUH-CCCB30	B	B.1.36.28 Alpha	6,529–6,955	B.1.36 Alpha	6,529–6,955
QEUH-CD0F1F	B	B.1.36.28 Alpha	6,529–6,955	B.1.36 Alpha	6,529–6,955
MILK-1166F52	C	Alpha B.1.221.1	25,997–27,443	Alpha B.1.221	25,997–27,973
MILK-11C95A6	C	Alpha B.1.221.1	25,997–27,443	Alpha B.1.221	25,997–27,973
QEUH-109B25C	C	Alpha B.1.221.1	25,997–27,443	Alpha B.1.221	25,997–27,973
MILK-126FE1F	D	B.1.36.39 Alpha	20,704–23,064	B.1.36.39 Alpha	22,445–23,064
RAND-12671E1	D	B.1.36.39 Alpha	20,704–23,064	B.1.36.39 Alpha	22,445–23,064
RAND-128FA33	D	B.1.36.39 Alpha	20,704–23,064	B.1.36.39 Alpha	22,445–23,064
CAMC-CBA018	n/a	B.1.177 Alpha	20,390–21,256	B.1.177 Alpha	17,616–21,256
CAMC-CB7AB3	n/a	Alpha B.1.177 Alpha	3,268–4,476 20,390–21,256	Alpha B.1.177 Alpha	3,268–5,389 17,616–21,256
MILK-103C712	n/a	B.1.177.17 Alpha	409–446 26,802–27,878	n/a	n/a
QEUH-1067DEF	n/a	Alpha B.1.177.9	10,524–10,871	Alpha B.1.177	7,729–10,871

Table S1: Recombinant sequences involving the Alpha (B.1.1.7) variant reported by Jackson et al. (2021) have recombinant ancestry in the Wide ARG. The breakpoint intervals and Pango lineage assignments of the parents were taken from Table 2 (3SEQ results) of Jackson et al., except the Pango lineage assignment of the parents of group B recombinants, which were taken from Table 1 (motif-based results). 3SEQ interval coordinate modifications are described in the caption for Table 2.

