Amplicon sequencing analyses & dadasnake

Anna Heintz-Buschart - 5 March 2024

a.u.s.heintzbuschart@uva.nl

github.com/a-h-b

twitter.com/_a_h_b_

dadasnake

University of Amsterdam
Life Sciences

Overview of today

- Short intro
- Scope

- From sample to data - challenges for interpretation and analysis
  - what happens to the representation of microbial cells
  - biases
  - sources of error
  - detection limits

ask me anything

- Data processing and dadasnake
  - demo
  - details

- Discussion/Questions

About me

projects | **software**

soil biodiversity

fungi in plant diversity gradients

plant-associated microbial communities

human & primate microbiomes

amplicon sequencing

genome binning

pangenomics

integrated metagenomics & -transcriptomics

iSBio consortium

dadasnake

THE JENA EXPERIMENT

gcef Global Change Experimental Facility

MicroHealth

binny

MICROP Harnessing the second genome of plants

DECRyPT

SPP 2089

GALORIOUS

PanAf Programme panafrican.eva.mpg

RPA Systems Biology Host-Microbiome Interactions

IMP3

RPA Microbiome Health

2008

MSc Biology (Microbiology, Botany, Molecular & Cell Biology)

2011

PhD: Fungal human pathogen
- compound screening, mode-of action
- gene expression analysis

2012

Postdoc: Gene regulatory network modelling

Postdoc: Integrated meta-omics
- human microbiome, wastewater treatment
- metagenomics, metatranscriptomics, metaproteomics
- lab automation
- bioinformatics pipelines

2017

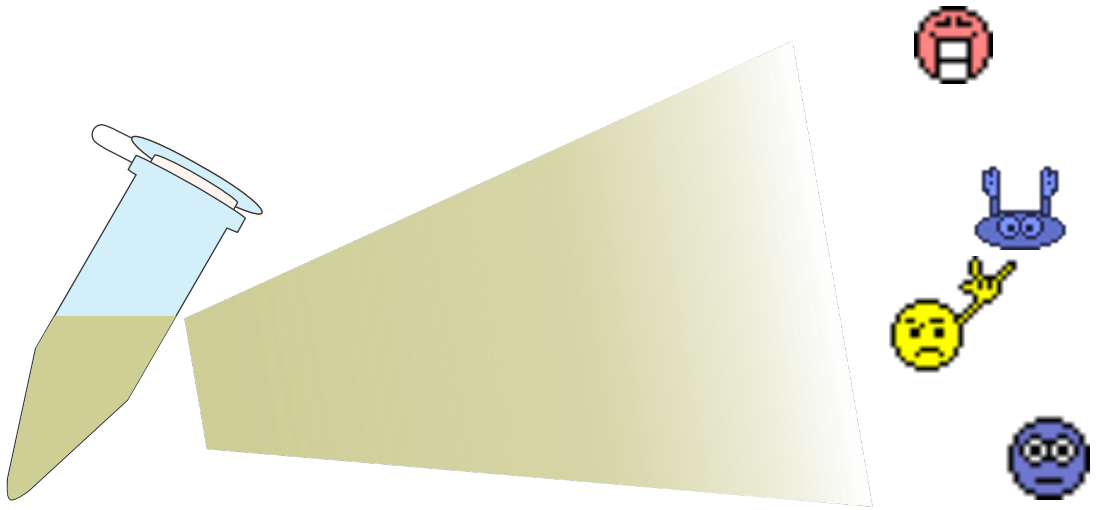
Metagenomics support:
- biodiversity
- soil, plants, animal microbiomes
- bioinformatics pipelines
- data integration

2021

Assistant Prof Microbial Metagenomics
- meta-omics integration
- human and plant microbiomes

Questions

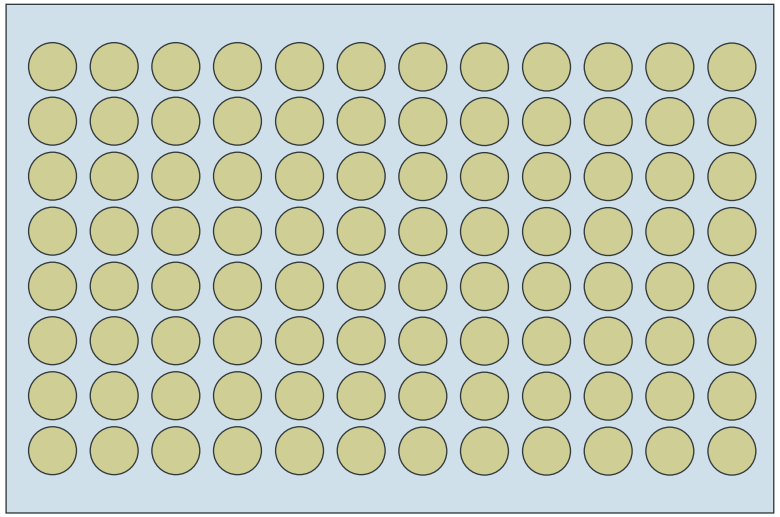
mixed community

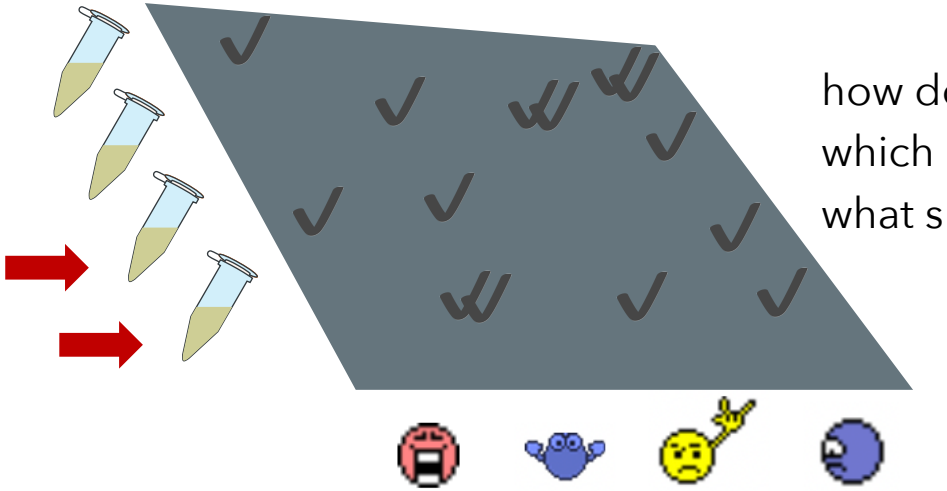


what is in my sample?

who is in my community?

what is in my samples?





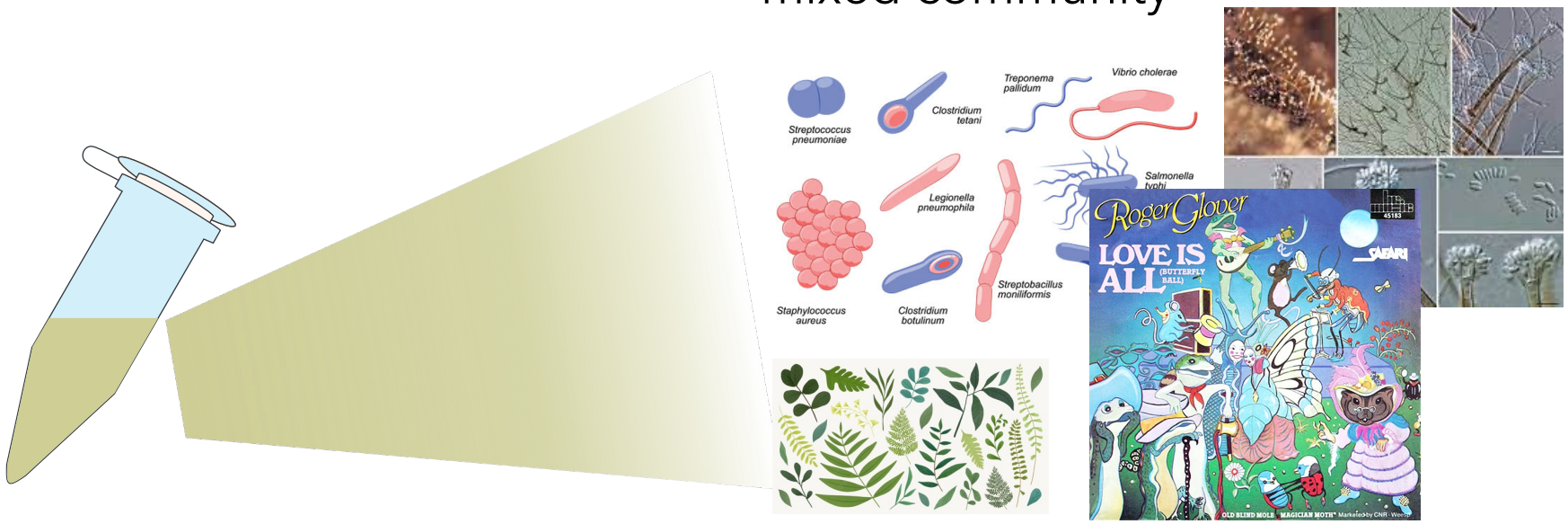
how do my samples compare?
which of my samples are similar?
what shapes my samples?

when are they there?

who is there often/in high numbers?
who is there with whom?

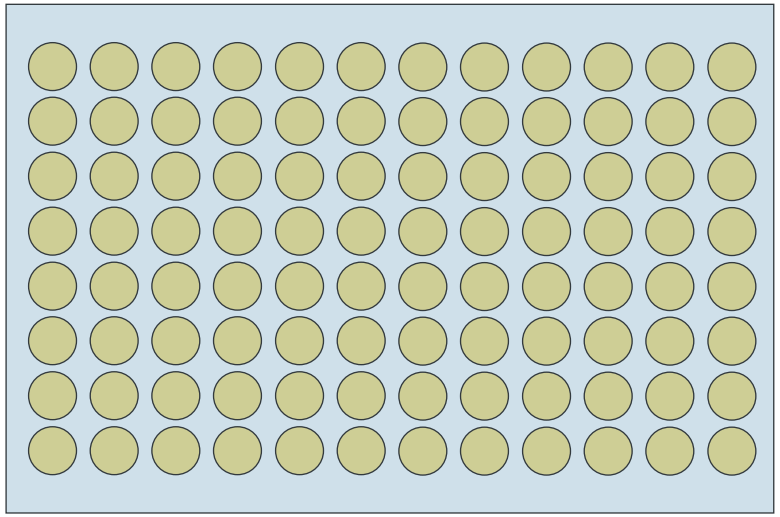
Questions

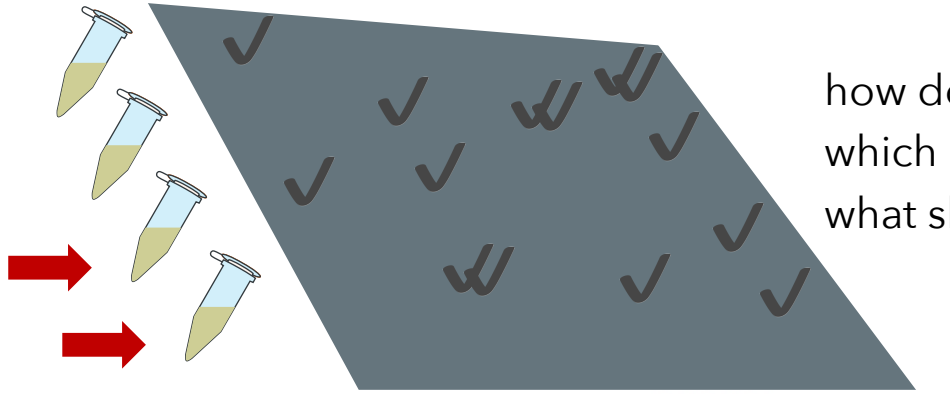
mixed community



what is in my sample?

what is in my samples?





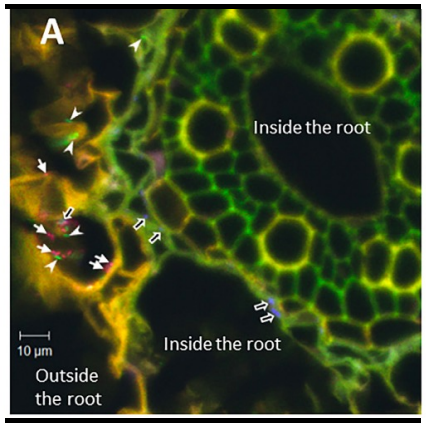
how do my samples compare?
which of my samples are similar?
what shapes my samples?

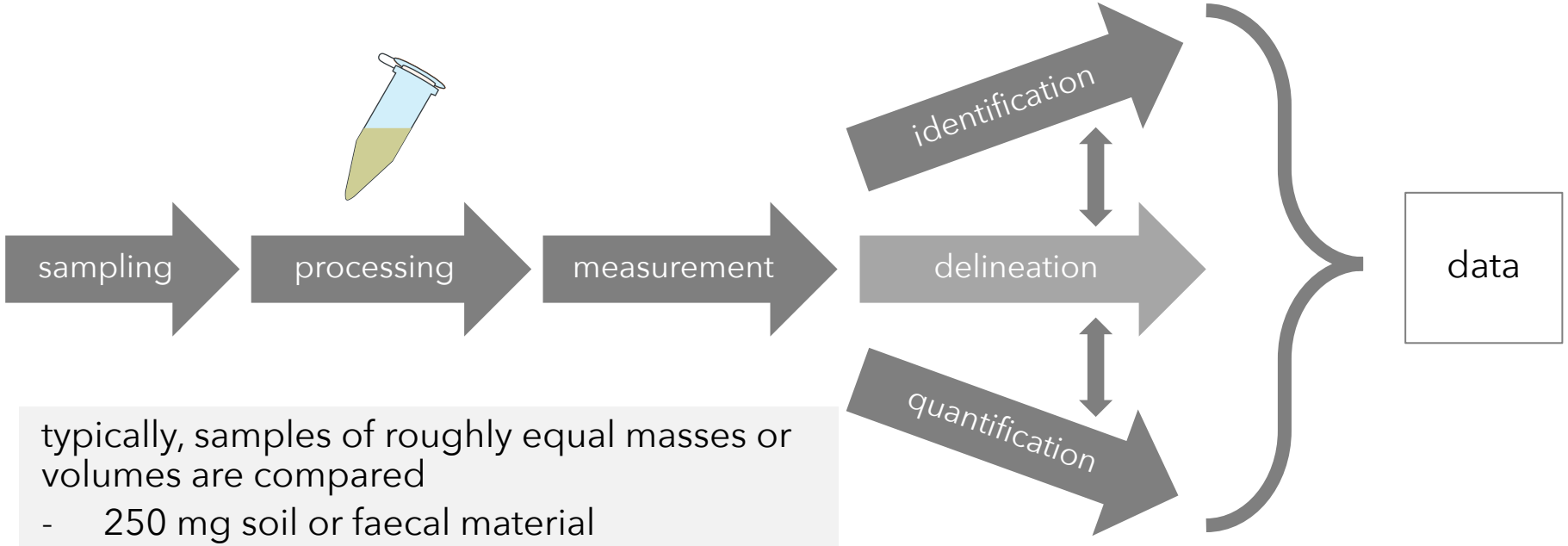
when are they there?

who is there often/in high numbers?
who is there with whom?

Workflow

N.L. Castanheira et al. / Microbiological Research 198 (2017) 47–55





typically, samples of roughly equal masses or volumes are compared

- 250 mg soil or faecal material
- 1 ml saliva
- bacteria on one oral/rectal/vaginal swap

similar cell numbers are **assumed**

- $2.5 * 10^7$-10^8 bacterial cells in soil sample
- $7 * 10^{10}$ bacteria in faecal sample
- DNA from about $2 * 10^6$ cells is input to measurements

Are samples representative?

typically, samples of roughly equal masses or volumes are compared

- 250 mg soil or faecal material
- 1 ml saliva
- bacteria on one oral/rectal/vaginal swap

similar cell numbers are **assumed**

- $2.5 * 10^7$-10^8 bacterial cells in soil sample
- $7 * 10^{10}$ bacte
- DNA from ab
  measurement

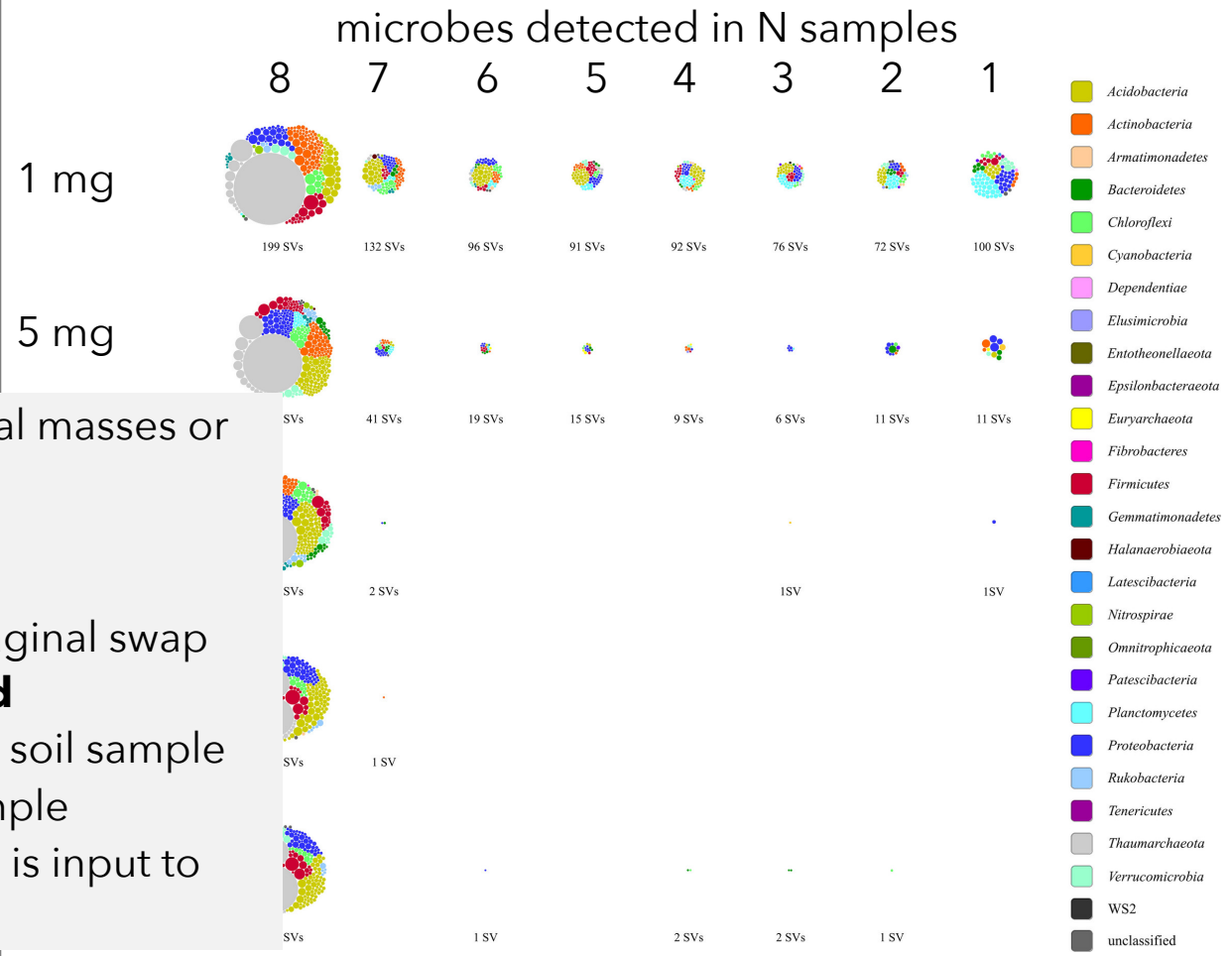
typically, samples of roughly equal masses or volumes are compared

