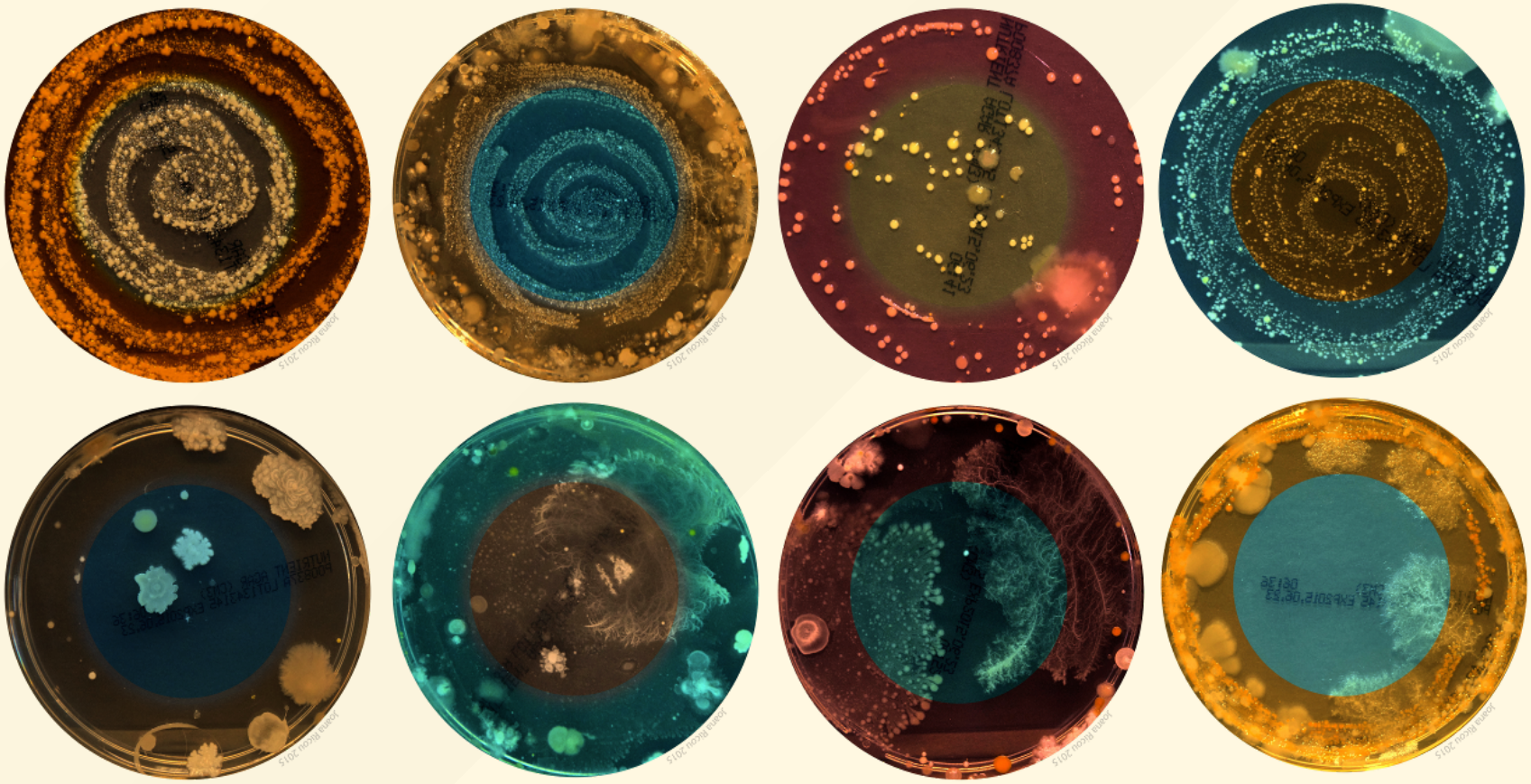
Simplifying shotgun metagenomics analysis with Nextflow

Alessia Visconti

***TwinsUK*, King's College London**

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
Agriculture
prebiotics
Metagenomics
fungi
Compost
Cheese
Type II Diabetes
Decomposition

