Hidden Markov Models Detect Recombination and Ancestry of SARS-CoV-2

Nobuaki Masaki[1,2] **and Trevor Bedford**[2,3]

[1]Department of Biostatistics, University of Washington, Seattle, WA, [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, [3]Howard Hughes Medical Institute, Seattle, WA

Abstract

When individuals are co-infected with distinct SARS-CoV-2 lineages, homologous recombination can generate mosaic genomes carrying mutations from both parental lineages. A variety of methods exist to detect recombinant sequences and their parental lineages in surveillance-scale datasets comprised of millions of SARS-CoV-2 genomes. However, these methods often rely on user-defined settings, such as the probability that a recombination breakpoint occurs between adjacent positions on the query sequence. In this study, we devise a hidden Markov model that detects recombinant SARS-CoV-2 sequences and identifies their parental lineages within a test set of sequences. Our method does not depend on user-defined parameters and can accommodate de novo mutations on the query sequence that are not present in the predicted parental lineages. To achieve this, we use maximum likelihood to estimate parameters that characterize the transition and emission probabilities in our hidden Markov model. Applying our method to 440,307 SARS-CoV-2 sequences sampled in England between September 2020 and March 2024, we detect 7,619 recombinant sequences corresponding to 1.73% (95% CI: [1.69%, 1.77%]) of all sampled sequences. We show a positive association between the proportion of query sequences detected as recombinant in each week and community SARS-CoV-2 prevalence. This is consistent with higher prevalence increasing co-infection risk and promoting the emergence of recombinant sequences. We further observe localized clusters of recombination breakpoints within spike and in intergenic regions.

1 Introduction

Recombination is thought to occur in coronaviruses via a copy-choice mechanism in which the viral RNA-dependent RNA polymerase switches template strands during negative strand synthesis [1]. When hosts are co-infected by multiple SARS-CoV-2 lineages, this template switching results in recombinant genomes sharing genetic material from both lineages [2].

One of the most significant recombinant lineages that emerged during the pandemic is XBB. Phylogenetic analysis indicates that this lineage was derived from a recombination event between two Omicron lineages (BJ.1 and BM.1.1.1) and resulted in significant reduction in neutralization from human serum samples [3]. The effective reproduction number of XBB was estimated to be 1.23 and 1.20 times higher than its parental lineages BJ.1 and BM.1.1.1, respectively, using epidemic data from late 2022 [3]. The derived lineage XBB.1.5 spread widely and reached a peak frequency of 55% globally in epidemiological week 12 of 2023 [4]. Because recombination can combine mutations from different SARS-CoV-2 lineages that jointly confer a growth advantage, systematic surveillance and robust statistical detection of recombinant lineages are crucial.

A wide range of computational approaches have been developed to detect recombination in viruses. Broadly, similarity methods such as SimPlot visualize how a query sequence's similarity shifts across the genome relative to putative parental lineages [5, 6]. RDP4 examines all triplets within a set of sequences and applies a suite of tests (e.g., GENECONV, MaxChi, Bootscan, 3SEQ) to detect recombination breakpoints and assign parental sequences [7–10]. However, the number of comparisons is cubic with respect to the sample size, which is infeasible for large-scale datasets.

Phylogeny-based methods such as GARD detect breakpoints by fitting independent phylogenies to alignment segments and comparing model fit across candidate partitions [11]. The repeated tree-fitting and model-comparison steps are computationally intensive, so GARD is generally applied to downsampled alignments instead of surveillance-scale datasets comprising millions of genomes.

More recently, SARS-CoV-2-specific tools have been designed to operate on surveillance-scale datasets. Bolotie uses a hidden Markov model (HMM) where the latent states represent SARS-CoV-2 lineages [12]. The Viterbi algorithm is used to assign a parental lineage to each position. RIPPLES identifies candidate recombinant sequences by scanning a global mutation-annotated phylogeny for unusually long branches that may represent recombination events [13]. For each candidate recombinant sequence, RIPPLES partitions the genome into multiple segments and re-places each onto the global phylogeny using maximum parsimony. RecombinHunt compares segment-wise mutation patterns on a query sequence to lineage-specific profiles [14]. It constructs a cumulative likelihood profile across the genome and uses the Akaike information criterion to choose between three models with zero, one, or two breakpoints.

Although these SARS-CoV-2–specific tools can be applied to surveillance-scale datasets, each has method-specific limitations. Bolotie's HMM does not model de novo mutations or genotyping errors, which can result in spurious state switches when the query sequence

harbors mutations absent from the mutation profile of its true lineage. The HMM's transition probability is also user-specified, making breakpoint detection sensitive to this choice. RIPPLES relies on a mutation-annotated phylogeny. Uneven sampling and sequencing artifacts can inflate or deflate the long-branch signal used to identify candidate recombinants. Moreover, the threshold for the long-branch signal is defined by the user, and the initial candidate set of recombinant sequences is sensitive to this chosen cutoff. RecombinHunt relies on several hard evidence gates (e.g., declaring a genome non-recombinant when it differs from the most likely lineage by two or fewer mutations). These thresholds are likewise sensitive to de novo mutations and genotyping errors. Finally, both RIPPLES and RecombinHunt permit at most two breakpoints, even though recombinant lineages with more breakpoints have been detected.

In this paper, we develop a method to detect recombinant SARS-CoV-2 sequences within a test set of sequences collected over a short interval (a few days to a week). Our method employs an HMM inspired by the Li and Stephens model [15] that accounts for de novo mutations and genotyping errors in both recombinant and non-recombinant sequences. For each test sequence, we estimate a pseudo-frequency for observing alleles absent from the parental lineage at a position, and the lineage-transition probability between consecutive sites. We implement an efficient version of the forward algorithm to speed up estimation of these frequencies (see Supplementary Materials). We then predict the local Pango lineage ancestry, defined as the sequence of Pango lineages ancestral at each genomic position of the test sequence, using lineage-specific nucleotide frequencies computed from prior sequences. We classify a test sequence as a recombinant if the predicted local Pango lineage ancestry contains one or more lineage transitions. Our method does not rely on a phylogeny or any user-defined parameters, and can accommodate any number of breakpoints.

We evaluate performance in a simulation where we generated synthetic recombinant and control genomes from SARS-CoV-2 sequences sampled between January and March 2022. We also apply our method to 440,307 SARS-CoV-2 genomes from GenBank, sampled in England between September 2020 and March 2024 [16] to identify recombinant sequences and measure their prevalence through time and occurrence across lineages.

2 Materials and methods

Figure 1 summarizes our workflow for detecting the local Pango lineage ancestry of SARS-CoV-2 sequences from GenBank. In this section, we describe each component of our method in detail.

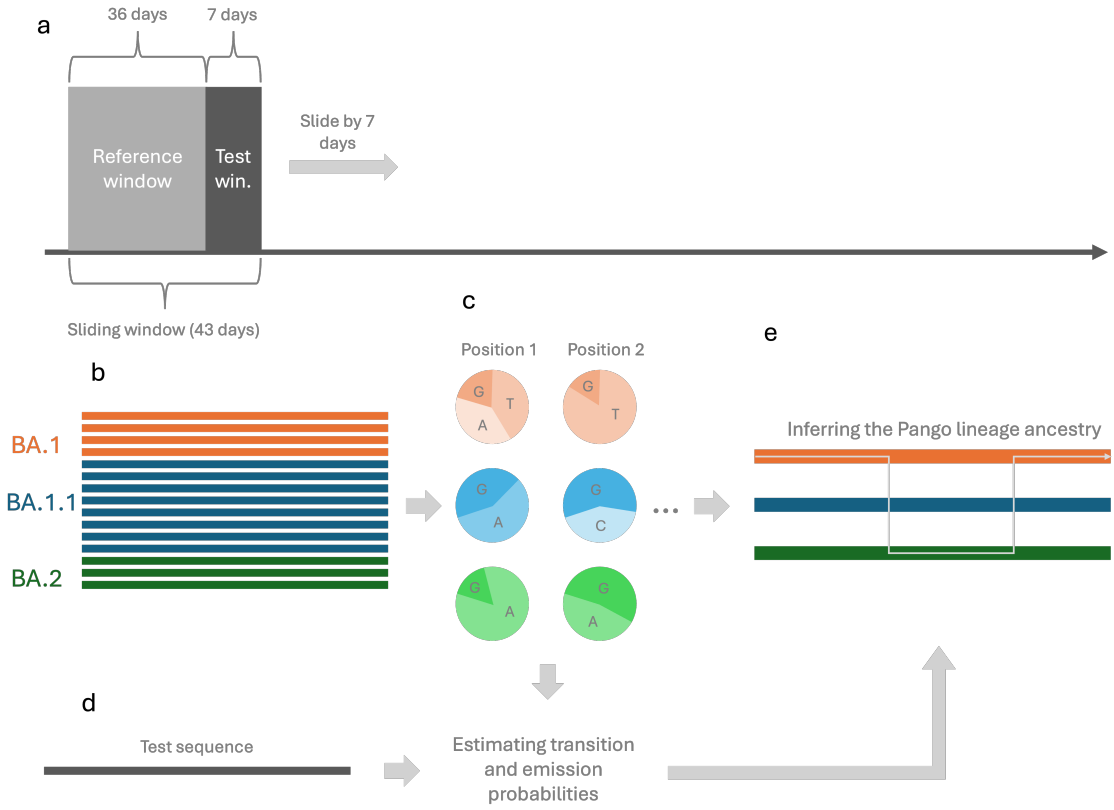

Figure 1. Overview of methods. (a) We first organize SARS-CoV-2 sequences collected in England between September 2020 and March 2024 into sliding windows of 43 days, which are advanced by 7 day increments. (b) In each sliding window, the first 36 days and last 7 days respectively comprise the reference window and test window. Sequences collected during the reference window comprise the reference set of sequences containing the mutational profile of each Pango lineage. (c) We next calculate the nucleotide frequency matrix containing per-position allele frequencies for each Pango lineage in the reference set. (d) For each test sequence collected during the test window, we use maximum likelihood to estimate frequencies that parameterize the transition and emission probabilities of our HMM. (e) We then use the Viterbi algorithm to predict the local Pango lineage ancestry for this sequence (panel e).

