A path to moderate the reproducibility crisis in Bioinformatics

Raffaele A Calogero, Francesca Cordero & Marco Beccuti

08/03/2019

Repeatability versus Reproducibility

- **What is repeatability?**
  - Repeatability practices were introduced by scientists Bland and Altman. For repeatability to be established, the following conditions must be in place:
    - the same location; the same measurement procedure; the same observer; the same measuring instrument, used under the same conditions; and repetition over a short period of time.
- **What is reproducibility?**
  - Reproducibility refers to the degree of agreement between the results of experiments conducted by different individuals, at different locations, with different instruments.
  - Reproducibility measures our ability to replicate the findings of others.

Reproducibility crisis

Most scientists 'can't replicate studies by their peers'

BBC NEWS

By Tom Feilden
Science correspondent, Today programme

22 February 2017 Share

Scientists attempting to repeat findings reported in five landmark cancer studies confirmed only two

Figure 1:

Reproducibility crisis

- The "reproducibility crisis" is the name given to the situation that a large percentage of the academic literature is not reproducible.
- The history of the reproducibility crisis is that in 2011, Glenn Begley, who ran the oncology division at Amgen, decided to try to reproduce 53 foundational papers in oncology.
  - He was unable to reproduce 47 of them, which is 89%.
- Bayer, pharmaceutical company, reported in the same year that it was unable to reproduce 65% of the papers in its sample of the biomedical literature.
- Reproducibility has also found to be an issue in psychology and computer science.

https://www.forbes.com/sites/quora/2017/02/09/how-the-reproducibility-crisis-in-academia-is-affecting-scientific-research/#24f9c4003dad

Richard Price, Founder of Academia.edu

Figure 2:

Keith A. Baggerly: "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics."

Using the NCI60 to Predict Sensitivity

ture.com/naturemedicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

Potti et al (2006), Nature Medicine, 12:1294-1300.

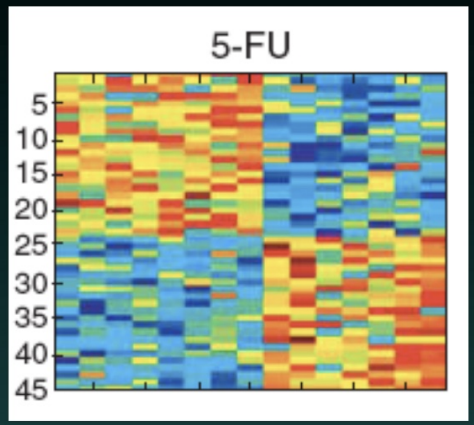
The main conclusion is that we can use microarray data from cell lines (the NCI60) to define drug response "signatures", which can be used to predict whether patients will respond.

They provide examples using 7 commonly used agents.

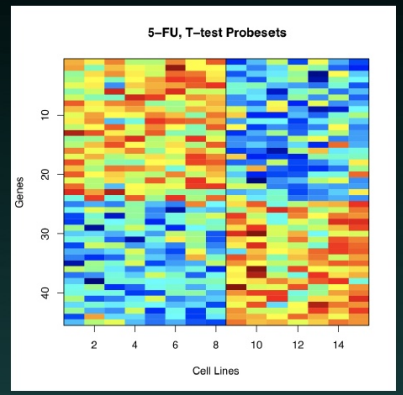
This got people at MDA very excited.

Keith A. Baggerly: "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics."

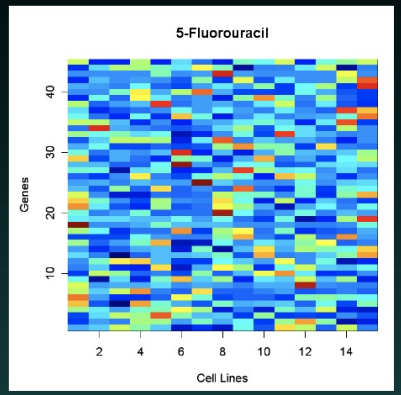
5-FU Heatmaps



Nat Med Paper

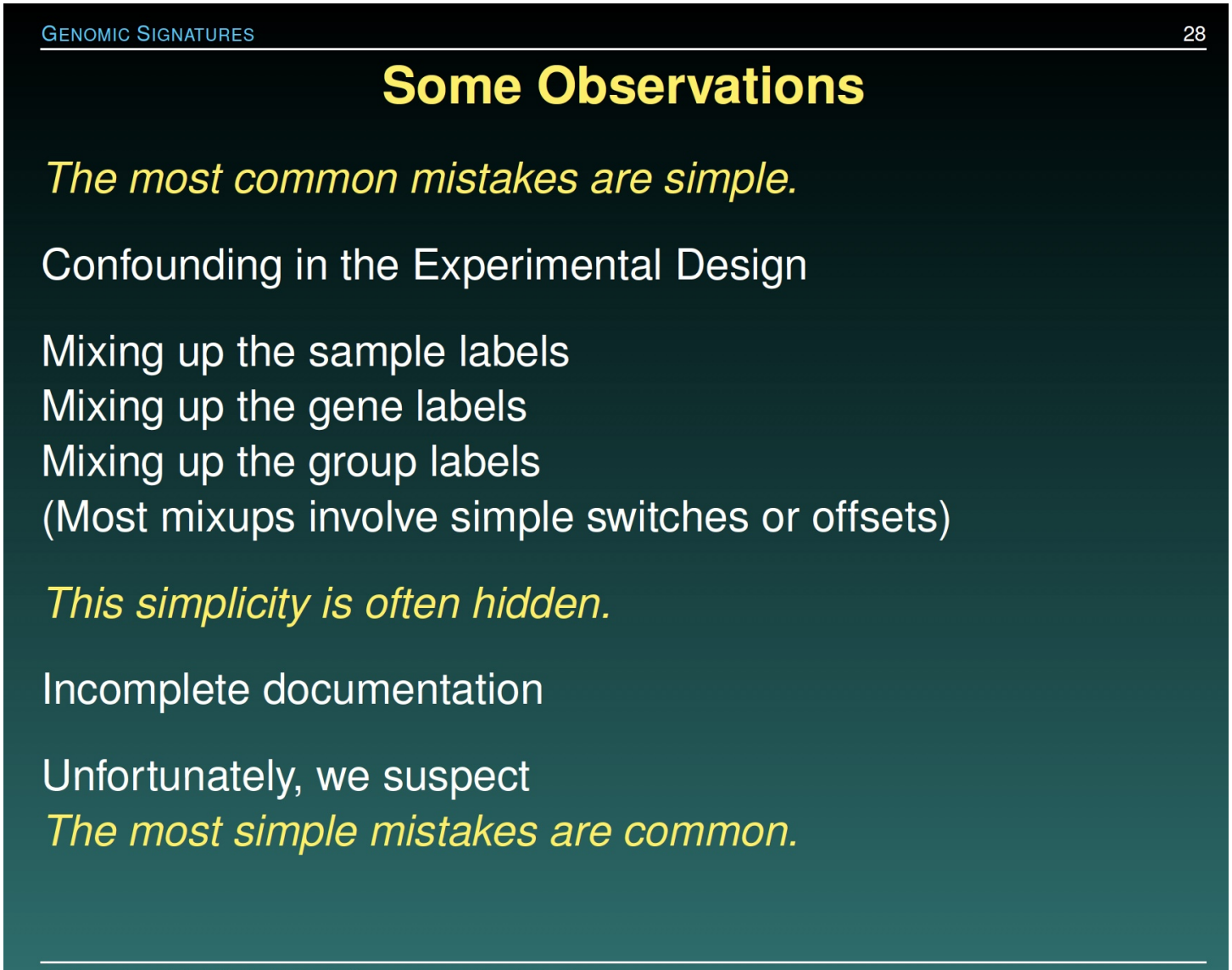


Our t-tests



Reported Genes

Keith A. Baggerly: "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics."



