RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets

Darren P. Martin,[1,*,†] Arvind Varsani,[2,3,‡] Philippe Roumagnac,[4] Gerrit Botha,[1] Suresh Maslamoney,[1] Tiana Schwab,[5] Zena Kelz,[1] Venkatesh Kumar,[1,6] and Ben Murrell[6]

[1]Department of Integrative Biomedical Sciences, Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Anzio Road Observatory, Cape Town 7549, South Africa, [2]The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85287-5001, USA, [3]Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Observatory, Cape Town 7701, South Africa, [4]BGPI, Univ Montpellier, CIRAD, INRAE, L'Institut Agro, Montpellier, France, [5]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, witzerland and [6]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

*Corresponding author: E-mail: darrenpatrickmartin@gmail.com

†http://orcid.org/0000-0002-8785-0870

‡https://orcid.org/0000-0003-4111-2415

Abstract

For the past 20 years, the recombination detection program (RDP) project has focused on the development of a fast, flexible, and easy to use Windows-based recombination analysis tool. Whereas previous versions of this tool have relied on considerable user-mediated verification of detected recombination events, the latest iteration, RDP5, is automated enough that it can be integrated within analysis pipelines and run without any user input. The main innovation enabling this degree of automation is the implementation of statistical tests to identify recombination signals that could be attributable to evolutionary processes other than recombination. The additional analysis time required for these tests has been offset by algorithmic improvements throughout the program such that, relative to RDP4, RDP5 will still run up to five times faster and be capable of analyzing alignments containing twice as many sequences (up to 5000) that are five times longer (up to 50 million sites). For users wanting to remove signals of recombination from their datasets before using them for downstream phylogenetics-based molecular evolution analyses, RDP5 can disassemble detected recombinant sequences into their constituent parts and output a variety of different recombination-free datasets in an array of different alignment formats. For users that are interested in exploring the recombination history of their datasets, all the manual verification, data management and data visualization components of RDP5 have been extensively updated to minimize the amount of time needed by users to individually verify and refine the program's interpretation of each of the individual recombination events that it detects.

1. Introduction

Recombination and genome component reassortment are processes that strongly impact the evolution of many virus species. The patterns of nucleotide sequence variation created in virus genomic sequence datasets by these processes (which are hereafter collectively referred to as recombination) can seriously undermine the accuracy of phylogenetics-based molecular evolution analyses (Schierup and Hein 2000a,b; Scheffler, Martin, and Seoighe 2006; Arenas and Posada 2010). It is therefore frequently desirable to test nucleotide sequence datasets for evidence of recombination and, when such evidence is found, take steps to minimize the impacts of detected recombination events on downstream analyses of these datasets.

Of the many available computer programs for analyzing recombination (see http://bioinf.man.ac.uk/robertson/recombina tion/programs.shtml; Martin et al. 2011), recombination detection program (RDP) is presently one of the most commonly used. Successive versions of RDP have applied an expanding array of recombination event detection, recombination breakpoint demarcation, and recombinant sequence identification methods, all applied in unison, to yield detailed descriptions of how recombination may have impacted the evolution of any given set of aligned nucleotide sequences (Martin et al. 2015). The accuracy of these descriptions, however, frequently depended on the amount of effort users were willing to put into exploring the many similarly plausible ways in which the detected patterns of recombination may have arisen. A guiding principle during the development of RDP5, the latest version of the RDP series, has therefore been a minimization of the amount of time that users need to invest in detecting and removing signals of recombination from nucleotide sequence datasets.

2. Generation of recombination-free datasets

RDP5 can take as input an aligned nucleotide sequence dataset, automatically identify and characterize individual recombination events that are evident within that dataset, and output modifications of the input alignment within which all signals of detectable recombination have been removed. The different types of modified recombination-free multiple sequence datasets that RDP5 can output include alignments where: (1) recombinant sequences have been removed; (2) fragments of sequence derived through recombination have been removed; (3), recombinant sequences are split up into their constituent parts; and (4) the input alignment is divided into multiple different gene/genome sub-region alignments based on the locations of detected recombination breakpoints (Supplementary Fig. S1).

3. Query vs reference scans for recombination

In addition to the fully exploratory recombination analysis modes found in previous versions of the program (Martin et al. 2015), RDP5 also includes a new highly automated 'query vs reference' analysis mode such as that found in the programs REGA (de Oliveira et al. 2005) and jpHMM (Schultz et al. 2006). Unlike with the default fully exploratory recombination analysis mode where every sequence is tested for evidence of recombination, the query vs reference mode will test a user-defined set of query sequences for evidence that they originated through recombination between a user-defined set of reference sequences. Such an analysis mode is well suited to analyzing patterns of recombination between two or more groups of viruses that have only recently had the opportunity to start recombining with one another: such as in an individual patient that has been infected with two distinct variants of a virus (Sheward et al. 2018), or within a geographical region where multiple distinct genetic variants of a virus have recently started co-circulating.

To minimize the amount of effort needed to define reference sequences and/or groups of reference sequences, RDP5 will automatically identify sequences as queries and references based on simple sequence naming rules (see the manual provided with RDP5 for these rules). These naming rules can also be used to group reference sequences into different reference types.

4. Automated sequence annotation

Given an accessible internet connection, RDP5 will automatically annotate the genomic features of input virus sequence datasets using the curated NCBI virus reference sequence database (https://www.ncbi.nlm.nih.gov/genome/viruses/). Annotation is useful in the context of generating recombination-free datasets because it enables RDP5 to output gene sequence alignments that are suitable for downstream codon-focused analyses of natural selection. The gene sequence alignments that RDP5 produces can account for variation in the positions of gene start and stop codons, intron splicing (wherever intron donor and acceptor sites are annotated in the NCBI reference sequence records), portions of genes that are expressed in two or more different open reading frames (these regions can be excluded) and real and/or artifactual frame-shift mutations (partial codons can be removed).

