NetTab 2016 - Roma

Ensuring reproducibility and portability of NGS analysis workflows

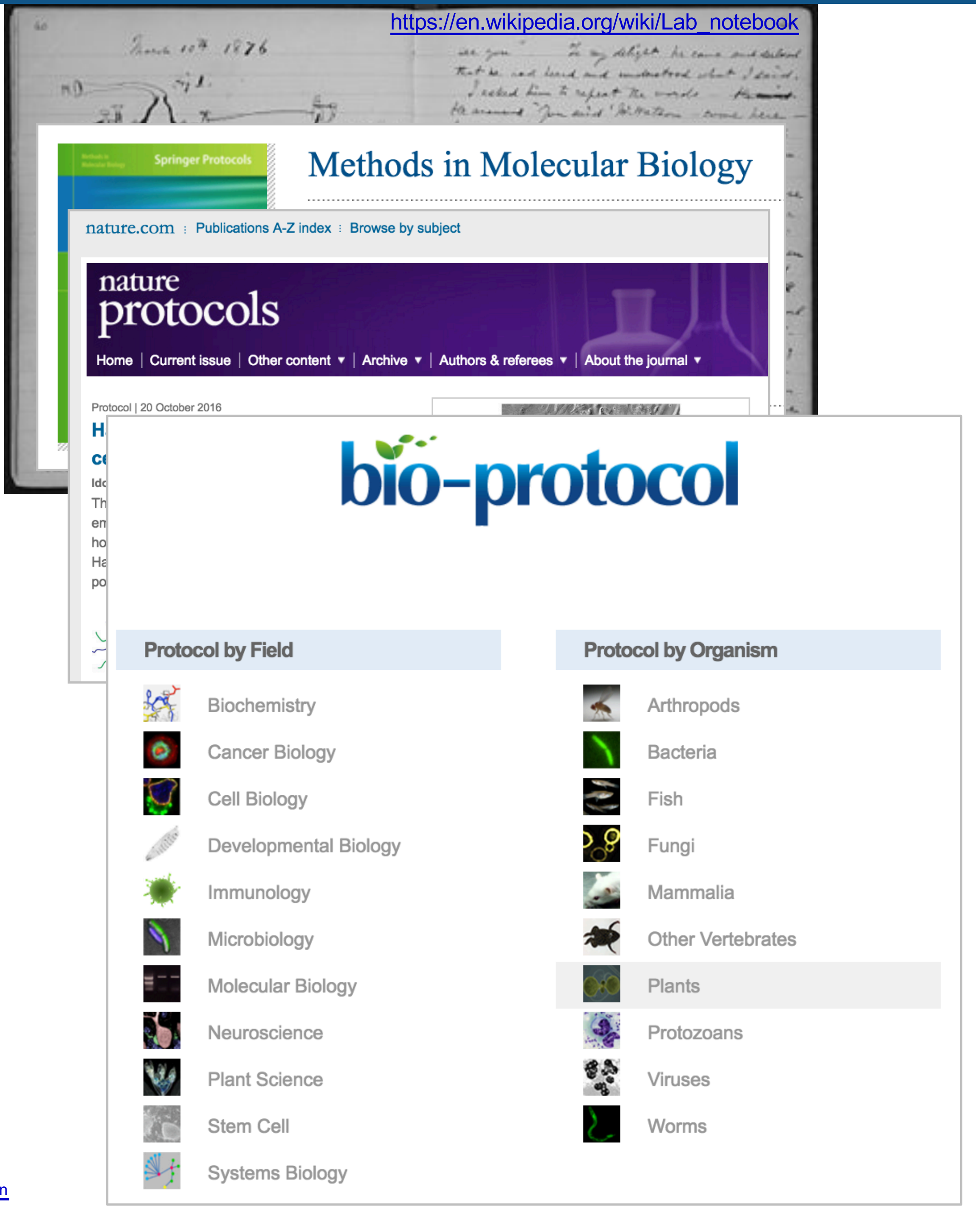
Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univ-amu.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)

How do experimental biologists ensure reproducibility?

- Materials and methods section in publications.
- Lab notebook
- Annotated protocols published in specialized journals.
  - Methods in Molecular Biology
  - Nature Protocols
- A structured resource for protocols http://www.bio-protocol.org/
  - Example: [Bio101] A General EMSA (Gel-shift) Protocol (http://www.bio-protocol.org/e24).
- Standard lab equipment (tools).
- Standard lab environment (infrastructure).
- Sharing mutant strains (Morgan, ~ 1915).
- Genetic strains libraries (Nikolai Vavilov).



https://en.wikipedia.org/wiki/Lab_notebook

Methods in Molecular Biology

nature.com : Publications A-Z index : Browse by subject

nature protocols

Home | Current issue | Other content ▾ | Archive ▾ | Authors & referees ▾ | About the journal ▾

Protocol | 20 October 2016

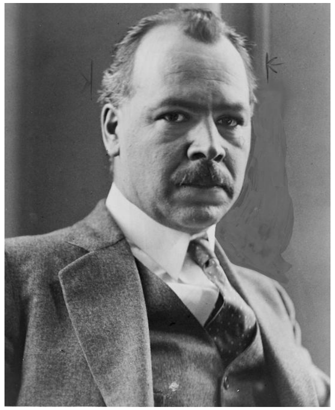
bio-protocol

Protocol by Field

- Biochemistry
- Cancer Biology
- Cell Biology
- Developmental Biology
- Immunology
- Microbiology
- Molecular Biology
- Neuroscience
- Plant Science
- Stem Cell
- Systems Biology

Protocol by Organism

- Arthropods
- Bacteria
- Fish
- Fungi
- Mammalia
- Other Vertebrates
- Plants
- Protozoans
- Viruses
- Worms



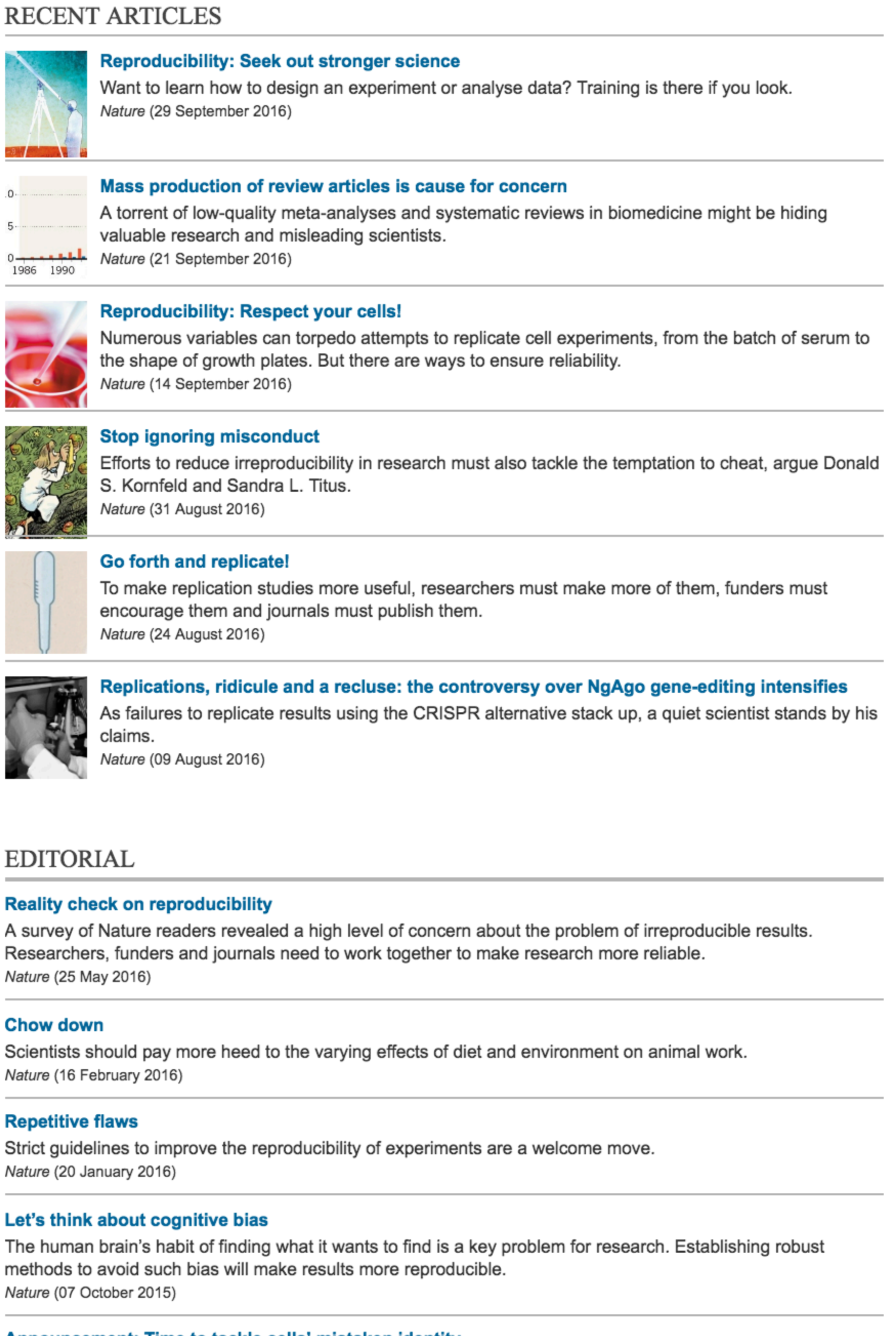
https://en.wikipedia.org/wiki/Nikolai_Vavilov

https://en.wikipedia.org/wiki/Thomas_Hunt_Morgan

Reproducing experiments - in practice

- What if you cannot reproduce an experimental technique?
- Try, fail, retry, re-fail, re-re-try, cry, …
- Check your reactants.
- Check your strains.
- Check anything you can think of.
- Search help in your neighborhood.
- Contact the authors of the method.
    - Sometimes they can provide some forgotten detail in the method.
- Ask to spend some weeks/months in the lab which developed or which masters the technique.
- Search another method.
- Search another research project.
- Stop experimental biology and become a bioinformatician, with the hope that computer procedures ensure reproducibility.

http://www.nature.com/news/reproducibility-1.17552



RECENT ARTICLES

Reproducibility: Seek out stronger science
Want to learn how to design an experiment or analyse data? Training is there if you look.
Nature (29 September 2016)

Mass production of review articles is cause for concern
A torrent of low-quality meta-analyses and systematic reviews in biomedicine might be hiding valuable research and misleading scientists.
Nature (21 September 2016)

Reproducibility: Respect your cells!
Numerous variables can torpedo attempts to replicate cell experiments, from the batch of serum to the shape of growth plates. But there are ways to ensure reliability.
Nature (14 September 2016)

Stop ignoring misconduct
Efforts to reduce irreproducibility in research must also tackle the temptation to cheat, argue Donald S. Kornfeld and Sandra L. Titus.
Nature (31 August 2016)

Go forth and replicate!
To make replication studies more useful, researchers must make more of them, funders must encourage them and journals must publish them.
Nature (24 August 2016)

Replications, ridicule and a recluse: the controversy over NgAgo gene-editing intensifies
As failures to replicate results using the CRISPR alternative stack up, a quiet scientist stands by his claims.
Nature (09 August 2016)

EDITORIAL

Reality check on reproducibility
A survey of Nature readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.
Nature (25 May 2016)

Chow down
Scientists should pay more heed to the varying effects of diet and environment on animal work.
Nature (16 February 2016)

Repetitive flaws
Strict guidelines to improve the reproducibility of experiments are a welcome move.
Nature (20 January 2016)

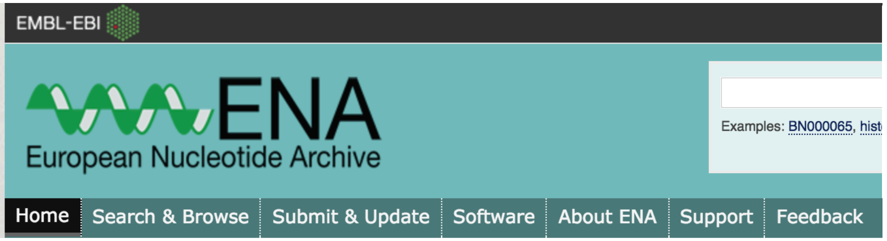
Let's think about cognitive bias
The human brain's habit of finding what it wants to find is a key problem for research. Establishing robust methods to avoid such bias will make results more reproducible.
Nature (07 October 2015)

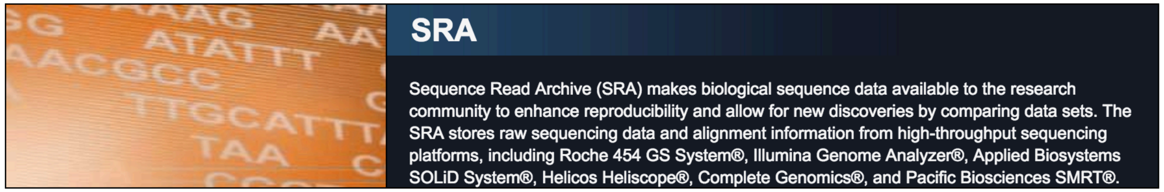
Reproducibility of bioinformatics results

- In contrast with biological experiments, which may require personal touch and skills, bioinformatics analyses are fully performed by computers.

- An example: Next Generation Sequencing Analysis
  - Any publication involving NGS data **implies** for the raw sequences (short reads) to be available in specialized databases.
    - Short Read Archive (SRA): http://www.ebi.ac.uk/ena
    - European Nucleotide Archive (ENA): https://www.ncbi.nlm.nih.gov/sra

- In principle, any bioinformatics analysis should thus be 100% reproducible.
- Is this the case?





Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

How many NGS published results can be reproduced from the raw sequences?

- Anyone has a clue? Probably not much.
- Why?
  - Mobility of active people (PhD, postdocs).
  - Bad practices are socially encouraged (publish or perish, not time to document).