Focal Pango	Num origins	Num focal samples (further split by origin, †=nested)	Official Pango parents	**Main clade**: sc2ts parents	**Main clade**: additional descendants
XA	1	**5**	B.1.1.7 + B.1.177	B.1.1.7 + B.1.177.18	
XF	1	**2**	B.1.617.2* + BA.1*	AY.4 + BA.1	
XG	1	**32**	BA.1* + BA.2*	BA.1.17 + BA.2	XAB: 1/48
XH	1	**11**	BA.1* + BA.2*	BA.1.20 + BA.2.9	XAF: 34/35, B.1.1.529: 2, XE: 3/163
XK	1	**3**	BA.1* + BA.2*	BA.1.1.1 + BA.2	
XL	1	**10**	BA.1* + BA.2*	BA.1.17.2 + BA.2	XAB: 1/48, XU: 1/3
XR	1	**8**	BA.1.1* + BA.2*	BA.1.1 + BA.2	XQ: 9/12, XAB: 2/48
XS	1	**4**	B.1.617.2* + BA.1.1*	AY.36 + BA.1.1	
XT	1	**1**	BA.1* + BA.2*	BA.2.23 + Unknown	BA.2.23: 1
XV	1	**1**	BA.1* + BA.2*	BA.1.1 + BA.2 + Unknown	BA.2: 8
XW	1	**11**	BA.1* + BA.2*	BA.1.1.15 + BA.2	XN: 2/13
XY	1	**6**	BA.1* + BA.2*	BA.1.1 + BA.2	XAF: 1/35
XAA	1	**8**	BA.1* + BA.2*	BA.1 + BA.2.9	XAB: 38/48, XAG: 17/17, XU: 1/3, XQ: 2/12
XAC	1	**27**	BA.1* + BA.2*	BA.1.17.2 + BA.2.3	
XAE	1	**9**	BA.1* + BA.2*	BA.1 + BA.2	
XAG	1	**17**	BA.1* + BA.2*	BA.1 + BA.2.9	XAB: 38/48, XAA: 8/8, XU: 1/3, XQ: 2/12
XB	2	57 (**57**, 1)†	B.1.631 + B.1.634	B.1 + B.1.627	B.1.634: 3, B.1.631: 7
XC	2	3 (1, **2**)	AY.29 + B.1.1.7	AY.103 + B.1.1.7	
XD	2	4 (1, **3**)	B.1.617.2* + BA.1*	AY.4 + BA.1.15	
XJ	2	2 (1, 1)	BA.1* + BA.2*	No main clade	No main clade
XAD	2	5 (1, **4**)	BA.1* + BA.2*	BA.1.1 + BA.2	
XAF	2	35 (**34**, 1)	BA.1* + BA.2*	BA.1.20 + BA.2.9	XH: 11/11, B.1.1.529: 2, XE: 3/163
XAH	2	12 (**10**, 2)	BA.1* + BA.2*	BA.1 + BA.2.10	XZ: 61/66, XAD: 1/5
XQ	3	12 (2, **9**, 1)	BA.1.1* + BA.2*	BA.1.1 + BA.2	XR: 8/8, XAB: 2/48
XU	3	3 (1, 1, 1)	BA.1* + BA.2*	No main clade	No main clade
XM	4	47 (2, 1, **40**, 4)	BA.1.1* + BA.2*	BA.1.1 + BA.2	
XN	4	13 (2, 2, **7**, 2)	BA.1* + BA.2*	BA.1 + BA.2	
XAJ	4	5 (1, 1, **3**, 1)†	BA.2.12.1* + BA.4*	Unknown	BA.5: 1
XE	5	163 (3, **155**, 1, 3, 2)†	BA.1* + BA.2*	BA.1.17.2 + BA.2	
XZ	5	66 (1, 1, **61**, 2, 1)	BA.1* + BA.2*	BA.1 + BA.2.10	XAH: 10/12, XAD: 1/5
XAB	8	48 (1, **38**, 1, 2, 2, 2, 18, 2)†	BA.1* + BA.2*	BA.1 + BA.2.9	XAA: 8/8, XAG: 17/17, XU: 1/3, XQ: 2/12

Table S2: Summary of the Pango X-lineages in the Long ARG (excluding XP and XAK whose samples are entirely filtered out, see text). In cases of multiple origins, most X-lineages have a single “main clade” (in bold). Sc2ts inferred parents for the main clade are based on Nextclade designations, imputed where necessary. Official parents are taken from Pango designation alias key (TODO: explain the asterisks). Note that B.1.617.2 is the origin of the Delta VoC, which includes all AY.* classes. Additional descendants within the main clade are summarised giving their Nextclade Pango designation and clade count / total ARG count (the latter being omitted for non-recombinant designations).

