Bioinformatics, 33(19), 2017, 3088–3090
doi: 10.1093/bioinformatics/btx346
Advance Access Publication Date: 29 May 2017
Applications Note

OXFORD

Genome analysis

karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data

Bernat Gel[1,*] and Eduard Serra[1,2]

[1]Hereditary Cancer Group, Program for Predictive and Personalized Medicine of Cancer – Germans Trias i Pujol Research Institute (PMPPC-IGTP), Campus Can Ruti, Badalona, Spain and [2]CIBERONC

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 3, 2017; revised on May 24, 2017; editorial decision on May 25, 2017; accepted on May 26, 2017

Abstract

Motivation: Data visualization is a crucial tool for data exploration, analysis and interpretation. For the visualization of genomic data there lacks a tool to create customizable non-circular plots of whole genomes from any species.

Results: We have developed karyoploteR, an R/Bioconductor package to create linear chromosomal representations of any genome with genomic annotations and experimental data plotted along them. Plot creation process is inspired in R base graphics, with a main function creating karyoplots with no data and multiple additional functions, including custom functions written by the end-user, adding data and other graphical elements. This approach allows the creation of highly customizable plots from arbitrary data with complete freedom on data positioning and representation.

Availability and implementation: karyoploteR is released under Artistic-2.0 License. Source code and documentation are freely available through Bioconductor (http://www.bioconductor.org/packages/karyoploteR) and at the examples and tutorial page at https://bernatgel.github.io/karyoploter_tutorial.

Contact: bgel@igtp.cat

1 Introduction

Data visualization is an important part of data analysis. It efficiently summarizes complex data, facilitates exploration and can reveal non-obvious patterns in the data. A natural representation for genomic data is positioned along the genome next to the ideograms of the different chromosomes. This type of representation is specially useful to identify the relation between different types of experimental data and genomic annotations. Various genomic visualization tools are available. Circos (Krzywinski *et al.*, 2009) produces highly customizable high quality circular plots, as does ites R counterpart RCircos (Zhang *et al.*, 2013). There are other R packages capable of plotting whole genome diagrams such as: ggbio (Yin *et al.*, 2012), based on the grammar of graphics that can produce different plot types including ideogram and karyogram plots; IdeoViz (Pai and Ren, 2014), to plot binned data along the genome either as lines or bars; or chromPlot (Oróstica and Verdugo, 2016), to plot up to four datasets given in a predefined format. These packages are either limited in the amount or type of data they can plot (IdeoViz and chromPlot) or have limited customization options (ggbio). In addition, the Bioconductor package Gviz (Hahne and Ivanek, 2016) is a powerful tool to create track-based plots of diverse biological data but it does not produce plots of the whole genome. There is a lack of a tool in R to create non-circular whole genome plots, able to plot arbitrary data in any organism and with ample customization capabilities.

Here we present karyoploteR, an extendable and customizable R/Bioconductor package to plot genome ideograms and genomic data positioned along them. It's inspired on the R base graphics, building plots with multiple successive calls to simple plotting functions.

2 Features

The interface of karyoploteR and the process to create a complete plot is very similar to that of base R graphics. We first create a simple or even empty plot with an initializing function and then add

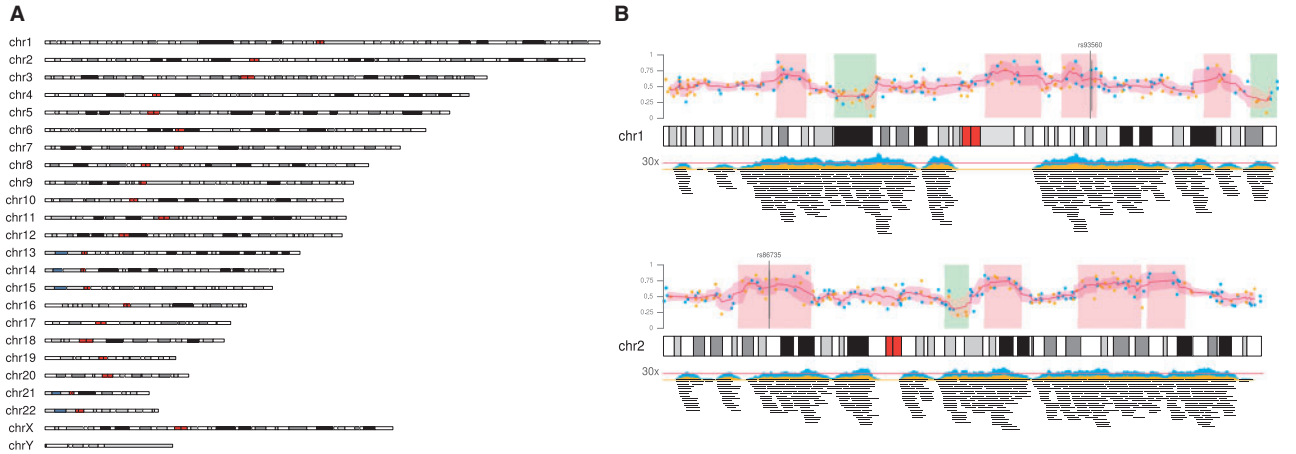

Fig. 1. (A) The complete human GRCh38 genome. This plot is created with the single command 'plotKaryotype(genome="hg38")'. **(B)** An example of a figure generated by karyoploteR representing different data types plotted in human chromosomes 1 and 2