2.1 Obtaining SARS-CoV-2 sequences and clustering Pango lineages

We obtained SARS-CoV-2 sequences and metadata from GenBank, processed using the Nextstrain pipeline [17]. After filtering for sequences collected in England between September 2020 and March 2024, we clustered Pango lineages based on their sequence count. We collapsed any Pango lineage with fewer than 10,000 sequences into its parental lineage, using unaliased Pango lineage names [18]. This was done iteratively to ensure that all collapsed Pango lineages contained at least 10,000 sequences. Lineages without a defined parent were grouped into a shared "other" category. We collapsed 2,304 Pango lineages that existed during this period to 41 collapsed lineages (including the "other" category). Unless otherwise specified, all mentions of Pango lineages refer to the collapsed lineages resulting from this procedure.

2.2 Reference and test sets

From the sequences collected in England between September 2020 and March 2024, we generated sliding windows of reference and test set pairs. Each sliding window consisted of 43 days, and these windows were incremented by 7 days at a time to generate 185 sliding windows.

In each 43-day sliding window, the first 36 days and last 7 days respectively comprise the reference window and test window. Sequences collected during the reference and test windows respectively comprise the reference and test sets for this sliding window.

If more than 100,000 sequences were available during the reference window, we drew a random sample of 100,000 sequences and used this as the reference set. Similarly, if more than 3,000 sequences were available during the test window, we drew a random sample of 3,000 sequences and used this as the test set.

This process results in 185 pairs of reference and test sets. In the following sections, we describe our process for calculating the nucleotide frequency matrix for each reference set. We then define our HMM, which uses the nucleotide frequency matrix to predict the local Pango lineage ancestry for each sequence in the paired test set.

2.3 Calculating the nucleotide frequency matrix

For each of the 185 reference sets, we calculated a nucleotide frequency matrix that contains the frequency of each nucleotide (A, C, G, and T) at every genomic position for each Pango lineage. Nucleotide frequencies were calculated by dividing nucleotide counts at each position by the total sequence count within each Pango lineage. When calculating frequencies, we excluded all non-standard nucleotides (i.e., those other than A, C, G, and T). If no sequences in a Pango lineage carried any of the standard nucleotides at a position, we assigned equal probabilities (0.25 each) to A, C, G, and T.

2.4 Predicting the local Pango lineage ancestry

The local Pango lineage ancestry of a SARS-CoV-2 sequence refers to the sequence of Pango lineages ancestral to each genomic position of the SARS-CoV-2 sequence.

If a sequence derives from a recombination event between two sequences in two distinct Pango lineages, its local Pango lineage ancestry will consist of segments from these distinct lineages, with transitions between segments marking recombination breakpoints. Conversely, for non-recombinant sequences, the local Pango lineage ancestry will only contain a single parental lineage. It is important to note that the true local Pango lineage ancestry of a sequence in a test set is defined in relation to the Pango lineages present in the paired reference set. For example, lineages L_1, L_2, and L_3 may all be present in the paired reference set, with lineage L_3 arising from a recombination event between two sequences in lineages L_1 and L_2 respectively. Suppose that there is a sequence from lineage L_3 in the test set. In this case, the true local Pango lineage ancestry of this sequence will have L_3 as the lineage contributing ancestry at all genomic positions.

We predict the local Pango lineage ancestry of all sequences in each test set using the

nucleotide frequency matrix calculated from the corresponding reference set and an HMM inspired by the Li and Stephens model [15].

2.5 Hidden Markov model to predict local Pango lineage ancestry

This HMM jointly models the latent local Pango lineage ancestry and the observed nucleotide sequence for each test sequence. It does so by considering three key components: (i) the probability of each lineage providing ancestry at the first position (initial state probabilities), (ii) the probability of transitioning between parental lineages from one position to the next (transition probabilities), and (iii) the probability of observing each nucleotide at a given position, conditional on the parental lineage (emission probabilities). Transitions between lineages correspond to recombination events.

Here, we define the HMM used to predict the local Pango lineage ancestry of a sequence in any given test set. We henceforth refer to this sequence as our test sequence. Let the genome length be denoted by N and let $t \in \{1, 2, \ldots, N\}$ index genomic positions. Our sequences are aligned, so all of our sequences have length N.

For our test sequence, we define the random variable for the parental Pango lineage at position t as Z_t. Z_t is supported on $\{1, 2, \ldots, M\}$, where M is the number of distinct Pango lineages contained in the paired reference set (the reference set paired with the test set from which the test sequence is drawn). Each value of $\{1, 2, \ldots, M\}$ corresponds to one of these Pango lineages.

We further define, for the test sequence, the random variable for the observed nucleotide at position t as O_t. O_t is supported on $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

In the following sections, we define the three key components of this HMM, which are the initial state probabilities, the transition probabilities, and the emission probabilities.

2.5.1 Initial state probabilities

The initial state probabilities give the probability of each Pango lineage being the parental lineage of the test sequence at the first genomic position. We define the initial state probability of Pango lineage i ($i \in \{1, 2, \ldots, M\}$) as $\pi_i = P(Z_1 = i)$. In our model, we set π_i to the frequency of lineage i in the paired reference set. Let n_i be the number of sequences assigned to lineage i in the reference set, and let $n_{\text{total}} = \sum_{j=1}^{M} n_j$ be the total number of sequences across all M lineages. Then,

$$\pi_i = \frac{n_i}{n_{\text{total}}}, \quad i \in \{1, 2, \ldots, M\}.$$

2.5.2 Transition probabilities

The transition probabilities give the probability of transitioning from one parental Pango lineage to another between consecutive positions on the test sequence. Here, transitions between Pango lineages correspond to recombination breakpoints. We define the transition probability from Pango lineage i to Pango lineage j as

$$a_{ij} = P(Z_{t+1} = j \mid Z_t = i), \quad i, j \in \{1, 2, \ldots, M\}, \quad t \in \{1, 2, \ldots, N-1\}.$$

Here, a_{ij} represents the probability that the parental Pango lineage of the test sequence changes from i to j between any consecutive positions on the genome. In our model, we set transition probabilities as

$$a_{ij} = \begin{cases} 1 - \lambda, & \text{if } i = j, \\ \frac{\lambda}{M-1}, & \text{if } i \neq j. \end{cases}$$

λ is the probability that there is a recombination breakpoint between consecutive positions on the genome. For the above formulation, we also assume that transitions between any two Pango lineages $i \neq j$ occur with the same probability. Because λ is an unknown parameter, we later describe our method for estimating λ.

2.5.3 Emission probabilities

The emission probabilities give the probability of observing each nucleotide (i.e., A, C, G, or T) at a particular position on the test sequence, conditional on the parental Pango lineage at that position. We define the emission probability of observing nucleotide k at position t, conditional on the parental Pango lineage being i at position t, as

$$b_{i,t}(k) = P(O_t = k | Z_t = i), \quad k \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}, \quad i \in \{1, 2, \ldots, M\}, \quad t \in \{1, 2, \ldots, N\}.$$

$b_{i,t}(k)$ depends on the nucleotide frequency matrix calculated from the paired reference set. We use $f_{i,t}(k)$ to denote the frequency of nucleotide k at position t in Pango lineage i in the paired reference set. To adjust for possible mutations and genotyping errors that could occur on the test sequence, we apply a pseudo-frequency ϵ. Specifically, we let

$$b_{i,t}(k) = \frac{f_{i,t}(k) + \epsilon}{1 + 4\epsilon}.$$

The pseudo-frequency ϵ assigns a non-zero probability of observing a nucleotide at position t, when the parental Pango lineage at t contains no sequences that have this nucleotide at t in the reference set. We want to allow for this non-zero probability in case the test sequence has acquired a mutation (or genotyping error) at position t that leads to an observed nucleotide that is not contained in the parental Pango lineage. A small value of ϵ allows occasional mutations or genotyping errors without forcing a lineage switch in the predicted local Pango lineage ancestry. Because ϵ is an unknown parameter, we describe our method for estimating ϵ in the following section.

We assume that positions with non-ACGT calls contain no information about the true nucleotide. Thus, we assign an emission probability of one to non-ACGT calls across all parental Pango lineages.

Symbol	Description
N	Genome length
M	Number of Pango lineages in the reference set
Z_t	Parental Pango lineage at position t
O_t	Observed nucleotide at position t
λ	Transition probability
ϵ	Pseudo-frequency for emissions (accounts for mutations and genotyping errors)
a_{ij}	Transition probability from lineage i to lineage j
$b_{i,t}(k)$	Emission probability of nucleotide k at t given $Z_t = i$
$f_{i,t}(k)$	Frequency of nucleotide k at t in lineage i in the reference set
π_i	Initial state probability that $Z_1 = i$

Table 1. Summary of symbols used in the HMM.

2.6 Maximum likelihood estimation of parameters in the hidden Markov model

We have two unknown parameters in our HMM. λ is the probability that the parental Pango lineage changes between consecutive positions and ϵ is our pseudo-frequency, which adjusts emission probabilities to accommodate mutations or genotyping errors on the test sequence.

To perform maximum likelihood estimation on these two parameters, we first obtain the probability of the observed nucleotide sequence of the test sequence, conditional on these two parameters. In this section, we describe the procedure we use to obtain this probability.

