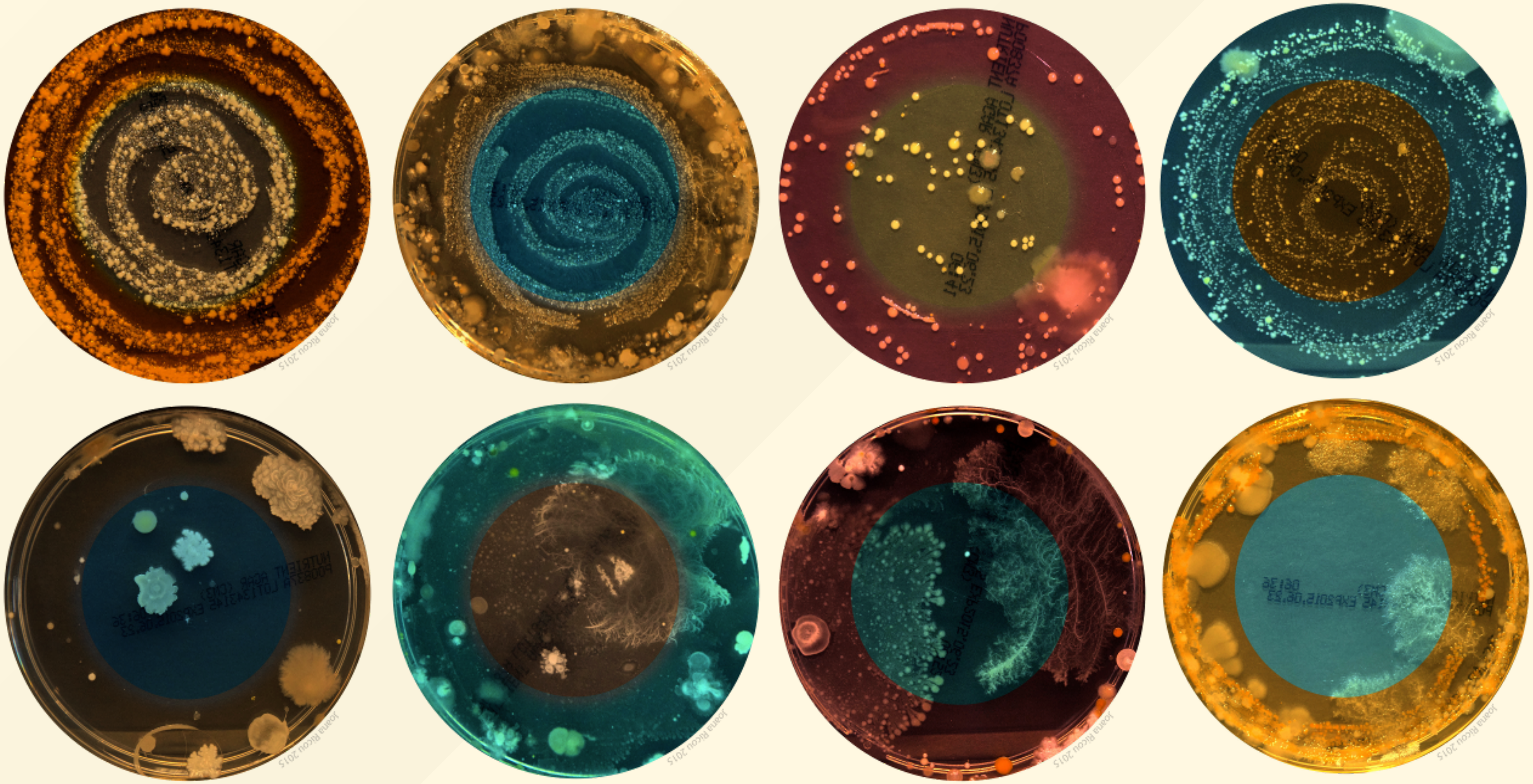
Reproducible shotgun metagenomic analysis with Nextflow and containers

