Rsubread: An R package for aligning next-generation sequencing reads

Wei Shi

27 July 2011

1 Introduction

This package provides functions for aligning next-generation sequencing reads to the reference genome using a novel mapping algorithm which is entirely different from the "seed-and-extend" algorithm. The new algorithm uses a number of 16bp substrings extracted from each read to find the candidate mapping locations and then use the consensus mapping location to determine the final location of each read without using the costly extension step. This new algorithm has been found to be much faster, more sensitive and accurate than the "seed-and-extend" aligners.

Read mapping functions in this package are R wrapper functions which call the underlying `C` functions to perform the alignment. A standalone C implementation of this new algorithm, which is called Subread, can be downloaded from `http://subread.sourceforge.net`.

In addition to the read mapping functions, this package provides other useful functions for processing next-gen sequencing data such as summarization of read counts to genomic features, quality assessment and so on.

2 Read alignment

There are two steps for mapping reads using Rsubread:

Step 1: Build an index for reference genome
Index building in Rsubread is simple and quick. Building an index for human genome takes about 1 hour.

The Rsubread package includes a sample reference sequence which was made up from 900 read sequences. 1000 reads were extracted from one of the Human Brain Reference RNA-seq datasets generated by the SEQC project (`http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm`). Read sequences from 900 of them were concatenated to form a synthetic reference

sequence, which is used as the reference for the mapping of these 1000 reads. This small set of reads were also included in this package including their sequences and quality scores. These reads are 100bp long and generated by Illumina Genome Analyzer.

Here is an example of building an index for a reference genome using the sample reference included in this package:

```
> library(Rsubread)
> ref <- system.file("extdata", "reference.fa", package = "Rsubread")
> path <- system.file("extdata", package = "Rsubread")
> buildindex(basename = file.path(path, "reference_index"), reference = ref)
```

The created index files are saved to the "extdata" folder in the directory where Rsubread package was installed. Rsubread creates a hash table for indexing the reference genome. Keys in the hash table are the 16bp sequences and hash values are their corresponding chromosomal locations. Color space index can be built by setting the `colorsapce` argument to `TRUE`.

A unique feature of Rsubread is that it allows users to control the computer memory usage in the read mapping process. Users can do this by specifying the amount of memory (in MB) to be used for mapping. By default, 3700MB of memory will be used. This will for example partition the index into two chunks for human genome. Only one chunk of index will be existent in the memory at any time. If larger memory is used and the entire index can be loaded into the memory in one go (e.g. 7400GB of memory is used for human genome), then the running time will be reduced by half. In a comparison with six "seed-and-extend" aligner, Subread was found to to be twice as fast as the second fastest aligner and 4-50 time faster than other aligners (the default setting of Subread aligner was used for comparison). When less memory is used, running time will increase accordingly.

Step 2: Map reads to the reference genome

Mapping reads using the 1000 reads included in this package and the index built in the last step:

```
> reads <- system.file("extdata", "reads.txt", package = "Rsubread")
> align(index = file.path(path, "reference_index"), readfile1 = reads,
+     output_file = file.path(path, "alignResults.SAM"))
```

Two key parameters used by this function are the number of subreads selected (`nsubreads` option) and the consensus threshold (`TH1` option) for determining mapping locations (`TH2` for the second read in a pair). We recommend using the default setting of these parameters because they were found to perform best in the evaluations using both simulation data and real data. Up to 16 indels are allowed in the mapping (`indels` option). Paired-end read mapping is also supported (`readfile2,TH2,min_distance,max_distance`

options). The mapping can be carried out in multithread mode (`nthreads` option). Mapping results are saved in a SAM format file. Please refer to the help page for this function for more details.

This function supports the read mapping for all major platforms including Illumina GA/HiSeq, ABI SOLiD, Roche 454 and Heliscope. It can map reads of both fixed length and variable length. It is capable for mapping long reads (>200bp) as well. When mapping long reads, the default setting for `nsubreads` and `TH1` are still recommended.

3 Counting mapped reads for genomic features

The `featureCounts` function summarizes read counts to genomic features. It uses the in-built annotation files or annotation provided by users to count the number of mapped reads for each feature. The in-built annotation files include annotation information for each exon in the mouse and human genomes (using NCBI build 37.2 annotation). Read counts can be summarized to gene level or exon level. This function can be used for general-purpose read summarization as well (for both DNA-seq reads and RNA-seq reads).

4 Quality assessment

Quality scores
Quality scores give the probability of base calling being wrong for each base in the reads, which is useful for examining quality of the sequencing data. The `qualityScores` function randomly extracts quality score information for a specified number of reads from a `FASTQ` format file.

```
> library(Rsubread)
> reads <- system.file("extdata", "reads.txt", package = "Rsubread")
> x <- qualityScores(filename = reads, nreads = 1000)
> boxplot(x)
```

GC content
Bias of GC content has been reported for the sequencing data. The `atgcContent` function can be used to get the fraction of each nucleotide (A,T,G,C) in the entire data file or at each base location across all the reads.

```
> library(Rsubread)
> reads <- system.file("extdata", "reads.txt", package = "Rsubread")
> x <- atgcContent(filename = reads, basewise = FALSE)
> xb <- atgcContent(filename = reads, basewise = TRUE)
```

5 Others

Percentage of mapped reads

Function `propmapped` counts the number of mapped reads in a SAM format file and gives their proportion in all the reads:

```
> library(Rsubread)
> results <- system.file("extdata", "alignResults.SAM", package = "Rsubread")
> propmapped(results)
```

6 Citation

Yang Liao and Wei Shi. "Subread: a superfast read aligner with high sensitivity and accuracy." In preparation.

7 Authors

Wei Shi (maintainer of Rsubread), Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Australia
Yang Liao (maintainer of Subread), Department of Computer Science and Software Engineering, The University of Melbourne, Australia
Jenny Zhiying Dai, Department of Computer Science and Software Engineering, The University of Melbourne, Australia

8 Contact

Please contact Wei Shi (shi@wehi.edu.au) if you have any questions or comments.