Using the transition and emission probabilities described in the previous sections, it is relatively straightforward to obtain the joint probability of a candidate local Pango lineage ancestry and the observed nucleotide sequence for a test sequence. Let $i_{1:N} = (i_1, i_2, \ldots, i_N) \in [M]^N$ be a candidate local Pango lineage ancestry and $k_{1:N} = (k_1, k_2, \ldots, k_N) \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^N$ be the observed nucleotide sequence of this test sequence. Finally, let $Z_{1:N} = (Z_1, Z_2, \ldots, Z_N)$ and $O_{1:N} = (O_1, O_2, \ldots, O_N)$. Then,

$$P(Z_{1:N} = i_{1:N}, O_{1:N} = k_{1:N} \mid \lambda, \epsilon) = \pi_{i_1} \left(\prod_{t=1}^{N-1} a_{i_t i_{t+1}} \right) \left(\prod_{t=1}^{N} b_{i_t,t}(k_t) \right).$$

To obtain the marginal probability of the observed nucleotide sequence, we can simply sum up this joint probability across all possible local Pango lineage ancestries, as shown below.

$$P(O_{1:N} = k_{1:N} \mid \lambda, \epsilon) = \sum_{i_{1:N} \in [M]^N} \pi_{i_1} \left(\prod_{t=1}^{N-1} a_{i_t i_{t+1}} \right) \left(\prod_{t=1}^{N} b_{i_t,t}(k_t) \right),$$

This procedure can be done efficiently using the forward algorithm described by Rabiner [19]. We implemented a fast version of this forward algorithm that computes the induction step in $\mathcal{O}(M)$ time compared to the normal $\mathcal{O}(M^2)$ time (see Supplementary Materials). We can maximize this marginal probability with respect to our two parameters to obtain our maximum likelihood estimates, as shown below.

$$\hat{\lambda}, \hat{\epsilon} = \arg\max_{\lambda,\epsilon} P(O_{1:N} = k_{1:N} \mid \lambda, \epsilon).$$

Maximum likelihood estimation of λ and ϵ is done for each test sequence. Optimization was carried out with the limited-memory BFGS algorithm subject to box constraints, using `scipy.optimize.minimize` (`method = "L-BFGS-B"`) [20]. Our version of the forward algorithm results in a large reduction in computation time, because the marginal likelihood is evaluated repeatedly during L-BFGS-B optimization.

During numerical optimization, we reparameterize λ to $\tau = \lambda(N-1)$, which represents the expected number of transitions for the test sequence. Furthermore, we optimized ϵ on the log scale and later exponentiated to obtain our estimate in the original scale. The search was initialized at $(\log(\epsilon), \tau) = (\log(0.005), 1)$ and restricted to the intervals $\log(\epsilon) \in [\log(10^{-8}), \log(0.02)]$ and $\tau \in [0, 3]$. The reparameterization of λ to τ was done to avoid possible numerical instabilities that might arise when trying to optimize λ directly, because we expect λ to be close to zero. We similarly optimized ϵ in the log scale because we expect ϵ to be close to zero.

We chose the upper bound of three for τ because most discovered recombinant lineages were detected to have three or fewer breakpoints. However, this does not prevent the predicted local Pango lineage ancestry from having more than three breakpoints.

2.7 Obtaining the most likely sequence of Pango lineage ancestry

We apply the Viterbi algorithm [19] to each test sequence to obtain the most probable sequence of parental Pango lineage states along the genome,

$$\hat{i}_{1:N} = (\hat{i}_1, \ldots, \hat{i}_N).$$

This represents the predicted local Pango lineage ancestry for this test sequence. Transitions between Pango lineage ancestries represent predicted recombination breakpoints.

When applying the Viterbi algorithm, we use our maximum likelihood estimates of the two frequencies, λ and ϵ, described in earlier sections. Specifically, we compute the sequence of Pango lineage ancestry that maximizes the joint probability of the ancestry path and the observed nucleotide sequence, given our maximum likelihood estimates. In other words,

$$\hat{i}_{1:N} = \arg\max_{i_{1:N} \in [M]^N} P(Z_{1:N} = i_{1:N}, O_{1:N} = k_{1:N} \mid \hat{\lambda}, \hat{\epsilon}).$$

Note that because $P(O_{1:N} = k_{1:N} \mid \hat{\lambda}, \hat{\epsilon})$ does not depend on the local Pango lineage ancestry, the above is equivalent to maximizing the posterior probability of the local Pango lineage ancestry given the observed nucleotide sequence and our maximum likelihood estimates. In other words,

$$\hat{i}_{1:N} = \underset{i_{1:N} \in [M]^N}{\arg\max}\, P(Z_{1:N} = i_{1:N} \mid O_{1:N} = k_{1:N}, \hat{\lambda}, \hat{\epsilon}).$$

2.8 Simulation study

To assess our method's ability to detect recombination and accurately predict local Pango lineage ancestries, we conducted a simulation study using synthetic SARS-CoV-2 sequences with known local Pango lineage ancestries. These synthetic sequences were generated from real SARS-CoV-2 genomes.

We generated these synthetic sequences using the reference set comprised of 14,599 SARS-CoV-2 sequences collected in England between November 6, 2022 and December 11, 2022. We simulated 1,000 recombinant sequences with two parental lineages and 1,000 control sequences with one parental lineage.

We generated 500 recombinant sequences using a single recombination breakpoint. To generate these sequences, we randomly sampled two parental sequences from two different Pango lineages in the reference set and copied nucleotides from one parent up to a breakpoint randomly chosen on the genome, and from the other parent thereafter. We generated the remaining 500 recombinant sequences using two breakpoints. For these sequences, we again sampled two parental sequences from different Pango lineages. We chose two breakpoints randomly from all possible breakpoint combinations on the genome and inserted a middle segment from one sequence between these breakpoints, replacing the corresponding region in the genome of the other sequence. If a synthetic recombinant sequence was identical to or differed by only one mutation from one of its parental sequences, we discarded this sequence and repeated the sequence generation process. We obtained 1,000 control sequences by sampling 1,000 sequences at random from the reference set.

To mimic mutations on all 2,000 synthetic sequences, we drew the mutation count from an empirical distribution obtained by tallying nucleotide substitution counts on each branch of a tree containing one tip per Pango lineage [17]. The empirical distribution of nucleotide substitution counts was right-skewed (n = 567; median = 2 [IQR 1–4]; mean = 3.45; 95th percentile = 8; range 1–70). Given the drawn mutation count m, we sampled m genomic positions uniformly at random and replaced the existing nucleotide at each position with one of the other four nucleotides (A, C, G, T, or N, excluding the original base) chosen at random. In our aligned sequences, N represents an unknown nucleotide.

For each of the 2,000 synthetic sequences, we applied the method described in Section 2.6 to estimate the transition probability λ and pseudo-frequency ϵ. We then predicted the local Pango lineage ancestry for each sequence using the method described in Section 2.7. Emission probabilities for the HMM were based on the nucleotide frequency matrix calculated from the reference set of sequences collected between November 6, 2022 and December 11, 2022.

To evaluate performance, we conducted several quantitative assessments. First, we estimated the sensitivity and specificity of our method for classifying sequences as recombinant or non-recombinant. We classified a synthetic sequence as a recombinant if the predicted local Pango lineage ancestry contained at least one lineage transition.

Second, we calculated the mean position-by-position accuracy of the predicted local Pango lineage ancestry across synthetic sequences by comparing the predicted parental lineage at each genomic position to the true parental lineage.

Third, we assessed whether the correct parental lineage (for control sequences) or lineage pair (for recombinant sequences) was correctly recovered. We calculated the proportion of synthetic sequences for which the parental lineage or lineage pair was correctly recovered. Both the mean position-by-position accuracy and the recovery rate of parental lineages were calculated separately for recombinant and non-recombinant sequences.

For the sensitivity, specificity, and recovery rate of parental Pango lineages, we report 95% exact binomial confidence intervals. For mean position-by-position accuracy, we calculated 95% bootstrap confidence intervals by sampling 500 times with replacement from synthetic sequences (either from the set of recombinants or the set of non-recombinant sequences), calculating the mean position-by-position accuracy in each bootstrap sample, and taking the 2.5th and 97.5th percentiles of the bootstrapped estimates.

We next quantified the association between the sensitivity to detect recombinants $s(d)$ and the genome-wide Hamming distance d between the parental sequences of recombinants using a logistic regression, which can be written as,

$$\text{logit}\left(s(d)\right) = \beta_0 + \beta_1 d.$$

It follows that $\exp(\beta_1)$ represents the multiplicative difference in the odds of detection for two synthetic recombinant sequences whose parental Hamming distances are one unit apart. To fit this logistic regression, we used the HMM's predicted label for all 1,000 synthetic recombinants (1 if the model detected the sequence as a recombinant and 0 otherwise) as the outcome.

To quantify the accuracy of predicted breakpoint positions, we calculated the distance between each predicted breakpoint position and its corresponding true genomic position. We restricted analysis to synthetic recombinants whose detected breakpoint count matched the number of true breakpoints. For synthetic recombinants with one true breakpoint, we calculated the distance between the true and detected breakpoint position. For synthetic recombinants with two true breakpoints, we ordered true and detected breakpoint positions 5' to 3', paired them positionally (first with first, second with second), and calculated the distance between each pair. We then calculated the mean breakpoint distance separately for recombinants with one and two breakpoints.

We obtained 95% confidence intervals for the mean breakpoint distance via nonparametric bootstrap. Specifically, we sampled recombinant sequences with replacement 500 times within each stratum (one and two breakpoints), calculated the mean breakpoint distance

for each bootstrap sample, and took the 2.5th and 97.5th percentiles of the bootstrapped estimates within each stratum. When detected and true breakpoints counts differed, we did not compute a distance. Instead we recorded the detected and true breakpoint counts per sequence and summarized mismatches in a contingency table.

2.9 Empirical data analysis

We applied our method to the full set of SARS-CoV-2 sequences collected in England between September 2020 and March 2024. We describe how we obtain these sequences in Section 2.1. As described in Section 2.2, sequences were divided into temporally matched reference and test sets using a 43-day sliding window. For each sequence in each test set, we estimated the transition probability λ and pseudo-frequency ϵ using maximum likelihood (see Section 2.6). We then predicted the local Pango lineage ancestry for each sequence using the Viterbi algorithm (see Section 2.7). There were 440,307 sequences across all test sets.

We classified any sequence with one or more lineage transitions in their predicted local Pango lineage ancestry as recombinant. For each 7-day test window, we calculated the estimated recombinant proportion, or the number of detected recombinants divided by the total number of tested sequences from this window. Recall that the total number of tested sequences can vary across test windows (see Section 2.2).

We hypothesized that the estimated recombinant proportion would be positively associated with community SARS-CoV-2 prevalence across test windows. This is because co-infection by two distinct Pango lineages, which is required for the emergence of detectable recombinant sequences, occurs more frequently when community prevalence of SARS-CoV-2 is high. To evaluate our hypothesis, we compared the estimated recombinant proportion in each test window with community prevalence estimated from the UK Office for National Statistics (ONS) Coronavirus Infection Survey [21]. ONS provides prevalence estimates by date. For comparability, we computed the mean ONS prevalence estimate within each test window and used this window-averaged prevalence estimate in our analysis.

