SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data

Hugo Varet[1,2], Jean-Yves Coppée[2] & Marie-Agnès Dillies[1,2]

[1]C3BI & [2]PF2 - Institut Pasteur, Paris, France

Contact: hugo.varet@pasteur.fr

Rencontres R francophones – Grenoble – June 25th, 2015

What is RNA-Seq?

DNA $\xrightarrow{\text{transcription}}$ RNA $\xrightarrow{\text{sequencing}}$ millions of reads $\xrightarrow{\text{mapping}}$ counts

Biostatistician point of view:

	sample1	sample2	...	sampleP
gene1	10	15	...	12
gene2	0	2	...	1
gene3	435	514	...	452
⋮	⋮	⋮	⋱	⋮
geneN	123	298	...	317

Main goal: find differentially expressed genes.

Background

Existing famous methods/R packages available on Bioconductor:

- DESeq2 (Love, 2014);
- edgeR (Robinson, 2009).

Why SARTools?

1. make systematic quality controls of the data;
2. prevent untrained users from misusing some functionalities of DESeq2 and edgeR;
3. keep track of all the parameters: reproducible research;
4. provide a report containing all the results of the analysis.

Structure of SARTools

R package containing:

- functions: data loading, quality control, differential analysis, data exportation...;
- a HTML vignette providing extensive help;
- toy example data files to test the pipeline;
- the skeleton of the final HTML report (rmarkdown file);
- two R script templates (DESeq2/edgeR).

Using SARTools: input files

Target file (design of the experiment):

```
label   files              condition
WT1     WT1.counts.txt     WT
WT2     WT2.counts.txt     WT
KO1     KO1.counts.txt     KO
KO2     KO2.counts.txt     KO
```

Count data files (one per sample):

```
gene1   23
gene2   355
gene3   0
:       :
geneN   3643
```

Using SARTools: R script templates

```
################################################################################
###                parameters: to be modified by the user                    ###
################################################################################
rm(list=ls())                                          # remove all the objects from the R session

workDir <- "C:/path/to/your/working/directory/"        # working directory for the R session

projectName <- "projectName"                           # name of the project
author <- "Your name"                                  # author of the statistical analysis/report

targetFile <- "target.txt"                             # path to the design/target file
rawDir <- "raw"                                        # path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",          # names of the features to be removed
                      "ambiguous", "no_feature",       # (specific HTSeq-count information and rRNA for example)
                      "not_aligned", "too_low_aQual")

varInt <- "group"                                      # factor of interest
condRef <- "WT"                                        # reference biological condition
batch <- NULL                                          # blocking factor: NULL (default) or "batch" for example

fitType <- "parametric"                                # mean-variance relationship: "parametric" (default) or "local"
cooksCutoff <- TRUE                                    # TRUE/FALSE to perform the outliers detection (default is TRUE)
independentFiltering <- TRUE                           # TRUE/FALSE to perform independent filtering (default is TRUE)
alpha <- 0.05                                          # threshold of statistical significance
pAdjustMethod <- "BH"                                  # p-value adjustment method: "BH" (default) or "BY"

typeTrans <- "VST"                                     # transformation for PCA/clustering: "VST" or "rlog"
locfunc <- "median"                                    # "median" (default) or "shorth" to estimate the size factors

colors <- c("dodgerblue","firebrick1",                 # vector of colors of each biological condition on the plots
            "MediumVioletRed","SpringGreen")
```

Output files: HTML report

Statistical report of project 2015-KOvsWT-RNASeq: pairwise comparison(s) of conditions with DESeq2

Author: Hugo Varet

Date: 2015-05-15

The SARTools R package which generated this report has been developped at PF2 - Institut Pasteur by M.-A. Dillies and H. Varet (hugo.varet@pasteur.fr). Thanks to cite H. Varet, J.-Y. Coppee and M.-A. Dillies, *SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-seq data*, 2015 (submitted) when using this tool for any analysis published.