Ten Simple Rules for Reproducible Computational Research

1. For Every Result, Keep Track of How It Was Produced
2. **Avoid Manual Data Manipulation Steps**
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Sandve et al. PLoS Comp Biol. 2013

Figure 6: Basic elements in reproducible bioinformatics

Sadve's rule 3: Archive exact versions of all external program used

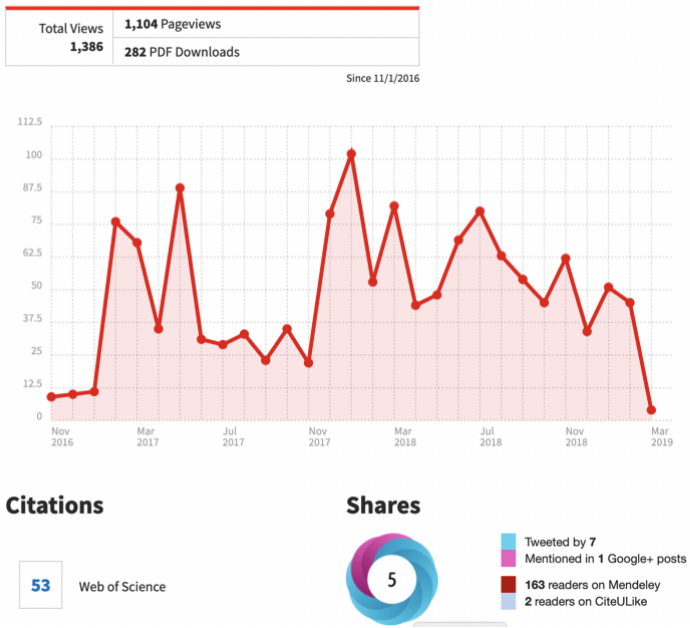
EXPANDS: expanding ploidy and allele frequency on nested subpopulations

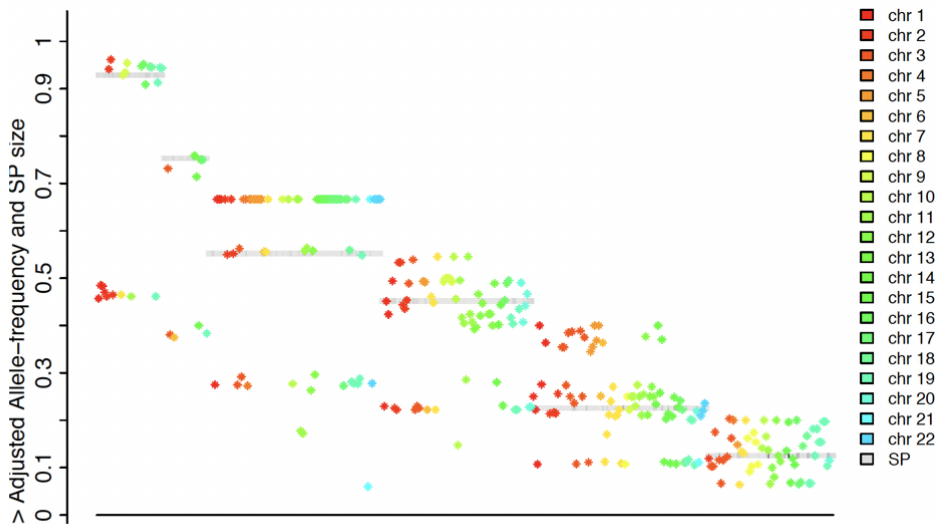
Noemi Andor, Julie V. Harness, Sabine Müller, Hans W. Mewes, Claudia Petritsch

Bioinformatics, Volume 30, Issue 1, 1 January 2014, Pages 50–60, https://doi.org/10.1093/bioinformatics/btt622

Published: 30 October 2013 **Article history** ▾

Metrics





```
library(expands)
c17 <-runExPANdS(SNV=as.matrix(snv), CBS=as.matrix(cnv),max_PM = 6,maxScore = 3)
…..
[1] "... Done."
[1] "tree saved under ./out.expands.tree»
….
[1] "Subpopulation specific point mutations  saved under  ./out.expands.sps"
[1] "Summary file of detected subpopulations  saved under  ./out.expands.spstats"

sessionInfo()
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] expands_1.7.2

loaded via a namespace (and not attached):
 [1] modeltools_0.2-21 mclust_5.2        nnet_7.3-9        matlab_1.0.2
 [5] flexmix_2.3-13    nlme_3.1-126      ape_3.4           grid_3.3.0
 [9] permute_0.9-0     moments_0.14      stats4_3.3.0      rJava_0.9-8
[13] lattice_0.20-33
```

SP	PM_B	SP_cnv	PM	PM_cnv	scenario	SP_0.245	SP_0.332	SP_0.39	SP_0.825	Clone
0.24502349	2	0.24502349	3	3	3	1	0	0	0	0.24502349
NA	1	NA	1	1	3	0	0	0	0	NA
NA	1	NA	1	1	3	0	0	0	0	NA
0.33203758	2	0.33203758	2	2	3	1	1	0	0	0.087
0.39004697	3	0.39004697	3	3	3	1	0	1	0	0.145
0.24502349	1	0.24502349	2	2	3	1	0	0	0	0.24502349

Figure 8: Packages versions before expands 1.7.2 release (2016-04-04)

```
library(expands)
c17 <-runExPANdS(SNV=as.matrix(snv), CBS=as.matrix(cnv),max_PM = 6,maxScore = 3)
…..
[1] "... Done.»
[1] "distance-matrix saved under ./out.expands.dist»
….
[1] "tree saved under ./out.expands.tree»
[1] "Subpopulation specific point mutations  saved under  ./out.expands.sps"
[1] "Summary file of detected subpopulations  saved under  ./out.expands.spstats"

sessionInfo()
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] expands_1.7.2

loaded via a namespace (and not attached):
 [1] compiler_3.5.1    modeltools_0.2-22 mclust_5.4.2      parallel_3.5.1
 [5] nnet_7.3-12       matlab_1.0.2      Rcpp_1.0.0        flexmix_2.3-14
 [9] nlme_3.1-137      ape_5.2           grid_3.5.1        permute_0.9-4
[13] moments_0.14      stats4_3.5.1      rJava_0.9-10      lattice_0.20-38
```

SP	PM_B	SP_cnv	PM	PM_cnv	scenario
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA

Figure 9: Latest version of packages

Reproducibility Projects

- **Finding issues**:
  - Investigating reproducibility in preclinical cancer research
  - Estimating the Reproducibility of Psychological Science
- **Fixing issues**:
  - Bioconductor
    - provides limited framework for complex pipelines
  - Reproducible Bioinformatics project
    - guarantee functional (i.e. information about data and the utilized tools are saved in terms of meta-data) and computational reproducibility (i.e. real image of the computation environment used to generate the data is stored)

Reproducible research in Bioinformatics

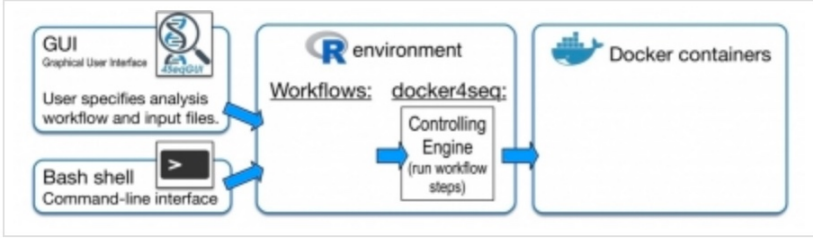
A project to provide reproducible results in Bioinformatics using Docker images

Home

Contact

Bx2M
Via Nizza 52
10126 c/o B&Gu@MBC Torino
Tel: +39 0116706454
info@reproducibile-bioinformatics.org

Welcome to the Reproducible Bioinformatics project



The aim of Reproducible Bioinformatics project is the creation of easy to use Bioinformatics workflows that fullfill the following roles (Sandve et al. PLoS Comp Biol. 2013):

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Figure 10: Kulkarni et al. BMC Bioinformatics 2018

Reproducible research in Bioinformatics

Reproducible Bioinformatics Project

(providing reproducible results using Docker images)

- RBP is a community open to anyone interested to share workflows under the umbrella of reproducibility.
- Our goal is to enable easy access to NGS data analysis pipelines for non-bioinformatics experts on any computing environment, whether a laboratory workstation, university computer cluster, or a cloud service provider.

www.reproducible-bioinformatics.org/

Figure 11:

Figure 12: Why using dockers' containers?