Metagenomics @TwinsUK

Pilot study

Cell Systems

CellPress

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

Article

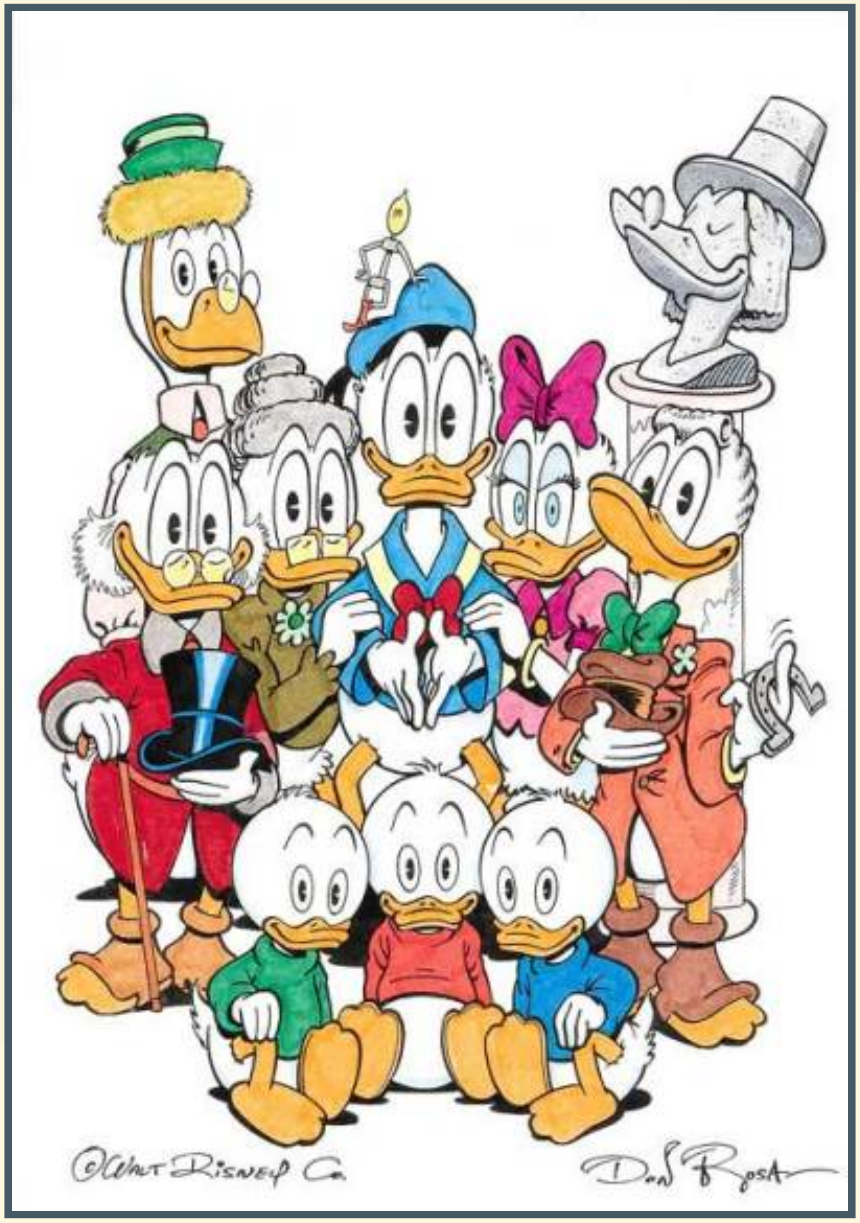
Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome

Let's go big!