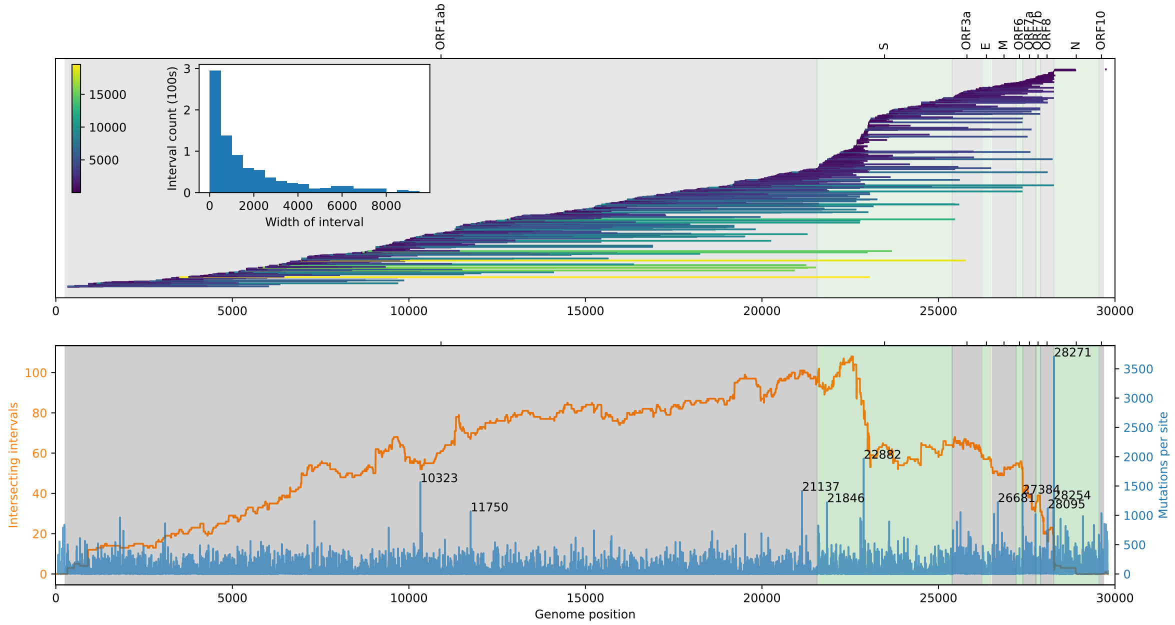


Figure S2: Distribution of recombination breakpoints and mutations along the genome in the Long ARG. Top panel shows the intervals for 851 breakpoints associated with 763 recombination nodes with at least two descending samples, plotted along the genome as line segments (coloured by interval width). Other details as described in Figure 4.