Table of contents

1 Introduction

The analyses reported in this document are part of the 2015-KOvsWT-RNASeq project. The aim is to find features that are differentially expressed between WT and KO. The statistical analysis process includes data normalization, graphical exploration of raw and normalized data, test for differential expression for each feature between the conditions, raw p-value adjustment and export of lists of features having a significant differential expression between the conditions.

Output files: HTML report

6 R session information and parameters

The versions of the R software and Bioconductor packages used for this analysis are listed below. It is important to save them if one wants to re-perform the analysis in the same conditions.

- R version 3.2.0 (2015-04-16), x86_64-w64-mingw32
- Locale: LC_COLLATE=French_France.1252, LC_CTYPE=French_France.1252, LC_MONETARY=French_France.1252, LC_NUMERIC=C, LC_TIME=French_France.1252
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.14.0, DESeq2 1.8.1, edgeR 3.10.0, GenomeInfoDb 1.4.0, GenomicRanges 1.20.3, IRanges 2.2.1, limma 3.24.3, Rcpp 0.11.6, RcppArmadillo 0.5.100.1.0, S4Vectors 0.6.0, SARTools 1.1.0, xtable 1.7-4
- Loaded via a namespace (and not attached): acepack 1.3-3.3, annotate 1.46.0, AnnotationDbi 1.30.1, Biobase 2.28.0, BiocParallel 1.2.1, cluster 2.0.1, codetools 0.2-11, colorspace 1.2-6, DBI 0.3.1, digest 0.6.8, evaluate 0.7, foreign 0.8-63, formatR 1.2, Formula 1.2-1, futile.logger 1.4.1, futile.options 1.0.0, genefilter 1.50.0, geneplotter 1.46.0, ggplot2 1.0.1, grid 3.2.0, gridExtra 0.9.1, gtable 0.1.2, Hmisc 3.16-0, knitr 1.10.5, lambda.r 1.1.7, lattice 0.20-31, latticeExtra 0.6-26, locfit 1.5-9.1, magrittr 1.5, MASS 7.3-40, munsell 0.4.2, nnet 7.3-9, plyr 1.8.2, proto 0.3-10, RColorBrewer 1.1-2, reshape2 1.4.1, rpart 4.1-9, RSQLite 1.0.0, scales 0.2.4, splines 3.2.0, stringi 0.4-1, stringr 1.0.0, survival 2.38-1, tools 3.2.0, XML 3.98-1.1, XVector 0.8.0

Parameter values used for this analysis are:

- workDir: C:/Users/hvaret/Desktop/slides_r2015/demo_SARTools/
- projectName: 2015-KOvsWT-RNASeq
- author: Hugo Varet
- targetFile: C:/Users/hvaret/Documents/R/win-library/3.2/SARTools/target.txt
- rawDir: C:/Users/hvaret/Documents/R/win-library/3.2/SARTools/raw
- featuresToRemove: alignment_not_unique, ambiguous, no_feature, not_aligned, too_low_aQual
- varInt: group
- condRef: WT
- batch: NULL
- fitType: parametric
- cooksCutoff: TRUE
- independentFiltering: TRUE
- alpha: 0.05
- pAdjustMethod: BH
- typeTrans: VST
- locfunc: median
- colors: dodgerblue, firebrick1, MediumVioletRed, SpringGreen