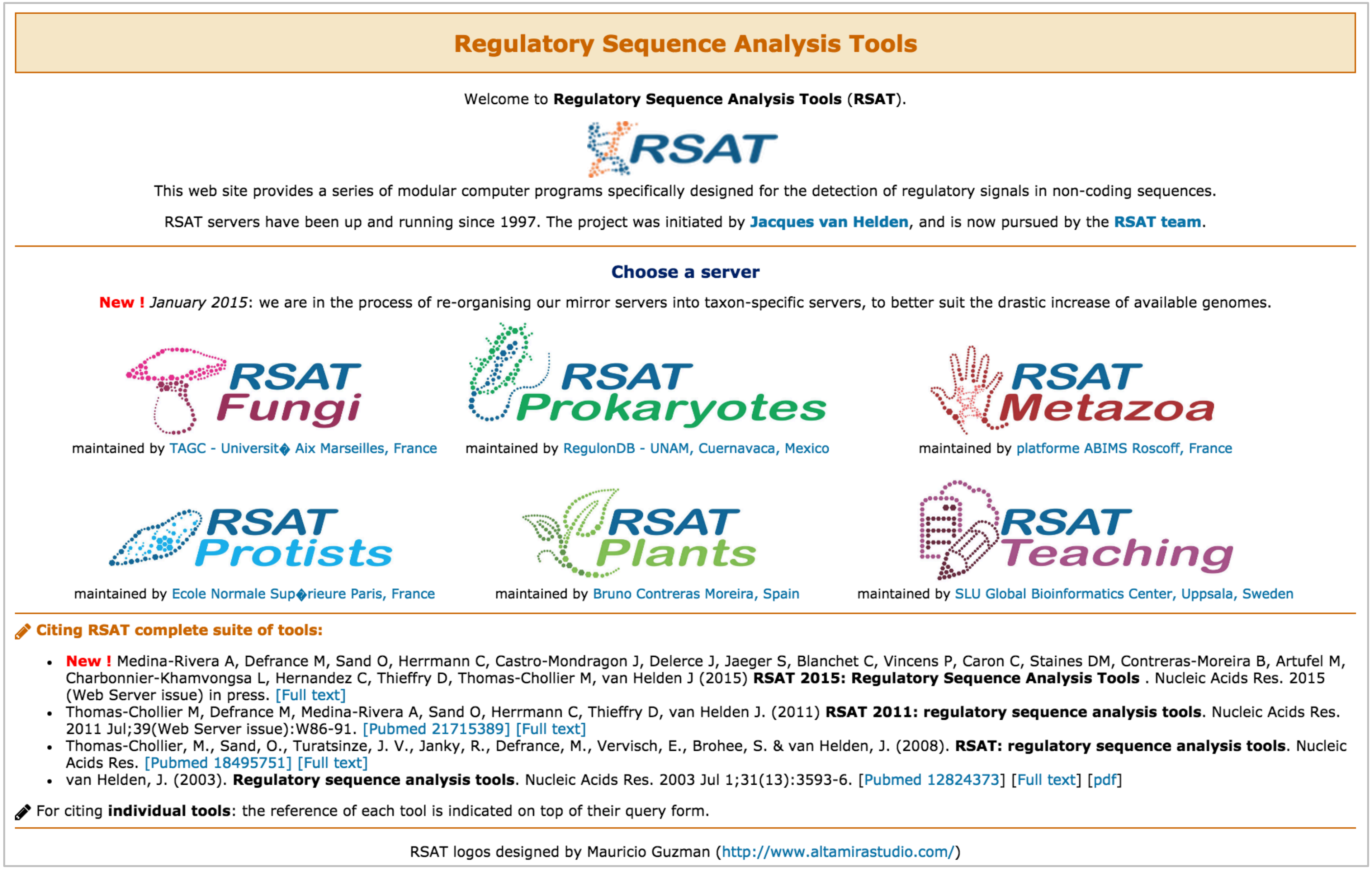
Some requirements and ingredients

- Code availability, portability, usability, evolvability, tractability.
  - Kleenex scripts vs bioinformatics tools.
- Backward compatibility or traceback capability
  - Run the analysis with the tools available when paper was published
- No hand-treatment of the Figures and tables.
  - Use notebooks (e.g. iPython, Sweave, Rmd, …)
  - Link the manuscript to the results. Ideally, all values in the manuscript should be injected from the analisis itself (Sweave, Rmd, …).
- Separation between data, scripts and results. Brave test at the end of the project (just before submission): rm –rf results, make all
  - Should regenerate everything from the

Development of bioinformatics tools

My standpoint

- Developing bioinformatics tools since 1997.
- Regulatory Sequence Analysis Tools (RSAT): http://rsat.eu/



Regulatory Sequence Analysis Tools

Welcome to **Regulatory Sequence Analysis Tools** (**RSAT**).

This web site provides a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences.

RSAT servers have been up and running since 1997. The project was initiated by **Jacques van Helden**, and is now pursued by the **RSAT team**.

Choose a server

New ! *January 2015*: we are in the process of re-organising our mirror servers into taxon-specific servers, to better suit the drastic increase of available genomes.

maintained by TAGC - Universit� Aix Marseilles, France

maintained by RegulonDB - UNAM, Cuernavaca, Mexico

maintained by platforme ABIMS Roscoff, France

maintained by Ecole Normale Sup�rieure Paris, France

maintained by Bruno Contreras Moreira, Spain

maintained by SLU Global Bioinformatics Center, Uppsala, Sweden

Citing RSAT complete suite of tools:

- **New !** Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon J, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) **RSAT 2015: Regulatory Sequence Analysis Tools** . Nucleic Acids Res. 2015 (Web Server issue) in press. [Full text]
- Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. (2011) **RSAT 2011: regulatory sequence analysis tools**. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W86-91. [Pubmed 21715389] [Full text]
- Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). **RSAT: regulatory sequence analysis tools**. Nucleic Acids Res. [Pubmed 18495751] [Full text]
- van Helden, J. (2003). **Regulatory sequence analysis tools**. Nucleic Acids Res. 2003 Jul 1;31(13):3593-6. [Pubmed 12824373] [Full text] [pdf]

For citing **individual tools**: the reference of each tool is indicated on top of their query form.

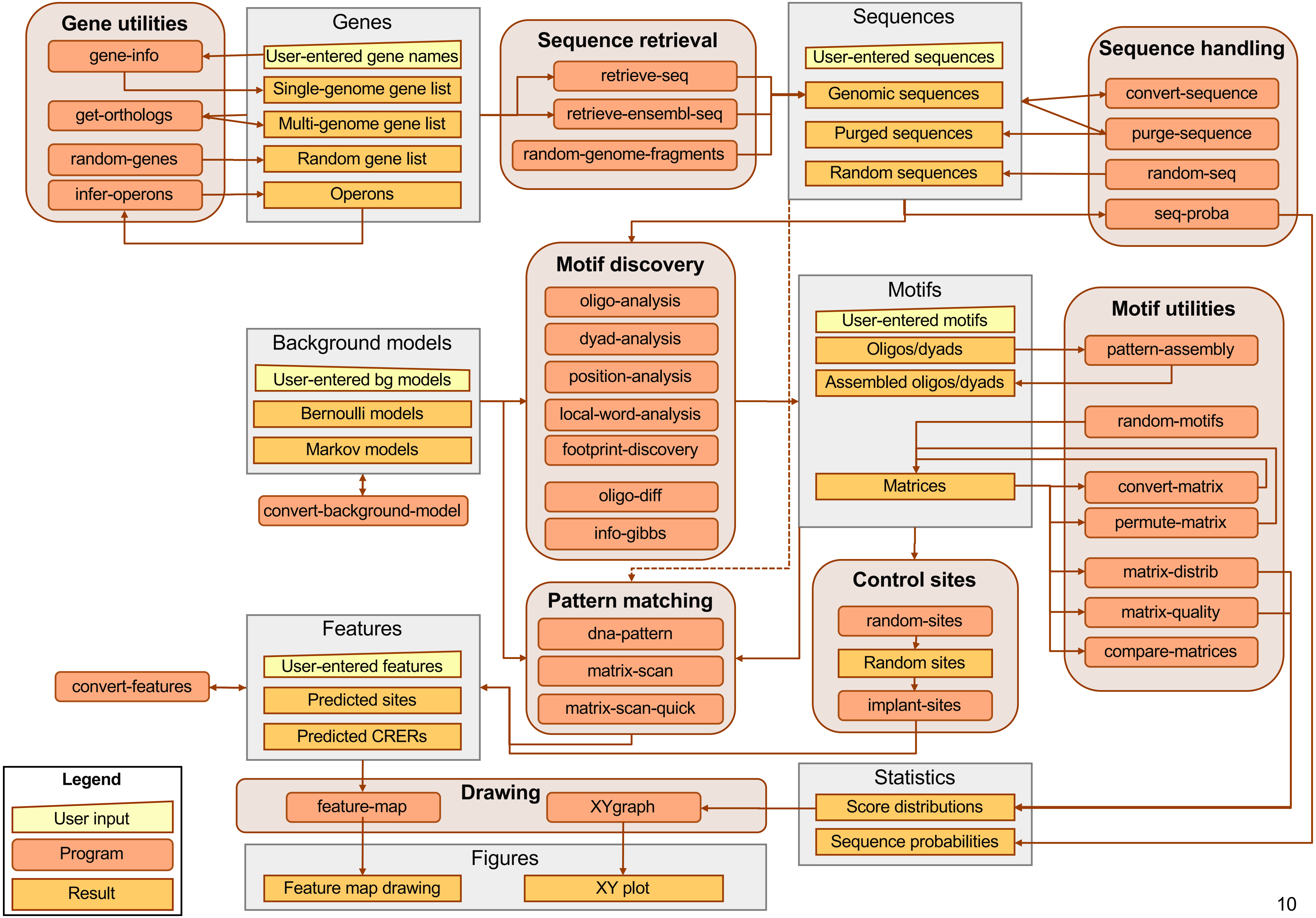
RSAT logos designed by Mauricio Guzman (http://www.altamirastudio.com/)

Accessibility

- Multiple ways to use the tools
  - 6 servers world-wide
  - Download the tar archive and install locally
  - Download a VirtualBox Virtual Machine.
  - Use a Virtual Machine on the cloud of the Institut Français de Bioinformatique (IFB).
  - In progress: run a Docker container.
  - Remote queries via SOAP/WSDL Web services.
- Each access mode has advantages and disadvantages

- Demo
  - Instantiate a VM on the IFB cloud.
    - https://cloud.france-bioinformatique.fr/cloud/instance/
  - Run the DEMO of some tools
  - Start peak-motifs demo (should be finished before my talk ends).

Regulatory Sequence Analysis Tools (RSAT) 2011/



Gene utilities

gene-info

get-orthologs

random-genes

infer-operons

Genes

User-entered gene names

Single-genome gene list

Multi-genome gene list

Random gene list

Operons

Sequence retrieval

retrieve-seq

retrieve-ensembl-seq

random-genome-fragments

Sequences

User-entered sequences

Genomic sequences

Purged sequences

Random sequences

Sequence handling

convert-sequence

purge-sequence

random-seq

seq-proba

Motif discovery

oligo-analysis

dyad-analysis

position-analysis

local-word-analysis

footprint-discovery

oligo-diff

info-gibbs

Background models

User-entered bg models

Bernoulli models

Markov models

convert-background-model

Motifs

User-entered motifs

Oligos/dyads

Assembled oligos/dyads

Matrices

Motif utilities

pattern-assembly

random-motifs

convert-matrix

permute-matrix

matrix-distrib

matrix-quality

compare-matrices

Pattern matching

dna-pattern

matrix-scan

matrix-scan-quick

Control sites

random-sites

Random sites

implant-sites

Features

User-entered features

Predicted sites

Predicted CRERs

convert-features

Drawing

feature-map

XYgraph

Figures

Feature map drawing

XY plot

Statistics

Score distributions

Sequence probabilities

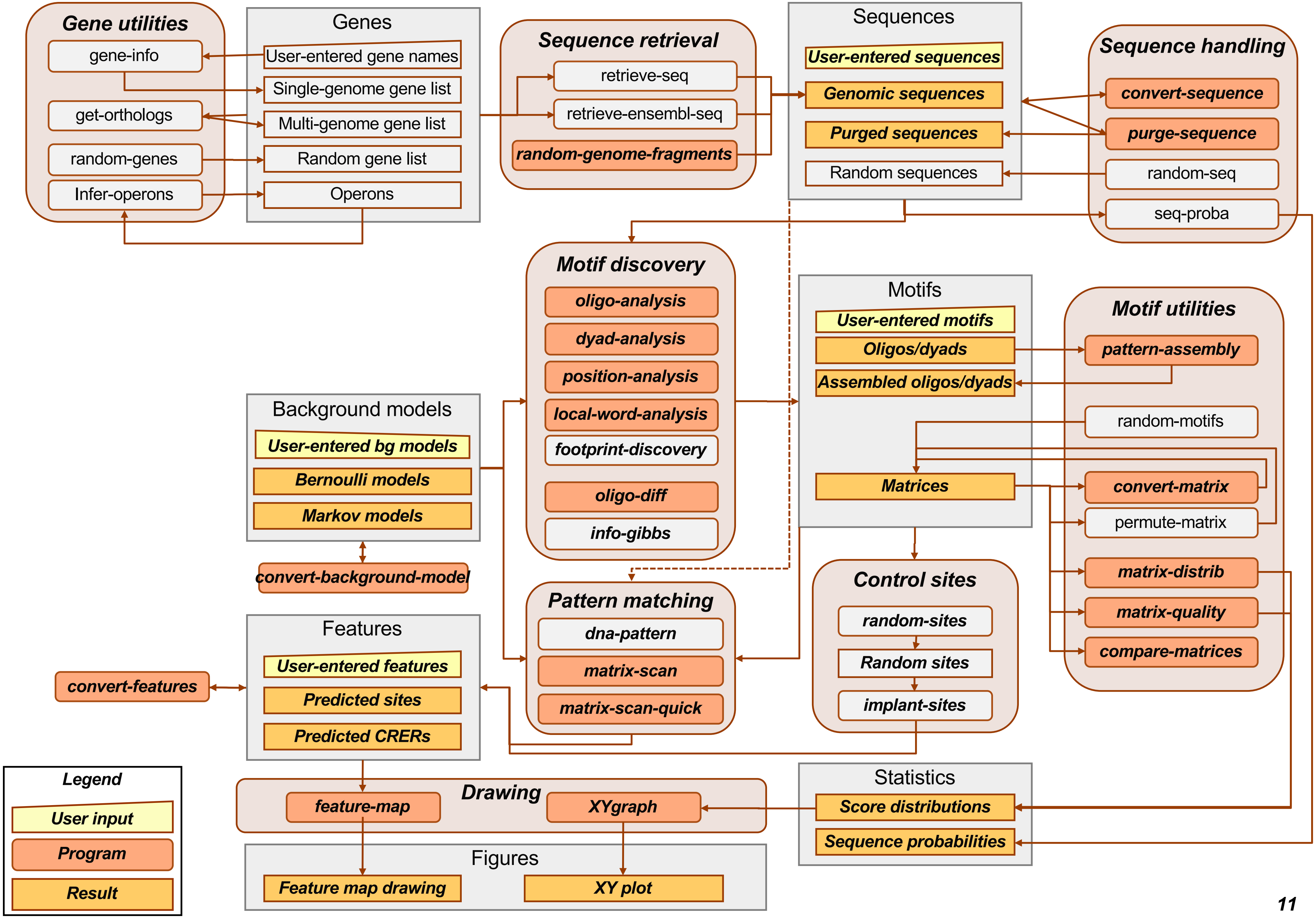
Legend

User input

Program

Result

Peak-motifs: discovering motifs in full-sized peak-sets



Gene utilities
gene-info
get-orthologs
random-genes
Infer-operons

Genes
User-entered gene names
Single-genome gene list
Multi-genome gene list
Random gene list
Operons

Sequence retrieval
retrieve-seq
retrieve-ensembl-seq
random-genome-fragments

Sequences
User-entered sequences
Genomic sequences
Purged sequences
Random sequences

Sequence handling
convert-sequence
purge-sequence
random-seq
seq-proba

Motif discovery
oligo-analysis
dyad-analysis
position-analysis
local-word-analysis
footprint-discovery
oligo-diff
info-gibbs

Motifs
User-entered motifs
Oligos/dyads
Assembled oligos/dyads
Matrices

Motif utilities
pattern-assembly
random-motifs
convert-matrix
permute-matrix
matrix-distrib
matrix-quality
compare-matrices

Background models
User-entered bg models
Bernoulli models
Markov models

convert-background-model

Pattern matching
dna-pattern
matrix-scan
matrix-scan-quick

Control sites
random-sites
Random sites
implant-sites

Features
User-entered features
Predicted sites
Predicted CRERs

convert-features

Drawing
feature-map
XYgraph

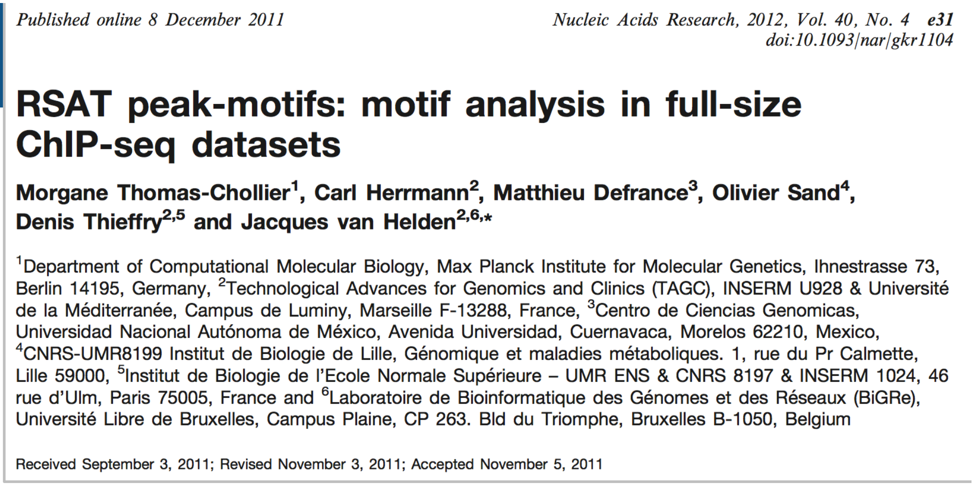
Figures
Feature map drawing
XY plot

Statistics
Score distributions
Sequence probabilities

Legend
User input
Program
Result

Tool documentation

- Each tool has a Web help page.
- Option –help for the command line use.
- Tutorials on the Web.
- Published description of methods/tools with their evaluation (e.g. NAR article about peak-motifs).
- Publication about the software suite as a whole (NAR Web software issue 2003, 2008, 2011, 2015).
- Published protocols (Nature Protocols, MiMB).