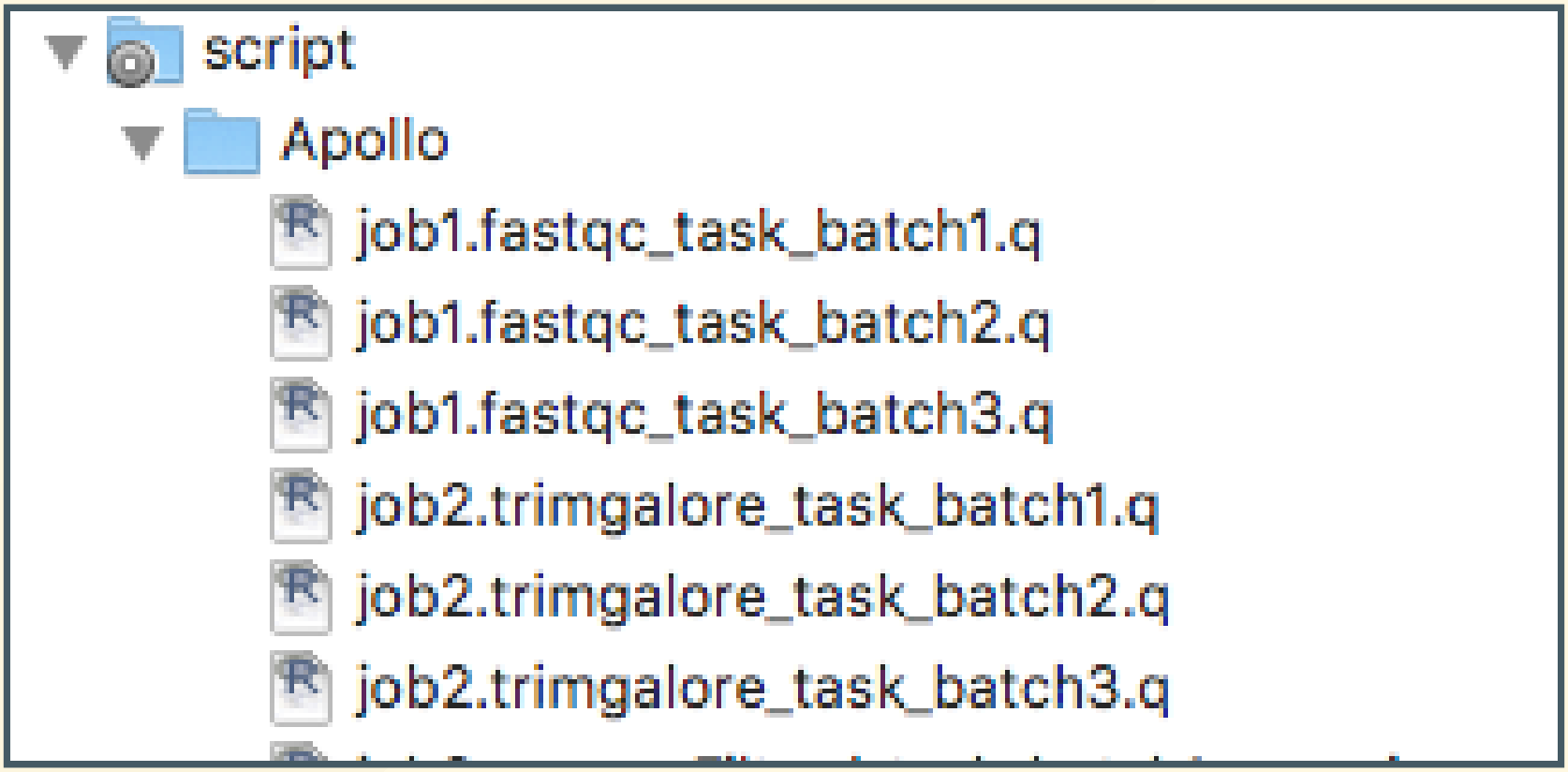
- 1300 samples
- 22M reads per sample
- 1.6 Gb of data per sample
- Novel in-house analysis workflow

THIS PAGE
INTENTIONALLY
LEFT BLANK

Why has this happened?