Docker is a shipping container for the code

Figure 13: Why using dockers' containers?

Reproducible research in Bioinformatics

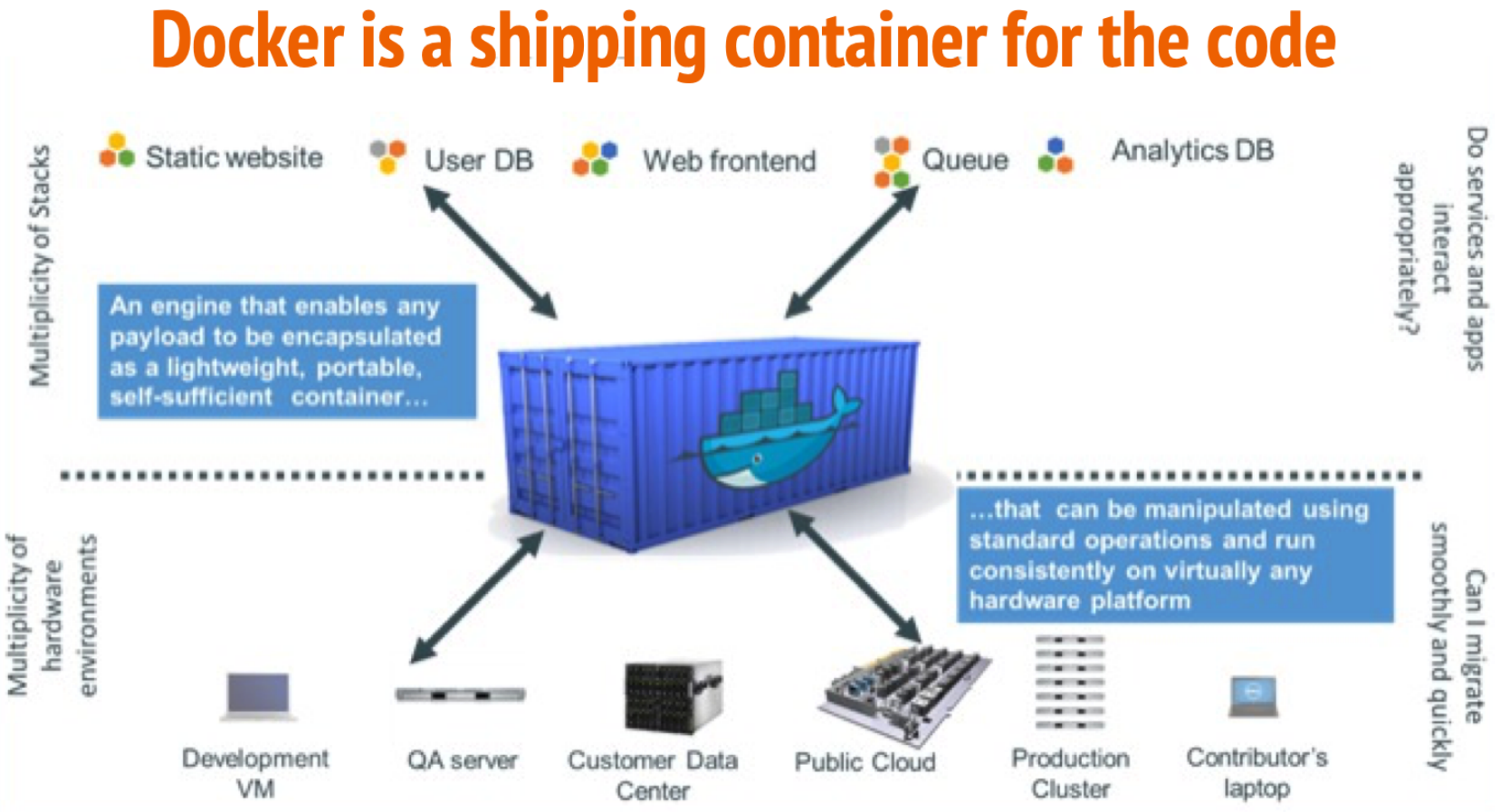
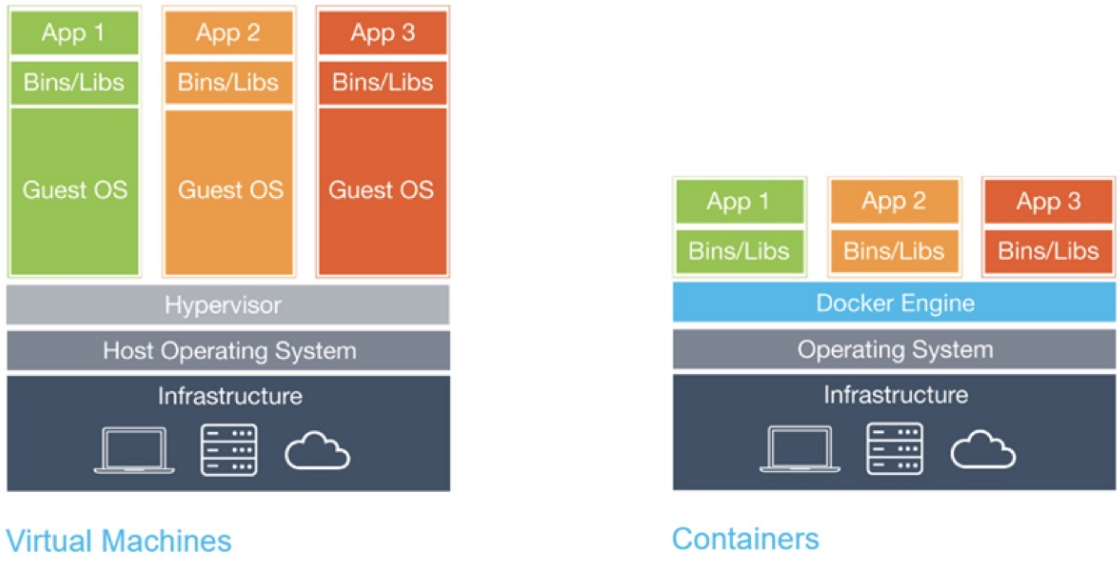


Figure 14: Why using dockers' containers?

