ShortRead R Package Report, for Computational Biology 2022/2023

L. Ganser[1], M. Fernandes[2], R. Moreira[2], and V. Dalovic[2]

[1]École Polytechnique Fédérale de Lausanne, Switzerland
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal

14 October 2022

1 Abstract

The ShortRead R package is a collection of functions for reading and manipulating short-read sequencing data. It is designed to be used in conjunction with the Bioconductor project. The package provides a unified interface to read data from a variety of sequencing platforms, including Illumina, 454, SOLiD, and Ion Torrent. It also provides functions for quality control, trimming, and filtering of reads. The package also provides functions for mapping reads to a reference genome, and for variant calling. The package is available on CRAN and Bioconductor. In this report, we explored the package while referring to

2 Introduction

ShortReads is an R package that provides a unified interface to read data from a variety of sequencing platforms, including Illumina, 454, SOLiD, and Ion Torrent. It also provides functions for quality control, trimming, and filtering of reads. The package also provides functions for mapping reads to a reference genome, and for variant calling. The package is available on CRAN and Bioconductor.

2.1 Motivation

ShortReads is a very used package and important for the Bioconductor project. It is used in many other packages and it is important to understand how it works. We decided to explore this package to learn more about it and to be able to use it in the future.

2.2 Background

ShortReads is an R package that provides a unified interface to read data from a variety of sequencing platforms, including Illumina, 454, SOLiD, and Ion Torrent. It also provides functions for quality control, trimming, and filtering of reads. The package also provides functions for mapping reads to a reference genome, and for variant calling. The package is available on CRAN and Bioconductor.

2.2.1 Phred Score

Phred score is a way to represent the probability of a base being wrong. It is used in the FASTQ format to represent the quality of a base. The Phred score is calculated as $-10 \log_{10} P$, where P is the probability of the base being wrong. The Phred score is an integer between 0 and 40. The higher the Phred score, the higher the probability of the base being wrong. The Phred score is represented by the ASCII character with the same value as the Phred score plus 33. For example, the Phred score 40 is represented by the ASCII character '!'. The Phred score is also represented by the ASCII character with the same value as the Phred score plus 64. For example, the Phred score 40 is represented by the ASCII character '@'.

2.2.2 FASTQ format

The FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. It is a text-based format for storing both a biological sequence and its corresponding quality scores. Each record contains four lines: the first line starts with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line); the second line is the raw sequence letters; the third line starts with a '+' character and is optionally followed by the same sequence identifier (and any description) again; and the fourth line encodes the quality values

for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence. The quality values are encoded using the Phred quality score plus 33. For example, a base with a quality score of 40 has a character '!' as its quality value. The quality values are also encoded using the Phred quality score plus 64. For example, a base with a quality score of 40 has a character '@' as its quality value.

2.2.3 FASTA format

The FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format is designed so that the file is human-readable, and can be parsed by software. The first line of the file must begin with a '¿', which is followed by the sequence identifier and an optional description. The sequence identifier and the description are separated by one or more spaces or tabs. The description is not used by most software. The second line contains the first line of the sequence. The sequence lines may be of any length. The sequence ends when the next line begins with a '¿'. The FASTA format is used in the ShortReads package to represent the reference genome.

3 Methods

To analyze the package, we looked into the Github repository[1] and the documentation[2]. We also used the RStudio IDE to run the examples and to test the package. We will now analyze the following functions:

- readFastq — Reads a FASTQ file and returns a `ShortRead` object.
- qa — Quality assessment of a `ShortRead` object.
- srFilter — Filters a `ShortRead` object. Pre-defined and bespoke filters are available.

3.1 The readFastq function and the ShortReadQ class

The `readFastq` function reads a FASTQ file and returns a `ShortReadQ` R object, a compact internal representation of the sequences. The `ShortReadQ`

class inherits from the `ShortRead` class and adds the quality scores to the sequences. The `ShortRead` class (see `AllClasses.R`) contains the following elements:

- `sread` — A character vector containing the sequences.
- `id` — ?A character vector containing the sequence identifiers.?