©Walt Disney Co.
Don Rosa

The analysis *'workflow'*



The infrastructure

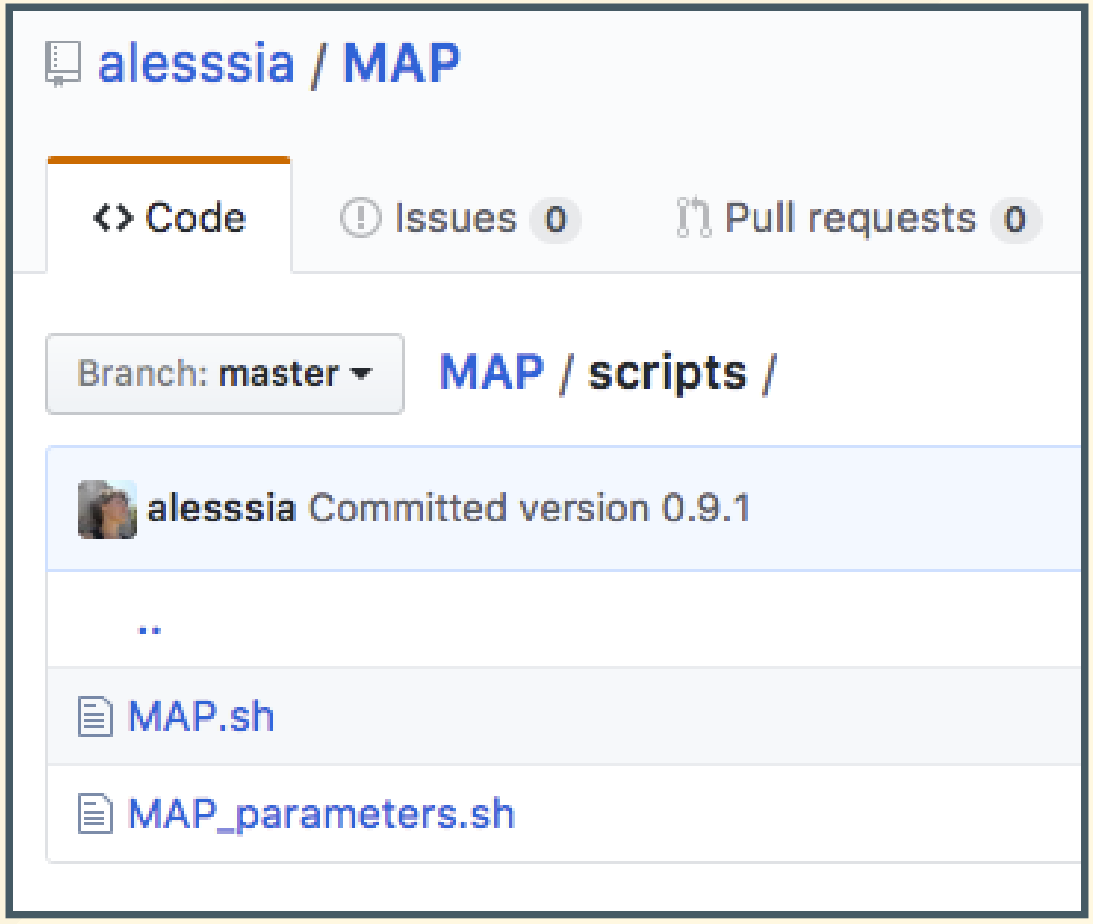
- Cluster1: `SGE`
- Cluster2: `PBS`
- Cluster3: `Slurm`

A recipe for disaster

- Limited computational literacy
- Fast-moving field
- Big(ger) data

Rewind

MAP, Metagenomics Analysis Pipeline



https://github.com/alesssia/MAP

This is not a solution

- No step parallelisation
- Limited portability
- Resources not fully exploited
- Software management still problematic

Rewind (again)

nextflow

docker

Is this the solution?

