Rsubread package: high-performance read alignment, quantification and mutation discovery

Wei Shi

2 December 2013

1 Introduction

This vignette provides a brief description to the Rsubread package. For more details, please refer to the Users Guide which can brought up in your R session via the following commands:

```
> library(Rsubread)
> RsubreadUsersGuide()
```

The Rsubread package provides facilities for processing the read data generated by next-gen sequencing technologies. These facilities include quality assessment, read alignment, read summarization, exon-exon junction detection, absolute expression calling and SNP discovery. They can be used to analyze data generated from all major sequencing platforms including Illumina GA/HiSeq, Roche 454, ABI SOLiD and Ion Torrent.

The Subread aligner (`align` function) is a highly efficient and accurate aligner for mapping genomic DNA and RNA sequencing reads. It adopts a novel mapping paradigm called "seed-and-vote". Under this paradigm, a number of 16mers (called seeds or subreads) are extracted from each read and they were mapped to the reference genome to vote for the mapping location of the read. Read mapping performed under this paradigm has been found to be more efficient and accurate than that carried out under the conventional "seed-and-extend" paradigm (Liao et al. 2013). This package also includes a program for detecting exon-exon junctions, `subjunc`, that makes use of the powerful "seed-and-vote" paradigm too.

An important step in processing next-gen sequencing data is to assign mapped reads to genomic features such as genes, exons, and promoters. This package includes a general-purpose read summarization function `featureCounts` that takes mapped reads as input and assigns them to genomic features. In-built annotations are provided for users convenience.

Different from microarray technologies, the next-gen sequencing technologies do not provide Present/Absent calls for genomic features such as genes. We have developed an algorithm to use the background noise measured from the RNA-seq data to call absolutely expressed genes. The function `detectionCall` returns a detection p value for each gene from the read mapping results.

We have also developed a new SNP calling algorithm which is being implemented in function `exactSNPs`. Our results showed that it compared favorably to competing methods, but was an order of magnitude faster.

This package also includes some other useful functions such as quality assessment (`qualityScores`, `atgcContent`), duplicate read removal (`removeDupReads`) and mapping percentage calculation (`propmapped`).

2 Read alignment

An index needs to be built first and then alignments can be carried out. Building the index is an one-off operation. The generated index can be re-used in subsequent read alignments.

Step 1: Index building

The Rsubread package includes a dummy reference sequence that was generated by concatenating 900 100bp reads that were taken from a pilot dataset generated from the SEquencing Quality Control (SEQC) project. We further extracted 100 reads from the same dataset and combine them with the 900 reads to make a read dataset for mapping. Below is the command for building an index for the reference sequence:

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

```
        ==========     _____ _    _ ____  _____  ______          _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
            ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
              ====    ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
       Rsubread 1.12.6

//=========================== indexBuilder setting ============================\\
||                                                                             ||
||                Index name : reference_index                                 ||
||               Index space : base-space                                      ||
||                    Memory : 3700 Mbytes                                     ||
||          Repeat threshold : 24 repeats                                      ||
||                                                                             ||
||               Input files : 1 file in total                                 ||
||                             o /tmp/RtmpB21K6b/Rinst159667f73b23/Rsubre ...  ||
||                                                                             ||
\\===================== http://subread.sourceforge.net/ =======================//

//================================== Running ==================================\\
||                                                                             ||
|| Check the integrity of provided reference sequences ...                     ||
|| No format issues were found                                                ||
|| Scan uninformative subreads in reference sequences ...                     ||
||    8%,   0 mins elapsed, rate=17.3k bps/s, total=0m                        ||
||   16%,   0 mins elapsed, rate=34.6k bps/s, total=0m                        ||
||   24%,   0 mins elapsed, rate=51.9k bps/s, total=0m                        ||
||   33%,   0 mins elapsed, rate=69.0k bps/s, total=0m                        ||
||   41%,   0 mins elapsed, rate=86.1k bps/s, total=0m                        ||
||   49%,   0 mins elapsed, rate=103.3k bps/s, total=0m                       ||
||   58%,   0 mins elapsed, rate=120.3k bps/s, total=0m                       ||
||   66%,   0 mins elapsed, rate=137.1k bps/s, total=0m                       ||
||   74%,   0 mins elapsed, rate=153.9k bps/s, total=0m                       ||
||   83%,   0 mins elapsed, rate=171.0k bps/s, total=0m                       ||
||   91%,   0 mins elapsed, rate=187.7k bps/s, total=0m                       ||
||   99%,   0 mins elapsed, rate=204.3k bps/s, total=0m                       ||
|| 1 uninformative subreads were found.                                       ||
|| These subreads were excluded from index building.                          ||
|| Build the index...                                                         ||
||    8%,   0 mins elapsed, rate=7456.5k bps/s, total=0m                      ||
||   16%,   0 mins elapsed, rate=4970.6k bps/s, total=0m                      ||
||   24%,   0 mins elapsed, rate=5592.1k bps/s, total=0m                      ||
||   33%,   0 mins elapsed, rate=4970.6k bps/s, total=0m                      ||
||   41%,   0 mins elapsed, rate=5325.7k bps/s, total=0m                      ||
||   49%,   0 mins elapsed, rate=4970.8k bps/s, total=0m                      ||
||   58%,   0 mins elapsed, rate=5219.2k bps/s, total=0m                      ||
||   66%,   0 mins elapsed, rate=5422.6k bps/s, total=0m                      ||
||   74%,   0 mins elapsed, rate=5161.8k bps/s, total=0m                      ||
||   83%,   0 mins elapsed, rate=5325.7k bps/s, total=0m                      ||
||   91%,   0 mins elapsed, rate=5126.0k bps/s, total=0m                      ||
||   99%,   0 mins elapsed, rate=5263.1k bps/s, total=0m                      ||
|| Save current index block...                                                ||
||  [ 0.0% finished ]                                                         ||
||  [ 10.0% finished ]                                                        ||
||  [ 20.0% finished ]                                                        ||
||  [ 30.0% finished ]                                                        ||
||  [ 40.0% finished ]                                                        ||
||  [ 50.0% finished ]                                                        ||
||  [ 60.0% finished ]                                                        ||
||  [ 70.0% finished ]                                                        ||
||  [ 80.0% finished ]                                                        ||
||  [ 90.0% finished ]                                                        ||
||  [ 100.0% finished ]                                                       ||
||                                                                            ||
||                      Total running time: 0.0 minutes.                      ||
||               Index reference_index was successfully built!                ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//
```

The generated index files were saved to the current working directory. 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 the human genome. Only one chunk of index will be existent in the memory at any time. To load the entire index into the memory, users can specify the requested amount of memory to be 8000MB for the human genome (the actual memory usage is up to 7.6GB). With this setting, Subread has the highest mapping speed.

Step 2: read mapping

After the index was successfully built, we map the read dataset (including 1,000 reads) to the reference sequence:

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

        ==========     _____ _    _ ____  _____  ______          _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
            ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
              ====    ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
       Rsubread 1.12.6

//========================== subread-align setting ===========================\\
||                                                                            ||
||           Function : Read alignment                                        ||
||            Threads : 1                                                     ||
||         Input file : /tmp/RtmpB21K6b/Rinst159667f73b23/Rsubread/extdat ... ||
||        Output file : alignResults.SAM (SAM)                                ||
||         Index name : reference_index                                       ||
||       Phred offset : 33                                                    ||
||                                                                            ||
||          Min votes : 3                                                     ||
||         Max indels : 5                                                     ||
||  # of Best mapping : 1                                                     ||
||     Unique mapping : yes                                                   ||
||   Hamming distance : yes                                                   ||
||     Quality scores : no                                                    ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//

//====================== Running (18-Dec-2013 21:22:47) ======================\\
||                                                                            ||
|| The input file contains base space reads.                                  ||
|| Load the 1-th index block...                                               ||
|| Map reads...                                                               ||
|| Detect indels...                                                           ||
|| Realign reads...                                                           ||
|| 1000 reads were processed. Save the mapping results for them...            ||
||                                                                            ||
||                          Completed successfully.                           ||
||                                                                            ||
\\============================================================================//