Alessia Visconti

TwinsUK, King's College London

🐦 @_alesssia

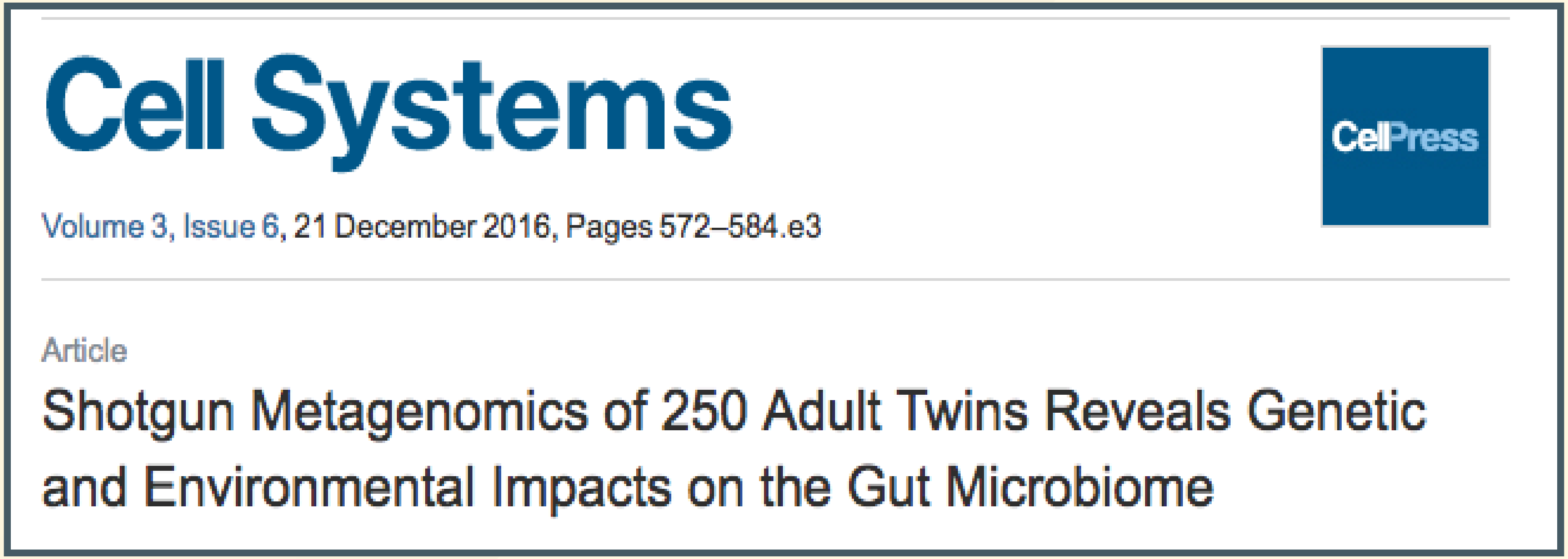
Metagenomics?!?