We further counted the number of detected recombinants across all windows, stratified by unique parental lineage pairs (e.g., BA.1.1–BA.2), and compared this to our estimated expected counts for each parental lineage pair based on co-infection dynamics, derived in Section 2.10. Finally, we aggregated predicted breakpoint positions across all detected recombinants to obtain the empirical genome-wide distribution of breakpoint positions.

2.10 Expected recombinant counts

In this section, we derive the expected number of detected recombinants with each parental lineage pair. For each lineage pair, this expected count depends on the lineage frequencies of the two parental lineages, community SARS-CoV-2 prevalence, and the number of sequences tested in each test window.

We denote the lineage frequency of lineage i in test window w, or the proportion of infections attributable to lineage i among all SARS-CoV-2 infections during w, as $p_i(w)$. Note that $p_i(w)$ is distinct from the lineage prevalence, which is the overall fraction of

the population infected by lineage i. We further denote the prevalence of SARS-CoV-2 in window w as $\text{prev}(w)$. Then, it follows that the lineage prevalence of lineage i in window w is $\text{prev}(w)p_i(w)$.

Under the null model of independent infections described by Chin et al. [22], the probability of co-infection by lineages i and j ($i < j$) is the product of their lineage prevalences. Thus,

$$P(\text{co-infected by } i \text{ and } j) = [\text{prev}(w)p_i(w)]\,[\text{prev}(w)p_j(w)] = \text{prev}(w)^2 p_i(w)p_j(w).$$

Conditioning on being infected (all of our sequences are from infected individuals) removes one factor of prevalence. We have,

$$P(\text{co-infected by } i \text{ and } j \mid \text{infected}) = \text{prev}(w)\, p_i(w)\, p_j(w).$$

Next, we denote the number of sequences for which we infer local Pango lineage ancestry in window w as $n(w)$. In our study, $n(w)$ is known. If $X_{i,j}(w)$ denotes the number of sequences among $n(w)$ that are from individuals co-infected by lineages i and j, it is reasonable that,

$$X_{i,j}(w) \sim \text{Binomial}\big(n(w), \text{prev}(w)\, p_i(w)\, p_j(w)\big).$$

However, not all sequences from individuals co-infected by lineages i and j will be i–j recombinants. Furthermore, not all i–j recombinants will be detected as recombinants using our method. To account for these two factors, we introduce two new parameters $\gamma_{i,j}(w)$ and $s_{i,j}(w)$:

$$\begin{aligned}\gamma_{i,j}(w) &= P\left(\text{is } i\text{–}j \text{ recombinant} \mid \text{sequence from individual co-infected by } i \text{ and } j\right),\\ s_{i,j}(w) &= P\left(\text{detected as recombinant} \mid \text{is } i\text{–}j \text{ recombinant}\right).\end{aligned}$$

Since $\gamma_{i,j}(w)$ and $s_{i,j}(w)$ are not separately identifiable, define the combined detection factor $\theta_{i,j}(w) = \gamma_{i,j}(w)s_{i,j}(w)$, and assume it is constant across lineage pairs and windows ($\theta_{i,j}(w) = \theta$).

Now let $R_{i,j}(w)$ denote the number of detected i–j recombinants in window w. We have,

$$R_{i,j}(w) \mid X_{i,j}(w) \sim \text{Binomial}\big(X_{i,j}(w), \theta\big).$$

It follows that,

$$R_{i,j}(w) \sim \text{Binomial}\big(n(w), \text{prev}(w)\, p_i(w)\, p_j(w)\, \theta\big).$$

Taking the expected value,

$$\text{E}\left[R_{i,j}(w)\right] = n(w)\, \text{prev}(w)\, p_i(w)\, p_j(w)\, \theta.$$

However, we still have not considered the possibility that our method will classify non-recombinant sequences as recombinant. The above $R_{i,j}(w)$ is only comprised of true positive cases, but our method can also misclassify non-recombinant sequences as i–j recombinants.

To account for this, we define,

$$\phi = P\left(\text{detected as recombinant} \mid \text{is non-recombinant}\right),$$

which is the false positive rate.

To simplify our derivations, we assume that $Y(w) = n(w) - \sum_{i<j} X_{i,j}(w)$ represents the number of true non-recombinant sequences tested in w. Although this assumption does not hold when $\gamma_{i,j}(w) < 1$ for some $i < j$ (when not all sequences from individuals co-infected by i and j are i–j recombinants), it is still reasonable because $n(w)$ is typically much larger than $\sum_{i<j} X_{i,j}$. This means that $Y(w) = n(w) - \sum_{i<j} X_{i,j}(w) \approx n(w) - \sum_{i<j} \gamma_{i,j}(w) X_{i,j}(w)$, with $n(w) - \sum_{i<j} \gamma_{i,j}(w) X_{i,j}(w)$ representing the correct number of non-recombinant sequences tested in w.

Then, denoting the set of sequences incorrectly classified as recombinants as $R^{FP}(w)$, we have that,

$$R^{FP}(w) \mid Y(w) \sim \text{Binomial}\left(Y(w), \phi\right).$$

Taking the expectation and using the tower rule,

$$\text{E}\left[R^{FP}(w)\right] = \text{E}\left[\text{E}[R^{FP}(w) \mid Y(w)]\right] = n(w)\phi \left[1 - \text{prev}(w) \sum_{i<j} p_i(w) p_j(w)\right].$$

Then, denoting the total number of recombinants detected in w as $R^{\text{total}}(w) = \sum_{i<j} R_{i,j}(w) + R^{FP}(w)$,

$$\begin{aligned}\mathrm{E}\left[R^{\text{total}}(w)\right] &= n(w)\theta\text{prev}(w)\sum_{i<j}p_i(w)p_j(w) + n(w)\phi\left[1-\text{prev}(w)\sum_{i<j}p_i(w)p_j(w)\right] \\ &= n(w)\left[\phi+(\theta-\phi)\,\text{prev}(w)\sum_{i<j}p_i(w)p_j(w)\right].\end{aligned}$$

Thus,

$$\mathrm{E}\left[\frac{R^{\text{total}}(w)}{n(w)}\right] = \phi+(\theta-\phi)\,\text{prev}(w)\sum_{i<j}p_i(w)p_j(w).$$

We next want to estimate θ and ϕ. Recall that $n(w)$ is known. We obtain our estimate of SARS-CoV-2 prevalence within window w, $\widehat{\text{prev}}(w)$, by averaging daily ONS SARS-CoV-2 prevalence estimates within window w. We obtain estimates for our lineage proportions, $\hat{p}_i(w)$ and $\hat{p}_j(w)$, by taking the number of sequences that belong to lineage i and j respectively in window w, and dividing this by the total number of sequences in window w.

Let $x_w = \widehat{\text{prev}}(w)\sum_{i<j}\hat{p}_i(w)\hat{p}_j(w)$ and then equate $\mathrm{E}\left[\frac{R^{\text{total}}(w)}{n(w)}\right]$ with $\frac{R^{\text{total}}(w)}{n(w)}$, which is the estimated recombinant proportion in w. Now we have,

$$\frac{R^{\text{total}}(w)}{n(w)} = \phi+\big(\theta-\phi\big)x_w.$$

We can estimate ϕ and θ by regressing $\frac{R^{\text{total}}(w)}{n(w)}$ on x_w. Denoting the fitted least squares intercept and slope as $\hat{\alpha}$ and $\hat{\beta}$ respectively, we calculate $\hat{\phi}=\hat{\alpha}$ and $\hat{\theta}=\hat{\alpha}+\hat{\beta}$.

Finally, we can use these estimates to obtain an estimate for the expected recombinant count by parental lineage pair. Recall that $R_{i,j}(w)$ represents the true positive i–j recombinant count in w. Then, letting $\mathcal{W}$ be our set of test windows, we can denote the total true positive i–j recombinant count across all windows as $R_{i,j}=\sum_{w\in\mathcal{W}}R_{i,j}(w)$. We have that $\mathrm{E}\left[R_{i,j}\right]=\theta\sum_{w\in\mathcal{W}}n(w)\text{prev}(w)p_i(w)p_j(w)$. Then,

$$\hat{\mathrm{E}}\left[R_{i,j}\right] = \hat{\theta}\sum_{w\in\mathcal{W}}n(w)\widehat{\text{prev}}(w)\hat{p}_i(w)\hat{p}_j(w).$$

In our results section, we compare this estimate with the number of detected recombinants that have i and j as parental lineages. Because we do not allocate false positives across lineage pairs, $\hat{\mathrm{E}}[R_{i,j}]$ represents a true positive expectation, whereas the observed counts may include false positives. Thus, it is reasonable to expect that the observed count of i–j recombinants will exceed $\hat{\mathrm{E}}[R_{i,j}]$. We are instead interested in the correlation between these two quantities across parental lineage pairs $i<j$.

3 Results

3.1 Simulation study

To evaluate the performance of our method for detecting recombinant SARS-CoV-2 sequences, we conducted a series of assessments based on the predicted local Pango lineage ancestry of synthetic SARS-CoV-2 sequences. The process used to generate synthetic sequences are described in Section 2.8.

First, we estimated the sensitivity and specificity of our method for classifying a sequence as a recombinant or non-recombinant. We classified a test sequence as a recombinant if the predicted local Pango lineage ancestry contained at least one lineage transition. Our method achieved a sensitivity of 0.801 (95% CI: [0.775, 0.825]) and a specificity of 0.989 (95% CI: [0.980, 0.994]).

To assess the accuracy of predicted local Pango lineage ancestries, we computed the mean position-by-position accuracy separately for recombinant and control sequences. On average, the inferred Pango lineage matched the true parental lineage at 86.9% (95% CI: [85.9%, 87.9%]) of genomic positions for recombinant sequences. Among control sequences, mean position-by-position accuracy was 99.2% (95% CI: [98.6%, 99.7%]).