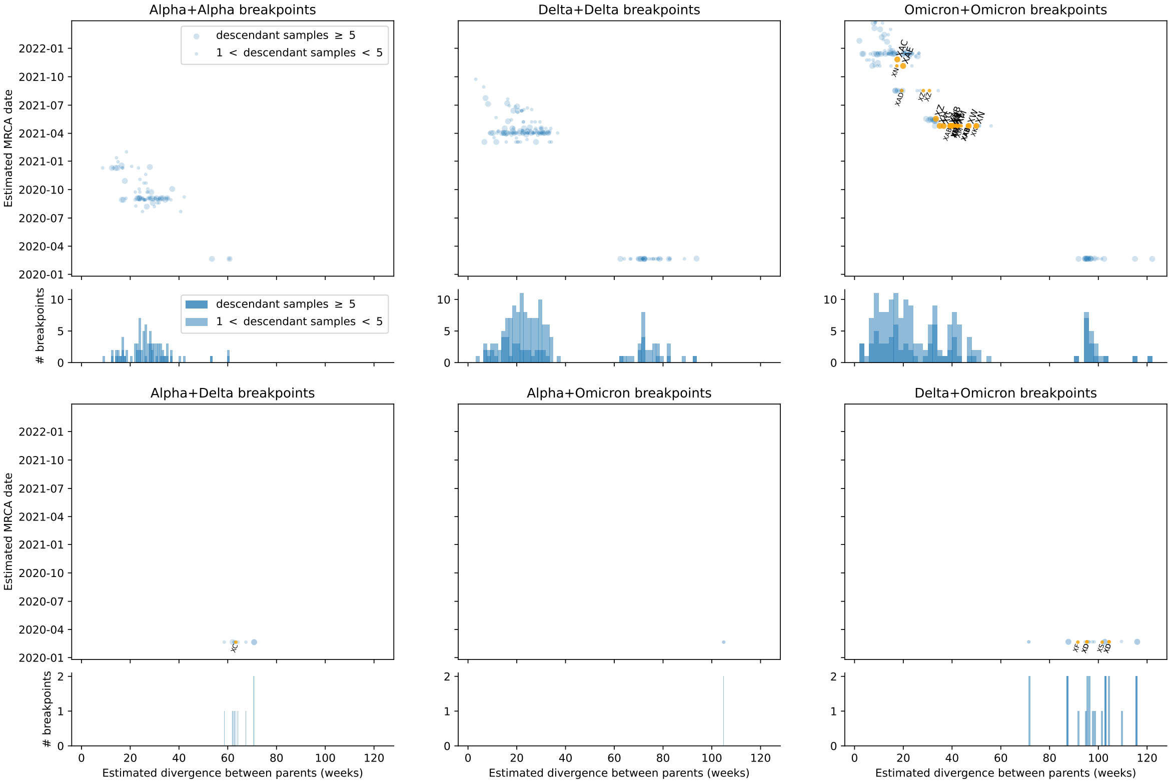


Figure S3: Divergence between parent lineages for recombination events within and among different VoC categories. There are 78 Alpha+Alpha recombination breakpoints corresponding to 75 recombination nodes (25 breakpoints / 24 nodes with ≥ 5 descendant samples). 148 breakpoints from 142 nodes are Delta+Delta recombinations (45 breakpoints / 45 nodes with ≥ 5 descendants), and 148 breakpoints from 142 nodes are Omicron+Omicron (71 breakpoints / 68 nodes with ≥ 5 descendants). The equivalent figures for Alpha+Delta are 9 / 8 (3 / 2), for Alpha+Omicron are 2 / 1 (0 / 0), and for Delta+Omicron are 20 / 13 (6 / 3). Note only recombination breakpoints involving lineages classified into Alpha, Delta, and Omicron VoC categories are plotted above: all other breakpoints are omitted.

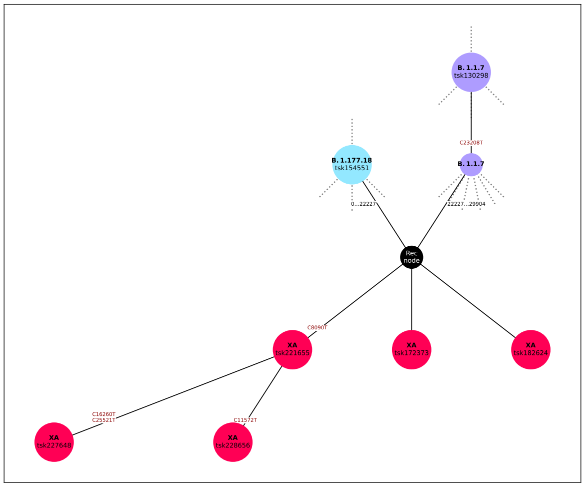


Figure S4: Detailed version of Figure 6A. All single nucleotide mutations are listed with the inherited nucleotide state, followed by the reference genome position, followed by the derived nucleotide state (after mutation). Recurrent mutations (see Section 4.1) are highlighted in bold, with reversions indicated by lowercase nucleotide letters. Sample nodes are shown with `tskit` IDs, which can be mapped to GISAID EPI ISL identifiers and strain names using supplementary file `Subgraph_sample_mapping.txt`. In contrast to Figure 6A, Pango lineages shown here are those assigned by GISAID rather than Nextclade; however, in the specific case of XA, Nextclade and GISAID exactly agree on the lineage designations.

