sapFinder User Guide

Bo Wen, Shaohang Xu

May 3, 2016

Contents

1 Introduction **1**

2 Variation-associated database construction **2**

3 Based on sample-specific SNV data **2**

3.1 Input data . . . 2

3.1.1 Input data . . . 2

3.1.2 Preparing annotation files from UCSC table brower . . . 2

3.1.3 The external cross reference file . . . 3

3.1.4 Example code . . . 4

3.2 Based on public SNV database . . . 5

4 MS/MS data searching **5**

5 Post-processing **6**

6 HTML-based report generation **6**

7 Integrated function `easyRun` **7**

8 Session Info **8**

1 Introduction

This vignette describes the functionality implemented in the *sapFinder* package. *sapFinder* is developed to automate

1. variation-associated database construction from public single nucleotide variations (SNVs) database or sample-specific genome-wide association studies (GWAS) and RNA-Seq data;
2. database searching;
3. post-processing;

4. HTML-based report generation.

2 Variation-associated database construction

Currently, two kinds of variation-associated databases can be constructed by using *sapFinder* package. One is the sample-specific variation-associated database and the other is the aggregate database that is created from the public SNV repositories, such as dbSNP [7] and COSMIC [2].

3 Based on sample-specific SNV data

3.1 Input data

3.1.1 Input data

To construct sample-specific variation-associated database by using *sapFinder*, three files are required as input. One is a Variant Call Format (VCF) file which can be generated from a BAM file using single nucleotide polymorphism calling tools such as SAMtools [4] and the Genome Analysis Toolkit (GATK) [5]. The other two files are gene annotation file and FASTA format mRNA sequence file which can be downloaded by users from the University of California, Santa Cruz (UCSC) table browser. For non-model organisms, users can manually provide these files in the format of NCBI or ENSEMBL.

3.1.2 Preparing annotation files from UCSC table brower

To map variation information to the protein level, numerous pieces of genome annotation information are needed, such as exon region boundary, CDS region boundary, mRNA sequence et al. It is possible to manually download these data from The Table Browser of UCSC (http://genome.ucsc.edu/cgi-bin/hgTables?command=start).

Currently,before the construction of variation-associated database,it requires users to download a tab-separated positional table annotation file and a corresponding mRNA sequence FASTA file from UCSC table brower. Since Refseq updates from time to time, we suggest generating those files in a same day as running.

The bullet list below summarizes the steps to download RefSeq genes annotation file and mRNA Sequence file.

- Go to UCSC Table browser
- Choose genome (e.g. ”Human”)
- Choose assembly (e.g. ”2009 GRCh37/hg19”)
- Choose group (e.g. ”Genes and Gene Predictions”)
- Choose track (e.g. ”RefSeq Genes”)
- Choose table (e.g. ”refGene”)

- Choose region (e.g. "genome")
- Choose output format "all fields from selected table" (retrieves annotation file)
  - Enter output filename(e.g. `"hg19_refGene.txt"`)
  - Press "get output" button
- Choose output format "sequence" (retrieves mRNA file)
  - Enter output filename(e.g. `"hg19_refGeneMrna.fa"`)
  - Press "get output" button
  - Select "mRNA" sequence type,and press "submit" button

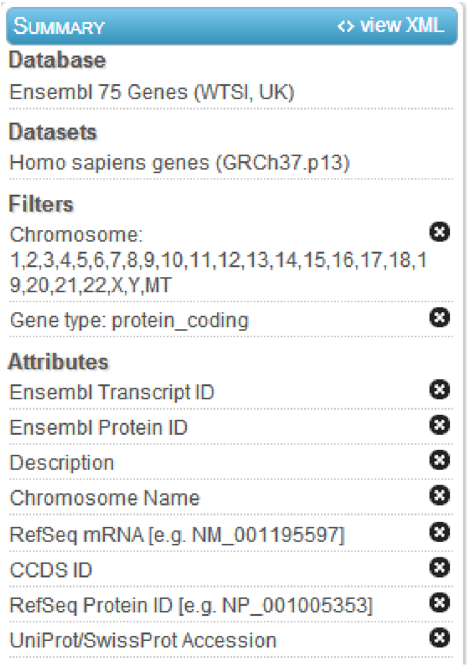
The bullet list below summarizes the steps to download Ensembl genes annotation file and mRNA Sequence file.