We further evaluated how often the true parental lineage pair or lineage was recovered for recombinant and control sequences respectively. In 69.9% (95% CI: [67.0%, 72.7%]) of synthetic recombinant sequences, we detected two parental lineages that matched the true parental lineage pair. There was an overlap between the true and detected lineages in 100% (95% CI: [99.6%, 100%]) of synthetic recombinant sequences. In 98.4% (95% CI: [97.4%, 99.1%]) of synthetic control sequences, we detected a single parental lineage that matched the true parental lineage, and there was an overlap between the true and detected lineages in 99.4% (95% CI: [98.7%, 99.8%]) of synthetic control sequences.

In Table 2, we report how often the true parental lineage pair was recovered for recombinant sequences, stratifying by the true parental lineage pair. We counted the number of times we detected two parental lineages that matched the true parental lineage pair, for recombinants with each true parental lineage pair. We restricted this analysis to true parental lineage pairs with at least ten synthetic recombinants.

True lineages	Num. samples	Recovered	Prop.	2.5 % CI	97.5 % CI
(BA.5.2, BQ.1.1)	81	76	0.938	0.862	0.980
(BA.5.2, CH.1.1)	13	12	0.923	0.640	0.998
(BQ.1.1, CH.1.1)	72	65	0.903	0.810	0.960
(BA.4, BQ.1.1)	18	16	0.889	0.653	0.986
(BA.2, BQ.1.1)	89	78	0.876	0.790	0.937
(BA.5.1, BQ.1.1)	16	14	0.875	0.617	0.984
(BQ.1.1, XBB.1)	31	27	0.871	0.702	0.964
(BA.5, BE.1.1)	11	9	0.818	0.482	0.977
(BA.5.2.1, BQ.1.1)	78	63	0.808	0.703	0.888
(BA.2, BA.5.2.1)	27	21	0.778	0.577	0.914
(BA.5.2.1, CH.1.1)	13	10	0.769	0.462	0.950
(BA.2, BA.5.2)	21	16	0.762	0.528	0.918
(BA.5.2, BE.1.1)	40	30	0.750	0.588	0.873
(BE.1.1, XBB.1)	11	8	0.727	0.390	0.940
(BA.2, BE.1.1)	46	33	0.717	0.565	0.840
(BA.2, XBB.1)	10	7	0.700	0.348	0.933
(BA.5.1, BE.1.1)	13	9	0.692	0.386	0.909
(BE.1.1, CH.1.1)	32	22	0.688	0.500	0.839
(BA.5.2.1, BE.1.1)	30	20	0.667	0.472	0.827
(BA.2, CH.1.1)	14	9	0.643	0.351	0.872
(BA.5, BQ.1.1)	22	12	0.545	0.322	0.756
(BA.5.2, BA.5.2.1)	19	9	0.474	0.244	0.711
(BQ.1.1, other)	23	10	0.435	0.232	0.655
(BE.1.1, BQ.1.1)	128	24	0.188	0.124	0.266

Table 2. Detection of parental lineages for recombinant sequences, stratified by true parental lineage pair. We report 95% exact binomial confidence intervals for the proportion of sequences for which we detected two parental lineages that matched the true parental lineage pair.

We then evaluated whether the sensitivity to detect synthetic recombinant sequences was associated with the Hamming distance between the two parental sequences of each synthetic recombinant sequence. Using logistic regression, we found a positive association between the parental Hamming distance and the sensitivity ($p < 2 \times 10^{-16}$ using a two-sided Wald test). We estimate that for two recombinant sequences that differ by one unit in their parental Hamming distances, the odds of detection is 1.11 times higher in the recombinant sequence with the higher parental Hamming distance (95% CI: [1.09, 1.13]). The relationship between the parental Hamming distance and detection probability is shown in Figure 2, which displays the fitted logistic regression and the associated 95% pointwise confidence band.

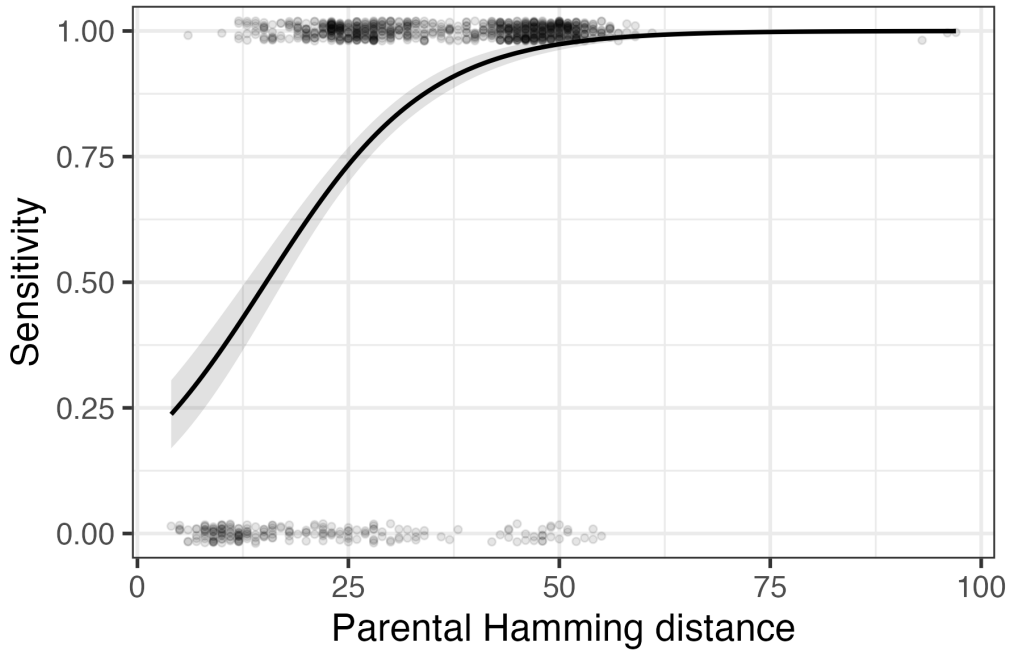

Figure 2. Sensitivity to detect recombinants as a function of the Hamming distance between their two parental sequences. Each point represents a synthetic recombinant sequence, with $y = 1$ indicating that it was classified as a recombinant using our method. y values are jittered vertically to avoid overplotting. The black line represents the logistic regression fit. The grey band represents pointwise 95% CIs.

We next assessed the accuracy of predicted breakpoint positions. Among sequences with one breakpoint (that had one predicted breakpoint), the mean breakpoint distance was 1238 nucleotides (95% CI: [1108, 1386]). For sequences with two breakpoints (that had two predicted breakpoints), the mean breakpoint distance was 1007 nucleotides (95% CI: [901, 1125]). For synthetic recombinants with two true breakpoints, we ordered true and detected breakpoint positions 5' to 3', paired them positionally (first with first, second with second), and calculated the distance between each pair. We then averaged the paired breakpoint distances across all sequences.

It is difficult to assess the accuracy of predicted breakpoint positions for a recombinant sequence whose predicted breakpoint count does not match its true breakpoint count. A confusion matrix of predicted and true breakpoint counts for recombinant sequences is shown in Table 3.

True breakpoints	**Predicted breakpoints**			
	0	1	2	3
1	77	417	6	0
2	122	119	256	3

Table 3. Confusion matrix of predicted versus true breakpoint counts.

3.2 Empirical data analysis

We used our method to predict the local Pango lineage ancestry for 440,307 SARS-CoV-2 sequences collected in England between September 2020 and March 2024. These sequences were sampled across 185 test windows that each consisted of a 7-day period with no gaps between successive windows. Of the 440,307 sequences, 7619 were detected to be recombinant sequences using our method, which corresponds to 1.73% (95% CI: [1.69%, 1.77%]) of the sequences.

In Figure 3, we plot the estimated recombinant proportion (the proportion of tested sequences predicted to be recombinant) in each test window and the SARS-CoV-2 prevalence estimates from the UK Office for National Statistics (ONS) Coronavirus Infection Survey [21], averaged within each test window. Shaded bands mark notable periods when ONS prevalence substantially exceeded the estimated recombinant proportion. We see a positive trend in the estimated recombinant proportion over time. Our confidence intervals widen in later windows because fewer sequences are available for analysis.

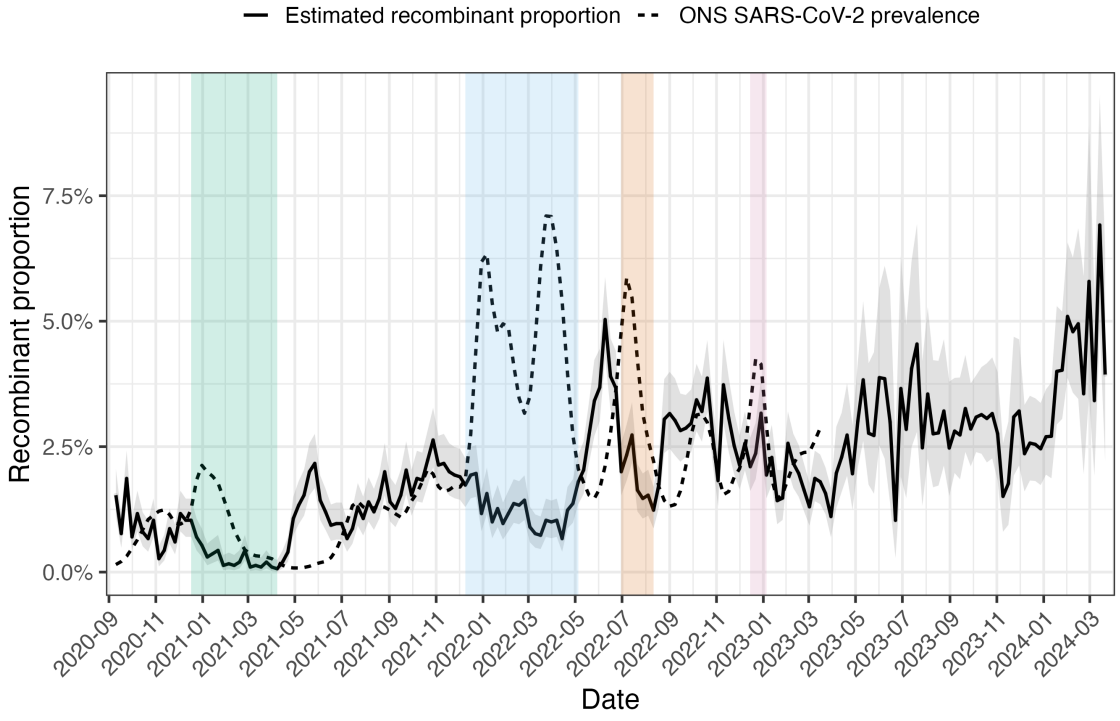

Figure 3. Estimated recombinant proportion and ONS prevalence estimates averaged within each test window. The solid line shows the estimated recombinant proportion in each test window. The shaded band indicates 95% exact binomial confidence intervals for the recombinant proportion in each test window. The dashed line shows daily SARS-CoV-2 prevalence estimates from the ONS survey averaged within each test window. Shaded bands mark periods when ONS prevalence substantially exceeded the estimated recombinant proportion.

