Bioinformatics, 31(10), 2015, 1663–1664
doi: 10.1093/bioinformatics/btu846
Advance Access Publication Date: 5 January 2015
Applications Note

OXFORD

Genome analysis

rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples

Wolfgang Kaisers[1,*,†], **Heiner Schaal**[1,2,†] **and Holger Schwender**[1,3]

[1]Center for Bioinformatics and Biostatistics, BMFZ, Heinrich Heine University Düsseldorf, [2]Institut für Virologie, and [3]Mathematical Institute, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

*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.

Associate Editor: John Hancock

Received on July 31, 2014; revised on December 18, 2014; accepted on December 19, 2014

Abstract

Summary: The open source environment R isf the most widely used software to statistically explore biological data sets including sequence alignments. BAM is the de facto standard file format for sequence alignment. With rbamtools, we provide now a full spectrum of accessibility to BAM for R users such as reading, writing, extraction of subsets and plotting of alignment depth where the script syntax closely follows the SAM/BAM format. Additionally, rbamtools enables fast accumulative tabulation of splicing events over multiple BAM files.

Availability and implementation: rbamtools is available on CRAN and on R-Forge.

Contact: kaisers@med.uni-duesseldorf.de

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

1 Introduction

The samtools format specifies various data slots for sequence alignments many of which are difficult to understand when sequencing experiments are to be analyzed. For analysis of sequencing data, detailed access to contents of BAM files is needed, especially when technical problems arise. *rbamtools* allows R users to investigate alignment results by reading the header section or retrieve and view alignments from regions of interest using basic R structures.

rbamtools provides functions for creation and modification of BAM file header or alignment section contents. *rbamtools* also facilitates writing of BAM files which is not possible in Bioconductor (Gentleman *et al.*, 2004; Morgan *et al.*, 2010).

Additionally *rbamtools* contains a framework for sequential and fast extraction of alignment gap positions (see Table 1) on RNA-seq data which are candidate sites for true splicing events. *rbamtools* is part of an analysis pipeline for analysis of splicing events in RNA-seq data which consists of three R packages: *rbamtools* and refGenome (Kaisers, 2013a) and spliceSites (Kaisers, 2013b).

The identification of splicing inaccuracies is a non trivial task on BAM files, since the positions of alignment gaps must be accounted on billions of reads. With *rbamtools*, processing data from, e.g. 60 RNA-seq samples (containing 8.37×10^9 alignments) can be done in 1.75 h on a standard workstation with minimal working memory demand.

Current versions of the samtools C library contain misalignment (bus) errors (http://en.wikipedia.org/wiki/Bus_error), which may cause program crashes on some architectures (e.g. SPARC). In *rbamtools*, these misalignment errors are corrected (see Supplementary Material).

2 Approach

2.1 Implementation

The package consists of three layers: the samtools C library, C based containers for alignments and alignment gaps as well as an S4 class library in R providing the user interface.

The samtools C library is a static copy of samtools (v1.4-r985). In order to meet CRAN policies, numerous changes had to be introduced into the source code (B.Ripley and K.Hornik, personal communication).

Table 1. Example for a gap site

Exon	Intron	Exon	Position	CIGAR
AG		CCTTGATG	3	2M6N8M
CAG		CCTTGAT	2	3M6N7M
CCAG		CCT	1	4M6N3M
CCCAG	GTCCAG	CCTTGATGTCC	(reference)	

A gap site defined by three alignments which share the same alignment gap site. The position values are 0-based (as described in the SAM file format[a]). The last row represents the (chromosomal) reference sequence.

[a] http://samtools.github.io/hts-specs/SAMv1.pdf

2.2 User interface

The S4 class library closely reflects the internal structures of BAM files. In order to provide detailed access to BAM file content, the API provides 14 classes and numerous functions.

Basic accessors

Basic accessors provide access to all parts of raw file content, header section and alignments for reading and writing. The following example opens the BAM file bam and copies alignments on chromosome 1 into a second BAM file.

```
rd <- bamReader(bam, idx = TRUE)
rg <- bamRange(rd, getRefCoords(rd, "chr1"))
wr <- bamWriter(getHeader(rd, "chr1.bam")
bamSave(wr, rg, refid = 0)
```

Inspecting ranges can be useful when downstream analysis indicates regions without any alignments or other technical flaws.

