Virus Evolution, 2023, **9(2)**, 1–14
DOI: https://doi.org/10.1093/ve/vead066
Advance access publication date: 23 November 2023
Research Article

Evaluation of recombination detection methods for viral sequencing

Frederick R. Jaya,[1,2,†,*] Barbara P. Brito,[1,3,‡] and Aaron E. Darling[1,4,§]

[1]Australian Institute for Microbiology & Infection, University of Technology Sydney, 15 Broadway, Ultimo, New South Wales 2007, Australia, [2]Ecology and Evolution, Research School of Biology, Australian National University, 134 Linnaeus Way, Acton, Australian Capital Territory 2600, Australia, [3]New South Wales Department of Primary Industries, Elizabeth Macarthur Agricultural Institute, Woodbridge Road, Menangle, New South Wales 2568, Australia and [4]Illumina Australia Pty Ltd, Ultimo, New South Wales 2007, Australia
[†]https://orcid.org/0000-0002-4019-7026
[‡]https://orcid.org/0000-0001-6122-2596
[§]https://orcid.org/0000-0003-2397-7925
*Corresponding author: E-mail: fredjaya1@gmail.com

Abstract

Recombination is a key evolutionary driver in shaping novel viral populations and lineages. When unaccounted for, recombination can impact evolutionary estimations or complicate their interpretation. Therefore, identifying signals for recombination in sequencing data is a key prerequisite to further analyses. A repertoire of recombination detection methods (RDMs) have been developed over the past two decades; however, the prevalence of pandemic-scale viral sequencing data poses a computational challenge for existing methods. Here, we assessed eight RDMs: PhiPack (Profile), 3SEQ, GENECONV, recombination detection program (RDP) (OpenRDP), MaxChi (OpenRDP), Chimaera (OpenRDP), UCHIME (VSEARCH), and gmos; to determine if any are suitable for the analysis of bulk sequencing data. To test the performance and scalability of these methods, we analysed simulated viral sequencing data across a range of sequence diversities, recombination frequencies, and sample sizes. Furthermore, we provide a practical example for the analysis and validation of empirical data. We find that RDMs need to be scalable, use an analytical approach and resolution that is suitable for the intended research application, and are accurate for the properties of a given dataset (e.g. sequence diversity and estimated recombination frequency). Analysis of simulated and empirical data revealed that the assessed methods exhibited considerable trade-offs between these criteria. Overall, we provide general guidelines for the validation of recombination detection results, the benefits and shortcomings of each assessed method, and future considerations for recombination detection methods for the assessment of large-scale viral sequencing data.

Keywords: recombination detection methods; recombination; bioinformatics.

1. Introduction

Recombination is the exchange of genetic material between different genomes. It is a fundamental process that occurs in many viruses and plays a significant role in their evolution and emergence. The interplay between recombination and mutation can increase the genetic diversity of viral populations. This enables the rapid acquisition of advantageous genotypes, leading to the selection of fitter strains and variants, the expansion of their host range, and the development of resistance to antiviral therapies and host immune responses (Duffy, Shackelton and Holmes, 2008; Pybus and Rambaut, 2009; Simon-Loriere and Holmes, 2011; Xiao et al., 2016). The process of recombination has been observed to create new lineages of fit viruses in agricultural and public health contexts (Wong et al., 2013; Yeşilbağ, Alpay and Becher, 2017; Brito et al., 2018; Boni et al., 2020; Lytras et al., 2022), highlighting the need to consider it in the understanding of viral evolution and the spread of diseases.

Failing to account for recombination in sequencing data can affect phylogenetic estimates, such as the relationships and ancestry of individuals (phylogenetic tree topology; (Arenas and Posada, 2010)) and the number of substitutions in lineages (branch lengths; (Schierup and Hein, 2000)). It can also impact estimations of site-rate variation (Anisimova, Nielsen and Yang, 2003), population structure, and positive selection (Castillo-Ramírez et al., 2011; Pérez-Losada et al., 2015; Rousselle et al., 2019). Therefore, it is important to test for and account for potential recombination before conducting evolutionary analyses to avoid misleading estimations and complicating the interpretation of results due to the conflicting intragenomic signals in the data.

Recombination detection methods (RDMs) are used to identify possible recombination breakpoints in sequence alignments. Specifically, homoplasy methods can be utilised to test for the presence or absence of recombination but do not identify the breakpoints. There are various RDMs available, which differ in their ability to identify recombination in specific sequences or across an entire alignment and the statistical tests and algorithms used (Martin, Lemey and Posada, 2011; Pérez-Losada et al., 2015). Previous studies have examined the accuracy and false positive

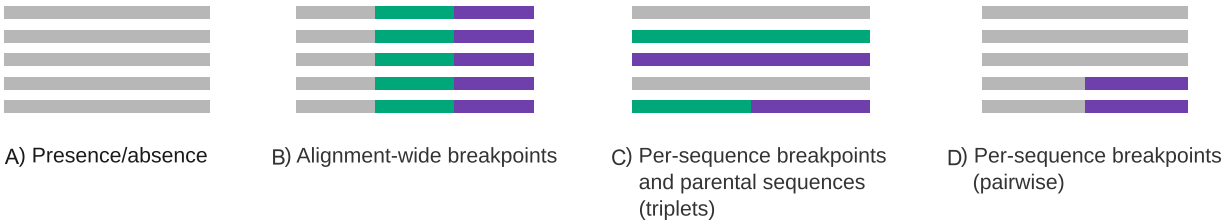

Figure 1. Modes of assessed RDMs. The resolution in which recombination is identified differs across methods. (A) Presence/absence of methods informs whether recombination is present in a sequence alignment, but provides no information where breakpoints are located. (B) Methods report the location of breakpoints across the entire sequence alignment but do not inform the exact sequences in which they occur. (C) Methods that compare every potential sequence triplet from the alignment identifies the specific breakpoint location within a putative recombinant and the parental sequences where those segments are inherited from. (D) Pairwise methods report regions shared by two sequences where recombination may have occurred between them.

(FP) rates of RDMs for detecting simulated recombination events in increasingly divergent sequences (Smith and Smith, 1998; Brown et al., 2001; Posada and Crandall, 2001; Boni, Posada and Feldman, 2007). However, the growing use of high-throughput viral sequencing (Loman et al., 2012; Goodwin, McPherson and McCombie, 2016; Pérez-Losada et al., 2020) and real-time genomic surveillance (Quick et al., 2016; Hadfield et al., 2018; Seemann et al., 2020) requires efficient and scalable RDMs, which has not been thoroughly evaluated.

In order to accurately determine the confidence of recombination tests, it is common practice to use multiple methods, as different approaches have respective benefits and limitations (Martin et al., 2015). However, the commonly used RDMs (Lole et al., 1999; Kosakovsky Pond et al., 2006; Martin et al., 2021) are not suitable for analysing pandemic-scale sequencing data. Here, we assess the impact of sequence diversity and recombination frequency on the performance of eight RDMs and identify which ones are scalable for analysis of thousands of viral sequences. Through the analysis of simulated and empirical data, we provide user guidelines and considerations for selecting and using RDMs.

2. Methods

2.1 Recombination detection methods

RDMs test sequencing data for the presence of recombination. A large variety of detection methods have been developed, implementing different algorithms and statistical tests to address different datasets and applications (Martin, Lemey and Posada, 2011; Pérez-Losada et al., 2015). The resolution of how recombination is reported varies across methods (Fig. 1). Methods such as PhiPack (Bruen, Philippe and Bryant, 2006) indicate the presence/absence of recombination within an entire alignment, whereas GARD (Kosakovsky Pond et al., 2006) identifies recombination breakpoint regions across the entire sequence alignment. MaxChi, Chimaera (Martin et al., 2015), 3SEQ (Boni, Posada and Feldman, 2007; Lam, Ratmann and Boni, 2018), and GENECONV (Sawyer, 1989; Padidam, Sawyer and Fauquet, 1999) identify recombination breakpoints within specific sequences and their putative parents.

The performance and scalability of eight RDMs were evaluated—PhiPack (Profile), 3SEQ, GENECONV, recombination detection program (RDP) (OpenRDP), MaxChi (OpenRDP), Chimaera (OpenRDP), UCHIME (VSEARCH) (Rognes et al., 2016), and gmos (Domazet-Lošo and Domazet-Lošo, 2016) (Table 1; Fig. 2). We prioritised selecting methods that were able to be incorporated in a Unix-based pipeline and can process over 1,000 sequences. The following section outlines the intended application for each selected RDM, the statistical method used, and the parameters used throughout this study. Although each RDM was developed for specific biological applications, all methods similarly identify signals of recombination between input sequences.

PhiPack evaluates sequence alignments for recombination using the pairwise homoplasy index, which examines site pairs across aligned sequences using the four gametes test. The 'Profile' function of PhiPack allows recombination hotspots to be searched across all sites in a sequence alignment with a sliding window. All analyses were conducted with default settings, according to recommendations based on previous analyses (Bruen, Philippe and Bryant, 2006). The window size spanned 1,000 nucleotides and moved every twenty-five nucleotides. *P*-values are reported for each window, where $P < 0.05$ indicates that recombination is likely to occur within the window.

