H3ABioNet

Pan African Bioinformatics Network for H3Africa

16S rRNA Intermediate Bioinformatics Online Course

Building portable, user-friendly pipelines using nextflow

Module website

About nextflow

Scalable and reproducible scientific workflows using software containers

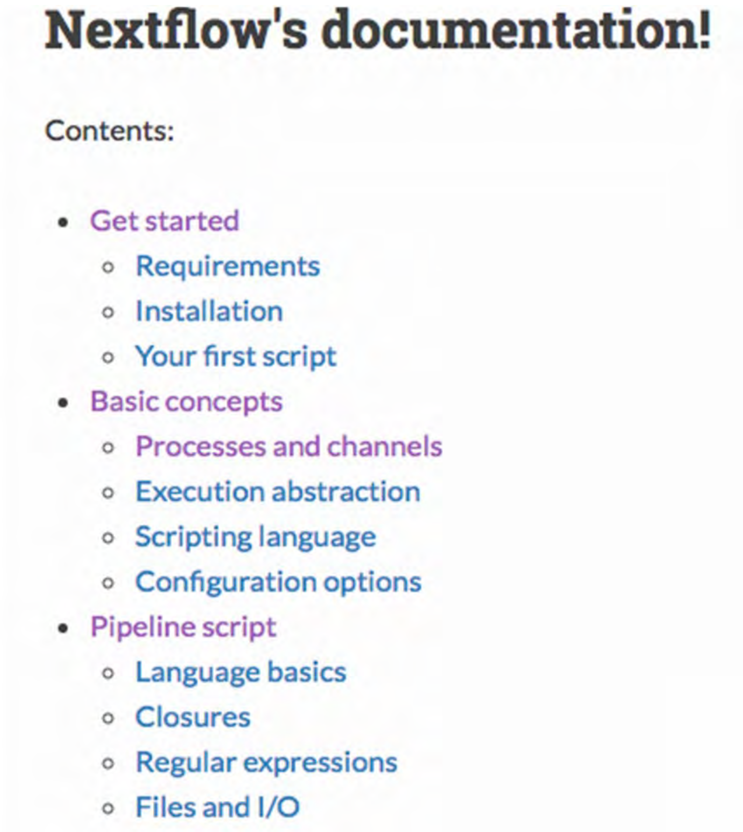
- Built-in GitHub support
- Compatibility with virtually all computational infrastructures, including all major cluster job schedulers
- Integrated software dependency management (Docker, Singularity, Conda)
- Portability so you can run your pipeline anywhere: laptop, cluster or cloud
- Reproducibility of analyses independent of time and computing platform

Why nextflow will save you time

- Reuse your existing scripts and tools (and you don't need to learn a new language or API to start using it)
  - Workflow 'processes' can be written in common scripting languages (R, python, bash, etc.)
- Resume pipeline execution from the last successfully executed step
  - All the intermediate results produced during the pipeline execution are automatically tracked
- Super easy setup
  - Check prerequisites (`java –version` ≥ Java 8)
  - Download Nextflow (curl -s https://get.nextflow.io | bash)
  - *Hello world!* (./nextflow run hello)

Why nextflow will save you time

- Good documentation

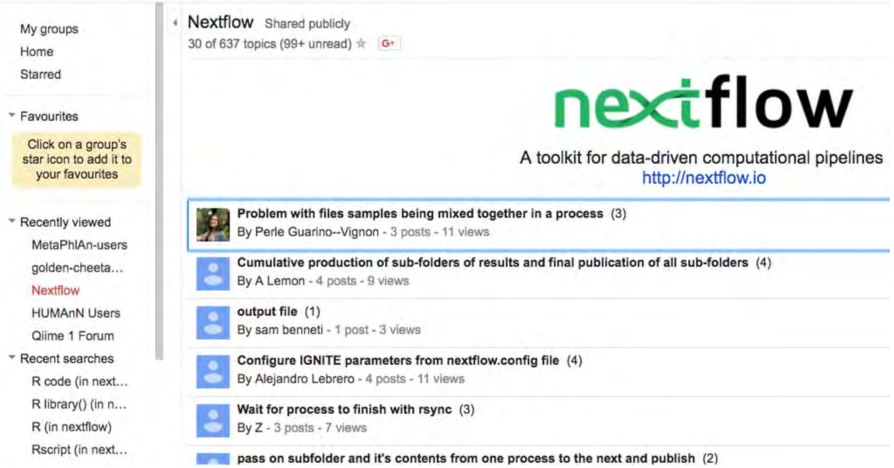


Nextflow's documentation!

Contents:

- Get started
  - Requirements
  - Installation
  - Your first script
- Basic concepts
  - Processes and channels
  - Execution abstraction
  - Scripting language
  - Configuration options
- Pipeline script
  - Language basics
  - Closures
  - Regular expressions
  - Files and I/O

- Google group support



Nextflow Shared publicly
30 of 637 topics (99+ unread)

nextflow
A toolkit for data-driven computational pipelines
http://nextflow.io

Problem with files samples being mixed together in a process (3)
By Perle Guarino--Vignon - 3 posts - 11 views

Cumulative production of sub-folders of results and final publication of all sub-folders (4)
By A Lemon - 4 posts - 9 views

output file (1)
By sam benneti - 1 post - 3 views

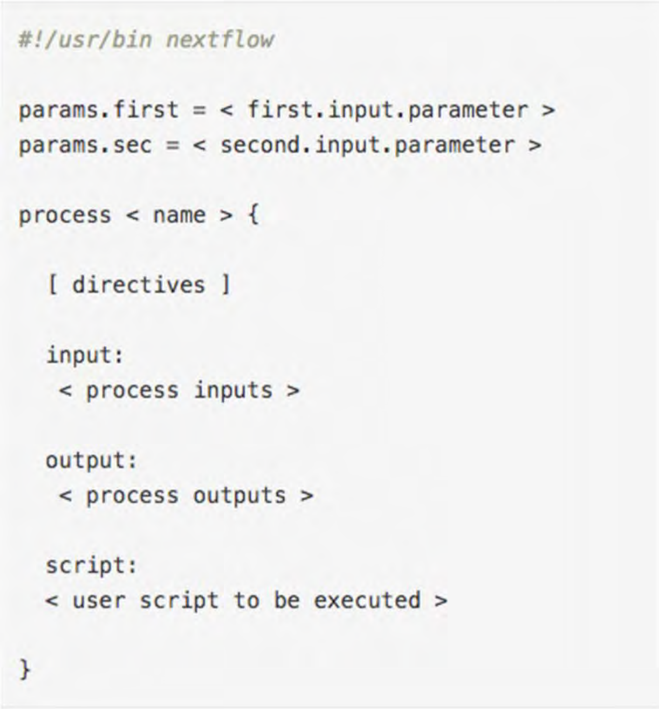
Configure IGNITE parameters from nextflow.config file (4)
By Alejandro Lebrero - 4 posts - 11 views

Wait for process to finish with rsync (3)
By Z - 3 posts - 7 views

pass on subfolder and it's contents from one process to the next and publish (2)