- Go to UCSC Table browser
- Choose genome (e.g. "Human")
- Choose assembly (e.g. "2009 GRCh37/hg19")
- Choose group (e.g. "Genes and Gene Predictions")
- Choose track (e.g. "Ensembl Genes")
- Choose table (e.g. "ensGene")
- Choose region (e.g. "genome")
- Choose output format "all fields from selected table" (retrieves annotation file)
  - Enter output filename(e.g. `"hg19_ensGene.txt"`)
  - Press "get output" button
- Choose output format "sequence" (retrieves mRNA file)
  - Enter output filename(e.g. `"hg19_ensGeneMrna.fa"`)
  - Press "get output" button
  - Select "genomic" sequence type,and press "submit" button
  - In Sequence Retrieval Region Options.Select three checkboxes("5'UTR Exons","CDS Exons" and "3'UTR Exons") and select one radiobutton("One FASTA record per gene.")
  - In Sequence Formatting Options.Select "Exons in upper case, everything else in lower case." button
  - Press "get sequence" button

Users need only to choose one type of annotation files above (Refseq or Ensembl) to download.

3.1.3 The external cross reference file

The xref file is an optional input for sapFinder.You can obtain it from BioMart Central Portal (http://central.biomart.org/martwizard/#!/Search_by_database_name?mart=Ensembl75Genes(WTSI,UK)) or MartView (http://biomart.intogen.org/biomart/martview/),The following table shows the major features must be selected:



SUMMARY <> view XML

Database

Ensembl 75 Genes (WTSI, UK)

Datasets

Homo sapiens genes (GRCh37.p13)

Filters

Chromosome: 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y,MT

Gene type: protein_coding

Attributes

Ensembl Transcript ID

Ensembl Protein ID

Description

Chromosome Name

RefSeq mRNA [e.g. NM_001195597]

CCDS ID

RefSeq Protein ID [e.g. NP_001005353]

UniProt/SwissProt Accession

Figure 1: Summary of biomart features

3.1.4 Example code

In additon to download the annotation files and xref file by yourself, you can also obtain them from the sapfinder_pipeline repository in bitbucket (https://bitbucket.org/xushaohang/sapfinder_pipeline) in the "annotation_files" directory.

The example data is extracted from a recently publised study [8].

```
> library(sapFinder)
> vcf <- system.file("extdata/sapFinder_test.vcf",
+                      package="sapFinder")
> annotation <- system.file("extdata/sapFinder_test_ensGene.txt",
+                      package="sapFinder")
> refseq <- system.file("extdata/sapFinder_test_ensGeneMrna.fa",
+                      package="sapFinder")
> xref       <- system.file("extdata/sapFinder_test_BioMart.Xref.txt",
+                      package="sapFinder")
> outdir <- "db_dir"
> prefix <- "sapFinder_test"
> db.files <- dbCreator(vcf=vcf, annotation=annotation,
+                      refseq=refseq, outdir=outdir,
+                      prefix=prefix,xref=xref)
```

Two files are outputed. One is a variation-associated database file which is written in FASTA format to the directory specified and contains the mutated peptides, the normal protein sequences and their reverse counterparts. The other is a tab-delimited file which contains the variant peptides information. Both files will be used in the following steps.

3.2 Based on public SNV database

The usage of creating variation-associated database from public SNV repositories is same as that based on sample-specific SNV data. Currently, *sapFinder* can be used to create variation-associated database based on the data from dbSNP [7] and COSMIC [2]. The required VCF files can be downloaded from their ftp sites.

4 MS/MS data searching

After the variation-associated database constructed, *rTANDEM* package [3] is adopted to search the database against tandem mass spectra to detect variant peptides. *rTANDEM* package interfaces with the popular used open source search engine *X!Tandem* [1] algorithm in R.

```
> outdir<-"."
> mgf.path <- system.file("extdata/sapFinder_test.mgf",
+                         package="sapFinder")
> protein.db <- db.files[1]
> xml.path <- runTandem(spectra=mgf.path, fasta=protein.db,
+                         outdir = outdir,tol=10, tolu="ppm",
+                         itol=0.1, itolu="Daltons")
```

```
2016-05-03 22:22:39
Loading spectra
 (mgf). loaded.
Spectra matching criteria = 331
Starting threads . started.
Computing models:
        t
                sequences modelled = 0 ks
Model refinement:
Creating report:
        initial calculations  ..... done.
        sorting  ..... done.
        finding repeats ..... done.
        evaluating results ..... done.
        calculating expectations ..... done.
        writing results ..... done.

Valid models = 329
```

```
Unique models = 208
Estimated false positives = 3 +/- 2
```

The results are written in xml format to the directory specified and will be loaded for further processing.