Belly Button Microbiome © Joana Ricou

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

Metagenomics @TwinsUK



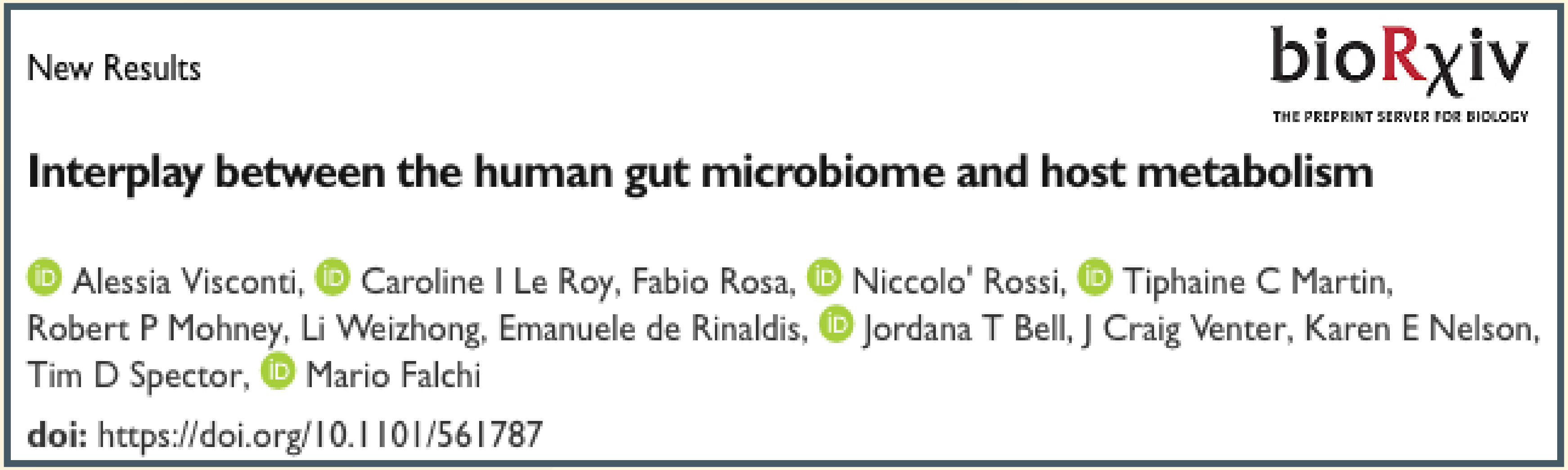
Cell Systems

CellPress

Volume 3, Issue 6, 21 December 2016, Pages 572–584.e3

Article

Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome

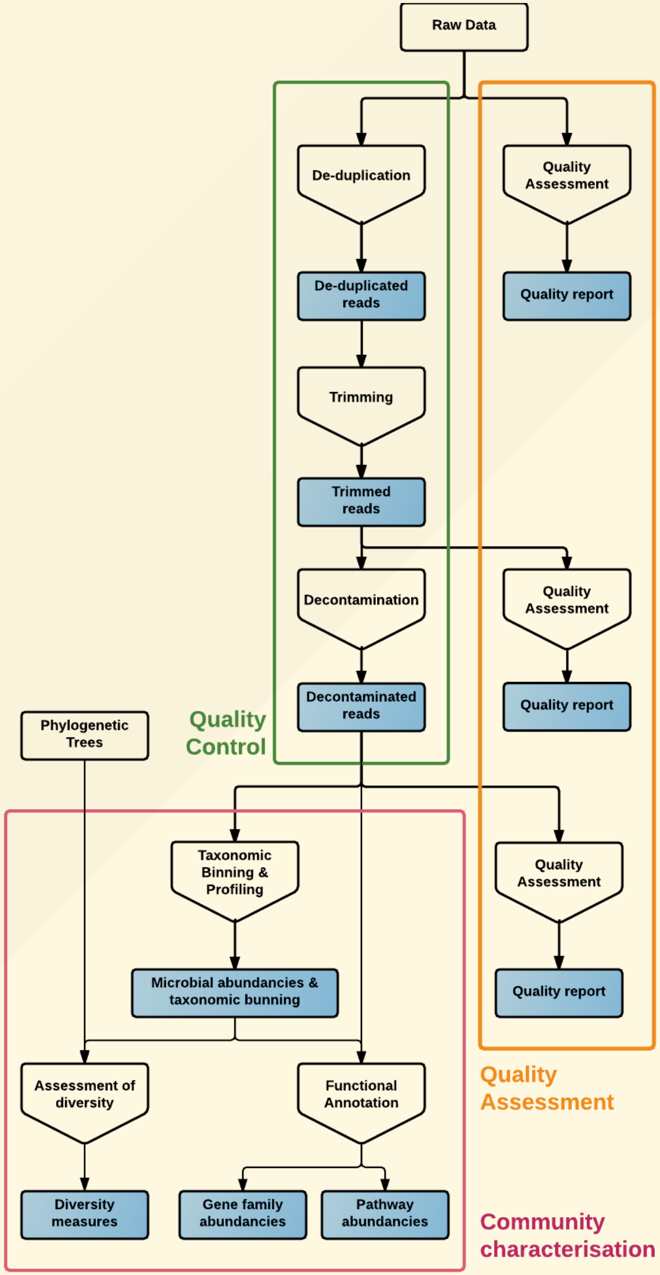


New Results

Interplay between the human gut microbiome and host metabolism

Alessia Visconti, Caroline I Le Roy, Fabio Rosa, Niccolo' Rossi, Tiphaine C Martin, Robert P Mohney, Li Weizhong, Emanuele de Rinaldis, Jordana T Bell, J Craig Venter, Karen E Nelson, Tim D Spector, Mario Falchi

doi: https://doi.org/10.1101/561787



Raw Data

De-duplication

De-duplicated reads

Trimming

Trimmed reads

Decontamination

Decontaminated reads

Quality Assessment

Quality report

Quality Assessment

Quality report

Quality Assessment

Quality report

Quality Control

Quality Assessment

Phylogenetic Trees

Taxonomic Binning & Profiling

Microbial abundancies & taxonomic bunning

Assessment of diversity

Diversity measures

Functional Annotation

Gene family abundancies

Pathway abundancies

Community characterisation

With great power comes great responsibility

(and great trouble)