- Existing pipelines & templates

nextflow: the basics

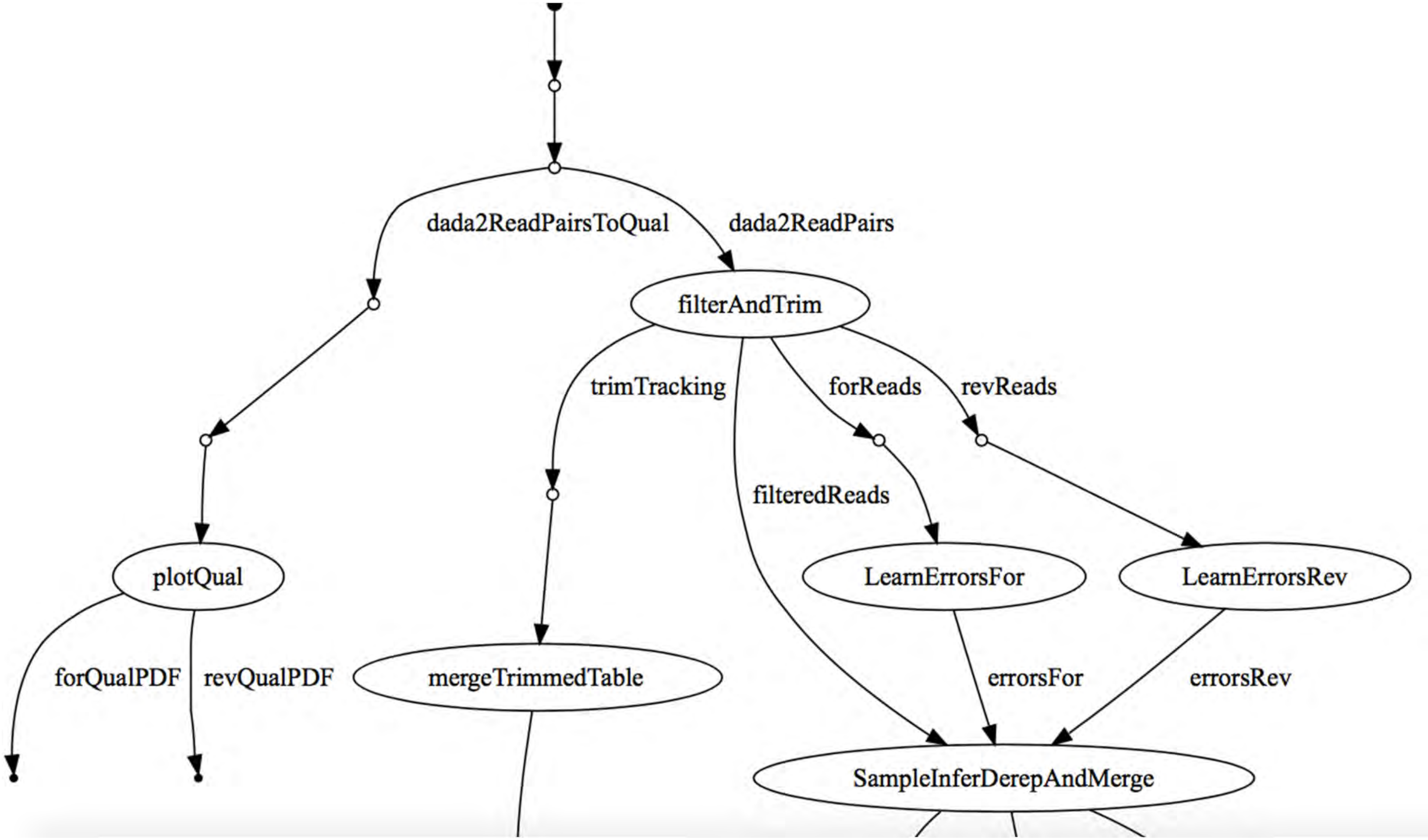


```
#!/usr/bin nextflow

params.first = < first.input.parameter >
params.sec = < second.input.parameter >

process < name > {

  [ directives ]

  input:
   < process inputs >

  output:
   < process outputs >

  script:
  < user script to be executed >

}
```

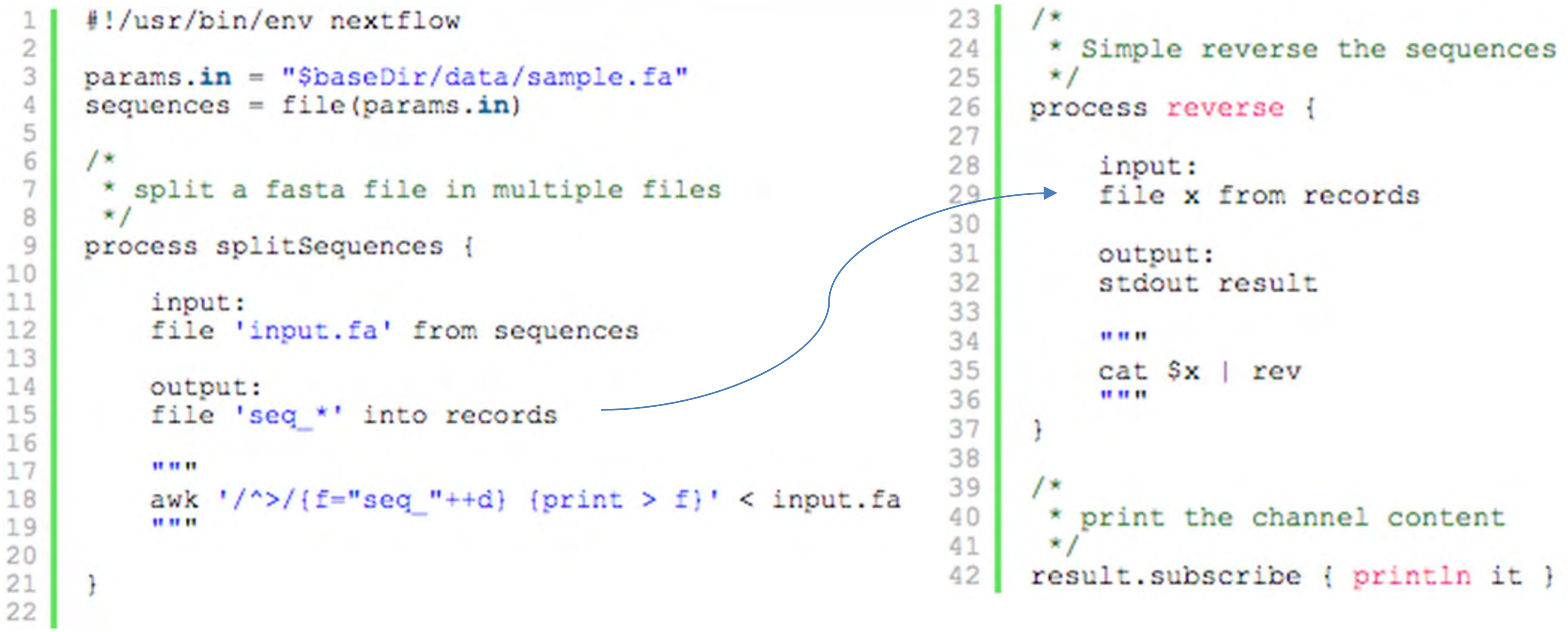
Nextflow terminology:

- ‘process’: one (independent) step in the pipeline
- ‘channel’: information flows from one process to another via ‘channels’ as defined in the input and output sections of each process
- ‘script’: each process contains a ‘script block’. This is where the executable coding happens
- ‘executor’: the component that determines the system where a pipeline process is run and supervises its execution
  - easy to change via config files