5 Post-processing

After the MS/MS data searching, the function tanparser can be used to parse the search result. It calculates the q-value for each peptide spectrum matches (PSMs) and then utilizes the Occam's razor approach [6] to deal with degenerated wild peptides by finding a minimum subset of proteins that covered all of the identified wild peptides.

```
> parserGear(file=xml.path, db=db.files[1],
+              outdir='parser_outdir', prefix=prefix)
```

It exports some tab-delimited files containing the peptide identification result and protein identification result. The annotated spectra for the identified variant peptides which pass the threshold are exported.

This function also accepts the "raw" Mascot result file as input(dat format). For instance,

```
> dat_file<-"mascot_raw.dat"
> parserGear(file=dat_file, db=db.files[1],
+              outdir='parser_outdir', prefix=prefix)
```

Unfortunately,we don't offer the wrapper function for Mascot search under current conditions. So you have to launch the independent identification by Mascot.

6 HTML-based report generation

The results are then summarised and compiled into an interactive HTML report.

```
> reportCreator(indir="parser_outdir",
+                  db= db.files[1], varInfor=db.files[2],prefix=prefix)
```

```
    Step 1: Reading the Info.
    Step 2: Spectrum plotting.
    Step 3: Creating the html pages.
```

After the analysis has completed, the file 'index.html' in the output directory can be opened in a web browser to access report generated.

7 Integrated function easyRun

The function easyRun automates the data analysis process. It will process the dataset in the following way:

1. Variation-associated database construction
2. MS/MS searching
3. Post-processing
4. HTML-based report generation

This function can be called as following:

```
> vcf        <- system.file("extdata/sapFinder_test.vcf",
+                          package="sapFinder")
> annotation <- system.file("extdata/sapFinder_test_ensGene.txt",
+                          package="sapFinder")
> refseq     <- system.file("extdata/sapFinder_test_ensGeneMrna.fa",
+                          package="sapFinder")
> mgf.path   <- system.file("extdata/sapFinder_test.mgf",
+                          package="sapFinder")
> xref       <- system.file("extdata/sapFinder_test_BioMart.Xref.txt",
+                          package="sapFinder")
> easyRun(vcf=vcf,annotation=annotation,refseq=refseq,
+         outdir="test",prefix="sapFinder_test",
+         spectra=mgf.path,cpu=0,tol=10, tolu="ppm",
+         itol=0.1,itolu="Daltons",xref=xref)
```

```
Stage 1. Variation-associated database construction.
Stage 2. MS/MS searching.
2016-05-03 22:22:51
Loading spectra
 (mgf). loaded.
Spectra matching criteria = 331
Starting threads . started.
Computing models:
        t
                sequences modelled = 0 ks
Model refinement:
Creating report:
        initial calculations  ..... done.
        sorting  ..... done.
        finding repeats ..... done.
        evaluating results ..... done.
        calculating expectations ..... done.
        writing results ..... done.
```

```
Valid models = 329
Unique models = 208
Estimated false positives = 3 +/- 2


Stage 3. Post-processing.
Stage 4. HTML-based report generation.
    Step 1: Reading the Info.
    Step 2: Spectrum plotting.
    Step 3: Creating the html pages.
```

After the analysis has completed, the file 'index.html' in the output directory can be opened in a web browser to access report generated.

8 Session Info

Here is the output of `sessionInfo`:

```
> toLatex(sessionInfo())
```

- R version 3.3.0 (2016-05-03), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Rcpp 0.12.4.5, XML 3.98-1.4, data.table 1.9.6, rTANDEM 1.12.0, sapFinder 1.10.0
- Loaded via a namespace (and not attached): BiocStyle 2.0.0, RColorBrewer 1.1-2, chron 2.3-47, colorspace 1.2-6, grid 3.3.0, gtable 0.2.0, munsell 0.4.3, pheatmap 1.0.8, plyr 1.8.3, scales 0.4.0, tools 3.3.0

References

[1] R Craig and R C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–7, Jun 2004. doi: 10.1093/bioinformatics/bth092.

[2] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res*, 39 (Database issue):D945–50, 2011.

[3] Frederic Fournier, Charles Joly Beauparlant, Rene Paradis, and Arnaud Droit. *rTANDEM: Encapsulates X!Tandem in R.*, 2013. R package version 1.2.0.

[4] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[5] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20(9):1297–303, 2010.

[6] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–58, 2003.

[7] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, 2001.

[8] G. M. Sheynkman, M. R. Shortreed, B. L. Frey, M. Scalf, and L. M. Smith. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J Proteome Res*, 13(1):228–40, 2014.