BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 18 2010, pages 2334–2335
doi:10.1093/bioinformatics/btq413

Genome analysis

Advance Access publication July 11, 2010

genoPlotR: comparative gene and genome visualization in R

Lionel Guy*, Jens Roat Kultima and Siv G. E. Andersson

Department of Molecular Evolution, Uppsala University, Norbyvägen 18C, 75236 Uppsala, Sweden

Associate Editor: John Quackenbush

ABSTRACT

Summary: The amount of gene and genome data obtained by next-generation sequencing technologies generates a need for comparative visualization tools. Complementing existing software for comparison and exploration of genomics data, genoPlotR automatically creates publication-grade linear maps of gene and genomes, in a highly automatic, flexible and reproducible way.

Availability: genoPlotR is a platform-independent R package, available with full source code under a GPL2 license at R-Forge: http://genoplotr.r-forge.r-project.org/

Contact: lionel.guy@ebc.uu.se

Received on May 10, 2010; revised on July 2, 2010; accepted on July 7, 2010

1 INTRODUCTION

Comparison of genes and genomes has increasingly been used to infer patterns and processes in evolution and to relate phenotypic differences to genomic changes. With the recent advent of modern high-throughput sequencing technologies, the need for methods and visualization tools in comparative genomics has vastly increased. As of 2010, more than 1000 bacterial and archaeal genomes are available in public databases, making the number of possible comparisons almost infinite. Several programs such as the Artemis Comparison Tool (ACT; Carver *et al.*, 2005), UCSC Genome Browser (Rhead *et al.*, 2010), Mauve (Darling *et al.*, 2004), M-GCAT (Treangen *et al.*, 2006) and Murasaki (Sakakibara *et al.*, 2007) have extensive visualization functions. However, these tools lack the ability (i) to directly add annotations to the publication quality graphics they produce or (ii) to fully automate the production of comparative figures. GenomeGraphs (Durinck *et al.*, 2009) is a R package, which allows the visualization of one genomic region with related datasets such as microarray data. It can display several annotations for the same region, but it cannot show several regions in a single plot.

The R package genoPlotR is an attempt to fill in those gaps, by providing a flexible, automatable tool. It allows the user to graphically represent the comparison between several segments or subsegments of genomes in a linear fashion. It reads data stored in commonly used formats (EMBL, Genbank, BLAST and Mauve outputs) or in user-created tabular files and allows comparisons of one or several subsegments of a genome. A tree can be added to show the phylogenetic relationships between the segments, as can also scales and annotations to each subsegment. The use of R (R Development Core Team, 2009) and its grid package enables the use of its graphical power and flexibility to manipulate data and to integrate gene and genome maps into more complex graphics. The results can be saved either in high-quality raster or vector formats for further editing.

2 INPUT DATA

genoPlotR uses two main objects: `dna_seg`, which represent segments of DNA containing genes or other features, and `comparison`, which represent the relationships between two `dna_seg`.

genoPlotR reads `dna_seg` objects from the widely used Genbank and EMBL formats (described here: http://www.ncbi.nlm.nih.gov/collab/FT/; Fig. 1A), or from tabular formats, either as protein table files (NCBI) or user-generated. Genbank or EMBL files can be generated by the user or downloaded from nucleotide databases, e.g. from NCBI's Nucleotide Entrez (http://www.ncbi.nlm.nih.gov/sites/entrez). By default, only CDS features are read, but any set of features can be specified for reading.

genoPlotR reads `comparison` objects using the tabular output of, e.g. BLAST (Altschul *et al.*, 1990) or from user-generated tabular files (Fig. 1A). Hit tables in text format produced by stand-alone or online BLAST programs are suitable for input in genoPlotR. For example, from the NCBI BLAST web page, sequence alignments of genes and genomes can be produced with the option 'align two or more sequences'. The hit table can then be downloaded. The backbone file produced by Mauve, a multiple-genome alignment tool (Darling *et al.*, 2004), can be transformed into both `dna_seg` and `comparison` objects (Fig. 1B).

Both `dna_seg` and `comparison` objects can be filtered either by using arguments to the reading functions, or by using R functions, for example, to remove short genes or low-significance comparisons. Objects can be modified to specify color, size or appearance (arrows, blocks, lines, etc.) for each element of the DNA segments and of the comparisons. Intron-containing genes in EMBL and Genbank files are automatically recognized (Fig. 1D).

In addition, a phylogenetic tree in Newick format can be parsed as a `phylog` object, using the package ade4 (Dray *et al.*, 2007; Fig. 1A). Finally, `annotation` objects can be designed to add a legend to the DNA segments.

3 VISUALIZATION

After being read and modified, objects can be passed to the main graphical function `plot_gene_map`.

The user can define, for each segment, several subsegments that will be represented in the plot. These subsegments can be represented in reverse orientation (e.g. see Fig. 1B, first subsegment on the top DNA segment). The placement of each subsegment on the plot is either automatically determined by minimizing the area of the

*To whom correspondence should be addressed.