nextflow automatically creates a DAG of your pipeline



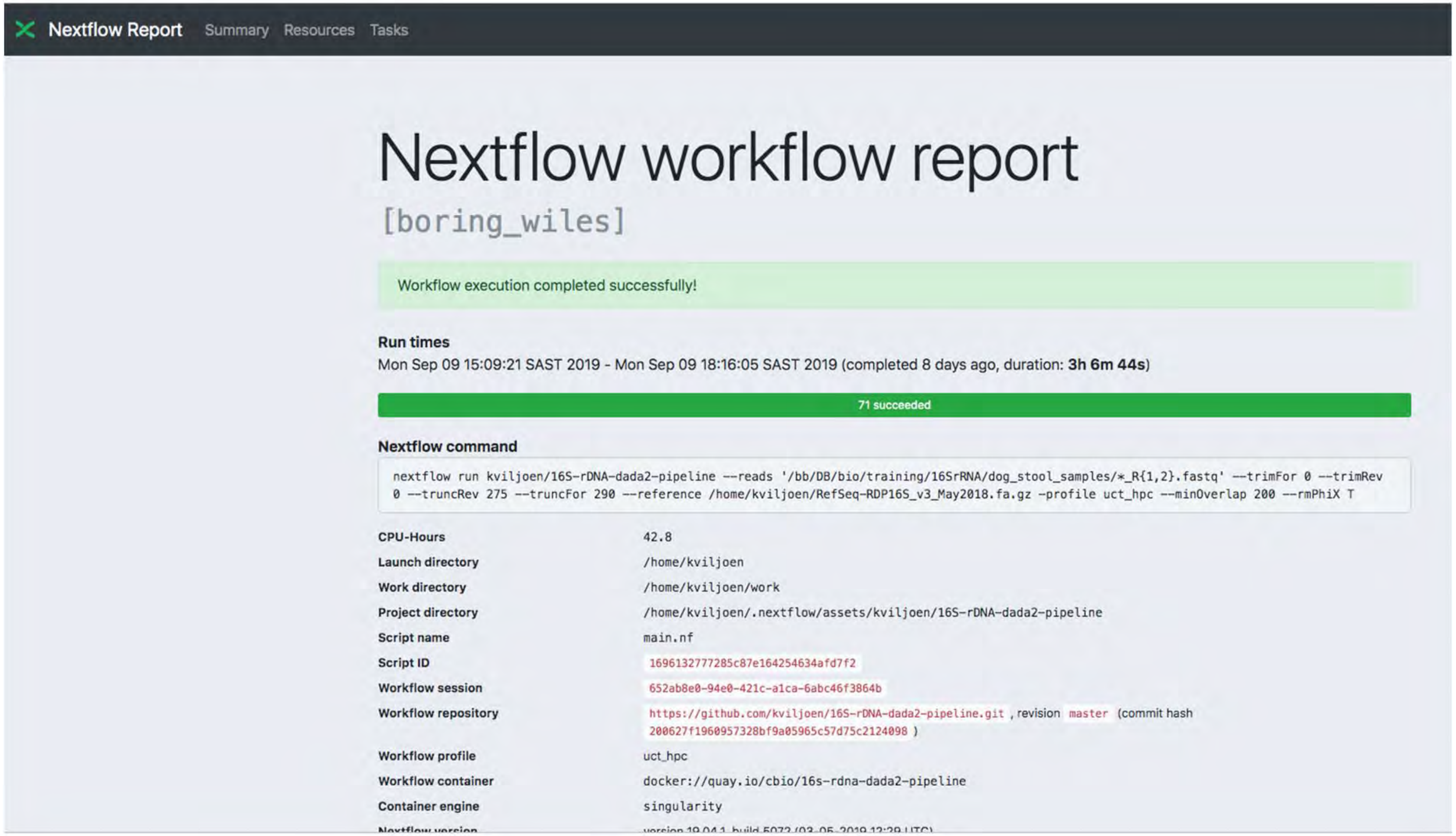
A nextflow script example



```
#!/usr/bin/env nextflow

params.in = "$baseDir/data/sample.fa"
sequences = file(params.in)

/*
 * split a fasta file in multiple files
 */
process splitSequences {

    input:
    file 'input.fa' from sequences

    output:
    file 'seq_*' into records

    """
    awk '/^>/{f="seq_"++d} {print > f}' < input.fa
    """

}

```

```
/*
 * Simple reverse the sequences
 */
process reverse {

    input:
    file x from records

    output:
    stdout result

    """
    cat $x | rev
    """
}

/*
 * print the channel content
 */
result.subscribe { println it }
```

nextflow workflow reports



Nextflow workflow report

[boring_wiles]

Workflow execution completed successfully!

Run times

Mon Sep 09 15:09:21 SAST 2019 - Mon Sep 09 18:16:05 SAST 2019 (completed 8 days ago, duration: **3h 6m 44s**)

71 succeeded

Nextflow command

```
nextflow run kviljoen/16S-rDNA-dada2-pipeline --reads '/bb/DB/bio/training/16SrRNA/dog_stool_samples/*_R{1,2}.fastq' --trimFor 0 --trimRev
0 --truncRev 275 --truncFor 290 --reference /home/kviljoen/RefSeq-RDP16S_v3_May2018.fa.gz -profile uct_hpc --minOverlap 200 --rmPhiX T
```

CPU-Hours	42.8
Launch directory	/home/kviljoen
Work directory	/home/kviljoen/work
Project directory	/home/kviljoen/.nextflow/assets/kviljoen/16S-rDNA-dada2-pipeline
Script name	main.nf
Script ID	1696132777285c87e164254634afd7f2
Workflow session	652ab8e0-94e0-421c-a1ca-6abc46f3864b
Workflow repository	https://github.com/kviljoen/16S-rDNA-dada2-pipeline.git , revision master (commit hash 200627f1960957328bf9a05965c57d75c2124098)
Workflow profile	uct_hpc
Workflow container	docker://quay.io/cbio/16s-rdna-dada2-pipeline
Container engine	singularity
Nextflow version	version 19.04.1 build 5072 (03-05-2019 12:29 UTC)

H3ABioNet

Pan African Bioinformatics Network for H3Africa

16S rRNA Intermediate Bioinformatics Online Course

nextflow : a hands on demonstration

Let’s process reads with dada2, as a Nextflow pipeline

- Data: Illumina paired reads .fastq files (Dog microbiome)
- Pipeline: https://github.com/grbot/16S-rDNA-dada2-pipeline

Running the DADA2 Nextflow pipeline on test data

- Log onto the cluster with your username e.g.

```
ssh gerrit@154.114.37.238
```

- Start an interactive job from a **worker node**
  - **NB:** Do **not** launch Nextflow from the head node (high java memory requirements)
  - **Instead start an Interactive job on a worker node:**

```
srun --nodes=1 --ntasks 1 --mem=8g --pty bash
```

- Pull the nextflow pipeline from Github

```
cd $HOME
git clone https://github.com/grbot/16S-rDNA-dada2-pipeline
cd $HOME/16S-rDNA-dada2-pipeline
```

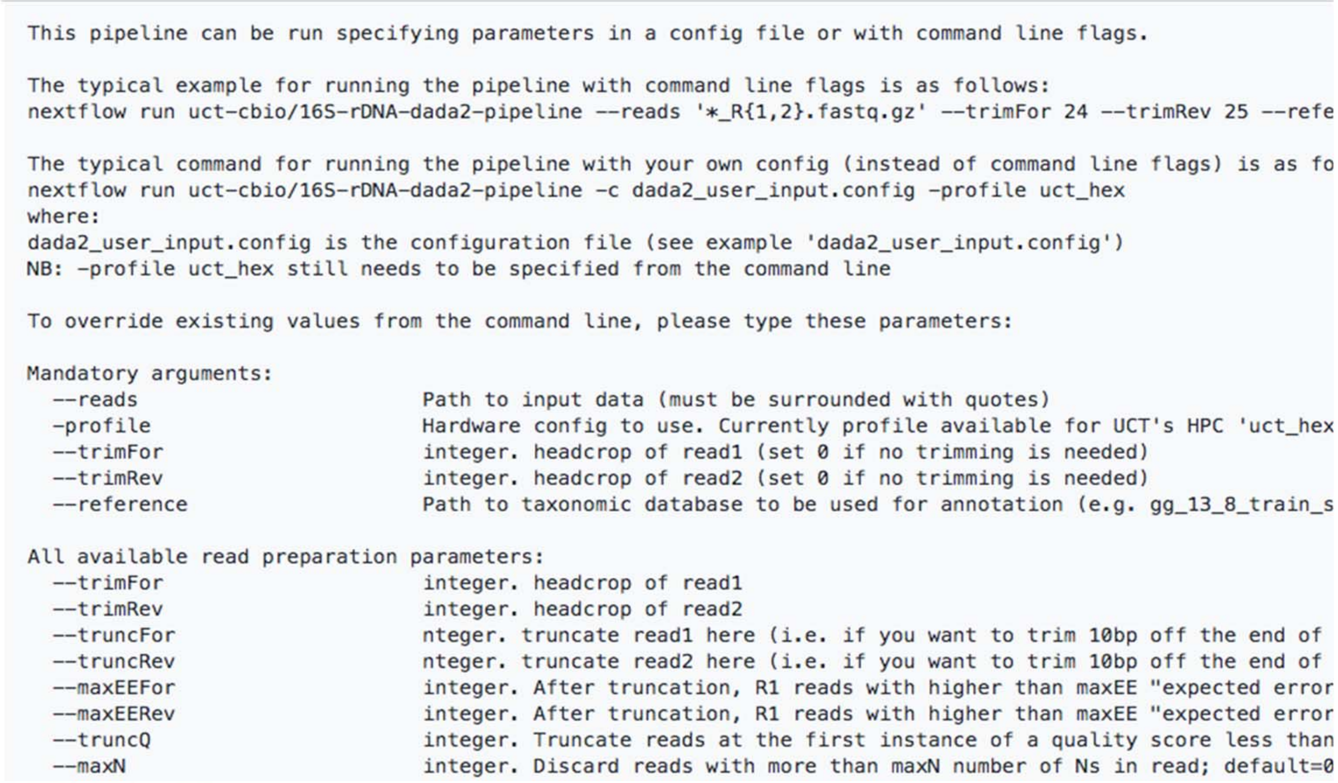
Running the DADA2 Nextflow pipeline on test data