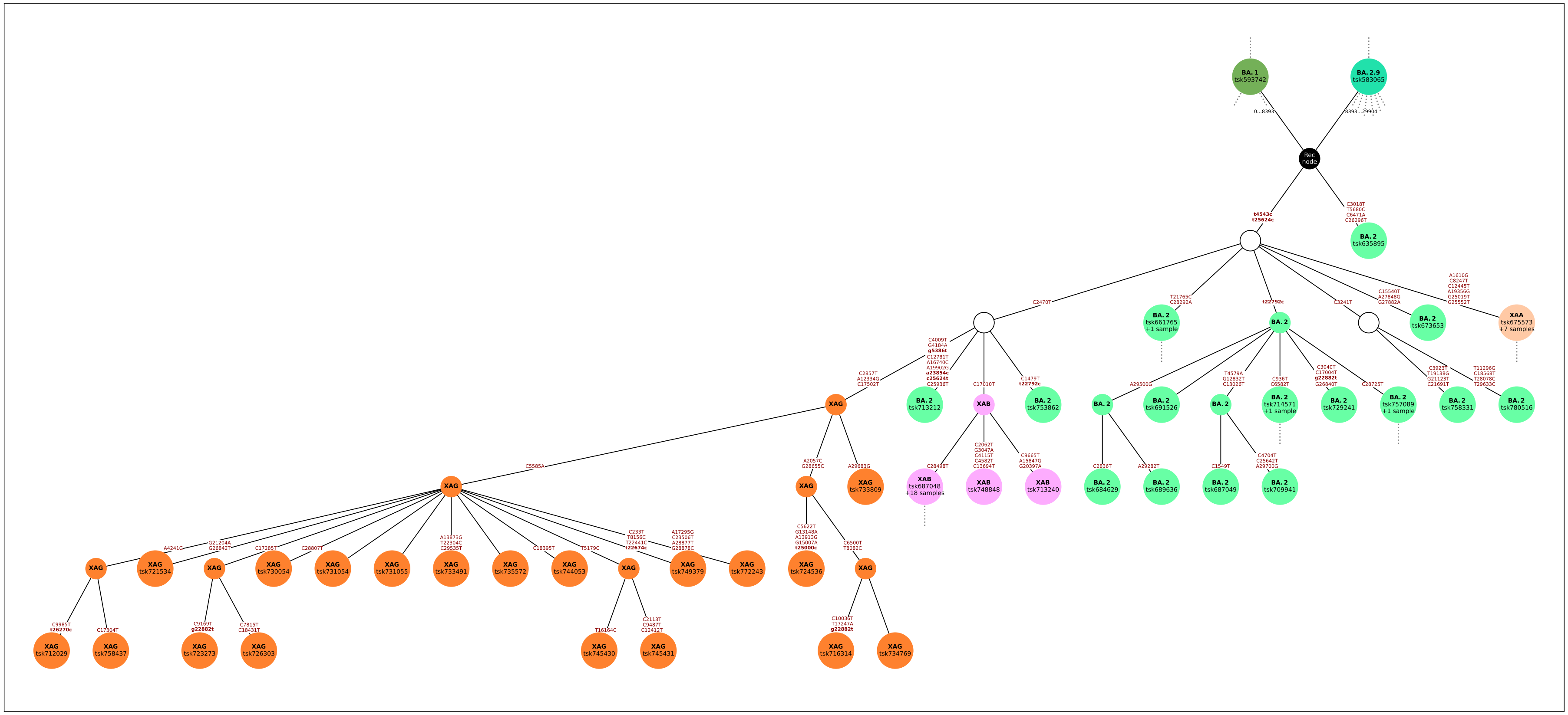


Figure S5: Detailed version of Figure 6B, with node and mutation labels as in Figure S4 (see supplementary file `Subgraph_sample_mapping.txt` to map tskID to EPI_ISL and strain name). Note that GISAID Pango designations assign an extra 2 samples to XAG, which are labelled as XAB by Nextclade in Figure 6B. This renders XAG fully monophyletic. However, unlike Nextclade, the GISAID designations also mark many of the samples in this subgraph as non-recombinants (designating them BA.2), and in general we find that Nextclade assignments agree more with our ARG structure than GISAID assignments.

