Bioinformatics, 2019, 1–2
doi: 10.1093/bioinformatics/btz681
Advance Access Publication Date: 30 August 2019
Applications Note

Genome analysis

breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data

David Porubsky[1,2,*,†], Ashley D. Sanders[3,4,†], Aaron Taudt [1,5], Maria Colomé-Tatché[1,5], Peter M. Lansdorp[1,3,6,‡] and Victor Guryev[1,‡]

[1]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, [2]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA, [3]Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC, Canada, [4]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany, [5]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany and [6]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint last Authors.

Associate Editor: John Hancock

Received on May 8, 2019; revised on August 12, 2019; editorial decision on August 23, 2019; accepted on August 27, 2019

Abstract

Motivation: Strand-seq is a specialized single-cell DNA sequencing technique centered around the directionality of single-stranded DNA. Computational tools for Strand-seq analyses must capture the strand-specific information embedded in these data.

Results: Here we introduce breakpointR, an R/Bioconductor package specifically tailored to process and interpret single-cell strand-specific sequencing data obtained from Strand-seq. We developed breakpointR to detect local changes in strand directionality of aligned Strand-seq data, to enable fine-mapping of sister chromatid exchanges, germline inversion and to support global haplotype assembly. Given the broad spectrum of Strand-seq applications we expect breakpointR to be an important addition to currently available tools and extend the accessibility of this novel sequencing technique.

Availability and implementation: R/Bioconductor package https://bioconductor.org/packages/breakpointR.

Contact: porubsky@uw.edu

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

1 Introduction

Strand-seq is a single-cell DNA sequencing technology that sequences only the template strands contained in a cell after DNA replication (Falconer *et al.*, 2012). The key advantage of sequencing a single strand for each chromosome is that the identity and structure of each homolog can be determined based on the strand-specific directionality of aligned Strand-seq data (Sanders *et al.*, 2017). To date, Strand-seq has been applied to many biologically relevant questions, such as mapping sister chromatid exchange events (SCEs) (Claussin *et al.*, 2017; Falconer *et al.*, 2012; van Wietmarschen and Lansdorp, 2016), locating germline inversions (Chaisson *et al.*, 2019; Sanders *et al.*, 2016), assembling chromosome-length haplotypes (Porubský *et al.*, 2016, 2017) and guiding *de novo* reference assemblies (Ghareghani *et al.*, 2018; O'Neill *et al.*, 2017).

A critical step in exploiting Strand-seq data for biological discovery is locating the coordinates of template-strand-state changes in each individual cell. Because canonical bioinformatics pipelines are not currently suited to exploit the DNA directionality embedded in Strand-seq data, we have developed breakpointR. breakpointR is a user-friendly R/Bioconductor package designed to track changes in read directionality in single-cell Strand-seq data at high resolution. In locating template-strand-state changes for individual cells and providing user-friendly output files to directly visualize these data, breakpointR represents an accessible tool to navigate many biological features made available through the Strand-seq technology.

2 Materials and methods

An overview of the breakpoint detection algorithm is provided as part of Supplementary Figure S1. In Strand-seq, sequencing reads are distinguished based on the direction they map to the reference genome: reads mapped to the positive (plus) strand are labeled as 'Crick' (C; teal), and reads mapped to the negative (minus) strand are labeled as 'Watson' (W; orange) (Supplementary Fig. S1A, i). We distinguish three possible strand states in a diploid daughter cell:

two Watson templates (WW), two Crick templates (CC) and one Watson and one Crick template strand (WC) (Supplementary Material). These states are evidenced by the proportion of W and C reads mapping to the given chromosome for the given cell (Supplementary Fig. S1A, i). Localized changes in template strand directionality (where for instance the pattern changes from a WC state to CC; see Supplementary Fig. S1A, i) are frequently observed, which notably mark important biological features in the cell (as described previously), are evidenced by changes in the proportion of W and C reads along the chromosome. Thus, breakpointR locates template-strand-state changes by pinpointing transition points in the relative proportion of W and C reads along each chromosome of the cell.

To calculate proportions of W and C reads, the chromosome data are first binned. In breakpointR, we have implemented a bi-directional read-based binning approach. We have previously shown that template-strand-state changes can be determined by calculating the ratio of W and C reads within defined genomic regions (bins) using the open-source software BAIT (Hills *et al.*, 2013). However, BAIT fragments the genome into a uniform and fixed-sized bins and, therefore, breakpoint resolution is limited to the bin length. Our read-based binning strategy scales each bin dynamically to accommodate a user-defined number of reads and slides the window position one read at a time to preserve the genomic context of each individual sequenced fragment (Supplementary Fig. S1A, ii). This approach accounts for biases that are caused by genome mappability (Baslan *et al.*, 2012) and variable or sparse sequence coverage typical to single-cell data. Overall, we found breakpointR being able to detect ~10–30% more simulated template-strand-changes than BAIT (Supplementary Material).