The estimated recombinant proportion was positively associated with ONS prevalence averaged within each test window ($p = 0.0436$ using a two-sided Pearson correlation test).

The Pearson correlation was 0.176. Figure 4 shows the estimated recombinant proportion against the window-averaged ONS prevalence. Windows within the shaded bands in Figure 3 are shown in matching colors. High prevalence windows correspond to late 2021 and early 2022, when ONS prevalence greatly exceeded the estimated recombinant proportion (see Figure 3). ONS prevalence was available for the first 132 of 185 test windows (until the window ending March 19, 2023). These windows were used to compute the Pearson correlation and p-value in Figure 4.

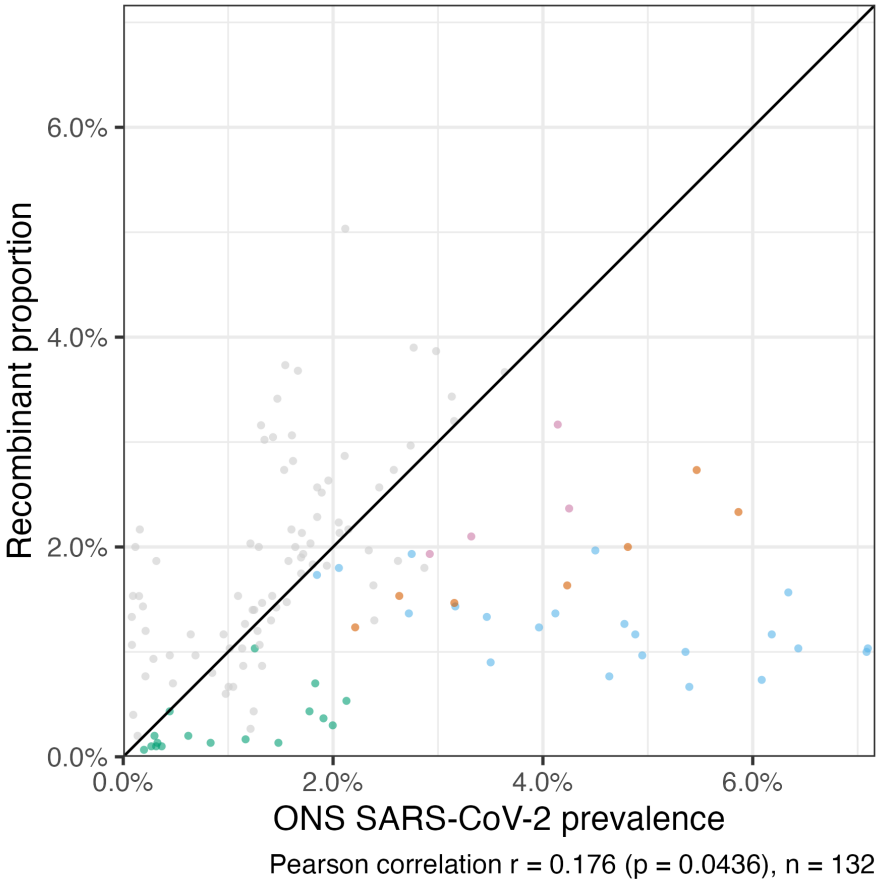

Figure 4. Scatterplot of estimated recombinant proportion against ONS prevalence estimates averaged within each test window. Each point represents one test window. Windows within the shaded bands in Figure 3 are shown in matching colors. The diagonal line is the identity $y = x$. The reported p-value ($p = 0.0436$) is from a two-sided Pearson correlation test.

Among 7619 detected recombinants, 7063 had exactly two parental lineages in their predicted local Pango lineage ancestry, corresponding to 92.7% (95% CI: [92.1%, 93.3%]) of detected recombinants. We counted the number of detected recombinants stratified by unique parental lineage pairs. For each parental lineage pair $i < j$, we compared the number of detected i–j recombinants (excluding cases in which one of the parental lineages was in the "other" category) to $\hat{\mathrm{E}}[R_{i,j}]$ derived in Section 2.10, which represents our estimate of the expected number of true positive i–j recombinants detected across all test windows.

For brevity, we henceforth refer to $\hat{\mathrm{E}}[R_{i,j}]$ as the expected true positive (TP) count for i–j recombinants.

ONS SARS-CoV-2 prevalence estimates are used to obtain expected TP counts for each lineage pair $i < j$. For comparability with these counts, detected recombinant counts are also restricted to windows with available ONS prevalence estimates (until the window ending March 19, 2023). For reference, test windows before March 19, 2023 contained 387,054 sequences, with 5006 sequences detected as recombinant with exactly two parental lineages in their predicted local Pango lineage ancestry (excluding cases in which one of the parental lineages were in the "other" category).

Across parental lineage pairs $i < j$, the number of detected i–j recombinants was positively associated with the expected TP count ($p = 8.88 \times 10^{-22}$ using a two-sided Pearson correlation test). We visualized this association in Figure 5.

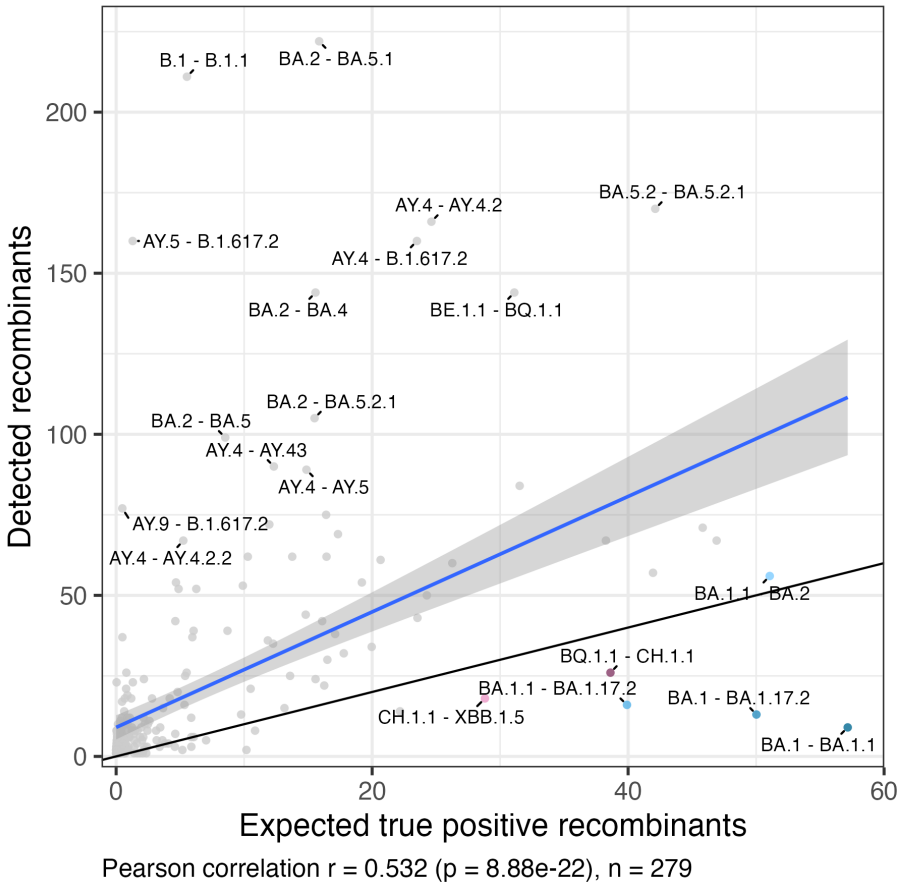

Figure 5. Scatterplot of detected recombinant counts against expected TP counts ($\hat{\mathrm{E}}[R_{i,j}]$). Each point represents a parental lineage pair. The blue line shows the least squares fit with the grey band indicating 95% confidence intervals for the mean response. The black line represents the identity $y = x$. We label the 20 parental lineage pairs with the highest absolute residual values from ordinary least squares. We color under-represented lineage pairs.

Because expected TP counts do not include contributions from false positives, we generally expect the number of detected recombinants to be larger across parental lineage pairs $i < j$. We observe this in Figure 5 (most points fall above the identity line $y = x$).

Several parental lineage pairs deviate markedly from the overall relationship between detected counts and expected TP counts. To identify these parental lineage pairs, we fit a least squares line with the detected count as the response and the expected TP count as the predictor. We identified the 20 parental lineage pairs with the highest absolute residual values and annotated them in Figure 5.

Six of these parental lineage pairs had negative residual values. We refer to these parental lineage pairs as under-represented lineage pairs. Recombinants with these parental lineage pairs (except for BA.1.1–BA.2) had lower detected counts than their expected TP counts (these recombinants fall below the identity line $y = x$ in Figure 5). This should be interpreted with caution. This apparent under-representation may be attributable to uncertainty in expected TP counts or sampling variability in detected counts. Alternatively, this may suggest recombinants with these parental lineage pairs occur less frequently in the population than expected under our co-infection model or we have lower sensitivity to detect these recombinants compared to our overall sensitivity (see Section 2.10).