1. Reproducible

2. Portable

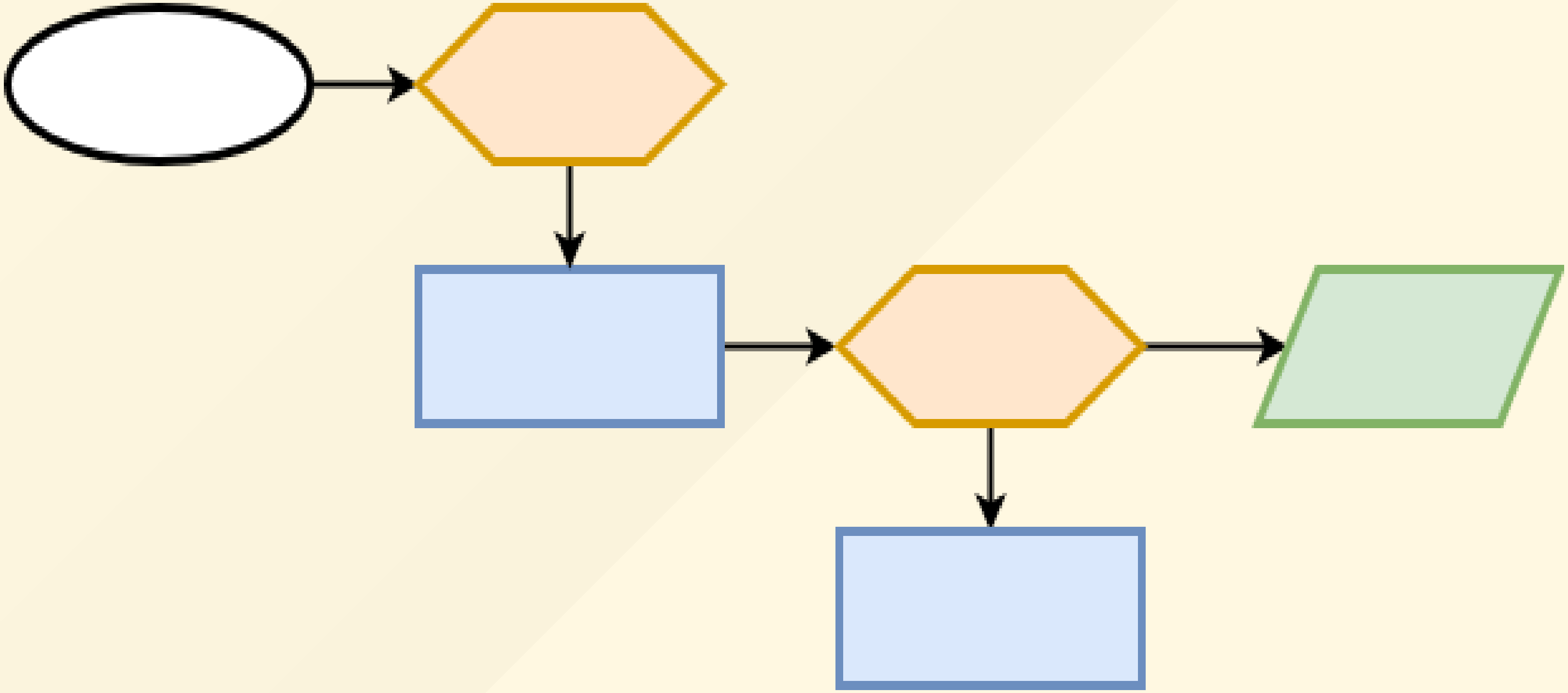
3. Flexible

4. Efficient

5. Easy to use

Reinventing the wheel?

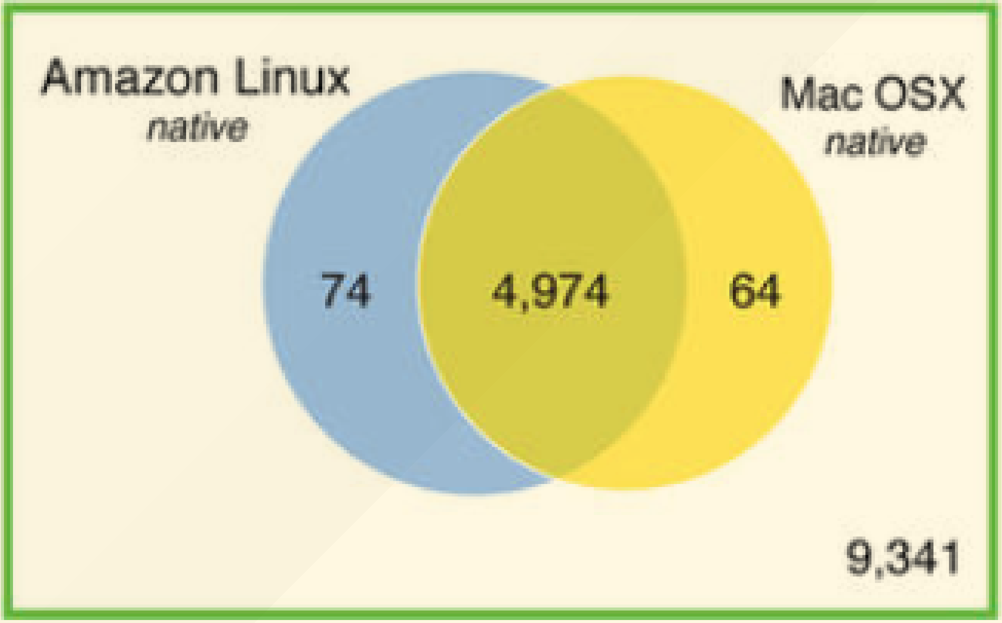
Well-structured workflows



nextflow

Are we there yet?

