Sequence analysis

Basic4Cseq: an R/Bioconductor package for analyzing 4C-seq data

Carolin Walter1,*, Daniel Schuetzmann2, Frank Rosenbauer2 and Martin Dugas1

1Institute of Medical Informatics, University of Münster, 48149 Münster, Germany, and 2Institute of Molecular Tumorbiology, University of Münster, 48149 Münster, Germany

ABSTRACT

Summary: Basic4Cseq is an R/Bioconductor package for basic filtering, analysis and subsequent near-cis visualization of 4C-seq data. The package processes aligned 4C-seq raw data stored in binary alignment/map (BAM) format and maps the short reads to a corresponding virtual fragment library. Functions are included to create virtual fragment libraries providing chromosome position and further information on 4C-seq fragments (length and uniqueness of the fragment ends, and blindness of a fragment) for any BSgenome package. An optional filter is included for BAM files to remove invalid 4C-seq reads, and further filter functions are offered for 4C-seq fragments. Additionally, basic quality controls based on the read distribution are included. Fragment data in the vicinity of the experiment’s viewpoint are visualized as coverage plot based on a running median approach and a multi-scale contact profile. Wig files or csv files of the fragment data can be exported for further analyses and visualizations of interactions with other programs.

Availability and implementation: Basic4Cseq is implemented in R and available at http://www.bioconductor.org/. A vignette with detailed descriptions of the functions is included in the package.

Contact: Carolin.Walter@uni-muenster.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 24, 2014; revised on June 6, 2014; accepted on July 17, 2014

1 INTRODUCTION

Circular chromosome conformation capture combined with high-throughput sequencing (4C-seq) is a method that allows the identification of chromosomal interactions between potential interaction partners, called viewpoint, and virtually any other part of the genome (Gheldof et al., 2012). Two rounds of digestion and ligation with two distinct restriction enzymes are performed to create a 4C-seq library. Unlike regular next-generation sequencing data, valid 4C-seq reads can only originate from precisely defined points in the genome. Any 4C-seq read will map to the end of a so-called 4C-seq fragment, i.e. a genomic region flanked by two primary restriction sites. When a secondary restriction site is not present, a fragment is called blind; a fragment end is defined as the region between a primary restriction site, and the ends of each fragment are checked for the presence of a secondary restriction site, and the ends of each fragment are checked for uniqueness. Fragment data are stored as a virtual fragment library file, which can be applied for 4C-seq experiments with the same underlying genome, restriction enzyme combination, and read length.

2 AVAILABLE FUNCTIONALITY

2.1 Preprocessing and virtual fragment library creation

Basic4Cseq expects 4C-seq data stored in BAM format. SAMtools (Li et al., 2009) and BEDtools (Quinlan and Hall, 2010) can be used to remove reads on very short or blind fragments directly from the BAM file. Basic4Cseq uses BSgenome packages to split a given genome at primary restriction sites. The resulting fragments are scanned for the presence of a secondary restriction site, and the ends of each fragment are checked for uniqueness. Fragment data are stored as a virtual fragment library file, which can be applied for 4C-seq experiments with the same underlying genome, restriction enzyme combination, and read length.

2.2 Filtering of 4C-seq data

Owing to sequencing errors or SNPs, the alignment of raw 4C-seq data needs to allow a certain number of mismatches. However, mismatches in the first restriction enzyme sequence can lead to mapped reads outside the predefined 4C-seq fragment ends. Overlaps of these reads with valid fragment ends may

*To whom correspondence should be addressed.
filter options and alternative tools are added as supplement. Further peaks suggest high contact intensities between the under-

with the typical high coverage at the experiment’s viewpoint. 

ment approach for the analysis of the 4C-seq fragment data 

ment are normalized to [0, 1], and the resulting intensity values 

mean windows with varying window sizes. Read counts per frag-

ment data and running medians, plus quantile data in a 

scale contact intensity profile. Fragment end-based resolution increases 

sequences caused by self-ligation. Basic4Cseq visualizes both raw 

experiment are removed to prevent bias through overrepresented 

options include a running median and a running mean approach; 

reads per million (RPM)-normalized and smoothed to counter 

line from van de Werken 

vides a visualization of the viewpoint region similar to the pipe-

from non-blind, unique fragment ends to prevent bias. It pro-

virtual fragment library. Basic4Cseq analyzes per default data 

After preprocessing, the aligned short reads are mapped to the 

Basic4Cseq analyzes per default data 

from non-blind, unique fragment ends to prevent bias. It pro-

Basic4Cseq visualizes both raw fragment data and running medians, plus quantile data in a 

plot. Additionally, multi-scale contact profiles of the 

Basic4Cseq can optionally discard 

any read with mismatches in the restriction enzyme sequence, if 

this sequence is present. Additional filtering options are available 

for the virtual fragment library. Very short or long fragments as 

well as blind fragments can be discarded; all repetitive fragment 

ends are removed per default to prevent ambiguous data.