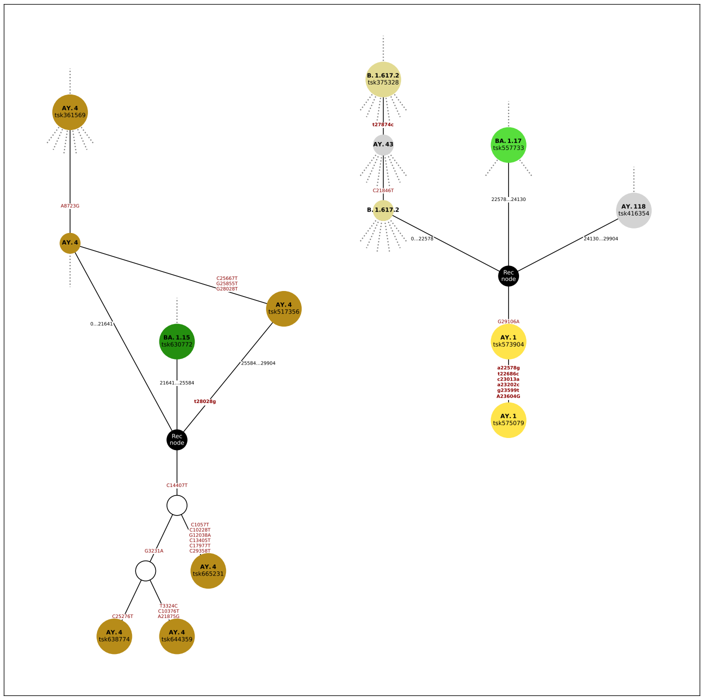


Figure S6: Detailed version of Figure 6C, with node and mutation labels as in Figure S4 (see supplementary file `Subgraph_sample_mapping.txt` to map tskID to EPI_ISL and strain name). Note that GISAID does not designate any nodes as XD in the Long ARG, hence no recombinant Pango lineages are marked in this plot. From inspection of the samples, we believe the GISAID designations to be erroneous in this case.