3SEQ is a non-parametric algorithm that uses a ranked clustering statistic to locate significant breakpoint regions. 3SEQ tests all combinations of sequence triplets within the alignment to determine if one sequence is a potential recombinant between the other two sequences. The 'maximum descent' metric reports the degree to which regions cluster to either parent. The more significant, or the 'steeper' the descent in a region, the more likely that the region indicates a breakpoint. 3SEQ was run using default settings using the 'full run' option, and duplicate sequences were automatically removed. A three-dimensional probability table is required for 3SEQ to determine the significance of putative recombination events. We generated and used a 700 x 700 x 700 probability table for all analyses.

GENECONV identifies possible gene conversion events by looking for similarities, or concordance, between aligned pairwise sites. Similar to 3SEQ, only polymorphic sites between the sequence pairs are assessed. Putative recombination between two sequences is assessed using a basic local alignment search tool (BLAST)-like statistic. Aligned regions between two sequences that are significantly similar are considered to be recombinant. Recombination can be detected between sequences present in the dataset (inner) or predicted to be occurring with a sequence absent from the dataset (outer). All analyses were conducted with default settings with 10,000 permutations.

OpenRDP is an open-source, command-line reimplementation of the commonly used RDP suite of programs (https://github.com/PoonLab/OpenRDP). All methods implemented in OpenRDP—RDP, MaxChi, and Chimaera test for recombination in polymorphic sites for each possible sequence triplet using a sliding window. Peaks in the respective *P*-values indicate potential recombination breakpoints. RDP (Martin and Rybicki, 2000) tests whether the sites between the two assumed parental sequences are more similar to the recombinant sequence (i.e. P1-R and P2-R), than that of each other (P1–P2). Significance is tested under a binomial distribution. On the other hand, MaxChi (Smith, 1992) and

Table 1. Properties of assessed RDMs. Schematics of the analysis and output resolutions for each method are detailed in Fig. S1.

Method	Version used	Statistical test	Alignment-free	Analysis resolution	Output resolution
PhiPack (Profile)	–	Pairwise homoplasy index	No	Alignment-wide windows	Alignment-wide breakpoints
3SEQ	v1.7	Mann–Whitney *U*-test	No	All possible sequence triplets	Per-sequence breakpoints
GENECONV	v1.8.1	BLAST-like	No	All possible sequence pairs	Per-sequence breakpoints
RDP (OpenRDP)	v0.1.0-rc2	Binomial distribution	No	All possible sequence triplets	Per-sequence breakpoints
MaxChi (OpenRDP)	v0.1.0-rc2	X^2 distribution	No	All possible sequence triplets	Per-sequence breakpoints
Chimaera (OpenRDP)	v0.1.0-rc2	X^2 distribution	No	All possible sequence triplets	Per-sequence breakpoints
UCHIME (VSEARCH)	v2.14.2	Numerical score according to 'diffs'	Yes	All possible sequence triplets	Per-sequence only
gmos	v1.0	BLAST-like	Yes	Query-subject sequence pairs	Per-sequence breakpoints

Chimaera (Posada and Crandall, 2001) utilise a X^2 distribution to test whether there is a significant difference in the proportions of polymorphic sites. MaxChi discards all monomorphic sites, whereas Chimaera additionally discards sites that differ between the recombinant and parental sequences. All OpenRDP methods were run using default settings on sequence triplets.

UCHIME (VSEARCH) is an alignment-free method developed to detect recombination in raw sequencing reads resulting from recombination during the sequencing or amplification processes. UCHIME (VSEARCH) assesses the likelihood for a query sequence to be a recombinant between two parental sequences by counting the number of sites that match either putative parental sequence ('diffs'). All 'diffs' are then scored to determine if the query sequence is a recombinant between the two. All analyses were executed with default settings using the de novo mode as implemented in VSEARCH (Rognes et al., 2016). A required dereplication step was conducted to count the frequency of unique sequence genotypes. Using the de novo mode, UCHIME (VSEARCH) only considers sequences as recombinants if the parental sequences are at least twice as abundant in the dataset.

gmos is an alignment-free method developed to identify recombination, or mosaicism, amongst prokaryotic sequences. gmos calculates the likelihood that local alignments between all query and subject sequences are recombinant with a BLAST-like algorithm. Two sequence files are required as input—a query file consisting of sequences to be tested against a file of subject sequences. Default settings were used for all analyses. To detect recombination between all sequence pairs in a dataset, we specified the identical, fasta file as both query and subject sequences.

2.2 Simulations

To assess the performance and scalability of the RDMs (Fig. 2), simulations were conducted using SANTA-SIM (Jariani et al., 2019). We modified SANTA-SIM to report the location of simulated recombination breakpoints (https://github.com/koadman/santa-sim). Viral sequences were simulated 'forwards-in-time' across discrete and non-overlapping generations and replicated according to various parameters. At each generation, the evolution and replication of sequences were determined by the mutation rate (m), recombination rate (r), and dual-infection probability (d). The mutation rate determined the probability of a point mutation occurring per site per generation. The recombination rate determined the probability of recombination occurring between two sequences per site per generation. The dual-infection probability determined the likelihood that one cell is co-infected with two viruses for recombination to occur.

All simulations utilised the complete hepatitis C virus (HCV) envelope-encoding gene regions from de-identified patient #37 from the study by Ho et al. (2017). For each replicate, one sequence from this alignment was randomly selected to form the starting population. All sequences ($n = 100$) in the starting population were identical. All sequences and intra-sequence regions were set to evolve neutrally, and hence, no selection occurred and all sequences had an equal chance of being inherited in the next generation (Jariani et al., 2019). We ensured that all methods were assessed on their ability to detect recombination events without subsequent mutations. These mutations can weaken the recombination signal and bias the evaluation of methods (Chan, Beiko and Ragan, 2006). Therefore, sequences were evolved for ninety-nine discrete generations with mutation only (no recombination), followed by a final generation of recombination alone (no mutation). It is important to note that this does not emulate empirical viral evolution, but to ensure that method performance was not impacted by recurrent mutations.

The performance of RDMs was assessed across a range of mutation rates $m = \{0, 1e-5, 1e-4, 0.001, 0.01, 0.1\}$ and recombination rates $r = \{0, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1\}$; d (dual-infection probability) = {0 and 1}. These parameter ranges were selected to encompass the observed rates of viral evolution (Drake and Holland, 1999; Sanjuán et al., 2010; Hedge, Lycett and Rambaut, 2013) and were extended beyond these observed rates to test the computational limits of the RDMs. Five replicates were conducted for each unique combination of these parameters (m, r, d). The sequence diversity and number of recombinations were reported for simulated populations. We calculated the pairwise sequence diversity for each population using seqinr::dist.alignment (Charif and Lobry, 2007).

2.3 Performance evaluation

To evaluate the performance of the methods, all recombination detection outputs from processed simulated data were classified according to a confusion matrix. Method outputs were classified as one of four cases—true positive, FP, true negative (TN), or false negative (FN). For PhiPack (Profile), the expected and observed cases were determined for each window. The expected case was whether recombination was simulated within the sliding window. The observed case was whether that window tested significant for recombination ($P \leq 0.05$). For the sequence-based methods (3SEQ, GENECONV, UCHIME (VSEARCH), and gmos), the expected and observed cases were determined for each sequence.

The normalised Matthews Correlation Coefficient (nMCC), power, and precision were calculated across simulation parameters according to the cases (Table S1). nMCC was chosen over *F*-score as it is more suited for highly unbalanced cases (Figs. S3–4; (Chicco and Jurman, 2020)). When division by zero occurred when calculating the power or precision, conditions were scored 1. For nMCC, cases with missing conditions were