- Now launch the pipeline from your interactive session as follows

```
nextflow run main.nf -profile training --reads="/cbio/data/test-data/*_R{1,2}.fastq.gz" --trimFor 24 --trimRev 25 --reference="/cbio/data/ref-data/silva_nr_v132_train_set.fa.gz" --species="/cbio/data/ref-data/silva_species_assignment_v132.fa.gz" --outdir="$HOME/out"
```

Nextflow pipeline parameter specification

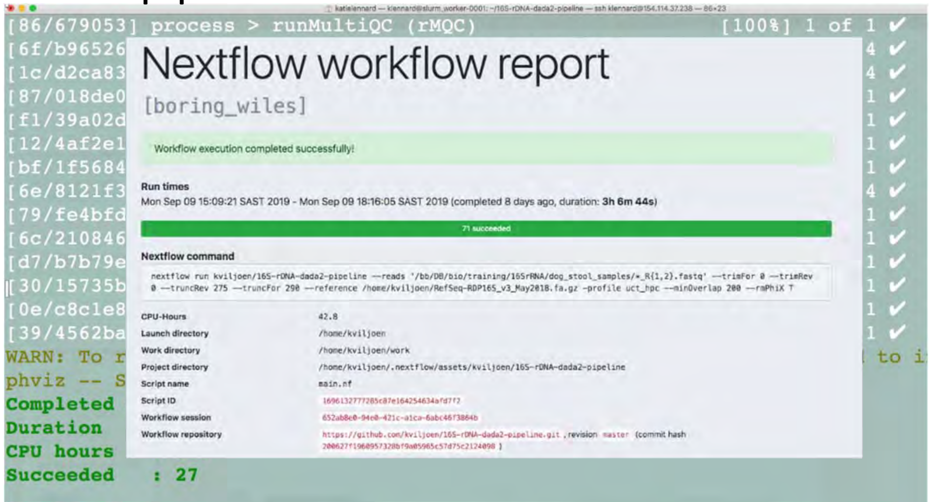
- Input parameters can be specified as
  - Command line flags, OR
  - In a user-defined config file



```
This pipeline can be run specifying parameters in a config file or with command line flags.

The typical example for running the pipeline with command line flags is as follows:
nextflow run uct-cbio/16S-rDNA-dada2-pipeline --reads '*_R{1,2}.fastq.gz' --trimFor 24 --trimRev 25 --refe

The typical command for running the pipeline with your own config (instead of command line flags) is as fo
nextflow run uct-cbio/16S-rDNA-dada2-pipeline -c dada2_user_input.config -profile uct_hex
where:
dada2_user_input.config is the configuration file (see example 'dada2_user_input.config')
NB: -profile uct_hex still needs to be specified from the command line

To override existing values from the command line, please type these parameters:

Mandatory arguments:
  --reads                      Path to input data (must be surrounded with quotes)
  -profile                     Hardware config to use. Currently profile available for UCT's HPC 'uct_hex
  --trimFor                    integer. headcrop of read1 (set 0 if no trimming is needed)
  --trimRev                    integer. headcrop of read2 (set 0 if no trimming is needed)
  --reference                  Path to taxonomic database to be used for annotation (e.g. gg_13_8_train_s

All available read preparation parameters:
  --trimFor                    integer. headcrop of read1
  --trimRev                    integer. headcrop of read2
  --truncFor                   nteger. truncate read1 here (i.e. if you want to trim 10bp off the end of
  --truncRev                   nteger. truncate read2 here (i.e. if you want to trim 10bp off the end of
  --maxEEFor                   integer. After truncation, R1 reads with higher than maxEE "expected error
  --maxEERev                   integer. After truncation, R1 reads with higher than maxEE "expected error
  --truncQ                     integer. Truncate reads at the first instance of a quality score less than
  --maxN                       integer. Discard reads with more than maxN number of Ns in read; default=0
```

DADA2 Nextflow pipeline output interpretation

Successful pipeline execution



```
katielennard — klennard@slurm_worker-0001: ~/16S-rDNA-dada2-pipeline — ssh klennard@154.114.37.238 — 86×23
[86/679053] process > runMultiQC (rMQC)                    [100%] 1 of 1 ✔
[6f/b96526                                                              4 ✔
[1c/d2ca83                                                              4 ✔
[87/018de0                                                              1 ✔
[f1/39a02d                                                              1 ✔
[12/4af2e1                                                              1 ✔
[bf/1f5684                                                              1 ✔
[6e/8121f3                                                              4 ✔
[79/fe4bfd                                                              1 ✔
[6c/210846                                                              1 ✔
[d7/b7b79e                                                              1 ✔
[30/15735b                                                              1 ✔
[0e/c8c1e8                                                              1 ✔
[39/4562ba                                                              1 ✔
WARN: To r                                                           to i
phviz -- S
Completed
Duration
CPU hours
Succeeded     : 27
```

Nextflow workflow report

[boring_wiles]

Workflow execution completed successfully!

Run times

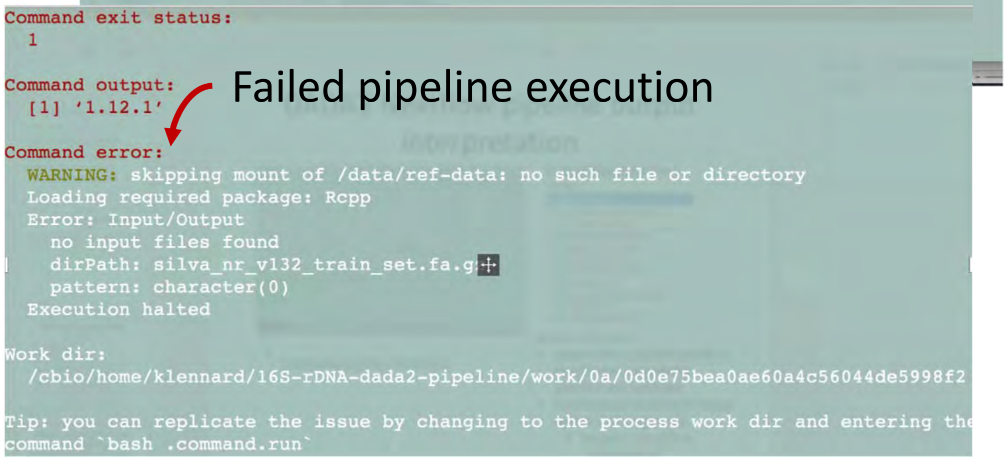
Mon Sep 09 15:09:21 SAST 2019 - Mon Sep 09 18:16:05 SAST 2019 (completed 8 days ago, duration: **3h 6m 44s**)

71 succeeded

Nextflow command

```
nextflow run kviljoen/16S-rDNA-dada2-pipeline --reads '/bb/DB/bio/training/16SrRNA/dog_stool_samples/*_R{1,2}.fastq' --trimFor 0 --trimRev
0 --truncRev 275 --truncFor 290 --reference /home/kviljoen/RefSeq-RDP16S_v3_May2018.fa.gz -profile uct_hpc --minOverlap 200 --rmPhiX T
```

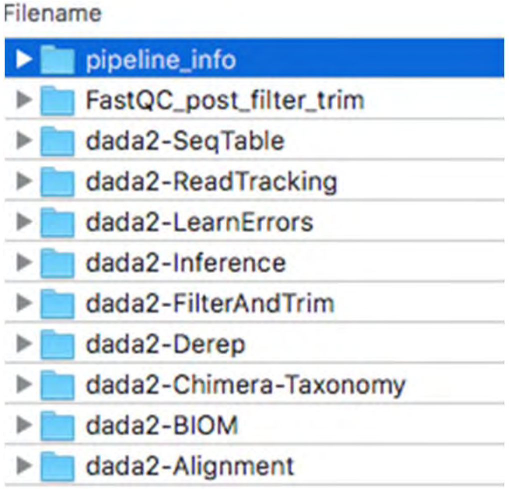
CPU-Hours	42.8
Launch directory	/home/kviljoen
Work directory	/home/kviljoen/work
Project directory	/home/kviljoen/.nextflow/assets/kviljoen/16S-rDNA-dada2-pipeline
Script name	main.nf
Script ID	1696132777285c87e164254634afd7f2
Workflow session	652ab8e0-94e0-421c-a1ca-6abc46f3864b
Workflow repository	https://github.com/kviljoen/16S-rDNA-dada2-pipeline.git , revision master (commit hash 200627f1960957328bf9a05965c57d75c2124098)