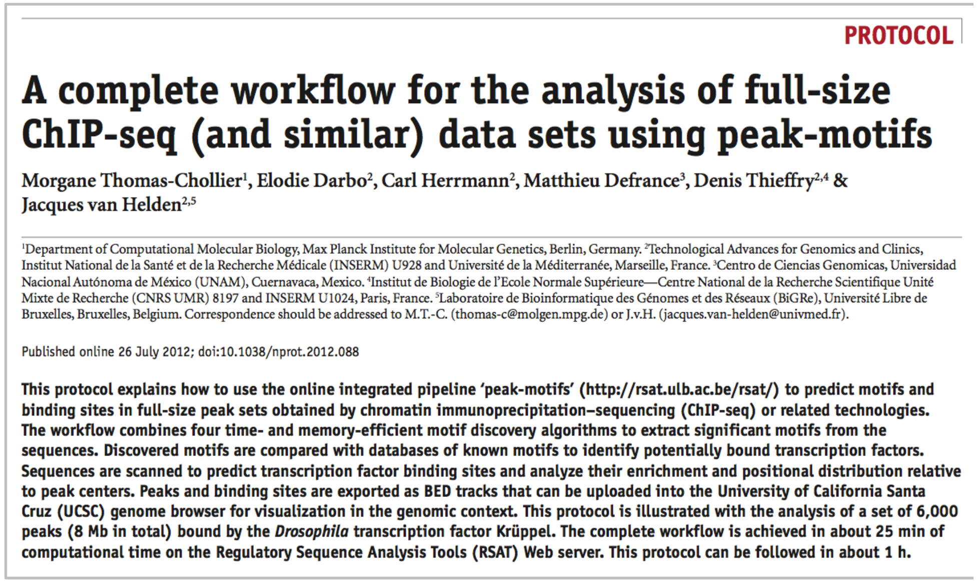
Published online 8 December 2011 Nucleic Acids Research, 2012, Vol. 40, No. 4 **e31**
doi:10.1093/nar/gkr1104

RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets

Morgane Thomas-Chollier[1], Carl Herrmann[2], Matthieu Defrance[3], Olivier Sand[4], Denis Thieffry[2,5] and Jacques van Helden[2,6,*]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, Berlin 14195, Germany, [2]Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée, Campus de Luminy, Marseille F-13288, France, [3]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México, Avenida Universidad, Cuernavaca, Morelos 62210, Mexico, [4]CNRS-UMR8199 Institut de Biologie de Lille, Génomique et maladies métaboliques. 1, rue du Pr Calmette, Lille 59000, [5]Institut de Biologie de l'Ecole Normale Supérieure – UMR ENS & CNRS 8197 & INSERM 1024, 46 rue d'Ulm, Paris 75005, France and [6]Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe, Bruxelles B-1050, Belgium

Received September 3, 2011; Revised November 3, 2011; Accepted November 5, 2011



PROTOCOL

A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs

Morgane Thomas-Chollier[1], Elodie Darbo[2], Carl Herrmann[2], Matthieu Defrance[3], Denis Thieffry[2,4] & Jacques van Helden[2,5]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Technological Advances for Genomics and Clinics, Institut National de la Santé et de la Recherche Médicale (INSERM) U928 and Université de la Méditerranée, Marseille, France. [3]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México (UNAM), Cuernavaca, Mexico. [4]Institut de Biologie de l'Ecole Normale Supérieure—Centre National de la Recherche Scientifique Unité Mixte de Recherche (CNRS UMR) 8197 and INSERM U1024, Paris, France. [5]Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, Bruxelles, Belgium. Correspondence should be addressed to M.T.-C. (thomas-c@molgen.mpg.de) or J.v.H. (jacques.van-helden@univmed.fr).

Published online 26 July 2012; doi:10.1038/nprot.2012.088

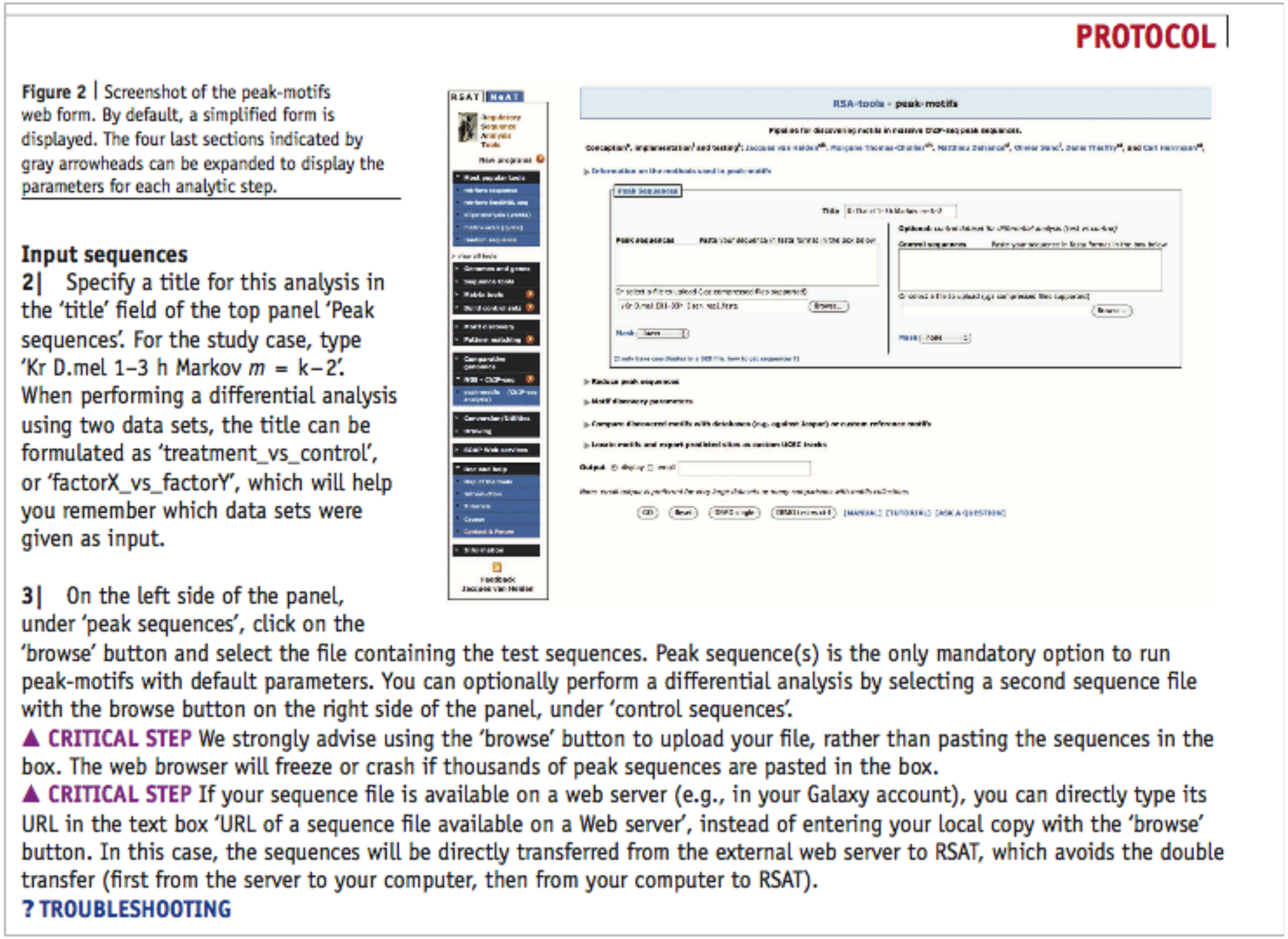
This protocol explains how to use the online integrated pipeline 'peak-motifs' (http://rsat.ulb.ac.be/rsat/) to predict motifs and binding sites in full-size peak sets obtained by chromatin immunoprecipitation–sequencing (ChIP-seq) or related technologies. The workflow combines four time- and memory-efficient motif discovery algorithms to extract significant motifs from the sequences. Discovered motifs are compared with databases of known motifs to identify potentially bound transcription factors. Sequences are scanned to predict transcription factor binding sites and analyze their enrichment and positional distribution relative to peak centers. Peaks and binding sites are exported as BED tracks that can be uploaded into the University of California Santa Cruz (UCSC) genome browser for visualization in the genomic context. This protocol is illustrated with the analysis of a set of 6,000 peaks (8 Mb in total) bound by the *Drosophila* transcription factor Krüppel. The complete workflow is achieved in about 25 min of computational time on the Regulatory Sequence Analysis Tools (RSAT) Web server. This protocol can be followed in about 1 h.

- **Selected recent publications with NGS analysis tools, evaluations, protocols and applications**
- Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*, **40**, e31–e31.
- Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols*, **7**, 1551–1568.
- Medina-Rivera,A., Defrance,M., Sand,O., Herrmann,C., Castro-Mondragon,J.A., Delerce,J., Jaeger,S., Blanchet,C., Vincens,P., Caron,C., *et al.* (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res*, **43**, W50–6.
- Castro-Mondragon,J.A., Rioualen,C., Contreras-Moreira,B. and van Helden,J. (2016) RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes. *Methods Mol Biol*, **1482**, 297–322.
- Cayrou,C., Ballester,B., Peiffer,I., Fenouil,R., Coulombe,P., Andrau,J.-C., van Helden,J. and Méchali,M. (2015) The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res*, **25**, 1873–1885.

Published protocols

More and more protocols

- Protocols are documented
  - Study cases
  - Critical steps
  - Caution notes
  - Troubleshooting
  - Illustrations of expected results
  - Interpretation of the results



PROTOCOL

Figure 2 | Screenshot of the peak-motifs web form. By default, a simplified form is displayed. The four last sections indicated by gray arrowheads can be expanded to display the parameters for each analytic step.

RSAT
NeAT
Regulatory Sequence Analysis Tools
RSA-tools - peak-motifs
Pipeline for discovering motifs in massive ChIP-seq peak sequences.
Peak Sequences
Title
Peak sequences
Control sequences
Browse...
Reduce peak sequences
Motif discovery parameters
Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs
Locate motifs and export predicted sites as custom UCSC tracks
Output
GO
Reset
[MANUAL] [TUTORIAL] [ASK A QUESTION]
Feedback
Jacques van Helden

Input sequences

2| Specify a title for this analysis in the 'title' field of the top panel 'Peak sequences'. For the study case, type 'Kr D.mel 1–3 h Markov *m* = k–2'. When performing a differential analysis using two data sets, the title can be formulated as 'treatment_vs_control', or 'factorX_vs_factorY', which will help you remember which data sets were given as input.

3| On the left side of the panel, under 'peak sequences', click on the 'browse' button and select the file containing the test sequences. Peak sequence(s) is the only mandatory option to run peak-motifs with default parameters. You can optionally perform a differential analysis by selecting a second sequence file with the browse button on the right side of the panel, under 'control sequences'.

▲ **CRITICAL STEP** We strongly advise using the 'browse' button to upload your file, rather than pasting the sequences in the box. The web browser will freeze or crash if thousands of peak sequences are pasted in the box.

▲ **CRITICAL STEP** If your sequence file is available on a web server (e.g., in your Galaxy account), you can directly type its URL in the text box 'URL of a sequence file available on a Web server', instead of entering your local copy with the 'browse' button. In this case, the sequences will be directly transferred from the external web server to RSAT, which avoids the double transfer (first from the server to your computer, then from your computer to RSAT).

? TROUBLESHOOTING

- Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size

Back to the past !

- Versioning systems

A workflow to study gene regulation

By Claire Rioualen & Jacques van Helden

Installing (or not) and running workflows

- Code and doc available at github
  - https://github.com/rioualen/gene-regulation/
- Environment
  - Install on your own computer.
  - Install a local VirtualBox virtual machine.
  - Use existing VM on the cloud of the Institut Français de Bioinformatique (IFB cloud).
  - Run as a Docker container.
- Tutorials
  - Demonstrator: mini-example with ChIP-seq from the budding yeast *Saccharomyces cerevisiae*.
  - Takes 10 minutes to run on a 2-CPU VM at the IFB cloud.
- Documentation
  - Installation tutorials.
  - Documentation of the snakemake libraries.

Using gene-regualtion at the IFB cloud

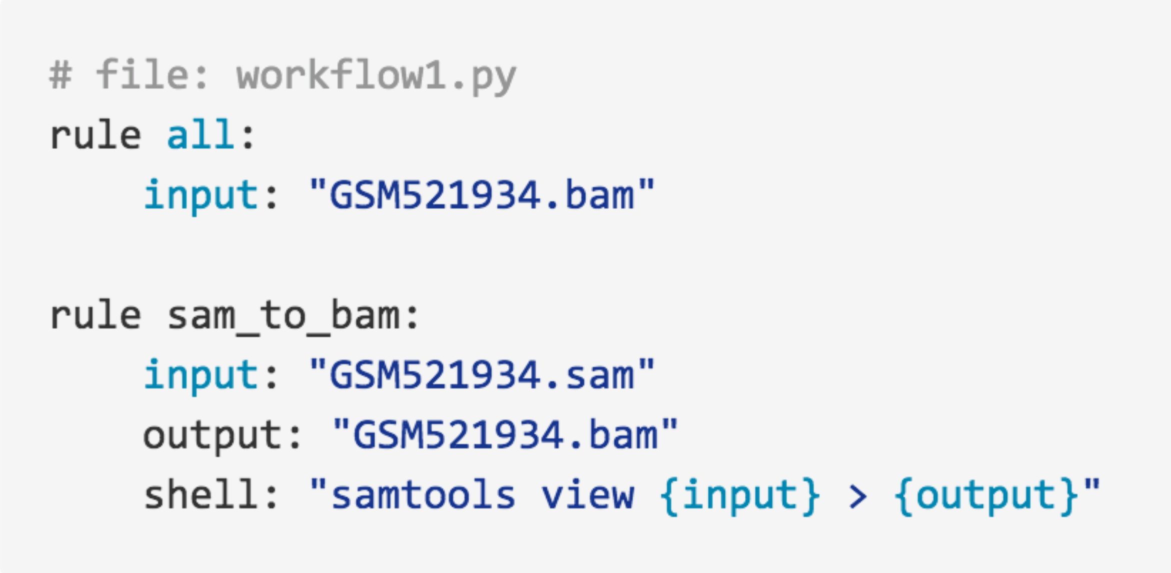
- Full tutorial at
  - https://github.com/rioualen/gene-regulation/blob/master/doc/gene-regulation_tutorials/gene-regulation_with_IFB_cloud.Rmd
  - Create a virtual disk to ensure persistence of your data.
  - Instantiate a virtual machine and mount the virtual disk.
  - Run the workflow.

- Live demo: run the brave test: rm -rf results; snakemake all (+ some parameters)
  - ## This address is only valid today, for me (my current instance)
  - ssh -A -p 22 root@192.54.201.196
  - cd ~/mydisk/GSE20870-analysis/
  - rm -rf results
  - date > start_date.txt ## not actually required
  - snakemake -p -s gene-regulation/scripts/snakefiles/workflows/factor_workflow.py --configfile gene-regulation/examples/GSE20870/GSE20870.yml
  - date > end_date.txt

- Before the end of this talk, the results should be re-generated (~10’ processing).

