Reconstruct clonal germline sequences

Kenneth B. Hoehn

2023-12-21

Contents

Before B cell lineage trees can be built, it is necessary to construct the unmutated germline sequence for each B cell clone. Typically the IGH D segment is masked, because the junction region of heavy chains often cannot be reliably reconstructed.

Identify clonal clusters

Before doing anything in Dowser, it is necessary to identify clonal clusters among B cells. This is not handled in Dowser, but is handled in our related package, `SCOPer`. More information about this can be found at the SCOPer documentation site.

Obtain IMGT-gapped sequences

The international ImMunoGeneTics information system (IMGT) reference database can be most easily obtained by downloading the Immcantation repository and running a script `fetch_imgtdb.sh` to download and format the IMGT reference database. The following commands are designed for Linux/Mac, but similar commands can be run for Windows. The `<data directory>` can be any directory you would like to place the Immcantation repository and IMGT germlines.

These commands will create a series of directories containing the IMGT reference directories of their respective species.

```
# Enter these commands in a terminal, not an R session!

# Move to the directory of interest
mkdir germlines

# Download the Immcantation repository
git clone https://bitbucket.org/kleinstein/immcantation

# Run script to obtain IMGT gapped sequences
immcantation/scripts/fetch_imgtdb.sh -o germlines

# View added directories
```

```
ls germlines
# human  IMGT.yaml  immcantation  mouse  rabbit  rat  rhesus_monkey
```

Construct clonal germlines

To reconstruct clonal germlines, read in the IMGT-gapped sequence directory and supply it, along with your data, to the createGermlines function.

Input data can be from multiple loci (this is different from older Dowser versions). However, the input reference sequences must be from one organism, such as human.

```
library(dowser)
library(dplyr)

data(ExampleAirr)

# Read in IMGT-gapped sequences
references = readIMGT(dir = file.path("germlines", "human", "vdj"))

# remove germline alignment columns for this example
db = select(ExampleAirr, -"germline_alignment",
    -"germline_alignment_d_mask")

# Reconstruct germline sequences
ExampleAirr = createGermlines(db, references, nproc=1)

# Check germline of first row
ExampleAirr$germline_alignment_d_mask[1]

# "CAGGTGCAGCTGGTGGAGTCTGGGGGA...GGCTTGGTCAAGCCTGGAGGGTCCCTGAGACTCTCCTGTGCAGCCTCTGGATTCACCTTC.
```