//================================= Summary ==================================\\
||                                                                            ||
||          Processed : 1000 reads                                            ||
||             Mapped : 902 reads (90.2%)                                     ||
||             Indels : 0                                                     ||
||                                                                            ||
||       Running time : 0.0 minutes                                           ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//
```

Map paired-end reads:

```
> reads1 <- system.file("extdata","reads1.txt",package="Rsubread")
> reads2 <- system.file("extdata","reads2.txt",package="Rsubread")
> align(index="reference_index",readfile1=reads1,readfile2=reads2,
+ output_file="alignResultsPE.SAM")

        ==========     _____ _    _ ____  _____  ______      _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
            ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
              ====    ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
       Rsubread 1.12.6

//========================== subread-align setting ===========================\\
||                                                                            ||
||           Function : Read alignment                                        ||
||            Threads : 1                                                     ||
||       Input file 1 : /tmp/RtmpB21K6b/Rinst159667f73b23/Rsubread/extdat ... ||
||       Input file 2 : /tmp/RtmpB21K6b/Rinst159667f73b23/Rsubread/extdat ... ||
||        Output file : alignResultsPE.SAM (SAM)                              ||
||         Index name : reference_index                                       ||
||       Phred offset : 33                                                    ||
||                                                                            ||
||    Min read1 votes : 3                                                     ||
||    Min read2 votes : 1                                                     ||
||  Max fragment size : 600                                                   ||
||  Min fragment size : 50                                                    ||
||                                                                            ||
||         Max indels : 5                                                     ||
||  # of Best mapping : 1                                                     ||
||     Unique mapping : yes                                                   ||
||   Hamming distance : yes                                                   ||
||     Quality scores : no                                                    ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//

//====================== Running (18-Dec-2013 21:22:48) ======================\\
||                                                                            ||
|| The input file contains base space reads.                                  ||
|| Load the 1-th index block...                                               ||
|| Map fragments...                                                           ||
|| Detect indels...                                                           ||
|| Realign fragments...                                                       ||
|| 1000 fragments were processed. Save the mapping results for them...        ||
||                                                                            ||
||                          Completed successfully.                           ||
||                                                                            ||
\\============================================================================//

//================================= Summary ==================================\\
||                                                                            ||
||          Processed : 1000 fragments                                        ||
||             Mapped : 907 fragments (90.7%)                                 ||
||   Correctly paired : 898 fragments                                         ||
||             Indels : 0                                                     ||
||                                                                            ||
||       Running time : 0.0 minutes                                           ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//
```

3 Counting mapped reads for genomic features

The `featureCounts` function is a general-purpose read summarization function that assigns mapped reads (RNA-seq or gDNA-seq reads) to genomic features such as genes, exons, promoters, gene bodies and genomic bins.

This function takes as input a set of files that contain read mapping results and an annotation file that includes genomic features. It automatically detects the format of input read files (SAM or BAM). It also automatically re-orders the paired reads if they are not in consecutive positions in the input. In-built NCBI RefSeq gene annotations for genomes mm9, mm10 and hg19 are provided for users convenience.

Below gives the example code of assigning reads and fragments generated in the last section to two genes. Assign single end reads to genes:

```
> ann <- data.frame(
+ GeneID=c("gene1","gene1","gene2","gene2"),
+ Chr="chr_dummy",
+ Start=c(100,1000,3000,5000),
+ End=c(500,1800,4000,5500),
+ Strand=c("+","+","-","-"),
+ stringsAsFactors=FALSE)
> ann
```

```
  GeneID       Chr Start  End Strand
1  gene1 chr_dummy   100  500      +
2  gene1 chr_dummy  1000 1800      +
3  gene2 chr_dummy  3000 4000      -
4  gene2 chr_dummy  5000 5500      -
```

