biorxiv
doi: xxxxx
Advance Access Publication Date: xxxx
Applications Note

Phylogenetics

RCandy: an R package for visualising homologous recombinations in bacterial genomes

Chrispin Chaguza[1]*, Gerry Tonkin-Hill[1], Stephanie W. Lo[1], James Hadfield[2], Nicholas J. Croucher[3], Simon R. Harris[4]‡, and Stephen D. Bentley[1]‡

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

[2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, USA

[3]Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom

[4]Microbiotica Ltd, Biodata Innovation Centre, Wellcome Genome Campus, Hinxton, United Kingdom

‡Contributed equally

*To whom correspondence should be addressed.

Abstract

Motivation: Homologous recombination is an important evolutionary process in bacteria and other prokaryotes, which increases genomic sequence diversity and can facilitate adaptation. Several methods and tools have been developed to detect genomic regions recently affected by recombination. Exploration and visualisation of such recombination events can reveal valuable biological insights, but it remains challenging. Here, we present RCandy, a platform-independent R package for rapid, simple, and flexible visualisation of recombination events in bacterial genomes.

Availability and implementation: RCandy is an R package freely available for use under the MIT license. It is platform-independent and has been tested on Windows, Linux, and MacOSX. The source code comes together with a detailed vignette available on GitHub at https://github.com/ChrispinChaguza/RCandy

Contact: cc19@sanger.ac.uk

Supplementary information: Supplementary data are available online.

1 Introduction

Homologous recombination is a fundamental evolutionary process, which allows for the exchange of genetic information between similar DNA sequences. Homologous recombination in bacteria may facilitate adaptation (Chewapreecha *et al.*, 2014), repair (Mongold, 1992), or removal of deleterious elements (Croucher *et al.*, 2016). A single recombination event from a donor of a different strain or species will typically introduce a cluster of single nucleotide polymorphisms (SNP) within a short genomic region, distinguishing the event from point mutations, which typically only affect a single base (Croucher *et al.*, 2015). For example, in the *S. pneumoniae* PMEN1 lineage, the ratio of SNPs imported through recombination relative to those arising due to random mutational processes is ≈7, highlighting the impact of genetic exchanges on genomic diversity. Several computational tools have been developed to analyse recombinations within a single strain of a diverse population including Gubbins (Croucher *et al.*, 2015) and ClonalFrame (Didelot and Wilson, 2015), which detect and reconstruct recombinations onto a phylogeny, and BRATNextGen (Marttinen *et al.*, 2012) and FastGEAR (Mostowy *et al.*, 2017), which identify recombinations without phylogenetic reconstruction.

Although the specific approaches used to delineate the recombination blocks vary, all these tools generate output showing positions in each genome where recombination is likely to have occurred. This information is crucial to understanding the distribution of recombination across bacterial isolates, and identifying the genetic features affected by these exchanges. Visual exploration of recombinant genomic regions may reveal genomic regions and genes containing either a high or a low number of unique recombination events which are known as recombination “hotspots” and “coldspots,” respectively. For example, in *S. pneumoniae*, recombination hotspots are associated with genes encoding the choline-binding protein A (*cbpA* or *pspA*), which is highly immunogenic, and appears to be under selective pressure to evade recognition by the host immune system.

Visualising recombinations alongside other metadata such as isolation source and country of origin can also aid in the interpretation of a lineage's evolutionary and epidemiological history.

Although some web-based tools exist for interactively visualising recombination events, such as Phandango (Hadfield *et al.*, 2018), additional standalone tools with extended functionality are required to facilitate the analysis of recombination data and to create publication-quality figure. Here, we describe RCandy, an R package for visualisation of recombination events in bacterial genomes. Although we have developed RCandy to primarily display the genomic location of putative recombination events identified by inferred by Gubbins and BRATNextGen, output from other tools can be reformatted by the user for visualisation using this package.

2 Implementation and usage scenarios

Four main input data files are required to use RCandy, namely: a Newick-formatted phylogenetic tree file or "phylo" object (Paradis *et al.*, 2004); a tab-delimited metadata file or loaded as a "data.frame" object for each isolate in the phylogeny; a file containing the genomic location of recombination events generated by Gubbins or BRATNextGen in General Feature Format (GFF) or loaded as a "data.frame" object; and a GFF gene annotation file for the reference genome used to create the alignments used as input for the recombination detection programs.