Reproducible research in Bioinformatics



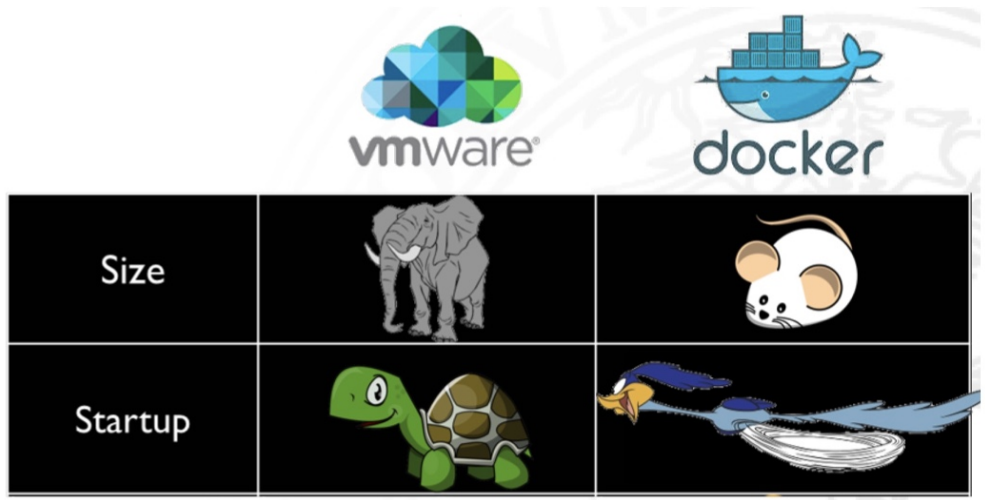
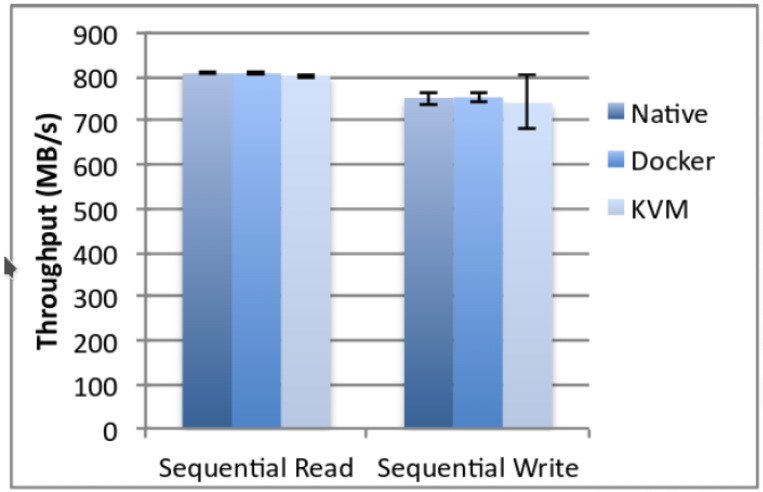
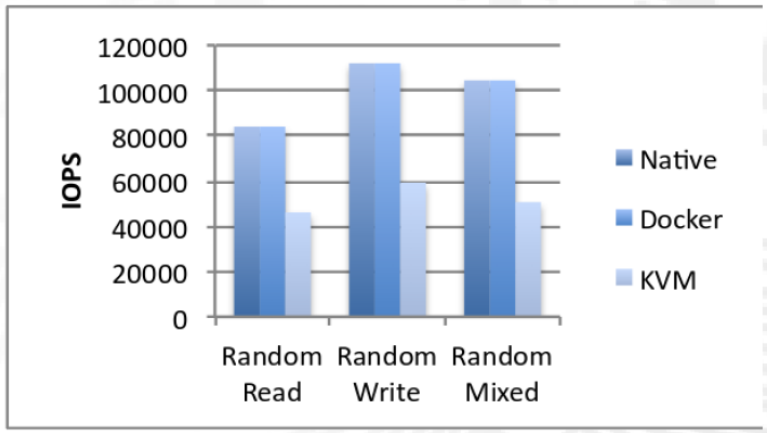


Figure 15: Virtual machine versus docker containers

Reproducible research in Bioinformatics

In [1] a comparison among physical server, KVM, and Docker is reported.





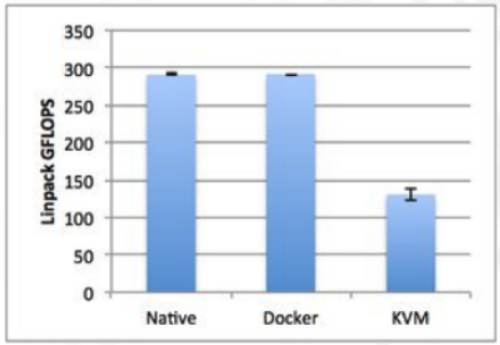


Figure 1. Linpack performance on two sockets (16 cores). Each data point is the arithmetic mean obtained from ten runs. Error bars indicate the standard deviation obtained over all runs.

[1] W. Felter, A. Ferreira, R. Rajamony and J. Rubio, ***An updated performance comparison of virtual machines and Linux containers***, 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Philadelphia, PA, 2015, pp. 171-172.

Figure 16: Docker container, VM and real server: a comparison

Reproducible research ecosystem

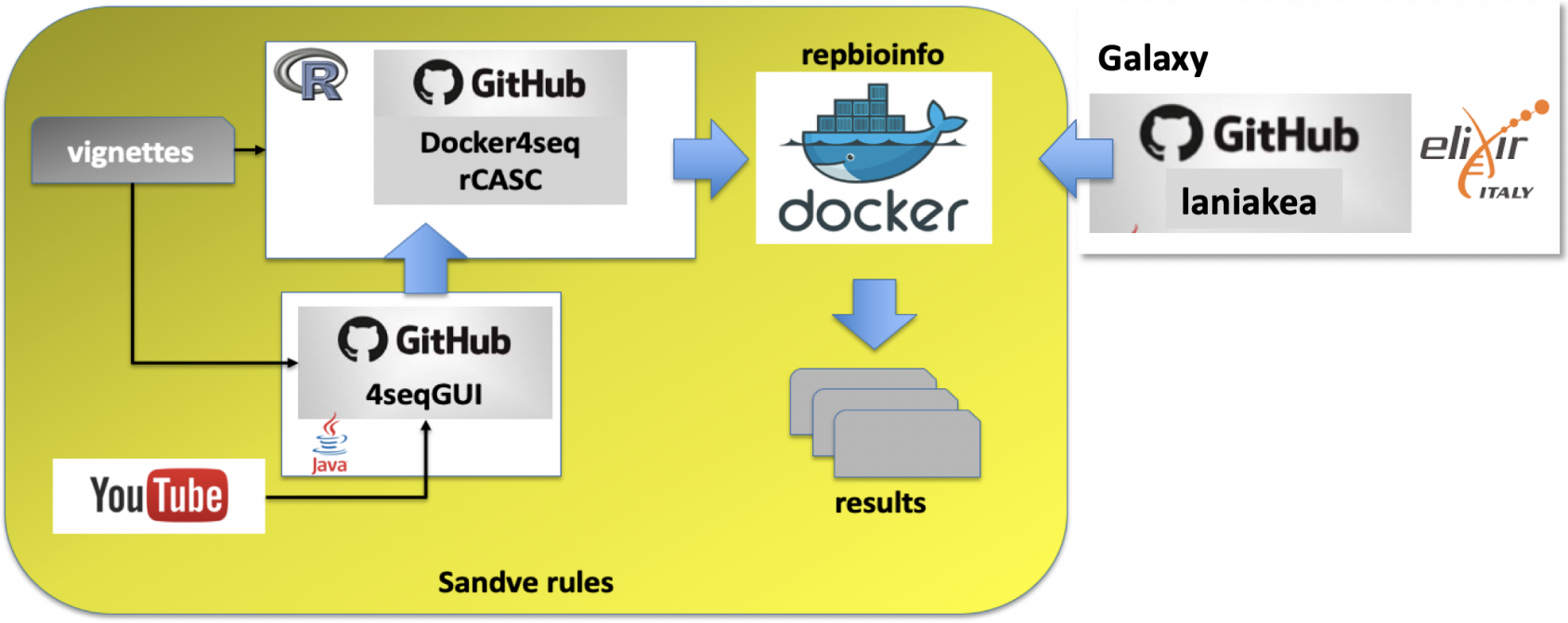


Figure 17:

Reproducible research in Bioinformatics

Reproducible Bioinformatics Project

- Each workflow is made of a set of R functions controlling one or more dockers containers.
- Any computation requiring specific software must be embedded in a docker image.
- Any workflow should be supported by an explanatory vignette