Snakemake features

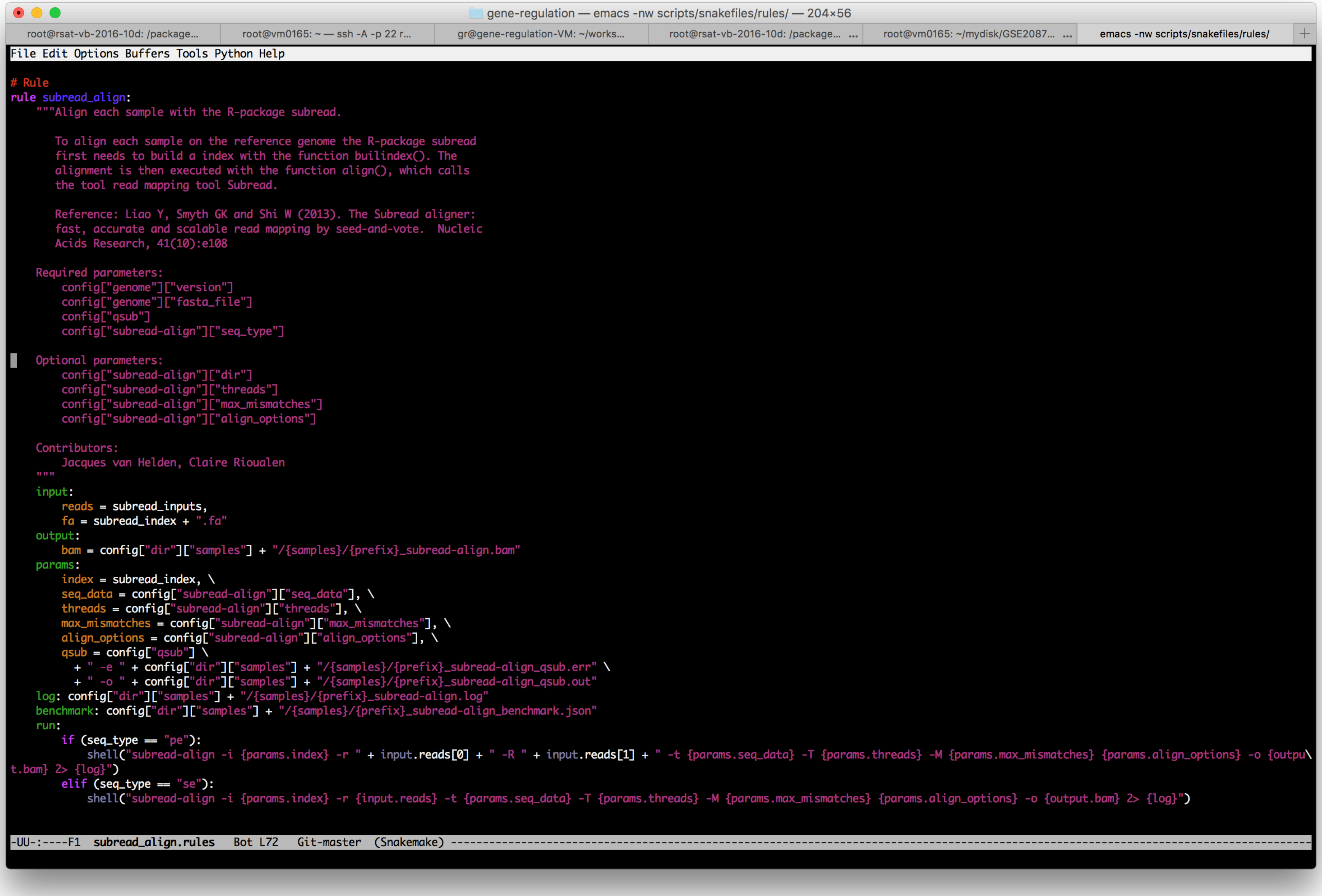
- Usual snakemake users write one script per workflows.
- Rules and targets (make inheritance).



```
# file: workflow1.py
rule all:
    input: "GSM521934.bam"

rule sam_to_bam:
    input: "GSM521934.sam"
    output: "GSM521934.bam"
    shell: "samtools view {input} > {output}"
```

Developing recyclable, modular rules for snakemake

- Claire Rioualen and myself developed a set of modular rules.
- Each rule describes one operation (use of a tool for one particular task).
- Rules can be combined *à la carte* by inclusion in a workflow.

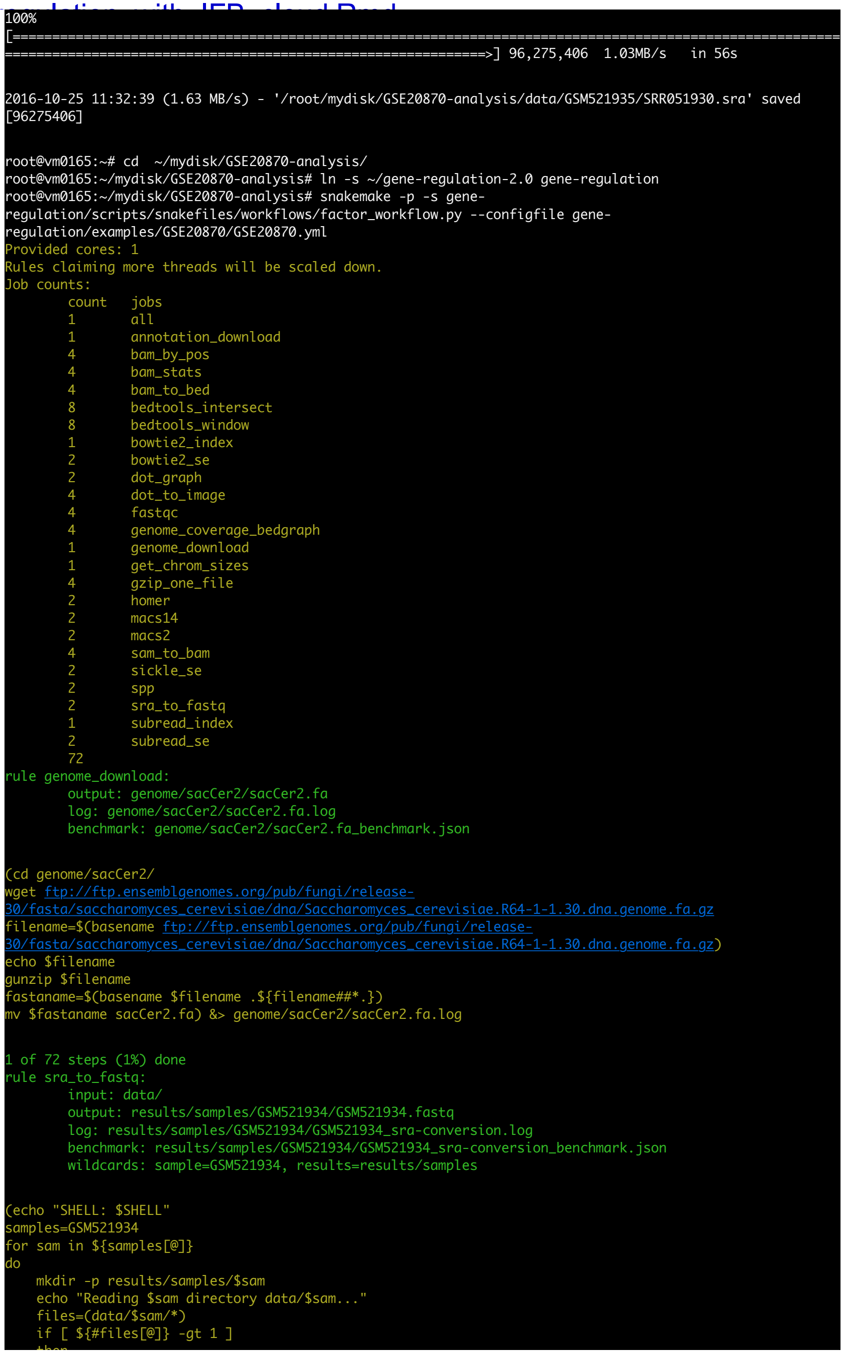

gene-regulation — emacs -nw scripts/snakefiles/rules/ — 204×56
root@rsat-vb-2016-10d: /package...
root@vm0165: ~ — ssh -A -p 22 r...
gr@gene-regulation-VM: ~/works...
root@rsat-vb-2016-10d: /package...
root@vm0165: ~/mydisk/GSE2087...
emacs -nw scripts/snakefiles/rules/
File Edit Options Buffers Tools Python Help

```
# Rule
rule subread_align:
    """Align each sample with the R-package subread.

       To align each sample on the reference genome the R-package subread
       first needs to build a index with the function builindex(). The
       alignment is then executed with the function align(), which calls
       the tool read mapping tool Subread.

       Reference: Liao Y, Smyth GK and Shi W (2013). The Subread aligner:
       fast, accurate and scalable read mapping by seed-and-vote.  Nucleic
       Acids Research, 41(10):e108

    Required parameters:
        config["genome"]["version"]
        config["genome"]["fasta_file"]
        config["qsub"]
        config["subread-align"]["seq_type"]

    Optional parameters:
        config["subread-align"]["dir"]
        config["subread-align"]["threads"]
        config["subread-align"]["max_mismatches"]
        config["subread-align"]["align_options"]

    Contributors:
        Jacques van Helden, Claire Rioualen
    """
    input:
        reads = subread_inputs,
        fa = subread_index + ".fa"
    output:
        bam = config["dir"]["samples"] + "/{samples}/{prefix}_subread-align.bam"
    params:
        index = subread_index, \
        seq_data = config["subread-align"]["seq_data"], \
        threads = config["subread-align"]["threads"], \
        max_mismatches = config["subread-align"]["max_mismatches"], \
        align_options = config["subread-align"]["align_options"], \
        qsub = config["qsub"] \
          + " -e " + config["dir"]["samples"] + "/{samples}/{prefix}_subread-align_qsub.err" \
          + " -o " + config["dir"]["samples"] + "/{samples}/{prefix}_subread-align_qsub.out"
    log: config["dir"]["samples"] + "/{samples}/{prefix}_subread-align.log"
    benchmark: config["dir"]["samples"] + "/{samples}/{prefix}_subread-align_benchmark.json"
    run:
        if (seq_type == "pe"):
            shell("subread-align -i {params.index} -r " + input.reads[0] + " -R " + input.reads[1] + " -t {params.seq_data} -T {params.threads} -M {params.max_mismatches} {params.align_options} -o {outpu\
t.bam} 2> {log}")
        elif (seq_type == "se"):
            shell("subread-align -i {params.index} -r {input.reads} -t {params.seq_data} -T {params.threads} -M {params.max_mismatches} {params.align_options} -o {output.bam} 2> {log}")

-UU-:----F1  subread_align.rules   Bot L72   Git-master  (Snakemake) ----------------------------------------
```

Gene-regulation demonstrator

https://github.com/rioualen/gene-regulation/blob/master/doc/gene-regulation_tutorials/ge

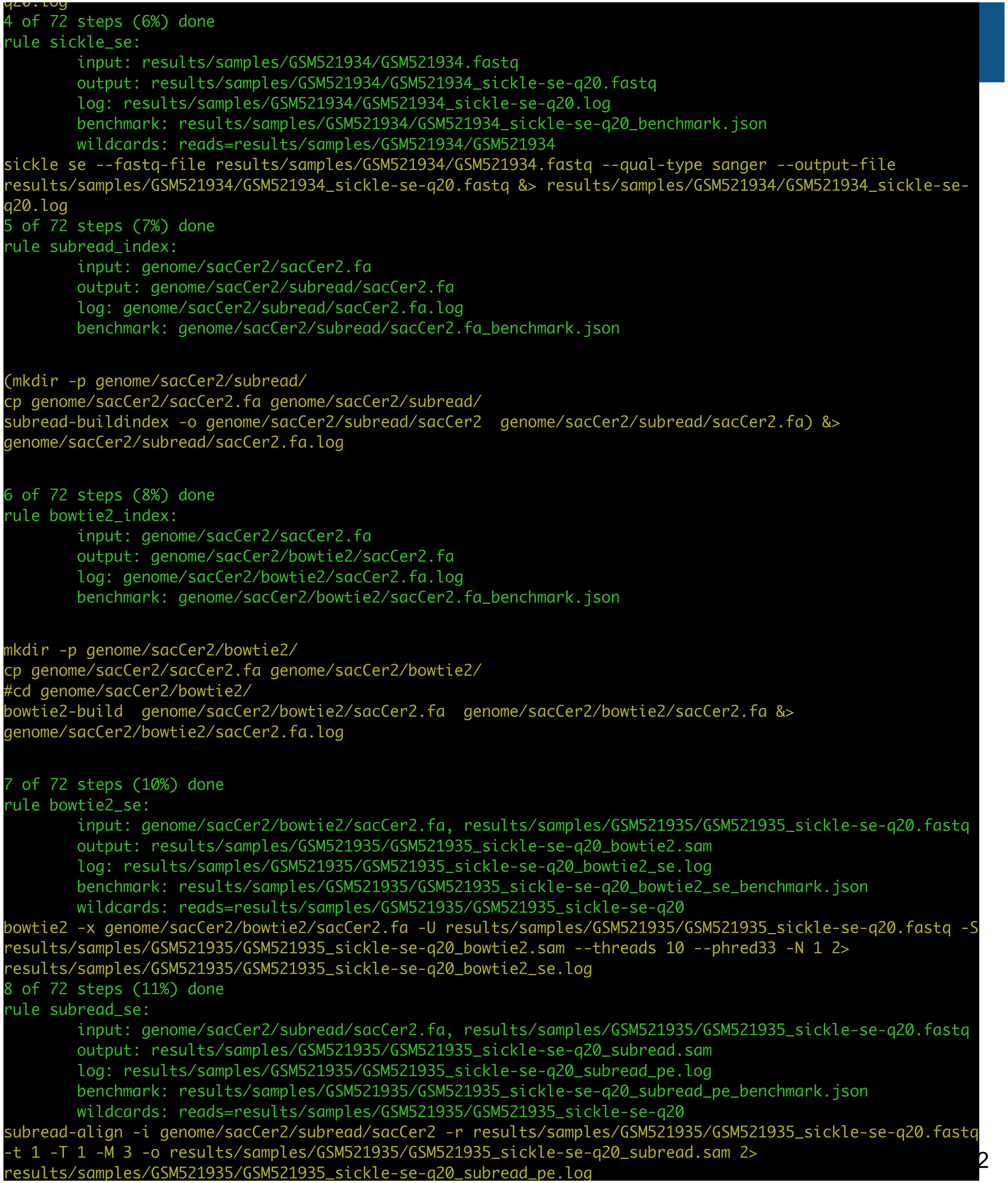
- A very small example of analysis: a small ChIP-seq experiment from the yeast Saccharomyces cerevisiae (GEO GSE20870).
- Create a Virtual Disk and VM instance on the IFB cloud
- Run the workflow
  - Download unaligned reads (172Mb for 2 samples).
  - Download and index reference genome.
  - Read mapping (bowtie2 + subread-align).
  - Peak calling
  - Motif analysis (RSAT peak-motifs via Web services).



```
100%
[=========================================================================================================
=============================================================>] 96,275,406  1.03MB/s   in 56s


2016-10-25 11:32:39 (1.63 MB/s) - '/root/mydisk/GSE20870-analysis/data/GSM521935/SRR051930.sra' saved
[96275406]


root@vm0165:~# cd  ~/mydisk/GSE20870-analysis/
root@vm0165:~/mydisk/GSE20870-analysis# ln -s ~/gene-regulation-2.0 gene-regulation
root@vm0165:~/mydisk/GSE20870-analysis# snakemake -p -s gene-
regulation/scripts/snakefiles/workflows/factor_workflow.py --configfile gene-
regulation/examples/GSE20870/GSE20870.yml
Provided cores: 1
Rules claiming more threads will be scaled down.
Job counts:
        count   jobs
        1       all
        1       annotation_download
        4       bam_by_pos
        4       bam_stats
        4       bam_to_bed
        8       bedtools_intersect
        8       bedtools_window
        1       bowtie2_index
        2       bowtie2_se
        2       dot_graph
        4       dot_to_image
        4       fastqc
        4       genome_coverage_bedgraph
        1       genome_download
        1       get_chrom_sizes
        4       gzip_one_file
        2       homer
        2       macs14
        2       macs2
        4       sam_to_bam
        2       sickle_se
        2       spp
        2       sra_to_fastq
        1       subread_index
        2       subread_se
        72
rule genome_download:
        output: genome/sacCer2/sacCer2.fa
        log: genome/sacCer2/sacCer2.fa.log
        benchmark: genome/sacCer2/sacCer2.fa_benchmark.json


(cd genome/sacCer2/
wget ftp://ftp.ensemblgenomes.org/pub/fungi/release-
30/fasta/saccharomyces_cerevisiae/dna/Saccharomyces_cerevisiae.R64-1-1.30.dna.genome.fa.gz
filename=$(basename ftp://ftp.ensemblgenomes.org/pub/fungi/release-
30/fasta/saccharomyces_cerevisiae/dna/Saccharomyces_cerevisiae.R64-1-1.30.dna.genome.fa.gz)
echo $filename
gunzip $filename
fastaname=$(basename $filename .${filename##*.})
mv $fastaname sacCer2.fa) &> genome/sacCer2/sacCer2.fa.log


1 of 72 steps (1%) done
rule sra_to_fastq:
        input: data/
        output: results/samples/GSM521934/GSM521934.fastq
        log: results/samples/GSM521934/GSM521934_sra-conversion.log
        benchmark: results/samples/GSM521934/GSM521934_sra-conversion_benchmark.json
        wildcards: sample=GSM521934, results=results/samples


(echo "SHELL: $SHELL"
samples=GSM521934
for sam in ${samples[@]}
do
    mkdir -p results/samples/$sam
    echo "Reading $sam directory data/$sam..."
    files=(data/$sam/*)
    if [ ${#files[@]} -gt 1 ]
```