To illustrate usage scenarios and features of the RCandy package, we identified recombination events in 170 serotype 19A *Streptococcus pneumoniae* isolates sequenced through the Global Pneumococcal Sequencing (GPS) consortium (https://www.pneumogen.net/gps/) (Gladstone *et al.*, 2019). These isolates belonged to a sequence type (ST) 320 clone defined using the multilocus sequence typing (MLST) scheme for the pneumococcus (https://pubmlst.org/organisms/streptococcus-pneumoniae). Supplementary Fig. 1 shows a multi-panel diagram generated by RCandy. Panel A shows the phylogenetic tree of the isolates. Panels B and C display metadata associated with the isolates in the tree and a legend describing the metadata shown in panel B. Panel D depicts genomic positions containing the inferred putative recombination events in each isolate present in the phylogenetic tree. By default, the identified recombination events are coloured differently to distinguish the events detected in more than one genome (red) and singleton events found in only one isolate (blue). This colouring distinguishes between recombinations shared through common descent, and those occurring independently in parallel across multiple isolates. Panel E shows the location of the genes in the reference genome used for sequence read mapping to generate the input pseudo-whole-genome alignment for the recombination detection tools. The arrows depict genes in the forward and reverse DNA strands. Two additional panels, F and G, show the frequency of unique recombinations per genomic position and per isolate. These panels can reveal recombination "hotspots" or "coldspots", and highly recombinogenic isolates.

RCandy implements several options to allow flexible visualisation of the phylogenetic tree, isolate metadata, and recombination events. For example, a user can specify a subset of isolates in the phylogenetic tree to display the recombination events and metadata. This option provides a zoom functionality allowing the user to analyse the distribution of recombination events in specific isolates; for example, those belonging to a particular clade. Similarly, the user can zoom in on certain genomic regions by specifying the start and end coordinates. By turning on the annotation labels, the user may identify the specific genes located in the selected region (see examples in the vignette). These extra options to customize the visualisation include, but are not limited to: ladderizing and midpoint tree rooting; colouring the phylogenetic tips based on isolate metadata; colouring the internal nodes of the phylogenetic tree based on the isolate metadata using the "ace" ancestral reconstruction function implemented in "ape" package (Paradis *et al.*, 2004), and hiding specific panels in Supplementary Fig. 1. Importantly, RCandy generates high-resolution figures, which can be saved to different file formats such as vector formats, including the portable document format (PDF) and scalable vector graphics (SVG) format. These figures can be edited further by the user to enhance aesthetics or highlight specific information for use in scientific publications.

Conclusions

RCandy is a user-friendly and platform-independent R package for rapid, simple, and flexible visualisation of genomic regions containing putative recombination events in genomes of a clonal bacterial population.

Acknowledgements

The authors would like to thank the Global Pneumococcal Sequencing (GPS) for providing published genomes to illustrate the functionality of RCandy.

Funding

The authors acknowledge funding from the Joint Initiative for Antimicrobial Resistance (grant no. MR/R003076/1), the Bill and Melinda Gates Foundation (grant no. OPP1034556), and Wellcome Trust (2016-2021 core award grant no. 206194).

Conflict of Interest: none declared.

References

Chewapreecha,C. *et al.* (2014) Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.*, **46**, 305–309.

Croucher,N.J. *et al.* (2016) Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. *PLoS Biol.*, **14**, e1002394–e1002394.

Croucher,N.J. *et al.* (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.*

Didelot,X. and Wilson,D.J. (2015) ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Comput. Biol.*, **11**, e1004041.

Gladstone,R.A. *et al.* (2019) International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, **43**, 338–346.

Hadfield,J. *et al.* (2018) Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, **34**, 292–293.

Marttinen,P. *et al.* (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.*, **40**, e6–e6.

Mongold,J.A. (1992) DNA repair and the evolution of transformation in Haemophilus influenzae. *Genetics*, **132**, 893–898.

Mostowy,R. *et al.* (2017) Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol. Biol. Evol.*, **34**, 1167–1182.

Paradis,E. *et al.* (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*.