YAMP: a framework enabling reproducibility in metagenomics research

Alessia Visconti

🐦 @_alesssia

Reproducibility

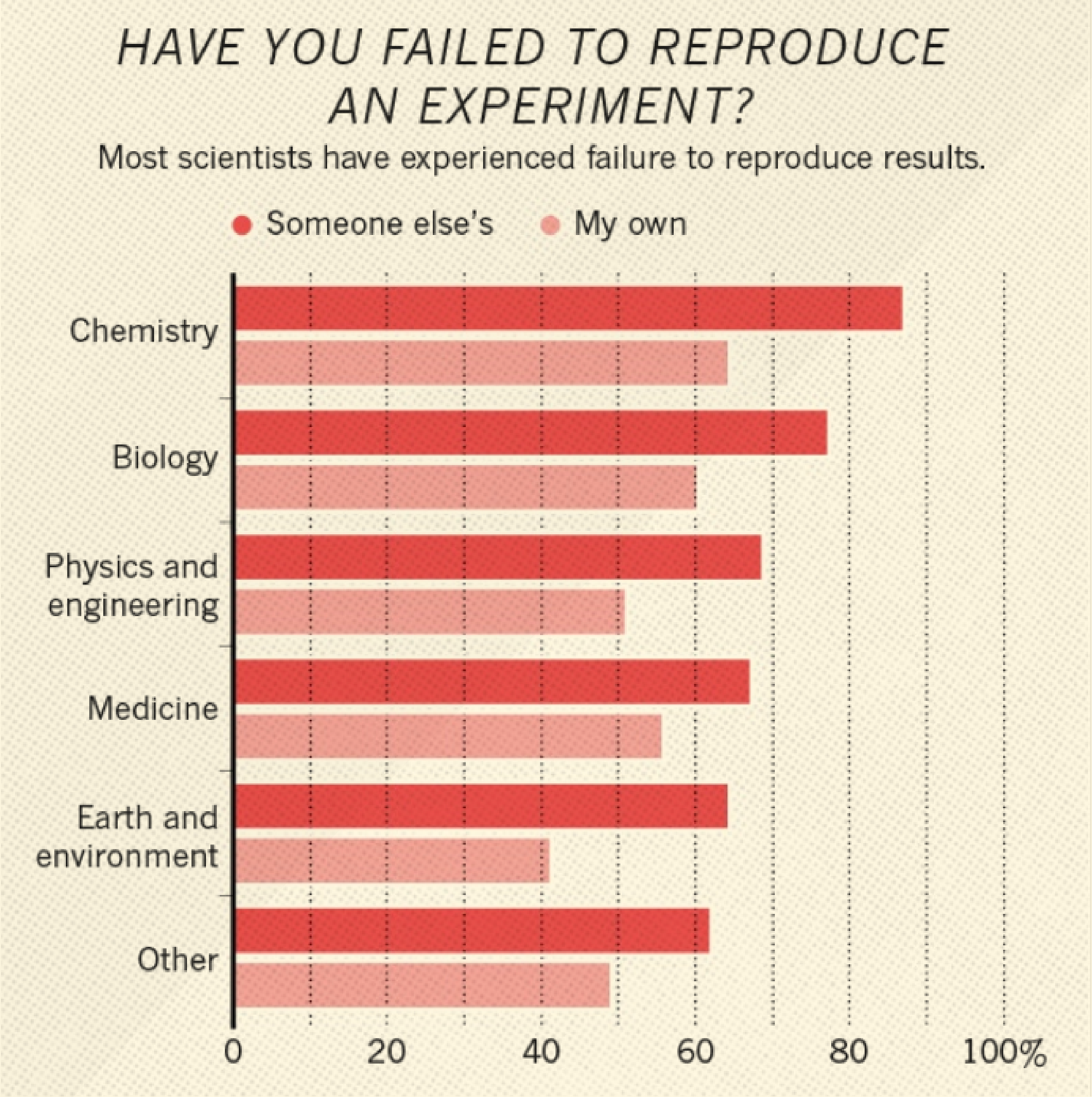
>50%

Repeatability

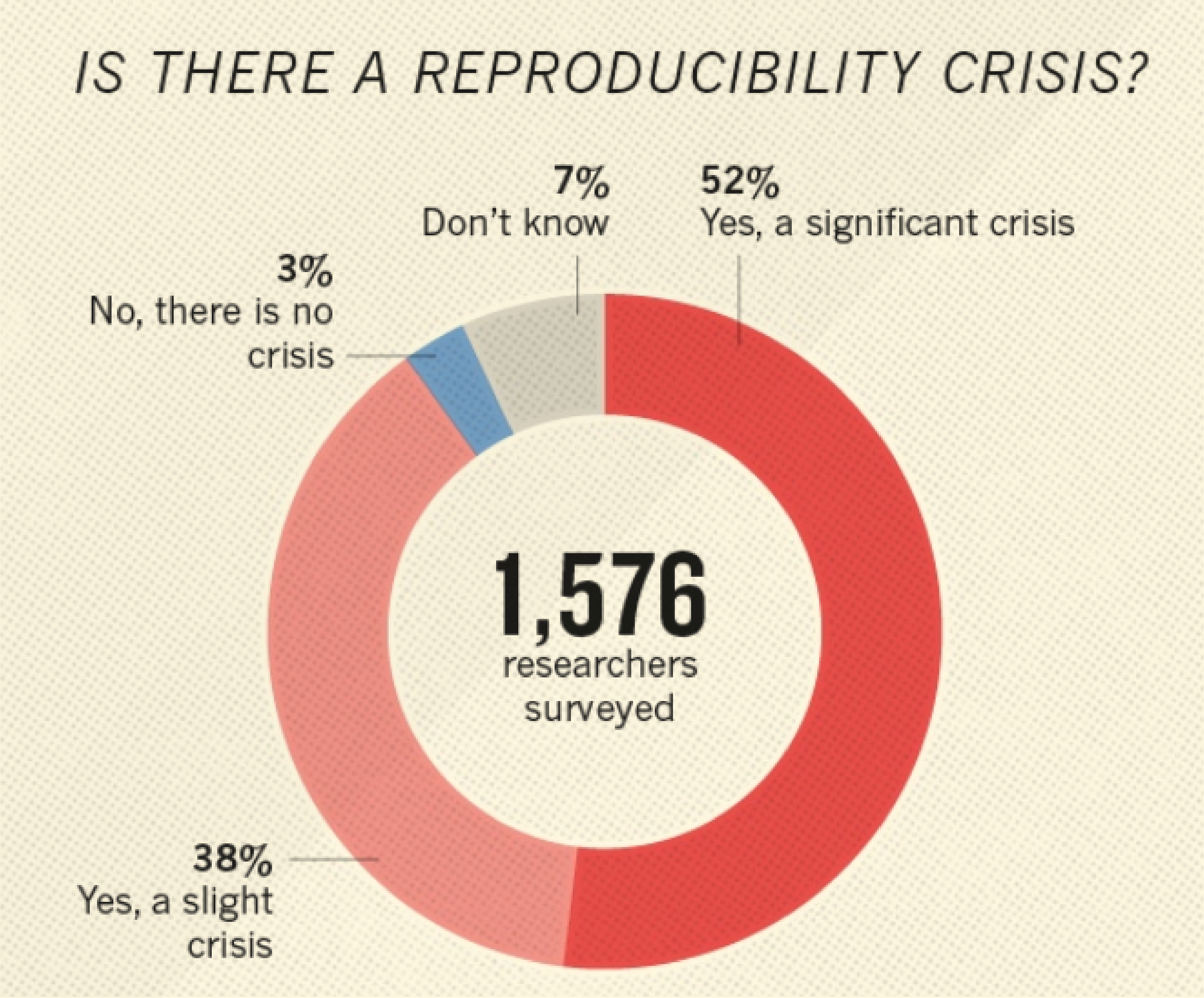
>70%

Reproducibility

Baker, Nature (2016)



Baker, Nature (2016)

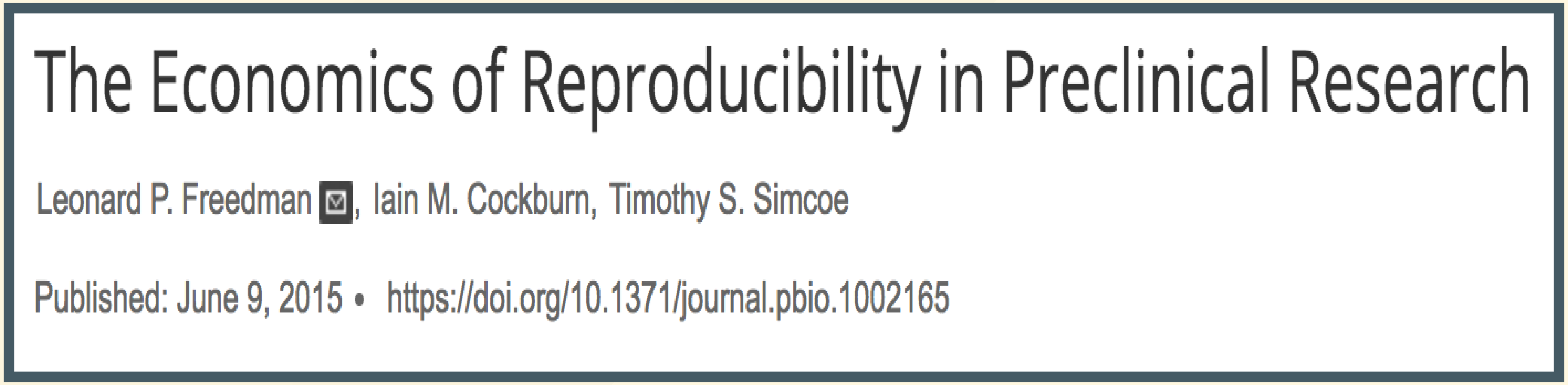


Baker, Nature (2016)

~31%

not-reproducible → non-trustable

Baker, Nature (2016)