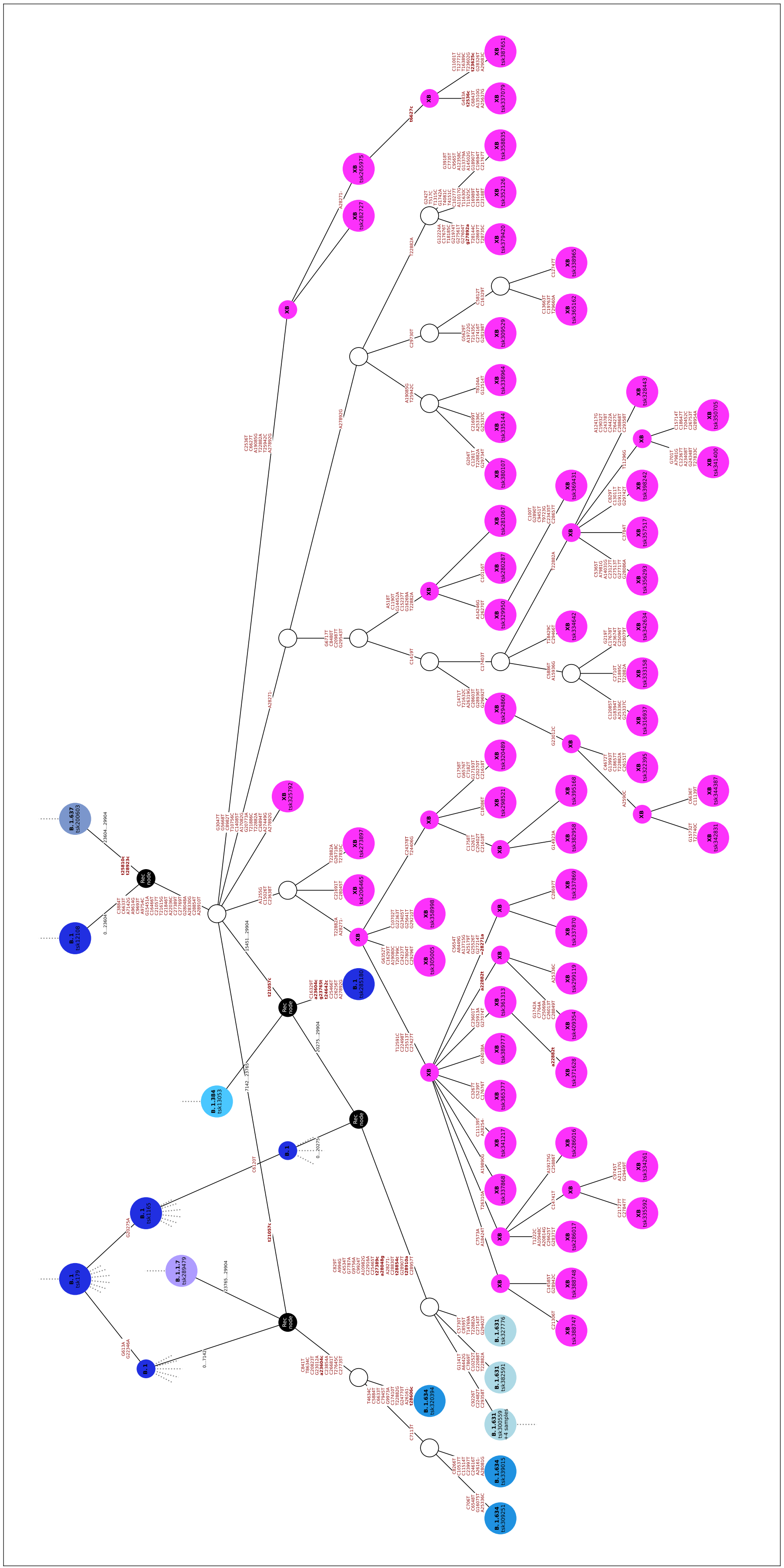


Figure S7: Detailed version of Figure 7, with node and mutation labels as in Figure S4 (see supplementary file `Subgraph_sample_mapping.txt` to map tskID to EPI_ISL and strain name). The GISAID XB designations agree with the Nextclade ones except in two cases: an unplotted singleton recombinant nested within the main grouping, and the tsk285180 node marked BA.1 in this plot, but which is labelled XB by Nextstrain.