Output files: list of DE genes

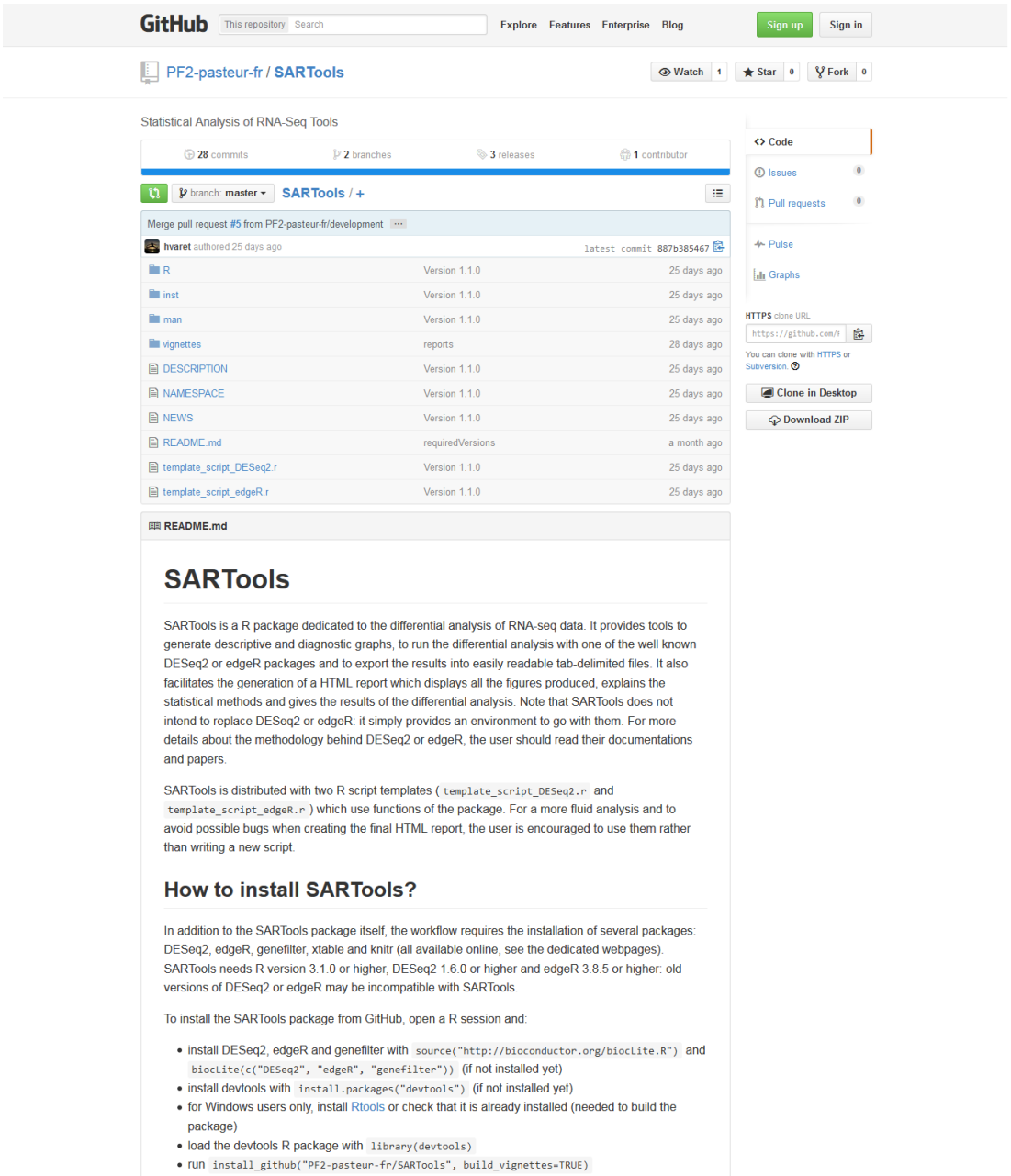
Three tab-delimited text files per comparison:

- `*.complete.txt`: all the genes;
- `*.up.txt`: up-regulated genes sorted by adj. p-value;
- `*.down.txt`: down-regulated genes sorted by adj. p-value;

Columns: gene identifier, $\log_2$(Fold-Change), adjusted p-value...

$\rightarrow$ Can be used as starting point for Gene Ontology terms studies.

Availability: github.com/PF2-pasteur-fr/SARTools/



HTML vignette

- use of markdown: lighter than LaTeX;
- better integration in GitHub;

