BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 22 2010, pages 2902–2903
doi:10.1093/bioinformatics/btq531

Sequence analysis
Advance Access publication September 21, 2010

girafe – an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads

Joern Toedling[1,2,3,4,5,*], Constance Ciaudo[1,4,5,6], Olivier Voinnet[6], Edith Heard[1,4,5] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, 26 rue d'Ulm, F-75248 Paris, [2]INSERM U900, F-75248 Paris, [3]Mines ParisTech, F-77300 Fontainebleau, [4]CNRS UMR3215, F-75248 Paris, [5]INSERM U934, F-75248 Paris and [6]Institut de Biologie Moléculaire des Plantes, CNRS UPR2357, Université Louis Pasteur, Strasbourg, France

Associate Editor: Alex Bateman

ABSTRACT

Summary: The R/Bioconductor package *girafe* facilitates the functional exploration of alignments of sequence reads from next-generation sequencing data to a genome. It allows users to investigate the genomic intervals together with the aligned reads and to work with, visualise and export these intervals. Moreover, the package operates within and extends the ever-growing Bioconductor framework and thus enables users to leverage a multitude of methods for their data in order to answer specific research questions.

Availability and Implementation: The R package *girafe* is available from the Bioconductor web site: http://www.bioconductor.org/packages/release/bioc/html/girafe.html

An extensive vignette and the Bioconductor mailing lists provide additional documentation and help for using the package.

Contact: joern.toedling@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 1, 2010; revised on August 20, 2010; accepted on September 13, 2010

1 INTRODUCTION

Next-generation sequencing (NGS) technologies provide users with millions of comparatively short RNA or DNA *reads* from biological samples of interest. The first step in the analysis of these data usually is to align the reads to the chosen reference genome, using powerful, specific alignment tools. For the secondary analysis that follows, there is a need for an integrated work environment which allows users to explore and annotate the genome intervals with the aligned reads. Besides providing functionality for data exploration, this environment must also provide interfaces to other available tools, especially as the field of NGS analysis software is comparatively new and rapidly evolving. Previously described, excellent tools that provide frameworks for working with aligned reads include *Galaxy* (Giardine *et al.*, 2005) and *SAMtools* (Li *et al.*, 2009).

Here, we describe the R/Bioconductor package *girafe* that enables users to investigate the genome intervals with the aligned reads, henceforth referred to as *aligned intervals*, and to work with, visualise and export these aligned intervals. One advantage of *girafe* over other tools for working with aligned reads is that the package operates within the open source, open development and constantly growing Bioconductor framework (Gentleman *et al.*, 2004). Thus, this package enables users to leverage a multitude of methods in the analysis of their data in order to answer specific research questions.

2 AVAILABLE FUNCTIONALITY

In the following, we present some functionalities of *girafe*. The package is built on, and greatly enhances, the functionalities of the Bioconductor packages *genomeIntervals* and *ShortRead* (Morgan *et al.*, 2009).

For this demonstration, we use example data downloaded from the Gene Expression Omnibus database (Edgar *et al.*, 2002, GSE10364). The data are Solexa reads obtained from small RNA profiling of mouse oocytes (Tam *et al.*, 2008).

Importing aligned reads: The reads were mapped to the mouse genome (assembly *mm9*) using the *Bowtie* aligner (Langmead *et al.*, 2009). The resulting file can be read into R, using the *ShortRead* package, and converted into an object of class *AlignedGenomeIntervals*, the core class of package *girafe*.

Exploring aligned intervals: Objects of this class can easily be explored using standard R functions to obtain summary statistics answering questions, such as (i) how long are the reads aligned to specific intervals? or (ii) how many intervals are located on each chromosome?

Processing the aligned intervals: Basic interval operations, such as sorting, shifting and determining intersections and unions of interval sets, are readily supported. Moreover, the function `reduce` provides a flexible way to combine, or merge, aligned intervals. One intention could be to combine aligned reads at exactly the same position, which only differ in their sequence due to sequencing errors. Another example objective could be to combine overlapping short reads that may be (degradation) products of the same primary transcript.

Visualisation: The package *girafe* contains functions for visualising aligned intervals with the powerful plotting facilities of R. These visualisation functions are a flexible alternative to those provided by genome browsers and may be especially relevant for sequencing data from organisms which are not well represented in genome browsers. Figure 1 shows aligned intervals from the example data in a 500 bp region on the X chromosome. The reads aligned in this region correspond to two miRNAs reported to be highly expressed in these data (Tam *et al.*, 2008).