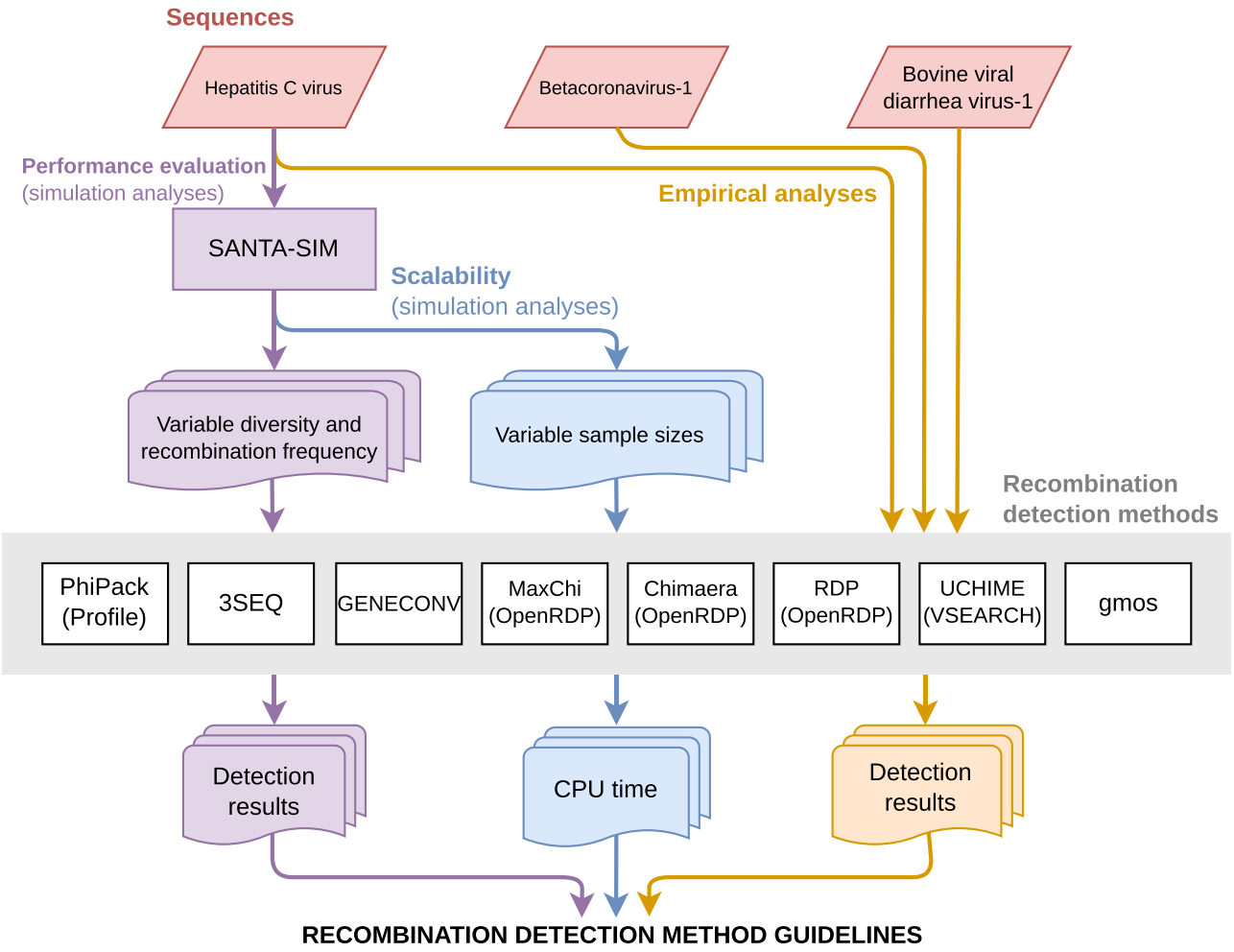

Figure 2. Workflow of simulated and empirical analyses conducted to assess the performance and scalability of eight RDMs.

treated according to the guidelines provided by Chicco and Jurman (2020).

2.4 Scalability

We assessed the scalability of all eight RDMs by recording the central processing unit (CPU) time required to analyse simulated datasets with variable sample sizes $n = 1,000, 5,000, 10,000, 50,000$. Datasets were simulated across a range of mutation rates $m = 0, 1e-5, 1e-4, 0.001, 0.01, 0.1$ and with and without recombination $r = 0, 0.1; d = 1$. Five replicates were conducted for each unique combination of these parameters, and all simulated populations were assessed by each RDM. Analyses were restricted to 24 CPU hours each. All individual analyses were run using a single-core on a 2.2 GHz Intel Xeon Gold 6240R CPU with 16 GB random-access memory (RAM). All analyses were conducted on the UTS eResearch High-Performance Computer Cluster.

2.5 Empirical data

In addition to the simulation analyses, we analysed three empirical datasets using all eight RDMs. The datasets include sequence alignments for positive-sense single-stranded RNA (+ssRNA) viruses. These include the envelope-encoding regions for the HCV ($n = 5479$), the open reading frames ORF1a and ORF1b of Betacoronavirus-1 (Beta-CoV-1; $n = 23$), and the whole genome of Bovine viral diarrhea virus-1 (BVDV-1; $n = 34$). The ORF1ab for Beta-CoV-1 was selected as the 3' can present challenges with alignment. We retrieved the Beta-CoV-1 and BVDV-1 sequences from National Center for Biotechnology Information (NCBI) GenBank and aligned them with MAFFT v7 using default settings (Katoh, Rozewicki and Yamada, 2019). We used the HCV dataset described earlier (from Ho et al. (2017)) from the simulation analyses. Datasets ranged in their sampling approach, sample size, and pairwise sequence distances to test a variety of applications (Table 2). We calculated the pairwise sequence distances with seqinr::dist.alignment.

We constructed separate phylogenies using the whole genome and recombination-free regions for Beta-CoV-1 and BVDV-1 using IQ-TREE v2.1.2 (Minh et al., 2020) with 1,000 ultrafast bootstrap replicates (Hoang et al., 2018) and 1,000 replicates for the Shimodaira–Hasegawa (SH)-like approximate likelihood ratio test (Guindon et al., 2010). Putative recombination-free regions were determined according to the location of detected recombination breakpoints across all methods. We compared the topologies between adjacent recombination-free trees to validate the recombination tests using phytools::cophylo (Revell, 2012).

3. Results and Discussion

3.1 The sequence diversity and recombination frequency of simulated datasets

We assessed the performance of the eight RDMs with simulated data across a range of sequence diversities and recombination frequencies. The pairwise sequence distance increased with

Table 2. Alignment properties of empirical datasets and the number of recombinant sequences detected by each method. Number in brackets indicates the number recombinants identified by gmos that were not between identical sequences and included in analyses. OpenRDP methods were not run on HCV due to computational limits.

	HCV	Beta-Cov-1	BVDV-1
Sampling	Within-host	Between-host	Between-host
Number of sequences	5479	23	34
Alignment length	1680	17,088	11,767
Mean pairwise distance	0.146	0.186	0.386
Max pairwise distance	0.194	0.194	0.482
3SEQ	1	5	25
GENECONV	2001	3	13
RDP (OpenRDP)	–	23	33
MaxChi (OpenRDP)	–	0	33
Chimaera (OpenRDP)	–	23	33
gmos	5479 (0)	23 (10)	34 (16)
UCHIME (VSEARCH)	0	0	0

higher mutation rates across simulated populations (Fig. 3A). A wide range of sequence diversity was simulated, with populations (sequence alignments) that consisted of all identical sequences ($m = 0$) to highly divergent populations with the largest pairwise distance of 0.89 ($m = 0.1$). This extends the pairwise sequence distance assessed in previous benchmarking studies of approximately 0.45 (Smith and Smith, 1998; Posada and Crandall, 2001; Brown et al., 2001; Boni, Posada and Feldman, 2007), allowing for novel insights at higher sequence diversities of methods that have been assessed prior (PhiPack, 3SEQ, GENECONV, RDP, MaxChi, and Chimaera).

3.2 Assessment of RDMs

Our analysis of simulated and empirical data shows that RDMs vary in their ability to handle sequencing data with different sequence diversities and recombination frequency. Notably, there is an inverse relationship between the power and precision of all methods (Figs. 4 and 5). This is due to the simulation scheme, where only a small proportion of recombinant sequences are simulated per population Fig. 3. For example, methods 3SEQ, GENECONV, RDP, MaxChi, Chimaera, and UCHIME generally detected more recombination as sequences become more diverse. As there are few recombinants, these were FP detections which decreased the precision.

While none of the assessed methods are perfectly suited for analysing large-scale viral sequencing data, we provide insights into the trade-offs between scalability, the analytical approach, and dataset-specific performance of the methods, Table 3.

3.2.1 *PhiPack (Profile)*

PhiPack (Profile) scored the highest average nMCC (≥ 0.75; Fig. 4A) when analysing similar sequences with minimal recombination ($m = \{2e-5, 1e-4\}, r \leq 0.005$). This may suggest that homoplasy methods could be suited for analysis of sequences with a pairwise distance of as little as 0.04 (Fig. 3a), extending a previously reported range for homoplasy methods of 0. 1–0.22 (Smith and Smith, 1998). Further analyses are needed to assess whether this limit is affected by fewer or more taxa. PhiPack (Profile) had low power between $m = \{1e-5, 1e-4\}$, congruent with observations that the pairwise homoplasy index (PHI) test is conservative within this range (Bruen, Philippe and Bryant, 2006).

When more recombination was present in the data ($r \geq 0.05$), the nMCC was highest when sequences were divergent ($m = \{1e-3, 1e-2\}$). The precision of PhiPack (Profile), defined as the rate of correctly identified breakpoints (Table S2), was unaffectd across recombination rates, and the average precision was highest (≥ 0.75) between $m = \{1e-5 - 1e-4\}$ and gradually declined as the mutation rate increased.

PhiPack (Profile) was the third most scalable method and did not complete any analyses within 24 h when $n = 50,000$ (Fig. 6).

3.2.2 *3SEQ and GENECONV*

3SEQ and GENECONV required a minimum pairwise distance of approximately 0.2 ($m \approx 0.001$) to detect recombination (Figs. 3A and 5). Recombination was detected most frequently at $m = 0.01$, with a few detections at $m = \{0.001, 0.1\}$ (Fig. S4). At $m = 0.1$, less recombination was detected as the recombination signal was weakened due to the high divergence between sequences. High precision was maintained between $m = \{0, 1e-5, 1e-04\}$ as no recombination was detected.

3SEQ and GENECONV performed well at $r = \{0, 0.001\}$ as no recombination was simulated nor detected. At $r = 0.005$, recombination was detected by both methods, with the nMCC highest at $m = 0.001$. GENECONV generally scored higher nMCC than 3SEQ at $r = \{0.005, 0.01, 0.05\}; m = \{0.01, 0.1\}$. At $r = \{0.005, 0.01\}$, 3SEQ is more powerful at higher mutation rates ($m = \{0.01, 0.1\}$), but GENECONV has higher precision.

We find that the average nMCC and power of 3SEQ and GENECONV declined with more simulated recombination in the sequence alignments (Fig. 5). This conflicts with previous benchmarks where detection methods increased in power with increasing recombination (Posada and Crandall, 2001; Boni, Posada and Feldman, 2007). This disparity could be due to the differences in the simulation process. Here, the recombination rate (r) and dual-infection probability (d) parameters control both the number of sequences that undergo recombination, as well as the number of recombination events in those sequences (Figs. 3B, C). This resulted in fragmented sequences (Fig. S2) that could be more difficult for programs to identify correctly due to a weakened recombination signal.