Failed pipeline execution

```
Command exit status:
  1

Command output:
  [1] '1.12.1'

Command error:
  WARNING: skipping mount of /data/ref-data: no such file or directory
  Loading required package: Rcpp
  Error: Input/Output
    no input files found
    dirPath: silva_nr_v132_train_set.fa.g
    pattern: character(0)
  Execution halted

Work dir:
  /cbio/home/klennard/16S-rDNA-dada2-pipeline/work/0a/0d0e75bea0ae60a4c56044de5998f2

Tip: you can replicate the issue by changing to the process work dir and entering the
command `bash .command.run`
```

Output folders/process



Filename

- pipeline_info
- FastQC_post_filter_trim
- dada2-SeqTable
- dada2-ReadTracking
- dada2-LearnErrors
- dada2-Inference
- dada2-FilterAndTrim
- dada2-Derep
- dada2-Chimera-Taxonomy
- dada2-BIOM
- dada2-Alignment

In case of errors:

- Inspect the .nextflow.log file in the directory where the pipeline was launched
- Find relevant working directory where error originated
  - Inspect/run individual .run.sh, .command.sh, command.log

H3ABioNet

Pan African Bioinformatics Network for H3Africa

H3ABioNet
Pan African Bioinformatics Network for H3Africa

16SrRNA Intermediate Bioinformatics Online Course

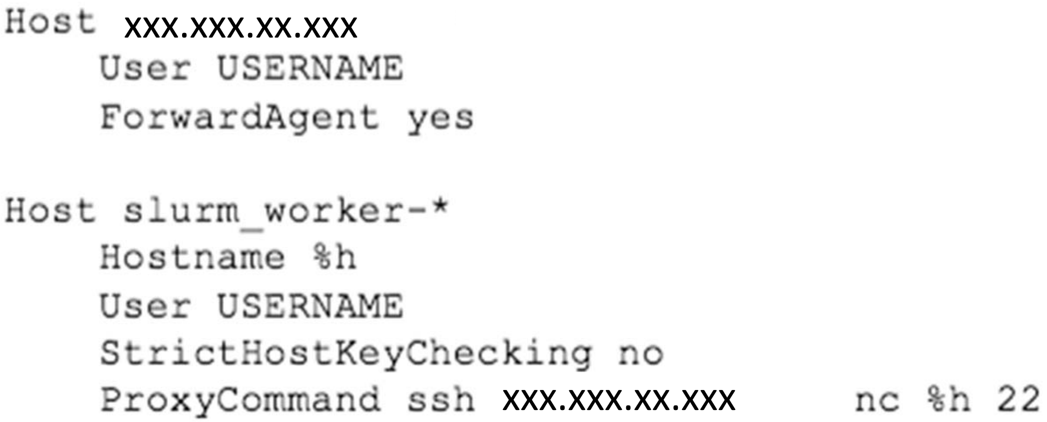
16S downstream analyses in R: importing data

Start RStudio from the Ilifu SLURM cluster

- Follow detailed instructions on Vula
  - Log in with ssh

ssh gerrit@xxx.xxx.xx.xxx

  - On your **local** machine add the following to your ~/.ssh/config file



```
Host  xxx.xxx.xx.xxx
    User USERNAME
    ForwardAgent yes

Host slurm_worker-*
    Hostname %h
    User USERNAME
    StrictHostKeyChecking no
    ProxyCommand ssh  xxx.xxx.xx.xxx      nc %h 22
```

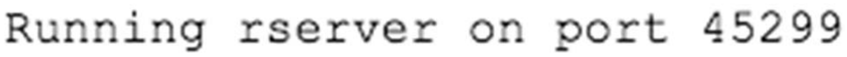
Start RStudio from the Ilifu SLURM cluster

– Start an interactive job on a worker node

```
srun --nodes=1 --ntasks 1 --mem=8g --pty bash
```

– Launch RStudio with:

```
USERNAME@slurm_worker-0002:~$
RSTUDIO_PASSWORD='Make your own secure
password here' /cbio/containers/bionic-R3.6.1-
RStudio1.2.1335-bio.simg
```



```
Running rserver on port 45299
```

Start RStudio from the Ilifu SLURM cluster

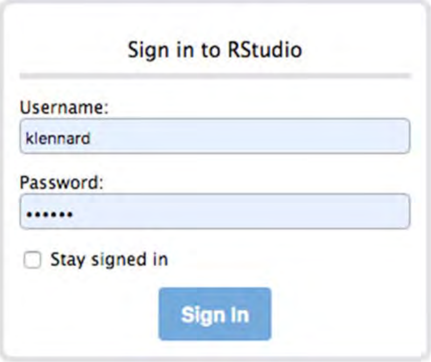
– From your local machine

```
$ ssh slurm_worker-0002 -L8082:localhost:45299
```