The Economics of Reproducibility in Preclinical Research

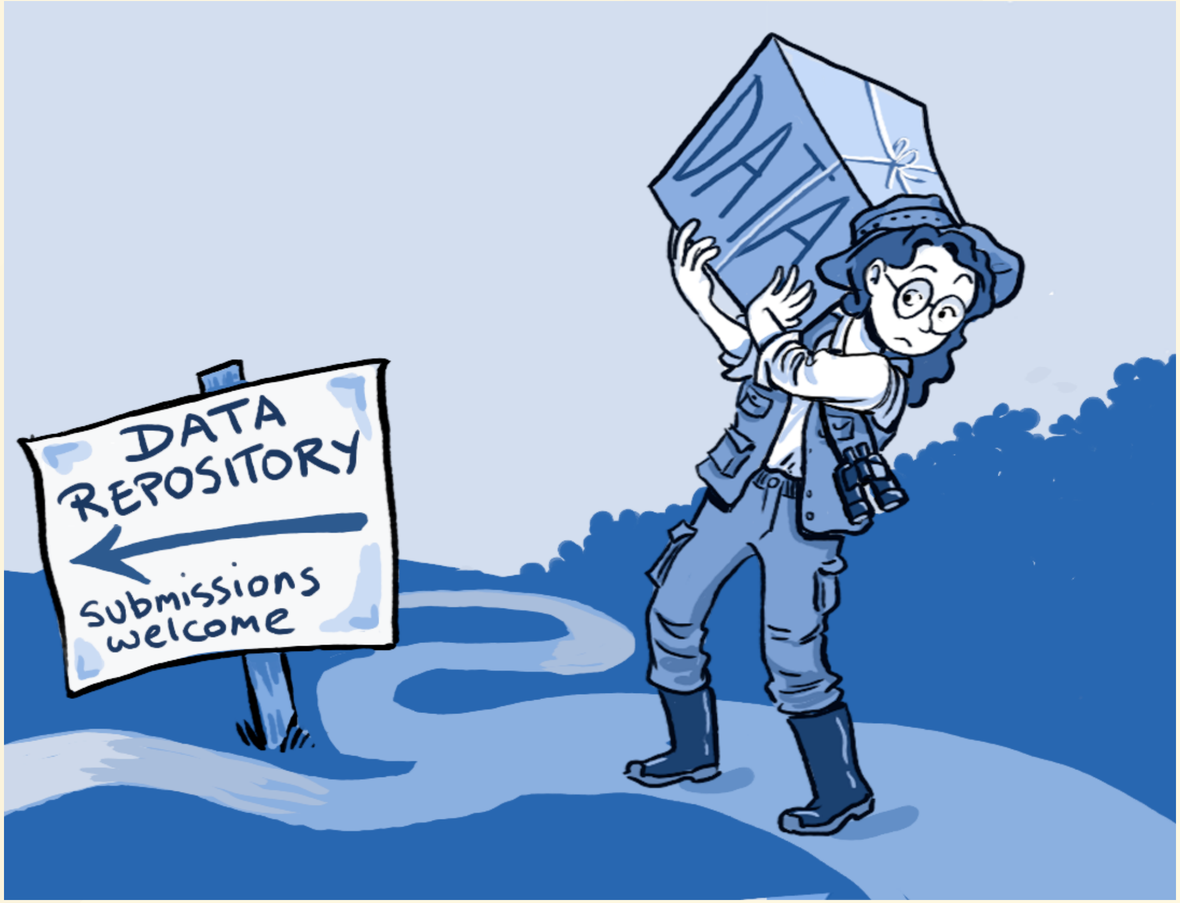
Leonard P. Freedman, Iain M. Cockburn, Timothy S. Simcoe

Published: June 9, 2015 • https://doi.org/10.1371/journal.pbio.1002165

What cause irreproducible research?

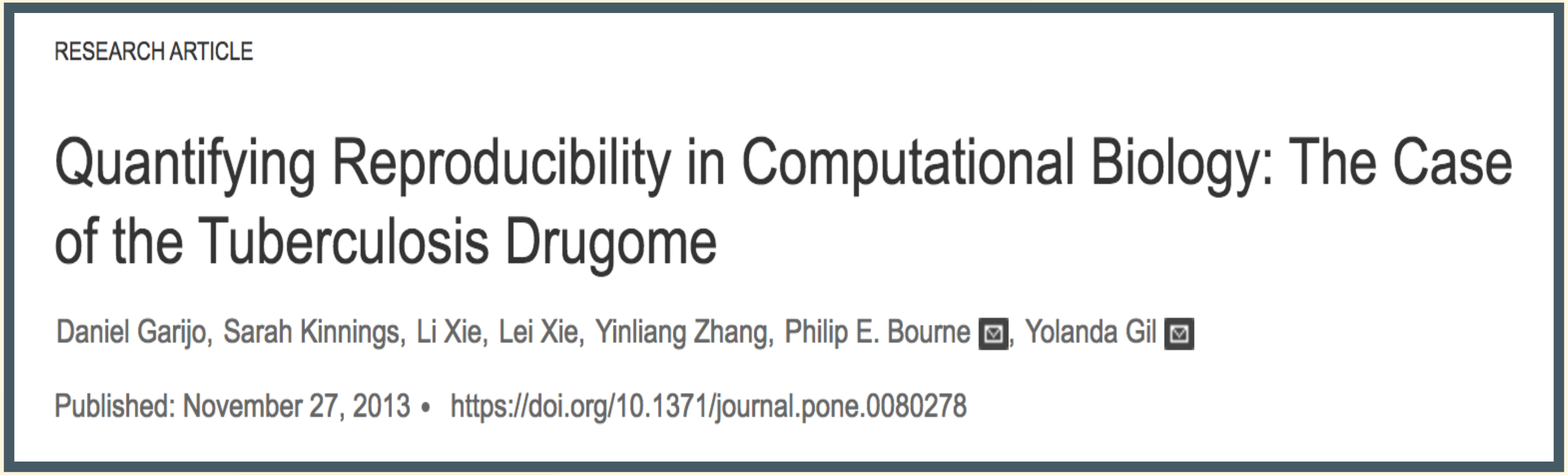
1. Unavailability of primary data

Solution: data repository



Data Sharing - Wikipedia

2. Unavailability of sufficient details on computational experimentation



RESEARCH ARTICLE

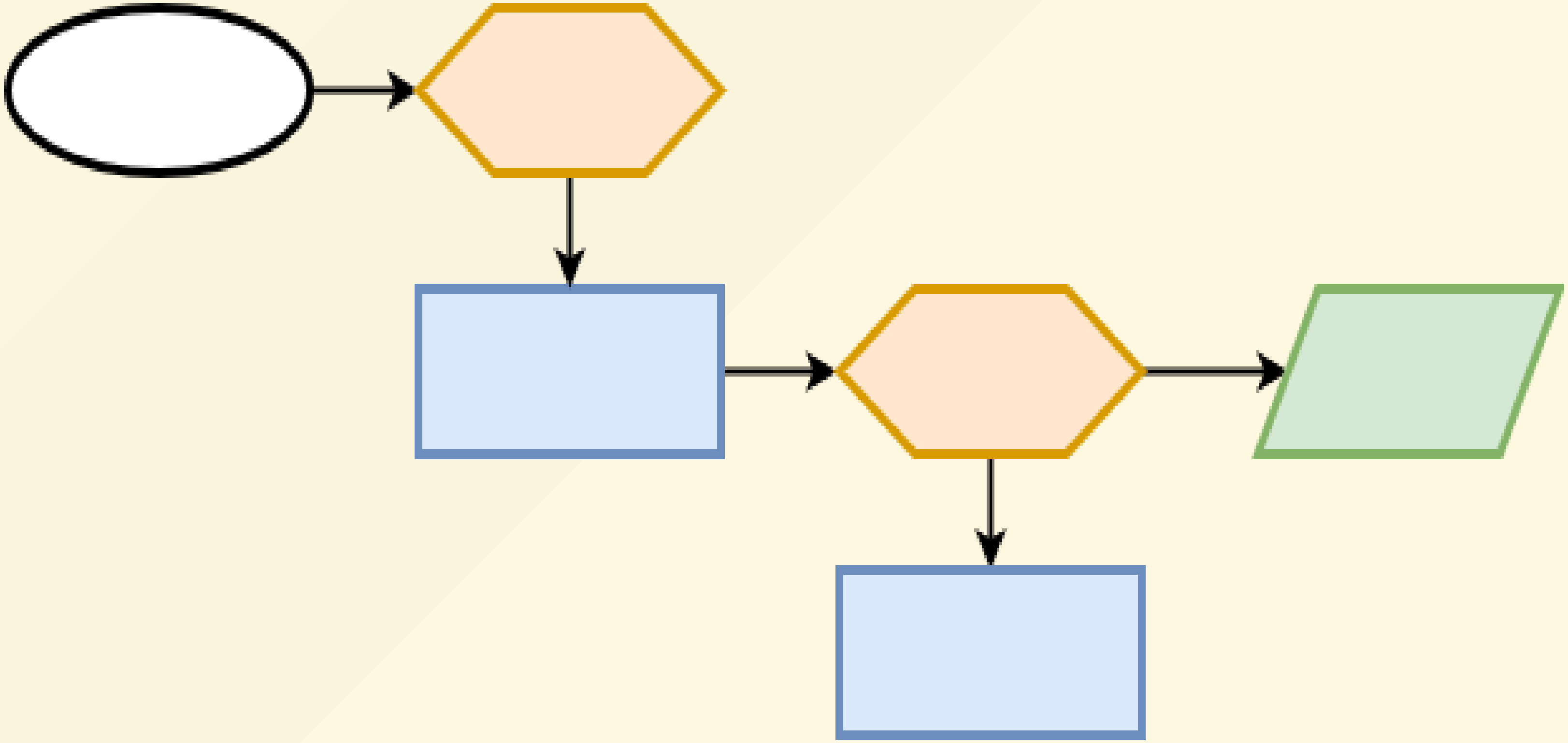
Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome

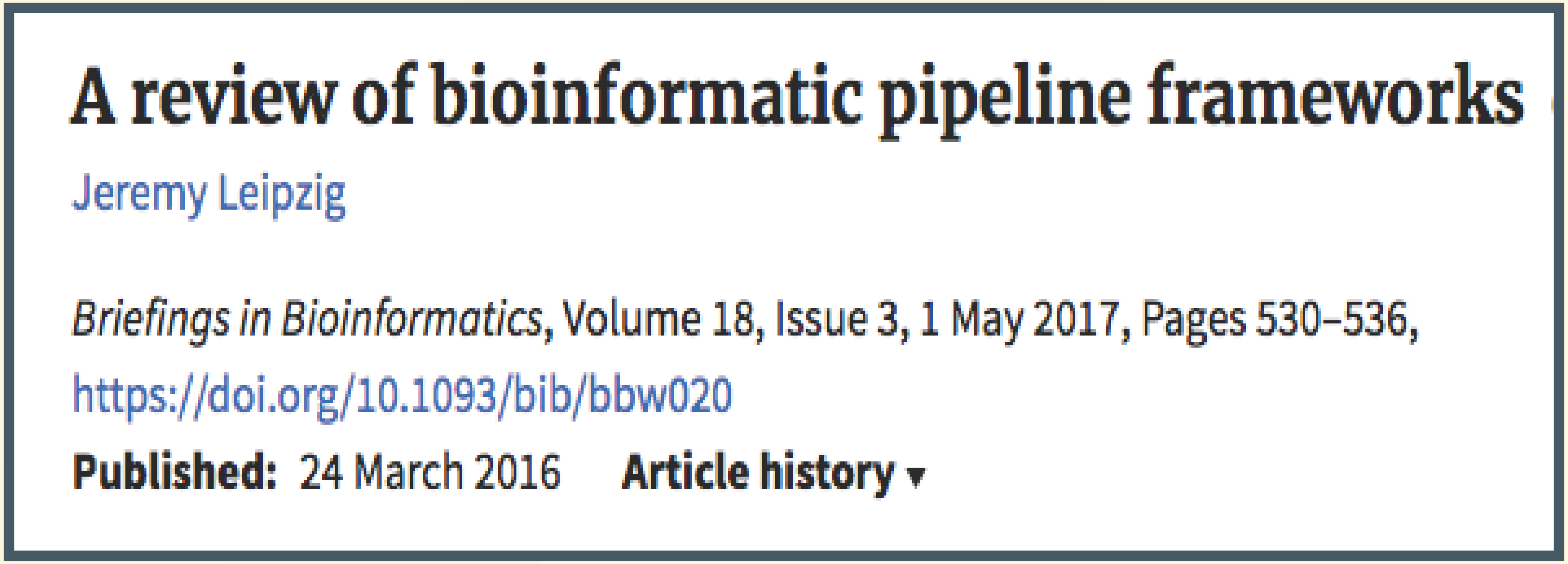
Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne ✉, Yolanda Gil ✉

Published: November 27, 2013 • https://doi.org/10.1371/journal.pone.0080278

Provenance

Solution: well-structured workflows





A review of bioinformatic pipeline frameworks

Jeremy Leipzig

Briefings in Bioinformatics, Volume 18, Issue 3, 1 May 2017, Pages 530–536, https://doi.org/10.1093/bib/bbw020

Published: 24 March 2016 **Article history ▾**

Galaxy

Snakemake

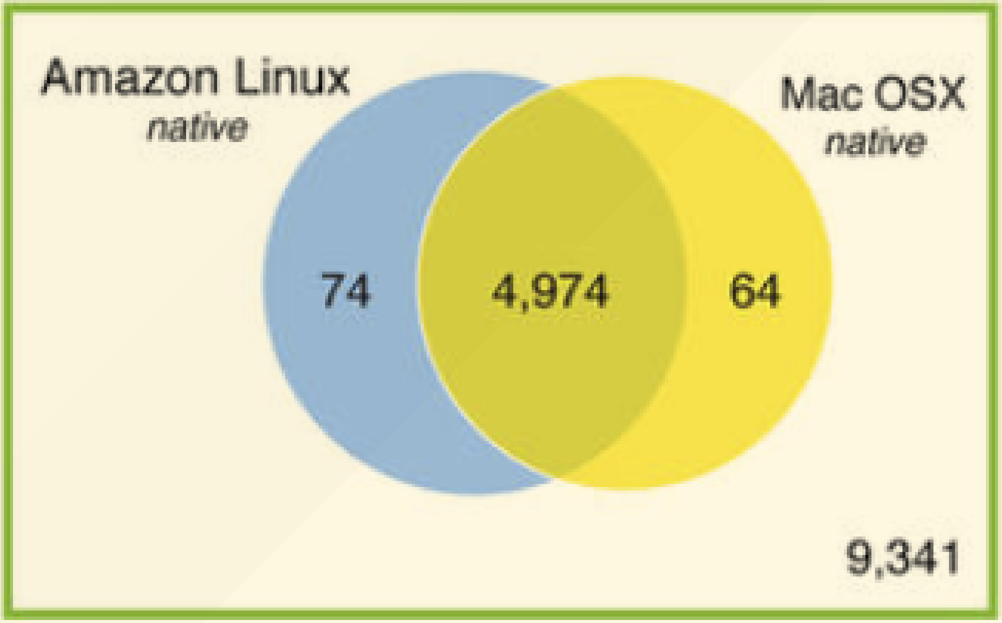
nextflow

Data Sharing +

well-structured workflow

3. Variations across workstations and operating systems

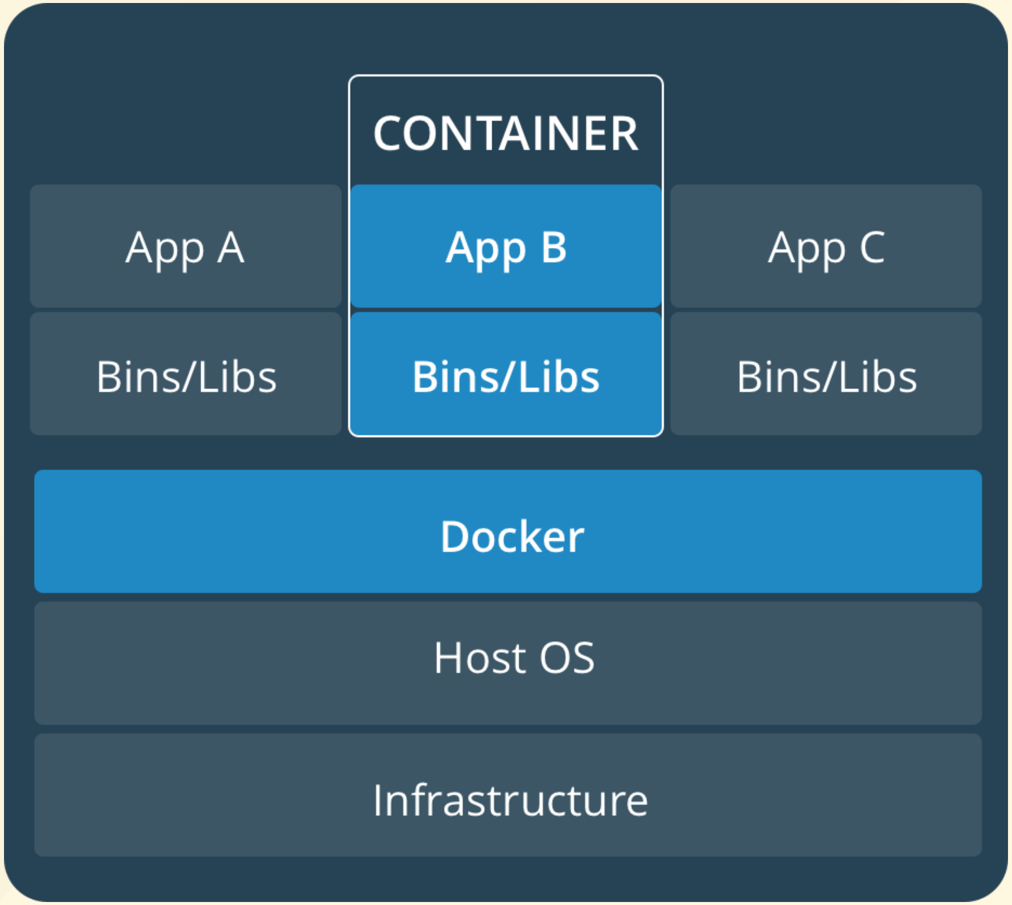
Same data & pipeline, different OS and results



Kallisto and Sleuth pipelines, applied to find differentially expressed genes (q-value < 0.01) in an RNA-seq experiment, using data from human lung fibroblasts.

[Di Tommaso et. al, Nat. Biotechnol (2017)]