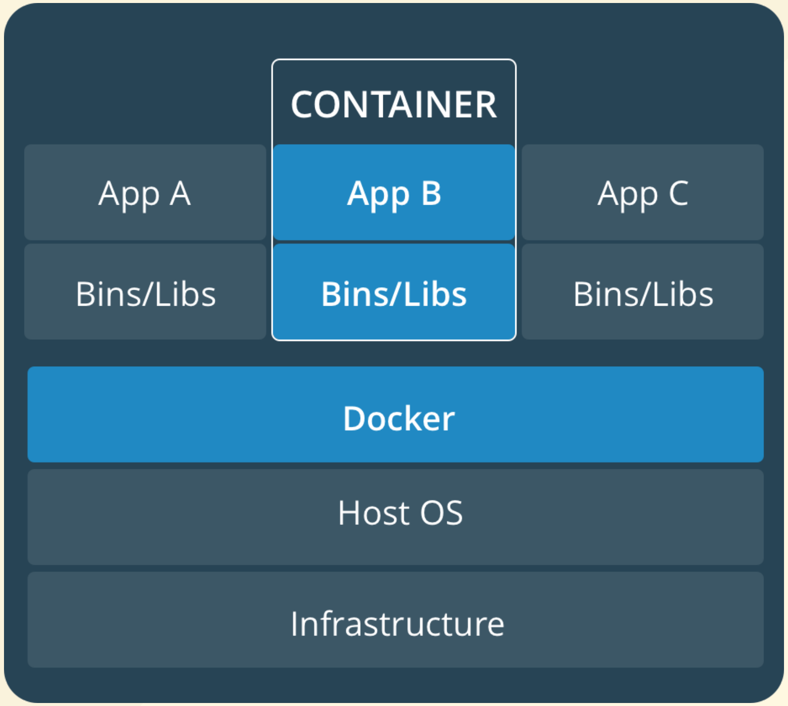
Same data & pipeline, different OS and results



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)

Containers



CONTAINER

App A

App B

App C

Bins/Libs

Bins/Libs

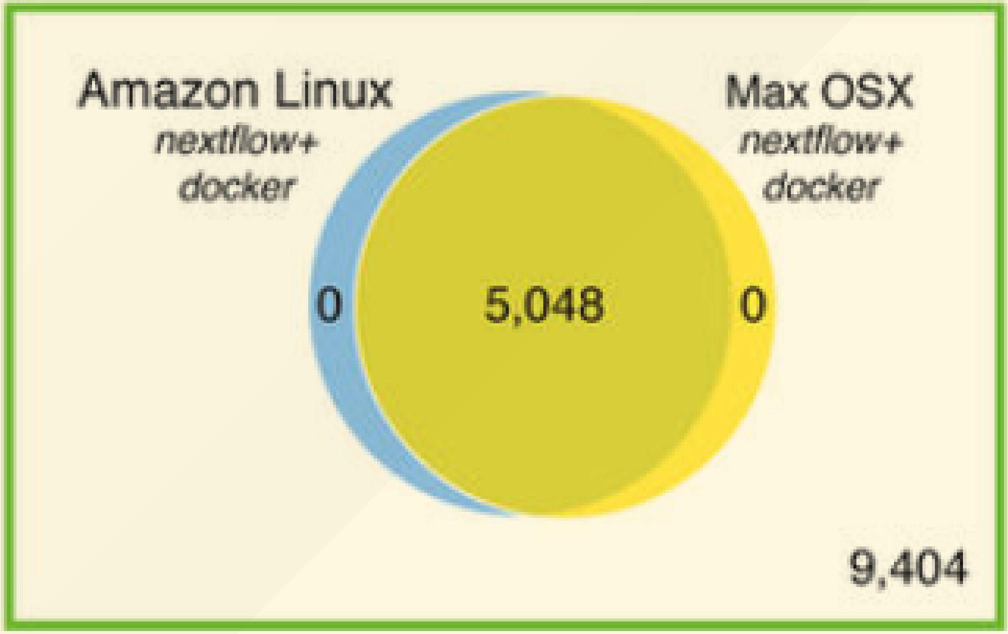
Bins/Libs

Docker

Host OS

Infrastructure

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)

YAMP: *Yet Another Metagenomics Pipeline*

1. Reproducible

- Parameters all in the same place

```
qin=33
kcontaminants = 23
phred = 10 trimmed
minlength = 60
mink = 11
hdist = 1
...
```