– From your browser

http://localhost:8082

– Enter your username and password



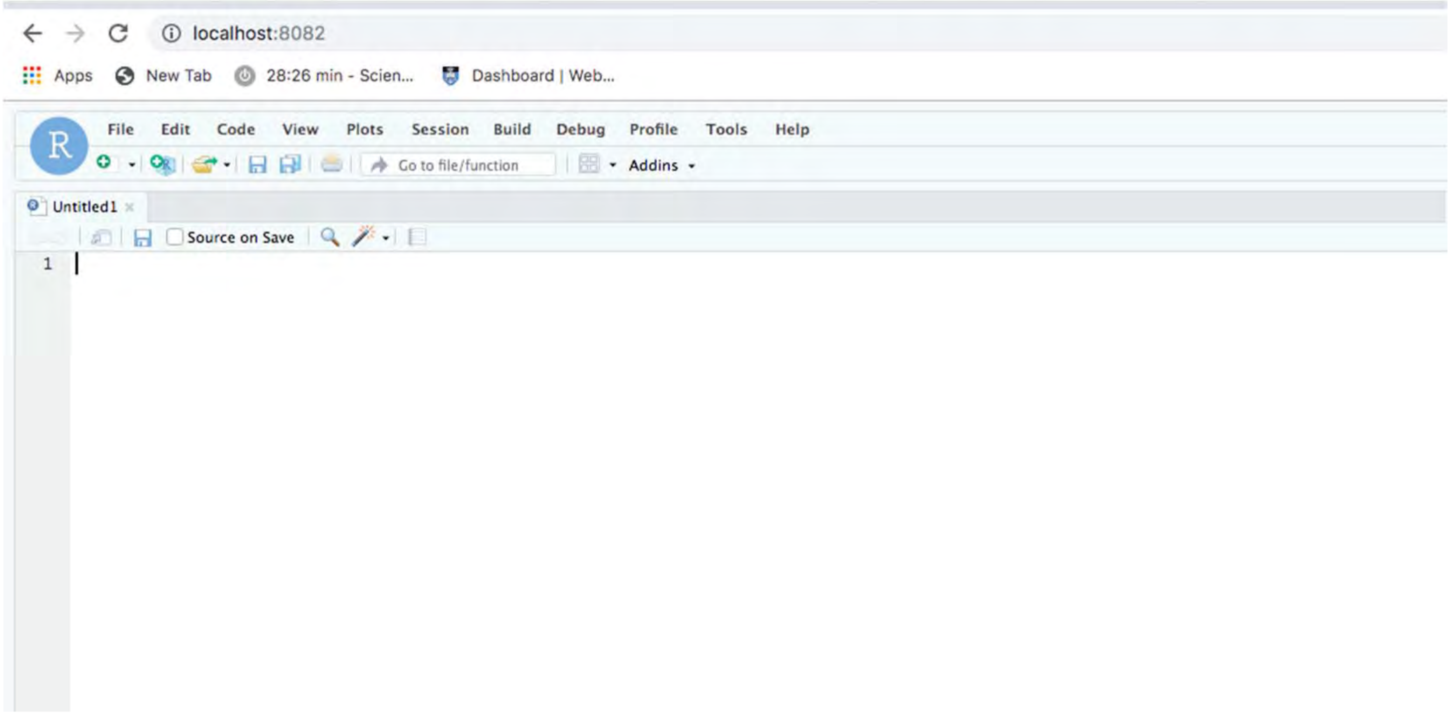
Sign in to RStudio

Username: klennard

Password: ••••••

☐ Stay signed in

Start RStudio from the Ilifu SLURM cluster

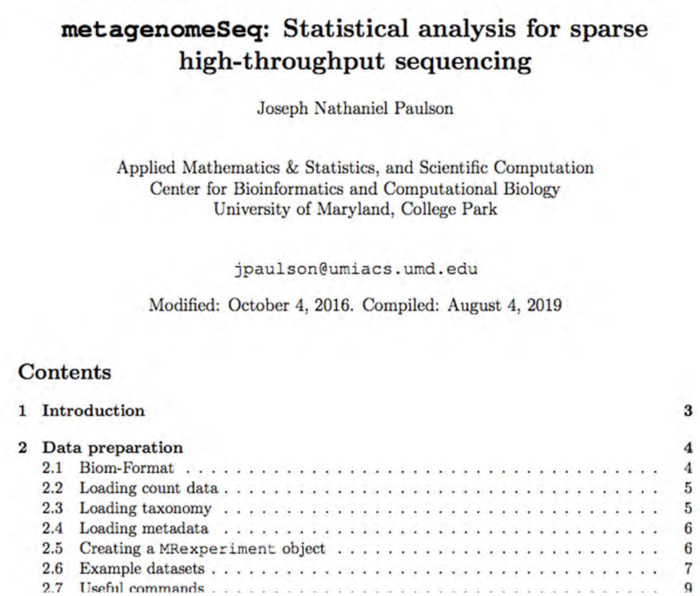


16S microbiome data characteristics

- Count data: skewed, zero-inflated distribution
- Differences in absolute read count between samples need normalization
- Redundant taxonomic information: merging?

16S downstream analyses in R

- Microbiome-specific packages in R
  - exploratory analyses: vegan, phyloseq
  - differential abundance testing: metagenomeSeq



metagenomeSeq: Statistical analysis for sparse high-throughput sequencing

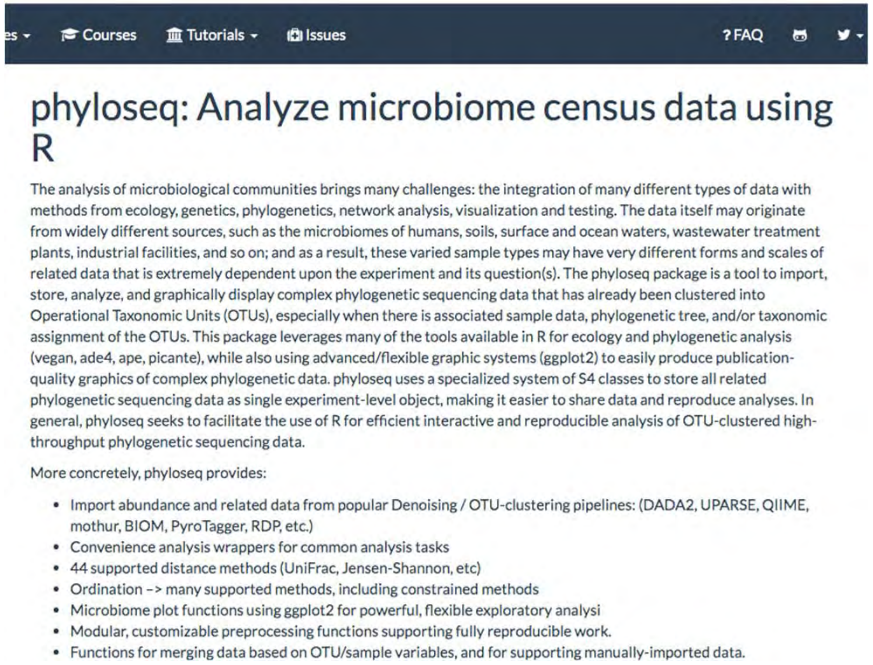
Joseph Nathaniel Paulson

Applied Mathematics & Statistics, and Scientific Computation
Center for Bioinformatics and Computational Biology
University of Maryland, College Park

jpaulson@umiacs.umd.edu

Modified: October 4, 2016. Compiled: August 4, 2019

Contents



phyloseq: Analyze microbiome census data using R

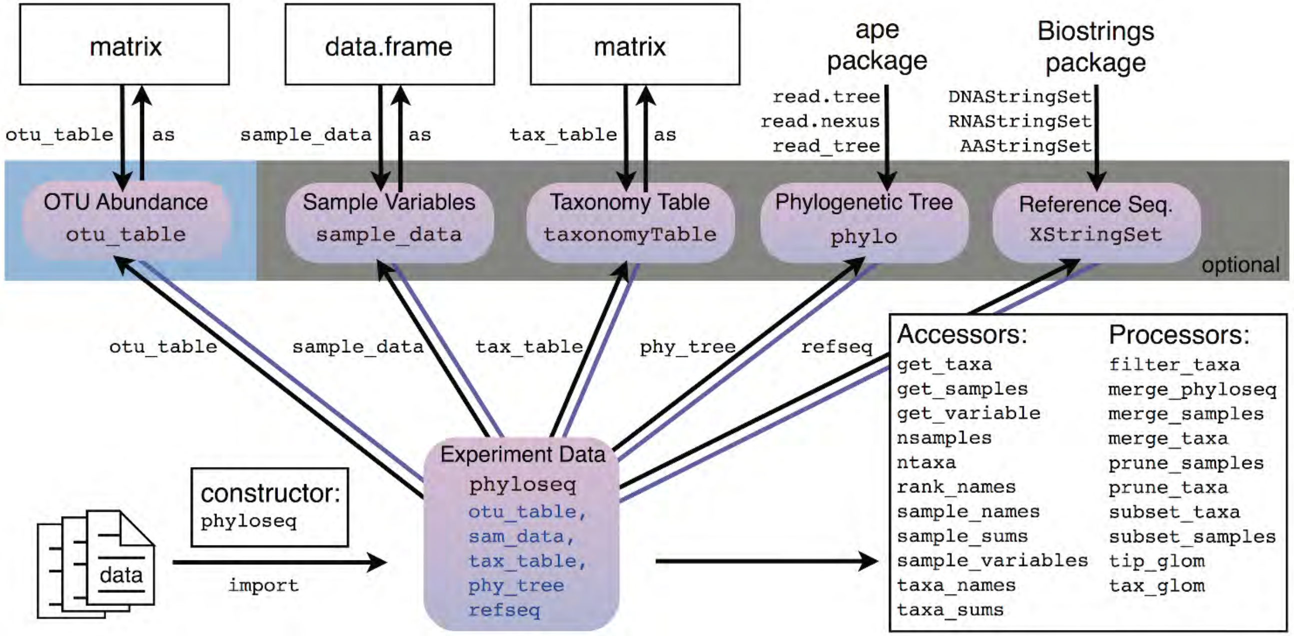
The analysis of microbiological communities brings many challenges: the integration of many different types of data with methods from ecology, genetics, phylogenetics, network analysis, visualization and testing. The data itself may originate from widely different sources, such as the microbiomes of humans, soils, surface and ocean waters, wastewater treatment plants, industrial facilities, and so on; and as a result, these varied sample types may have very different forms and scales of related data that is extremely dependent upon the experiment and its question(s). The phyloseq package is a tool to import, store, analyze, and graphically display complex phylogenetic sequencing data that has already been clustered into Operational Taxonomic Units (OTUs), especially when there is associated sample data, phylogenetic tree, and/or taxonomic assignment of the OTUs. This package leverages many of the tools available in R for ecology and phylogenetic analysis (vegan, ade4, ape, picante), while also using advanced/flexible graphic systems (ggplot2) to easily produce publication-quality graphics of complex phylogenetic data. phyloseq uses a specialized system of S4 classes to store all related phylogenetic sequencing data as single experiment-level object, making it easier to share data and reproduce analyses. In general, phyloseq seeks to facilitate the use of R for efficient interactive and reproducible analysis of OTU-clustered high-throughput phylogenetic sequencing data.