The `ShortReadQ` class (see `AllClasses.R`) inherits from the `ShortRead` class and adds the quality scores to the sequences which are represented in a character vector. The `readFastq` function (see `methods-ShortReadQ.R`) has the following arguments:

- `dirPath` (Character vector) — The path to the directory containing the FASTQ files.
- `pattern` — The pattern describing the FASTQ file names to be read. Default is `"*.fastq"`, which results in line counts for all FASTQ files in the directory.

The `readFastq` function (see `methods-ShortReadQ.R`) has the following steps:

1. The function searches for the FASTQ files in the directory specified by the `dirPath` argument matching the pattern specified by the `pattern` argument.
2. If no FASTQ files are found, the function throws an error.
3. The function reads elements using the `.read_solexa_fastq` function (?see `methods-ShortReadQ.R`?) and creates a `ShortReadQ` object.

3.2 Installation

The ShortReads package can be installed from CRAN or Bioconductor. To install the package from CRAN, run the following command in R:

```
install.packages("ShortRead")
```

To install the package from Bioconductor, run the following command in R:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("ShortRead")
```

3.3 Usage

The ShortReads package provides a unified interface to read data from a variety of sequencing platforms, including Illumina, 454, SOLiD, and Ion Torrent. It also provides functions for quality control, trimming, and filtering of reads. The package also provides functions for mapping reads to a reference genome, and for variant calling. The package is available on CRAN and Bioconductor.

3.3.1 Reads

The ShortReads package provides a unified interface to read data from a variety of sequencing platforms, including Illumina, 454, SOLiD, and Ion Torrent.

3.3.2 Quality Control

The ShortReads package provides functions for quality control, trimming, and filtering of reads.

3.3.3 Mapping

The ShortReads package provides functions for mapping reads to a reference genome.

3.3.4 Variant Calling

The ShortReads package provides functions for variant calling.

4 Results

4.1 The readFastq function

4.2 Evolution of the package over time

The package was first added to Bioconductor in version BioC 2.3 (R-2.8) in 2008. The first commit on Github was made in April 2008. The package has been updated regularly since then. The number of commits per month is shown in Figure 1.

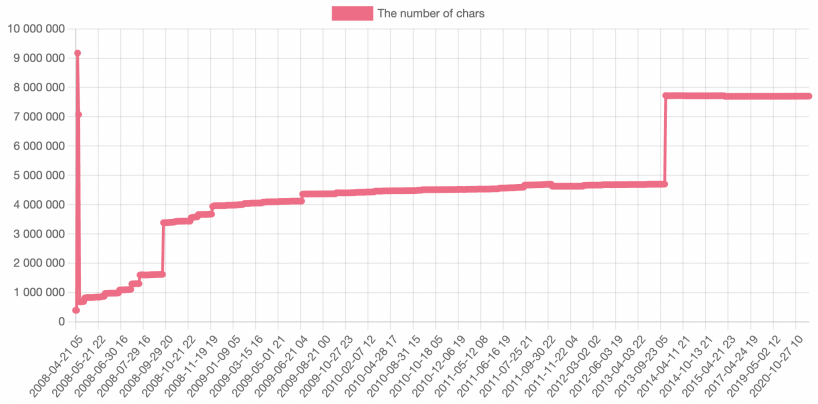

Figure 1: Number of commits per month. Created using Writer-Stock by PenguinCabinet

References

[1] ShortReads on Github. `https://github.com/Bioconductor/ShortRead`

[2] ShortReads documentation. `https://bioconductor.org/packages/release/bioc/vignettes/ShortRead/inst/doc/ShortRead.pdf`

[3] Morgan, M. T., et al. (2009). Shortread: an r/bioconductor package for input, quality control, and preprocessing of short read sequence data. *Bioinformatics*, 25(16):2078–2079.