BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2462–2463
doi:10.1093/bioinformatics/btq467

Sequence analysis

Advance Access publication August 26, 2010

RDP3: a flexible and fast computer program for analyzing recombination

Darren P. Martin[1,2,*], Philippe Lemey[3], Martin Lott[1,2,4], Vincent Moulton[4], David Posada[5] and Pierre Lefeuvre[1,6]

[1]Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, [2]Centre for High Performance Computing, Rosebank, Cape Town, South Africa, [3]Department of Microbiology and Immunology, Rega Institute, K.U. Leuven, Belgium, [4]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK, [5]Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain and [6]CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Ligne Paradis, La Réunion

Associate Editor: Martin Bishop

ABSTRACT

Summary: RDP3 is a new version of the RDP program for characterizing recombination events in DNA-sequence alignments. Among other novelties, this version includes four new recombination analysis methods (3SEQ, VISRD, PHYLRO and LDHAT), new tests for recombination hot-spots, a range of matrix methods for visualizing over-all patterns of recombination within datasets and recombination-aware ancestral sequence reconstruction. Complementary to a high degree of analysis flow automation, RDP3 also has a highly interactive and detailed graphical user interface that enables more focused hands-on cross-checking of results with a wide variety of newly implemented phylogenetic tree construction and matrix-based recombination signal visualization methods. The new RDP3 can accommodate large datasets and is capable of analyzing alignments ranging in size from 1000×10 kilobase sequences to 20×2 megabase sequences within 48 h on a desktop PC.

Availability: RDP3 is available for free from its web site http://darwin.uvigo.es/rdp/rdp.html

Contact: darrenpatrickmartin@gmail.com

Supplementary information: The RDP3 program manual contains detailed descriptions of the various methods it implements and a step-by-step guide describing how best to use these.

Received on July 5, 2010; revised on August 8, 2010; accepted on August 10, 2010

RDP3 is a computer program for statistical identification and characterization of historical recombination events. Given a set of aligned nucleotide sequences, RDP3 will rapidly analyze these with a range of powerful non-parametric recombination detection methods (including BOOTSCAN, MAXCHI, CHIMAERA, 3SEQ, GENECONV, SISCAN, PHYLPRO and VISRD; Boni *et al*., 2007; Gibbs *et al*., 2000; Lemey *et al*., 2009; Padidam *et al*., 1999, Posada and Crandall, 2001; Weiller, 1998). It will provide a detailed breakdown of recombination breakpoint locations, and the identities of recombinant and parental sequences. For further downstream analyses, the program enables users to save edited sequence alignments with (i) recombinant sequences removed; (ii) recombinationally derived tracts of sequence removed; or (iii) recombinant sequences split into their constituent parts.

An important strength of RDP3 that makes it applicable to a variety of recombination analysis problems is that, unlike many other recombination detection programs such as SIMPLOT (Lole *et al*.,1999), DUAL BROTHERS (Minin *et al*., 2005), jpHMM (Schultz *et al*., 2006) or SCUEAL (Kosakovsky *et al*., 2009), it does not screen predefined sets of potentially recombinant (or query) sequences against other predefined sets of non-recombinant (or reference) sequences. RDP3 instead treats every sequence within an input alignment as a potential recombinant and systematically screens large numbers of sequence triplets and/or quartets to identify sets of three or four sequences that contain a recombinant and two sequences resembling its parents. Such an approach means that RDP3 can simultaneously detect the entire scope of recombination evident within a dataset (i.e. not just that occurring between the reference strains or species) enabling its use in the characterization of complex recombinants such as those derived through recombination between parental sequences that were themselves recombinant. The drawback of such a flexible, exploratory framework is that it can often be difficult to assess the uncertainty associated with inferred recombination patterns. However, with its wide range of cross-checking tools, RDP3 is complementary to probabilistic recombination analysis approaches.

1 NEW FEATURES IN RDP3

Although the graphically intensive and highly interactive RDP3 interface remains superficially unchanged from that of its predecessor, RDP2 (Martin *et al*., 2005a, b), it includes simple point-and-click access to a multitude of powerful new features. Among these are three new non-parametric recombination detection methods (3SEQ, VISRD and PHYLPRO; Boni *et al*., 2007; Lemey *et al*., 2009; Weiller, 1998), a parametric recombination rate estimation method (LDHAT; McVean *et al*., 2004), two new tree construction methods (Maximum likelihood with PHYML and Bayesian with MRBAYES; Guindon and Gascual, 2003; Ronquist and Huelsenbeck, 2003), two recombination hotspot-tests (Heath *et al*., 2006), a test of recombination induced protein mis-folding (Lefeuvre *et al*., 2007;

*To whom correspondence should be addressed.

Voigt *et al*., 2002), recombination-aware methods for reconstructing ancestral sequences (Arenas and Posada *et al*., 2010) and a range of matrix methods for visualizing overall patterns of recombination within datasets (Jakobsen and Easteal, 1996; Lefeuvre *et al*., 2009; McVean *et al*., 2004).