Given sufficient diversity in the dataset, both 3SEQ and GENECONV were likely to identify recombination when it was present, but identification of the exact sequences was poor. In total, 86.26 per cent of sequences identified as recombinant by both methods were FPs (Fig. S4). 3SEQ has been shown to have a low FP rate and be able to recover most recombination events given sufficient diversity (polymorphic sites) in a dataset (Boni, Posada and Feldman, 2007). Interestingly, 99.74 per cent of sequences identified as FPs by 3SEQ had recombinant parents (Table S2), suggesting that 3SEQ could identify recombination in the putative parental sequences (Table S2). However, further evaluation with known parental-recombinant triplets is required to confirm this. A similar trend is observed in the analysis of the empirical BVDV-1 data, where 3SEQ repeatedly identified JN644055_ChinaXinJiang_2011 and JN704144_China_3156 as the parental sequences for putative recombinants (Fig. S6). 3SEQ and GENECONV were the least scalable methods and scaled poorly when analysing

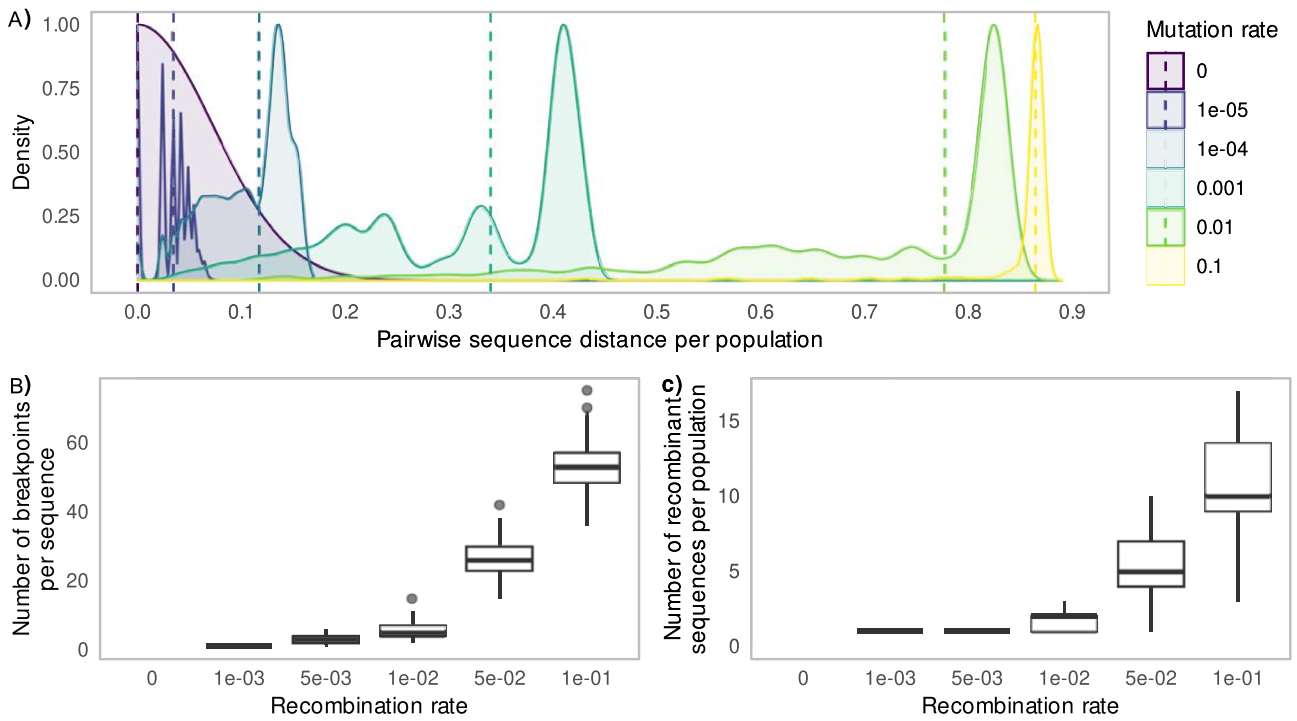

Figure 3. Pairwise distance and breakpoint frequency distributions of simulated sequences. (A) Pairwise sequence distances across mutation rates. Dashed lines indicate the median pairwise distance per mutation rate. (B) Number of breakpoints simulated per sequence across recombination rates. (C) Number of recombinant sequences per population across recombination rates.

divergent sequences, as more informative sites are analysed (Fig. 6).

3.2.3 *RDP, MaxChi, and Chimaera (OpenRDP)*

The performance of all three OpenRDP methods was impacted considerably by sequence divergence, whereas the recombination frequency had a negligible effect. In particular, detection amongst divergent sequences was poor. Recombination was detected in nearly all sequences (Supplementary Fig. X), resulting in high power and low precision ($m \geq 0.001$; Fig. 5). In more similar sequences ($m \leq 1e-4$), the performance was comparable to 3SEQ and GENECONV. At $m = 0.001$, MaxChi detected less recombination and resulted in fewer FPs than RDP and Chimaera (Supp. Fig X). This was reflected in the empirical analyses where MaxChi did not detect any recombination in BCoV-1, where the sequence diversity corresponds to approximately $m = 1e-4$ in the simulated data (Fig. S5). Ample recombination events were detected in BVDV-1 where the sequence diversity is approximate to $m \approx 1e-3$ data.

To improve the precision of the OpenRDP method (therefore reducing the FP detections), a post-processing step on the inferred *P*-values is recommended. Previously suggested techniques include selecting *P*-value peaks to delineate breakpoints or using a hidden Markov model like BURT (Martin et al., 2015). Utilising the boundaries output from another RDM that directly compares sites across sequences, such as GENECONV, has previously been recommended (Smith, 1992; Posada and Crandall, 2001); however, we advise against this as it can overlook the data-specific performance of RDMs (Section 4.4).

Although these approaches will improve the accuracy of the methods, it will further limit the capacity of the OpenRDP methods to analyse moderately sized data sets. RDP, MaxChi, and Chimaera scaled poorly with an increasing number of sequences and polymorphic sites. All three methods were unable to complete analyses of 1,000 sequences above $m = 1e-5$ within 24 h (Fig. 6). However, ongoing testing and development of the OpenRDP package aims to improve both the accuracy and speed of the methods.

3.2.4 *UCHIME (VSEARCH) and gmos*

UCHIME (VSEARCH) and gmos are both alignment-free methods. Interestingly, they were the most scalable methods but both encountered analytical issues. UCHIME (VSEARCH) was the fastest assessed method, completing all analyses within the allocated 24-h restriction (Fig. 6). The longest analysis elapsed 15m 55s (CPU time) at the largest sample size ($n = 50,000$). gmos was the next fastest method after UCHIME (VSEARCH). Similar to PhiPack (Profile), gmos did not complete any analyses within 24 h when $n = 50,000$. Furthermore, the time and resources required for sequence assembly and alignment were not considered in this study, but are expected to provide a further speed advantage in comparison to methods that require aligned sequences.

UCHIME (VSEARCH) did not detect any recombination in any simulated or empirical dataset. As a result, UCHIME (VSEARCH) scored low power and a high average precision in the simulation analyses (Fig. 5). However, the precision would decrease if a higher proportion of recombinant sequences were present in an alignment. UCHIME (VSEARCH) was originally developed to identify chimeric sequences during amplification and was run using default settings. Given the promising scalability of UCHIME (VSEARCH), further work should aim on identifying the optimal parameters (particularly the abundance skew) suited for recombination detection of assembled (viral) sequences, rather than raw reads.

gmos considered all identical subject and query sequences as recombinant, and a thorough assessment could not be conducted

