RDP3: a flexible and fast computer program for analyzing recombination

Darren P. Martin1,2,*, Philippe Lerney3, Martin Lott1,2,4, Vincent Moulton4, David Posada5 and Pierre Lefeuvre1,6

1Computational Biology Group, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, 2Centre for High Performance Computing, Rosebank, Cape Town, South Africa, 3Department of Microbiology and Immunology, Rega Institute, K.U. Leuven, Belgium, 4School of Computing Sciences, University of East Anglia, Norwich, NR4 7TT, UK, 5Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain and 6CIRAD, UMR 53 PVBM 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 (SEQ, VISRD, PHYLR0 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/rdp3.html

Contact: darrenmartin@uvigo.es

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, Marche, Chimaera, 3SEQ, Geneconv, Siscan, PhyloSr and Visrd; Boni et al., 2007; Gibbs et al., 2000; Lemey et al., 2009; Padidan et al., 1999; Posada and Crandall, 2001; Weiler, 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), DualBrothers (Minin et al., 2005), PhyML (Schultze 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 PhyloSr; Boni et al., 2007; Lemey et al., 2009; Weiler, 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), G_ru and Gcausal, 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;...
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; Lefèvre et al., 2009; McVean et al., 2004).

In addition to the new methods implemented in RDP3, another important improvement over RDP2 is the way in which RDP3 automatically scans alignments for recombination signals and then infers the minimum numbers of recombination events needed to account for these signals. RDP3 implements a range of heuristic recombinant sequence identification methods based on the HYPhLPo (Weiller, 1998), VISOR (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 RDP3 by separately realigning recombinant sequences with each of their identified parents (RDP3 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 breakpoints and discounting recombination signals attributable to sequence misalignments, RDP3 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, RDP3 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 RDP3 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 P.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 CRV0 (grant NPIRAO/AIRD/CRV0/08/03 to P.L.); Wellcome Trust (grant number GR079127MA).

Conflict of Interest: none declared.

REFERENCES