We likely have low sensitivity to detect BA.1–BA.1.1, BA.1–BA.1.17.2, and BA.1.1–BA.1.17.2 recombinants, given that only a few mutations separate these lineage pairs. Using consensus sequences and ignoring non-standard nucleotides, pairwise Hamming distances were one for BA.1–BA.1.1, three for BA.1–BA.1.17.2, and four for BA.1.1–BA.1.17.2. Recall that in our simulation study, the sensitivity of our method was low when parental sequences only differed by a few mutations (see Figure 2).

Pairwise Hamming distances were relatively high for BQ.1.1–CH.1.1 (39 nucleotides), BA.1.1–BA.2 (41 nucleotides), and CH.1.1–XBB.1.5 (34 nucleotides). However, the sensitivity to detect recombinants with these parental lineage pairs may still be low, depending on where recombinantion breakpoints occur. The resulting recombinant sequence may only have a few mutations relative to one of its parents, if most of its genome is inherited from this parent. Furthermore, these parental lineage pairs lie close to the identity line $y = x$ (slightly above the line for BA.1.1–BA.2). Thus, their under-representation may be explained by uncertainty in expected TP counts or sampling variability in detected counts.

Interestingly, every under-represented parental lineage pair co-circulated in England during two intervals, late 2021 to early 2022 (BA.1–BA.1.1, BA.1–BA.1.17.2, BA.1.1–BA.1.17.2, BA.1.1–BA.2) and late 2022 to early 2023 (BQ.1.1–CH.1.1, CH.1.1–XBB.1.5), when the estimated recombinant proportion was low relative to ONS prevalence (see Figure 3). For under-represented parental lineage pairs, Figure 6 shows the pairwise product of their lineage frequencies across test windows. Recombinants with these parental lineage pairs had relatively few detected counts during these two periods, indicating that these under-represented lineage pairs contributed to the lower frequency of recombinants detected during this period. However, further work is needed to determine the cause of this apparent under-representation.

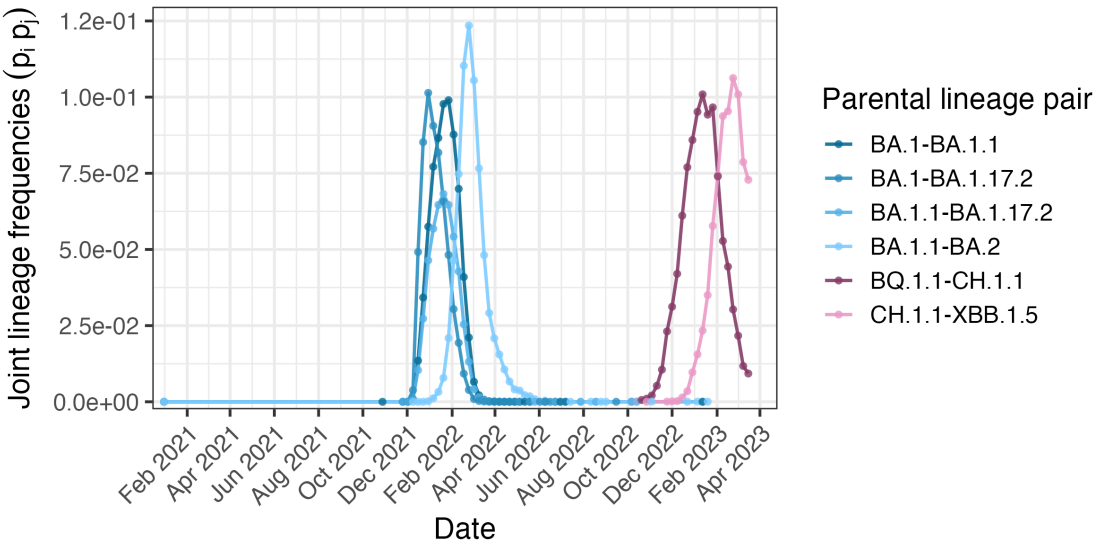

Figure 6. Joint frequencies across test windows for under-represented lineage pairs.

Of the 20 annotated recombinant pairs with the highest residual values (see Figure 5), 14 parental lineage pairs had positive residual values, indicating that recombinants with these lineage pairs were detected more frequently relative to the overall trend between detected counts and expected TP counts. It is difficult to pinpoint whether this reflects the population frequency of recombinant sequences with these parental lineage pairs, higher detection sensitivity, disproportionate false positive assignments to these parental lineage pairs, or some combination of these factors.

In the process of estimating expected TP counts, we estimated the detection factor θ and false positive rate ϕ to be 0.557 (95% CI: [0.309, 0.804]) and 0.011 (95% CI: [0.009, 0.013]) respectively. Recall that θ equals sensitivity times the probability that a sample from a co-infected individual is a recombinant (see Section 2.10). Confidence intervals are Wald intervals from the linear regression model described in Section 2.10, treating x_w as fixed. Remarkably, our estimated false positive rate closely matches what we estimated in the simulation (see Section 3.1).

Across 7619 detected recombinants, we inferred 9105 recombination breakpoints. 6324 (83.0%) had one breakpoint, 1146 (15.0%) had two, 118 (1.5%) had three, 23 (0.3%) had four, and 8 (0.1%) had five or above. In Figure 7, we plot the genomic position of each detected breakpoint. We observed a recombination hotspot within spike, and an enrichment of breakpoints at the 3' end of the genome corresponding to accessory proteins. However, we do not account for potential variation in breakpoint detection sensitivity across positions.

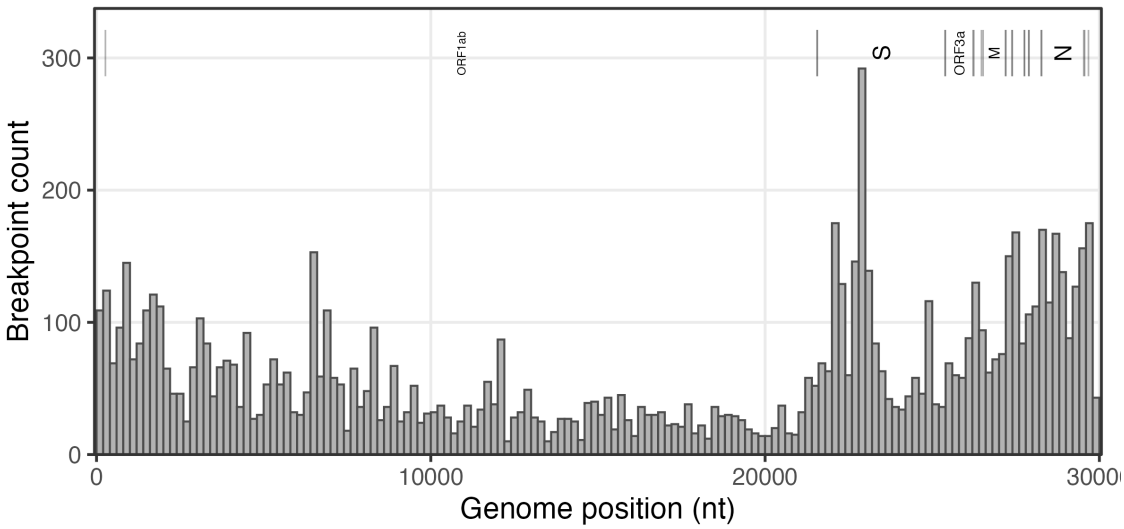

Figure 7. Histogram of detected recombination breakpoints.

We also observed enrichment of recombination breakpoints in intergenic regions. Gene boundaries were defined using the Wuhan-Hu-1 reference genome (GenBank: MN908947) [16, 23]. Although intergenic regions comprise only around 0.5% of the genome, 1.1% of all detected breakpoints were localized in these regions (90/8767). When calculating the proportion of detected breakpoints in intergenic regions, we excluded 339 breakpoints mapping to the ends of the genome, specifically from the 5' end to ORF1ab and from ORF10 to the 3' end. Using a two-sided binomial test, we found this enrichment to be highly statistically significant ($p = 1.63 \times 10^{-9}$).

4 Discussion

Genomic surveillance of recombinant SARS-CoV-2 sequences is important, given that mutations from each parental lineage can provide a growth advantage to the recombinant sequence. In this study, we developed a HMM to detect the local Pango lineage ancestry of query SARS-CoV-2 sequences based on lineage-specific nucleotide frequencies calculated using a reference set of recent sequences. Our method does not depend on an existing phylogeny or any user-defined parameters such as the mutation rate or recombination rate. Instead, we use maximum likelihood to estimate the lineage-transition probability between consecutive sites, and the probability of observing alleles absent from the parental lineage at each site, which accounts for mutations on the recombinant sequence.

We validated our method using synthetic sequences generated from real SARS-CoV-2 genomes. In our simulation, our method achieved a sensitivity of 0.801 (95% CI: [0.775, 0.825]) for classifying recombinant sequences and a specificity of 0.989 (95% CI: [0.980, 0.994]) for classifying non-recombinant sequences. In 69.9% (95% CI: [67.0%, 72.7%]) of synthetic recombinant sequences, we detected two parental lineages that matched the true parental lineage pair. In 98.4% (95% CI: [97.4%, 99.1%]) of synthetic control sequences,

we detected a single parental lineage that matched the true parental lineage. We found the sensitivity of our method to be positively associated with the number of mutations separating the parental sequences of the recombinant (see Figure 2). Finally, we estimated the mean distance between true and inferred breakpoints to be 1238 nucleotides (95% CI: [1108, 1386]) and 1007 nucleotides (95% CI: [901, 1125]) for synthetic recombinants with one and two breakpoints respectively.

Applying our model to real SARS-CoV-2 sequences collected in England between September 2020 and March 2024, we found 7619 recombinant sequences across 440,307 sequences, corresponding to 1.73% (95% CI: [1.69%, 1.77%]) of sequences. These 440,307 sequences were sampled across 185 test windows, each window corresponding to a 7-day period with no gaps between successive windows.

We hypothesized that across our test windows, the fraction of tested sequences detected as recombinant using our method would be positively associated with community SARS-CoV-2 prevalence, because higher prevalence raises co-infection opportunities, which should result in a higher rate of recombinant sequences in the population. There was a positive association between the estimated recombinant proportion in each test window and SARS-CoV-2 prevalence from the ONS survey, averaged within each test window ($r = 0.176$, $p = 0.0436$ using a two-sided Pearson correlation test).