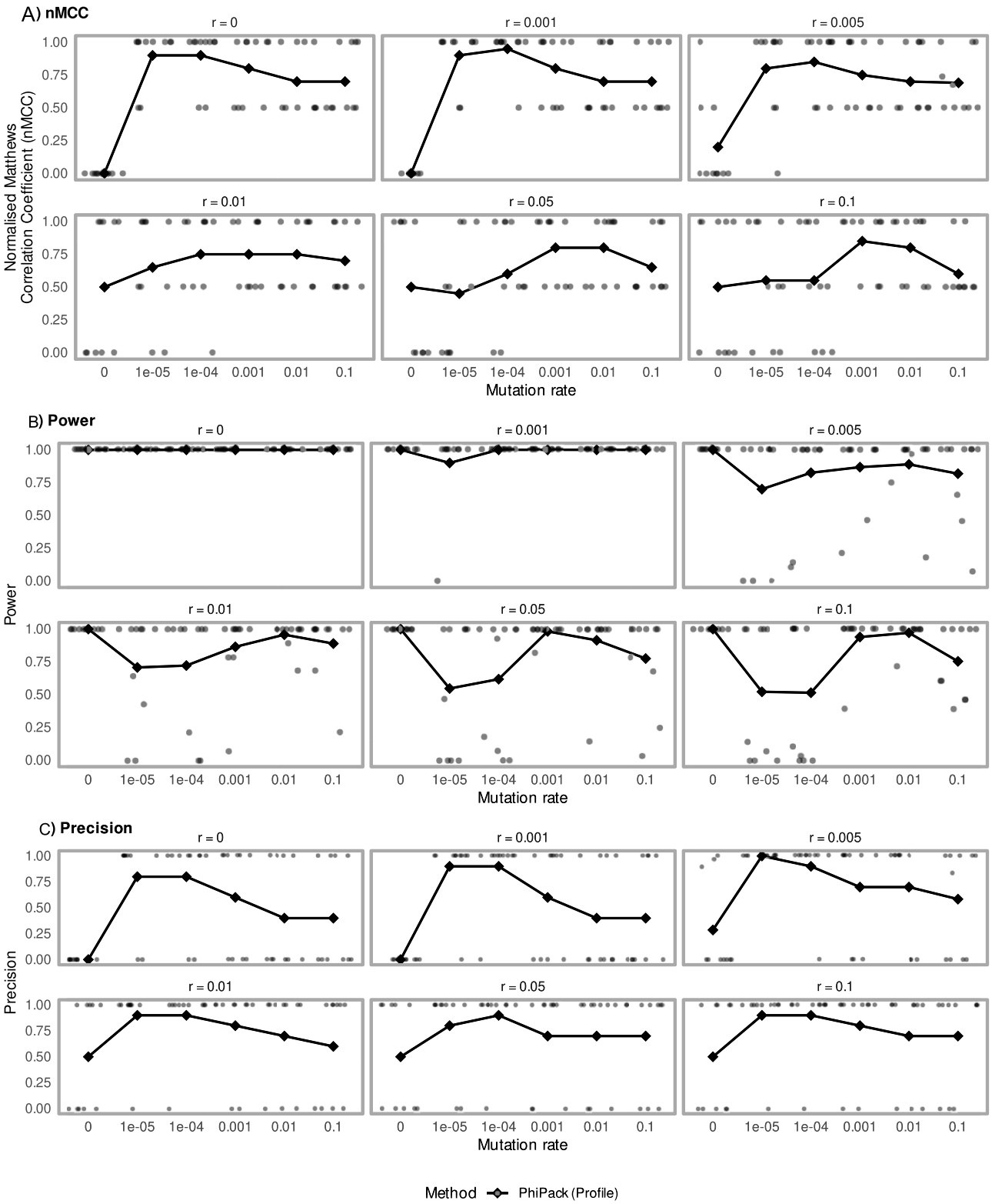

Figure 4. Performance of PhiPack (Profile) across mutation rates and recombination rates (r). Scores for the (A) nMCC, (B) power, and (C) precision are presented. Each point represents individual replicate scores, and lines and diamonds represent the mean score across the mutation and recombination rates.

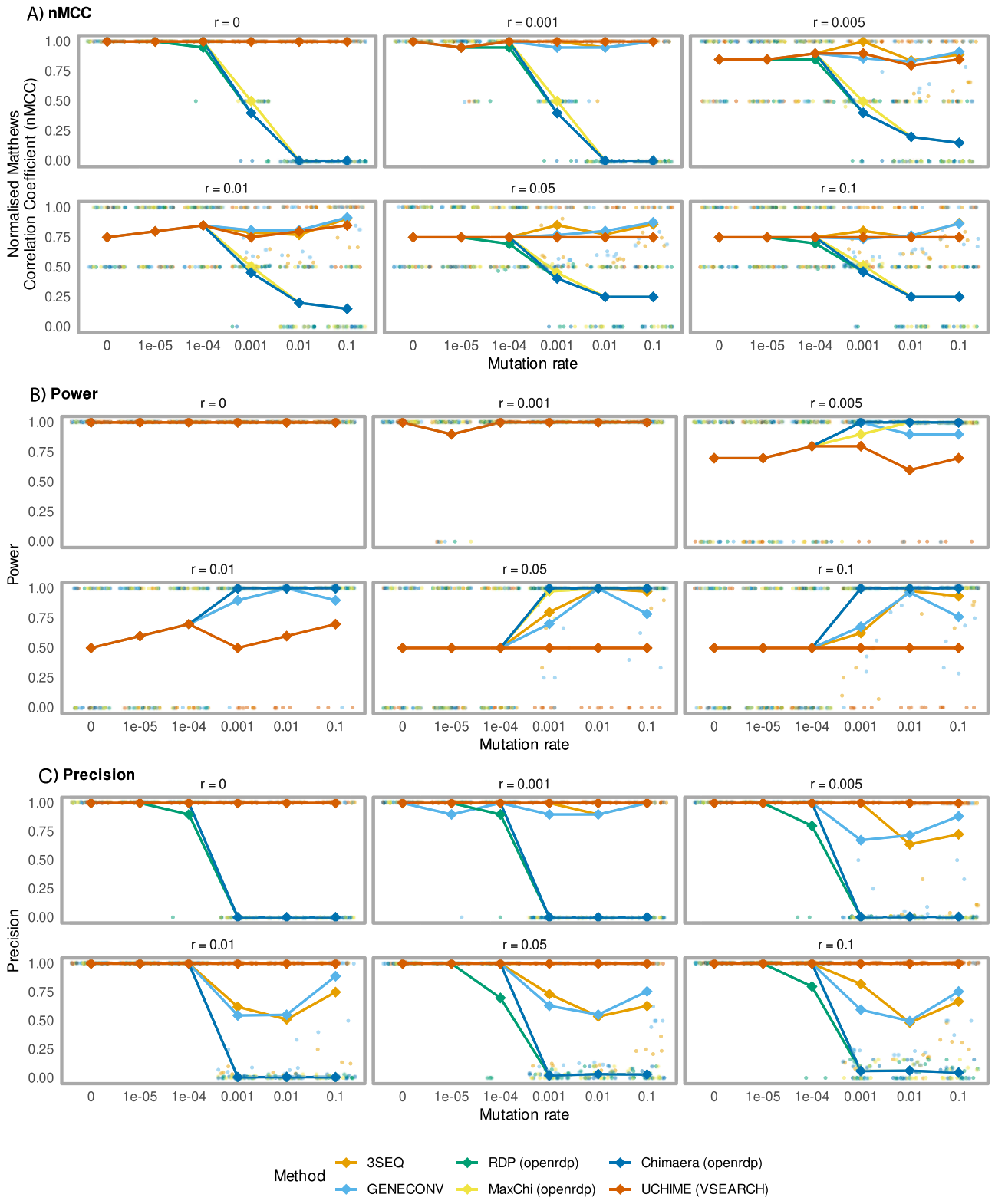

Figure 5. Performance of the sequence-based methods (3SEQ, GENECONV, RDP (OpenRDP), MaxChi (OpenRDP), Chimaera (OpenRDP) and UCHIME (VSEARCH)) across mutation rates and recombination rates (r). Scores for the (A) nMCC, (B) power, and (C) precision are presented. Each point represents individual replicate scores and lines and diamonds represent the mean score across the mutation and recombination rates.

Table 3. Summary of RDM performance and scalability. ★ indicates the method performed well at this sequence divergence (average *nMCC* ≥ 0.75; Figs. 4 and 5) or completed all analyses within 24 h (Fig. 6). Triangles indicate that the method performed well at either lower (▽; $r \leq 0.005$) or higher (△; $r \geq 0.05$) recombination rates.

Simulated mutation rate	Median sequence distance	PhiPack (Profile)	3SEQ	GENECONV	RDP (OpenRDP)	MaxChi (OpenRDP)	Chimaera (OpenRDP)	UCHIME (VSEARCH)	gmos
$m = 0$	0								
$m = 10^{-5}$	0.04	▷	★	★ ★ ▷	★ ★	★ ★	★ ★	★ ★ ★ ★	
$m = 10^{-4}$	0.11	▷	★	★ ★ ▷	★ ★	★ ★	★ ★	★ ★ ★ ★	
$m = 10^{-3}$	0.34	◁	★	▷ ★ ★					
$m = 10^{-2}$	0.79	◁	▷	★ ★					
$m = 10^{-1}$	0.87		★	★					
Sequence number									
$n = 1000$		★	★	★					★
$n = 5000$		★	★						★
$n = 10,000$		★							★
$n = 50,000$									

with simulated data. For empirical analyses, only recombination results between unique sequences were considered. gmos may not be suited for analysis of datasets with an abundance of unique sequences, such as for within-host or dense sampling and better suited for analysis of distinct subject-query sequences as intended.

3.3 Empirical analyses

We analysed three empirical datasets of HCV, Beta-CoV-1, and BVDV-1 (Table 2) using all eight RDMs. 3SEQ, and GENECONV detected recombination in sequences above a pairwise sequence distance of 0.1, whereas gmos detected recombination with a distance below 0.1 (Fig. S5). Phylogenetically, 3SEQ and GENECONV detections were between paraphyletic sequences and gmos detections were common between monophyletic sequences (Fig. S6).

The specific performance of each method is reflected in the vastly different locations of inferred breakpoints. This highlights the need to consider the dataset-specific performance of RDMs when interpreting their results. To illustrate this, we assess the results for each empirical dataset according to the simulation analyses.

