Sequence analysis

ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data

Martin Morgan1,∗, Simon Anders2, Michael Lawrence1, Patrick Aboyoun1, and Robert Gentleman1

1Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, USA and
2European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK

Received on April 16, 2009; revised on June 23, 2009; accepted on July 15, 2009
Advance Access publication August 3, 2009
Associate Editor: Dmitrij Frishman

SUMMARY
Summary: ShortRead is a package for input, quality assessment, manipulation and output of high-throughput sequencing data. ShortRead is provided in the R and Bioconductor environments, allowing ready access to additional facilities for advanced statistical analysis, data transformation, visualization and integration with diverse genomic resources.

Availability and Implementation: This package is implemented in R and available at the Bioconductor web site; the package contains a ‘vignette’ outlining typical work flows.

Contact: mtmorgan@hrcrc.org

High-throughput DNA sequencing technologies include Illumina (Solexa) (Bentley et al., 2008), Roche 454 (Torres et al., 2008) and other platforms. These technologies produce millions of DNA sequences of tens to hundreds of nucleotides each. Biological questions addressed with this data include SNP calling, ChIP-sequencing of tens to hundreds of nucleotides each. Biological and other platforms. These technologies produce millions of DNA sequence data.

1 AVAILABLE FUNCTIONALITY

1.1 Input and output

ShortRead provides mechanisms for input of diverse high-throughput sequence data. A major starting point is reads aligned to references, as from manufacturer software or aligners such as MAQ (Li et al., 2008) and Bowtie (Langmead et al., 2009). ShortRead parses additional formats (e.g. fasta, fastq and arbitrary column-oriented text files). Resulting R data structures allow manipulation of sequence, quality, alignment and other information. Input functions transparently parse compressed (.gz) files; most file types can be read as ‘chunks’, to allow processing of data subsets. ShortRead inputs but does not specially represent fine-grained alignment descriptions (e.g. in Stockholm format). Facilities for data output include fasta and fastq text formats, arbitrary column-oriented output of reads and auxiliary information, serialization of objects in native R format, and through use of additional R packages such as rtracklayer export to common genome browser formats such as wiggle, bed and gff (Kuhn et al., 2008).

1.2 Quality assessment

ShortRead includes facilities for assessment (QA) of read quality, sample processing and sequencing artifacts, and alignment characteristics. The QA pipeline can start with base calls and their quality scores (e.g. fastq or quq files), as well as aligned data formats from special-purpose aligners. The result is an HTML report with embedded narrative to facilitate interpretation; a sample report is included with the package. Illustrative results are shown in Figure 1. Highlights include: (i) The number of raw, filtered and aligned reads; (ii) Base call frequencies. (iii) Cycle-specific base calls and read qualities (e.g. Fig. 1A). (iv) Tabulation of read occurrences (how often reads are represented once, twice,..., n times). For instance, reads occurring once or a few times (to the left in Fig. 1B) may be unique due to base call errors, whereas reads occurring very frequently (at the extreme right in Fig. 1B) typically reflect PCR or resequencing issues. (v) Preliminary alignment quality score summaries. Technology-specific quality measures are also generated, especially for Illumina’s Genome Analyzer (e.g. tile-specific read counts and qualities).

1.3 Transformation and downstream analysis

ShortRead provides facilities for data exploration, transformation, and down-stream analysis. For example, alphabetByCycle summarizes cycle-specific nucleotide counts (Fig. 1A) and base qualities. The alphabetFrequency function summarizes nucleotide use over all cycles, on a per-read basis or over the entire set of reads. The table function summarizes commonly occurring sequences, as illustrated in Figure 1B. ShortRead contains facilities for sorting reads, finding duplicates, trimming left and right ends and for exploiting the extensive string pattern matching functions of Biostrings.

The features described here are generally fast, operating on tens of millions of short reads in a few seconds; input of large text files...
The function reduces the data volume in the file, e.g. from be applied to the file. A typical use takes a list of file names and a function to conducted on a per-lane basis. The srapply function exploits this work flow. A typical use takes a list of file names and a function to be applied to the file. srapply applies the function to each file. Usually the function reduces the data volume in the file, e.g. from a large collection of reads to a compact summary of lane quality. The distinguishing feature of srapply is that the calculation is distributed across processors or nodes in a computer cluster, if such resources exist.

2 CONCLUSIONS

This note introduces the Bioconductor ShortRead package for analysis of resequencing data. The package allows input into R of diverse sequence-related files, and output of common data formats. It provides quality assessment tools and HTML-based report-generating functionality. ShortRead data structures allow convenient manipulation of data, such as filtering reads based on sequence characteristics. The package work flow represents an entry point for down-stream analysis using Bioconductor or other software. Future plans include improved support for longer and paired-end reads, and development of additional quantitative measures of quality; the challenge of incorporating the SOLiD color space model into standard work flows precludes support for this format beyond that available from data transformed to sequence and Phred-like quality scores.

ACKNOWLEDGEMENTS

We are grateful to early adopters and Bioconductor course participants for their helpful input.

Funding: National Human Genome Research Institute (grant P41HG004059 R.G.); EU (Research and Training Network ‘Chromatin Plasticity’ to S.A).

Conflict of Interest: none declared.

REFERENCES


Fig. 1. Quality assessment. (A) Unlikely directional nucleotide change and base calls (cycle 26) from a Short Read Archive accession. (B) Left and right ‘tails’ correspond to infrequently and highly repeated reads, respectively, in a xN174 control lane.