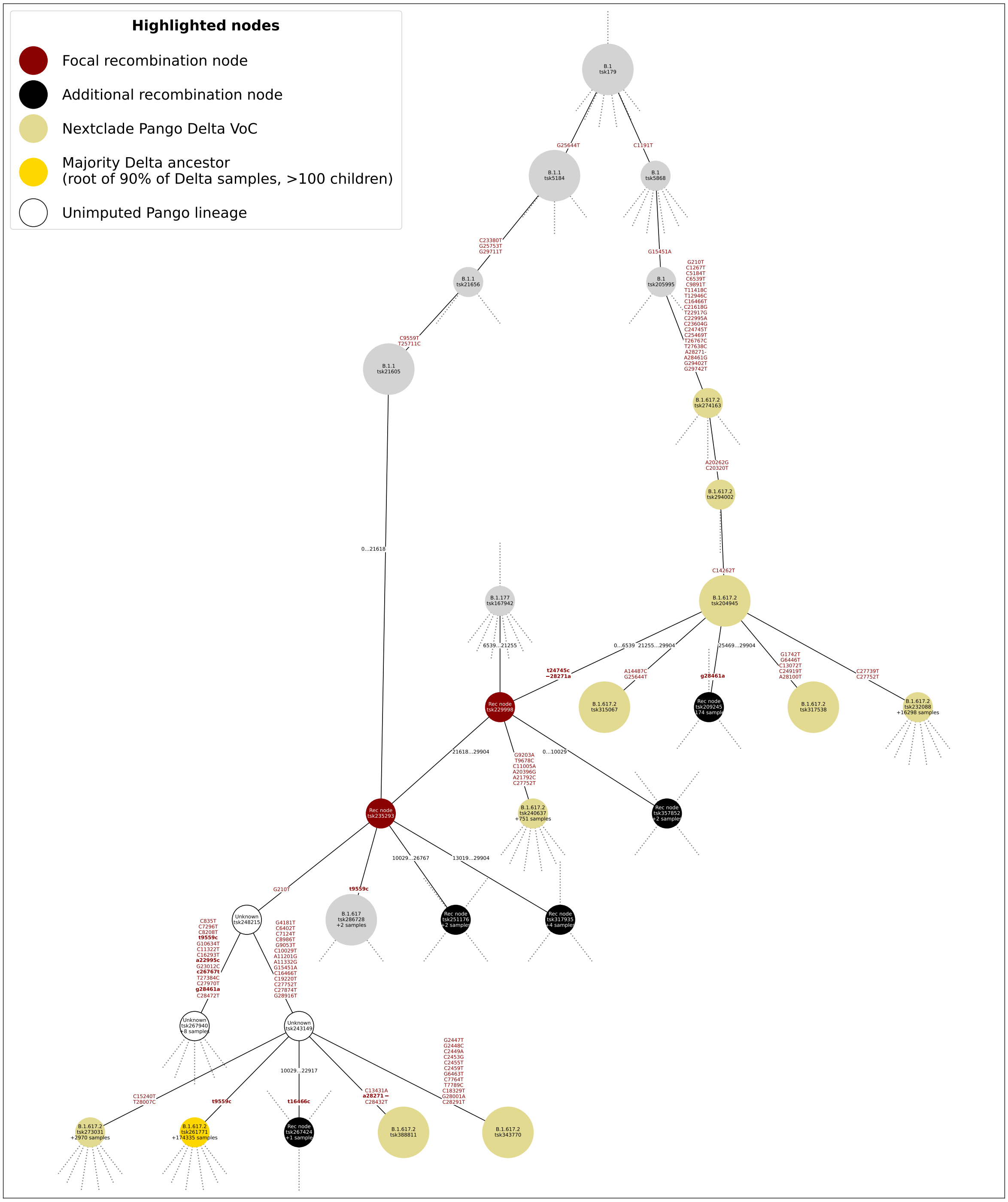

Highlighted nodes
● Focal recombination node
● Additional recombination node
● Nextclade Pango Delta VoC
● Majority Delta ancestor (root of 90% of Delta samples, >100 children)
○ Unimputed Pango lineage
B.1 tsk179
G25644T
C1191T
B.1.1 tsk5184
B.1 tsk5868
C23380T G25753T G29711T
G15451A
B.1.1 tsk21656
B.1 tsk205995
G210T C1267T C5184T C6539T C9891T T11418C T12946C C16466T C21618G T22917G C22995A C23604G C24745T C25469T T26767C T27638C A28271- A28461G G29402T G29742T
C9559T T25711C
B.1.1 tsk21605
B.1.617.2 tsk274163
A20262G C20320T
B.1.617.2 tsk294002
0...21618
C14262T
B.1.177 tsk167942
B.1.617.2 tsk204945
6539...21255
0...6539
21255...29904
25469...29904
t24745c −28271a
A14487C G25644T
g28461a
G1742T G6446T C13072T C24919T A28100T
C27739T C27752T
Rec node tsk229998
B.1.617.2 tsk315067
Rec node tsk209245 174 sample
B.1.617.2 tsk317538
B.1.617.2 tsk232088 +16298 samples
21618...29904
G9203A T9678C C11005A A20396G A21792C C27752T
0...10029
Rec node tsk235293
B.1.617.2 tsk240637 +751 samples
Rec node tsk357852 +2 samples
10029...26767
13019...29904
G210T
t9559c
Unknown tsk248215
B.1.617 tsk286728 +2 samples
Rec node tsk251176 +2 samples
Rec node tsk317935 +4 samples
C835T C7296T C8208T t9559c G10634T C11322T C16293T a22995c G23012C c26767t T27384C C27970T g28461a C28472T
G4181T C6402T C7124T C8986T G9053T C10029T A11201G A11332G G15451A C16466T C19220T C27752T C27874T G28916T
Unknown tsk267940 +8 samples
Unknown tsk243149
10029...22917
G2447T G2448C C2449A C2453G C2455T C2459T G6463T C7764T T7789C C18329T G28001A C28291T
C15240T T28007C
t9559c
t16466c
C13431A a28271 − C28432T
B.1.617.2 tsk273031 +2970 samples
B.1.617.2 tsk261771 +174335 samples
Rec node tsk267424 +1 sample
B.1.617.2 tsk388811
B.1.617.2 tsk343770

Figure S8: Subgraph of the Long ARG, focusing on two likely false positive recombination nodes at the start of the Delta wave ("focal" nodes, in red, corresponding to rows A and B in Table 3). The path to node `tsk261771` has been expanded: this node (in gold) is the ancestor of ∼89.8% of Delta samples in the Long ARG, and represents a large polytomy with 107 immediate children. Of the remaining Delta samples, most (8.4%) are descendants of the node `tsk232088` on the far right, a sibling of the earliest focal recombination node. The suspected-incorrect recombination path to an early B.1 MRCA is also shown. Note the large amount of additional recombination (black nodes) among close descendants of the focal nodes.