- 250 mg soil or faecal material
- 1 ml saliva
- bacteria on one oral/rectal/vaginal swap

similar cell numbers are **assumed**

- $2.5 * 10^7$-10^8 bacterial cells in soil sample
- $7 * 10^{10}$ bacteria in faecal sample
- DNA from about $2 * 10^6$ cells is input to measurements

Example: Soil aggregates vs. regular samples:



Szoboszlay & Tebbe (2020) MicrobiologyOpen doi: 10.1002/mbo3.1144)

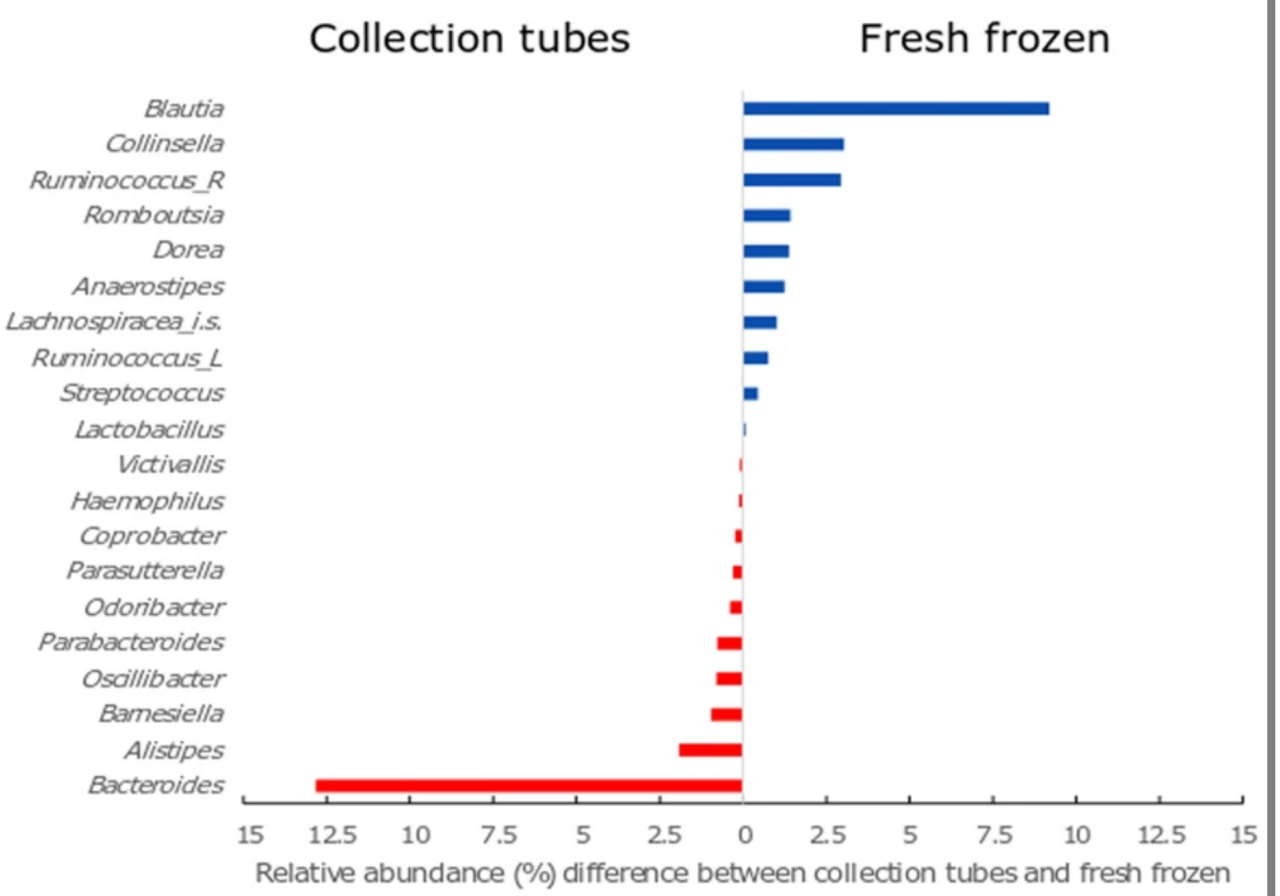
Is DNA representative?

typically, samples of roughly equal masses or volumes are compared

- 250 mg soil or faecal material
- 1 ml saliva
- bacteria on one oral/rectal/vaginal swap

similar cell numbers are **assumed**

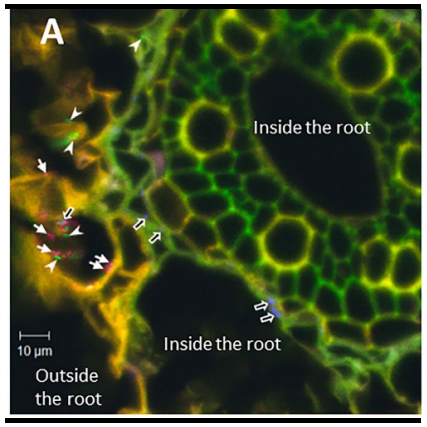
- $2.5 * 10^7$-10^8 bacterial cells in soil sample
- $7 * 10^{10}$ bacteria in faecal sample
- DNA from about $2 * 10^6$ cells is input to measurements

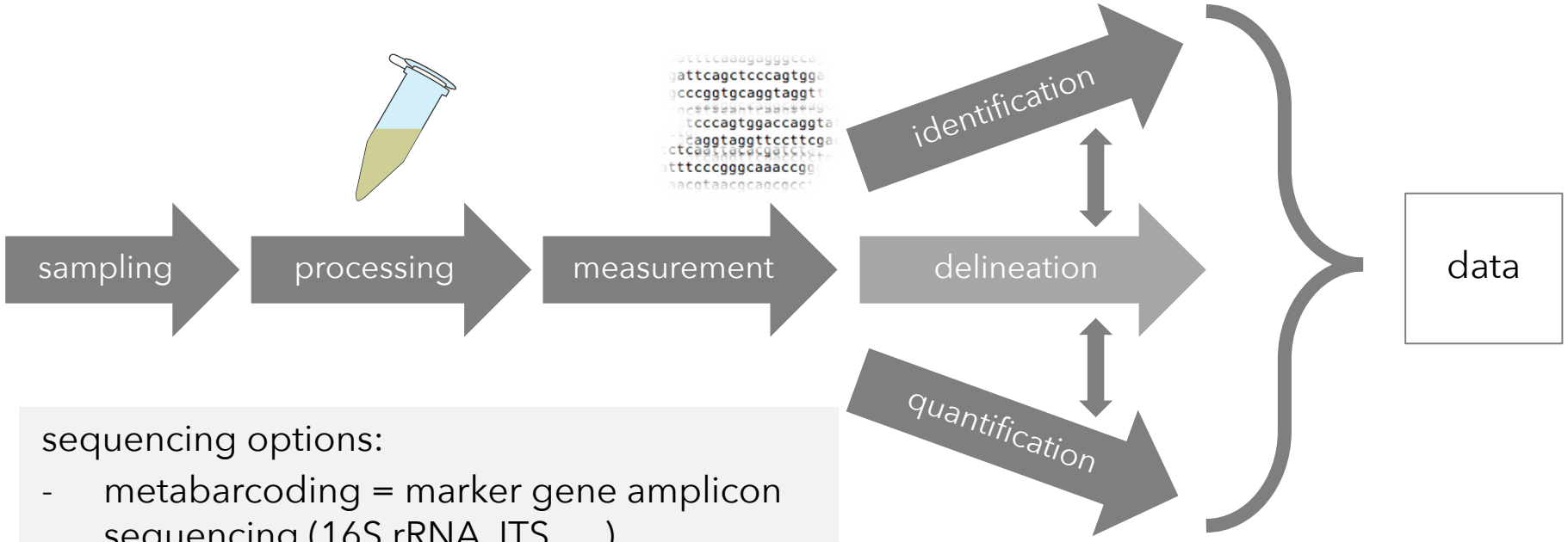


Watson *et al.*, 2019, Scientific Reports

Workflow

N.L. Castanheira et al. / Microbiological Research 198 (2017) 47–55



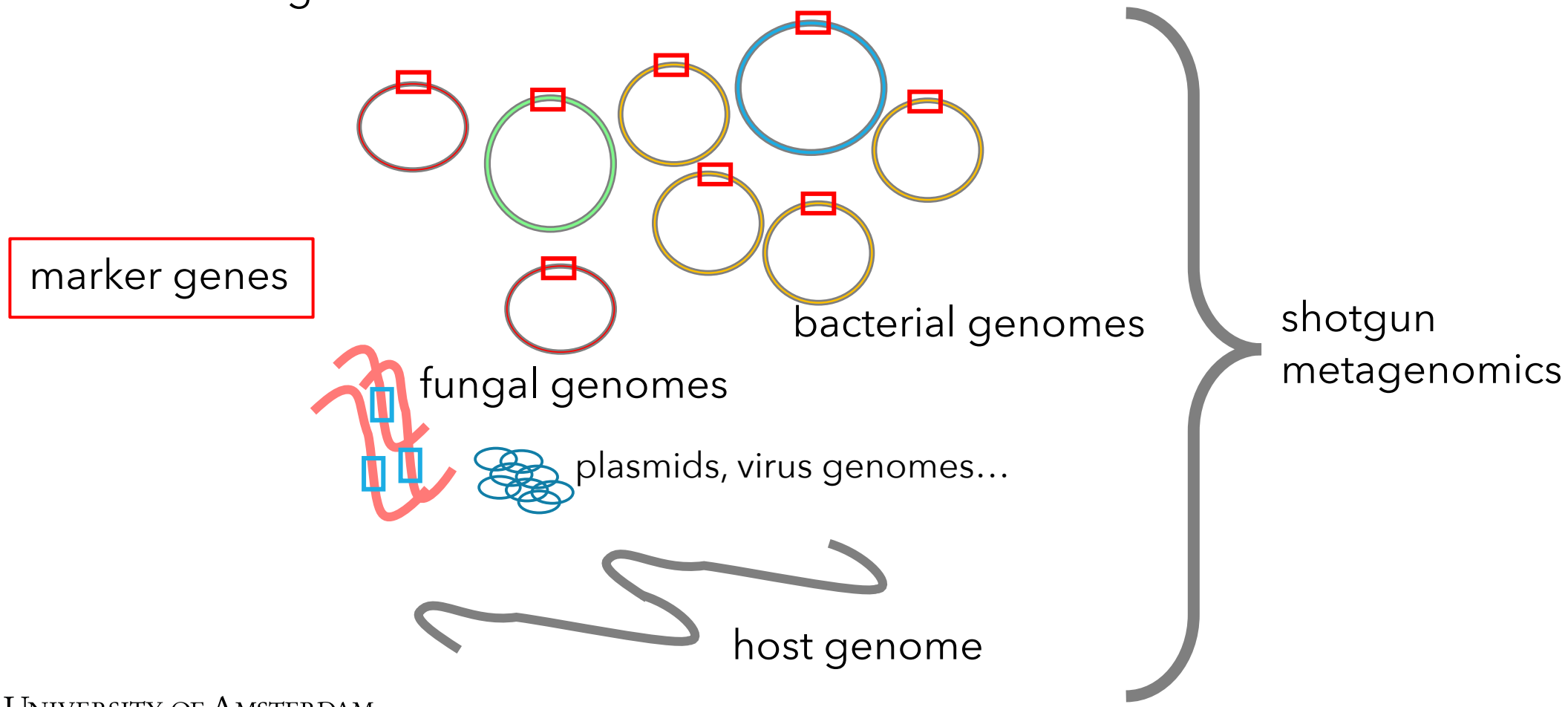


sequencing options:

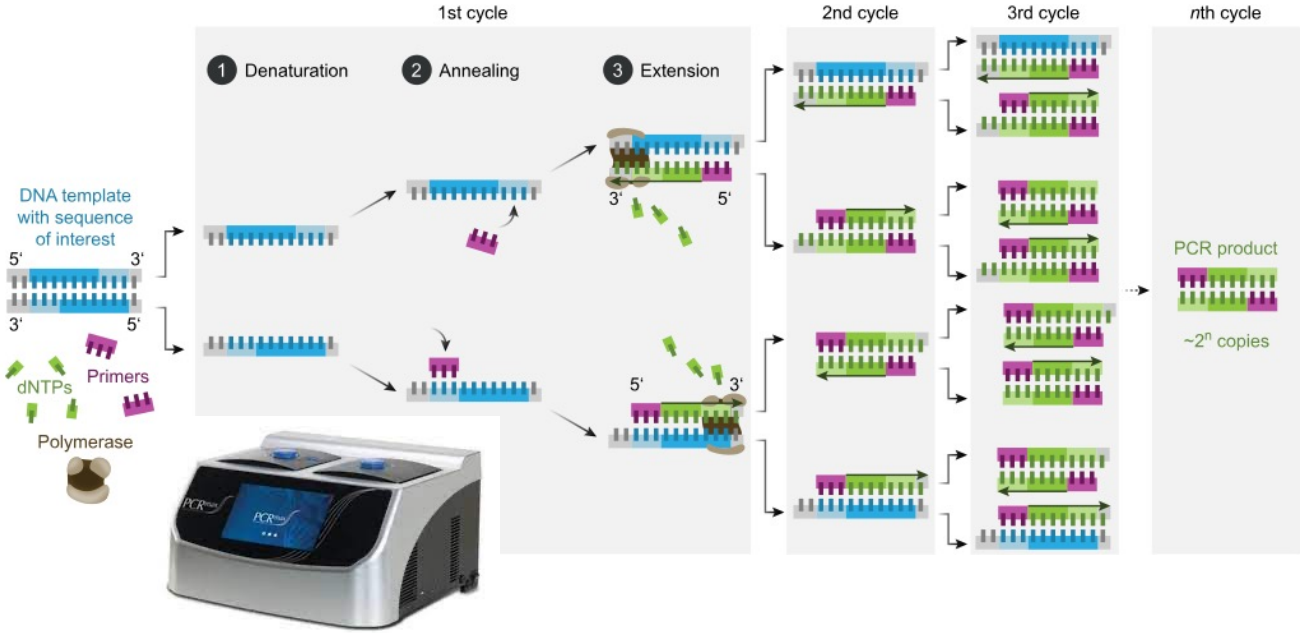
- metabarcoding = marker gene amplicon sequencing (16S rRNA, ITS, …)
- shotgun metagenomics

Amplicon sequencing vs metagenomics

the 'metagenome':



Marker gene amplification



marker gene pre-requisites:

- conserved regions for primers to bind
- variable regions with suitable phylogenetic resolution
- similar mutation rates across all measurable taxa
- no horizontal gene transfer
- suitable length for sequencing

classical target: 16S rRNA

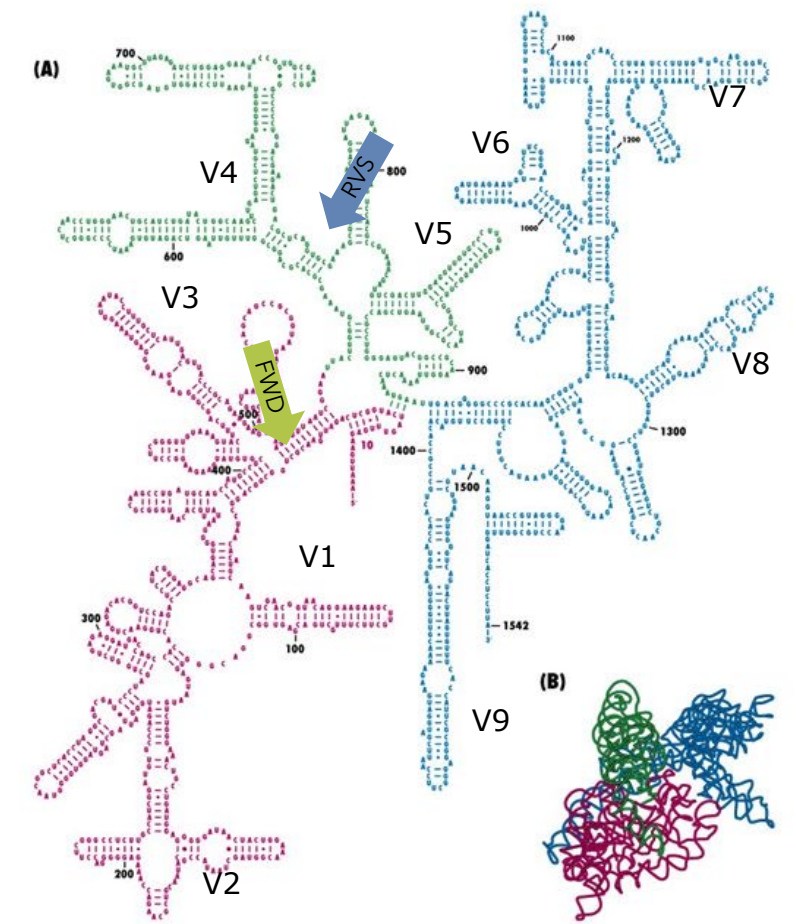
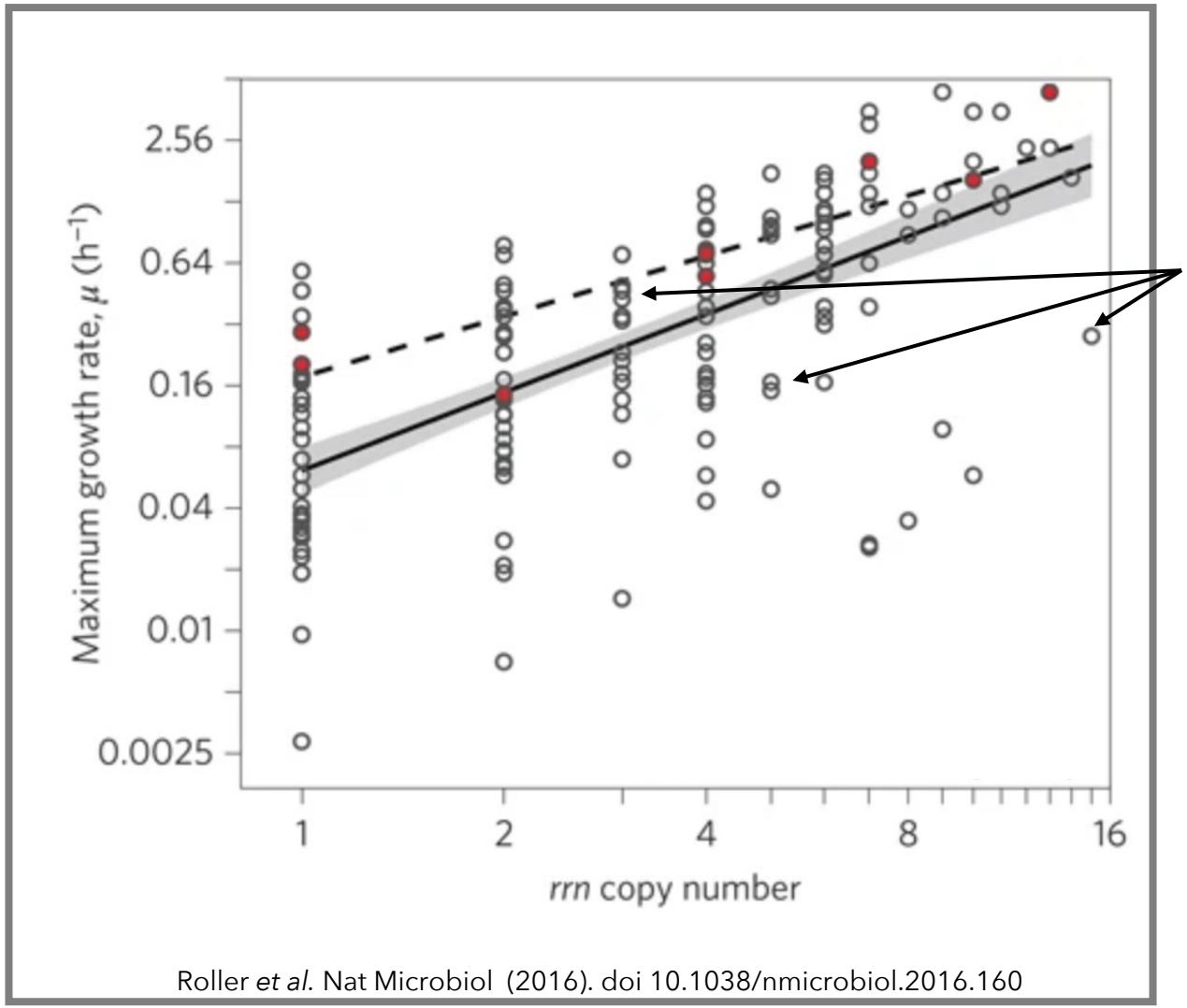


Figure 30-14
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

Are amplicons representative (i)?