Snakemake log

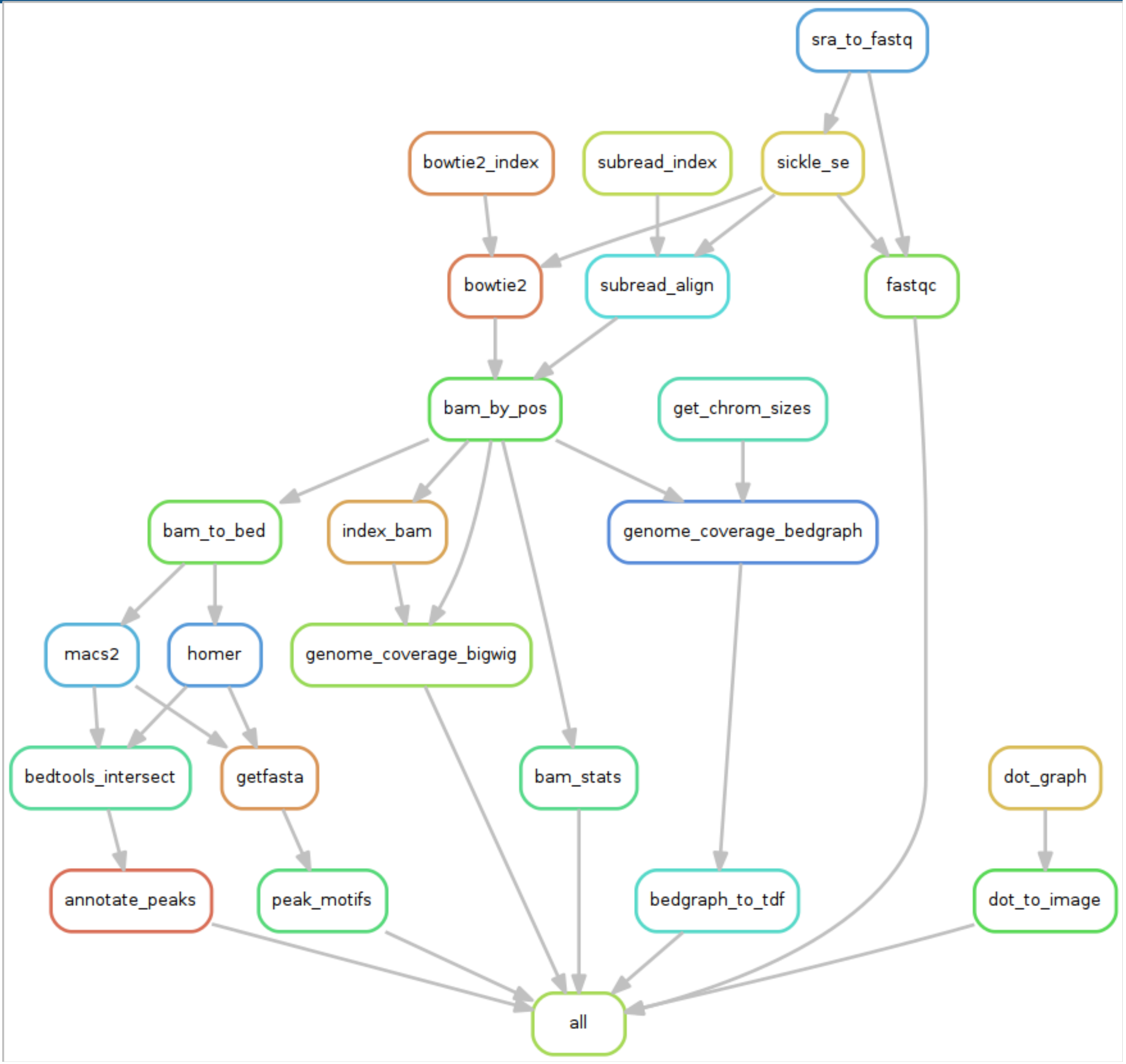
- Colored log indicating if everything goes fine.



```
q20.log
4 of 72 steps (6%) done
rule sickle_se:
        input: results/samples/GSM521934/GSM521934.fastq
        output: results/samples/GSM521934/GSM521934_sickle-se-q20.fastq
        log: results/samples/GSM521934/GSM521934_sickle-se-q20.log
        benchmark: results/samples/GSM521934/GSM521934_sickle-se-q20_benchmark.json
        wildcards: reads=results/samples/GSM521934/GSM521934
sickle se --fastq-file results/samples/GSM521934/GSM521934.fastq --qual-type sanger --output-file
results/samples/GSM521934/GSM521934_sickle-se-q20.fastq &> results/samples/GSM521934/GSM521934_sickle-se-
q20.log
5 of 72 steps (7%) done
rule subread_index:
        input: genome/sacCer2/sacCer2.fa
        output: genome/sacCer2/subread/sacCer2.fa
        log: genome/sacCer2/subread/sacCer2.fa.log
        benchmark: genome/sacCer2/subread/sacCer2.fa_benchmark.json


(mkdir -p genome/sacCer2/subread/
cp genome/sacCer2/sacCer2.fa genome/sacCer2/subread/
subread-buildindex -o genome/sacCer2/subread/sacCer2  genome/sacCer2/subread/sacCer2.fa) &>
genome/sacCer2/subread/sacCer2.fa.log


6 of 72 steps (8%) done
rule bowtie2_index:
        input: genome/sacCer2/sacCer2.fa
        output: genome/sacCer2/bowtie2/sacCer2.fa
        log: genome/sacCer2/bowtie2/sacCer2.fa.log
        benchmark: genome/sacCer2/bowtie2/sacCer2.fa_benchmark.json


mkdir -p genome/sacCer2/bowtie2/
cp genome/sacCer2/sacCer2.fa genome/sacCer2/bowtie2/
#cd genome/sacCer2/bowtie2/
bowtie2-build  genome/sacCer2/bowtie2/sacCer2.fa  genome/sacCer2/bowtie2/sacCer2.fa &>
genome/sacCer2/bowtie2/sacCer2.fa.log


7 of 72 steps (10%) done
rule bowtie2_se:
        input: genome/sacCer2/bowtie2/sacCer2.fa, results/samples/GSM521935/GSM521935_sickle-se-q20.fastq
        output: results/samples/GSM521935/GSM521935_sickle-se-q20_bowtie2.sam
        log: results/samples/GSM521935/GSM521935_sickle-se-q20_bowtie2_se.log
        benchmark: results/samples/GSM521935/GSM521935_sickle-se-q20_bowtie2_se_benchmark.json
        wildcards: reads=results/samples/GSM521935/GSM521935_sickle-se-q20
bowtie2 -x genome/sacCer2/bowtie2/sacCer2.fa -U results/samples/GSM521935/GSM521935_sickle-se-q20.fastq -S
results/samples/GSM521935/GSM521935_sickle-se-q20_bowtie2.sam --threads 10 --phred33 -N 1 2>
results/samples/GSM521935/GSM521935_sickle-se-q20_bowtie2_se.log
8 of 72 steps (11%) done
rule subread_se:
        input: genome/sacCer2/subread/sacCer2.fa, results/samples/GSM521935/GSM521935_sickle-se-q20.fastq
        output: results/samples/GSM521935/GSM521935_sickle-se-q20_subread.sam
        log: results/samples/GSM521935/GSM521935_sickle-se-q20_subread_pe.log
        benchmark: results/samples/GSM521935/GSM521935_sickle-se-q20_subread_pe_benchmark.json
        wildcards: reads=results/samples/GSM521935/GSM521935_sickle-se-q20
subread-align -i genome/sacCer2/subread/sacCer2 -r results/samples/GSM521935/GSM521935_sickle-se-q20.fastq
-t 1 -T 1 -M 3 -o results/samples/GSM521935/GSM521935_sickle-se-q20_subread.sam 2>
results/samples/GSM521935/GSM521935_sickle-se-q20_subread_pe.log
```

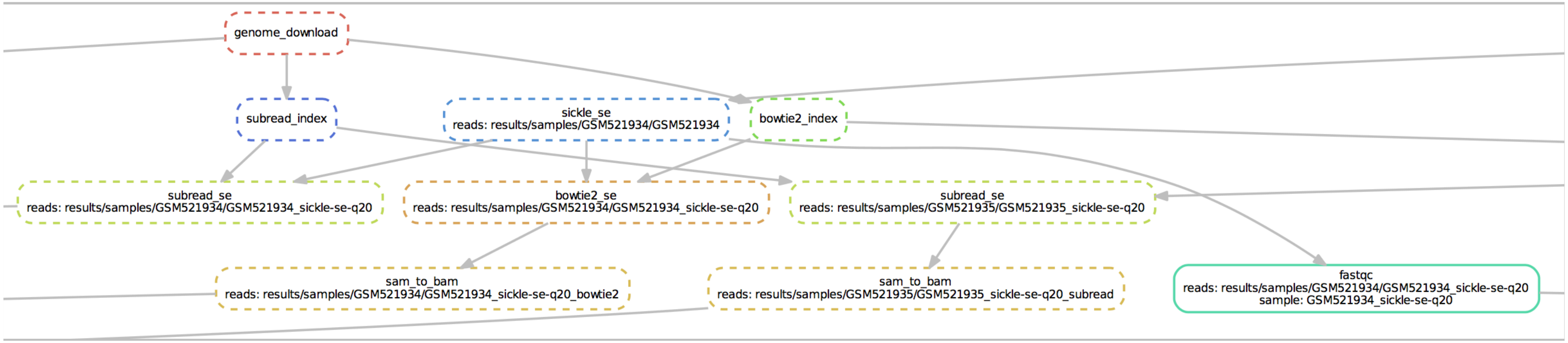
Snakemake rule graph - the logics of the workflow

- Snakemake automatically generate the flow chart of operations from the script itself.



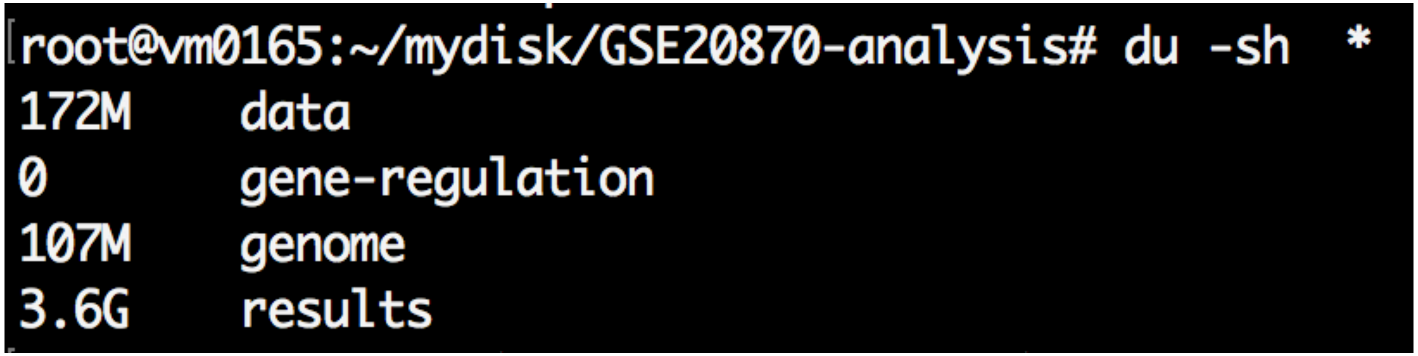
Snakemake DAG- live report of the state of each result file

- Automatically generated by snakemake.
- The DAG indicates the state of each result file.
  - Dotted border: completed.
  - Plain border: to be done.


genome_download
subread_index
sickle_se
reads: results/samples/GSM521934/GSM521934
bowtie2_index
subread_se
reads: results/samples/GSM521934/GSM521934_sickle-se-q20
bowtie2_se
reads: results/samples/GSM521934/GSM521934_sickle-se-q20
subread_se
reads: results/samples/GSM521935/GSM521935_sickle-se-q20
sam_to_bam
reads: results/samples/GSM521934/GSM521934_sickle-se-q20_bowtie2
sam_to_bam
reads: results/samples/GSM521935/GSM521935_sickle-se-q20_subread
fastqc
reads: results/samples/GSM521934/GSM521934_sickle-se-q20
sample: GSM521934_sickle-se-q20

The final result

- The result folder contains all the intermediate files + final results.
- Could be cleaned for the intermediate files (.sam, .fastq, …).
- No problem to run the brave final test before publishing
  - rm –rf results; snakemake all



```
root@vm0165:~/mydisk/GSE20870-analysis# du -sh  *
172M	data
0	gene-regulation
107M	genome
3.6G	results
```

Other snakemake features

- Regulating the number of jobs sent to job scheduler (qsub).
- Managing dependencies between jobs (only sends jobs for which input files are present).
- Management of interrupted processes: remove all the files if the job did not complete.
- Regenerating the
- Benchmarking: automated measurement of the time spent in each task for each sample.

The Century-lasting P-value controversy

The P-value controversy (selected sample)

1. Leek,J.T. and Peng,R.D. (2015) ***Statistics: P values are just the tip of the iceberg.*** *Nature*, **520**, 612–612.
2. Halsey,L.G., Curran-Everett,D., Vowler,S.L. and Drummond,G.B. (2015) ***The fickle P value generates irreproducible results.*** *Nature Methods*, **12**, 179–185.
3. Trafimow,D. and Marks,M. (2015) Editorial. *http://dx.doi.org/10.1080/01973533.2015.1012991*, **37**, 1–2. (Banning the p-value)
4. McIlroy,D. (2014) ***Seduced by a P-value...*** *Anaesth Intensive Care*, **42**, 551–554.
5. Annesley,T.M. and Boyd,J.C. (2014) ***The P value: probable does not mean practical.*** *Clin Chem*, **60**, 1021–1023.
6. Gaudart,J., Huiart,L., Milligan,P.J., Thiebaut,R. and Giorgi,R. (2014) ***Reproducibility issues in science, is P value really the only answer?*** *Proc Natl Acad Sci U S A*, **111**, E1934–E1934.
7. Wood,J., Freemantle,N., King,M. and Nazareth,I. (2014) ***Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data.*** *BMJ*, **348**, g2215.
8. Pandis,N. (2013) ***The P value problem.*** *Am J Orthod Dentofacial Orthop*, **143**, 150–151.
9. Pollet,T.V. (2013) ***Much ado about p. What does a p-value mean when testing hypotheses with aggregated cross-cultural data in the field of evolution and human behavior?*** *Front Psychol*, **4**, 734.
10. Yin,Y. and Zhao,J. (2013) ***Testing normal means: the reconcilability of the P value and the Bayesian evidence.*** *ScientificWorldJournal*, **2013**, 381539–7.