- Container Integration

```
nextflow run <script> -with-docker <docker>
nextflow run <script> -with-singularity <docker>
```

- Detailed logs

2. Portable

- Cluster1

```
executor 'sge'
```

- Cluster2

```
executor 'pbs'
```

- Personal laptop

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

- The *Cloud*

3. Flexible

```
process dedup {

        input:
        set file(in1), file(in2) from todedup

        output:
        file("${params.prefix}_dedupe*.fq") into totrim
        file("${params.prefix}_dedupe*.fq") into topublishdedupe

        when:
        (params.mode == "QC" || params.mode == "complete") && params.dedup

        script:
        """
            ...

        """
}
```

4. Efficient

- User-transparent parallelisation
- Resources fully exploited

```
$trim
{
    time '1h'
    cpus 4
    memory '32 GB'
}

$qualityAssessmentTrimmed
{
    time '15m'
    cpus 4
    memory '4 GB'
}
```

5. 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 the analysis

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

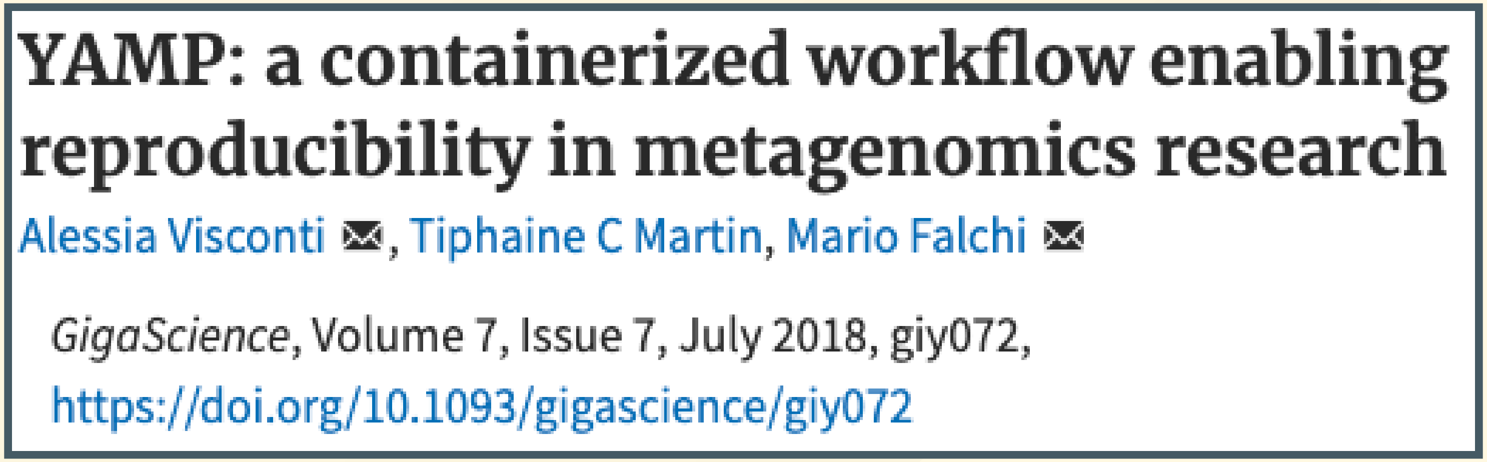
6. Not only for reproducibility

7. Not only for metagenomics

nf-core

https://nf-co.re

Nextflow + containers = 🎉



YAMP: a containerized workflow enabling reproducibility in metagenomics research

Alessia Visconti ✉, Tiphaine C Martin, Mario Falchi ✉

GigaScience, Volume 7, Issue 7, July 2018, giy072,
https://doi.org/10.1093/gigascience/giy072

GitHub

https://github.com/alesssia/YAMP

docker hub

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

Acknowledgements

Mario Falchi
Tiphaine Martin

Paolo Di Tommaso
Richard Davies

twinsUK

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