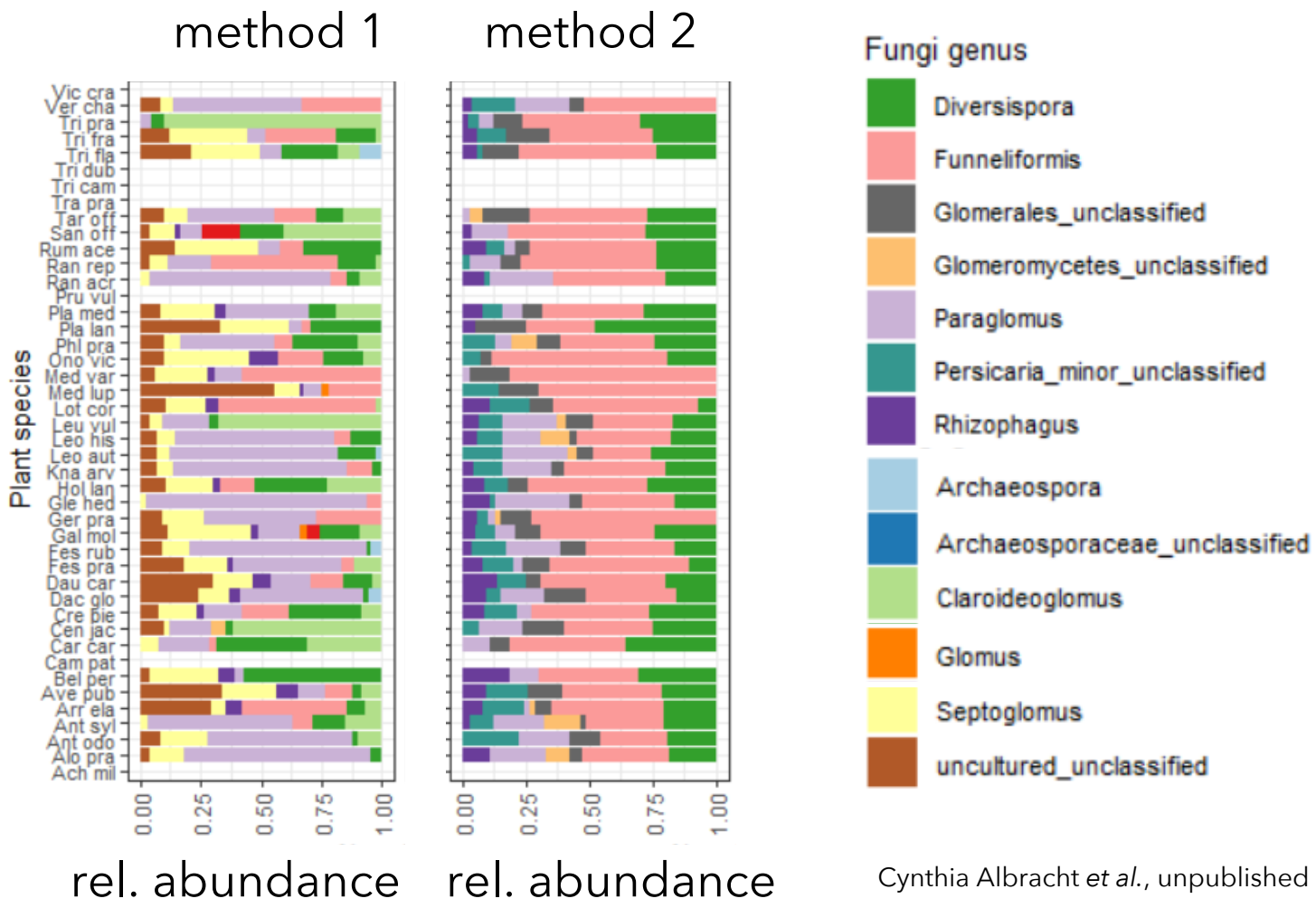


Roller *et al.* Nat Microbiol (2016). doi 10.1038/nmicrobiol.2016.160

1 copy ≠ 1 cell

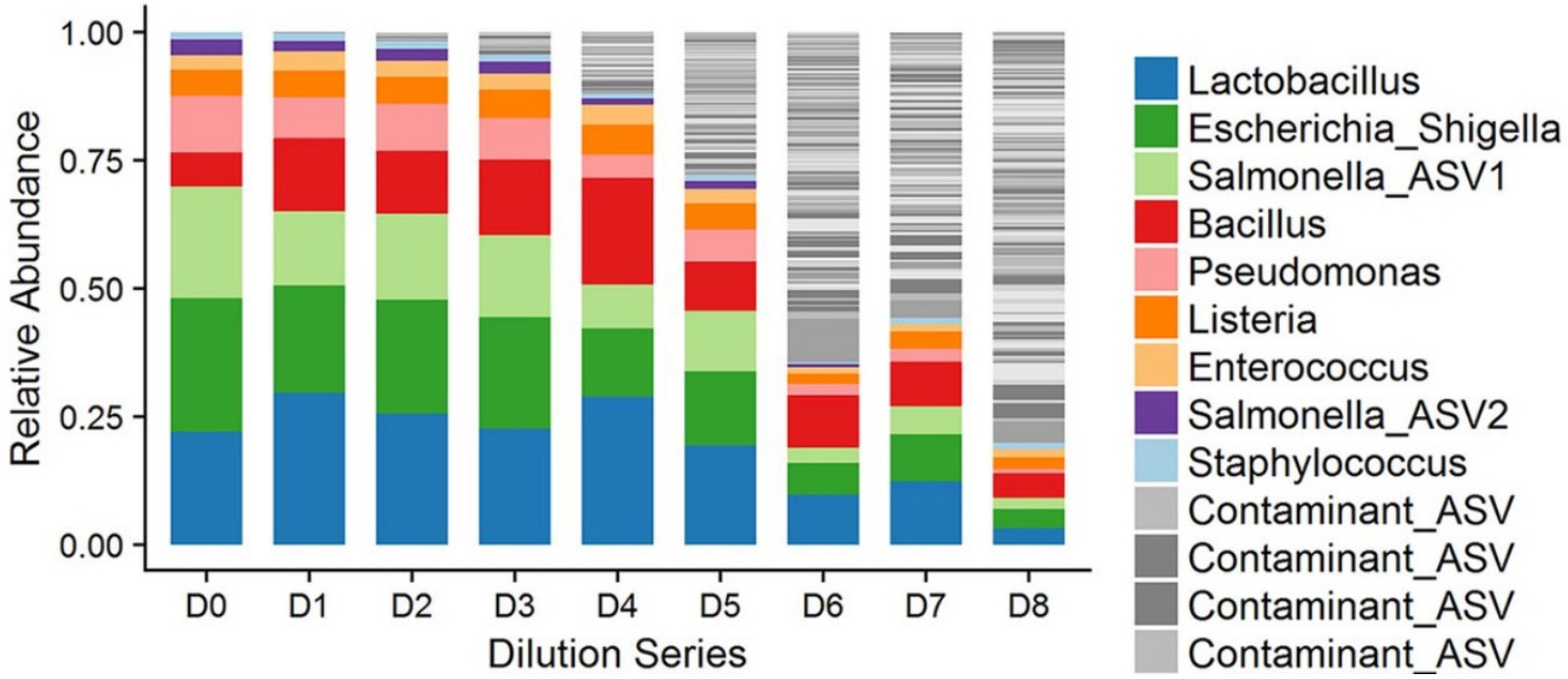
Are amplicons representative (ii)?

amplification bias



Cynthia Albracht *et al.*, unpublished

Are amplicons representative (iii)?

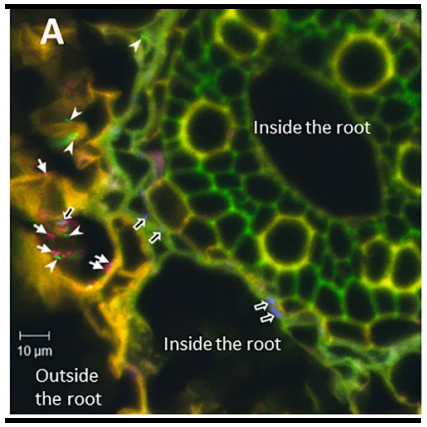


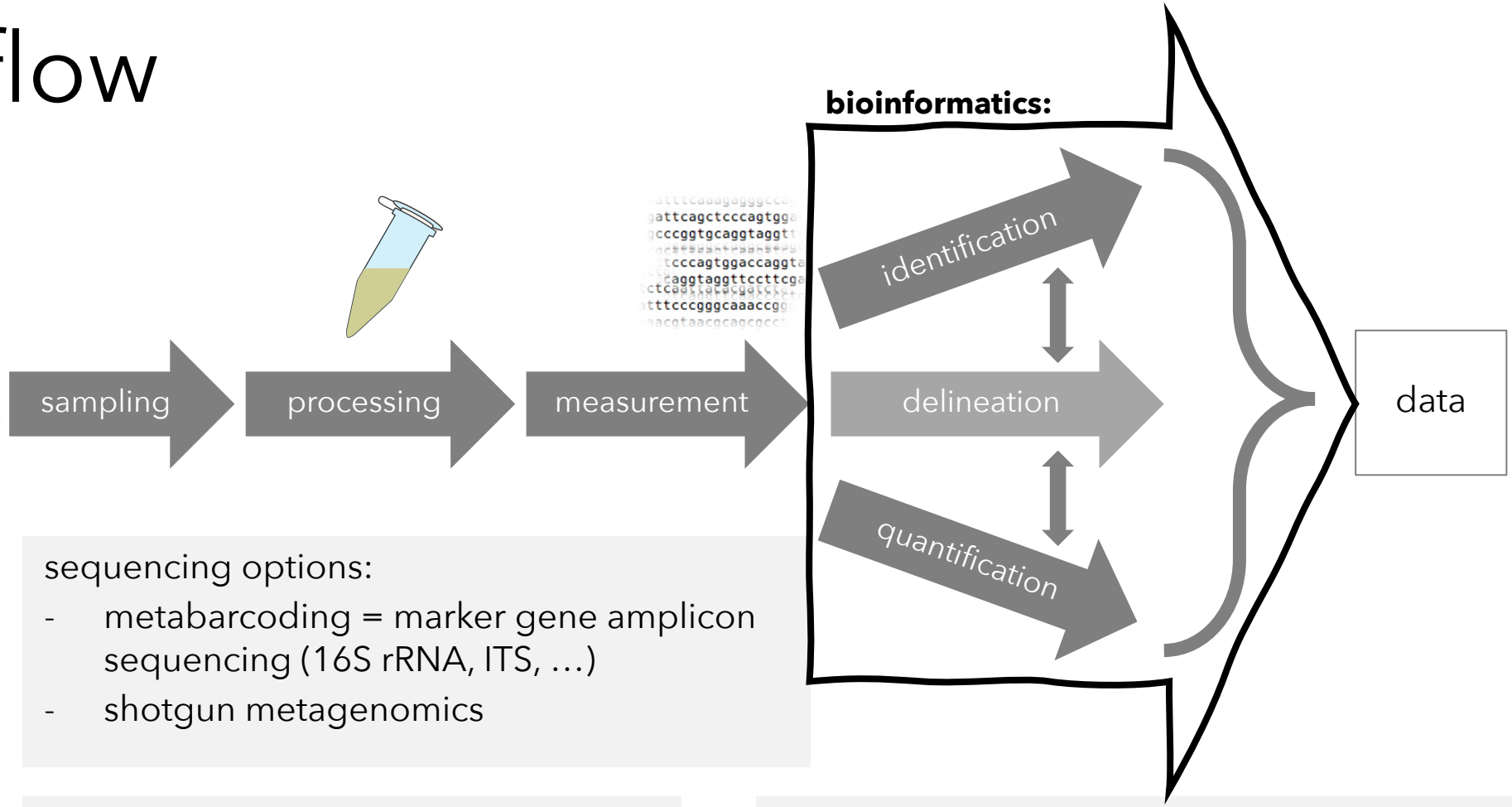
Karstens L et al. mSystems. 2019 doi: 10.1128/mSystems.00290-19.

amplification bias + contaminants

Workflow

N.L. Castanheira et al. / Microbiological Research 198 (2017) 47–55





sequencing options:

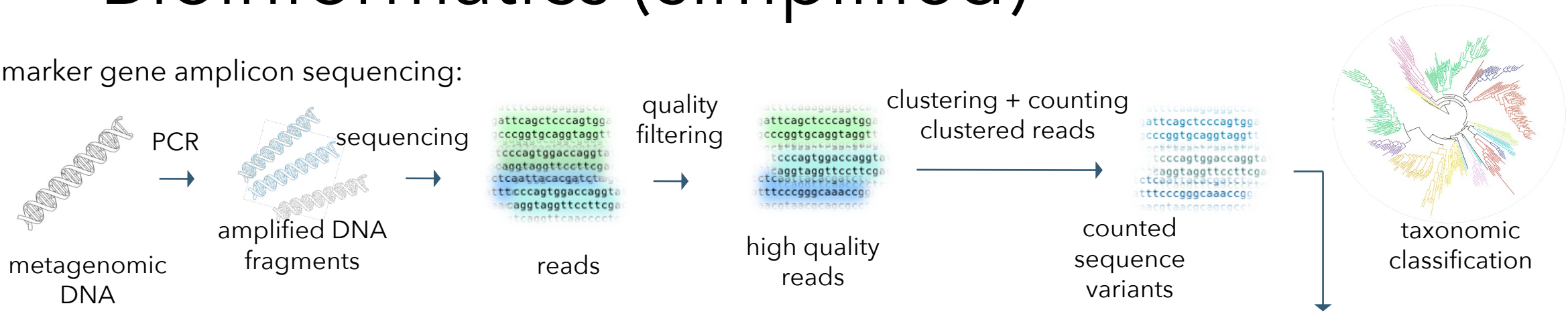
- metabarcoding = marker gene amplicon sequencing (16S rRNA, ITS, …)
- shotgun metagenomics

marker gene amplicon sequencing:
- 10,000 - 200,000 reads per sample
~ 1,000 - 200,000 cells

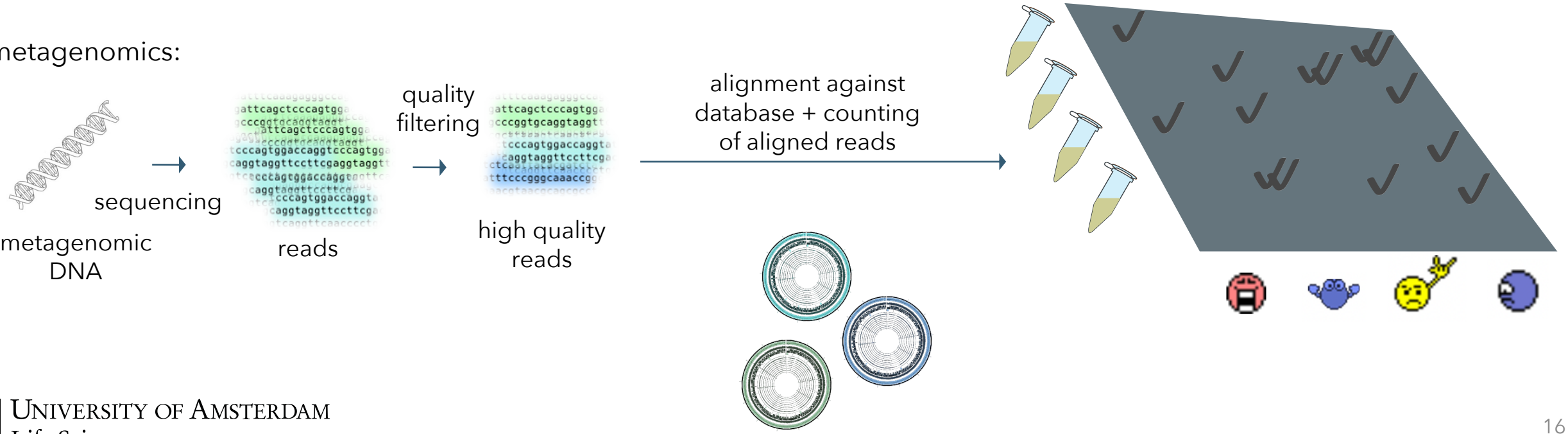
shotgun metagenomics:
- 1,000,000 - 100,000,000 reads per sample
(~ 100 - 10,000 cells)

Bioinformatics (simplified)