In addition to the new methods implemented in RPD3, another important improvement over RPD2 is the way in which RPD3 automatically scans alignments for recombination signals and then infers the minimum numbers of recombination events needed to account for these signals. RPD3 implements a range of heuristic recombinant sequence identification methods based on the PHYLPRO (Weiller, 1998), VISRD (Lemey *et al*., 2009) and subtree-prune and regraft methods (that identify recombinants sequences as those which 'jump' between the branches of phylogenetic trees constructed from different fragments of the same sequence alignment; Beiko and Hamilton, 2006; Heath *et al*., 2006). RDP3 also automatically checks detected recombination signals to determine whether they might not be better accounted for by sequence misalignment than recombination. Misalignments introduce homoplasy and are a common cause of false positive recombination signals. Misalignments are automatically detected in RPD3 by separately realigning recombinant sequences with each of their identified parents (RPD3 uses CLUSTALW to do this; Chenna *et al*., 2003) and comparing these pair-wise alignments to those of the corresponding sequence pairs in the full multiple sequence alignment. By more accurately identifying recombinant sequences and discounting recombination signals attributable to sequence misalignments, RPD3 significantly outperforms RDP2 for overall quantitative assessments of recombination patterns such as those carried out in the new breakpoint hot-spot and protein folding disruption tests.

In addition to streamlined tools for managing, testing and editing information on detected recombination events, RPD3 also provides a range of new tools for users to cross-check how accurately the program has identified (i) groups of recombinants supposedly sharing traces of the same recombination events; (ii) recombinant and parental sequences; and (iii) recombination breakpoint positions. These include heat-plots indicating how closely the recombination patterns in two recombinants resemble one another in relation to their supposed parental sequences, color coded phylogenetic trees for identifying recombinants and parental sequences and MAXCHI (Maynard Smith, 1992) and LARD (Holmes *et al*., 1999) breakpoint matrices for manually identifying breakpoint positions.

All of the automated recombination detection methods in RPD3 have been rigorously speed optimized and as a result the program is able to analyze datasets containing up to 40 million nt within 48 h on a standard 2 GHz processor with 2 GB of RAM. Such large datasets might, for example, consist of 20 full bacterial genome sequences, or 1000 full viral genome sequences. With default program settings datasets containing 100 10 kb long sequences can be analyzed within 10 min.

Funding: Wellcome Trust (to D.P.M.); Postdoctoral fellowship from the Fund for Scientific Research (FWO) Flanders (to Ph.L.); South African Centre of High Performance Computing bursary (to M.L.); European Research Council (ERC-2007-Stg 203161-PHYGENOM to D.P.); Spanish Ministry of Science and Education (BFU2009-08611 to D.P.); GIS CRVOI (grant NPRAO/AIRD/CRVOI/08/03 to Pi.L.); Wellcome Trust (grant number GR079127MA).

Conflict of Interest: none declared.

REFERENCES

Arenas,M. and Posada,D. (2010) The effect of recombination on the reconstruction of ancestral sequences. *Genetics*, **184**, 1133–1139.

Beiko,R.G. and Hamilton,N. (2006) Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.*, **6**, 15.

Boni,M.F. *et al*. (2007) An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, **176**, 1035–1047.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Gibbs,M.J. *et al*. (2000) Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, **16**, 573–582.

Guindon,S. and Gascuel,O. (2003) A. simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Heath,L. *et al*. (2006) Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J. Virol.*, **80**, 11827–11832.

Holmes,E.C. *et al*. (1999) Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.*, **16**, 405–409.

Jakobsen,I.B. and Easteal,S. (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.*, **12**, 291–295.

Kosakovsky Pond,S.L. *et al*. (2009) An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput. Biol.*, **5**, e1000581.

Lefeuvre,P. *et al*. (2007) Avoidance of protein fold disruption in natural virus recombinants. *PLoS Pathog.*, **3**, e181.

Lefeuvre,P. *et al*. (2009) Widely conserved recombination patterns among single-stranded DNA viruses. *J. Virol.*, **83**, 2697–2707.

Lemey,P. *et al*. (2009) Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. *BMC Bioinformatics*, **10**, 126.

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. *J. Virol.*, **73**, 152–160.

Martin,D.P. *et al*. (2005a) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.

Martin,D.P. *et al*. (2005b) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res. Hum. Retrovir.*, **21**, 98–102.

Maynard Smith,J. (1992) Analyzing the mosaic structure of genes. *J. Mol. Evol.*, **34**, 126–129.

McVean,G.A. *et al*. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Minin,V.N. *et al*. (2005) Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, **21**, 3034–3042.

Padidam,M. *et al*. (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology*, **265**, 218–225.

Posada,D. and Crandall,K.A. (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl Acad. Sci. USA*, **98**, 13757–13762.

Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

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.

Voigt,C.A. *et al*. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol.*, 9, 553–558.

Weiller,G.F. (1998) Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol. Biol. Evol.*, **15**, 326–335.