More concretely, phyloseq provides:

- Import abundance and related data from popular Denoising / OTU-clustering pipelines: (DADA2, UPARSE, QIIME, mothur, BIOM, PyroTagger, RDP, etc.)
- Convenience analysis wrappers for common analysis tasks
- 44 supported distance methods (UniFrac, Jensen-Shannon, etc)
- Ordination -> many supported methods, including constrained methods
- Microbiome plot functions using ggplot2 for powerful, flexible exploratory analysi
- Modular, customizable preprocessing functions supporting fully reproducible work.
- Functions for merging data based on OTU/sample variables, and for supporting manually-imported data.

The R package ‘phyloseq’

- phyloseq uses a specialized system of S4 classes to store all related phylogenetic sequencing data as single experiment-level object



16S data import in R

- To import:
  - dada2 output: ASV table, taxonomic annotation
  - metadata: user-defined sample data (.csv or .txt)

H3ABioNet
Pan African Bioinformatics Network for H3Africa

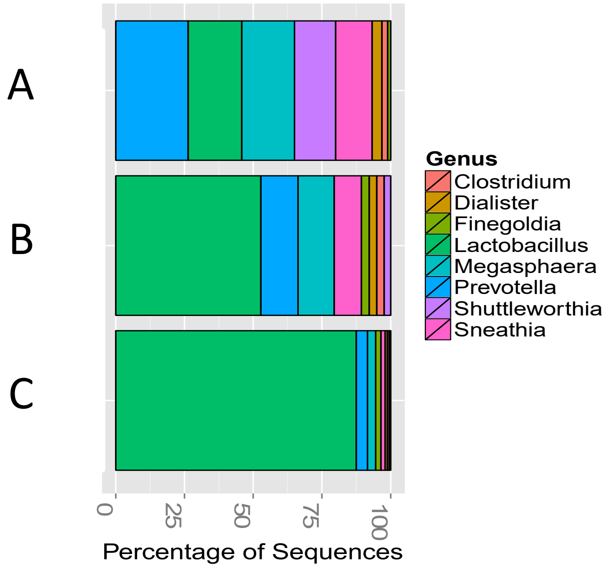
16SrRNA Intermediate Bioinformatics Online Course

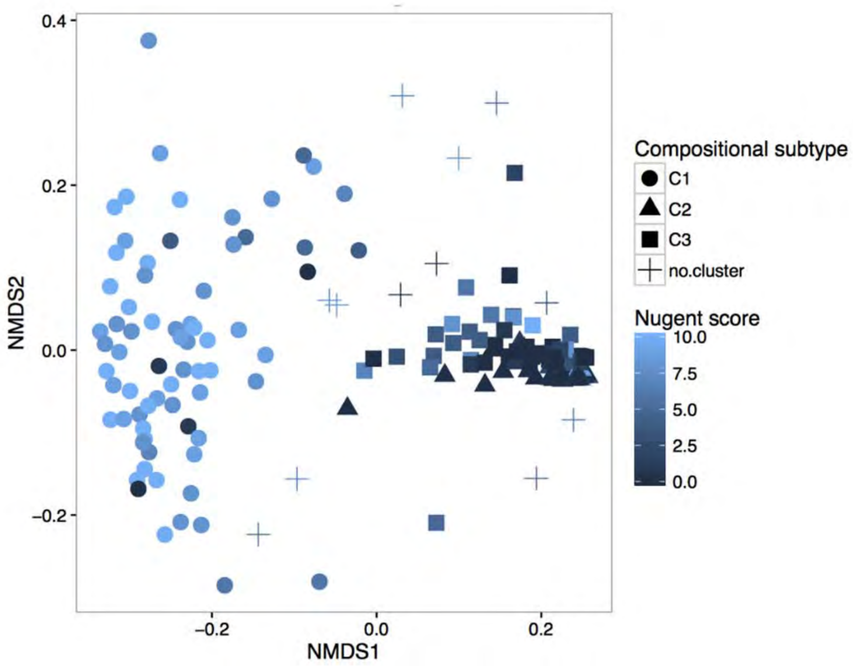
16S downstream analyses in R

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Microbial diversity estimates

- Bacterial diversity can be estimated
  - single sample measure (alpha diversity) or
  - between samples similarity (beta diversity)

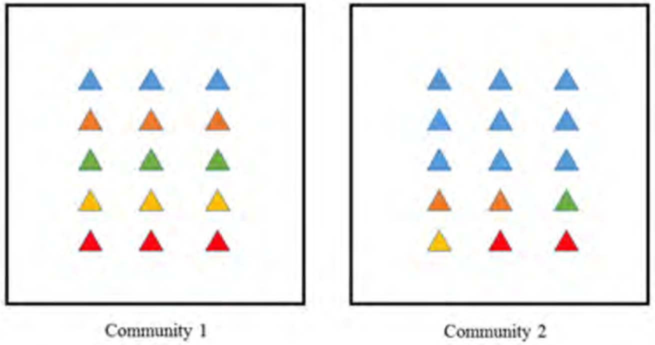




Alpha diversity metrics

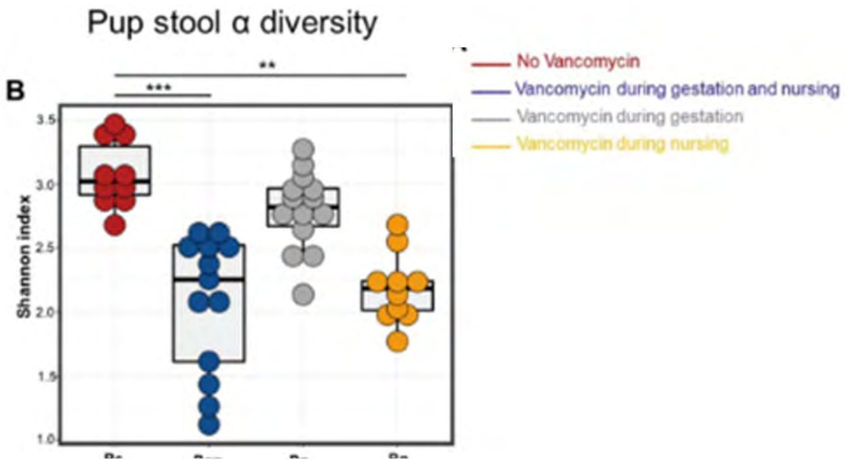
- Measure of within-sample richness and evenness

Community 1 vs. 2: Same richness (number of distinct taxa), different evenness



[1] http://www.jmb.or.kr/submission/Journal/027/JMB027-12-02_FDOC_2.pdf

- Why useful?
  - Microbiota diversity often related to biological outcomes



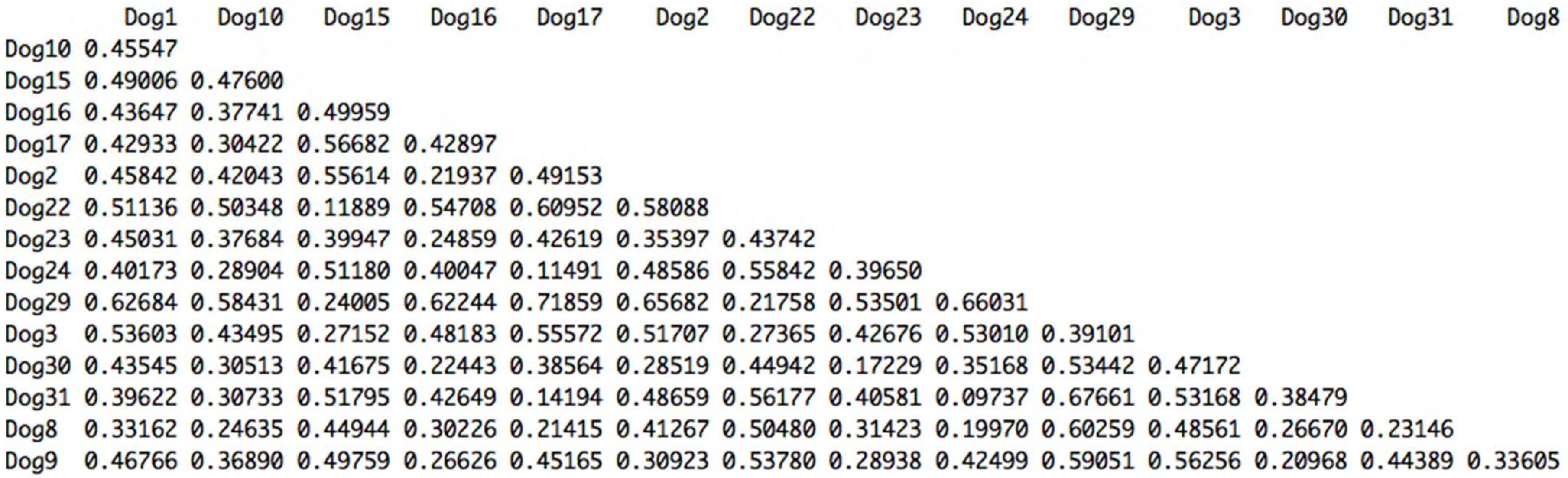
https://doi.org/10.1186/s40168-018-0511-7