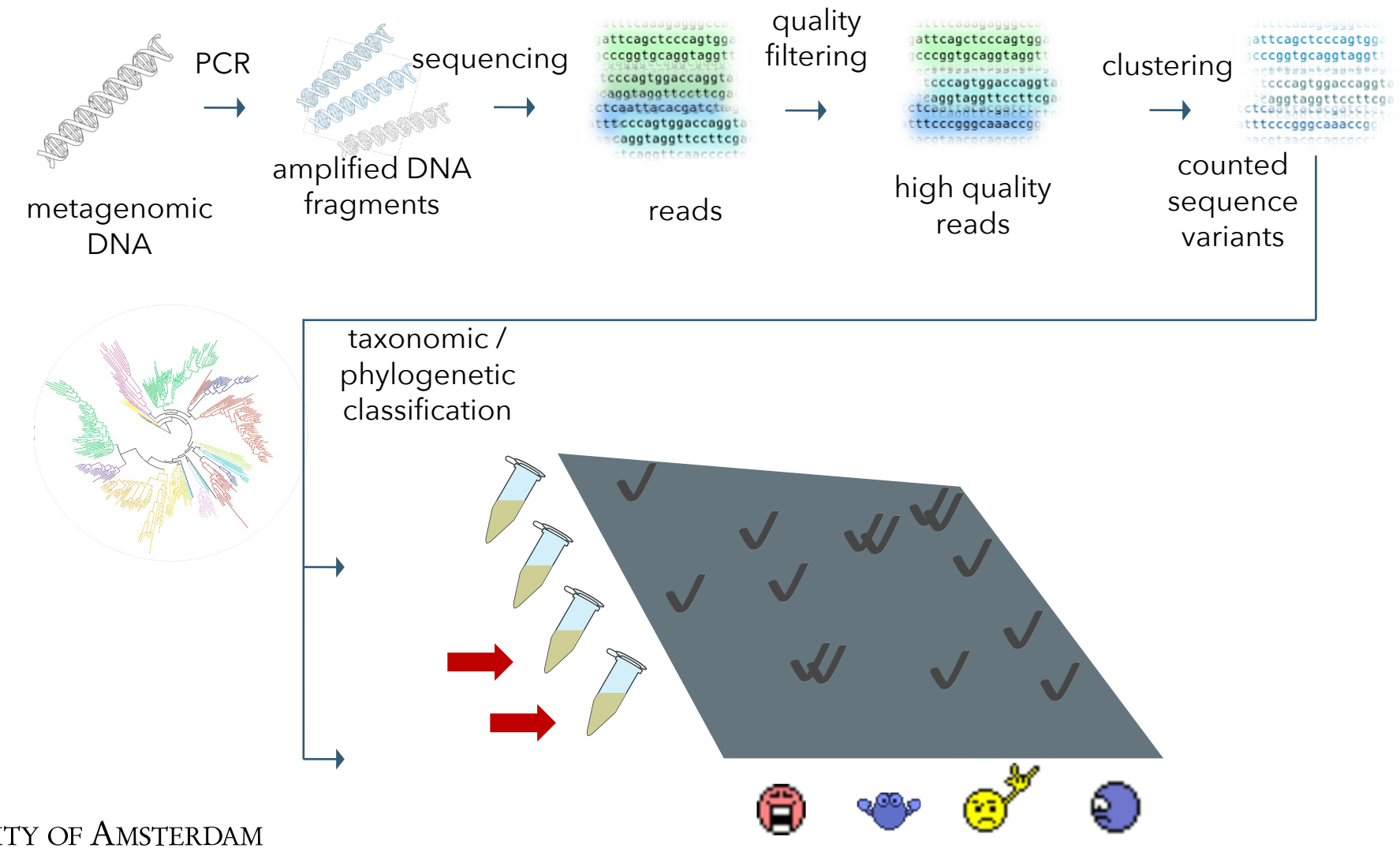
marker gene amplicon sequencing:



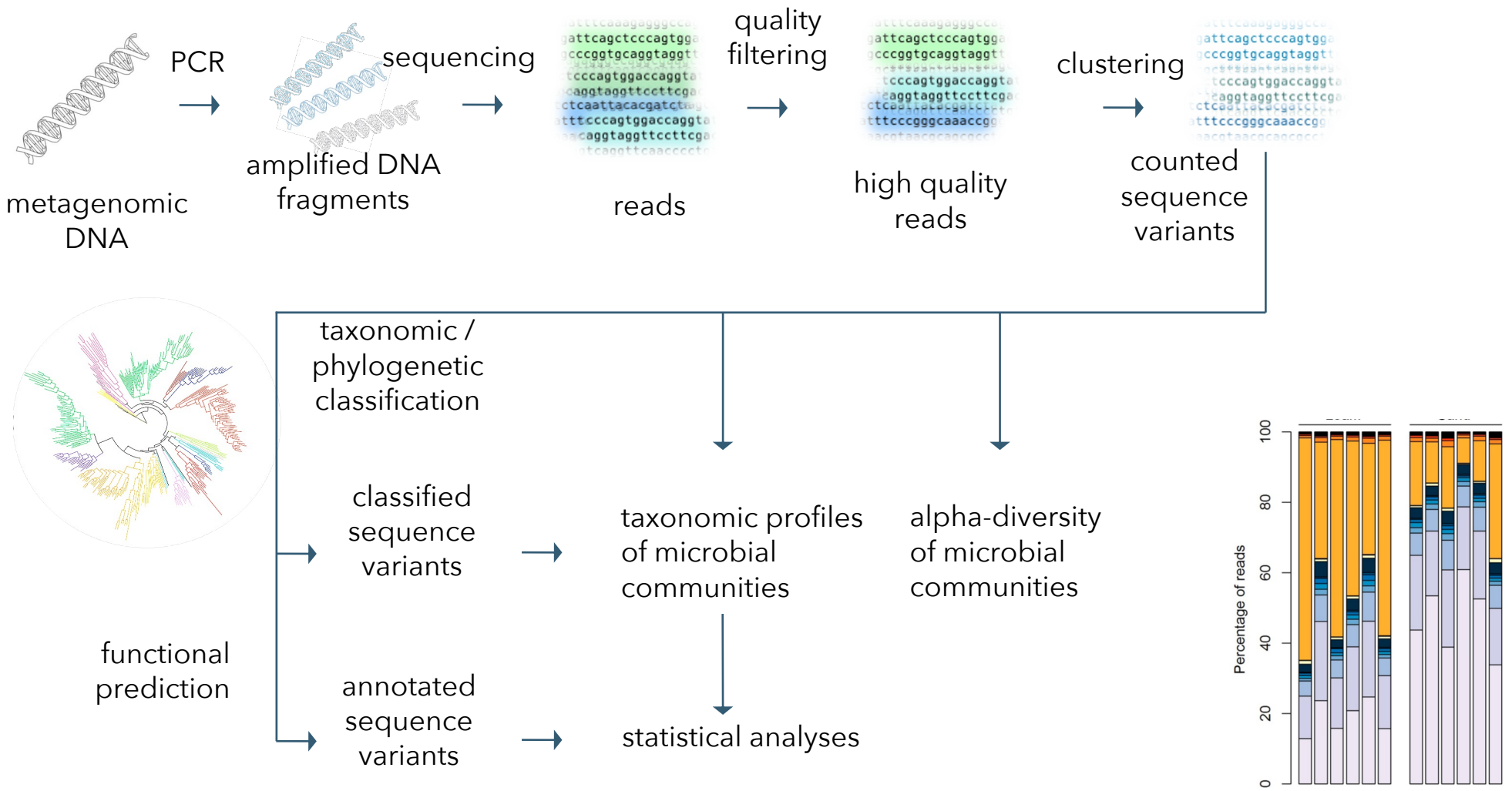
metagenomics:



Metabarcoding workflow

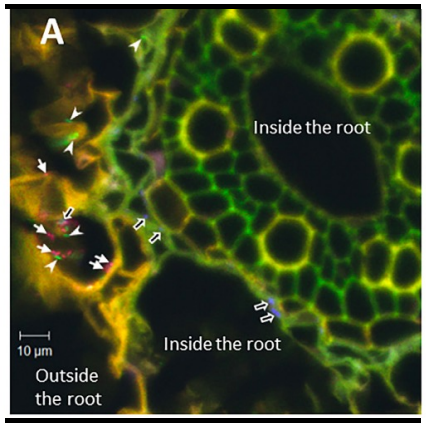


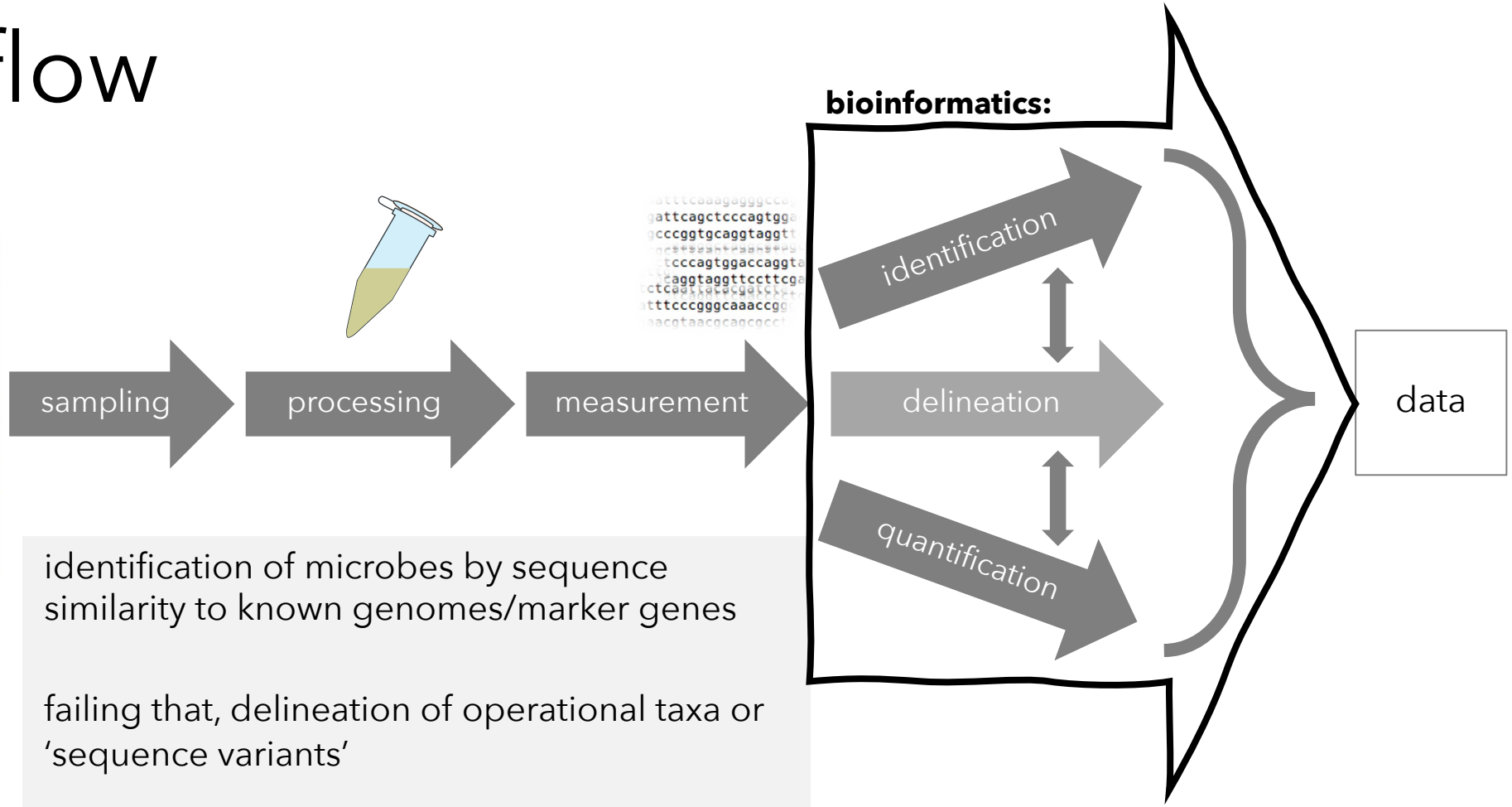
Metabarcoding workflow



Workflow

N.L. Castanheira et al. / Microbiological Research 198 (2017) 47–55





identification of microbes by sequence similarity to known genomes/marker genes

failing that, delineation of operational taxa or 'sequence variants'

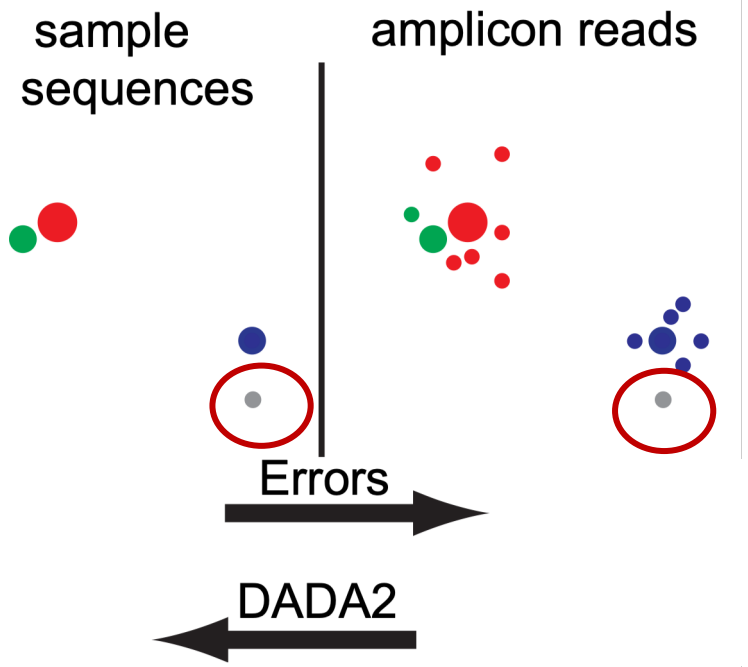
counting of reads covering genomes/marker genes

Amplicon sequencing units: ASVs (aka. ESVs, zOTUs)

- **A**mplicon **S**equence **V**ariants
  = **E**xact **S**equence **V**ariants
  = **z**ero-radius **OTU**s

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                 |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

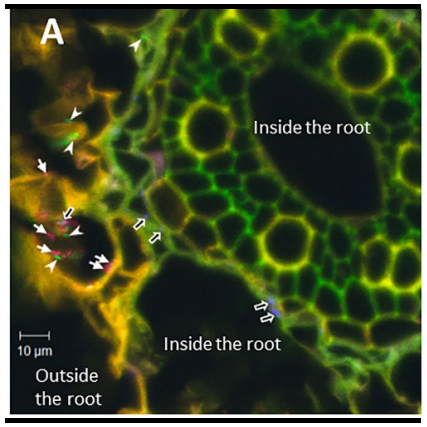
$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

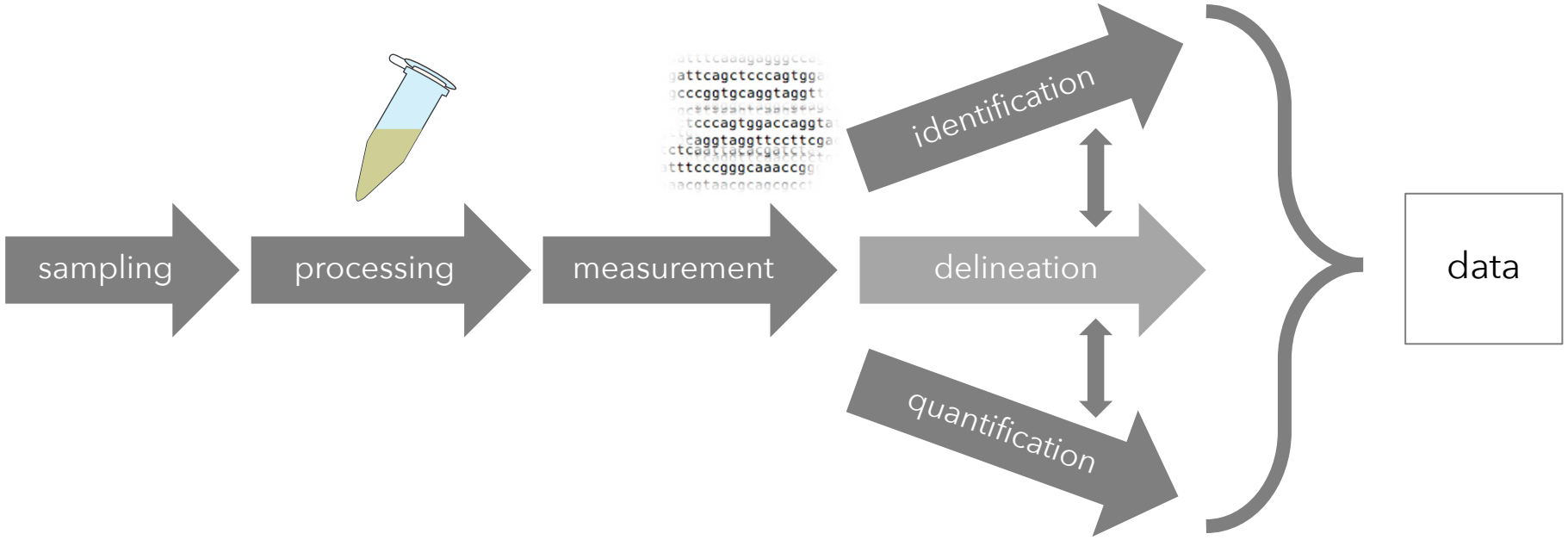


© B. Callahan (DADA2 documentation)

Workflow

N.L. Castanheira et al. / Microbiological Research 198 (2017) 47–55



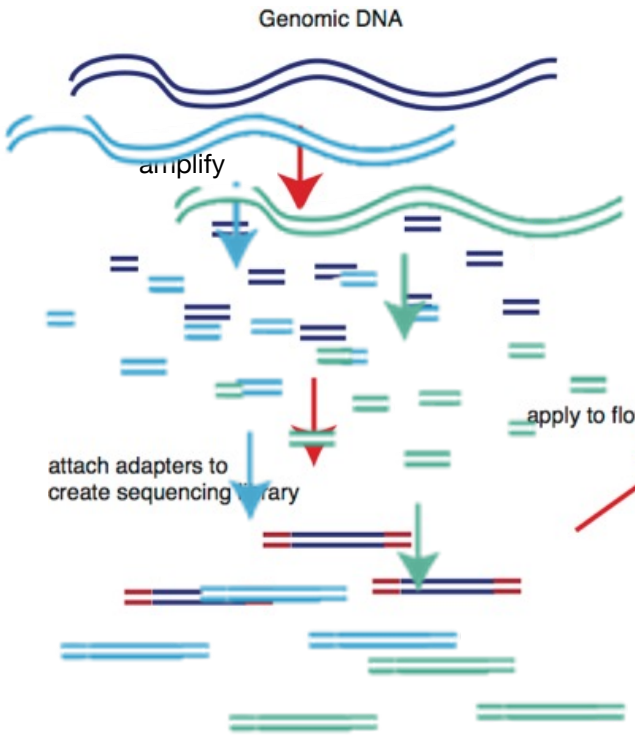


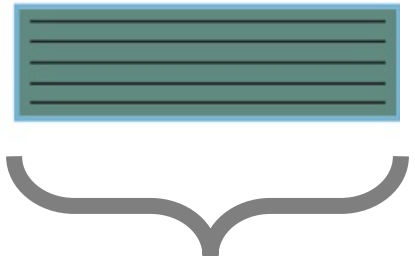
marker gene amplicon sequencing:
- 10,000 - 200,000 reads per sample
< 1,000 - 200,000 cells

shotgun metagenomics:
- 1,000,000 - 100,000,000 reads per sample
< 100 - 10,000 cells

low-abundant taxa can end up below the detection limit

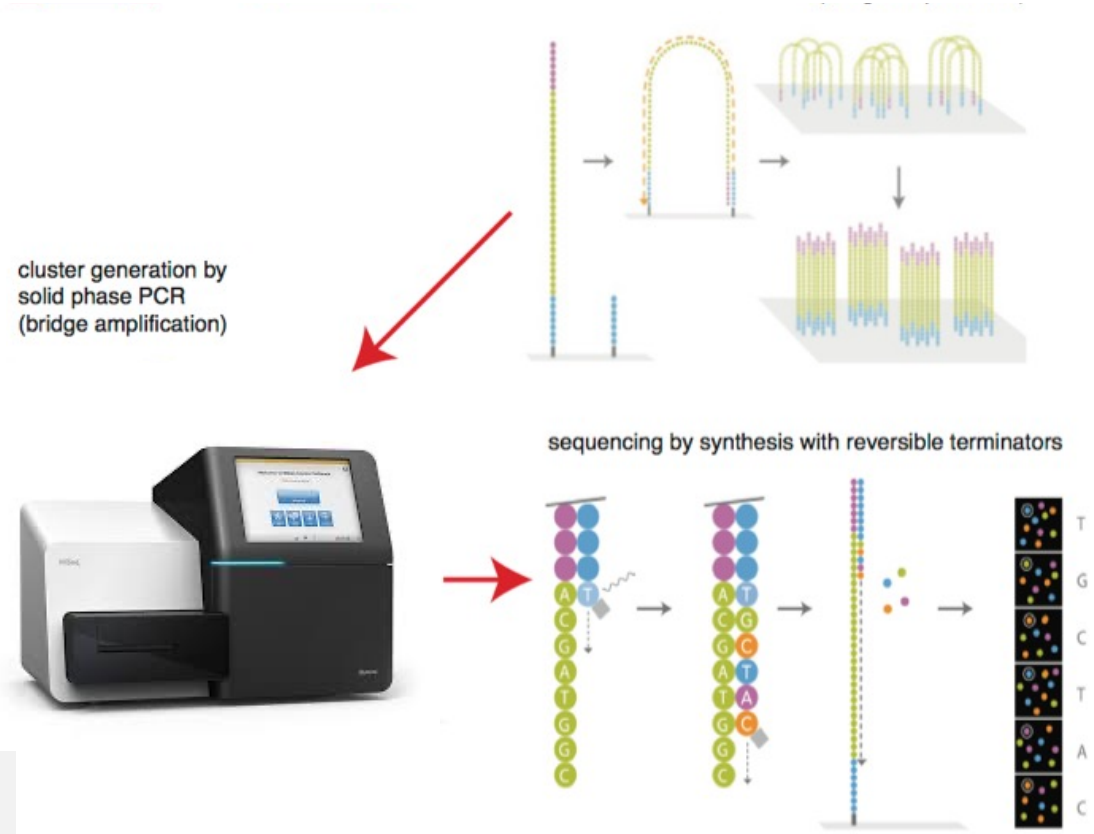
Sequencing depth





limited space: maximum number of reads per run

- e.g.: 14,000,000



total number of reads per sample is a choice (+ result of imprecise dilution / mixing)