The P-value controversy (continued)

11. Zeichner,S.B. (2012) ***Holding rhetoric to a new standard: what's the P value of that statement, senator?*** *J Am Osteopath Assoc*, **112**, 648–649.
12. Keriazes,G.A. (2012) ***Misuse of the p value for baseline characteristics.*** *Pharmacotherapy*, **32**, e172–3.
13. Sullivan,G.M. and Feinn,R. (2012) ***Using Effect Size-or Why the P Value Is Not Enough.*** *J Grad Med Educ*, **4**, 279–282.
14. Ranstam,J. (2012) ***Why the P-value culture is bad and confidence intervals a better alternative.*** *Osteoarthr. Cartil.*, **20**, 805–808.
15. Youngquist,S.T. (2012) Part 19: ***What is a P value?*** *Air Med. J.*, **31**, 56–71.
16. Kuss,O. and Stang,A. (2012) ***The p-value - a well-understood and properly used statistical concept?*** *Contact Derm.*, **66**, 1–3.
17. Falissard,B. (2012) ***Statistics in brief: when to use and when not to use a threshold p value***. *Clin. Orthop. Relat. Res.*, **470**, 315–316.
18. Dalman,M.R., Deeter,A., Nimishakavi,G. and Duan,Z.-H. (2012) ***Fold change and p-value cutoffs significantly alter microarray interpretations***. *BMC Bioinformatics*, **13 Suppl 2**, S11.
19. Gadbury,G.L. and Allison,D.B. (2012) ***Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature***. *PLoS ONE*, **7**, e46363.
20. Casson,R.J. (2011) ***The pesty P value***. *Clin. Experiment. Ophthalmol.*, **39**, 849–850.

The P-value controversy (continued)

21. Singh,K. (2011) ***A latent p-value in testing by bootstrap***. *J Biopharm Stat*, **21**, 1232–1235.
22. Gerke,O., Høilund-Carlsen,P.F. and Vach,W. (2011) [***Different meaning of the p-value in exploratory and confirmatory hypothesis testing***]. *Ugeskr. Laeg.*, **173**, 2261–2264.
23. Ghazarian,S.R. (2011) ***Distinguishing statistical significance from clinical importance: the value of the P value.*** *Pediatr Rev*, **32**, 73–74.
24. Lee,J.J. (2011) ***Demystify statistical significance--time to move on from the p value to bayesian analysis.*** *J Natl Cancer Inst*, **103**, 2–3.
25. Lewis,E.F. (2010) ***Beyond the P value: The quest for improving health status in patients with ischemic heart disease.*** *Circulation*, **122**, 1664–1666.
26. Cook,C. (2010) ***Five per cent of the time it works 100 per cent of the time: the erroneousness of the P value.*** *J Man Manip Ther*, **18**, 123–125.
27. Mitchell,M.S., Yu,M.C. and Whiteside,T.L. (2010) ***The tyranny of statistics in medicine: a critique of unthinking adherence to an arbitrary p value***. *Cancer Immunol. Immunother.*, **59**, 1137–1140.
28. Pitak-Arnnop,P., Dhanuthai,K., Hemprich,A. and Pausch,N.C. (2010) ***Misleading p-value:do you recognise it?*** *Eur J Dent*, **4**, 356–358.
29. Niculescu,A.B. and Le-Niculescu,H. (2010) ***The P-value illusion: how to improve (psychiatric) genetic studies***. *Am J Med Genet B Neuropsychiatr Genet*, **153B**, 847–849.
30. Connelly,N.R. and Raghunathan,K. (2010) ***The value of the P-value.*** *J Clin Anesth*, **22**, 154–155.

The P-value controversy (continued)

31. Hooper,R. (2009) ***The Bayesian interpretation of a P-value depends only weakly on statistical power in realistic situations***. *J Clin Epidemiol*, **62**, 1242–1247.
32. Huak,C.Y. (2009) ***Are you a p-value worshipper?*** *Eur J Dent*, **3**, 161–164.
33. Prel,du,J.-B., Hommel,G., Röhrig,B. and Blettner,M. (2009) ***Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications.*** *Dtsch Arztebl Int*, **106**, 335–339.
34. Grunkemeier,G.L., Wu,Y. and Furnary,A.P. (2009) ***What is the value of a p value?*** *Ann. Thorac. Surg.*, **87**, 1337–1343.
35. Atenafu,E.G., Hamid,J.S., Stephens,D., To,T. and Beyene,J. (2009) ***A small p-value from an observed data is not evidence of adequate power for future similar-sized studies: a cautionary note.*** *Contemp Clin Trials*, **30**, 155–157.
36. Goodman,S. (2008) ***A dirty dozen: twelve p-value misconceptions***. *Semin. Hematol.*, **45**, 135–140.
37. Dahiru,T. (2008) ***P - value, a true test of statistical significance? A cautionary note***. *Ann Ib Postgrad Med*, **6**, 21–26.
38. Adedokun,B.O. (2008) ***P - value and confidence intervals - facts and farces.*** *Ann Ib Postgrad Med*, **6**, 33–34.
39. Henry,R.E. (2008) ***The great p value: we the people.*** *Am Health Drug Benefits*, **1**, 3.
40. Fundel,K., Küffner,R., Aigner,T. and Zimmer,R. (2008) ***Normalization and gene p-value estimation: issues in microarray data processing.*** *Bioinform Biol Insights*, **2**, 291–305.

The P-value controversy (continued)

41. Lesaffre,E. (2008) ***Use and misuse of the p-value.*** *Bull NYU Hosp Jt Dis*, **66**, 146–149.
42. Morgan,J.F. (2007) ***p Value fetishism and use of the Bonferroni adjustment.*** *Evid Based Ment Health*, **10**, 34–35.
43. Halloran,P.F., Reeve,J. and Kaplan,B. (2006) ***Lies, damn lies, and statistics: the perils of the P value.*** *Am. J. Transplant.*, **6**, 10–11.
44. Kain,Z.N. (2005) ***The legend of the P value.*** *Anesth. Analg.*, **101**, 1454–1456.
45. Koehnle,T. (2005) ***The proof is not in the P value***. *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, **288**, R777–author reply R777–8.
46. Wilhelmus,K.R. (2004) ***Beyond the P: III: Possible insignificance of the nonsignificant P value.*** *J Cataract Refract Surg*, **30**, 2425–2426.
47. Verhagen,A.P., Ostelo,R.W.J.G. and Rademaker,A. (2004) ***Is the p value really so significant?*** *Aust J Physiother*, **50**, 261–262.
48. Goodman,S. (2003) ***Commentary: The P-value, devalued.*** *International journal of epidemiology*, **32**, 699–702.
49. Trafimow,D. (2003) ***Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem.*** *Psychological Review*, **110**, 526–535.
50. Osterwalder,J.J. (2002) ***The P value as the guardian of medical truth -- illusion or reality?*** *Eur J Emerg Med*, **9**, 283–286.
51. Taube,A. and Malmquist,J. (2001) ***[Count on your beliefs. Bayes--not the P value--measures credence]***. *Lakartidningen*, **98**, 3208–3211.

The P-value controversy (continued)

51. Weinberg,C.R. (2001) ***It's time to rehabilitate the P-value.*** *Epidemiology*, **12**, 288–290.
52. Sandweiss,D.A. (2000) ***P value out of control.*** *Arch Intern Med*, **160**, 1872–author reply 1877–8.
53. Lang,J.M., Rothman,K.J. and Cann,C.I. (1998) ***That confounded P-value.*** *Epidemiology*, **9**, 7–8.
54. Hopkins,B.L., Cole,B.L. and Mason,T.L. (1998) ***A critique of the usefulness of inferential statistics in applied behavior analysis.*** *Behav Anal*, **21**, 125–137.
55. Kanter,M.H., Poole,G. and Garratty,G. (1997) ***Misinterpretation and misapplication of p values in antibody identification: the lack of value of a p value.*** *Transfusion*, **37**, 816–822.
56. Chia,K.S. (1997) ***'Significant-itis'--an obsession with the P-value***. *Scand J Work Environ Health*, **23**, 152–154.
57. Barnett,M.L. and Mathisen,A. (1997) ***Tyranny of the p-value: the conflict between statistical significance and common sense.*** *J. Dent. Res.*, **76**, 534–536.
58. Nurminen,M. (1997) ***[Slaves of the p-value?]***. *Duodecim*, **113**, 277–280.
59. Marino,P. (1995) ***In survival curves is the P-value enough?*** *Lung Cancer*, **12**, 87–89.
60. Hadler,N.M. (1994) ***The imperious p value.*** *J. Rheumatol.*, **21**, 1787–1788.

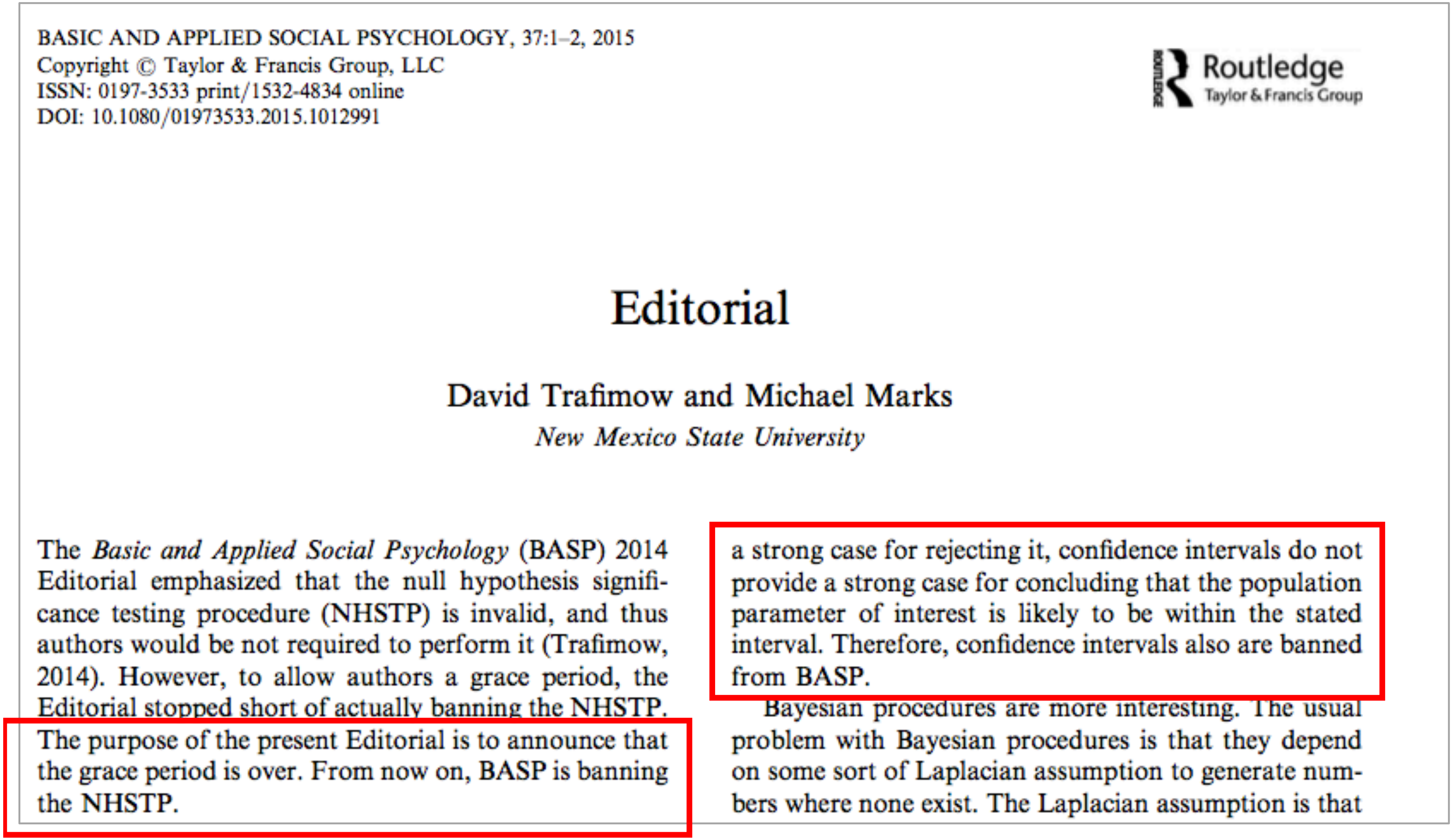
The P-value controversy (continued)

61. Mikulecký,M. (1990) ***Not the p value but the practice has the last word in the search for the truth.*** *J. Lab. Clin. Med.*, **115**, 526–527.
62. Boen,J.R. (1989) ***Understanding p-value misuse.*** *Statistics in medicine*, **8**, 1413–1414.
63. Mikulecký,M. (1989) ***The scientific truth, the significance level alpha, and the p value of a test.*** *J. Lab. Clin. Med.*, **113**, 759–760.
64. Evans,S.J., Mills,P. and Dawson,J. (1988) ***The end of the p value?*** *Br Heart J*, **60**, 177–180.
65. Gordon,I. (1985) ***Misconception concerning the ubiquitous p value.*** *J Occup Med*, **27**, 403.
66. Cook,D. (1978) ***We should stop misuse of P value.*** *Pediatrics*, **61**, 502.
67. Jekel,J.F. (1977) ***Should we stop using the P value in descriptive studies?*** *Pediatrics*, **60**, 124–126.
68. Rozeboom,W.W. (1960) ***The fallacy of the null-hypothesis significance test***. *Psychol Bull*, **57**, 416–428.
69. … and many others !

Recurrent criticisms

- Misuse of signficance testing.
- Mis-interpretation of the P-value.
- Bayesian inference between “frequentist” interpretation of probabilities.
- Instability of the parameters.

The p-value is banned from a scientific journal



BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

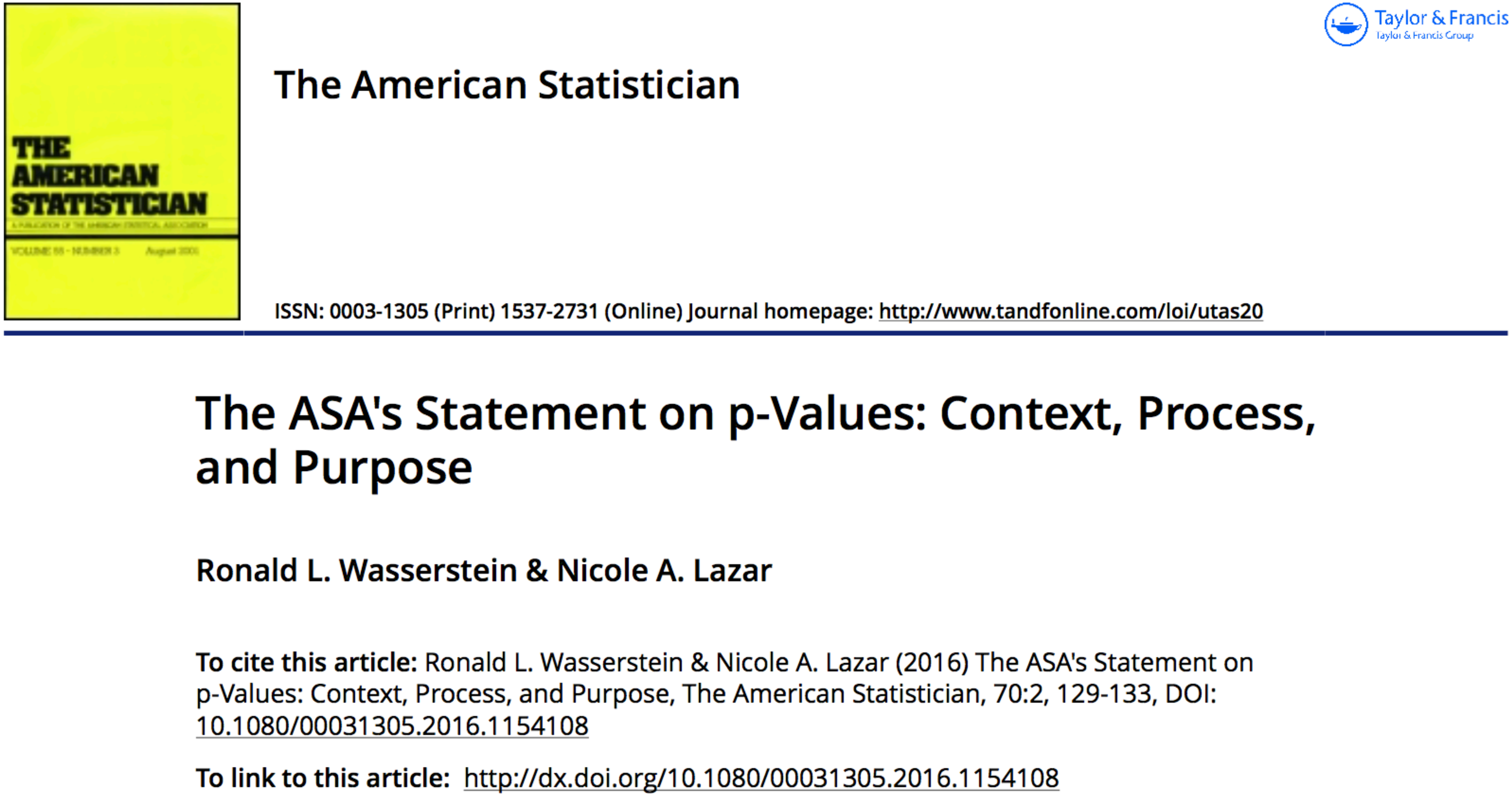
David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that

- Trafimow,D. and Marks,M. (2015) BASP Editorial. *http://dx.doi.org/10.1080/01973533.2015.1012991*, **37**, 1–2.