Figure 18: s

Reproducible research in Bioinformatics

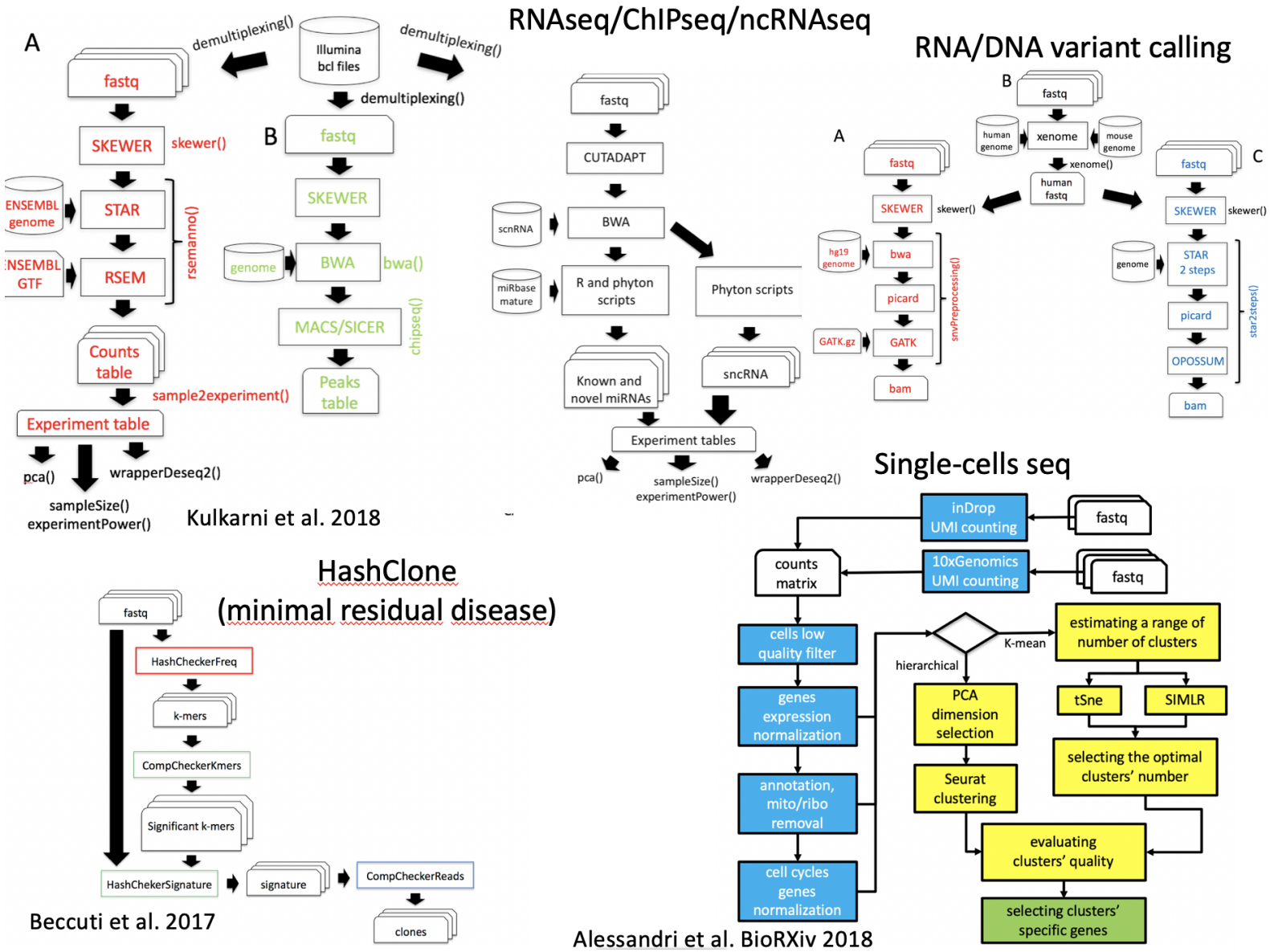


Figure 19: Implemented workflows

circRNA workflow

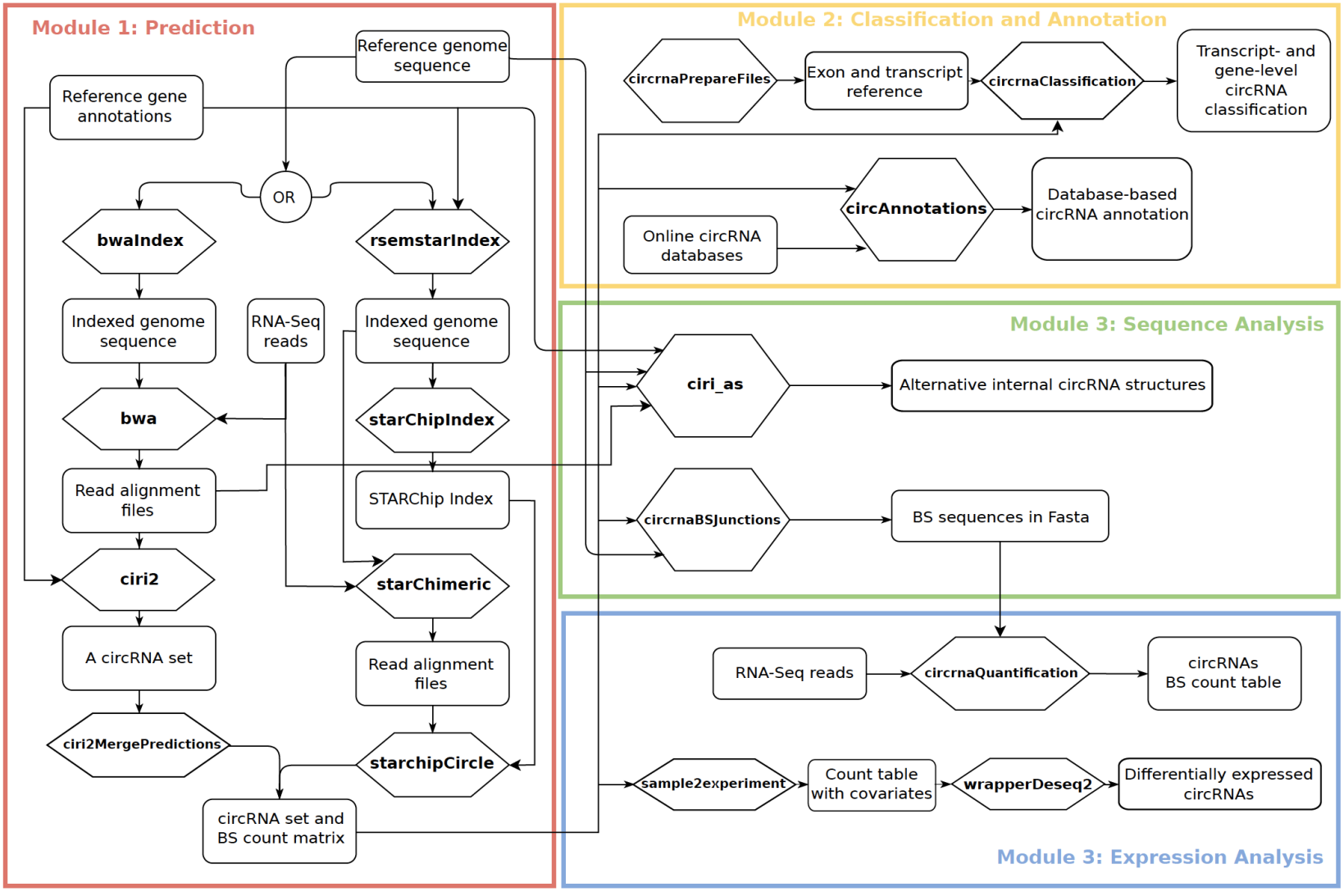


Figure 20:

Docker container nomenclature

- **docker.io/repbioinfo/name.YYYY.ZZ:**
  - the field ZZ will be updated in case of updates required to solve bugs, which do not affect the calculation.
  - The field YYYY will be updated in case of updates which affect the calculation (e.g. new release of Bioconductor libraries).
  - Docker image tag shows the version of the embedded softwares
    - docker.io/repbioinfo/bwa.2017.01:bwa-0.7.12_picard-tools-1.133_samtools-1.3.1_R-3.3.2_jdk-1.8.0_111

Figure 21:

Reproducible research in Bioinformatics

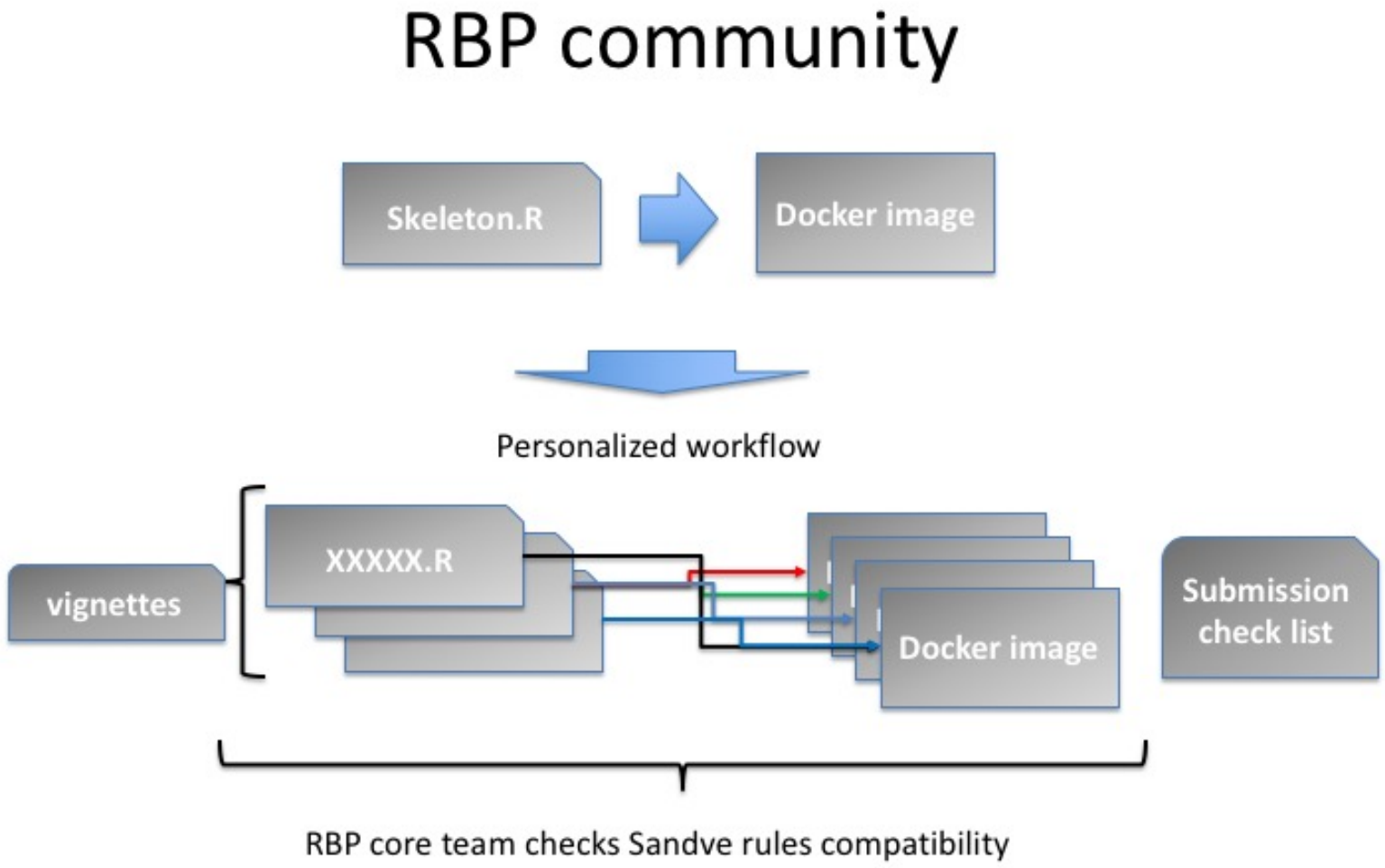


Figure 22:

Bringing back data analysis in the hands of Biologists

- The majority of the NGS experiments are very simple.
- The analysis of such experiments should be done by wet researchers
- How we can cope with:
  - Lack of experience in scripting?
    - **GUI (4SeqGUI)**
    - Training
  - Lack of high-end servers?
    - SeqBox (Beccuti et al. Bioinformatics 2017)

4SeqGUI

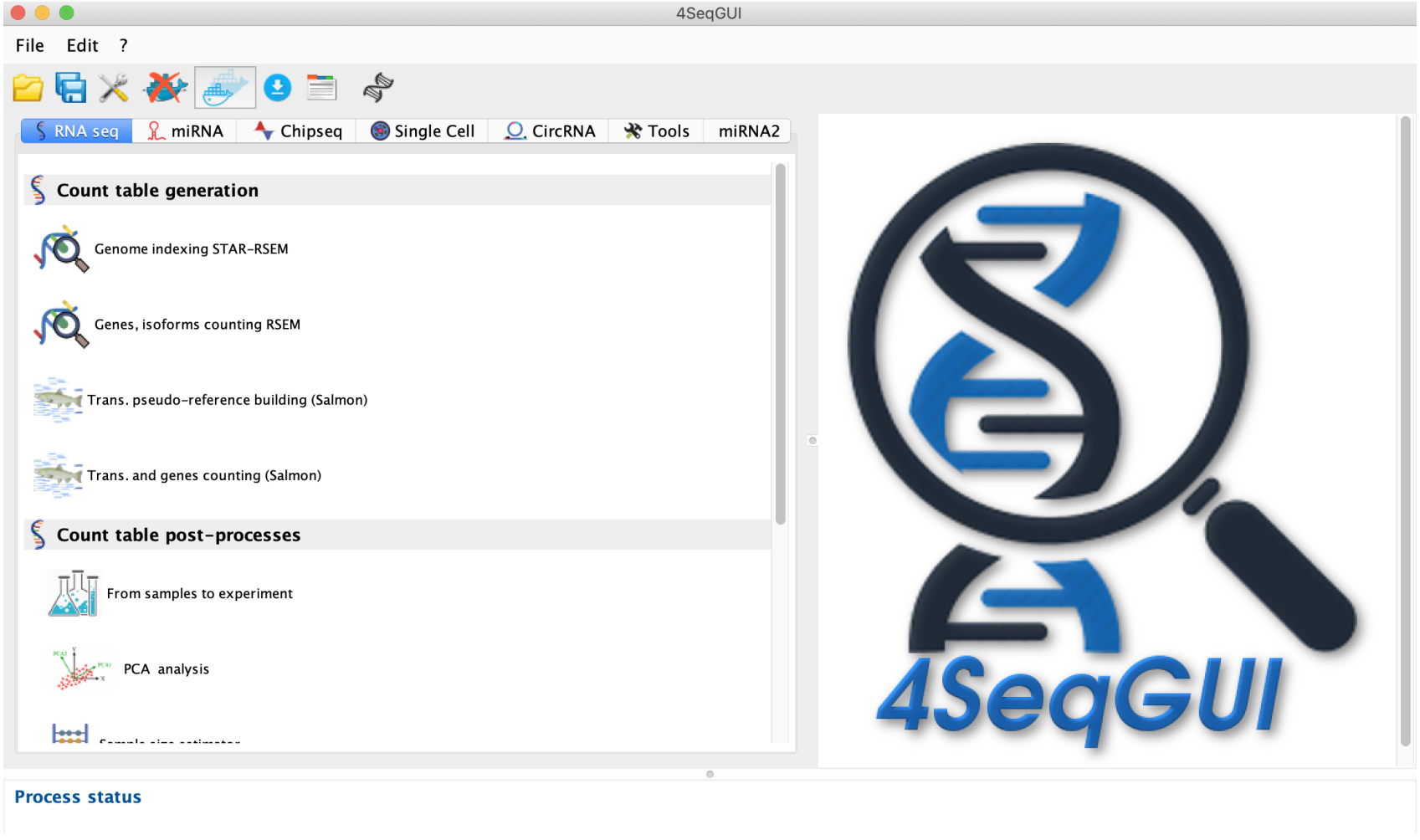


Figure 23:

Bringing back data analysis in the hands of Biologists

- The majority of the NGS experiments are very simple.
- The analysis of such experiments should be done by wet researchers
- How we can cope with:
  - Lack of experience in scripting?
    - GUI (4SeqGUI)
    - **Training**
  - Lack of high-end servers?
    - SeqBox (Beccuti et al. Bioinformatics 2017)

