clonotypeR: Identify and analyse B and T cell receptors at a high throughput

Author: Charles Plessy <plessy@riken.jp>
Date: 15 Apr 2013

clonotypeR is a R package and accompanying scripts to identify and analyse clonotypes from high-throughput T cell receptors sequence libraries. *clonotypeR* is suited to process and organise very large number of clonotypes, in the order of millions, typically produced by Roche 454 instruments, and to prepare these sequences for differential expression analysis with the typical transcriptomics tools as well as for statistical analysis using existing R packages.

The home page of *clonotypeR* is `http://clonotyper.branchable.com/`.

clonotypeR's workflow

Typically, the user receives the output of a next-generation sequencer and runs some shell commands that are not part of the *clonotypeR* R package, but that are distributed with it on `http://clonotyper.branchable.com/`.

This workflow summarises the different commands to run. Other examples are available on line at `http://clonotyper.branchable.com/doc/workflow/`.

This example analysis assumes a unix system (Linux, Mac OS, . . .)

Example data

The data provided on-line at `http://clonotyper.branchable.com/example_data/` is a sub-sample of three sequence librairies (2,000 reads each) made on the 454 Titanium or the 454 junior platforms. The original libraries will be deposited in public databanks after publication in a peer-reviewed journal.

These example libraries are called `A`, `B` and `C`, and are in FASTQ format, with entries like the following (the sequence was truncated for the convenience of the display).

```
@HKTLYLP01BOMTM
gactGTCCATCTTCCTTTTATCGGACACTGAAGTATGGATATCAGAAGTGCAgggccttcccacgggaacg
+
IIIIIIIIIIIHHFF::::G&gt;IIIGGGIIIIIIIIGGIIIIIIIFEBDCDC&lt;//-5522------
```

Detection of V segments

Run the command `clonotyper detect A.fastq` in the same directory as a copy of the file `A.fastq`.

The result is stored in a temporary directory called `extraction_files`, that will be created if it does not already exist.

`clonotyper detect` compares the sequences to the reference V segments using BWA, and produces output like the following.

```
[bsw2_aln] read 2000 sequences/pairs (843395 bp)...
[samopen] SAM header is present: 167 sequences.
[main] Version: 0.6.2-r126
[main] CMD: bwa bwasw -t8 /usr/share/clonotypeR/references/V/index A.fastq
[main] Real time: 1.099 sec; CPU: 8.225 sec
```

This indicates that 2,000 reads have been processed, representing 843,395 base pairs in total. There were 167 reference V segments, and the version number of BWA was `0.6.2-r126`. The whole process took less than 10 seconds.

Process the example libraries `B` and `C` similarly with the commands `clonotyper detect C.fastq` and `clonotyper detect C.fastq`.

Extraction of CDR3 regions

Run the command `clonotyper extract A` in the same directory as where you ran `clonotyper detect A.fastq`. The result is a table stored in a directory called `clonotypes`, that will be created if it does not already exist.

The output is quite voluminous, and indicates which V / J combinations are being found, like on the following.

```
TRAV14-3    233
    TRAJ61  0
    TRAJ60  0
    TRAJ59  0
    TRAJ58  1
    TRAJ57  39
    TRAJ56  2
    TRAJ55  0
```

The format of the table is explained in the manual page of the function `read_clonotypes()` of the R package.

For each library (`A`, `B` and `C`), one file is available in the `clonotypes` directory. With BWA `0.6.2-r126`, the following numbers of clonotypes are found.

```
  1072 clonotypes/A.tsv
   924 clonotypes/B.tsv
   689 clonotypes/C.tsv
```

The files need to be concatenated before analysis in `R`, with the following command.

```
find clonotypes/ -name '*tsv' | xargs cat > clonotypes.tsv
```

Data analysis in R

Load the clonotypeR library: `library(clonotypeR)`

Load the data in a R object called *clonotypes*: `clonotypes <- read_clonotypes('clonotypes.tsv')`

The command `summary(clonotypes)` already provides useful information.

```
> summary(clonotypes)
 lib                    V             J            read
 A:1072   TRAV14-1       :944   TRAJ31 : 380   Length:2684
 B: 924   TRAV14-2       :237   TRAJ23 : 270   Class :character
 C: 688   TRAV14-3       :251   TRAJ22 : 257   Mode  :character
          TRAV14D-3/DV8  :242   TRAJ37 : 156
          TRAV14N-1_14D-1:604   TRAJ34 : 141
          TRAV14N-2_14D-2:235   TRAJ40 : 104
          TRAV14N-3      :171   (Other):1376
     dna                qual               pep             unproductive
 Length:2684        Length:2684        Length:2684        Mode :logical
 Class :character   Class :character   Class :character   FALSE:2130
 Mode  :character   Mode  :character   Mode  :character   TRUE :554
                                                          NA's :0
```

Identify unique clonotypes, count their sequences in the libraries `A`, `B` and `C`, and store the result as a table arbitrarily named `abc`.

```
> abc <- clonotype_table(c('A','B','C'))

> head(abc)
                            A B C
TRAV14-1 AAASSGSWQLI TRAJ22 1 0 0
TRAV14-1 AACNNRIF TRAJ31    1 0 0
TRAV14-1 AAGAKLT TRAJ39     3 0 0
TRAV14-1 AAGGSWQLI TRAJ22   1 0 0
TRAV14-1 AAGTNTGKLT TRAJ27  1 0 0
```

```
TRAV14-1 AAHDTNAYKVI TRAJ30 1 0 0

> summary(abc)
       A                 B                 C
 Min.   : 0.0000   Min.   : 0.0000   Min.   :  0.0000
 1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:  0.0000
 Median : 0.0000   Median : 0.0000   Median :  0.0000
 Mean   : 0.7599   Mean   : 0.6606   Mean   :  0.5018
 3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.:  0.0000
 Max.   :18.0000   Max.   :22.0000   Max.   :121.0000
```

The summary shows that the most frequent clonotype is in `C`. Using `R` index vectors, we can see that its CDR3 sequence is AASDSNNRIF and that it was not found in the other libraries.

```
> abc[C == 121,]
                  A B   C
TRAV14N-1_14D-1 AASDSNNRIF TRAJ31 0 0 121
```

The `clonotype_table` function can also produce a count table for and combination of *V*, *CDR3* or *J* segments.

```
> clonotype_table(c('A','B','C'), "V")
                  A   B   C
TRAV14-1        239 493   0
TRAV14-2        131  61   0
TRAV14-3         79   9 113
TRAV14D-3/DV8   140  50   4
TRAV14N-1_14D-1  78  24 388
TRAV14N-2_14D-2  81  61  49
TRAV14N-3        94  34   2

> head(clonotype_table(c('A','B','C'), c("V","J")))
                 A  B C
TRAV14-1 TRAJ11  1  1 0
TRAV14-1 TRAJ12  2  2 0
TRAV14-1 TRAJ13  2  1 0
TRAV14-1 TRAJ15 10 11 0
TRAV14-1 TRAJ16  4  3 0
TRAV14-1 TRAJ18  1 21 0
```