Additional to a sensitive binning strategy, breakpointR implements a 'ΔW function' to calculate changes in the relative abundance of W and C reads. For a given bi-directional bin, a delta W (ΔW) value is calculated as the absolute difference in the total number of W reads found in the first half of the bin to the total number of W reads found in the second half of the bin (Supplementary Fig. S1A, ii). In simple terms, the ΔW represents the template-strand-state of the region; a consistent ('unchanged') template-strand-state will produce a low ΔW value, whereas a template-strand-state change will produce a high ΔW value. Thus, by calculating ΔW values for each sliding bin, template-strand-state changes, referred to as 'breakpoints', can be located as peaks in the ΔW values (Supplementary Fig. S1A, iii).

Described in detail in the Supplementary Material, breakpointR takes as input strand-specific sequencing reads aligned to the reference genome in BAM file format. By, implementing the read-based binning strategy to calculate a ΔW value for each window, a vector of ΔWs is produced for each chromosome. The algorithm interrogates the ΔWs to locate values significantly above a defined threshold. Each ΔW peak represents a putative breakpoint that marks a template-strand-state change for that chromosome. It assigns the breakpoint to the end position of the first read in the peak and the start position of the last read in the peak (Supplementary Fig. S1A, iv). To validate the breakpoint, breakpointR then compares strand-states of neighboring segments and the breakpoint is retained only if a bona fide strand-state change is observed. From this, a list of breakpoint coordinates for each chromosome is produced for each input BAM file.

An accompanying package vignette illustrates the basic and some advanced features of breakpointR. The outputs of a breakpointR analysis include a 'BreakPoint' class object that stores the raw directional reads, the vector of ΔW values, as well as all detected breakpoints for the input file. Additionally, breakpointR prepares bed-formatted files of these data, which enables the user to visualize their results directly in the UCSC Genome Browser. An example of these output data are shown in Supplementary Figure S1B. Last, breakpointR outputs genome-wide and chromosome-specific plots of all the template-strand-state changes located per input single cell, as well as a population-scale heatmap representing a summary of all template-strand-states detected in the input dataset. Accordingly, breakpointR provides the user with ample ways to interpret the Strand-seq data for biological discoveries.

3 Discussion

Here, we introduce an easy-to-use tool called breakpointR that directly detects template-strand-state changes in Strand-seq data. Locating template-strand-state changes is required for locating SCEs, mapping germline inversions and defining segments for haplotype phasing. Given the many biological applications already made possible by strand-specific sequencing, we expect a steady increase in the number of Strand-seq users in coming years. Therefore, the development of user-friendly computational tools tailored to the unique features of Strand-seq datasets is of paramount importance. With the end user in mind, we designed breakpointR to facilitate easy navigation and visualization of the results. We believe this makes the tool widely accessible, irrespective of the user's computational background or specific biological question. In this way, breakpointR represents an important tool that helps make Strand-seq more accessible to the single-cell genomics community.

Acknowledgements

We thank Ester Falconer, Mark Hills and Diana Spierings for helpful discussions and Tonia Brown for proofreading this manuscript.

Funding

This work was supported by a European Research Council Advanced grant to PML.

Conflict of Interest: none declared.

References

Baslan,T. *et al.* (2012) Genome-wide copy number analysis of single cells. *Nat. Protoc.*, **7**, 1024–1041.

Chaisson,M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.

Claussin,C. *et al.* (2017) Genome-wide mapping of sister chromatid exchange events in single yeast cells using Strand-seq. *Elife*, **6**, 1–17.

Falconer,E. *et al.* (2012) DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods*, **9**, 1107–1112.

Ghareghani,M. *et al.* (2018) Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics*, **34**, i115–i123.

Hills,M. *et al.* (2013) BAIT: organizing genomes and mapping rearrangements in single cells. *Genome Med.*, **5**, 82.

O'Neill,K. *et al.* (2017) Assembling draft genomes using contiBAIT. *Bioinformatics*, **33**, 2737–2739.

Porubsky,D. *et al.* (2017) Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.*, **8**, 1293.

Porubský,D. *et al.* (2016) Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res.*, **26**, 1565–1574.

Sanders,A.D. *et al.* (2016) Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.*, **26**, 1575–1587.

Sanders,A.D. *et al.* (2017) Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.*, **12**, 1151–1176.

van Wietmarschen,N. and Lansdorp,P.M. (2016) Bromodeoxyuridine does not contribute to sister chromatid exchange events in normal or Bloom syndrome cells. *Nucleic Acids Res.*, **44**, 6787–6793.