3.3.1 *HCV*

GENECONV identified 2,001 sequences as recombinant (Table 2). The sequence diversity range detected (Fig. S5) corresponded approximately with $m = 1e-04$ in the simulation analyses (Fig. 3A), where precision is high across all recombination rates, although some recombination may be missed (Fig. 5). **Interestingly, all the inferred breakpoints spanned the 5' end of the E1 gene region (between positions 110 and 175) and the 3' end of the E1 gene region and hypervariable region within E2 (HVR1; positions 546-621). The E1 and E2 regions are known to be under strong purifying selection, and in turn reduce the sequence diversity (Raghwani et al., 2019). On the other hand, the HVR1 region undergoes strong positive selection, allowing for the rapid acc-quisition of new, diverse variants. Here, GENECONV is likely identifying the regions to be more significantly more similar than the highly diverse HVR1 (Sawyer, 1989).** 3SEQ detected one putative **recombination event**. There may be insufficient variation in serially sampled sequencing for an adequate parental signal.

3.3.2 *Betacoronavirus-1*

The pairwise distances of Beta-CoV-1 sequences were between 0.02 and 0.2 (Fig. S5), corresponding to $m = \{1e-5, 1e-4\}$ in the simulation analyses (Fig. 3). Within this range, recombination detected by PhiPack (Profile), 3SEQ, and GENECONV is likely to be correct, but lack the power to recover all recombination breakpoints (Figs. 4 and 5). Furthermore, the phylogenies delimited by breakpoints identified by 3SEQ and GENECONV all produced incongruent topologies ($t1-t2$, $t4-t5$, $t5-t6$ signal topologies; Fig. S6). On the contrary, comparison of trees $t2$ and $t3$, delimited by a breakpoint identified by gmos, had identical topologies.

3.3.3 *BVDV-1*

In comparison to Beta-CoV-1, PhiPack (Profile) detected a higher proportion of recombinant windows in BVDV-1 (Fig. 7). This aligns with the simulated data ($m \approx 1e-3$; Fig. 3A), where PhiPack (Profile) is powerful across divergent sequences; however, it is uncertain if the detected windows are correct as the precision varies across recombination rates (Fig. 4). According to simulated analyses, 3SEQ and GENECONV detections may be incorrect due to the

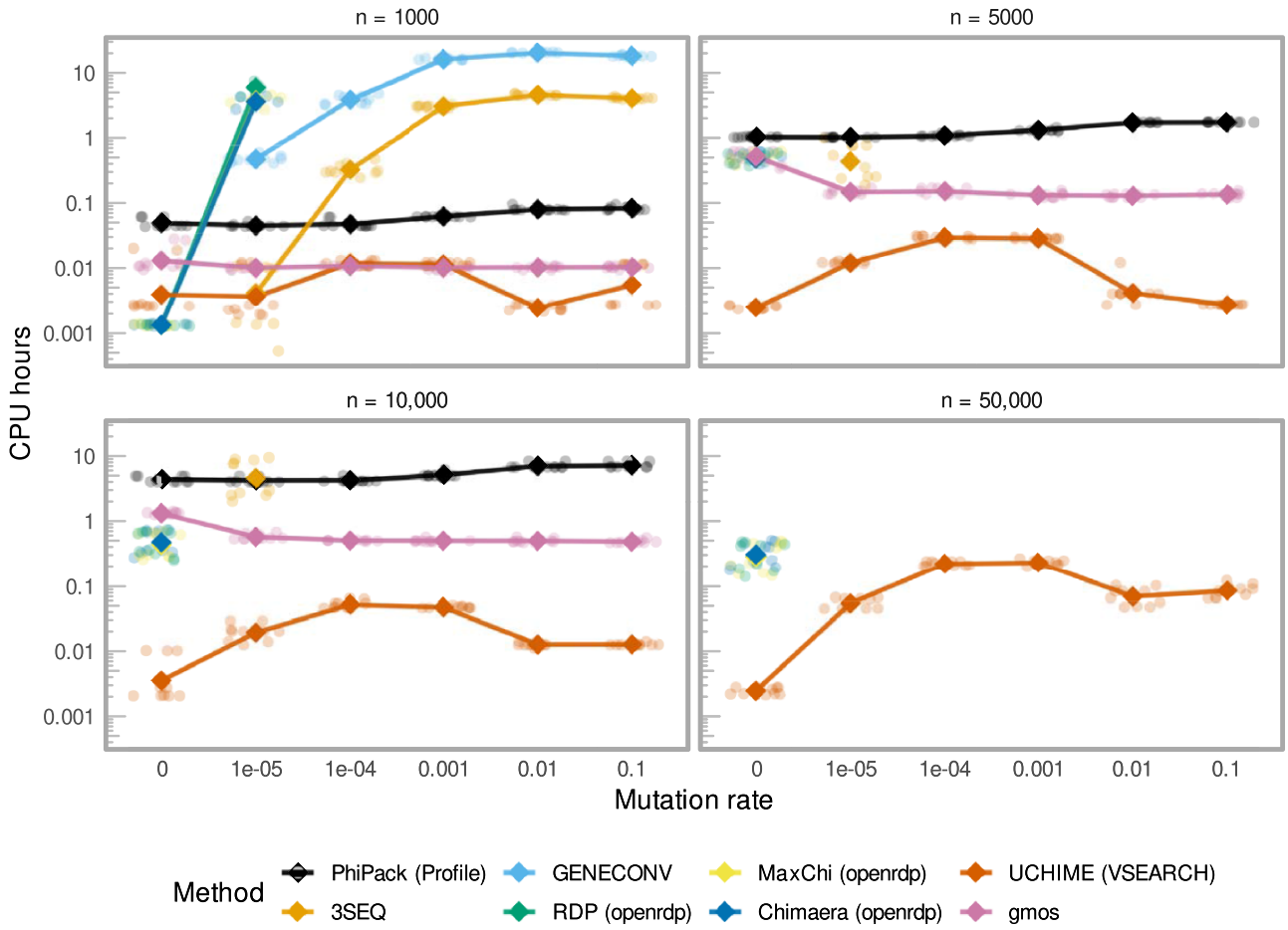

Figure 6. Scalability of the RDMs. The CPU time (log-scaled) required to process simulated populations with a varying number of sequences (n) across mutation rate is presented. Points represent the CPU time required to process individual replicates and lines, and diamonds represent the mean time per sample size (n) and mutation rate. Failed analyses (i.e. 3SEQ and GENECONV at $m = 0$) and analyses that were not completed within 24 h were omitted.

decreased precision when analysing divergent sequences (pairwise sequence distance 0.4; Fig. 5). However, topologies were incongruent between trees delimited by breakpoints identified 3SEQ and GENECONV ($t2 - t3$, $t3 - t4$; Fig. S8).

Strikingly, 80 per cent of sequences identified as recombinant by 3SEQ had both parental sequences as JN644055 and JN704144. Analysis of non-recombinant regions delimited by the breakpoints identified by 3SEQ revealed that sequences JN644055 and JN704144 clustered in different clades (Fig. S8), suggesting that recombination has occurred in the NS4B or NS5A region in this lineage.

3.3.4 *Detecting recombination in highly similar sequences*

Detecting recombination between highly similar sequences has been a persistent challenge, as existing methods lack the required power (Posada and Crandall, 2001; Martin, Lemey and Posada, 2011; Pérez-Losada et al., 2020; Richard et al., 2020). **Although it is impossible for sequence-based methods to identify recombination between identical sequences, it is crucial to note the erroneous behaviour of some methods.** When all sequences in an alignment were identical, PhiPack (Profile) considered all windows as recombinant (Fig. S3); 3SEQ and GENECONV were unable to run due to a lack of polymorphic sites (Fig. S4). gmos deemed all identical pairs of subject and query sequences as recombinants and was not evaluated further for simulated data.

New methods have been developed for the analysis of viral sequencing data with low diversity by incorporating genealogical information (Van Insberghe et al., 2020; Ignatieva, Hein and Jenkins, 2021; Varabyou et al., 2021; Turakhia et al., 2022). However, these methods may only be effective for the analysis of densely sampled sequences where homoplasy is unlikely to be present in the data. Recombination between highly similar regions may continue to elude genomic approaches and is important to continue incorporating known biological information to inform recombination analyses. For example, considering in vivo studies (Giorgi et al., 2021; Ignatieva, Hein and Jenkins, 2021) and drawing upon epidemiological information such as the plausibility for co-infection (Ingle, Howden and Duchene, 2021; Lytras et al., 2022).

3.3.5 *Impact of selection on recombination detection*

Recombination analysis of HCV revealed a high number of recombinant sequences detected, with breakpoints spanning regions of contrasting per-site diversity mediated by opposing selective pressures. In general, the location of recombination breakpoints in empirical populations is non-random and concentrated in similar positions. Such examples include the HCV HVR1 region (Raghwani et al., 2019) and Spike region in coronaviruses (Klerk et al., 2022).