2.3 Visualization

After preprocessing, the aligned short reads are mapped to the 

virtual fragment library. Basic4Cseq analyzes per default data 

from non-blind, unique fragment ends to prevent bias. It pro-

vides a visualization of the viewpoint region similar to the pipe-

line from van de Werken et al. (2012a): Fragment-based data are 

reads per million (RPM)-normalized and smoothed to counter 

the effects of single, overrepresented fragment ends. Smoothing 

options include a running median and a running mean approach; 

quantiles are further smoothed and interpolated with R’s loess 

function. Fragments adjacent to the viewpoint of the 4C-seq 

experiment are removed to prevent bias through overrepresented 

sequences caused by self-ligation. Basic4Cseq visualizes both raw 

fragment data and running medians, plus quantile data in a 

single plot. Additionally, multi-scale contact profiles of the 

chosen fragment data are created for running median or running 

mean windows with varying window sizes. Read counts per frag-

ment are normalized to [0, 1], and the resulting intensity values 

are expressed on a log 2 scale to allow for better visibility of 

distant interactions. Basic4Cseq can annotate custom regions of interest for easier interpretation and comparison of experi-

ments. Import and near-cis visualization of fragment-based data (e.g. fourSig fragment data, which are easily convertible) 

is possible as well. Figure 1 shows a near-cis interaction profile 

with the typical high coverage at the experiment’s viewpoint. 

Further peaks suggest high contact intensities between the under-

lying genomic regions and the viewpoint. Comparisons between 

filter options and alternative tools are added as supplement. 

For regions located either more remote from the viewpoint or 

on other chromosomes, it is advisable to use a statistical enrich-

ment approach for the analysis of the 4C-seq fragment data

(Splinter et al., 2012). Basic4Cseq can export filtered 4C-seq data as csv or wig files for use with other programs, e.g. 

Splinter et al.’s domainogram and spider-plot functions 

(Splinter et al., 2012). Additionally, a visualization routine based on RCircos (Zhang et al., 2013) is included for imported 

trans interaction data intervals.

2.4 Read distribution and quality control

The distribution of reads throughout 4C-seq fragment ends can 

provide information on the quality of the experiment data (van 

de Werken et al., 2012b). Basic4Cseq provides the following 

quality control statistics: number of total reads, cis/overall 

ratio of reads and the percentage of covered fragment ends 

within a certain distance around the experiment’s viewpoint. 

Reference values for high-quality experiments are >1 million 

reads total, a cis/overall ratio of >40% and a large fraction of 

covered fragment ends in the viewpoint’s vicinity (van de Werken et al., 2012b).

3 CONCLUSION

Basic4Cseq enables users to analyze 4C-seq experimental data 

in the R/Bioconductor environment. Quality control statistics 

can be calculated, and virtual fragment libraries with relevant fragment data generated. Functions for near-cis visualizations 

are included; both coverage profiles and heatmap-like multi-

scale contact intensity visualizations allow the exploration of 

contact profiles around the viewpoint. The package allows 

users to import fragment data for visualization, and to export 

filtered 4C-seq reads as csv or wig files for visualization or fur-

ther analysis of significant interactions.

Conflicts of Interest: none declared.

REFERENCES

Ghedin,N. et al. (2012) Detecting long-range chromatin interactions using the 

chromosome conformation capture sequencing (4c-seq) method. Methods 


Li H. et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. 

Bioinformatics, 25, 2078–2079.


genomic features. Bioinformatics, 26, 841–842.

Splinter,E. et al. (2012) Determining long-range chromatin interactions for selected 

genomic sites using 4c-seq technology: from fixation to computation. Methods, 

58, 221–230.

Stadhouders,R. et al. (2012) Dynamic long-range chromatin interactions control 

Myb proto-oncogene transcription during erythroid development. EMBO J., 

31, 969–999.

Thongjuea,S. et al. (2013) r3Cseq: an R/Bioconductor package for the discovery of 

long-range genomic interactions from chromosome conformation capture and 


van de Werken,H. et al. (2012a) Robust 4C-seq data analysis to screen for regula-

tory DNA interactions. Nat. Methods, 9, 969–971.

van de Werken,H. et al. (2012b) 4C technology: protocols and data analysis. 


Williams,R. et al. (2014) fourSig: a method for determining chromosomal inter-


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

Bioinformatics, 14, 244.