Solution: containers

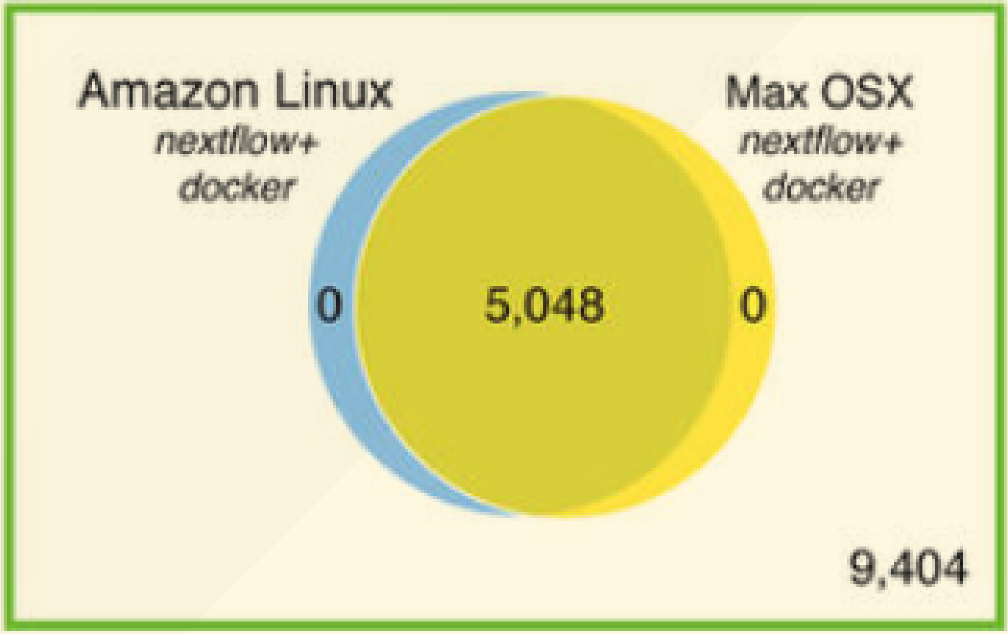


What is a container - Docker

Data sharing +

well-structured containerised workflow

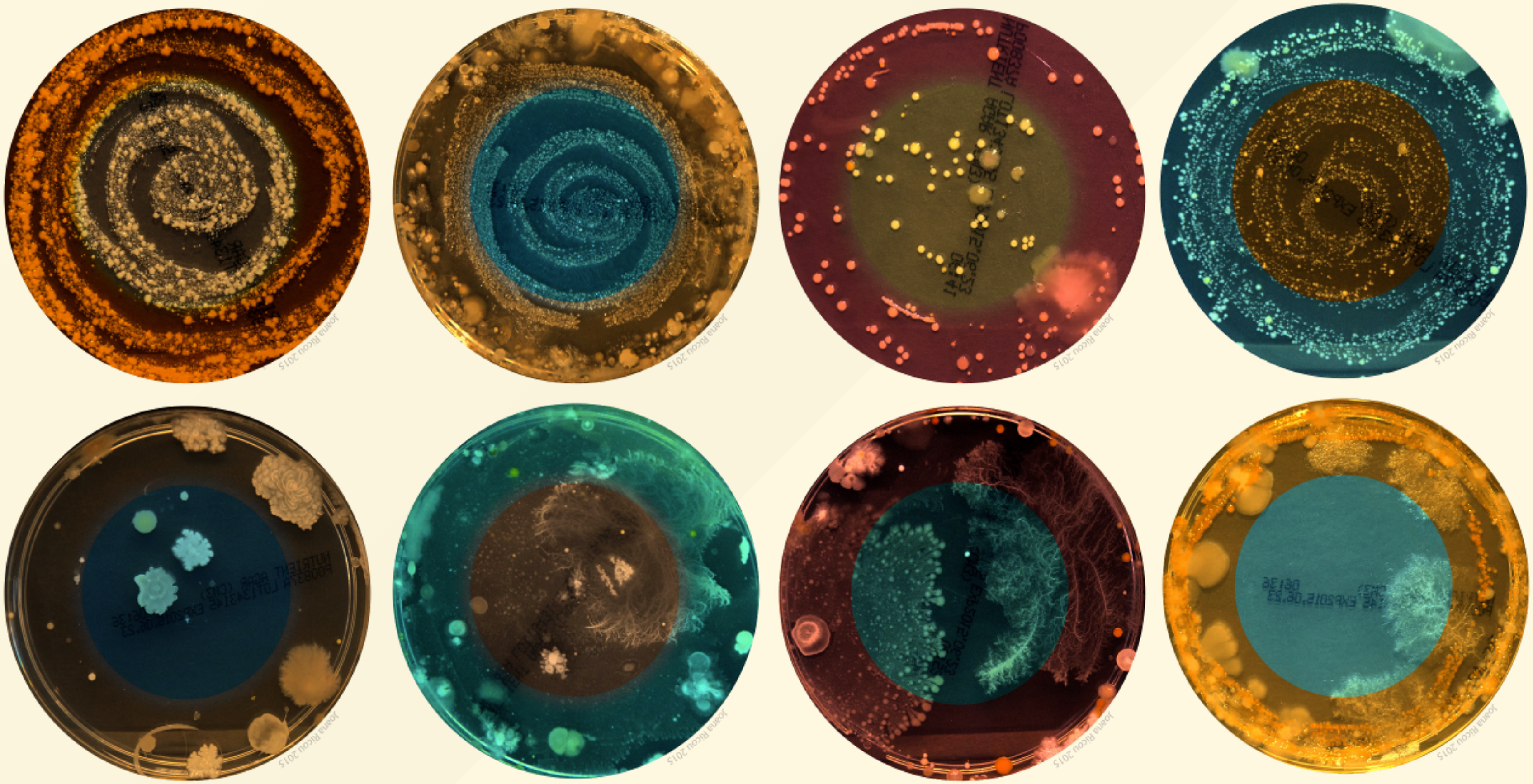
Same data, pipeline, container, and results



Kallisto and Sleuth pipelines, applied to find differentially expressed genes (q-value < 0.01) in an RNA-seq experiment, using data from human lung fibroblasts.

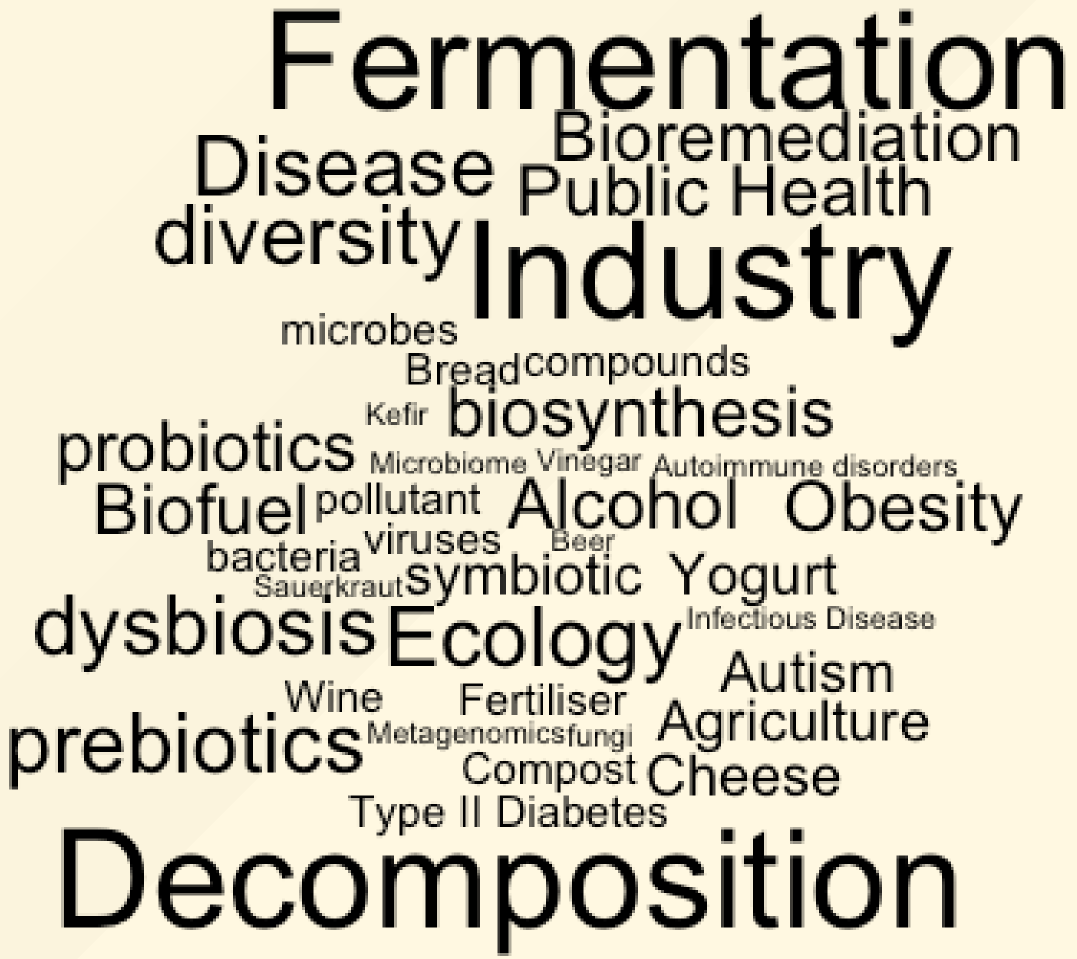
Di Tommaso et. al, Nat. Biotechnol (2017)

Metagenomics



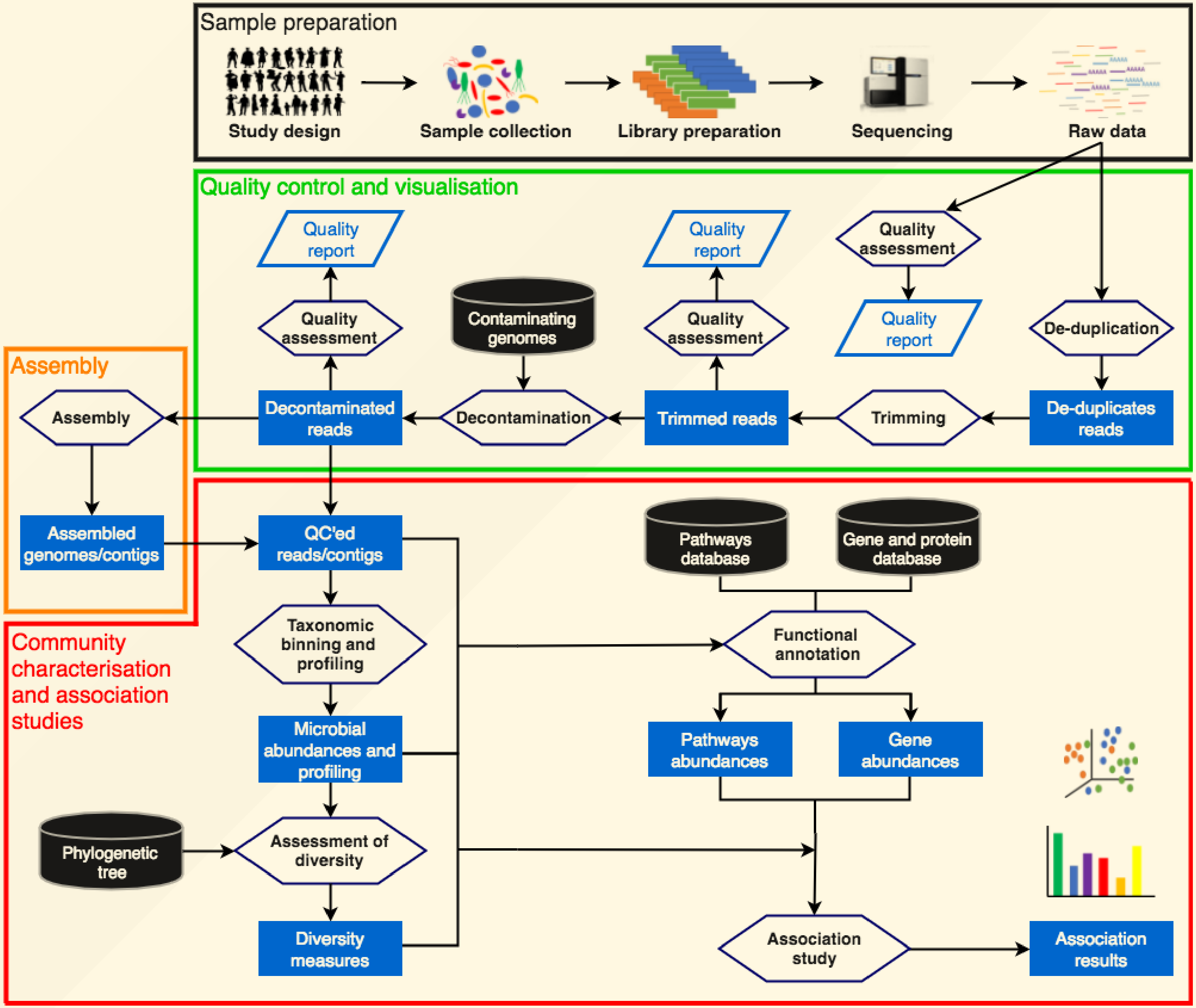
Belly Button Microbiome © Joana Ricou

> “ Metagenomics is the study of genetic material recovered directly from environmental samples.
>
> *Metagenomics - Wikipedia* ”