We next modeled the number of recombinants in our sample as a function of community SARS-CoV-2 prevalence to derive, for each parental lineage pair, the expected number of true positive recombinants across the 440,307 sequences analyzed (Section 2.10). For brevity, we refer to this as the expected true positive (TP) count for each parental lineage pair. This derivation relies on two key assumptions. First, we assume independent infections in our co-infection model, which means that the probability of co-infection by two lineages equals the product of their marginal prevalences. Second, we assume that the false positive rate and the detection factor (the product of sensitivity and the probability that a sequence from a co-infected individual is a recombinant) are constant across parental lineage pairs and test windows. Our second assumption ensures identifiability of the false positive rate and detection factor in our model. We estimated the false positive rate and detection factor to be 0.011 (95% CI: [0.009, 0.013]) and 0.557 (95% CI: [0.309, 0.804]) respectively for the real data analysis. Our estimated false positive rate closely matched the false positive rate in our simulation study.

We estimated the expected TP count for each parental lineage pair. We found that the number of detected recombinants with each parental lineage pair exceeded the corresponding expected TP count for most parental lineage pairs. This is not surprising, because our expected TP count does not include potential contributions from non-recombinant sequences that were detected as recombinant using our method. However, we cannot reliably allocate expected false positive counts across specific parental lineage pairs.

Recall that during the period when ONS SARS-CoV-2 prevalence estimates were available (until the test window ending March 19, 2023), we analyzed 387,054 sequences, with 5006 sequences detected as recombinant with two parental lineages. The vast majority of analyzed sequences should be non-recombinant. If our estimated false positive rate is correct, we would expect around $380,000 \times 0.01 = 3800$ of these sequences to be false

positive cases. The expected TP count summed across all lineage pairs was 1390, which is only slightly higher than the implied TP count ($5006 - 3800 = 1206$).

Although many detections are likely false positives, across parental lineage pairs, we found a strong positive association between the expected TP count and the number of detected recombinants ($r = 0.532$, $p = 8.88 \times 10^{-22}$ using a two-sided Pearson correlation test). This indicates that our method is detecting recombinants at a rate that is predictable based on SARS-CoV-2 prevalence and co-infection dynamics between lineages under the null model of independent infections [22].

We then identified parental lineage pairs whose detected counts deviated from the overall trend between detected counts and expected TP counts. We identified six under-represented parental lineage pairs (BA.1–BA.1.1, BA.1–BA.1.17.2, BA.1.1–BA.1.17.2, BQ.1.1–CH.1.1, BA.1.1–BA.2, CH.1.1–XBB.1.5). These lineage pairs, except for BA.1.1–BA.2, had lower detected counts than expected TP counts.

Under-representation of BA.1–BA.1.1, BA.1–BA.1.17.2, and BA.1.1–BA.1.17.2 recombinants is likely explained by low sensitivity to detect these recombinants. Only a few mutations separate each of these lineage pairs, making recombinant detection difficult (see Figure 2). On the contrary, pairwise Hamming distances are high for BQ.1.1–CH.1.1, BA.1.1–BA.2, and CH.1.1–XBB.1.5. However, if most of the genome is inherited from a single parent, the recombinant can be very similar to one of the parental lineages, so detection sensitivity may still be low.

Under-representation can also occur if parental lineage pairs were segregated to different geographical locations or subpopulations in England, which would make co-infection by these lineage pairs unlikely. Co-infection rates may be lower than expected even under homogeneous mixing, due to within-host interference. Moreover, lineage pairs may differ in their propensity to produce viable recombinants. Finally, the apparent under-representation of these parental lineage pairs could result from estimation error in expected TP counts or sampling variability in detected counts. Estimating the variability of expected TP counts is challenging. We do not have standard errors for ONS SARS-CoV-2 prevalence estimates. Additionally, this would require accounting for correlations in lineage prevalences across test windows.

We found that these six under-represented lineage pairs co-circulated in England during two intervals, late 2021 to early 2022 (BA.1–BA.1.1, BA.1–BA.1.17.2, BA.1.1–BA.1.17.2, BA.1.1–BA.2) and late 2022 to early 2023 (BQ.1.1–CH.1.1, CH.1.1–XBB.1.5). These two intervals coincide with periods when the estimated recombination proportion was low relative to ONS SARS-CoV-2 prevalence estimates. This indicates that these under-represented lineage pairs contributed to the lower frequency of recombinants detected during these periods.

Aggregating all detected recombination breakpoints, we observed a recombination hotspot within spike, which is consistent with previous work on recombination hotspots in sarbecoviruses [24]. Additionally, we found that breakpoints were enriched in intergenic regions, consistent with their high colocalization with TRS-B sites [25].

Using RIPPLES, Turakhia et al. found 2.7% of sampled genomes inferred to have de-

tectable recombinant ancestry [13]. This is higher than the proportion of detected recombinants using our method (1.73%; 95% CI: [1.69%, 1.77%]). This discrepancy is likely attributable to many factors. Turakhia et al. only analyze sequences up to May 2021, before the emergence of XBB. Furthermore, our method cannot detect recombination between sequences in the same lineage, which explains the lower proportion of detected recombinants using our method. Finally, even a modest difference in false positive rates would affect the estimated proportion.

Future work could estimate SARS-CoV-2 prevalence from the frequency of detected recombinants across test windows. In this study, we developed a statistical framework linking disease prevalence and lineage frequencies to the expected number of detected recombinants (see Section 2.10). We further showed that these expected counts were correlated with observed recombinant counts. Estimating prevalence is feasible if the method's sensitivity and false positive rate were known for the set of query sequences. In our study, we estimated the detection factor and false positive rate using ONS prevalence estimates (see Section 2.10), so these rates are not generalizable outside England or beyond March 2023, when ONS prevalence estimates are no longer available. To estimate the prevalence of SARS-CoV-2 outside of England or beyond March 2023, we would need reliable sensitivity and specificity estimates in these settings.

Although we focused on SARS-CoV-2 in this study, our HMM is broadly applicable to other RNA and DNA viruses for detecting recombinants. Moreover, by limiting lineage transitions to predefined genome positions, our HMM can be easily adapted to detect reassortment events in segmented viruses such as influenza. Our detection method should also perform well on rapidly evolving viruses because we explicitly model novel alleles on recombinant sequences via a pseudo-frequency.

5 Implementation

The hidden Markov model and detection of recombinant sequences were implemented in Python 3.12.2. Results files were processed and plotted in R version 4.4.1.

6 Availability

All code for the analysis is available at github.com/nobuakimasaki/HMM-recombination.

7 Acknowledgments

We thank Professor Brian Browning for helpful suggestions on accelerating the forward algorithm and Professor Nicola Mueller for discussions of co-infection dynamics.

We gratefully acknowledge the investigators and laboratories that generated, submitted, and shared sequence data and metadata via GenBank (NCBI), which form the basis of this research.

8 Funding

This work is supported by NIH NIGMS R35 GM119774 to T.B. T.B. is a Howard Hughes Medical Institute Investigator.

References

1. Chrisman BS, Paskov K, Stockham N, Tabatabaei K, Jung JY, et al. (2021) Indels in SARS-CoV-2 occur at template-switching hotspots. BioData Mining 14: 20.
2. Trémeaux P, Latour J, Ranger N, Ferrer V, Harter A, et al. (2023) SARS-CoV-2 Co-Infections and Recombinations Identified by Long-Read Single-Molecule Real-Time Sequencing. Microbiology Spectrum 11: e0049323.
3. Tamura T, Ito J, Uriu K, Zahradnik J, Kida I, et al. (2023) Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. Nature Communications 14: 2800.
4. Erkihun M, Ayele B, Asmare Z, Endalamaw K (2024) Current Updates on Variants of SARS-CoV-2: Systematic Review. Health Science Reports 7: e70166.
5. Samson S, Lord Makarenkov V (2022) SimPlot++: a Python application for representing sequence similarity and detecting recombination. Bioinformatics (Oxford, England) 38: 3118–3120.
6. Salminen MO, Carr JK, Burke DS, McCutchan FE (1995) Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS research and human retroviruses 11: 1423–1425.
7. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evolution 1: vev003.
8. Sawyer S (1989) Statistical tests for detecting gene conversion. Molecular Biology and Evolution 6: 526–538.
9. Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences of the United States of America 98: 13757–13762.
10. Lam HM, Ratmann O, Boni MF (2018) Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. Molecular Biology and Evolution 35: 247–251.
11. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics (Oxford, England) 22: 3096–3098.
12. Varabyou A, Pockrandt C, Salzberg SL, Pertea M (2021) Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. Genetics 218: iyab074.

13. Turakhia Y, Thornlow B, Hinrichs A, McBroome J, Ayala N, et al. (2022) Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature 609: 994–997.

14. Alfonsi T, Bernasconi A, Chiara M, Ceri S (2024) Data-driven recombination detection in viral genomes. Nature Communications 15: 3313.

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

16. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, et al. (2013) GenBank. Nucleic Acids Research 41: D36–D42.

17. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, et al. (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics (Oxford, England) 34: 4121–4123.

18. O'Toole Scher E, Underwood A, Jackson B, Hill V, et al. (2021) Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evolution 7: veab064.

19. Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77: 257–286.

20. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17: 261–272.

21. Pouwels KB, House T, Pritchard E, Robotham JV, Birrell PJ, et al. (2021) Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. The Lancet Public Health 6: e30–e38.

22. Chin T, Foxman EF, Watkins TA, Lipsitch M (2024) Considerations for viral co-infection studies in human populations. mBio 15: e0065824.

23. Wu F, Zhao S, Yu B, Chen YM, Wang W, et al. (2020) A new coronavirus associated with human respiratory disease in China. Nature 579: 265–269.

24. Lytras S, Hughes J, Martin D, Swanepoel P, de Klerk A, et al. (2022) Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination. Genome Biology and Evolution 14: evac018.

25. Yang Y, Yan W, Hall AB, Jiang X (2021) Characterizing Transcriptional Regulatory Sequences in Coronaviruses and Their Role in Recombination. Molecular Biology and Evolution 38: 1241–1248.