```
> fc_SE <- featureCounts("alignResults.SAM",annot.ext=ann)
```

```
        ==========     _____ _    _ ____  _____  ______          _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
            ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
              ====    ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
       Rsubread 1.12.6

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 SAM file                                       ||
||                           o alignResults.SAM                               ||
||                                                                            ||
||             Output file : ./.Rsubread_featureCounts_pid8071                ||
||             Annotations : ./.Rsubread_UserProvidedAnnotation_pid8071 ( ... ||
||                                                                            ||
||                 Threads : 1                                                ||
||                   Level : meta-feature level                               ||
||              Paired-end : no                                               ||
||         Strand specific : no                                               ||
||      Multimapping reads : not counted                                      ||
|| Multi-overlapping reads : not counted                                      ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//

//================================= Running ==================================\\
||                                                                            ||
|| Load annotation file ./.Rsubread_UserProvidedAnnotation_pid8071 ...        ||
||    Number of features is 4                                                 ||
||    Number of meta-features is 2                                            ||
||    Number of chromosomes is 1                                              ||
||                                                                            ||
|| Process SAM file alignResults.SAM...                                       ||
||    Assign reads to features...                                             ||
||    Total number of reads is : 1000                                         ||
||    Number of successfully assigned reads is : 29 (2.9%)                    ||
||    Running time : 0.00 minutes                                             ||
||                                                                            ||
||                         Read assignment finished.                          ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//

> fc_SE

$counts
      alignResults.SAM
gene1               13
gene2               16

$annotation
  GeneID                 Chr     Start       End Strand Length
1  gene1 chr_dummy;chr_dummy  100;1000  500;1800    +;+   1202
2  gene2 chr_dummy;chr_dummy 3000;5000 4000;5500    -;-   1502

$targets
[1] "alignResults.SAM"

$stat
                      Status alignResults.SAM
1                   Assigned               29
2       Unassigned_Ambiguity                0
3    Unassigned_MultiMapping                0
4      Unassigned_NoFeatures              873
5        Unassigned_Unmapped               98
6  Unassigned_MappingQuality                0
7 Unassigned_FragementLength                0
8         Unassigned_Chimera                0
```

Assign fragments (read pairs) to genes:

```
> fc_PE <- featureCounts("alignResultsPE.SAM",annot.ext=ann,isPairedEnd=TRUE)

        ==========     _____ _    _ ____  _____  ______          _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====      | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
            ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
              ====    ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
       Rsubread 1.12.6

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 SAM file                                       ||
||                           o alignResultsPE.SAM                             ||
||                                                                            ||
||             Output file : ./.Rsubread_featureCounts_pid8071                ||
||             Annotations : ./.Rsubread_UserProvidedAnnotation_pid8071 ( ... ||
||                                                                            ||
||                 Threads : 1                                                ||
||                   Level : meta-feature level                               ||
||              Paired-end : yes                                              ||
||         Strand specific : no                                               ||
||      Multimapping reads : not counted                                      ||
|| Multi-overlapping reads : not counted                                      ||
||                                                                            ||
||          Chimeric reads : counted                                          ||
||        Both ends mapped : not required                                     ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//

//================================= Running ==================================\\
||                                                                            ||
|| Load annotation file ./.Rsubread_UserProvidedAnnotation_pid8071 ...        ||
||    Number of features is 4                                                 ||
||    Number of meta-features is 2                                            ||
||    Number of chromosomes is 1                                              ||
||                                                                            ||
|| Process SAM file alignResultsPE.SAM...                                     ||
||    Assign fragments (read pairs) to features...                            ||
||    Each fragment is counted once.                                          ||
||    Total number of fragments is : 1000                                     ||
||    Number of successfully assigned fragments is : 34 (3.4%)                ||
||    Running time : 0.00 minutes                                             ||
||                                                                            ||
||                         Read assignment finished.                          ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//

> fc_PE

$counts
      alignResultsPE.SAM
gene1                 15
gene2                 19

$annotation
  GeneID                 Chr     Start       End Strand Length
1  gene1 chr_dummy;chr_dummy  100;1000  500;1800    +;+   1202
2  gene2 chr_dummy;chr_dummy 3000;5000 4000;5500    -;-   1502