Bringing back data analysis in the hands of Biologists

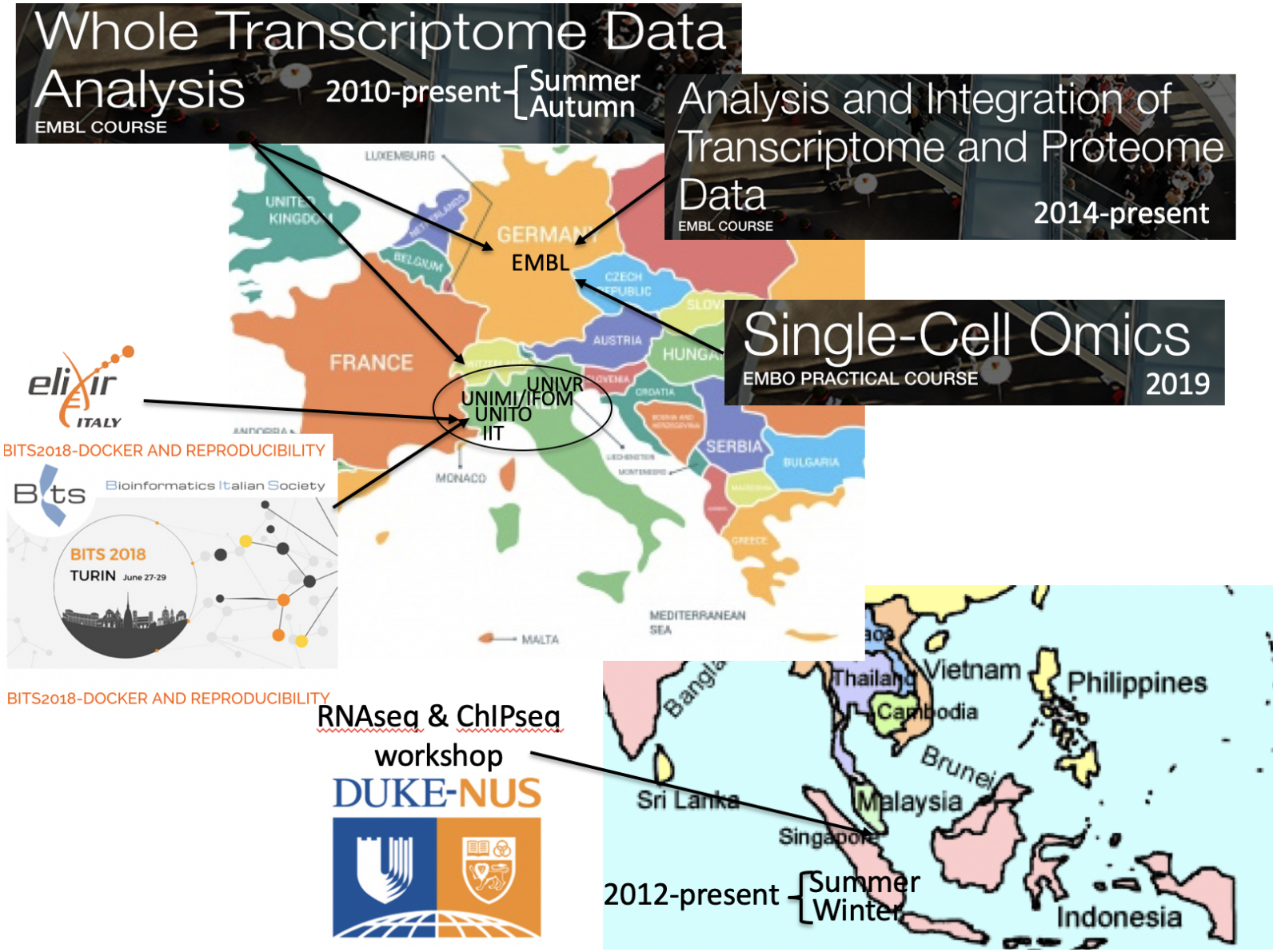


Figure 24: Training

Bringing back data analysis in the hands of Biologists

- The majority of the NGS experiments are very simple.
- The analysis of such experiments should be done by wet researchers
- How we can cope with:
  - Lack of experience in scripting?
    - GUI (4SeqGUI)
    - Training
  - Lack of high-end servers?
    - **SeqBox (Beccuti et al. Bioinformatics 2017)**

Bringing back data analysis in the hands of Biologists

Sequence analysis

SeqBox: RNAseq/ChIPseq reproducible analysis on a consumer game computer

Marco Beccuti[1,*], Francesca Cordero[1], Maddalena Arigoni[2], Riccardo Panero[2], Elvio G. Amparore[1], Susanna Donatelli[1] and Raffaele A. Calogero[2]

[1]Department of Computer Sciences, University of Torino, 10124 Turin, Italy and [2]Department of Molecular Biotechnology and Health Sciences, University of Torino, 10124 Turin, Italy

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on July 28, 2017; revised on October 12, 2017; editorial decision on October 17, 2017; accepted on October 19, 2017