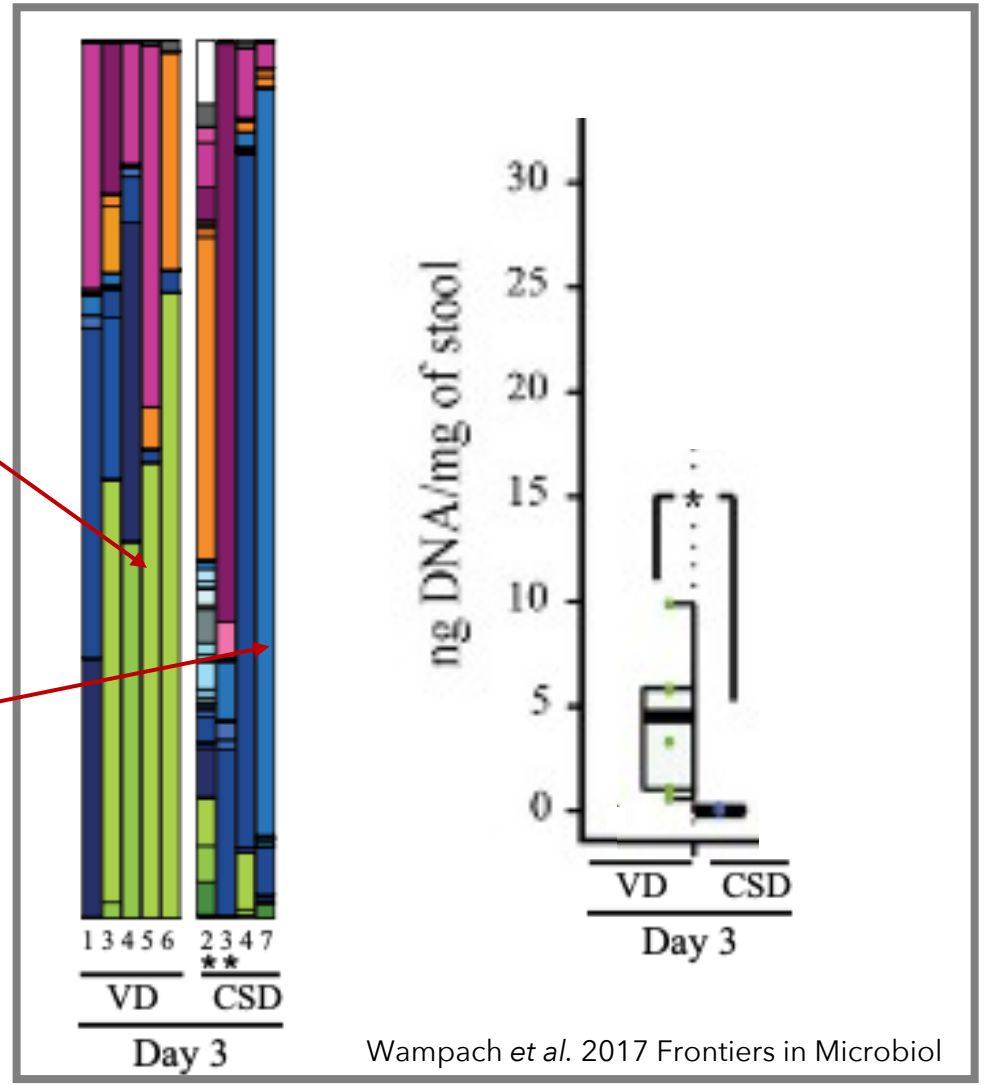
adapted from https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/

(most of the time)
Microbiome data are frequency estimates

are these missing in the other samples?

OR

are these a lot more abundant in here?

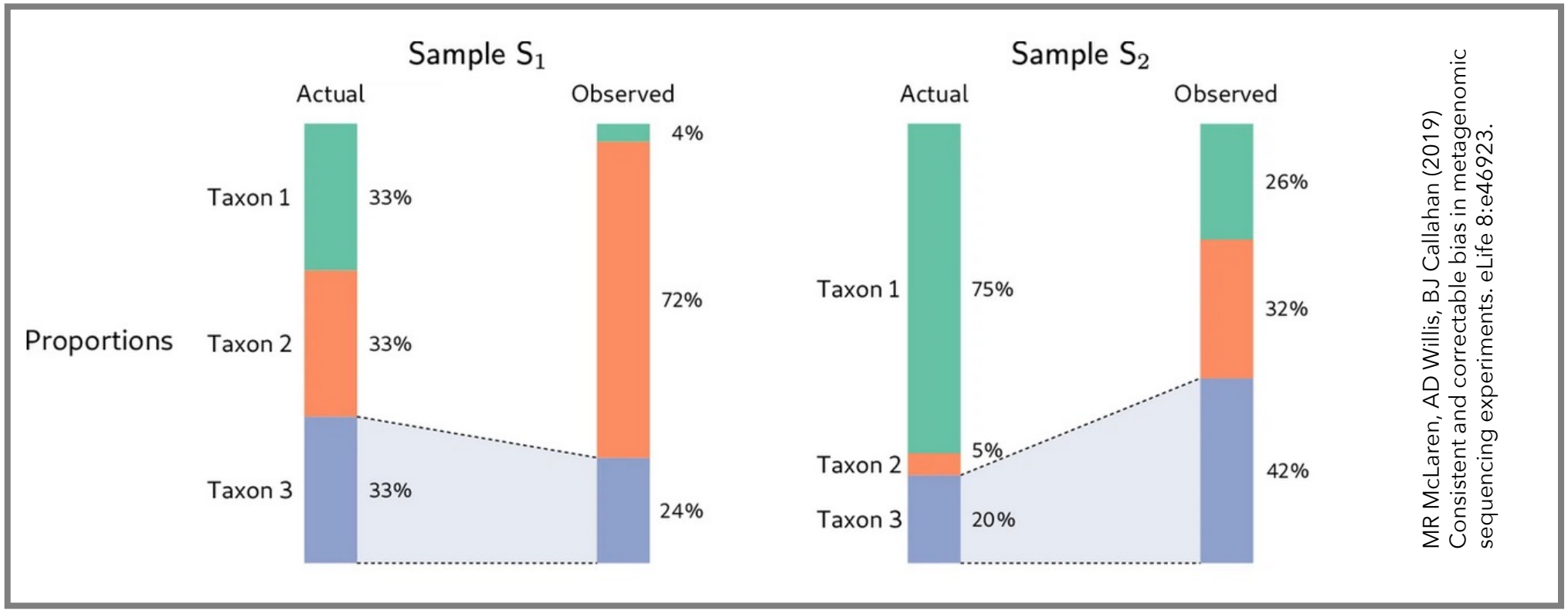


Wampach *et al.* 2017 Frontiers in Microbiol

overgrowth of a (set of) microbe(s) can induce correlations between the others

Remember the amplification bias?

assuming consistent bias per taxon:

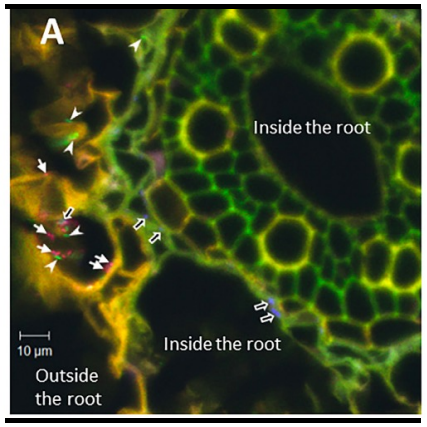


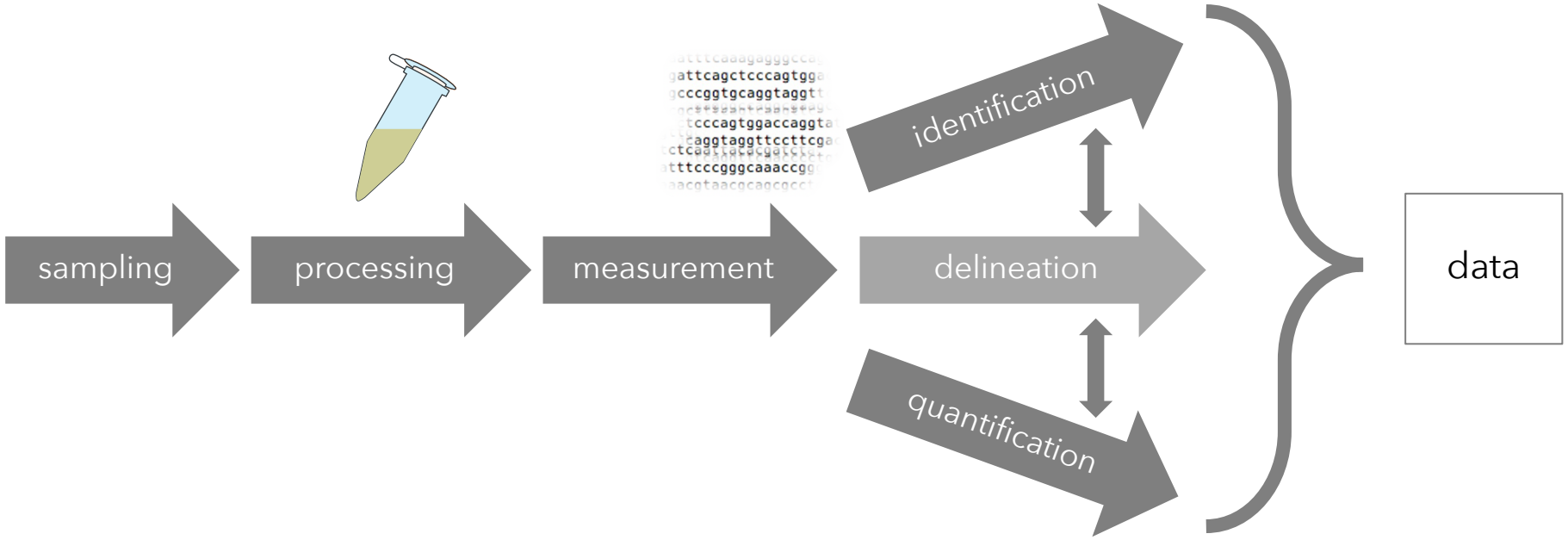
MR McLaren, AD Willis, BJ Callahan (2019) Consistent and correctable bias in metagenomic sequencing experiments. eLife 8:e46923.

the context can make a declining population look increasing

Workflow - possible improvements

N.L. Castanheira et al. / Microbiological Research 198 (2017) 47–55





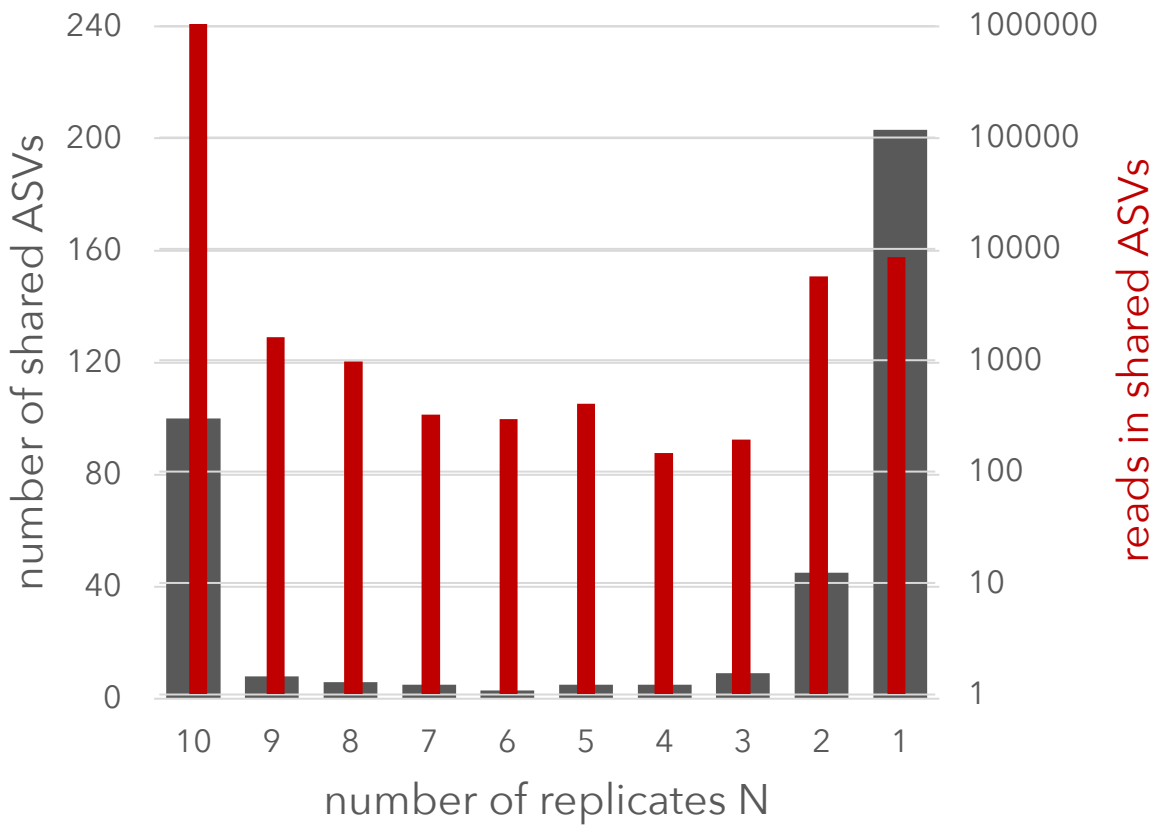
quantitative add-ons, e.g., spike-in standards, unique molecular identifiers, qPCR

copy-number correction

ratios or CLR

Where do 0's come from?

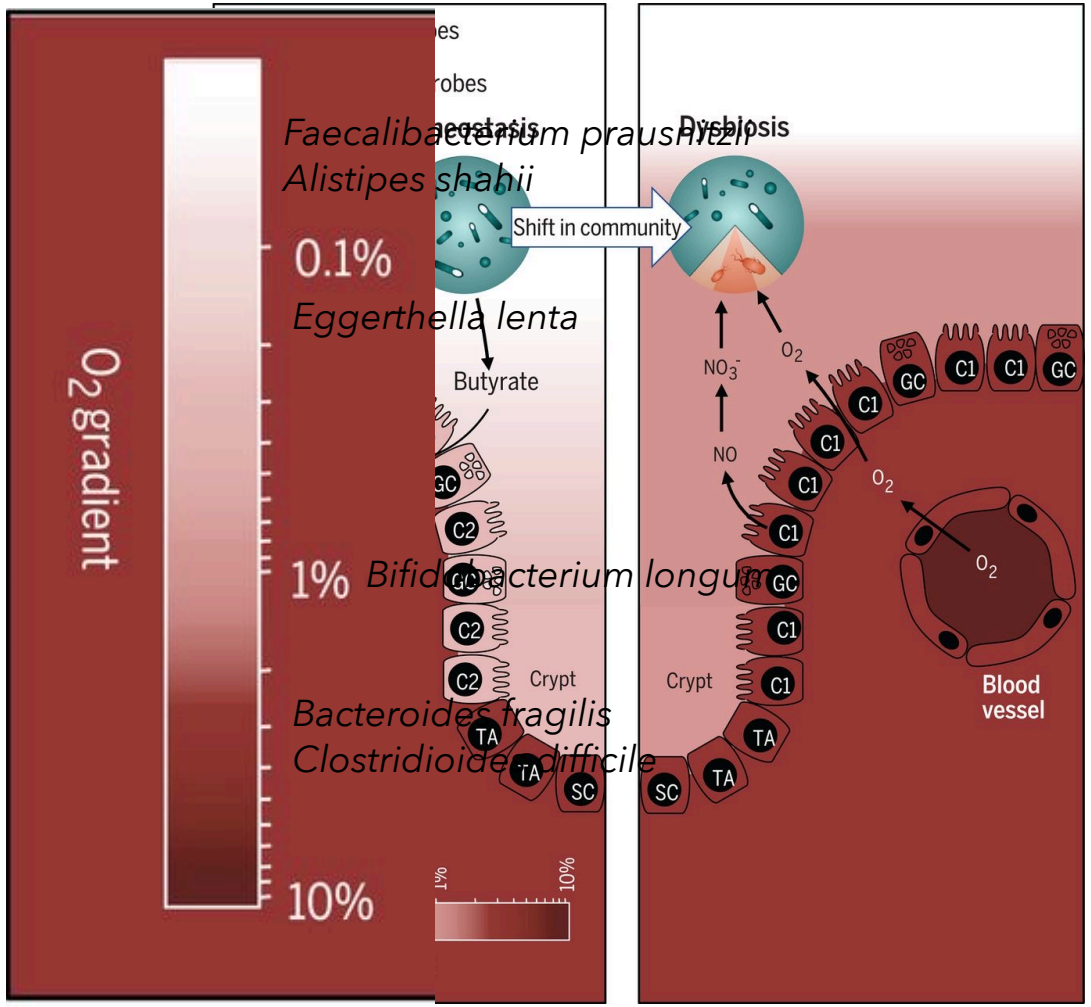
- there are always low-abundant taxa that randomly fall under the detection limit in some samples
- sequencing errors can lead to spurious ASVs



data from Heintz-Buschart et al. *Movement Disorders* (2018) doi 10.1002/mds.27105

Where do 0's come from?