Fermentation
Bioremediation
Disease
Public Health
diversity
Industry
microbes
Bread
compounds
Kefir
biosynthesis
probiotics
Microbiome
Vinegar
Autoimmune disorders
Biofuel
pollutant
Alcohol
Obesity
viruses
Beer
bacteria
Sauerkraut
symbiotic
Yogurt
dysbiosis
Ecology
Infectious Disease
Autism
Wine
Fertiliser
prebiotics
Metagenomics
fungi
Agriculture
Compost
Cheese
Type II Diabetes
Decomposition

How to conduct metagenomics studies (simplified)


Sample preparation
Study design
Sample collection
Library preparation
Sequencing
Raw data
Quality control and visualisation
Quality report
Quality assessment
Contaminating genomes
Quality report
Quality assessment
Quality assessment
Quality report
De-duplication
Decontaminated reads
Decontamination
Trimmed reads
Trimming
De-duplicates reads
Assembly
Assembly
Assembled genomes/contigs
QC'ed reads/contigs
Pathways database
Gene and protein database
Community characterisation and association studies
Taxonomic binning and profiling
Functional annotation
Pathways abundances
Gene abundances
Microbial abundances and profiling
Phylogenetic tree
Assessment of diversity
Diversity measures
Association study
Association results

YAMP: *Yet Another Metagenomics Pipeline*

Why?

1. Easy to use

2. Portable

3. Flexible

4. Reproducible

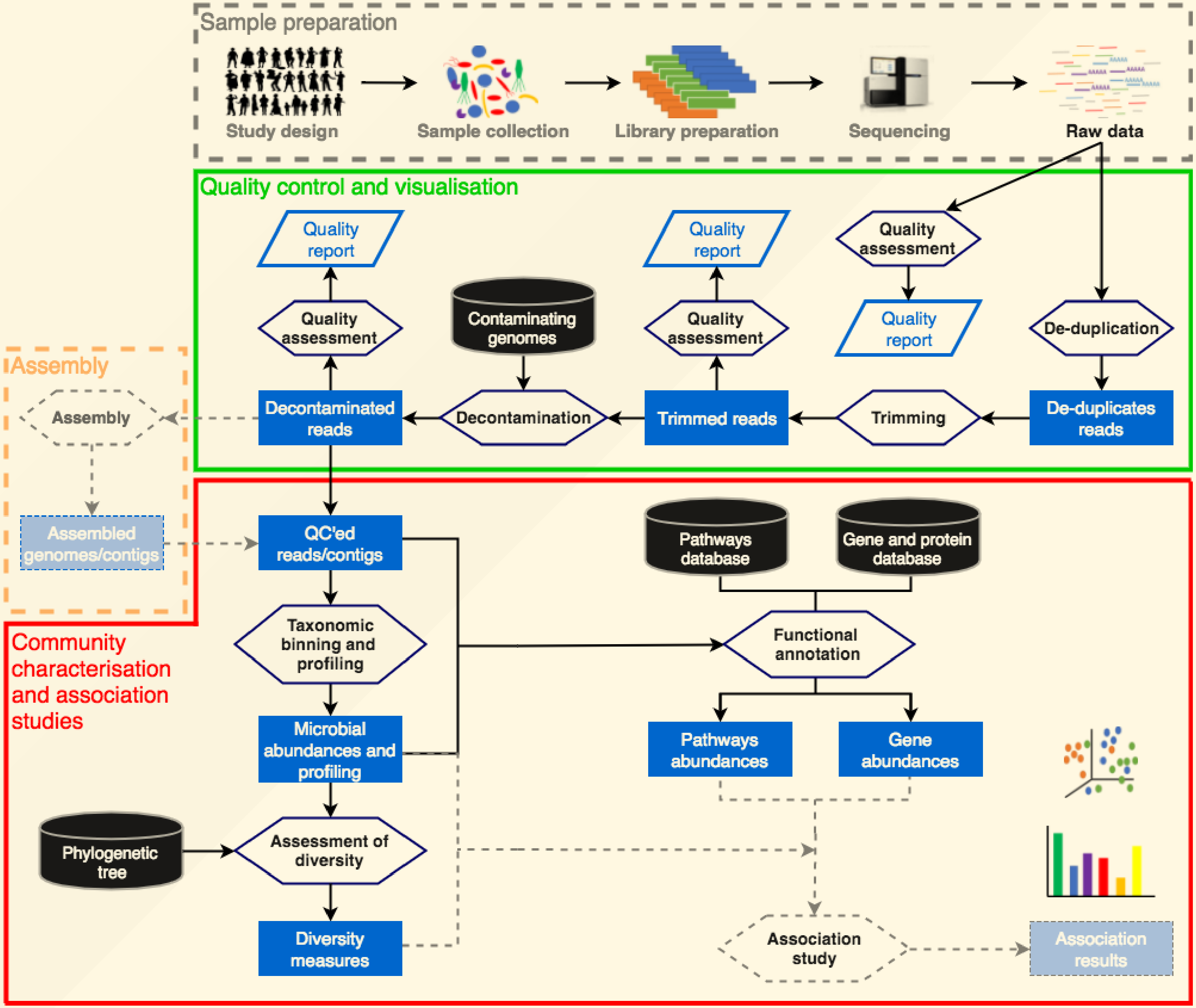
How?

nextflow

docker

S
Singularity

What?