Parallelisation

```
process decontaminate {

...

output:
file "${params.prefix}_clean.fq" into assessdecontaminated
file "${params.prefix}_clean.fq" into toprofiletaxa
file "${params.prefix}_clean.fq" into toprofilefunction
```

Disclaimer: I know I could use the 'into' operator to duplicate the channel output

Portability

- Cluster1

```
executor = 'sge'
```

- Cluster2

```
executor = 'pbs'
queue = 'metagenome'
```

- My laptop

```
# executor = 'sge'
```

Resources fully exploited

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


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

Reproducibility

- All parameters in one place

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

...
```

- Docker Integration

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

Flexibility

A single parameter in the configuration file

```
dedup = true
```

and a test in the main script

```
process dedup {

...

when:
  params.dedup
```

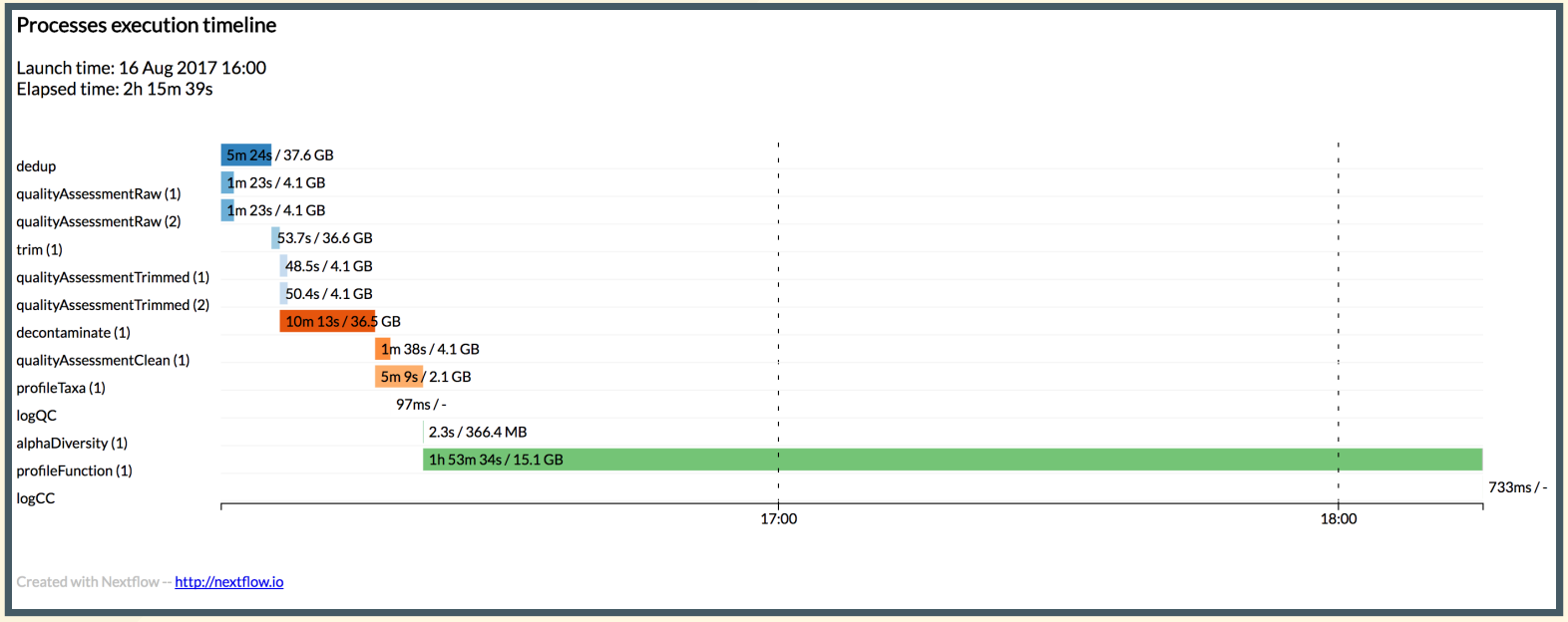
allow selecting whether dedup should be performed

A few more extra

- File management

```
publishDir wdir, mode: 'copy', pattern: "*.{html,txt}"
```

- Profiling



A recipe for success

- Simplicity
- Flexibility
- Portability
- Reproducibility

All's well that ends well

YAMP, Yet Another Metagenomics Pipeline

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

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

Acknowledgements

Mario Falchi
Tiphaine Martin

Paolo Di Tommaso

@_alesssia
alessia.visconti@kcl.ac.uk