The impact of selective pressures was not explored in our simulation scheme, and the random distribution of simulated

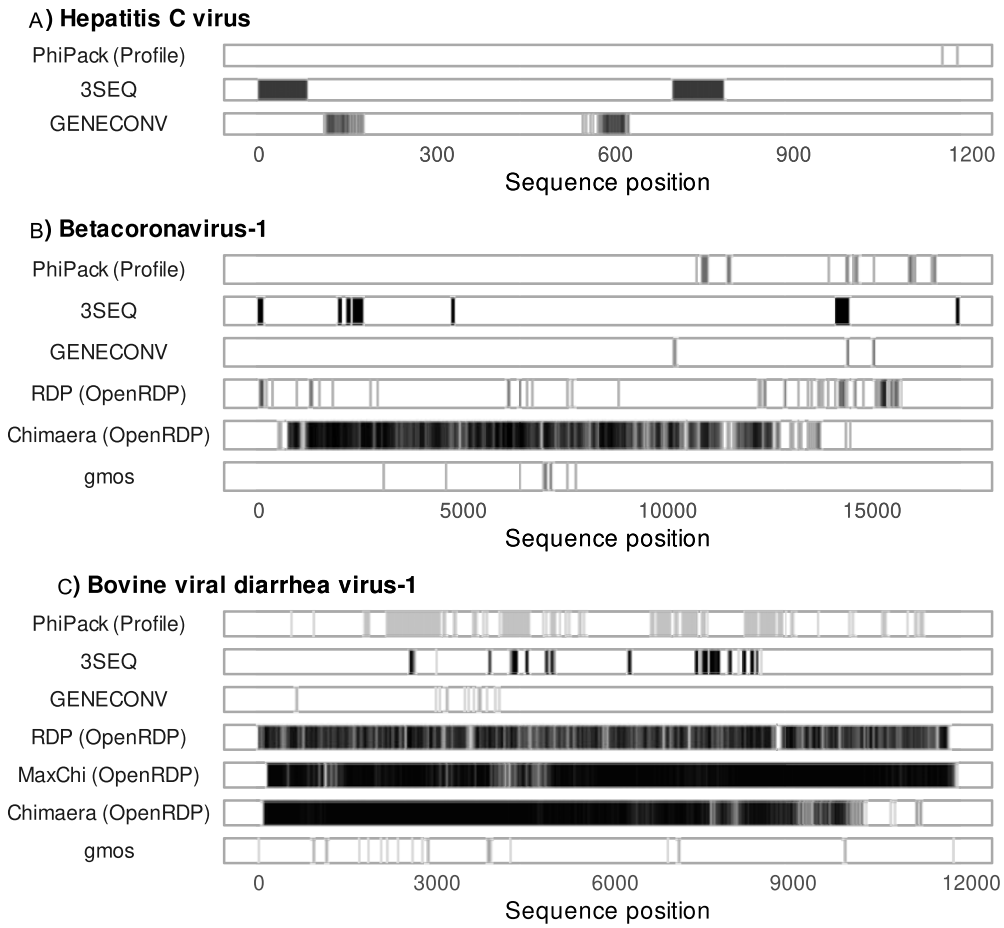

Figure 7. Breakpoint detection of empirical datasets by the eight RDMs. Vertical bars indicate the detected breakpoint locations for each method with significance $P < 0.05$. Opaque regions indicate a high frequency of breakpoints. UCHIME did not identify any recombination breakpoints. The three OpenRDP methods (RDP, MaxChi, Chimaera) were not run on HCV due to computational limitations. MaxChi did not detect any recombination in BCoV-1.

breakpoints may have limited the power of the methods that rely on a strong recombination signal between neighbouring sites. Potentially, this could be causing the misidentification of 3SEQ parentals as the recombinant.

Purifying selection is known to restrict the diversity of the genetic regions it acts upon and can bias broader evolutionary estimations based on the genetic distance (Ewing and Jensen, 2016; Wertheim and Kosakovsky Pond, 2011). Simulation schemes that account for selection and its impact of per-site variability are therefore important areas of future research.

3.4 Practical guidelines for using RDMs in a pandemic-scale era

Using multiple RDMs has been the predominant approach in determining the confidence of inferred recombination breakpoints (Martin et al., 2015). However, this approach can overlook the dataset-specific performance and suitability of methods. Therefore, we propose a guideline for selecting and validating RDMs according to the three key properties of methods: (1) the scalability, (2) analytical approach, and (3) expected performance for a given dataset. A summary of the performance and scalability of each method is presented in Table 3.

First, the size of the dataset should be considered. For example, UCHIME (VSEARCH), gmos and PhiPack (Profile) are more suited for analysis of a dataset consisting of 10, 000–50,000 sequences (Fig. 6). Users can alternatively select a reduced number of representative sequences for analysis with less scalable methods. For example, in densely sampled viruses such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), sequences can be selected according to the known parental sequences (Lytras et al., 2022; Tamura et al., 2023). However, downsampling may not be feasible for methods such as UCHIME (VSEARCH), which rely on sequence abundance data.

Second, the analytical approach of the method should be appropriate for the intended research question or application. A window-based method like PhiPack (Profile) is sufficient to identify alignment-wide breakpoints, such as for partitioning an alignment to analyze recombination-free regions. Sequence-based methods (3SEQ, GENECONV, UCHIME (VSEARCH), and gmos) are necessary when breakpoints need to be identified in exact sequences, such as when recombinant lineages need to be inferred. The method-specific limitations discussed earlier should also be considered.

Lastly, the sequence diversity and expected recombination frequency of the dataset should be determined (by drawing upon *in vivo* studies) to inform the validity of recombination detection outputs. For example, 3SEQ and GENECONV results are expected to be more accurate than PhiPack (Profile) results for sequences with a pairwise distance of 0.4 and two recombination breakpoints (Figs. 3, 4 and 5). Additionally, knowing the

diversity of the dataset can account for FNs as methods may be restricted to analysis of sequence diversity ranges (Figs. 4 and 5, Fig. S5).

These guidelines may apply for analysis of non-viral sequencing data as well. The parameters for the simulated datasets were not specific to +ssRNA viral evolution, encompassing a wide range of mutation and recombination rates that may fall within the expected ranges of other organisms.

3.5 Future work

Further work in evaluating the performance of these eight methods should focus on the accuracy of the methods such as the impact of sequencing error (Turakhia et al., 2020), the accuracy of breakpoint detection, and the power to recover recombination with recurrent mutation (Chan, Beiko and Ragan, 2006). Furthermore, there remains a plethora of RDMs that have not been benchmarked (Table S5).

Future studies should particularly focus on benchmarking recently developed methods intended to be scalable and are powerful at lower sequence diversities (Van Insberghe et al., 2020; Ignatieva, Hein and Jenkins, 2021; Varabyou et al., 2021; Turakhia et al., 2022). However, comparison of methods may continue to be challenging due to the fundamental differences between them (Martin, Lemey and Posada, 2011).

Traditional RDMs that require aligned sequences can be computationally expensive and time-consuming, limiting their utility for large-scale analyses. Alignment-free approaches provide a scalable option for recombination detection. Therefore, we suggest future development of RDMs to implement alignment-free approaches. Further research could focus on identifying the optimal parameters for UCHIME (VSEARCH) and exploring modifications to gmos, or implement 'shustrings', which enable it to work for pairwise sequence comparisons.

Additionally, we recommend an updated review or resource of current RDMs to assist with the selection of a suitable tool. Previous reports, such as that by Martin et al. (2011) and the website http://bioinf.man.ac.uk/robertson/recombination/programs.shtml, do not include newly developed methods and retain methods that are no longer available. We also note an underestimation between the reported speed of methods and the maximum number of sequences that can be analysed.

It is important to note that the recombination frequency for most viruses is assumed to be severely underestimated, particularly in understudied viral families. Future biological studies focusing on characterising the complex evolutionary drivers of viral evolution will further assist in informing the selection and application of appropriate bioinformatic tools.

4. Conclusion

The prevalence of pandemic-scale viral sequencing data poses a computational challenge for existing RDMs. Ideally, methods need to be scalable and have appropriate statistical and analytical approaches depending on the dataset and research question. Evaluation of five RDMs using simulated and empirical data revealed critical trade-offs between these criteria, finding that none of the assessed methods are suited for the analysis of large-scale viral sequencing data. The performance impact of recombination frequency, and especially sequence diversity, varies depending on the RDM. Therefore, we emphasise the importance of understanding the particular scenarios where each method is accurate for, in place of accepting putative recombination events according to the number of methods that jointly identify it. Accordingly, guidelines for selecting and validating methods are provided through application to real viral data, a broad simulation space which extends the sequence diversity range than previously explored, and the first unified evaluation of these methods at scale. Continued work to improve how recombination detection is conducted involves the development of scalable methods. Alignment-free approaches provide a promising approach for analysis of large viral sequencing data that are understudied, or cannot be effectively downsampled. On the other hand, there remains a repertoire of methods that have yet to be assessed.

Data availability

Simulated sequence alignments are available at https://doi.org/10.5061/dryad.d7wm37q6f.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

This research was supported by the Australian Government Research Training Program. Computational facilities and support were provided by the University of Technology eResearch High Performance Computer Cluster. The authors would like to thank Sebastian Duchene, Cheong Xin Chan, and two anonymous reviewers for their valuable comments and suggestions.

Conflict of interest: A.E.D. is employed by Illumina Australia Pty Ltd and holds a financial interest in its parent company Illumina Inc.

References

Anisimova, M., R., Nielsen and Z., Yang (2003) 'Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites', *Genetics*, 164: 1229–36.

Arenas, M. and D., Posada (2010) 'The Effect of Recombination on the Reconstruction of Ancestral Sequences', *Genetics*, 184: 1133–39.

Boni, M. F. et al. (2020) 'Evolutionary Origins of the SARS-CoV-2 Sarbecovirus Lineage Responsible for the COVID-19 Pandemic', *Nature Microbiology*, 5: 1408–17.

Boni, M. F., D., Posada and M. W., Feldman (2007) 'An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets', *Genetics*, 176: 1035–47.

Brito, B. et al. (2018) 'A Traditional Evolutionary History of Foot-and-Mouth Disease Viruses in Southeast Asia Challenged by Analyses of Non-Structural Protein Coding Sequences', *Scientific Reports*, 8: 6472.

Brown, C. J. et al. (2001) 'The Power to Detect Recombination Using the Coalescent', *Molecular Biology and Evolution*, 18: 1421–24.

Bruen, T. C., H., Philippe and D., Bryant (2006) 'A Simple and Robust Statistical Test for Detecting the Presence of Recombination', *Genetics*, 172: 2665–81.

Castillo-Ramírez, S. et al. (2011) 'The Impact of Recombination on dN/dS within Recently Emerged Bacterial Clones', *PLoS Pathogens*, 7: e1002129.

Chan, C. X., R. G., Beiko and M. A., Ragan (2006) 'Detecting Recombination in Evolving Nucleotide Sequences', *BMC Bioinformatics*, 7: 412.

Charif D. and J. R. Lobry 2007. 'SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis', In U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo (eds.) *Structural Approaches to Sequence*

Evolution: Molecules, Networks, Populations, pp. 207–32. New York: Springer Verlag.

Chicco, D. and G., Jurman (2020) 'The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation', *BMC Genomics*, 21: 6.

de Klerk, A. et al. (2022) 'Conserved Recombination Patterns across Coronavirus Subgenera', *Virus Evolution*, 8: veac054.

Domazet-Lošo, M. and T., Domazet-Lošo (2016) 'Gmos: Rapid Detection of Genome Mosaicism over Short Evolutionary Distances', *PLoS One*, 11: e0166602.

Drake, J. W. and J. J., Holland (1999) Mutation Rates among RNA Viruses, *Proceedings of the National Academy of Sciences*, 96: 13910–13.

Duffy, S., L. A., Shackelton and E. C., Holmes (2008) 'Rates of Evolutionary Change in Viruses: Patterns and Determinants', *Nature Reviews Genetics*, 9: 267–76.

Ewing, G. B. and J. D., Jensen (2016) 'The Consequences of Not Accounting for Background Selection in Demographic Inference', *Molecular Ecology*, 25: 135–41.

Giorgi, E. E. et al. (2021) 'Recombination and Low-Diversity Confound Homoplasy-Based Methods to Detect the Effect of SARS-CoV-2 Mutations on Viral Transmissibility', *bioRxiv*, 2021.01.29.428535.

Goodwin, S., J. D., McPherson and W. R., McCombie (2016) 'Coming of Age: Ten Years of Next-Generation Sequencing Technologies', *Nature Reviews Genetics*, 17: 333–51.

Guindon, S. et al. (2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology*, 59: 307–21.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–23.

Hedge, J., S. J., Lycett and A., Rambaut (2013) 'Real-Time Characterization of the Molecular Epidemiology of an Influenza Pandemic', *Biology Letters*, 9: 20130331.

Hoang, D. T. et al. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.

Ho, C. K. Y. et al. (2017) 'Characterization of Hepatitis C Virus (HCV) Envelope Diversification from Acute to Chronic Infection within a Sexually Transmitted HCV Cluster by Using Single-Molecule, Real-Time Sequencing', *J. Virol.*, 91.

Ignatieva, A., J., Hein and P. A., Jenkins (2021) 'Investigation of Ongoing Recombination Through Genealogical Reconstruction for Sars-Cov-2', *bioRxiv*, 2021.01.21.427579.

Ingle, D. J., B. P., Howden and S., Duchene (2021) 'Development of Phylodynamic Methods for Bacterial Pathogens', *Trends in Microbiology*, 29, 788–97.

Jariani, A. et al. (2019) 'SANTA-SIM: Simulating Viral Sequence Evolution Dynamics under Selection and Recombination', *Virus Evolution*, 5, vez003.

Katoh, K., J., Rozewicki and K. D., Yamada (2019) 'MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization', *Briefings in Bioinformatics*, 20: 1160–66.

Kosakovsky Pond, S. L. et al. (2006) 'GARD: A Genetic Algorithm for Recombination Detection', *Bioinformatics*, 22: 3096–98.

Lam, H. M., O., Ratmann and M. F., Boni (2018) 'Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm', *Molecular Biology and Evolution*, 35: 247–51.

Lole, K. S. et al. (1999) 'Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination', *Journal of Virology*, 73: 152–60.

Loman, N. J. et al. (2012) 'Performance Comparison of Benchtop High-Throughput Sequencing Platforms', *Nature Biotechnology*, 30: 434–39.

Lytras, S. et al. (2022) 'Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination', *Genome Biology and Evolution*, 14: evac018.

Martin, D. P. et al. (2015) 'RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes', *Virus Evolution*, 1, vev003.

Martin, D. P. et al. (2021) 'RDP5: A Computer Program for Analyzing Recombination in, and Removing Signals of Recombination from, Nucleotide Sequence Datasets', *Virus Evolution*, 7, veaa087.

Martin, D. P., P., Lemey and D., Posada (2011) 'Analysing Recombination in Nucleotide Sequences', *Molecular Ecology Resources*, 11: 943–55.

Martin, D. and E., Rybicki (2000) 'RDP: Detection of Recombination amongst Aligned Sequences', *Bioinformatics*, 16: 562–63.

Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–34.

Padidam, M., S., Sawyer and C. M., Fauquet (1999) 'Possible Emergence of New Geminiviruses by Frequent Recombination', *Virology*, 265: 218–25.

Pérez-Losada, M. et al. (2015) 'Recombination in Viruses: Mechanisms, Methods of Study, and Evolutionary Consequences', *Infection, Genetics and Evolution*, 30: 296–307.

Pérez-Losada, M. et al. (2020) 'High-Throughput Sequencing (HTS) for the Analysis of Viral Populations', *Infection, Genetics and Evolution*, 80: 104208.

Posada, D. and K. A., Crandall (2001) 'Evaluation of Methods for Detecting Recombination from DNA Sequences: Computer Simulations', *Proceedings of the National Academy of Sciences*, 98: 13757–62.

Pybus, O. G. and A., Rambaut (2009) 'Evolutionary Analysis of the Dynamics of Viral Infectious Disease', *Nature Reviews Genetics*, 10: 540–50.

Quick, J. et al. (2016) 'Real-Time, Portable Genome Sequencing for Ebola Surveillance', *Nature*, 530: 228–32.

Raghwani, J. et al. (2019) 'High-Resolution Evolutionary Analysis of Within-Host Hepatitis C Virus Infection', *The Journal of Infectious Diseases*, 219: 1722–29.

Revell, L. J. (2012) 'Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things)', *Methods in Ecology and Evolution*, 3: 217–23.

Richard, D. et al. (2020) 'No Detectable Signal for Ongoing Genetic Recombination in SARS-CoV-2', *bioRxiv*, 2020.12.15.422866.

Rognes, T. et al. (2016) 'VSEARCH: A Versatile Open Source Tool for Metagenomics', *PeerJ*, 4, e2584.

Rousselle, M. et al. (2019) 'Influence of Recombination and GC-Biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals Versus Birds', *Molecular Biology and Evolution*, 36: 458–71.

Sanjuán, R. et al. (2010) 'Viral Mutation Rates', *Journal of Virology*, 84: 9733–48.

Sawyer, S. (1989) 'Statistical Tests for Detecting Gene Conversion', *Molecular Biology and Evolution*, 6: 526–38.

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

Seemann, T. et al. (2020) 'Tracking the COVID-19 Pandemic in Australia Using Genomics', *Nature Communications*, 11: 4376.

Simon-Loriere, E. and E. C., Holmes (2011) 'Why Do RNA Viruses Recombine?', *Nature Reviews Microbiology*, 9: 617–26.

Smith, J. M. (1992) 'Analyzing the Mosaic Structure of Genes', *Journal of Molecular Evolution*, 34: 126–29.

Smith, J. M. and N. H., Smith (1998) 'Detecting Recombination from Gene Trees', *Molecular Biology and Evolution*, 15: 590–99.

Tamura, T. et al. (2023) 'Virological Characteristics of the SARS-CoV-2 XBB Variant Derived from Recombination of Two Omicron Subvariants', *Nature Communications*, 14: 2800.

Turakhia, Y. et al. (2020) 'Stability of SARS-CoV-2 Phylogenies', *PLoS Genetics*, 16: e1009175.

Turakhia, Y. et al. (2022) 'Pandemic-Scale Phylogenomics Reveals the SARS-CoV-2 Recombination Landscape', *Nature*, 609: 994–97.

Van Insberghe, D. et al. (2020) 'Identification of SARS-CoV-2 Recombinant Genomes', *bioRxiv*, 2020.08.05.238386.

Varabyou, A. et al. (2021) 'Rapid Detection of Inter-Clade Recombination in SARS-CoV-2 with Bolotie', *Genetics*, 218: iyab074.

Wertheim, J. O. and S. L., Kosakovsky Pond (2011) 'Purifying Selection Can Obscure the Ancient Age of Viral Lineages', *Molecular Biology and Evolution*, 28: 3355–65.

Wong, T. H. N. et al. (2013) 'Whole Genome Sequencing and De Novo Assembly Identifies Sydney-Like Variant Noroviruses and Recombinants during the winter 2012/2013 Outbreak in England', *Virology Journal*, 10: 1–10.

Xiao, Y. et al. (2016) 'RNA Recombination Enhances Adaptability and Is Required for Virus Spread and Virulence', *Cell Host & Microbe*, 19: 493–503.

Yeşilbağ, K., G., Alpay and P., Becher (2017) 'Variability and Global Distribution of Subgenotypes of Bovine Viral Diarrhea Virus', *Viruses*, 9: 128.