- there are always low-abundant taxa that randomly fall under the detection limit in some samples
- sequencing errors can lead to spurious ASVs
- differences between microbiomes are much greater than between other biological samples



Litvak et al. Science (2018) doi 10.1126/science.aat9076

Overview of today

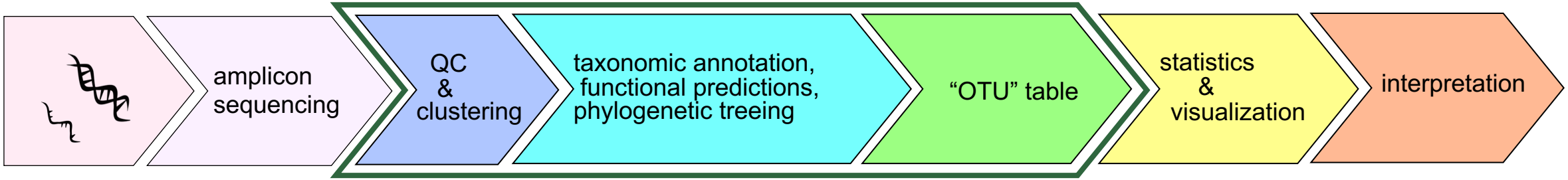
- Short intro
- Scope

- From sample to data - challenges for interpretation and analysis
  - what happens to the representation of microbial cells
  - biases
  - sources of error
  - detection limits

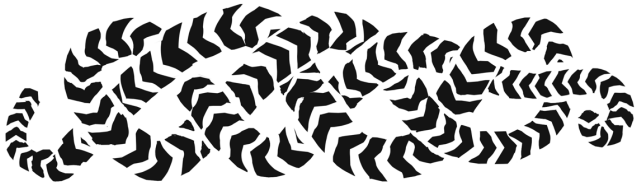
- Data processing and dadasnake
  - demo
  - details

- Discussion/Questions

dadasnake pipeline



dadasnake



- https://github.com/a-h-b/dadasnake

OXFORD (GIGA)n SCIENCE

GigaScience, 9, 2020, 1–8

doi: 10.1093/gigascience/giaa135
Technical Note

TECHNICAL NOTE

Dadasnake, a Snakemake implementation of DADA2 to process amplicon sequencing data for microbial ecology

Christina Weißbecker[1], Beatrix Schnabel[1] and Anna Heintz-Buschart[1,2,*]

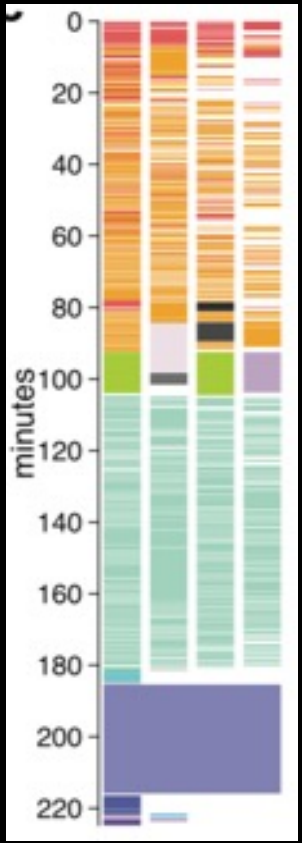
- doi: 10.1093/gigascience/giaa135

dadasnake pipeline – aim & ambition

- wrap DADA2 + pre-/post-processing
- be more configurable than qiime2
- be able to use high-performance compute clusters
  = parallelisation, module-based, use big-mem
- be reproducible
- be open-source
- be low-maintenance for the developer

- be really easy to use

pipeline



Thanks to:

DADA2 Amplicon Sequencing. Exactly. Version 1.12

CONDA

git

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

UFZ
Christina Weißbecker
Bea Schnabel
Julia Moll
Kezia Goldmann

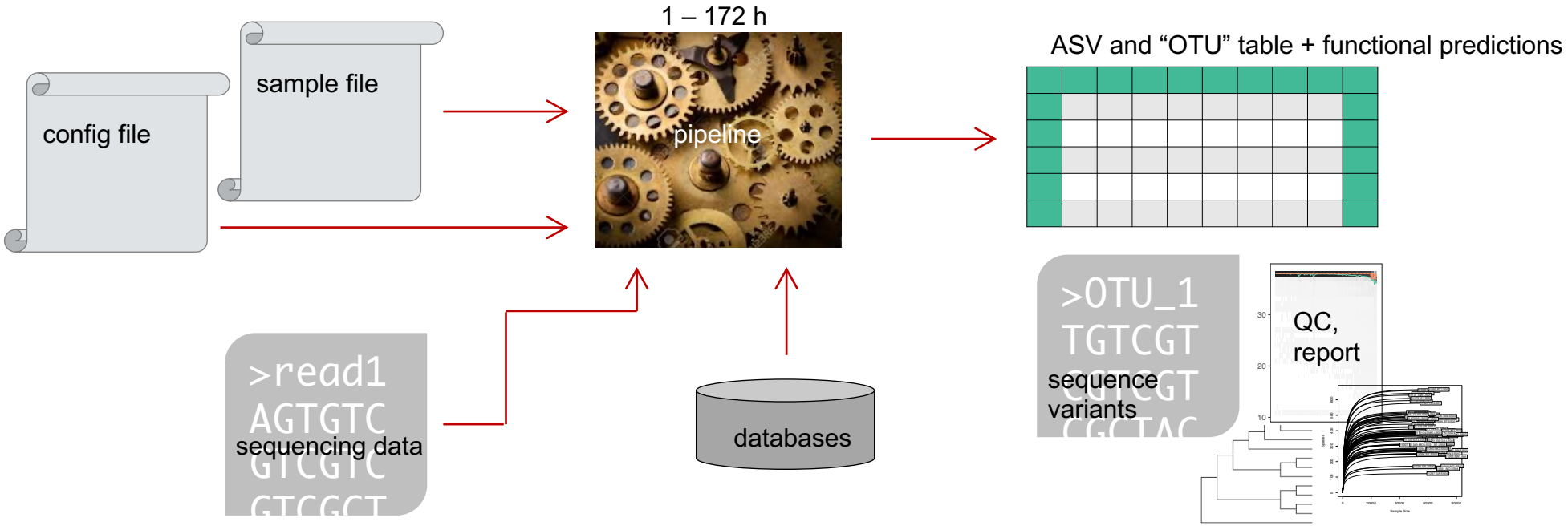
DFG Deutsche Forschungsgemeinschaft

iDiv
Christian Krause

iDiv

University of Amsterdam
Life Sciences

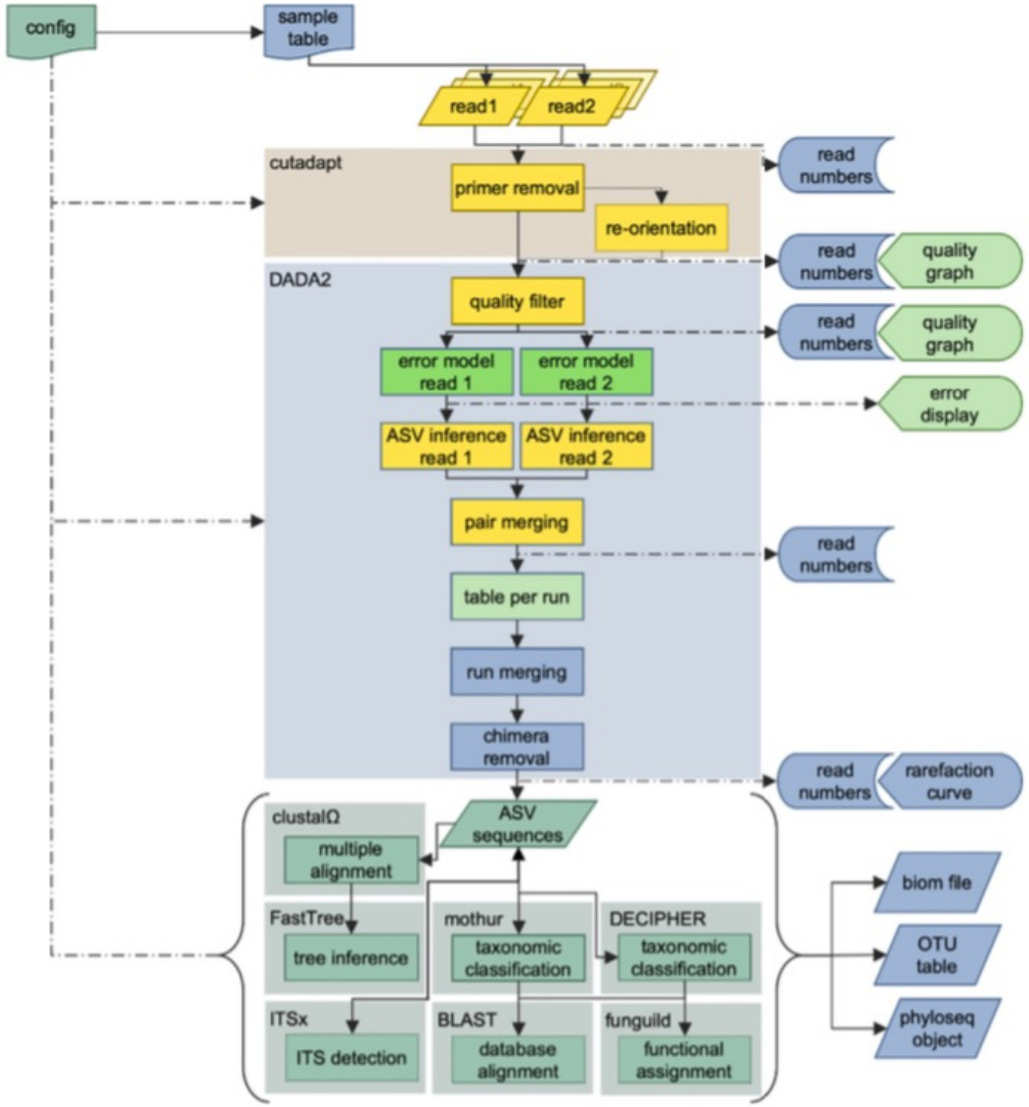
dadasnake pipeline



dadasnake

What does dadasnake do?

- optional primer removal
- quality filtering and trimming
- optional down-sampling
- error estimation & denoising
- optional paired-ends assembly
- ASV table generation
- optional chimera removal
- optional clustering of ASVs at user-defined similarity
- taxonomic classification (& ITS detection)
- optional length check, taxonomic filtering
- optional functional annotation/prediction, treeing…
- reporting of stats and quality measures



Overview of today

- Short intro
- Scope

- From sample to data - challenges for interpretation and analysis
  - what happens to the representation of microbial cells
  - biases
  - sources of error
  - detection limits

- Data processing and dadasnake
  - demo
  - details

- Discussion/Questions

How to run dadasnake? – demo

- download (and prepare) databases for your project

- set up your files:
  - reads
  - sample file
  - configuration file

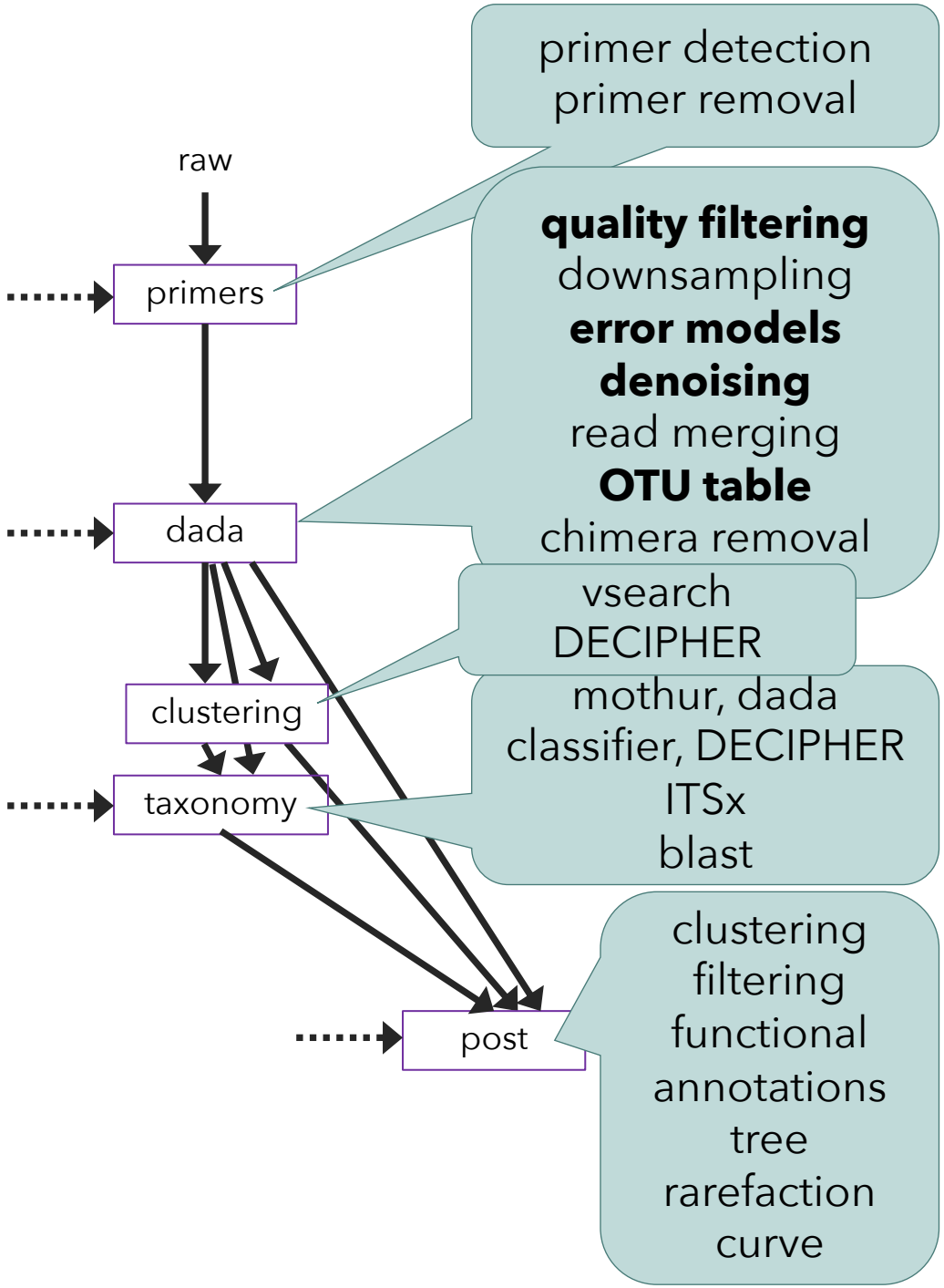
- run dadasnake:

```
./dadasnake -d /path/to/your/configuration/file
```

Steps

- by default, all steps are done
- but this can be configured:

```
do_primers: true
do_dada: true
do_taxonomy: true
do_postprocessing: true
```



Input

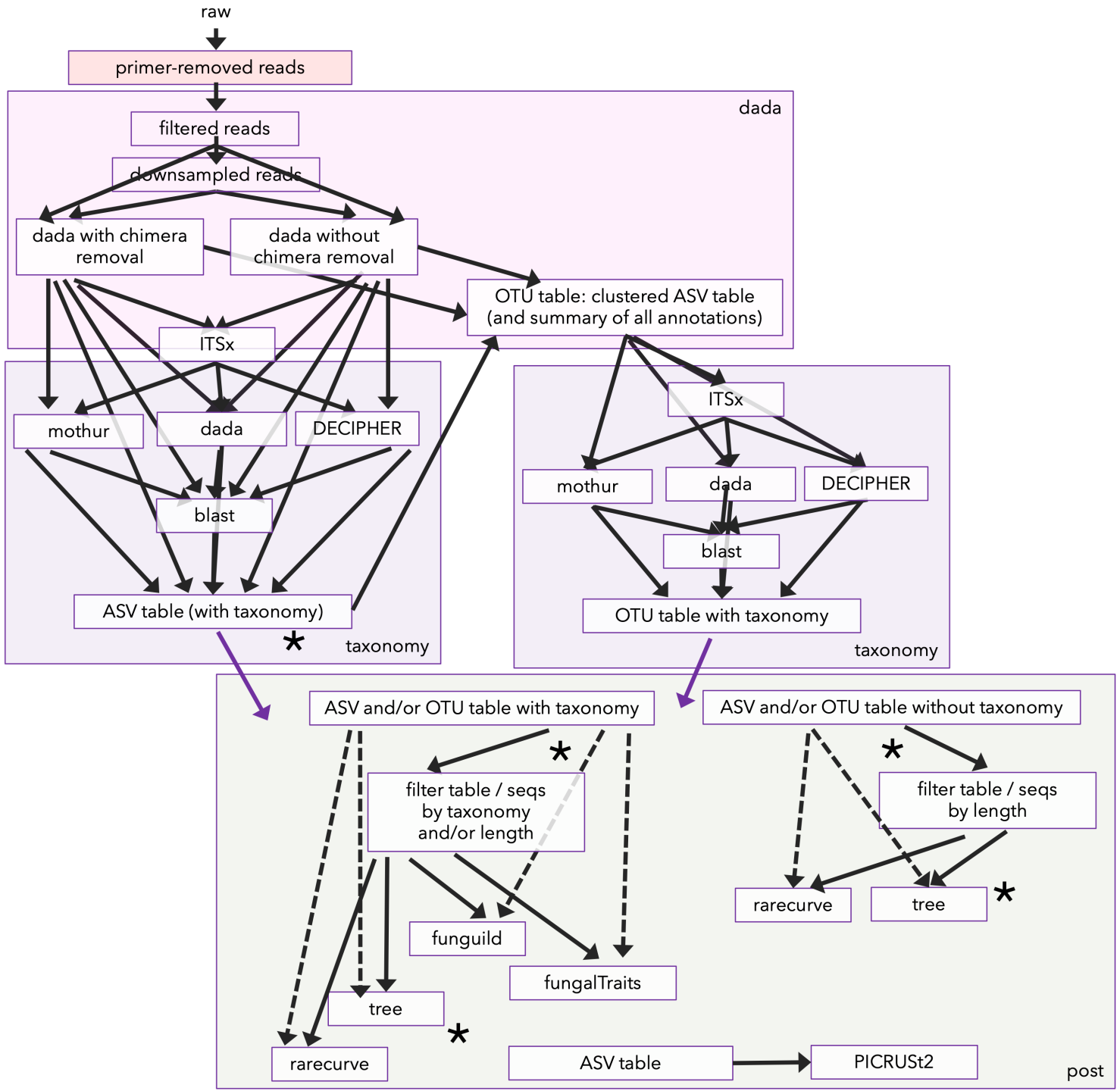
- raw (or pre-processed) reads (.fastq format – compressed or not)
- a sample table
- a configuration file (.yaml format)
- dadasnake has been tested on data sets of all sizes:
  - between 1 and 27,000 samples
  - between a few hundred and a few million reads per sample
  - between 10 and 800,000 ASVs

Output

- ASV (and OTU) table with taxonomy and comments
  - .tsv
  - .RDS (optional phyloseq object)
  - optional .biom
- ASV sequences
  - .fasta
- optional phylogenetic tree (.newick)
- optional functional annotation data
- stats (reads at every step, visualization: QC, errors, rarefaction curve)
- configuration, report

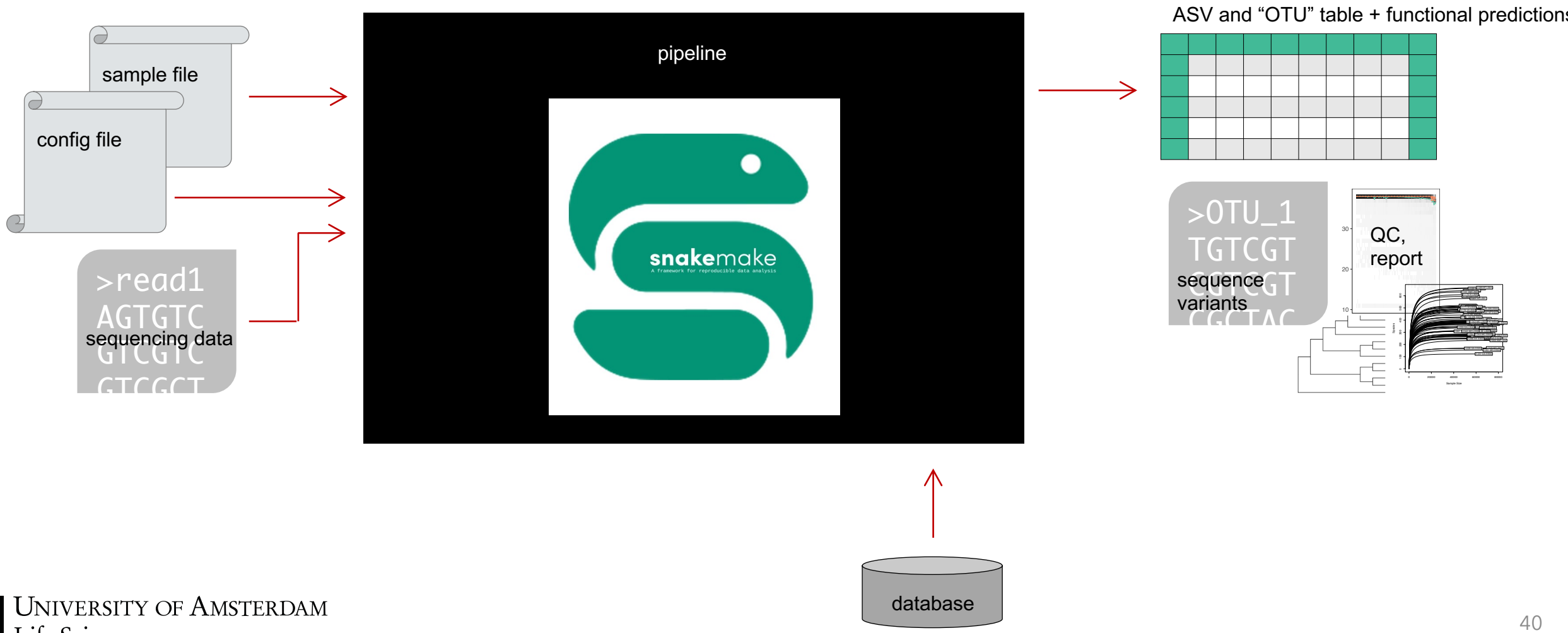
Workflows

dadasnake



dadasnake

How does dadasnake work?