Sample preparation

Study design

Sample collection

Library preparation

Sequencing

Raw data

Quality control and visualisation

Quality assessment

Quality report

De-duplication

De-duplicates reads

Trimming

Trimmed reads

Quality assessment

Quality report

Contaminating genomes

Decontamination

Decontaminated reads

Quality assessment

Quality report

Assembly

Assembly

Assembled genomes/contigs

Community characterisation and association studies

QC'ed reads/contigs

Taxonomic binning and profiling

Microbial abundances and profiling

Phylogenetic tree

Assessment of diversity

Diversity measures

Pathways database

Gene and protein database

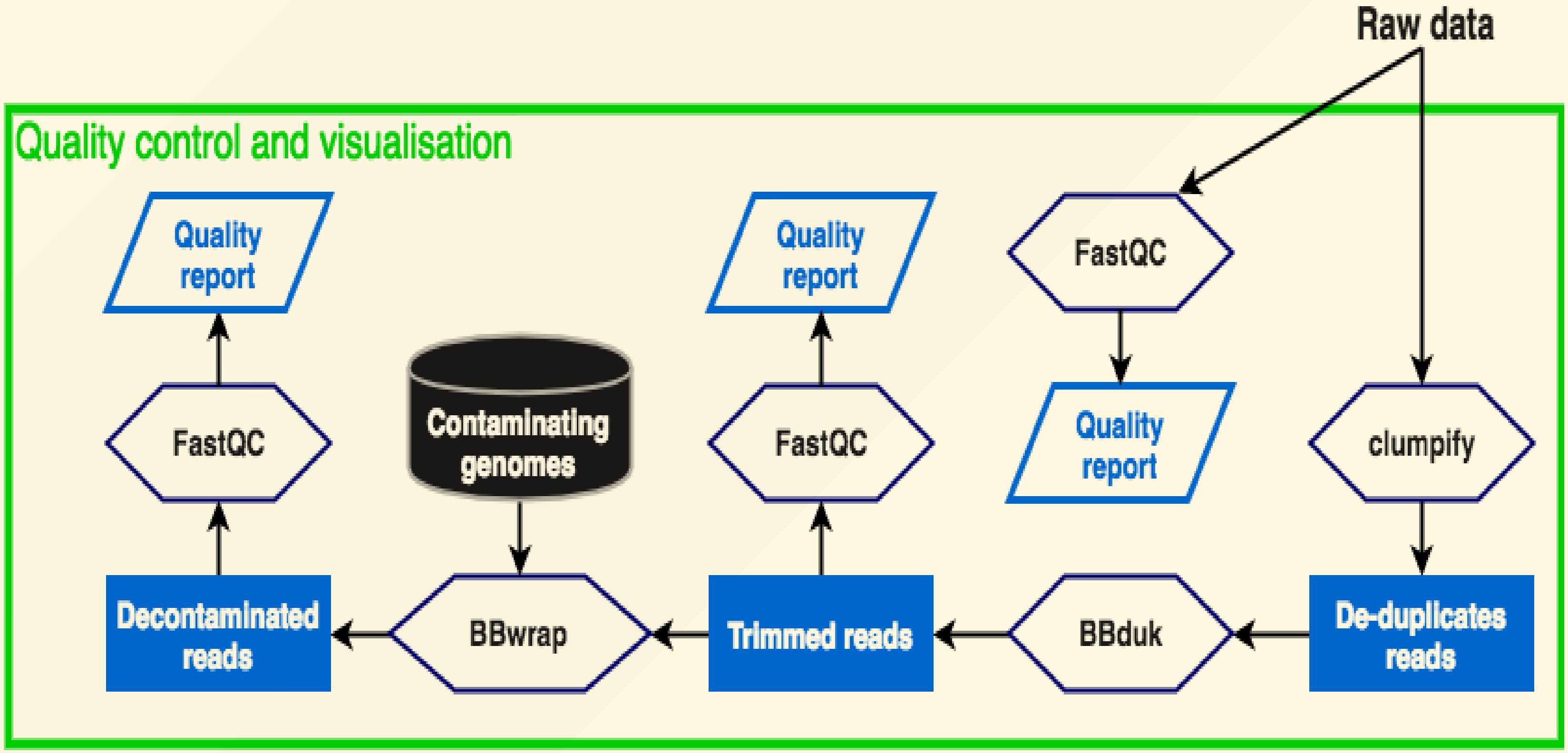
Functional annotation

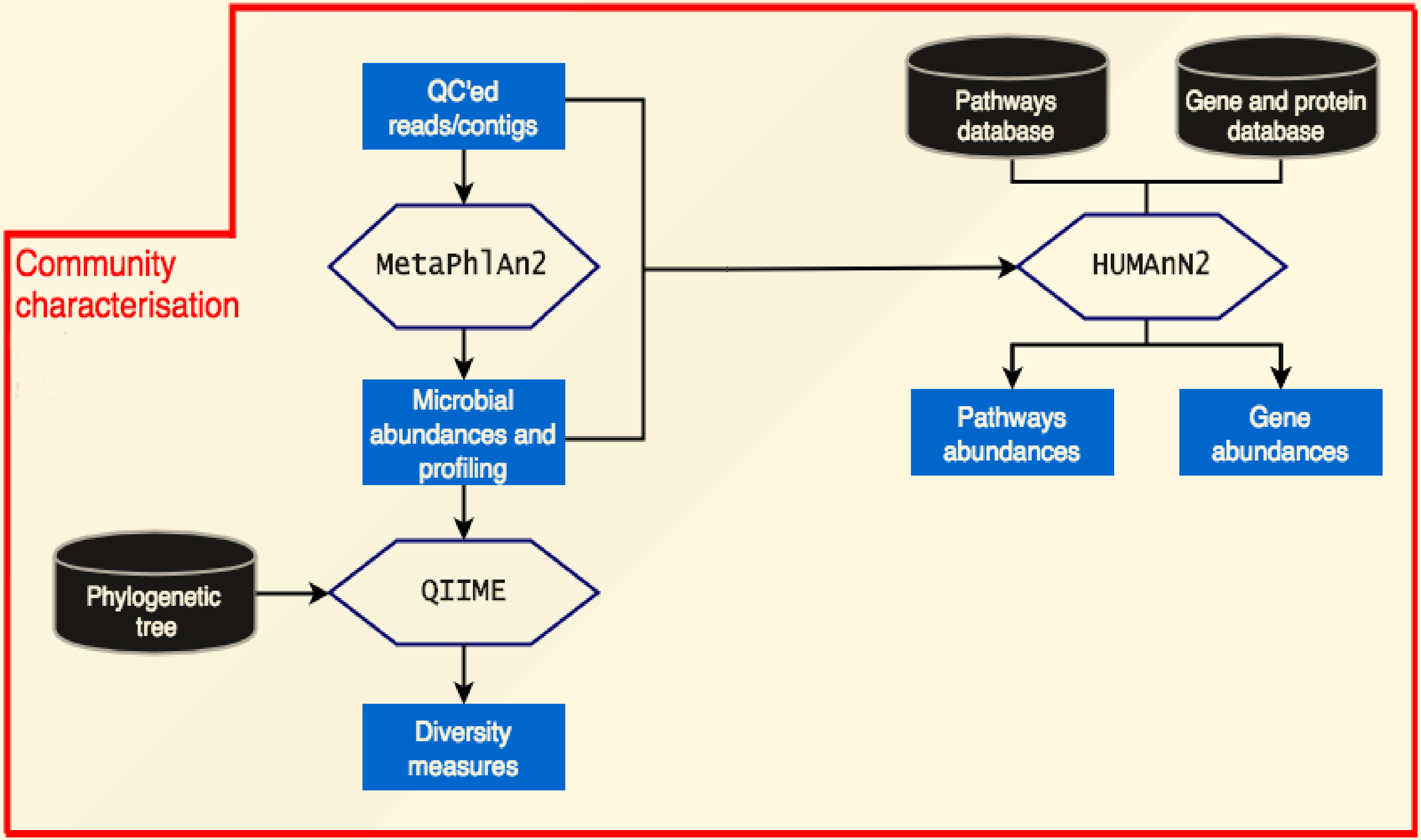
Pathways abundances

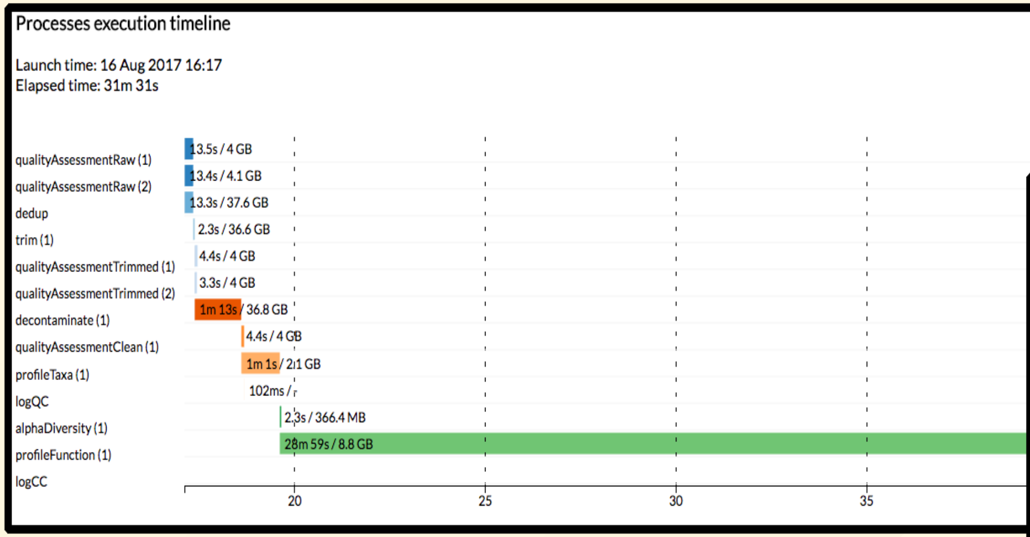
Gene abundances

Association study

Association results

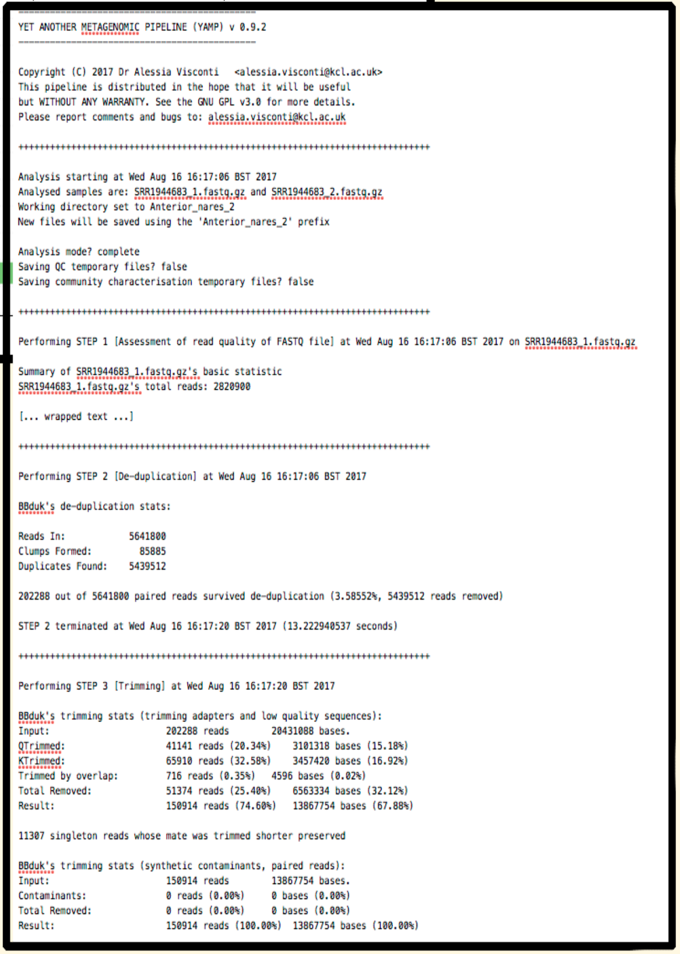

Raw data
Quality control and visualisation
Quality report
FastQC
Contaminating genomes
Decontaminated reads
BBwrap
Quality report
FastQC
Trimmed reads
FastQC
Quality report
BBduk
clumpify
De-duplicates reads


Community characterisation
QC'ed reads/contigs
MetaPhlAn2
Microbial abundances and profiling
Phylogenetic tree
QIIME
Diversity measures
Pathways database
Gene and protein database
HUMAnN2
Pathways abundances
Gene abundances