[*]To whom correspondence should be addressed.

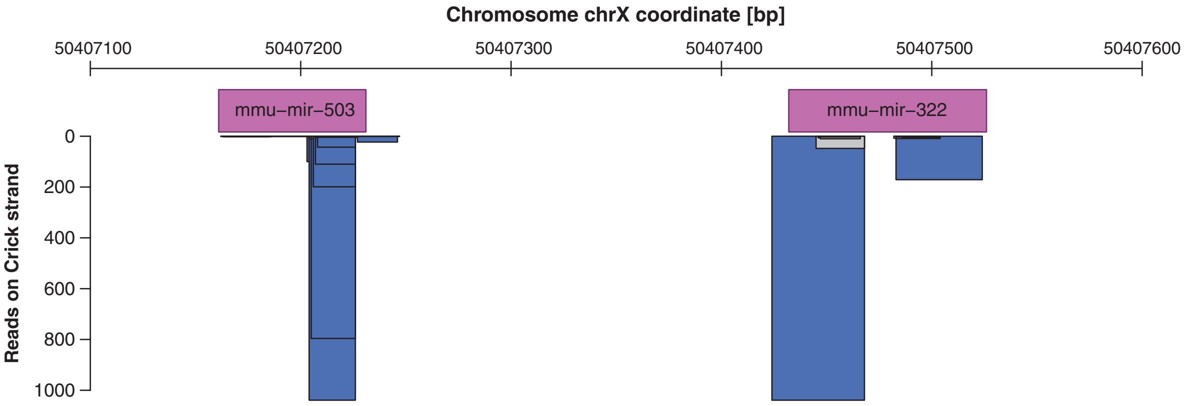

Fig. 1. Visualisation of genome intervals with aligned reads. Two miRNAs (shown in pink) have been annotated on the Crick strand within this region on the X chromosome. Reads were exclusively aligned to the Crick strand in this region and mostly correspond to the mature and star sequences of the miRNAs. Intervals with uniquely aligned reads are shown in blue while grey marks non-unique reads. For *mmu-mir-503* (left), the original data are shown while for *mmu-mir-322* (right) the overlapping *AlignedGenomeIntervals* have been merged using the function `reduce`.

Summarising the data using sliding windows: The data can be searched for genome regions of defined interest using a sliding-window approach. For each window, the number of intervals with aligned reads, the total number of reads aligned, the number of unique reads aligned, the fraction of intervals on the Watson strand, and the higher number of aligned reads at a single interval within the window are reported.

Overlap with annotated genome features: A frequent task is to determine the overlap of the aligned intervals with genome elements that are described in databases, in order to annotate the aligned reads. *girafe* includes functions for efficiently determining these overlaps and allows the user to specify custom requirements, such as a minimum fraction of the total interval length, for considering intervals and features to be truly overlapping.

Exporting the data: The *girafe* package contains methods for exporting the data into tab-delimited text files, which can be uploaded to genome browsers for further visualisation and exploration. Currently supported formats include ‘bed’, ‘bedGraph’ and ‘wiggle’.

Vignette: The package vignette provides more detailed examples together with the corresponding R source code and discusses memory usage and interactions with other Bioconductor packages.

3 CONCLUSION

The R/Bioconductor package *girafe* provides users with a powerful, flexible and extensible framework to explore NGS data, following alignment of the reads to a reference genome.

The package interacts with other Bioconductor packages and allows export of the data in various formats for exploring them in other software, with the aim of not restricting the user to a limited set of analysis tools.

The field of NGS analysis software is growing rapidly. Thus, future developments of the package will include adding further methods for working with aligned genome intervals, reducing the memory footprint, and providing additional interfaces to other R/Bioconductor packages and other software.

ACKNOWLEDGEMENTS

Many thanks to Nicolas Servant, Valérie Cognat, Nicolas Delhomme and Patrick Aboyoun for suggestions and feedback on the package. Special thanks to Julien Gagneur and Richard Bourgon for writing *genomeIntervals*. The plotting functions in package *girafe* are based on the function `plotAlongChrom` from package *tilingArray*, written by Wolfgang Huber.

Funding: Institut Curie, INCa “GepiG”.

Conflict of Interest: none declared.

REFERENCES

Edgar,R. *et al.* (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Giardine,B. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Morgan,M. *et al.* (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607–2608.

Tam,O.H. *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.