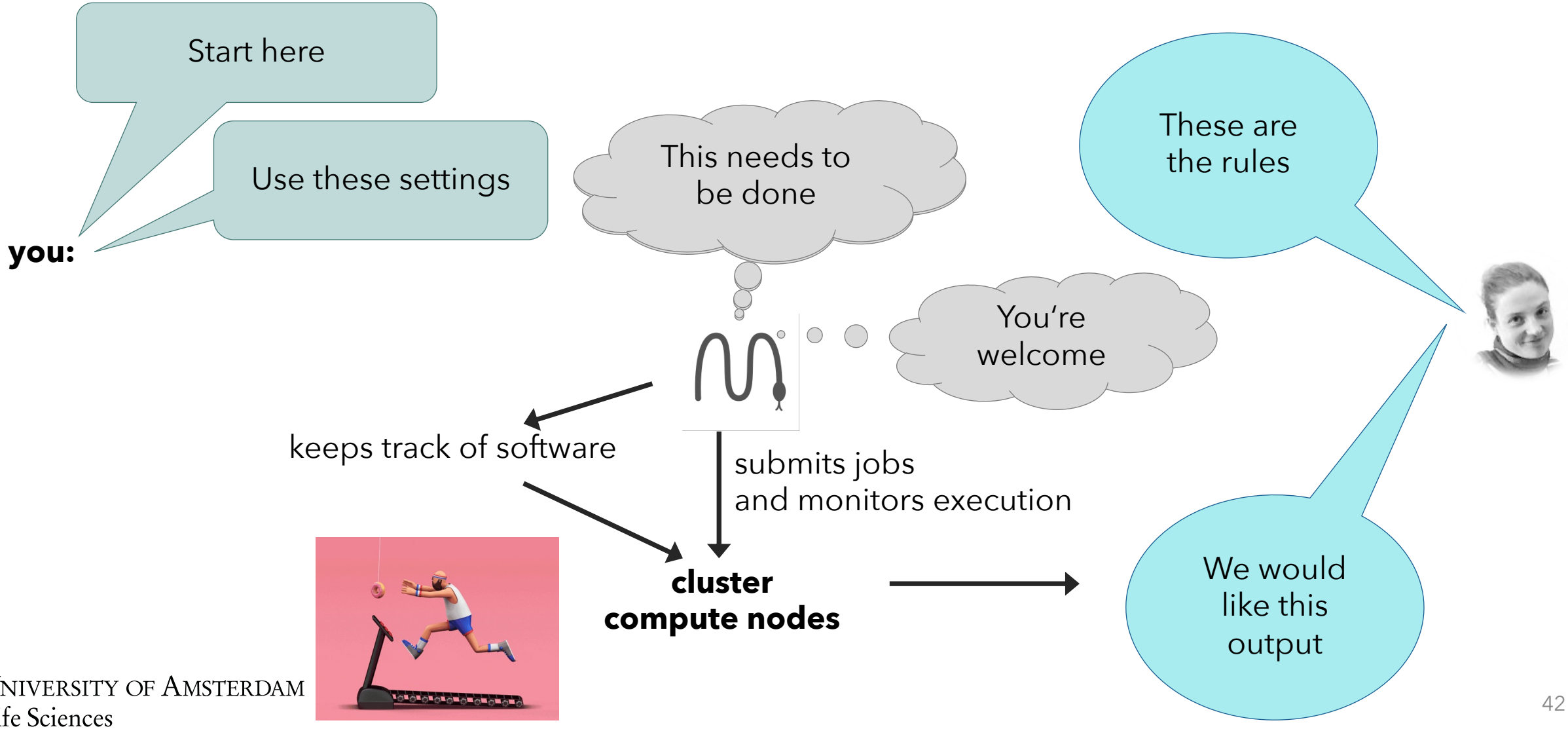
How does snakemake work?

Job execution

A job is executed if and only if

- output file is target and does not exist
- output file needed by another executed job and does not exist
- input file newer than output file
- input file will be updated by other job

How does snakemake make dadasnake work?



How can I re-start the pipeline?

Job execution

A job is executed if and only if

- output file is target and does not exist
- output file needed by another executed job and does not exist
- input file newer than output file
- input file will be updated by other job

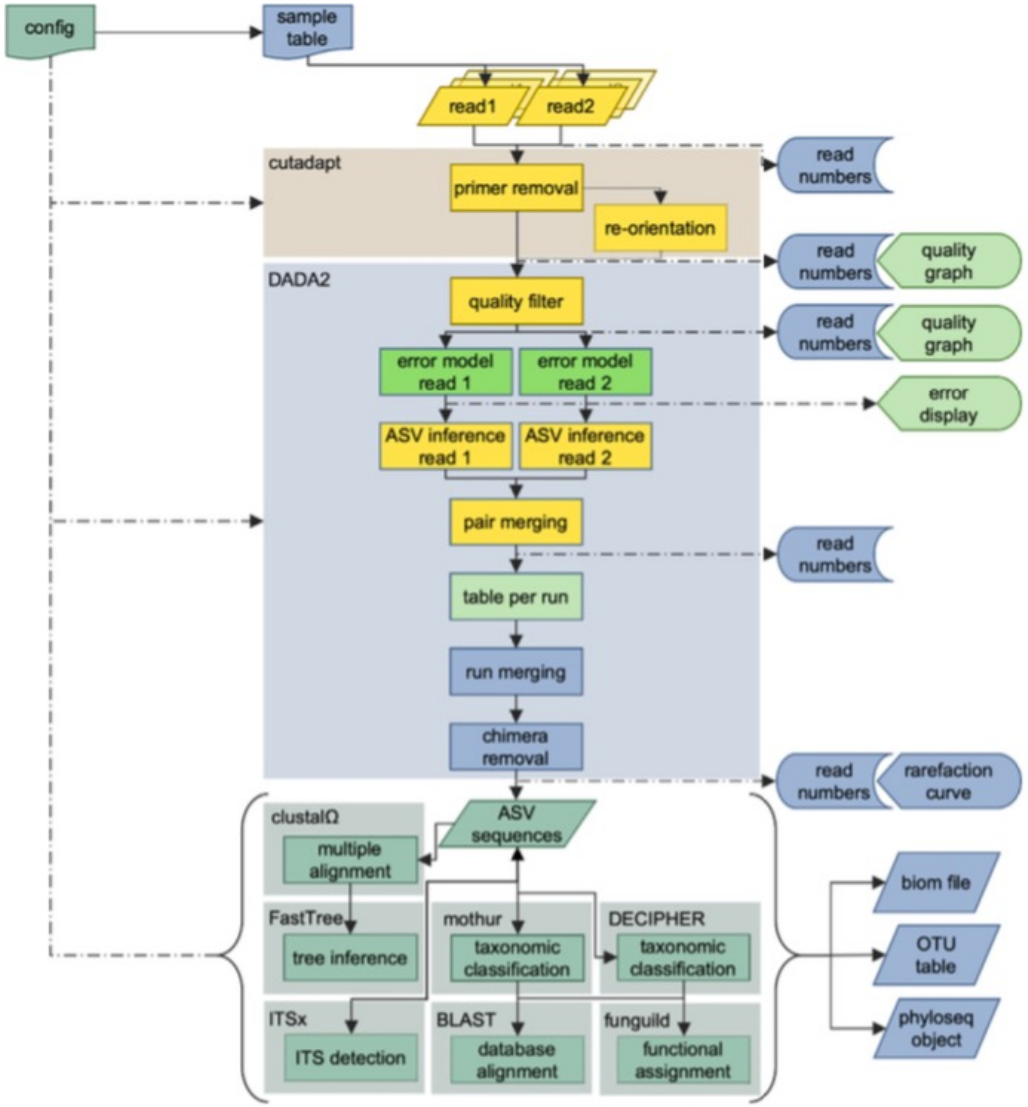
How can I re-start the pipeline?

- if the pipeline failed:
  - you can usually just repeat the start command, once the error is fixed

- if you want to re-do something: you have to delete all the file that you want to have redone. Then you can restart.

Options

- dadasnake defaults are for 16S rRNA V4 amplicons (515-806) - paired end
- it was also extensively benchmarked for fungal ITS2
- suggestions available:
  - for other targets: AMF, archaea, nematodes, trnL, several protist markers
  - for other techniques: single end, 454, pacbio CCS data settings



Where do I get more information?

- primer removal: cutadapt - http://gensoft.pasteur.fr/docs/cutadapt/1.18/guide.html
- DADA2 – steps: http://benjjneb.github.io/dada2/index.html
  - quality filtering and trimming, error estimation & denoising, paired-ends assembly, OTU table generation, chimera removal, taxonomic annotation
- taxonomic classification (& ITS detection):
  - DECIPHER: http://www2.decipher.codes/Bioinformatics.html
  - mothur classification: https://www.mothur.org/wiki/Classify.seqs
  - ITSx: https://microbiology.se/software/itsx/
  - BASTA: https://github.com/timkahlke/BASTA/wiki
- functional annotation, treeing…
  - funguild: https://github.com/UMNFuN/FUNGuild
  - fungalTraits: https://github.com/traitecoevo/fungaltraits
  - tax4fun2: https://github.com/bwemheu/Tax4Fun2
  - GTDB: https://gtdb.ecogenomic.org/
  - treeing: http://www.microbesonline.org/fasttree/ http://www.clustal.org/omega/

How to get help

- read the manual
  - https://github.com/a-h-b/dadasnake/

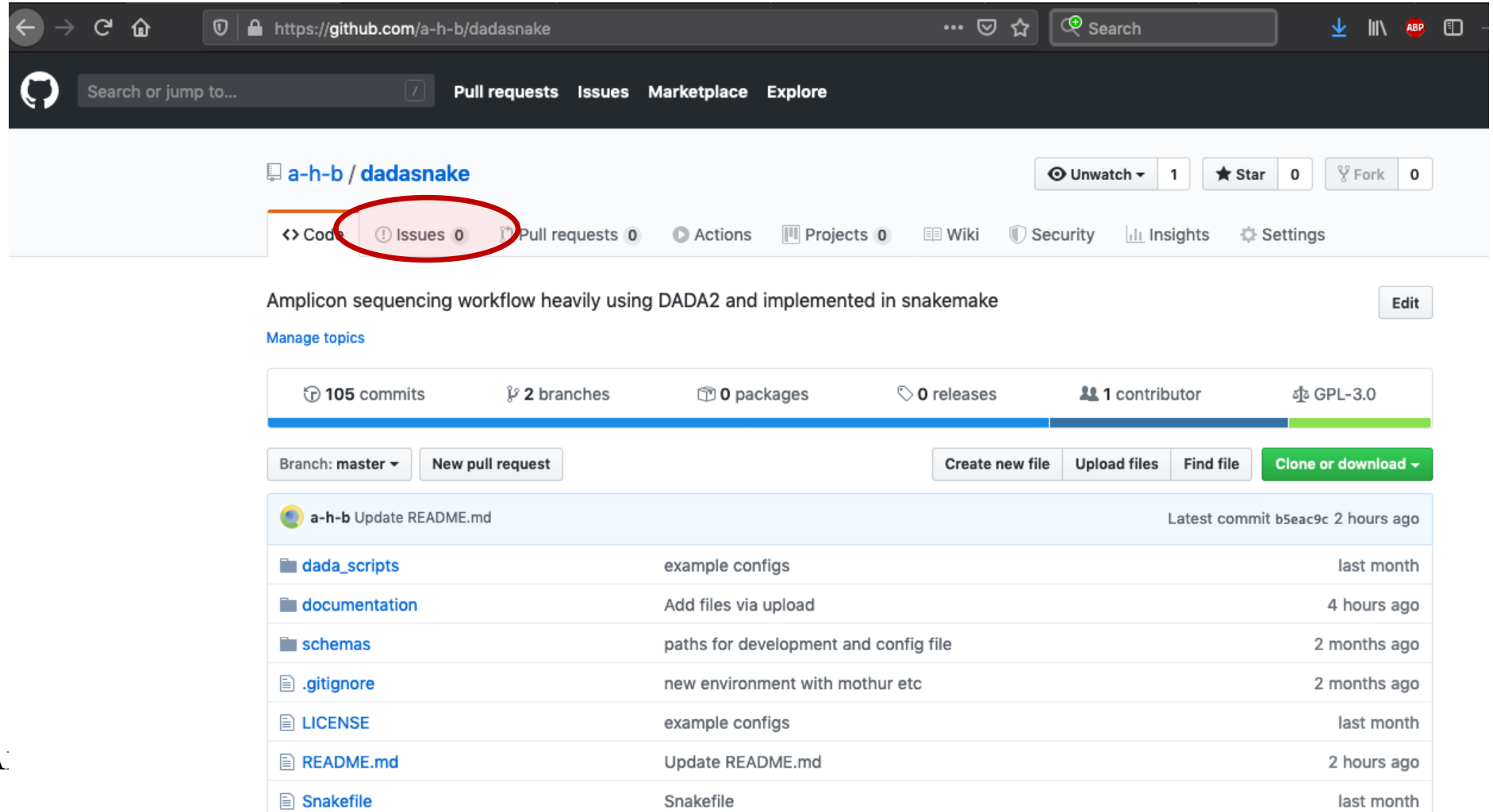
- use the github issue tracker:
  - public
  - permanent
  - searchable

  - you can attach files (logs, screenshots)
  - I can reply, you can reply

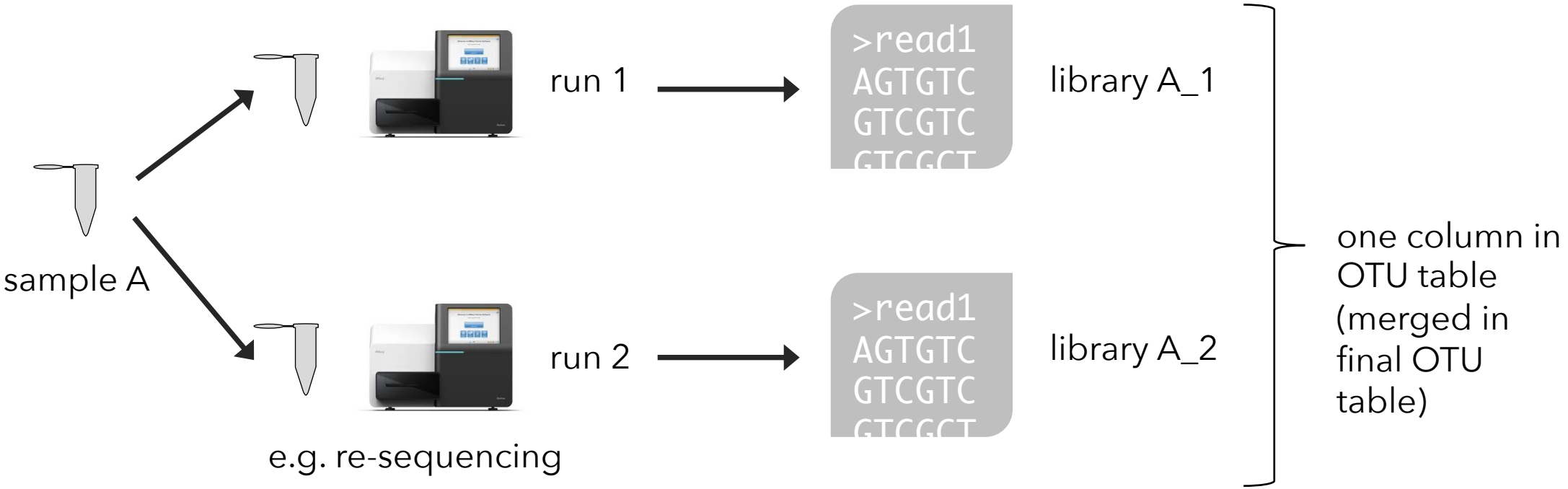
  - fixes can be linked directly to versioning