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

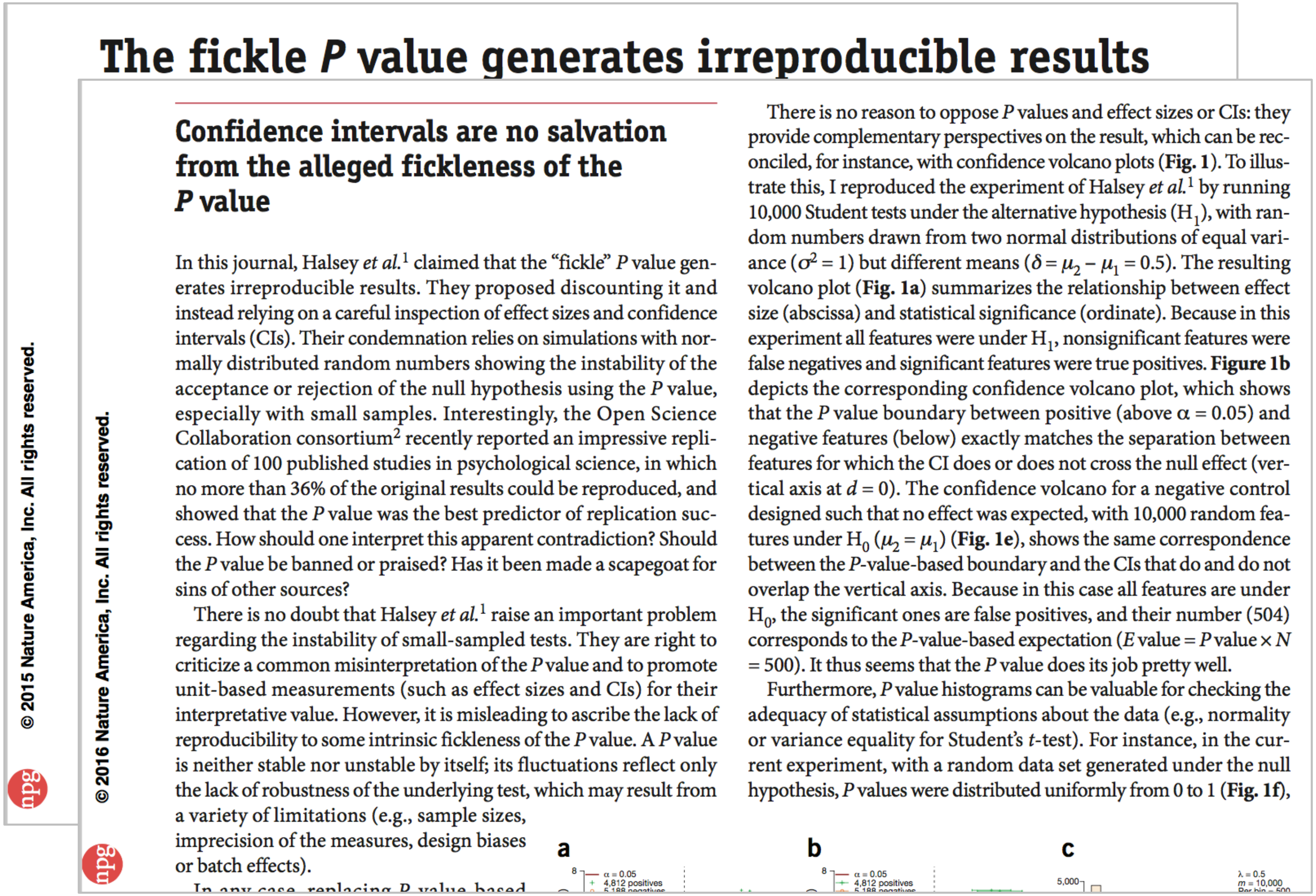
Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1154108

- Wasserstein,R.L. and Lazar,N.A. (2016) The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

P-values and irreproducible results



The fickle *P* value generates irreproducible results

Confidence intervals are no salvation from the alleged fickleness of the *P* value

In this journal, Halsey *et al.*[1] claimed that the "fickle" *P* value generates irreproducible results. They proposed discounting it and instead relying on a careful inspection of effect sizes and confidence intervals (CIs). Their condemnation relies on simulations with normally distributed random numbers showing the instability of the acceptance or rejection of the null hypothesis using the *P* value, especially with small samples. Interestingly, the Open Science Collaboration consortium[2] recently reported an impressive replication of 100 published studies in psychological science, in which no more than 36% of the original results could be reproduced, and showed that the *P* value was the best predictor of replication success. How should one interpret this apparent contradiction? Should the *P* value be banned or praised? Has it been made a scapegoat for sins of other sources?

There is no doubt that Halsey *et al.*[1] raise an important problem regarding the instability of small-sampled tests. They are right to criticize a common misinterpretation of the *P* value and to promote unit-based measurements (such as effect sizes and CIs) for their interpretative value. However, it is misleading to ascribe the lack of reproducibility to some intrinsic fickleness of the *P* value. A *P* value is neither stable nor unstable by itself; its fluctuations reflect only the lack of robustness of the underlying test, which may result from a variety of limitations (e.g., sample sizes, imprecision of the measures, design biases or batch effects).

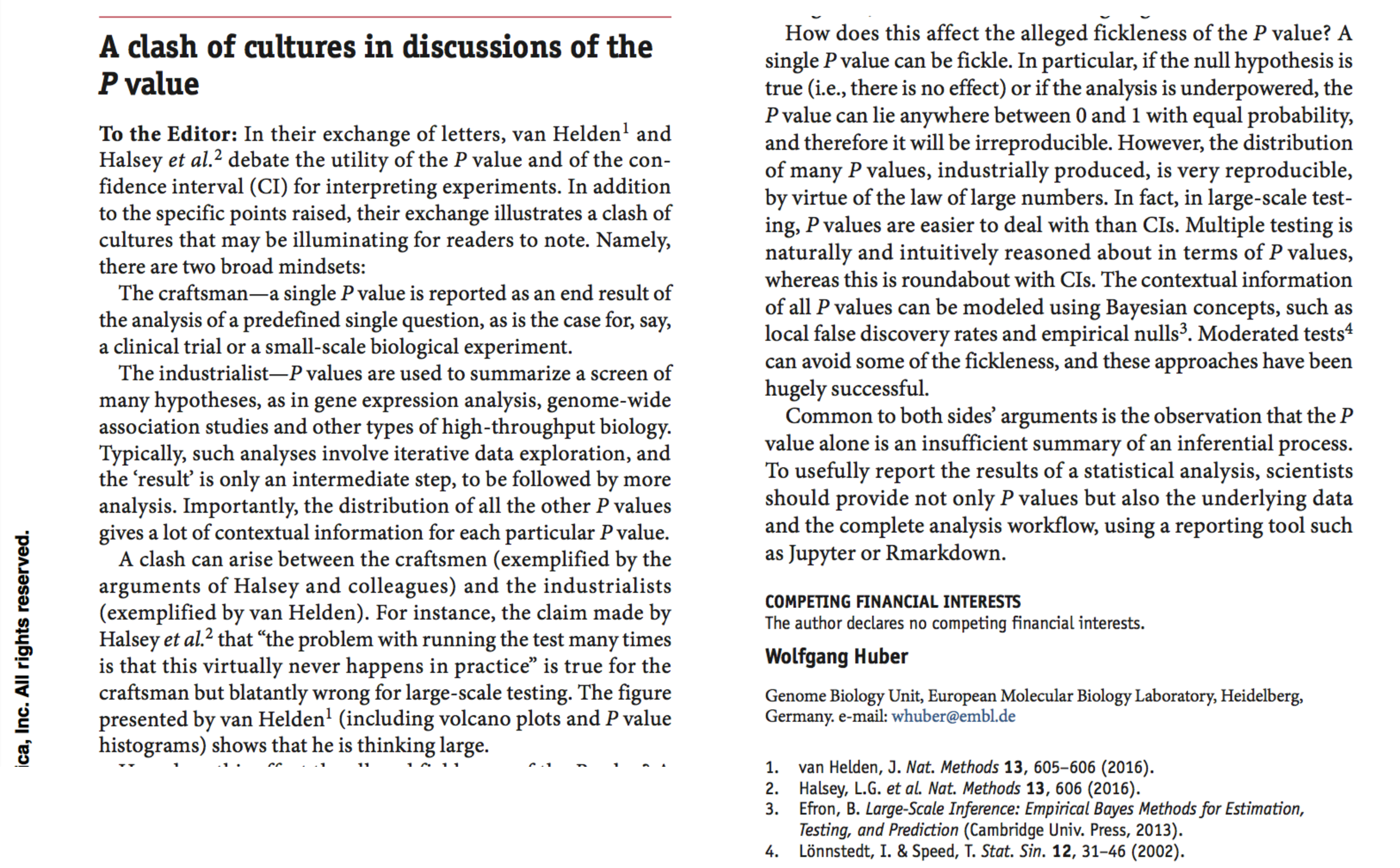
In any case, replacing *P* value based

There is no reason to oppose *P* values and effect sizes or CIs: they provide complementary perspectives on the result, which can be reconciled, for instance, with confidence volcano plots (**Fig. 1**). To illustrate this, I reproduced the experiment of Halsey *et al.*[1] by running 10,000 Student tests under the alternative hypothesis (H_1), with random numbers drawn from two normal distributions of equal variance ($\sigma^2 = 1$) but different means ($\delta = \mu_2 - \mu_1 = 0.5$). The resulting volcano plot (**Fig. 1a**) summarizes the relationship between effect size (abscissa) and statistical significance (ordinate). Because in this experiment all features were under H_1, nonsignificant features were false negatives and significant features were true positives. **Figure 1b** depicts the corresponding confidence volcano plot, which shows that the *P* value boundary between positive (above $\alpha = 0.05$) and negative features (below) exactly matches the separation between features for which the CI does or does not cross the null effect (vertical axis at $d = 0$). The confidence volcano for a negative control designed such that no effect was expected, with 10,000 random features under H_0 ($\mu_2 = \mu_1$) (**Fig. 1e**), shows the same correspondence between the *P*-value-based boundary and the CIs that do and do not overlap the vertical axis. Because in this case all features are under H_0, the significant ones are false positives, and their number (504) corresponds to the *P*-value-based expectation (*E* value = *P* value × *N* = 500). It thus seems that the *P* value does its job pretty well.

Furthermore, *P* value histograms can be valuable for checking the adequacy of statistical assumptions about the data (e.g., normality or variance equality for Student's *t*-test). For instance, in the current experiment, with a random data set generated under the null hypothesis, *P* values were distributed uniformly from 0 to 1 (**Fig. 1f**),

- Halsey,L.G., Curran-Everett,D., Vowler,S.L. and Drummond,G.B. (2015) The fickle P value generates irreproducible results. *Nature Methods*, **12**, 179–185.
- van Helden,J. (2016) Confidence intervals are no salvation from the alleged fickleness of the P value. *Nature Methods*, **13**, 605–606

A clash of cultures?



A clash of cultures in discussions of the *P* value

To the Editor: In their exchange of letters, van Helden[1] and Halsey *et al.*[2] debate the utility of the *P* value and of the confidence interval (CI) for interpreting experiments. In addition to the specific points raised, their exchange illustrates a clash of cultures that may be illuminating for readers to note. Namely, there are two broad mindsets:

The craftsman—a single *P* value is reported as an end result of the analysis of a predefined single question, as is the case for, say, a clinical trial or a small-scale biological experiment.

The industrialist—*P* values are used to summarize a screen of many hypotheses, as in gene expression analysis, genome-wide association studies and other types of high-throughput biology. Typically, such analyses involve iterative data exploration, and the 'result' is only an intermediate step, to be followed by more analysis. Importantly, the distribution of all the other *P* values gives a lot of contextual information for each particular *P* value.

A clash can arise between the craftsmen (exemplified by the arguments of Halsey and colleagues) and the industrialists (exemplified by van Helden). For instance, the claim made by Halsey *et al.*[2] that "the problem with running the test many times is that this virtually never happens in practice" is true for the craftsman but blatantly wrong for large-scale testing. The figure presented by van Helden[1] (including volcano plots and *P* value histograms) shows that he is thinking large.

How does this affect the alleged fickleness of the *P* value? A single *P* value can be fickle. In particular, if the null hypothesis is true (i.e., there is no effect) or if the analysis is underpowered, the *P* value can lie anywhere between 0 and 1 with equal probability, and therefore it will be irreproducible. However, the distribution of many *P* values, industrially produced, is very reproducible, by virtue of the law of large numbers. In fact, in large-scale testing, *P* values are easier to deal with than CIs. Multiple testing is naturally and intuitively reasoned about in terms of *P* values, whereas this is roundabout with CIs. The contextual information of all *P* values can be modeled using Bayesian concepts, such as local false discovery rates and empirical nulls[3]. Moderated tests[4] can avoid some of the fickleness, and these approaches have been hugely successful.

Common to both sides' arguments is the observation that the *P* value alone is an insufficient summary of an inferential process. To usefully report the results of a statistical analysis, scientists should provide not only *P* values but also the underlying data and the complete analysis workflow, using a reporting tool such as Jupyter or Rmarkdown.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

Wolfgang Huber

Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. e-mail: whuber@embl.de

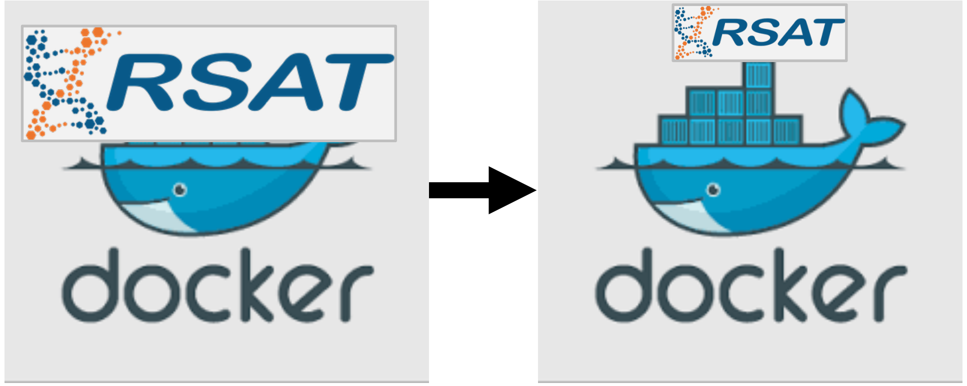
1. van Helden, J. *Nat. Methods* **13**, 605–606 (2016).
2. Halsey, L.G. *et al. Nat. Methods* **13**, 606 (2016).
3. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Cambridge Univ. Press, 2013).
4. Lönnstedt, I. & Speed, T. *Stat. Sin.* **12**, 31–46 (2002).