Processes execution timeline

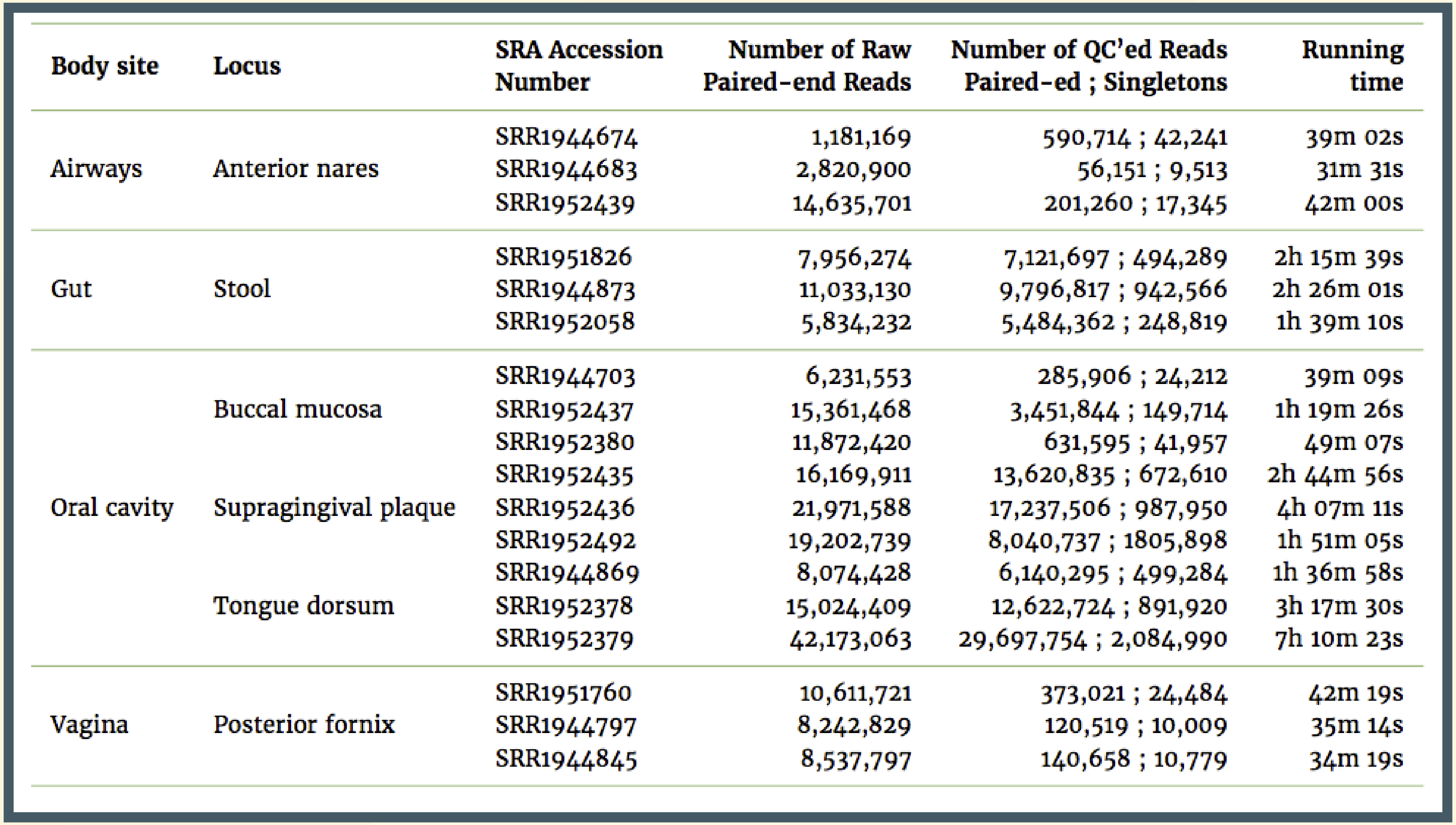
Launch time: 16 Aug 2017 16:17
Elapsed time: 31m 31s



```
----------------------------------------
YET ANOTHER METAGENOMIC PIPELINE (YAMP) v 0.9.2
----------------------------------------

Copyright (C) 2017 Dr Alessia Visconti   <alessia.visconti@kcl.ac.uk>
This pipeline is distributed in the hope that it will be useful
but WITHOUT ANY WARRANTY. See the GNU GPL v3.0 for more details.
Please report comments and bugs to: alessia.visconti@kcl.ac.uk

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Analysis starting at Wed Aug 16 16:17:06 BST 2017
Analysed samples are: SRR1944683_1.fastq.gz and SRR1944683_2.fastq.gz
Working directory set to Anterior_nares_2
New files will be saved using the 'Anterior_nares_2' prefix

Analysis mode? complete
Saving QC temporary files? false
Saving community characterisation temporary files? false

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Performing STEP 1 [Assessment of read quality of FASTQ file] at Wed Aug 16 16:17:06 BST 2017 on SRR1944683_1.fastq.gz

Summary of SRR1944683_1.fastq.gz's basic statistic
SRR1944683_1.fastq.gz's total reads: 2820900

[... wrapped text ...]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Performing STEP 2 [De-duplication] at Wed Aug 16 16:17:06 BST 2017

BBduk's de-duplication stats:

Reads In:           5641800
Clumps Formed:        85885
Duplicates Found:   5439512

202288 out of 5641800 paired reads survived de-duplication (3.58552%, 5439512 reads removed)

STEP 2 terminated at Wed Aug 16 16:17:20 BST 2017 (13.222940537 seconds)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Performing STEP 3 [Trimming] at Wed Aug 16 16:17:20 BST 2017

BBduk's trimming stats (trimming adapters and low quality sequences):
Input:                      202288 reads        20431088 bases.
QTrimmed:                   41141 reads (20.34%)    3101318 bases (15.18%)
KTrimmed:                   65910 reads (32.58%)    3457420 bases (16.92%)
Trimmed by overlap:         716 reads (0.35%)   4596 bases (0.02%)
Total Removed:              51374 reads (25.40%)    6563334 bases (32.12%)
Result:                     150914 reads (74.60%)   13867754 bases (67.88%)

11307 singleton reads whose mate was trimmed shorter preserved

BBduk's trimming stats (synthetic contaminants, paired reads):
Input:                      150914 reads        13867754 bases.
Contaminants:               0 reads (0.00%)     0 bases (0.00%)
Total Removed:              0 reads (0.00%)     0 bases (0.00%)
Result:                     150914 reads (100.00%)  13867754 bases (100.00%)
```

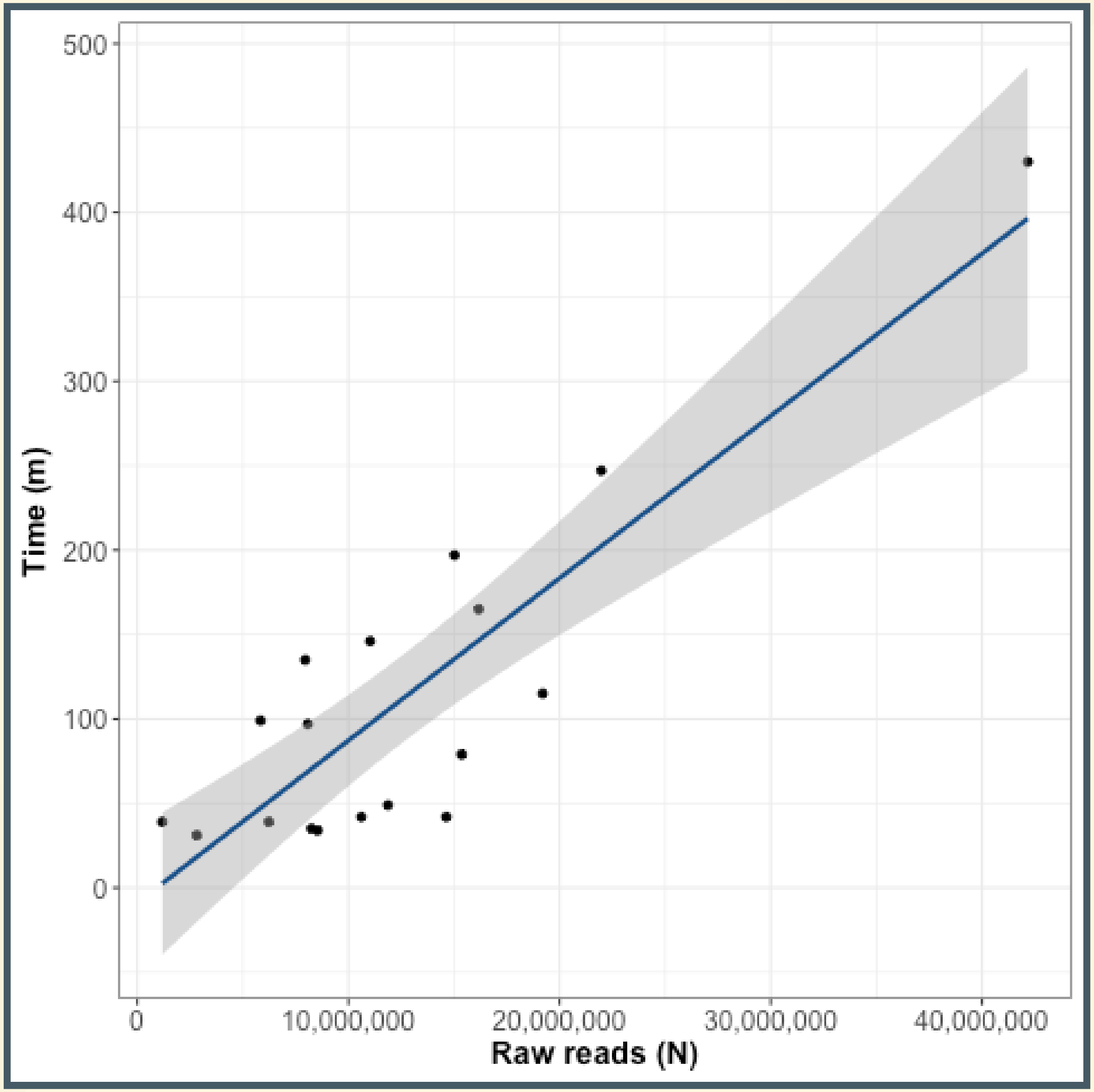
Additional output: detailed log file &
statistics of memory usage and time of execution