$targets
[1] "alignResultsPE.SAM"

$stat
                      Status alignResultsPE.SAM
1                   Assigned                 34
2       Unassigned_Ambiguity                  0
3    Unassigned_MultiMapping                  0
4      Unassigned_NoFeatures                873
5        Unassigned_Unmapped                 93
6  Unassigned_MappingQuality                  0
7 Unassigned_FragementLength                  0
8         Unassigned_Chimera                  0
```

4 Finding exon junctions

The RNA-seq technology provides a unique opportunity to identify the alternative splicing events that occur during the gene transcription process. The subjunc function can be used to detect exon-exon junctions. It first extracts a number of subreads (16mers) from each read, maps them to the reference genome and identifies the two best mapping locations for each read (representing potential locations of exons spanned by the read). Then, it builds a junction table including all putative junctions. Finally, it carries out a verification step to remove false positives in junction detection by realigning all the reads. The donor ('GT') and receptor sites('AG'), are required to be present when calling exon-exon junctions. Output of this function includes the discovered exon-exon junctions and also read mapping results.

5 Base quality scores

Quality scores give the probabilities of read bases being incorrectly called, which is useful for examining the quality of sequencing data. The `qualityScores` function can be used to quickly retrieve and display the quality score data extracted from a read file.

```
> x <- qualityScores(filename=reads,input_format="FASTQ",offset=64,nreads=1000)

qualityScores Rsubread 1.12.6

Scan the input file...
Totally 1000 reads were scanned; the sampling interval is 1.0.
Now extract read quality information...

Program finished successfully. 999 reads were scored.

> x[1:10,1:10]

       1  2  3  4  5  6  7  8  9 10
 [1,] 33 33 33 20 20 24 31 15 21 16
 [2,] 33 33 30 33 33 30 34 30 32 28
 [3,] 32 33 33 32 33 33 33 20 32 24
 [4,] 33 33 33 33 33 30 29 34 31 25
 [5,] 33 33 33 33 33 34 34 34 33 30
 [6,] 33 30 31 24 24 28 33 33 30 32
 [7,] 33 33 33 33 30 28 17 25 31 33
 [8,] 33 32  2  2  2  2  2  2  2  2
 [9,] 33 33 33 34 33 33 31 33 33 33
[10,] 33 33 33 33 28 24 33 33 33 28
```

6 GC content

The `atgcContent` function returns the fraction of A, T, G and C bases in all the reads or at each base location of the reads.

```
> ## Fraction of A,T,G and C bases in all the reads
> x <- atgcContent(filename=reads)
> x

          A      T      G      C     N
[1,] 0.1983 0.2192 0.3343 0.2482 1e-04

> ## Fraction of A,T,G and C bases at each read location
> xb <- atgcContent(filename=reads,basewise=TRUE)
> xb[,1:5]
```

```
      1     2     3     4     5
A 0.069 0.189 0.214 0.286 0.263
T 0.113 0.265 0.213 0.153 0.202
G 0.386 0.309 0.299 0.320 0.321
C 0.432 0.237 0.274 0.241 0.214
N 0.000 0.000 0.000 0.000 0.000
```

7 Mapping percentage

Function propmapped returns the proportion of mapped reads include in a SAM/BAM file.

```
> propmapped("alignResults.SAM")
```

```
          Samples NumTotal NumMapped PropMapped
1 alignResults.SAM     1000       902      0.902
```

8 Citation

Yang Liao, Gordon K Smyth and Wei Shi (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108.

Yang Liao, Gordon K Smyth and Wei Shi (2013). featureCounts: an efficient general-purpose read summarization program. Bioinformatics, In Press, accepted Nov 7.

9 Authors

Wei Shi and Yang Liao
Bioinformatics Division
The Walter and Eliza Hall Institute of Medical Research
1G Royal Parade, Parkville, Victoria 3052
Australia

10 Contact

Please contact Wei Shi (shi at wehi dot edu dot au) if you have any inquires.