Sequence annotation is also useful for testing the selective processes that impact patterns of recombination detected within sequence datasets. Given a set of annotated genome sequences recombination breakpoint distribution tests in RDP5 will now automatically detect whether observed recombination breakpoint distributions vary between: (1) coding and noncoding regions; (2) different genes; and (3) the edges and internal regions of genes.

5. Detection of potential false-positive recombination signals

A crucial facilitator of the highly automated recombination analyses that RDP5 can perform is the inclusion of tests that are specifically designed to detect and flag potential false-positive signals of recombination. Besides automatically testing whether each of the recombination signals that RDP5 detects might be attributable to sequence misalignment (which is a major contributor to false positive signals of recombination), the program will also detect and flag as suspicious any detectable recombination signals that may have arisen through evolutionary processes other than recombination. Specifically, RDP5 uses the PHI test (Bruen, Philippe, and Bryan 2006) and 4-gamete test (McVean, Awadalla, and Fearnhead 2002) adapted versions of the homoplasy test (Maynard Smith and Smith 1998) to flag apparent recombination signals that are potentially attributable to a combination of inter-lineage and inter-site mutation-rate variation rather than recombination (Bertrand, Johansson, and Norberg 2016).

6. RDP5CL: a command-line version of RDP5

For instances where users would like to integrate RDP5 into an analysis pipeline, a separate command-line driven version of the program, RDP5CL, is distributed with RDP5. RDP5CL will take

as input a multiple sequence alignment in any standard alignment file format and, contingent on command line switches, output any of various different types of recombination-free alignments, recombination breakpoint distribution plots and/or maximum likelihood (Stamatakis 2014) phylogenetic trees accounting for recombination.

7. Improved computational performance

Despite the additional tests that RDP5 carries out during the characterization of detected recombination events, it is still able to analyze any given dataset between two and five times faster than RDP4 could (Supplementary Table S1). To achieve this, RDP5 implements a multitude of algorithmic improvements such as multi-core CPU-level parallelization and intensive use of lookup tables for bootstrap, permutation, likelihood and probability calculations. Additionally, while previous versions of RDP were restricted to using 2 GB of available RAM (which is standard for 32-Bit Windows programs), RDP5 can utilize 4 GB of RAM even under 32-bit Windows operating systems. This, together with improvements in memory management, means that RDP5 can analyze datasets that contain up to twice as many sequences that are each five times longer than those which RDP4 could manage.

The improved computational performance of RDP5 also extends to the manual recombination-signal verification, data management and data visualization components of the program. These enhancements will substantially decrease the amount of time that it takes users to manually verify and refine RDP5's interpretation of each of the individual recombination events that it detects.

8. Operational limits

RDP5 will work on computers with Windows 7/Vista/8/10 operating systems and can also be installed to run under Windows 7 emulators running on computers with MacOSx and UNIX operating systems.

RDP5 can be used to productively analyze datasets containing up to 400 million nucleotides within 24 h on a standard 8-core 2.5 GHz processor with >4 GB of RAM. Such datasets might, for example, consist of 120 3.0-Mb-long bacterial genome sequences, or 4000 10-kb-long viral genome sequences. With default program settings, RDP5 can analyze 100 10-kb-long sequences in under 5 min on a standard desktop computer.

Availability

RDP5 is available for free download from http://web.cbio.uct.ac.za/∼darren/rdp.html. It is distributed with an extensive manual that contains (1) detailed descriptions of the various methods implemented in the program; (2) a step-by-step guide describing how to create and analyze datasets for recombination detection; (3) instructions on how to run completely automated analyses from the command-line using RDP5CL; and (4) information on how to run RDP5 on Mac and Linux computers.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Arenas, M., and Posada, D. (2010) 'The Effect of Recombination on the Reconstruction of Ancestral Sequences', *Genetics*, 184: 1133–9.

Bertrand, Y. J. K., Johansson, M., and Norberg, P. (2016) 'Revisiting Recombination Signal in the Tick-Borne Encephalitis Virus: A Simulation Approach', *PLoS One*, 11: e0164435.

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

de Oliveira, T. et al. (2005) 'An Automated Genotyping System for Analysis of HIV-1 and Other Microbial Sequences', *Bioinformatics*, 21: 3797–800.

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

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

Maynard Smith, J., and Smith, N. H. (1998) 'Detecting Recombination from Gene Trees', *Molecular Biology and Evolution*, 15: 590–9.

McVean, G., Awadalla, P., and Fearnhead, P. (2002) 'A Coalescent-Based Method for Detecting and Estimating Recombination from Gene Sequences', *Genetics*, 160: 1231–41.

Scheffler, K., Martin, D. P., and Seoighe, C. (2006) 'Robust Inference of Positive Selection from Recombining Coding Sequences', *Bioinformatics*, 22: 2493–9.

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

——, and —— (2000b) 'Recombination and the Molecular Clock', *Molecular Biology and Evolution*, 17: 1578–9.

Schultz, A.-K. et al. (2006) 'A Jumping Profile Hidden Markov Model and Applications to Recombination Sites in HIV and HCV Genomes', *BMC Bioinformatics*, 7: 265.

Sheward, D. J. et al. (2018) 'HIV Superinfection Drives De Novo Antibody Responses and Not Neutralization Breadth', *Cell Host & Microbe*, 24: 593–9.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–3.