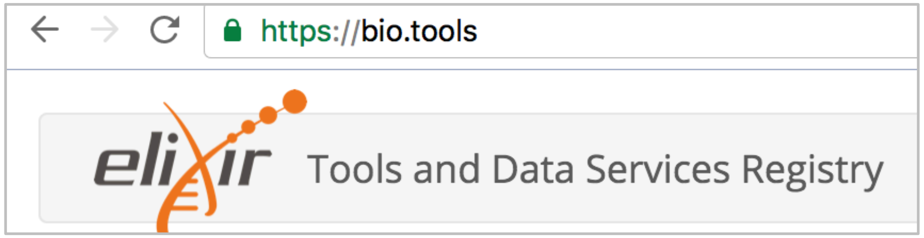
- Huber,W. (2016) A clash of cultures in discussions of the P value. *Nature Methods*, **13**, 607–607.

Conclusions and perspectives

On-going projects

- Packaging: apt-get install RSAT.
- Docker with modular components: bioconda, mulled.
- RSAT access via Galaxy
  - Current project: Galaxy calls RSAT Web services: Jocelyn Brayet (Institut Curie, Paris, France).
  - Later perspectives: Galaxy runs RSAT docker container.
- Documenting all tools in bio.tools.



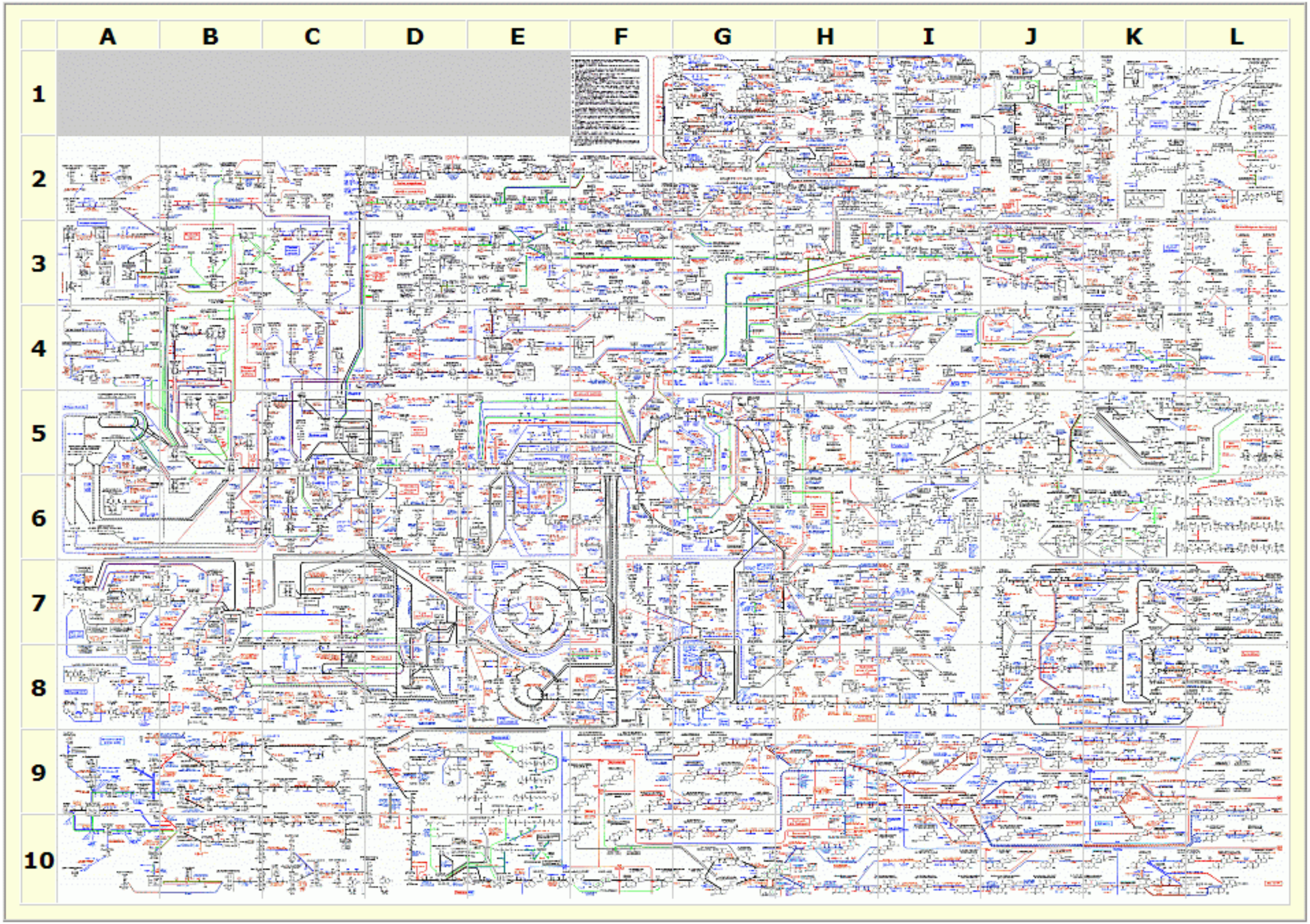

https://bio.tools

elixir Tools and Data Services Registry

Snakemake community building (currently starting in France)

- CoBRAS: user community
  COmmunauté de Bioinformaticiens Rassemblés Autour de Snakemake
  [*Community Of Bioinformaticians Regrouped Around Snakemake*]
- SnaTaF: Snakemake Task Force: workflow designers and library developers.

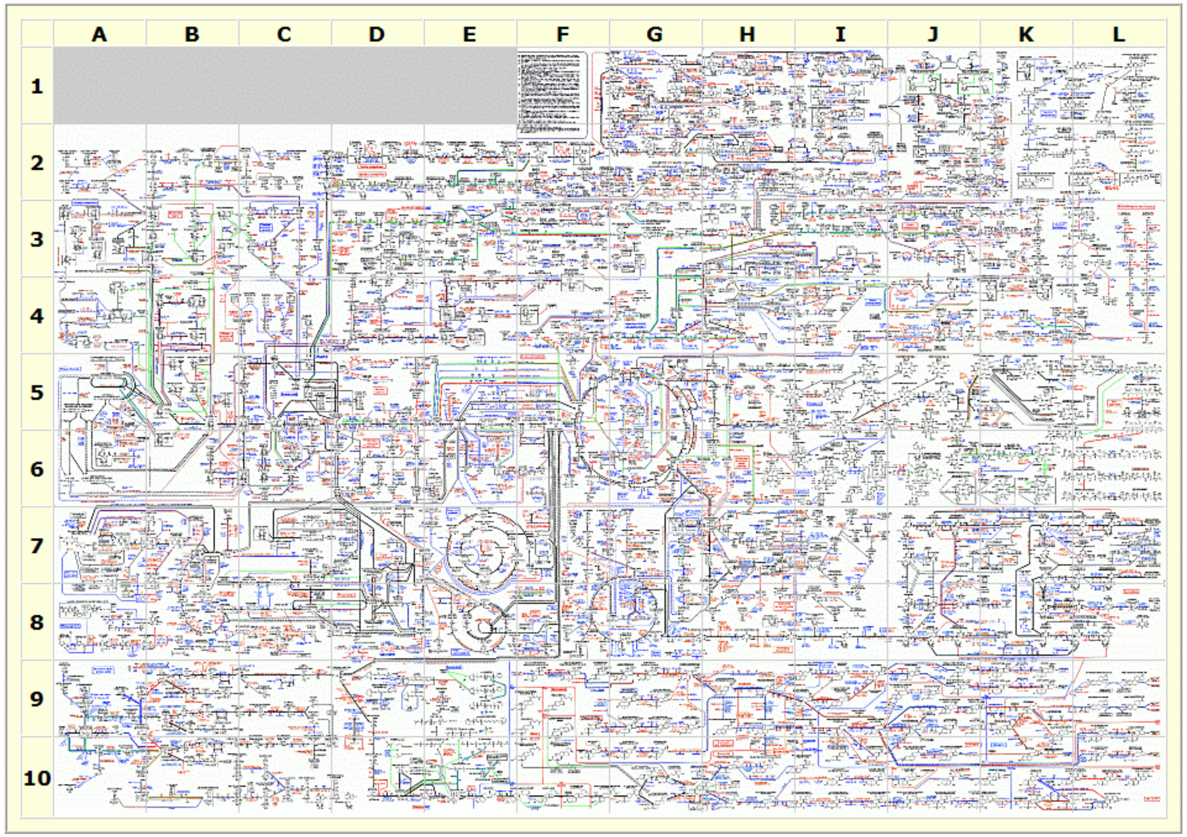
Bio.tools - towards a wall chart of all bioinformatics tools ?



http://www.expasy.ch/cgi-bin/show_thumbnails.pl

Bio.tools - towards a wall chart of all bioinformatics tools ?

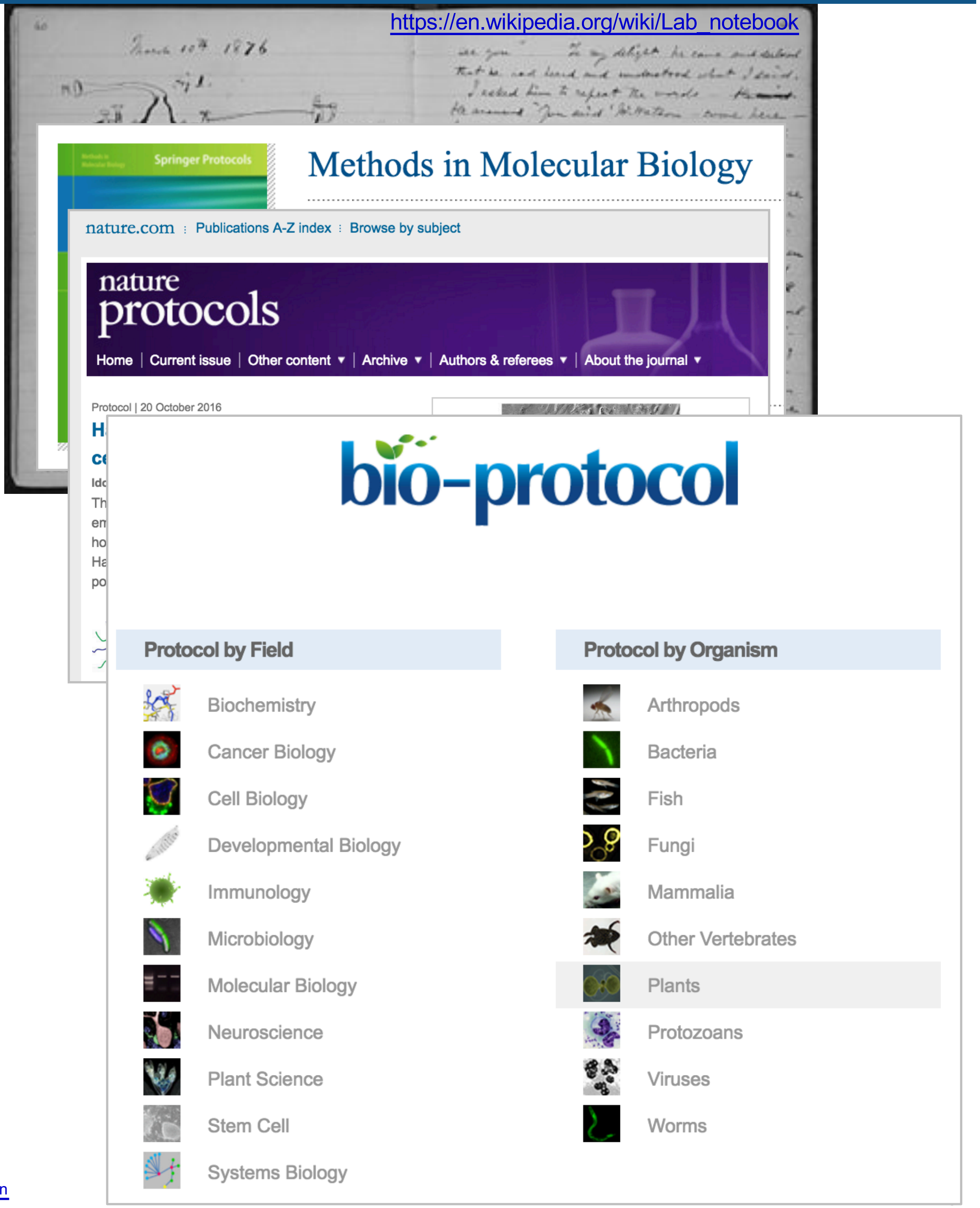
- Each tool is as a reaction
- Each data/result file as a metabolite
- Any workflow should be a sub-graph of the whole wall chart.
- Note: the wall chart actually never fits on a wall, at least a 2-dimensional wall.



http://www.expasy.ch/cgi-bin/show_thumbnails.pl

How do experimental biologists ensure reproducibility?

- Materials and methods section in publications.
- Lab notebook
- Annotated protocols published in specialized journals.
  - Methods in Molecular Biology
  - Nature Protocols
- A structured resource for protocols http://www.bio-protocol.org/
  - Example: [Bio101] A General EMSA (Gel-shift) Protocol [(http://www.bio-protocol.org/e24)](http://www.bio-protocol.org/e24).
- Standard lab equipment (tools).
- Standard lab environment (infrastructure).
- Sharing mutant strains (Morgan, ~ 1915).
- Genetic strains libraries (Nikolai Vavilov).



https://en.wikipedia.org/wiki/Lab_notebook

March 10th 1876

Fig. 1.

see you" — To my delight he came and declared
that he had heard and understood what I said.
I asked him to repeat the words — He ~~[illegible]~~
He answered "You said 'Mr. Watson — come here

Methods in Molecular Biology

Protocol | 20 October 2016

H
c
Id
Th
en
ho
Ha
po

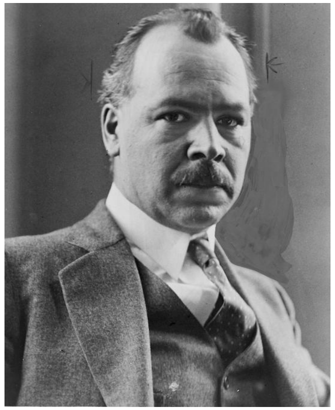
bio-protocol

Protocol by Field

- Biochemistry
- Cancer Biology
- Cell Biology
- Developmental Biology
- Immunology
- Microbiology
- Molecular Biology
- Neuroscience
- Plant Science
- Stem Cell
- Systems Biology

Protocol by Organism

- Arthropods
- Bacteria
- Fish
- Fungi
- Mammalia
- Other Vertebrates
- Plants
- Protozoans
- Viruses
- Worms



https://en.wikipedia.org/wiki/Nikolai_Vavilov

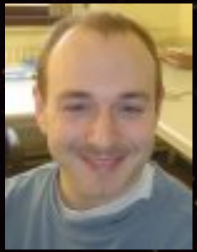
https://en.wikipedia.org/wiki/Thomas_Hunt_Morgan

RSAT TREK

Jacques van Helden

Julio Collado-Vides

Rekin’ s Janky

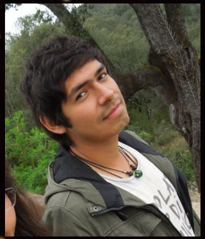

Sylvain Brohée

Matthieu Defrance

Carl Herrmann

Lionel Spinelli

Alejandra
Medina-Rivera


Jaime
Castro Mondragon

STARRING

BY ORDER OF APPEARANCE

JACQUES VAN HELDEN
JULIO COLLADO-VIDES
BRUNO ANDRE
OLIVIER SAND
REKIN'S JANKY
JEAN-VALERY TURATSINZE
MORGANE THOMAS-CHOLLIER
SYLVAIN BROHEE
ERIC VERVISCH
MATTHIEU DEFRANCE
KAROLINE FAUST
GIPSI LIMA-MENDEZ
ALEJANDRA MEDINA-RIVERA
CARL HERRMANN
DENIS THIEFFRY
JEREMY DELERCE
GAIME CASTRO-MONTRAGON
MARIE ARTUFEL
LUCIE KHAMVONGSA
SéBASTIEN JAEGER
CLAIRE RIOUALEN
BRUNO CONTRERAS-MOREIRA

HTTP://RSAT.EU/

Olivier Sand

Morgane
Thomas-Chollier

Jean Valéry
Turatsinze

Eric Vervisch

Karoline Faust

Gipsi Lima

Denis Thieffry

Bruno
Contreras-Moreira