Supplementary method of the manuscript

A scalable method for identifying recombinants from unaligned sequences

Qian Feng [1], Kathryn Tiedje [2], Shazia Ruybal [2,3], Gerry Tonkin-Hill [4], Michael Duffy [2], Karen Day [2], Heejung Shim [1], and Yao-ban Chan [1,†]

[1]Melbourne Integrative Genomics / School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3052, Australia

[2]Department of Microbiology and Immunology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville VIC 3052, Australia

[3]Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Parkville VIC 3052, Australia

[4]Parasites and Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom

[†]To whom correspondence should be addressed.

Contents

1 Simulation Details

To simulate amino acid sequences based on the genealogy created in the previous step, two different situations are taken into account.

- Indel rate = 0 (which takes up majority of simulation cases), we use software *pyvolve* [1].
    - We adopt argument *scale_tree* in *pyvolve.read_tree* function to control the level of mutation. It takes a numeric value and multiplies all branch lengths in the tree by this scalar.
    - We employ argument *models* in function *pyvolve.Partition* to take one of empirical amino-acid substitution models, *size* specifies the length of simulated protein sequence.
- Indel rate $\neq 0$, sequence evolver *INDELible* [2] is utilized. It accounts for simulating insertions and deletions except general substitutions in Pyvolve. Although it would be useful to use only one program (e.g. Pyvolve only or INDELible only), no program currently exists for taking all listed factors and all possible values of listed factors. The main components of an indel model include indel rates and base indel fragment size distributions [3].
    - For simplicity, insertion and deletion instantaneous rate are the same, changing from 0.01 to 0.04.
    - Suppose fragment size distribution follows a negative binomial distribution. By increasing the value of distribution parameter q from 0.1 to 0.5 without varying indel rate and fragment size variance, we decrease mean of fragment size from around 10 to 2 gradually.

2 Methods for three-category classification of ups using protein-protein BLAST

References

[1] Stephanie J Spielman and Claus O Wilke. Pyvolve: a flexible python module for simulating sequences along phylogenies. *PloS one*, 10(9):e0139047, 2015.

[2] William Fletcher and Ziheng Yang. Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888, 2009.

[3] Heejung Shim and Bret Larget. Bayescat: Bayesian co-estimation of alignment and tree. *Biometrics*, 74(1):270–279, 2018.