additional graphic elements with successive calls to other plotting functions. The first call creates and initializes the graphical device and returns a *karyoplot* object with all the information needed to add data to it. The *karyoplot* object contains a coordinate change function mapping genomic coordinates into plotting coordinates, which is used by all plotting functions. Plotting functions are classified into three groups: the ones adding non-data elements to the plot and two data plotting groups, low-level functions and high-level functions. karyoploteR also takes some ideas from Circos, such as not defining fixed tracks but leaving complete freedom to the user with respect to data positioning using the *r0* and *r1* parameters. All non-data elements in the karyoplot (main title, chromosome names, ...) are drawn by specific functions. These functions accept standard graphical parameters but it's also possible to swap them for custom functions if a higher level of customization is needed.

2.1 Ideogram plotting

Ideogram plotting is the basic functionality of karyoploteR. Default ideograms can be plotted with a single function call (Fig. 1A). However, itwe possible to customize them: positioning the chromosomes in different arrangements, representing just a subset of chromosomes or change whether the cytobands are included and how they are represented. It is also possible to create different data plotting regions either above or below the ideograms as well as customizing all sizings and margins by changing the values stored in *plot.params*.

2.2 Not only human

karyoploteR is not restricted to human data in any way. It is possible to specify other organisms when creating a karyoplot. Genome data for a small set of organisms is included with the package and it will use functionality from regioneR (Gel *et al.*, 2016) to get it from UCSC or Bioconductor for other genomes. If an organism is not available anywhere, it is possible to plot it providing its genome information. Therefore, if required, it's possible to create custom genomes for specific purposes.

2.3 Data plotting

Data plotting functions are divided in two groups: low-level and high-level. Low-level data plotting functions plot graphical primitives such as points, lines and polygons. Except for the additional *chr* parameter, they mimic the behaviour of their base graphics counterparts including the usage of most of the standard graphical parameters. These plotting functions offer a flexible signature and are completely data agnostic: they know nothing about biological concepts, giving the user total freedom on how to use them. High-level functions, in contrast, are used to create more complex data representations. They understand some basic concepts such as 'genomic region' and they usually perform some kind of computation prior to data plotting (Fig. 1B).

2.4 Customization and extensibility

In addition to customizing sizings and margins and the using custom genomes, karyoploteR can be extended with custom plotting functions. All internal functions, including the main coordinate change function, are exported and documented in the package vignette. With this it is possible to create custom plotting functions adapted to specific data types and formats.

3 Conclusion

We have developed an R/Bioconductor package, karyoploteR, to plot arbitrary genomes with data positioned on them. It offers a flexible API inspired in R base graphics, with low-level functions to plot graphical primitives and high-level functions to plot complex data. The plots are highly customizable in data positioning and appearance and it is possible to extend the package functionality with custom plotting functions. karyoploteR requires R $\geq$ 3.4 and Bioconductor $\geq$ 3.5. More information and examples can be found at the package Bioconductor page and at https://bernatgel.github.io/karyoploter_tutorial

Acknowledgements

We thank Roberto Malinverni for his insightful comments and the IGTP HPC Core Facility and Iñaki Martínez de Ilarduya for his help.

Funding

This work has been supported by: the Spanish Ministry of Science and Innovation, Carlos III Health Institute (ISCIII) (PI11/1609; PI14/00577)(RTICC RD12/0036/008) Plan Estatal de I + D + I 2013–16, and co-financed by the FEDER program; the Government of Catalonia (2014 SGR 338); and the Spanish Association Against Cancer (AECC).

Conflict of Interest: none declared.

References

Gel,B. *et al*. (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289–291.

Hahne,F. and Ivanek,R. (2016) Visualizing genomic data using Gviz and bioconductor. *Methods Mol. Biol*., **1418**, 335–351.

Krzywinski,M. *et al*. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*., **19**, 1639–1645.

Oróstica,K.Y. and Verdugo,R.A. (2016) ChromPlot: Visualization of genomic data in chromosomal context. *Bioinformatics*, **32**, 2366–2368.

Pai,S. and Ren,J. (2014) IdeoViz: Plots data (continuous/discrete) along chromosomal ideogram. *R package version 1.8.0*.

Yin,T. *et al*. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol*., **13**, R77.

Zhang,H. *et al*. (2013) RCircos: an R package for Circos 2D track plots. *BMC Bioinformatics*, **14**, 244.