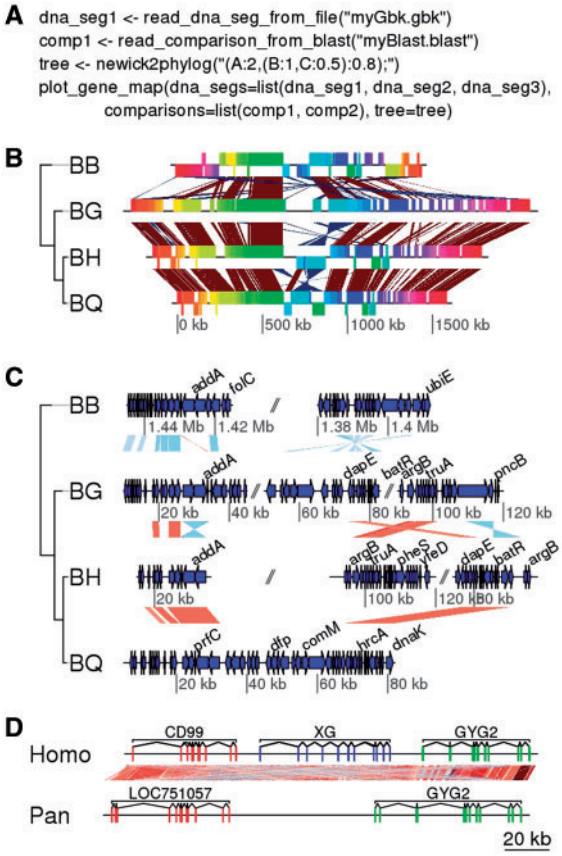

Fig. 1. Examples of gene and genome comparisons with genoPlotR. The code used to generate this figure is available at http://genoplotr.r-forge.r-project.org/screenshots.php. (**A**) Example of minimal code to represent a three-way comparison in genoPlotR. (**B**) Comparison of four *Bartonella* genomes using the backbone output of Mauve. The features are colored according to their order along the second segment (BG), following a rainbow palette. (**C**) BLAST comparison of genes in several subsegments of the same four *Bartonella* genomes. Note the scale of the first subsegment on the top row, which is in reverse orientation. (**D**) Comparison of a 220 kb segment of the Y chromosome in *Homo sapiens* and *Pan troglodytes*. In (C and D), the *E*-value of the BLAST alignment is represented with shades of blue and red. All datasets used here are present in the package.

comparisons, or fixed by the user. `annotation` and `phylog` objects are also passed to the main function `plot_gene_map`, and a scale can be added to any of the DNA segments, or placed at the bottom right of the plot.

Colors of the comparisons can be set to gray scales or to shades of blue and red, depending on the *e*-value or the bit score of the hit, as in ACT.

In addition to DNA segments and comparisons, other objects can be added to the plot, such as a phylogenetic tree—parsed into R using the package ade4 (Dray *et al.*, 2007)—or annotations for each DNA segment.

Using R embedded graphic generation functions, the figure can then be displayed on the screen or saved to various graphical formats, including both raster (functions `png, jpeg, tiff`, etc.) and vector formats (functions `postscript, svg, pdf`, etc.). The output of the graphical function can also be embedded in larger plots, allowing the user to create multi-panel figures.

Numerous examples and datasets, including the ones in Figure 1, are included in the package, to guide the users' first steps.

4 CONCLUSIONS

By using the graphical power and flexibility of R, the package genoPlotR generates reproducible maps of genes and genomes that can be used to generate publication-ready figures, starting from a wide range of formats. Since all the instructions for drawing the figures are contained in R code, it is highly flexible, and thus it is straightforward to automate the process of drawing very similar figures for different datasets. The use of a scripting language makes it particularly suitable for integration into annotation and comparative genomics pipelines.

ACKNOWLEDGEMENTS

The authors wish to thank the users that contributed to test the package and suggested improvements and new features.

Funding: the Swedish Research Council (623-2009-743 to L.G.); the Swiss National Science Foundation (PBLA33-119626 to L.G.); the Swedish Research Council, the Göran Gustafsson Foundation and the European Union (to S.G.E.A.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Carver,T.J. *et al.* (2005) ACT: the Artemis comparison tool. *Bioinformatics*, **21**, 3422–3423.

Darling,A.C.E. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.

Dray,S. *et al.* (2007) The ade4 Package: implementing the duality diagram for ecologists. *J. Stat. Softw.*, **22**, 1–20.

Durinck,S. *et al.* (2009) GenomeGraphs: integrated genomic data visualization with R. *BMC Bioinformatics*, **10**, 2.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rhead,B. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.

Sakakibara,Y. *et al.* (2007) [Development of a large-scale comparative genome system and its application to the analysis of mycobacteria genomes]. *Nihon Hansenbyo Gakkai Zasshi*, **76**, 251–256.

Treangen,T. *et al.* (2006) M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics*, **7**, 433.