Github issue tracker

https://github.com/a-h-b/dadasnake/issues/new

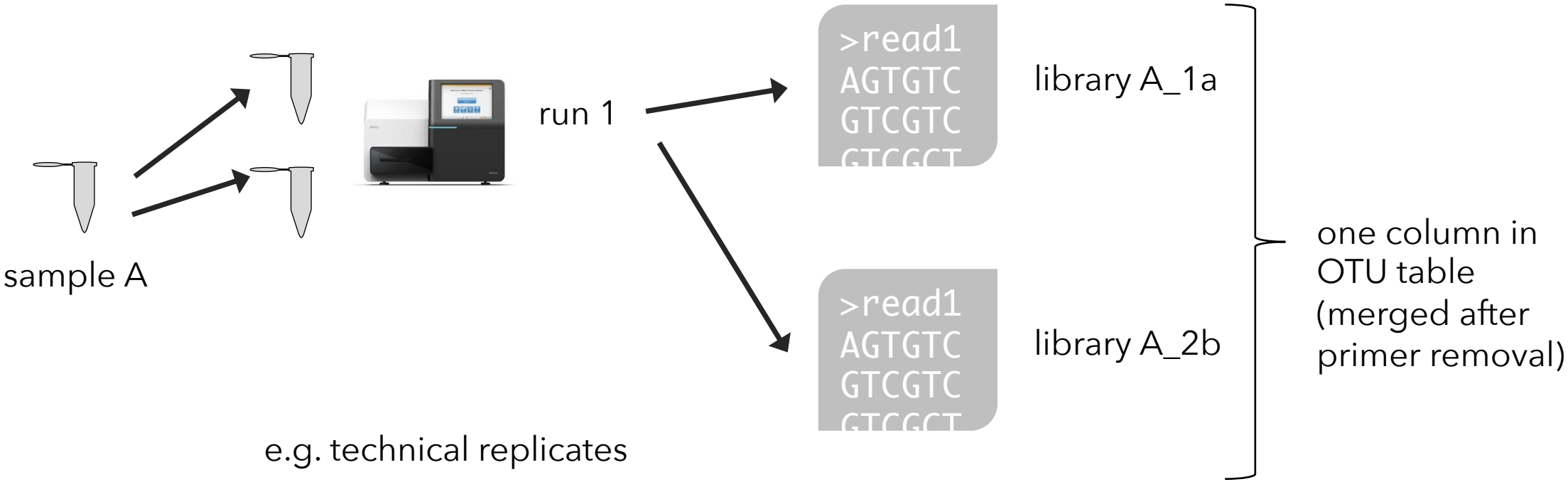


Raw data options



sample	library	run	r1_file	r2_file
A	A_1	1	myExp.A_R1.fastq.gz	myExp.A_R2.fastq.gz
A	A_2	2	myExp.A.reseq_R1.fastq.gz	myExp.A.reseq_R2.fastq.gz

Raw data options



sample	library	run	r1_file	r2_file
A	A_1a	1	myExp.A1_R1.fastq.gz	myExp.A1_R2.fastq.gz
A	A_2b	1	myExp.A2_R1.fastq.gz	myExp.A2_R2.fastq.gz

The samples file

- contains all the information on your samples
- must be tab-separated

- should not contain DOS-style end-of-line

- you can change the encoding by opening the config file using vi, then type

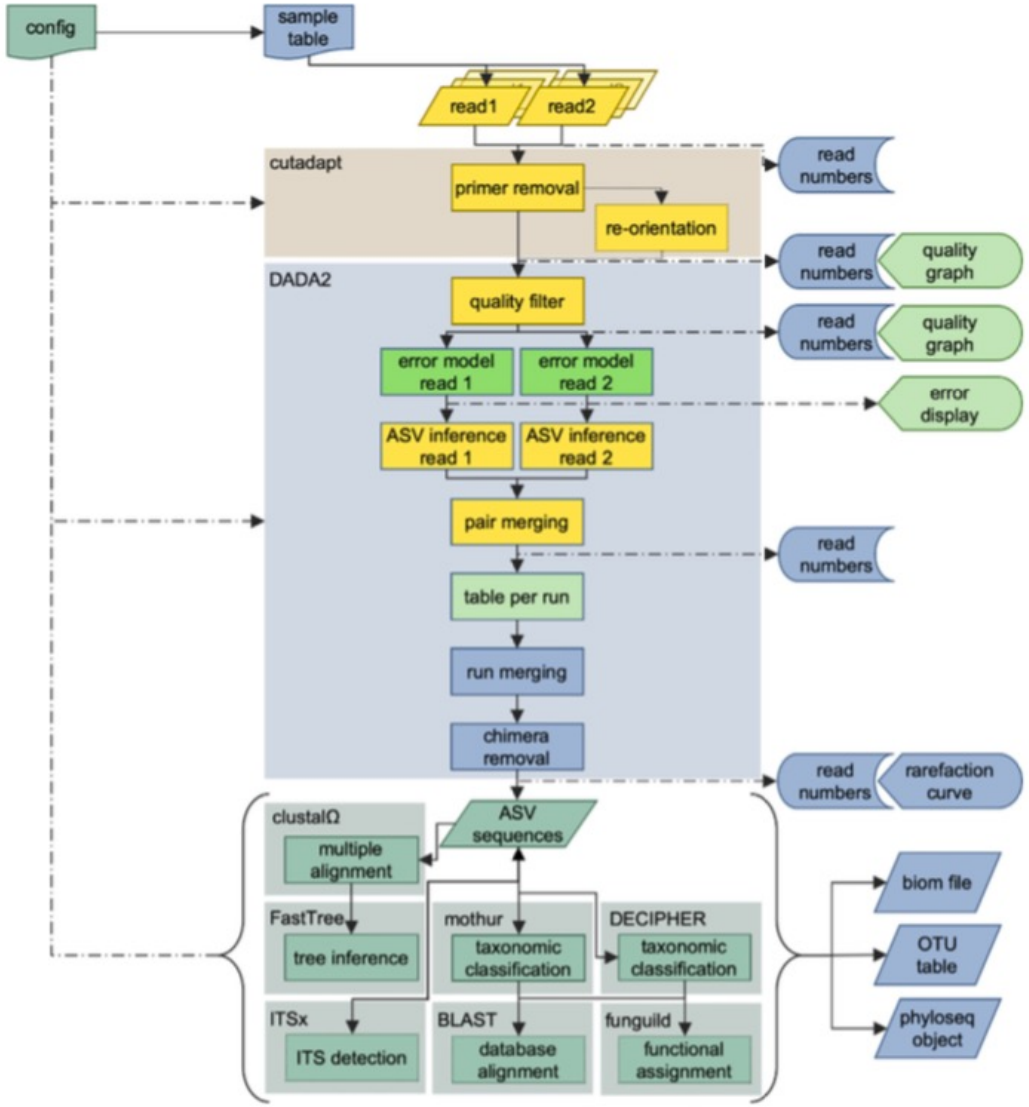
```
:set ff=unix
:wq
```

- must contain named columns: library and r1_file
- can contain named columns: r2_file, sample, run

- libraries and samples should not have the same name, if there are libraries that have different names

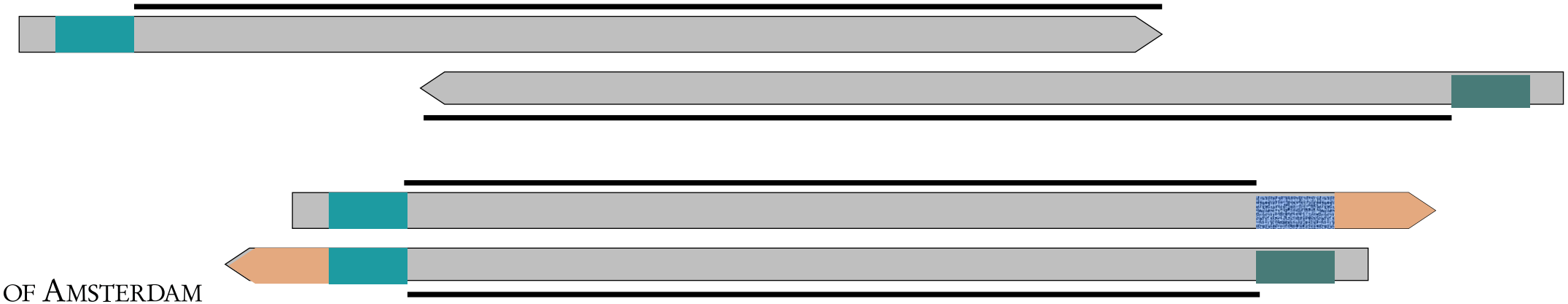
Steps in detail

- please interrupt me at any point to ask questions/comment on the steps and options



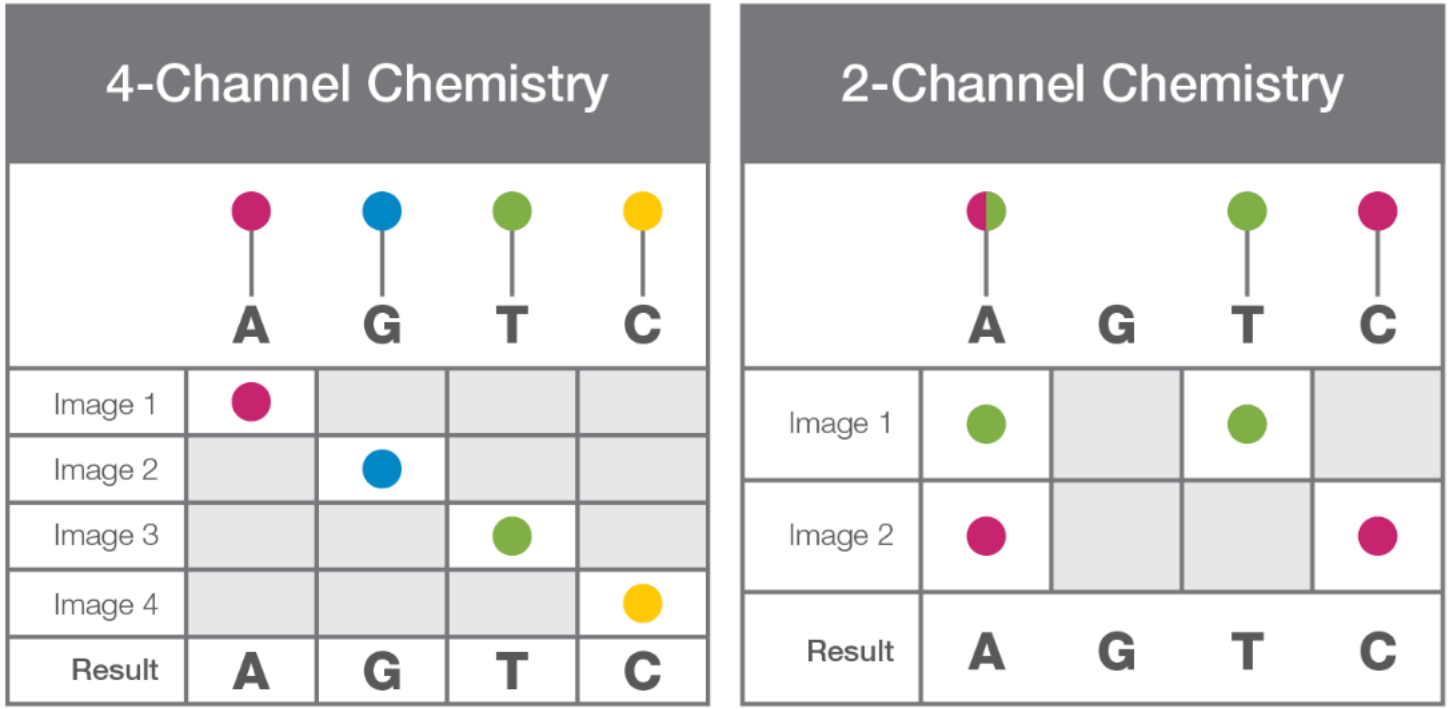
Primer removal

- using cutadapt
- flexible minimal overlap (default 10)
- flexible mismatches (default 20%)
- flexible AND/OR matching (default “any”, i.e. both reads need primers)
- flexible sequencing direction, or automatic detection
- removal of reverse-complement second primer



Quality filtering / trimming

- removal of trailing Gs (dark-cycle) for novaseq/nextseq



https://www.illumina.com/content/dam/illumina-marketing/documents/products/techspotlights/cmos-tech-note-770-2013-054.pdf

Quality filtering / trimming

- removal of trailing Gs (dark-cycle) for novaseq/nextseq
- rest is part of DADA2 pipeline:
- visualization of quality before and after – including fastQC/multiQC
- options:
  - minimum length
  - maximum length
  - truncation at specific length (too short kicked out)
  - truncation before first position with low quality (cut-off user-defined)
  - maximum overall error (based on quality)
  - trim positions from the left

Down-sampling

- quality-filtered/trimmed data can be down-sampled (rarefied) to a specified or minimum number of reads
- if reads of one sample are split into several libraries, the number of reads is adjusted to that

Error profile & denoising

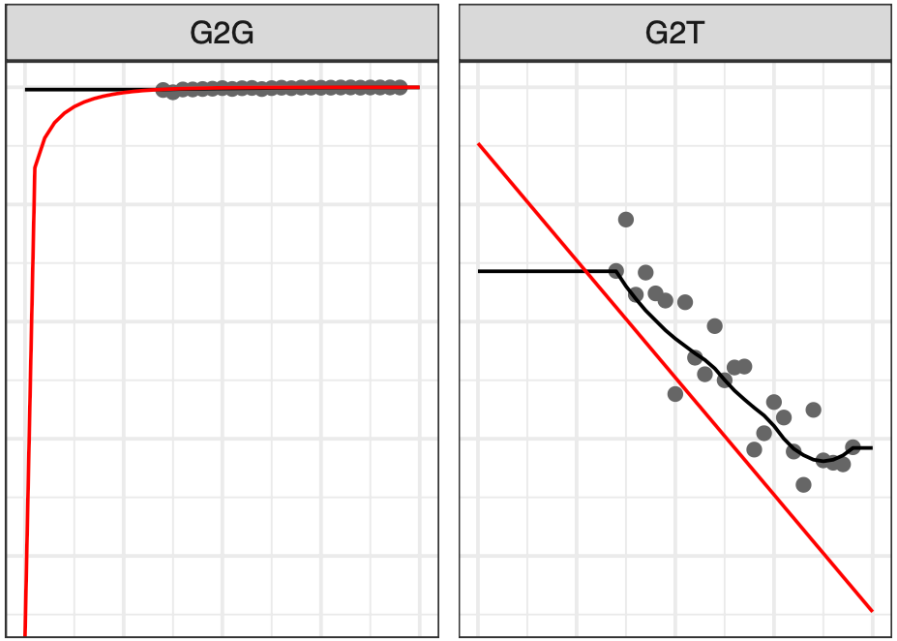
- part of DADA2 pipeline
- build ASVs per sample, per run, or for the whole study
- visualization
- experimental error-models for novaseq data
- settings can be adjusted for non-Illumina data

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Reminder: error models

- model substitutions for every run




```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                 |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Error rates depend on....

- Substitution (eg. A->C)
- Quality score (eg. Q=30)
- Batch effect (eg. run)

Paired-ends assembly

- part of DADA2 pipeline
- options:
  - minimum overlap (can be 0)
  - number of mismatches

- single-end data can also be used

Chimera removal

- part of DADA2 pipeline
- is done after the ASV table is made
- options:
  - consensus
  - pool

- chimera removal is optional

Clustering

- optional
- by VSEARCH or DECIPHER
- is done based on ASVs
- cut-off can be set by user

```
# SETTINGS FOR CLUSTERING ASV TABLE AT e.g. 97%
post_clustering:
  do: true
# do is only used if no taxonomy is done to trigger clustering
  cutoff: 0.97
# similarity cut-off
  method: vsearch
# method can be vsearch or decipher
  strand: plus
# strand only works for vsearch
```

Taxonomic annotation/classification

- choices:
  - DECIPHER algorithm
  - works better than DADA2-native algorithm
  - annotation to genus level
  - but doesn't scale (don't use for large datasets)
- and/or Bayesian classifier from mothur or from dada2 (slower than mothur)
- optional BLAST for unclassified sequences or all sequences, best hit and LCA can be added to ASV table, thanks to BASTA
- options:
  - databases
  - direction
  - before or after optional ITSx

Database choices

- dadasnake does not provide databases
- go get them from the people who make them

- dadasnake comes with a script to prune databases for the mothur classifier
  - select taxa (e.g. Fungi, Bacteria etc.)
  - select based on primer sequences
  - cut to region of interest

Functional annotation/prediction

- dadasnake does not provide databases
- go get them from the people who make them

- fungalTraits

- picrust2

Other functional information

- bacterial traits DB

- https://github.com/bacteria-archaea-traits/bacteria-archaea-traits
- https://www.nature.com/articles/s41597-020-0497-4

scientific **data**

Data Descriptor | Open Access | Published: 05 June 2020

A synthesis of bacterial and archaeal phenotypic trait data

Joshua S. Madin, Daniel A. Nielsen, [...]Mark Westoby

Scientific Data **7**, Article number: 170 (2020) | Cite this article

Abstract

A synthesis of phenotypic and quantitative genomic traits is provided for bacteria and archaea, in the form of a scripted, reproducible workflow that standardizes and merges 26 sources. The resulting unified dataset covers 14 phenotypic traits, 5 quantitative genomic traits, and 4 environmental characteristics for approximately 170,000 strain-level and 15,000 species-aggregated records. It spans all habitats including soils, marine and fresh waters and sediments, host-associated and thermal. Trait data can find use in clarifying major dimensions of ecological strategy variation across species. They can also be used in conjunction with species and abundance sampling to characterize trait mixtures in communities and responses of traits along environmental gradients.

Questions/comments on snakemake?

Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

https://github.com/a-h-b/dadasnake

How to run dadasnake? – installation of dependencies

- install/set up conda

- install mamba:

```
conda install -n base -c conda-forge mamba
```

- install snakemake:

```
mamba install -c conda-forge -c bioconda snakemake=6.9.1
mamba tabulate=0.8
```

thanks @ Nina

How to run dadasnake? – installation

- clone dadasnake and prepare run script

```
git clone https://github.com/a-h-b/dadasnake.git
cd dadasnake
cp auxiliary_files/dadasnake_tmux dadasnake
chmod 755 dadasnake
```

- adjust VARIABLE_CONFIG to your computer (if necessary)

How to run dadasnake? – initialization and testing

- initialize dadasnake

```
./dadasnake -i config/config.init.yaml
```

- test dadasnake

```
./dadasnake -l -n "TESTRUN" -r config/config.test.yaml
```

How to run dadasnake? – set up your files

- your reads:
  - all of your reads need to be in the same directory. Alternatively, you can set links to all of your reads into one directory. Reads can be gzipped or not (fastq.gz or fastq)
- config file:
  - you can copy one of the files in dadasnake/config and adjust the settings
- sample file*:
  - you can quickly generate a sample table like this:

```
paste <(ls *_R1_*fastq.gz | sed "s#_R.*##g") <(ls *_R1_*fastq.gz) \
 <(ls *_R1_*fastq.gz | sed "s#_R1_#_R2_#g")  >> samples.new.tsv
```

  - then, open in vi and introduce a header, containing:

library, r1_file, r2_file, (run) - separated by tabs

  - fix sample names, if you wish

*for multiple runs in the sample file, you can do this for the first run from the first run's directory:

```
paste <(ls *_R1_*fastq.gz | \
sed "s#_R.*##g") <(ls
*_R1_*fastq.gz) \
<(ls *_R1_*fastq.gz | \
sed "s#_R1_#_R2_#g") | \
sed 's#$#\trun1#' >>
../samples.2run.tsv
```

and then from the second run's directory:

```
paste <(ls *_R1_*fastq.gz | \
sed "s#_R.*##g") <(ls
*_R1_*fastq.gz) \
<(ls *_R1_*fastq.gz | \
sed "s#_R1_#_R2_#g") | \
sed 's#$#\trun2#' >>
../samples.2run.tsv
```

then fix header in vi

How to run dadasnake? – run

- connect to your server, navigate to your config file

```
/path/2/dadasnake/dadasnake -d /path/to/your/configuration/file
```

- check output
- then start dadasnake, e.g.:

```
/path/2/dadasnake/dadasnake -c -r \
-n ANYNAME /path/to/configuration/file
```

- wait, check status in output folder
- download results

How to find the error?

- start from outside to inside, from back to beginning

How to say that something went wrong

- meaningful summary
  - ☹️ "it doesn't work"
  - ✔ "error when output of step X is empty"
- what did you do?
  - ☹️ "I ran the pipeline"
  - ✔ add your config file
- what did you expect to happen?
  - ☹️ "there's nothing there"
  - ✔ I am looking for the output of step Y
- what happened?
  - ☹️ "I don't know what happened"
  - ✔ add the error messages and logs