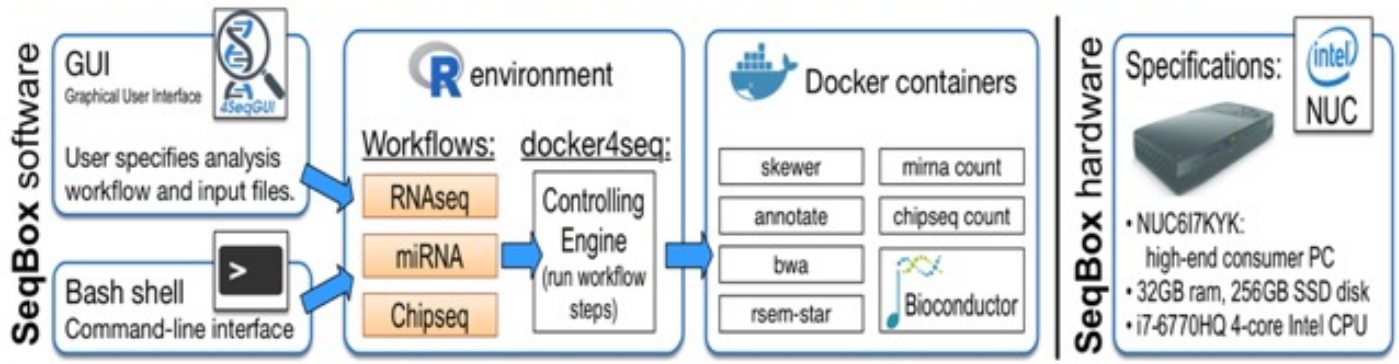


Figure 25: SeqBox

SeqBox

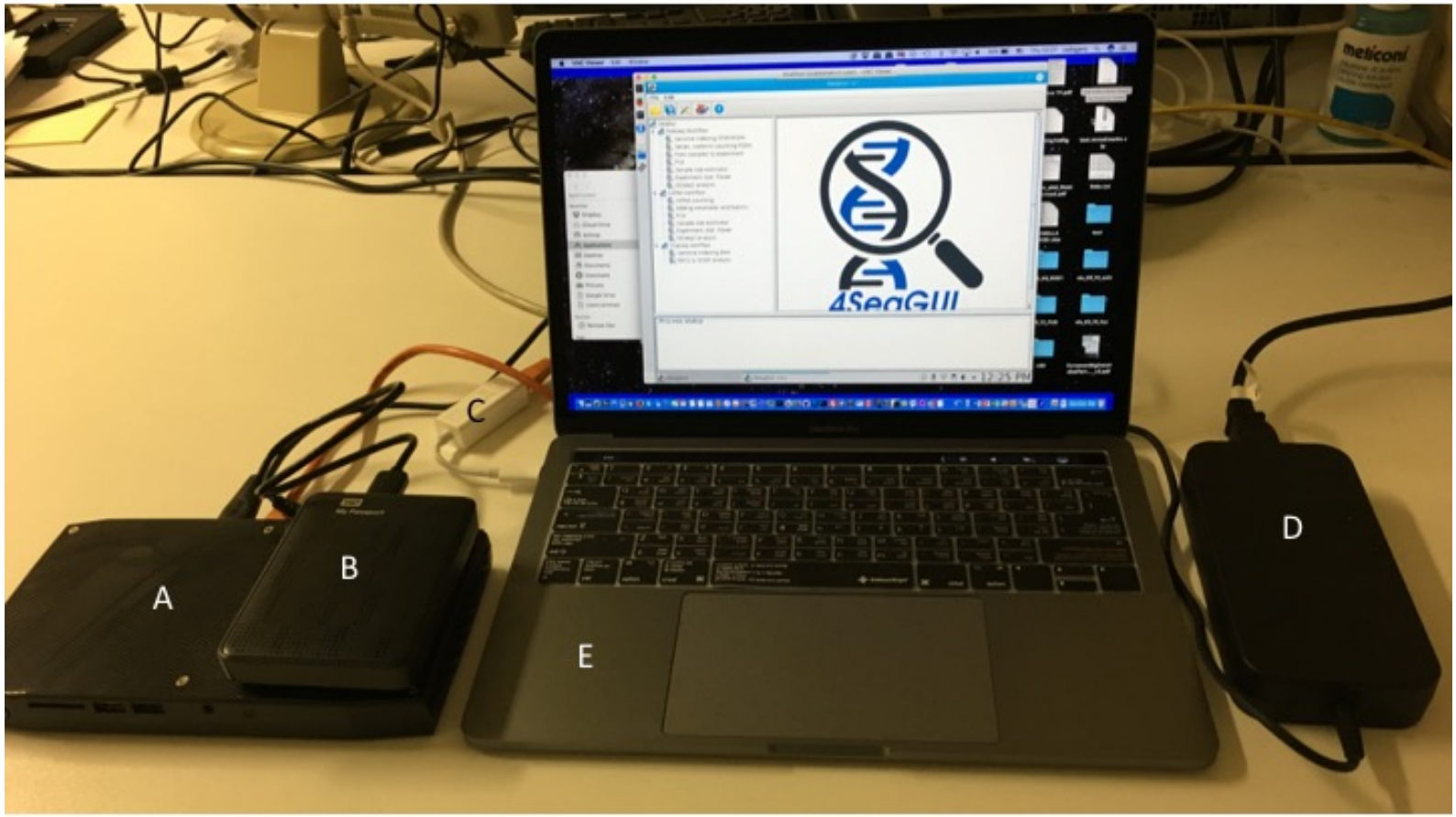
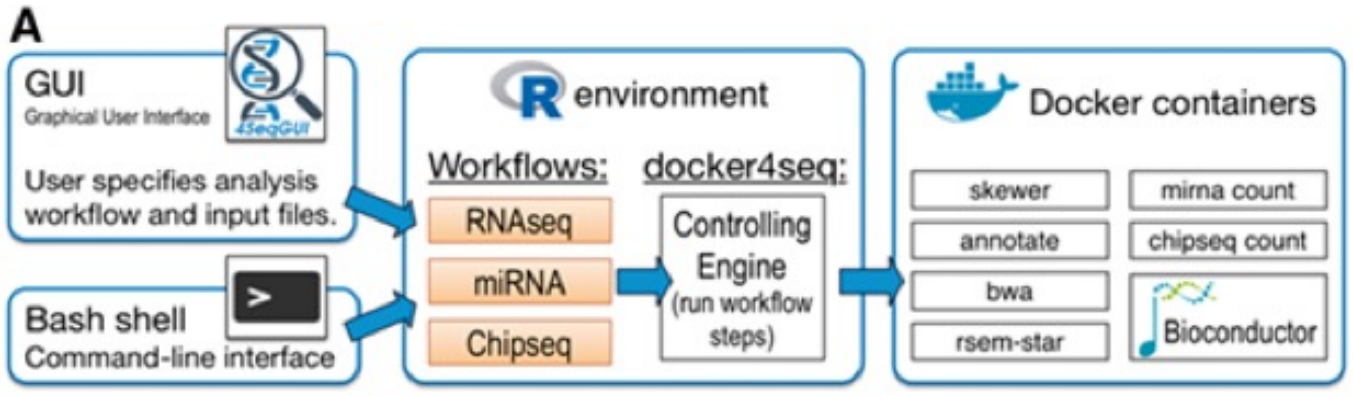


Figure 26: Processing 450 milion di reads, standard NextSeq500 throughout, in 20 hours

SeqBox

RBP workflows for the biological/medical community



B

	N° of CPU cores	CPU type	Clock Speed	Total RAM	Storage type	Hardware type	Software configuration
SeqBox	8	i7-6770HQ	3.5 GHz	32 GB	SSD	Mini-pc (1)	Fedora
SGI UV2000	16	E5-4650	2.4 GHz	1024 GB	SATA	Server (10)	Red Hat

Beccuti et al. Bioinformatics 2017

Figure 27:

SeqBox

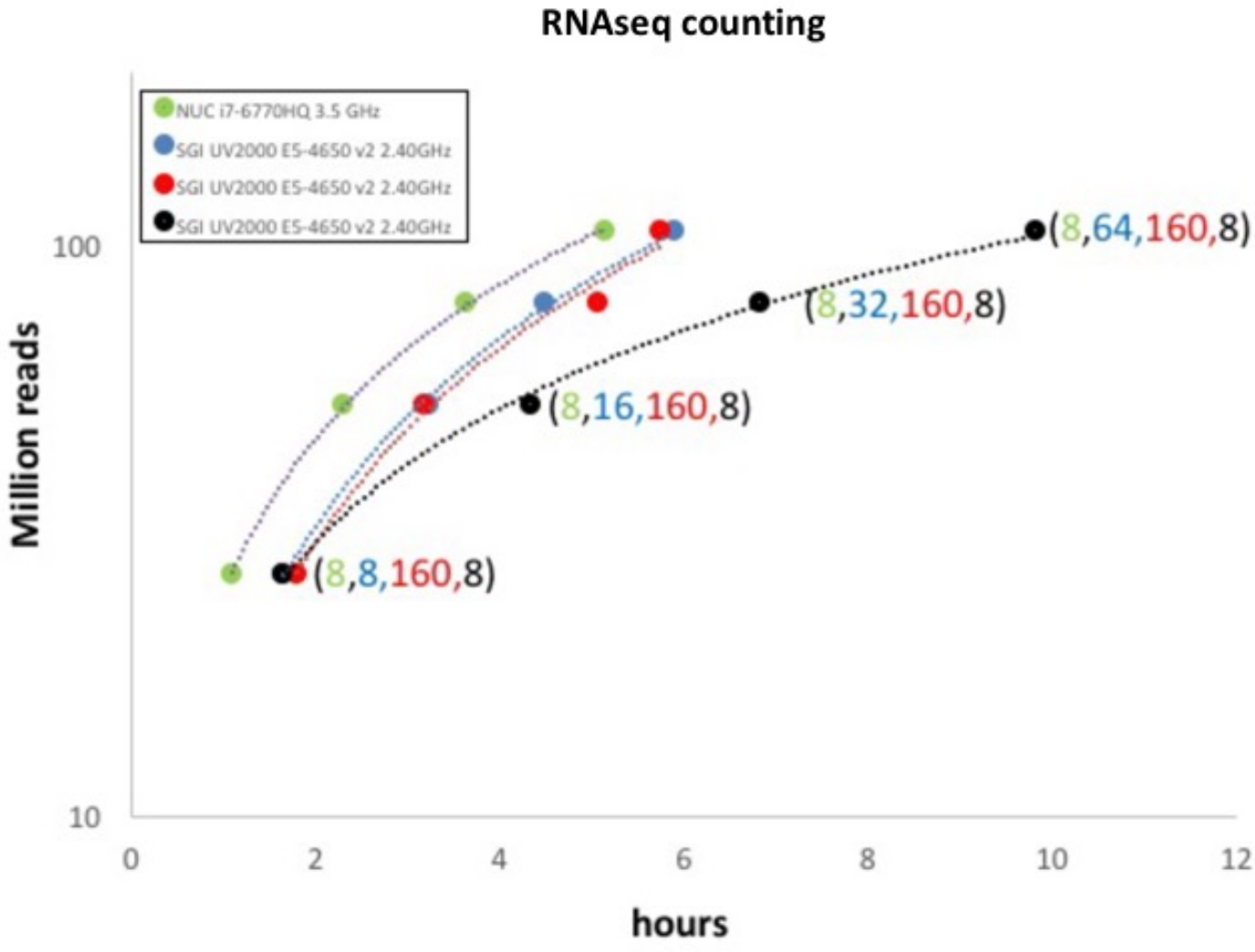


Figure 28: Counting RNAseq data

Conclusions

- Reproducibility is an important topic in many research fields
- reproducible-bioinformatics project is a community providing an ecosystem for developing life science reproducible workflows
- Docker technology combined to Sandve's rules for good bioinformatics practice guarantee functional and computational reproducibility within the reproducible-bioinformatics project ecosystem
- reproducible-bioinformatics project provides an infrastructure to allow access to complex workflows to life science researchers lacking computation skills

Thank you

Figure 29: