Evaluation of recombination detection methods for analysis of viral sequences

Frederick Jaya, Barbara Brito Rodriguez, Aaron Darling
University of Technology Sydney

1 Recombination-free regions in sequences are identified using recombination detection methods

Unaccounted recombination in sequencing data can mislead evolutionary analyses. For example, phylogenetic branch lengths and topologies can be distorted.

60+ METHODS

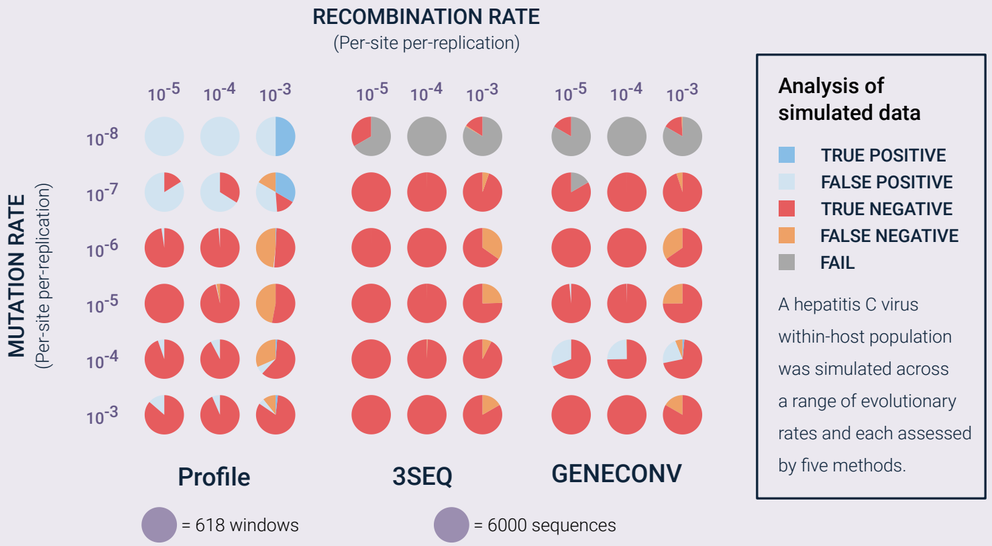
Previous studies have aimed to provide guidelines for the selection of suitable methods, but whether these methods are suitable for analysis of recent sequencing data is unclear.

2 Large viral sequencing datasets

Recent data requiring analysis of many sequences include within-host hepatitis C virus deep-sequencing (*n* ~ 6000) and global SARS-CoV-2 sequences (*n* ~ 60,000). Long-read platforms such as PacBio's Sequel II HiFi are capable of producing ~ 4,000,000 reads.

~6K ~60K ~4M

3 Sequence diversity has an overwhelming effect on recombination detection



Analysis of simulated data

- TRUE POSITIVE
- FALSE POSITIVE
- TRUE NEGATIVE
- FALSE NEGATIVE
- FAIL

A hepatitis C virus within-host population was simulated across a range of evolutionary rates and each assessed by five methods.

Low sequence diversity is problematic

PhiPack (Profile) detected many false positives, whereas 3SEQ and GENECONV were unable to process filles due to a lack of polymorphic sites.

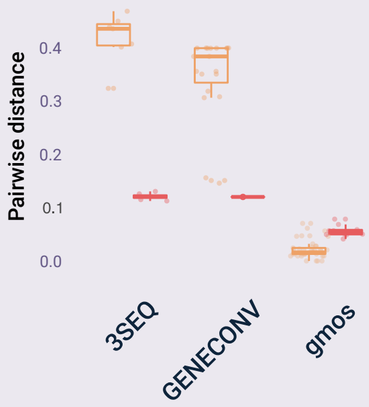
Methods were unable to recover simulated recombination events

Within-host simulations may yield weak recombination signals and parental sequences are omitted due to sequence subsampling.

Identified method-specific behaviour

3SEQ requires both parental sequences in the subsampled dataset. GENECONV output many false positives at a restricted range of sequence diversity.

4 Methods may only detect recombination in limited ranges of sequence diversity

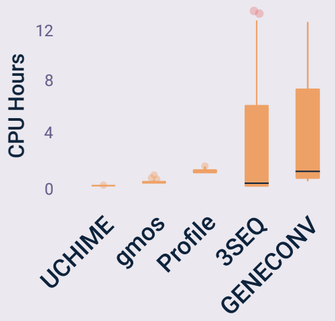


Analysis of empirical data

- BETACORONAVIRUS
- BOVINE VIRAL DIARRHEA VIRUS

Two empirical datasets of bovine RNA viruses were assessed for recombination breakpoints. The pairwise distance of detected recombinants are reported.

5 Scalable methods are not suitable or analysis of within-host data



Simulated alignments (n = 216) containing 5000 sequences were processed by each method and the duration per analysis reported. Although the most scalable, UCHIME failed to detect any recombination, and gmos predicts all identical sequences as recombinant.