

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

Hugo Varet<sup>1,2</sup>, Jean-Yves Coppée<sup>2</sup> & Marie-Agnès Dillies<sup>1,2</sup>

<sup>1</sup>C3BI & <sup>2</sup>PF2 - Institut Pasteur, Paris, France

Contact: [hugo.varet@pasteur.fr](mailto:hugo.varet@pasteur.fr)

Rencontres R francophones – Grenoble – June 25<sup>th</sup>, 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
K01	K01.counts.txt	KO
K02	K02.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")
```

## 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 developed at PF2 - Institut Pasteur by M.-A. Dillies and H. Varet ([hugo.varet@pasteur.fr](mailto: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
2. Description of raw data
3. Variability within the experiment: data exploration
4. Normalization
5. Differential analysis
6. R session information and parameters
7. Bibliography

---

## 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, GenomInfoDb 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...

→ Can be used as starting point for [Gene Ontology](#) terms studies.

GitHub  [Explore](#) [Features](#) [Enterprise](#) [Blog](#) [Sign up](#) [Sign in](#)

PF2-pasteur-fr / SARTools [Watch](#) [★ Star](#) [V Fork](#)

Statistical Analysis of RNA-Seq Tools

🔍 [Commits](#) [Branches](#) [Releases](#) [1 contributor](#)

🔍 [Search: master](#) SARTools / +

Merge pull request #9 from PF2-pasteur-fr/development (3)

📁 R	Version 1.1.0	25 days ago
📁 inst	Version 1.1.0	25 days ago
📁 man	Version 1.1.0	25 days ago
📁 vignettes	reports	25 days ago
📄 DESCRIPTION	Version 1.1.0	25 days ago
📄 NAMESPACE	Version 1.1.0	25 days ago
📄 NEWS	Version 1.1.0	25 days ago
📄 README.md	requiredVersions	a month ago
📄 template_script_DESeq2.r	Version 1.1.0	25 days ago
📄 template_script_edgeR.r	Version 1.1.0	25 days ago

📄 README.md

## SARTools

SARTools is a R package dedicated to the differential analysis of RNA-seq data. It provides tools to generate descriptive and diagnostic graphs, to run the differential analysis with one of the well known DESeq2 or edgeR packages and to export the results into easily readable tab-delimited files. It also facilitates the generation of a HTML report which displays all the figures produced, explains the statistical methods and gives the results of the differential analysis. Note that SARTools does not intend to replace DESeq2 or edgeR: it simply provides an environment to go with them. For more details about the methodology behind DESeq2 or edgeR, the user should read their documentations and papers.

SARTools is distributed with two R script templates (`template_script_DESeq2.r` and `template_script_edgeR.r`) which use functions of the package. For a more fluid analysis and to avoid possible bugs when creating the final HTML report, the user is encouraged to use them rather than writing a new script.

### How to install SARTools?

In addition to the SARTools package itself, the workflow requires the installation of several packages: DESeq2, edgeR, genefilter, stable and knitr (all available online, see the dedicated webpages). SARTools needs R version 3.1.0 or higher, DESeq2 1.6.0 or higher and edgeR 3.8.5 or higher: old versions of DESeq2 or edgeR may be incompatible with SARTools.

To install the SARTools package from GitHub, open a R session and:

- install DESeq2, edgeR and genefilter with `source("http://bioconductor.org/biocLite.R")` and `biocLite(c("DESeq2", "edgeR", "genefilter"))` (if not installed yet)
- install devtools with `install.packages("devtools")` (if not installed yet)
- for Windows users only, install Rtools or check that it is already installed (needed to build the package)
- load the devtools R package with `library(devtools)`
- run `install_github("PF2-pasteur-fr/SARTools", build_vignettes=FALSE)`

🔍 [Code](#) [Issues](#) [Pull requests](#) [+ Pulse](#) [Graphs](#)

HTTPS clone URL  
`https://github.com/PF2-pasteur-fr/SARTools`  
You can also clone with HTTP or Subversion

[Clone in Desktop](#) [Download ZIP](#)

# HTML vignette

- ▶ use of markdown: lighter than  $\text{\LaTeX}$ ;
- ▶ 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):

The screenshot displays the Galaxy web-based platform interface for running the SARTools DESeq2 tool. The main panel shows the tool configuration form with the following fields:

- Name of the project used for the report:** 2015-T048 (with a subtext: (-P, --projectName))
- Name of the report author:** Hugo Varet (with a subtext: (-A, --author))
- Design / target file:** 62: targetT048.txt (with a subtext: (-t, --targetFile) See the help section below for details on the required format.)
- Zip file containing raw counts files:** 182: t048.zip (with a subtext: (-r, --rawDir) See the help section below for details on the required format.)
- Names of the features to be removed:** alignment\_not\_unique,ambiguous\_no\_feature\_not\_aligned,too\_low\_aQual (with a subtext: (-F, --featuresToRemove) Separate the features with a comma, no space allowed. More than once can be specified. Specific: HTSeq-count information and rRNA for example. Default: are 'alignment\_not\_unique,ambiguous\_no\_feature\_not\_aligned,too\_low\_aQual'.)
- Factor of interest:** time (with a subtext: (-v, --varInt) Biological condition in the target file. Default is 'group'.)
- Reference biological condition:** T0 (with a subtext: (-C, --condRef) Reference biological condition used to compute fold-changes, must be one of the levels of 'Factor of interest'.)
- Advanced Parameters:** Hide

An **Execute** button is located at the bottom of the configuration form. On the right side, a **History** panel shows a list of datasets:

- DESeq2 (4 shown, 203 deleted, 175 hidden, 73.4 MB)
- 182: t048.zip
- 62: targetT048.txt
- 2: targetAnonymise.txt
- 1: rawAnonymises.zip

The left sidebar contains a **Tools** menu with various categories like **Get Data**, **MICRHODE WORKFLOW**, **ABIMS WORKFLOWS**, **W4H WORKFLOWS**, **COMMON TOOLS**, and **Send Data**.

# 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!