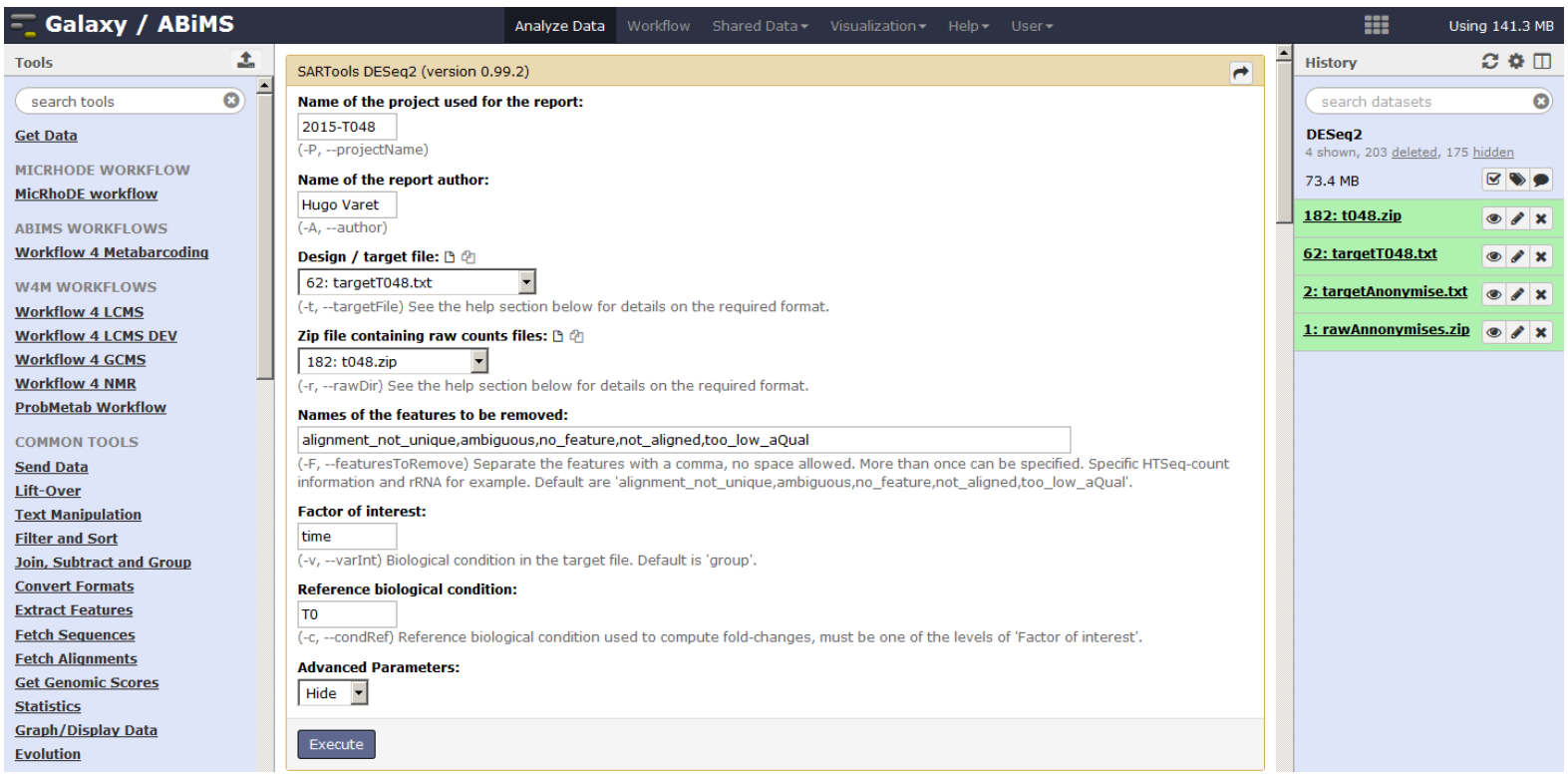
- provides extensive help on the use of SARTools:
  - installation;
  - input files required;
  - definition of the parameters;
  - troubleshooting cases: detection of potential sequencing/technical problems such as inversion of samples, batch effects, outliers...

Other ways of running SARTools

Command line with a dedicated script which requires optparse:

```
Rscript template_script_DESeq2_CL.r --help
```

Galaxy web-based platform (by Loraine Guéguen from SB-Roscoff):



Conclusion

SARTools...

- facilitates the use of DESeq2 and edgeR for untrained R users/biologists;
- performs quality controls and helps to detect potential problems;
- fits the requirements of the reproducible research.

Thanks for your attention!