Specialized analysis routines

Specialized analysis routines for visualization of phred score distribution as well as for tabulation of nucleotide content and calculation of GC content and AT/GC ratio are provided.

Alignment depth

This information can be retrieved from a *bamRange* object. Figure 1 shows an example where alignment depth is plotted for Gene CHMP2A (ENSG00000130724).

```
ad <- alignDepth(range)
plotAlignDepth(ad)
```

Gap sites

Gap sites are kept in containers of class `bamGapSite`. The data is gathered from BAM files using the `bamGapList` function which directly operates on `bamReader` objects as shown below.

```
bgl1 <- bamGapList(bamReader(bam1, idx = TRUE)
bgl2 <- bamGapList(bamReader(bam2, idx = TRUE)
bgl  <- merge(bgl1, bgl2)
```

Gap site positions and numbers of crossing read alignments can be obtained from multiple BAM files as data.frame by executing:

```
gap <- readPooledBamGapDf(fileNames)
```

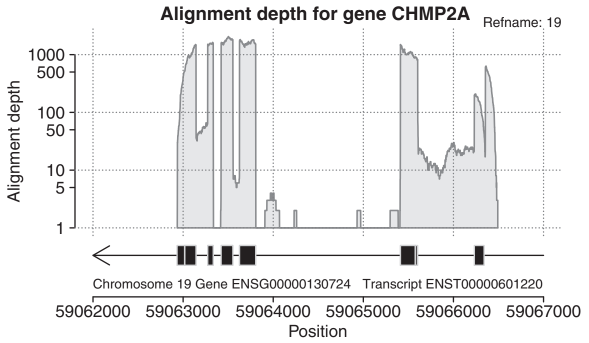

Fig. 1. Number of alignments per genomic position for gene CHMP2A

The algorithm processes 1 196 149± 536 alignments per second or 4.3 billion alignments per hour. The data inside `bamGapSites` objects can directly extracted into a `data.frame`. For each gap site, alignment (read) counts are provided which can be used for differential expression analysis and for differential splicing analysis. Gene annotation can be added by using a specialized annotation procedure for gap sites provided by the CRAN refGenome package. Further information on gap sites for example identification of non canonical splice sites, MaxEnt (Yeo and Burge, 2004) and HBond (Freund *et al.*, 2003) scores as well as information on alternative splicing events can be obtained using the Bioconductor spliceSites package.

Application of these rbamtools functions to data from an RNA-seq experiment on 60 human fibroblast samples resulted in 115 968 gap sites which are present in all samples. Thereof, 98.1 % exactly lie on annotated (Ensembl Release 74) splice sites while 1.98 % (2210 gap sites) are located on not yet annotated positions.

Acknowledgements

We thank the R Core Team (R Core Team, 2014), and in particular Profs. Brian Ripley and Kurt Hornik, very much for their fruitful comments and for their valuable help in improving *rbamtools*.

Funding

RNA-seq datasets were obtained from RNA deep sequencing projects which were partly funded by the German Ministry of Research and Education (Network Gerontosys), DFG [SCHA 909/3-1], the Heinz Ansmann Foundation for AIDS Research, Düsseldorf, Germany (He.S), and the Jürgen Manchot Stiftung (H.S.). The financial support of the Deutsche Forschungsgemeinschaft [SCHW 1508/3-1 to H.S.] is gratefully acknowledged.

Conflict of Interest: none declared.

References

Freund,M. *et al.* (2003) A novel approach to describe a u1 snrna binding site. *Nucleic Acids Res.*, **31**, 6963–6975.

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

Kaisers,W. (2013a) *refGenome: Gene and Splice Site Annotation Using Annotation Data From Ensembl and UCSC Genome Browsers.* CRAN R package version 1.3.0.

Kaisers,W. (2013b) *spliceSites: A Bioconductor Package for Exploration of Alignment Gap Positions from RNA-Seq Data.* Bioconductor R package version 1.3.3.

Morgan,M. *et al.* (2010) *Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import.* R package version 1.16.1.

R Core Team (2014) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Yeo,G. and Burge,C. (2004). Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *J. Comput. Biol*, **11**, 377–394.