A case study

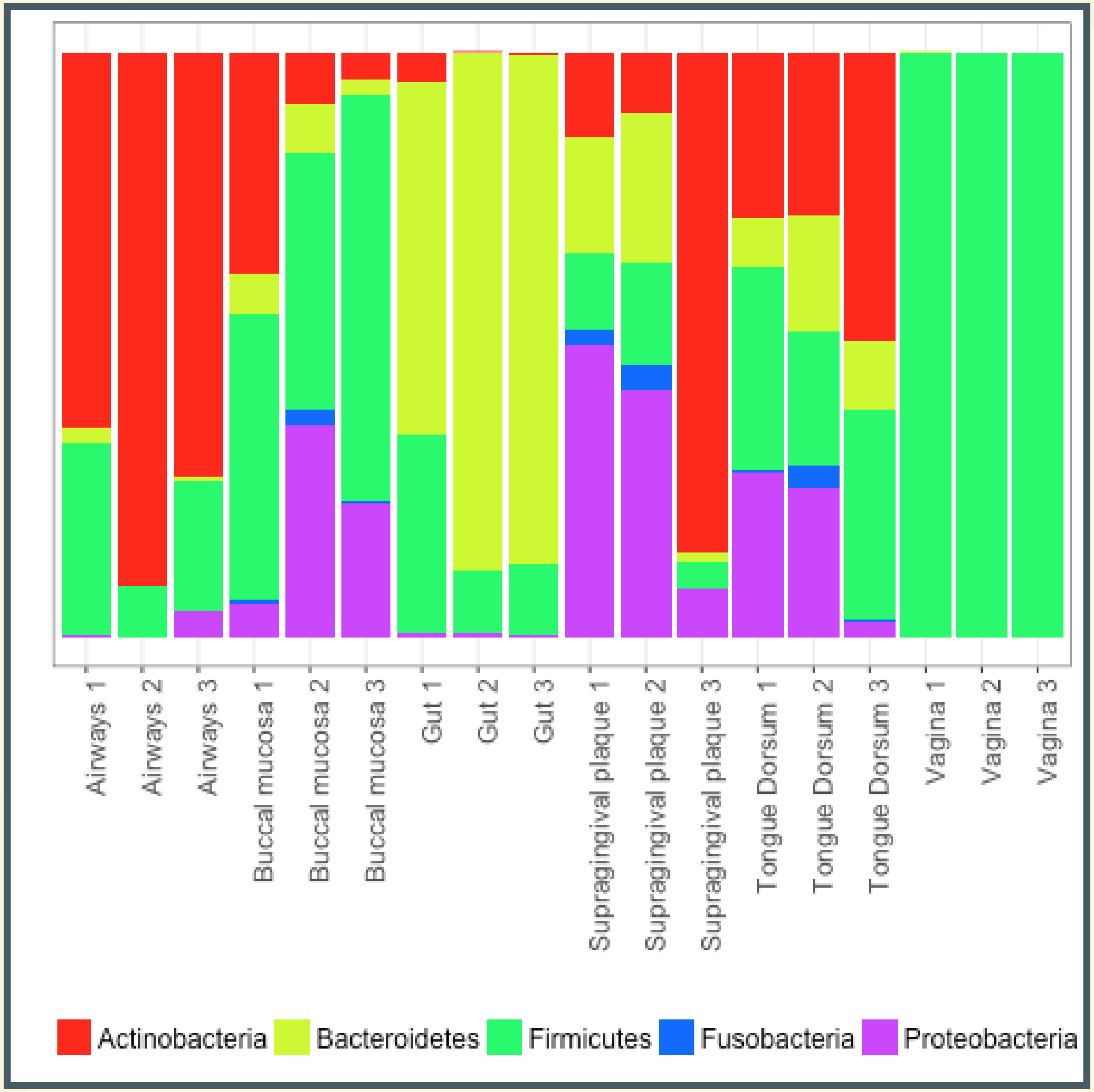


Body site	Locus	SRA Accession Number	Number of Raw Paired-end Reads	Number of QC'ed Reads Paired-ed ; Singletons	Running time
Airways	Anterior nares	SRR1944674	1,181,169	590,714 ; 42,241	39m 02s
		SRR1944683	2,820,900	56,151 ; 9,513	31m 31s
		SRR1952439	14,635,701	201,260 ; 17,345	42m 00s
Gut	Stool	SRR1951826	7,956,274	7,121,697 ; 494,289	2h 15m 39s
		SRR1944873	11,033,130	9,796,817 ; 942,566	2h 26m 01s
		SRR1952058	5,834,232	5,484,362 ; 248,819	1h 39m 10s
Oral cavity	Buccal mucosa	SRR1944703	6,231,553	285,906 ; 24,212	39m 09s
		SRR1952437	15,361,468	3,451,844 ; 149,714	1h 19m 26s
		SRR1952380	11,872,420	631,595 ; 41,957	49m 07s
	Supragingival plaque	SRR1952435	16,169,911	13,620,835 ; 672,610	2h 44m 56s
		SRR1952436	21,971,588	17,237,506 ; 987,950	4h 07m 11s
		SRR1952492	19,202,739	8,040,737 ; 1805,898	1h 51m 05s
	Tongue dorsum	SRR1944869	8,074,428	6,140,295 ; 499,284	1h 36m 58s
		SRR1952378	15,024,409	12,622,724 ; 891,920	3h 17m 30s
		SRR1952379	42,173,063	29,697,754 ; 2,084,990	7h 10m 23s
Vagina	Posterior fornix	SRR1951760	10,611,721	373,021 ; 24,484	42m 19s
		SRR1944797	8,242,829	120,519 ; 10,009	35m 14s
		SRR1944845	8,537,797	140,658 ; 10,779	34m 19s

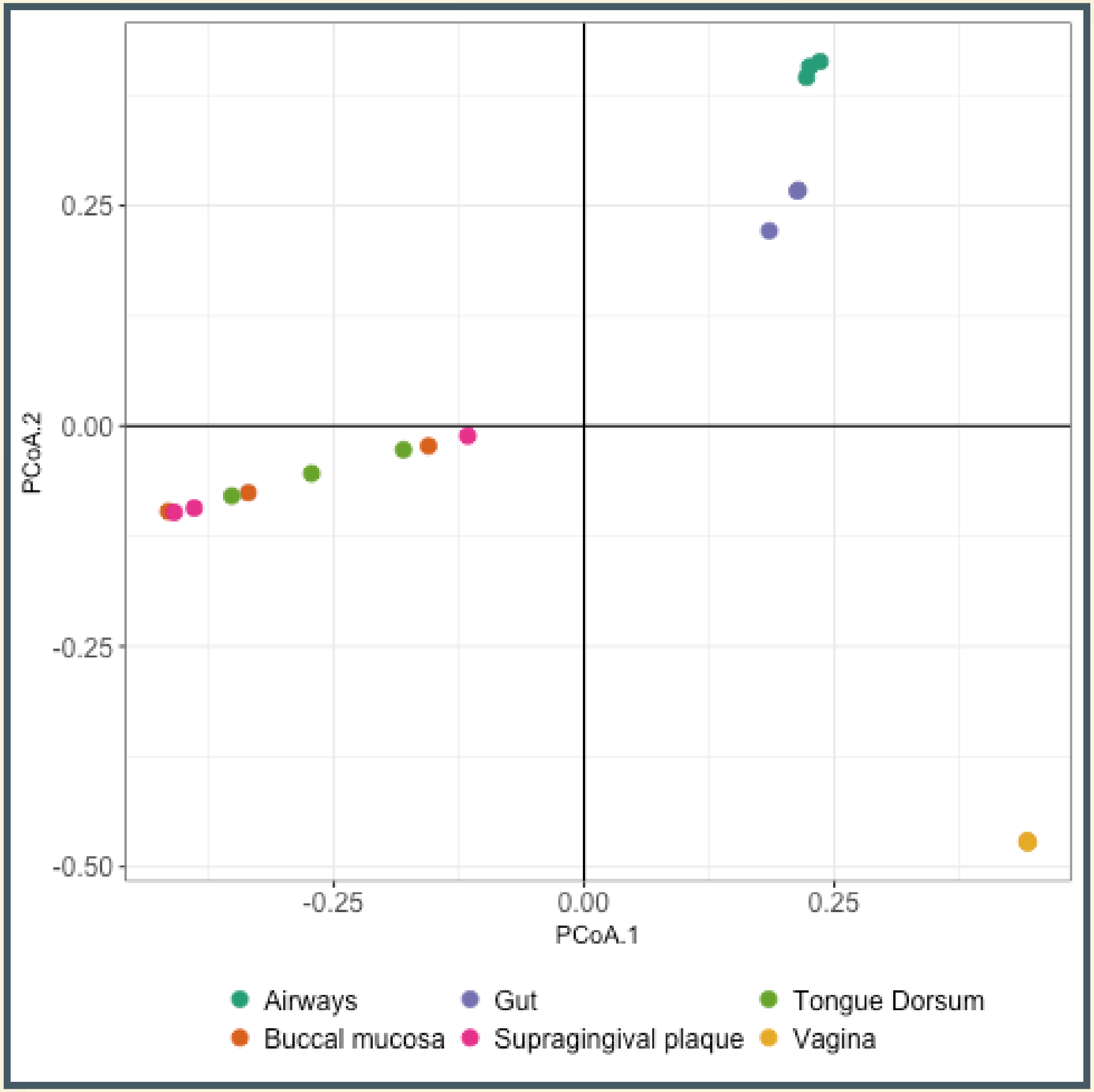
HMP Phase III, PRJNA275349



2.60GHz Intel® Xeon® processor with 32 GB of RAM



Phylum level relative abundances



PCoA evaluated on the Bray-Curtis dissimilarity
among species relative abundances

Conclusion

1. Easy to use

Install YAMP (with preset default parameters):

```
git clone https://github.com/alesssia/YAMP.git
```

Download the supporting data:

```
wget https://zenodo.org/record/1068229/files/YAMP_resources_20171128.tar.gz
tar -xzf YAMP_resources_20171128.tar.gz
```

Run your analysis:

```
nextflow run YAMP.nf --reads1 R1.fq.gz --reads2 R2.fq.gz
        --prefix mysample --outdir outdir --mode complete
        --with-singularity docker://alessia/yampdocker
```

2. Portable

- SGE cluster

```
executor = 'sge'
```

- PBS/Torque cluster

```
executor = 'pbs'
```

- Your laptop

```
// executor = 'sge'
```

3. Flexible

4. Reproducible

5. Not only for metagenomics

6. Not only for reproducibility

Where?

GitHub

https://github.com/alesssia/YAMP
https://github.com/alesssia/YAMP/wiki

docker hub

https://hub.docker.com/r/alesssia/yampdocker

bioRχiv beta

https://www.biorxiv.org/content/early/2017/11/21/223016

Who?

Acknowledgements

Mario Falchi
Tiphaine Martin

Paolo Di Tommaso
Brian Bushnell

Richard Davies

KING'S College LONDON

TwinsUK

🐦 @_alesssia
✉️ alessia.visconti@kcl.ac.uk