Alpha diversity metrics

- Alpha **diversity** metrics that incorporate **richness & evenness**
  - Shannon
    - places greater weight on richness
  - Simpson
    - places greater weight on evenness
- Abundance-based measures of **richness**
  - Chao1: non-parametric method for estimating the number of species in a community
    - uses singletons, doubletons to estimate number of missing species
      - does not account for misclassification uncertainty?
    - NB: don’t use with dada2 as dada2 automatically removes singletons: “DADA2 does not call singletons because of how difficult it is to robustly distinguish between real singletons and singleton errors”

Beta diversity metrics

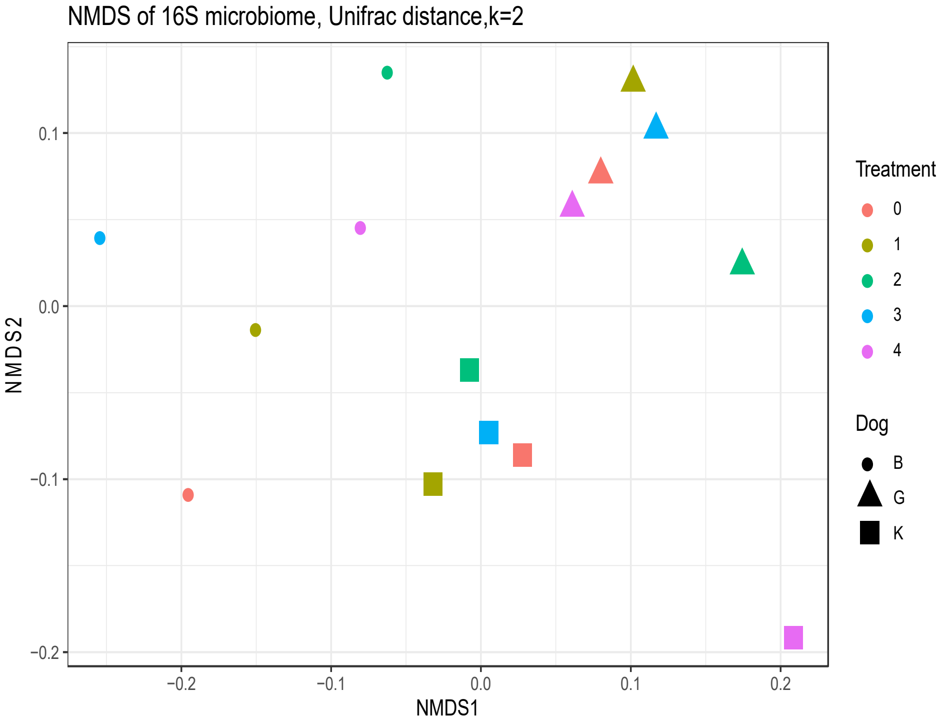
- First calculate **pairwise dissimilarity** between samples



```
         Dog1   Dog10   Dog15   Dog16   Dog17    Dog2   Dog22   Dog23   Dog24   Dog29    Dog3   Dog30   Dog31    Dog8
Dog10 0.45547
Dog15 0.49006 0.47600
Dog16 0.43647 0.37741 0.49959
Dog17 0.42933 0.30422 0.56682 0.42897
Dog2  0.45842 0.42043 0.55614 0.21937 0.49153
Dog22 0.51136 0.50348 0.11889 0.54708 0.60952 0.58088
Dog23 0.45031 0.37684 0.39947 0.24859 0.42619 0.35397 0.43742
Dog24 0.40173 0.28904 0.51180 0.40047 0.11491 0.48586 0.55842 0.39650
Dog29 0.62684 0.58431 0.24005 0.62244 0.71859 0.65682 0.21758 0.53501 0.66031
Dog3  0.53603 0.43495 0.27152 0.48183 0.55572 0.51707 0.27365 0.42676 0.53010 0.39101
Dog30 0.43545 0.30513 0.41675 0.22443 0.38564 0.28519 0.44942 0.17229 0.35168 0.53442 0.47172
Dog31 0.39622 0.30733 0.51795 0.42649 0.14194 0.48659 0.56177 0.40581 0.09737 0.67661 0.53168 0.38479
Dog8  0.33162 0.24635 0.44944 0.30226 0.21415 0.41267 0.50480 0.31423 0.19970 0.60259 0.48561 0.26670 0.23146
Dog9  0.46766 0.36890 0.49759 0.26626 0.45165 0.30923 0.53780 0.28938 0.42499 0.59051 0.56256 0.20968 0.44389 0.33605
```

Beta diversity metrics

- …then **plot** this dissimilarity matrix as an ordination
- Ordination: “a term used in ecology to refer to several multivariate techniques for visualization of species abundance in a low-dimensional space” [1]
- Position samples in a space of reduced dimensionality while preserving their distance relationships as well as possible

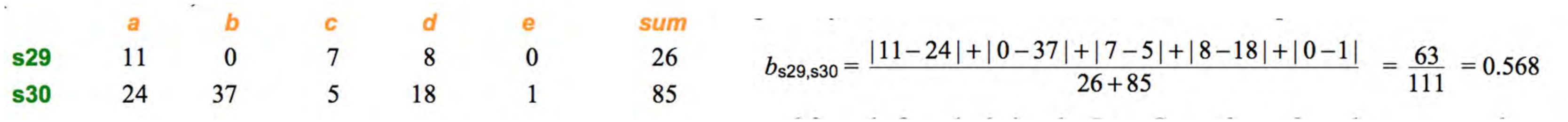


Beta diversity metrics

- Suitable distance metrics for 16S (ecological) data should consider:
  - Abundance
  - Composition
  - Phylogenetic relatedness
  - Why not Euclidean?
    - if you're comparing two samples certain species may be absent/zero in both samples – Euclidean distance would make these two samples look more similar even though this may not be the case

Beta diversity metrics

- Bray-Curtis dissimilarity $$b_{ii'} = \frac{\sum_{j=1}^{J} | n_{ij} - n_{i'j} |}{n_{i+} + n_{i'+}}$$

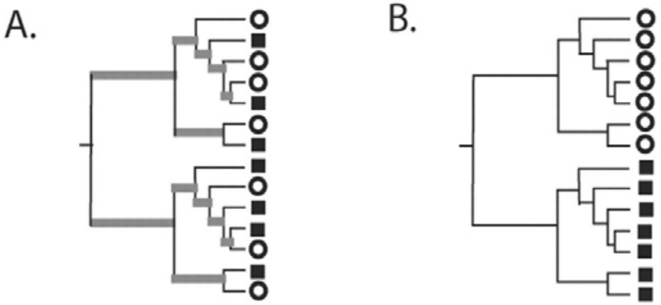


	a	b	c	d	e	sum
s29	11	0	7	8	0	26
s30	24	37	5	18	1	85

$$b_{s29,s30} = \frac{|11-24|+|0-37|+|7-5|+|8-18|+|0-1|}{26+85} = \frac{63}{111} = 0.568$$

- Unifrac distance
  - Measure of phylogenetic relatedness
  - Weighted Unifrac:

phylogeny + abundance (quantitative)



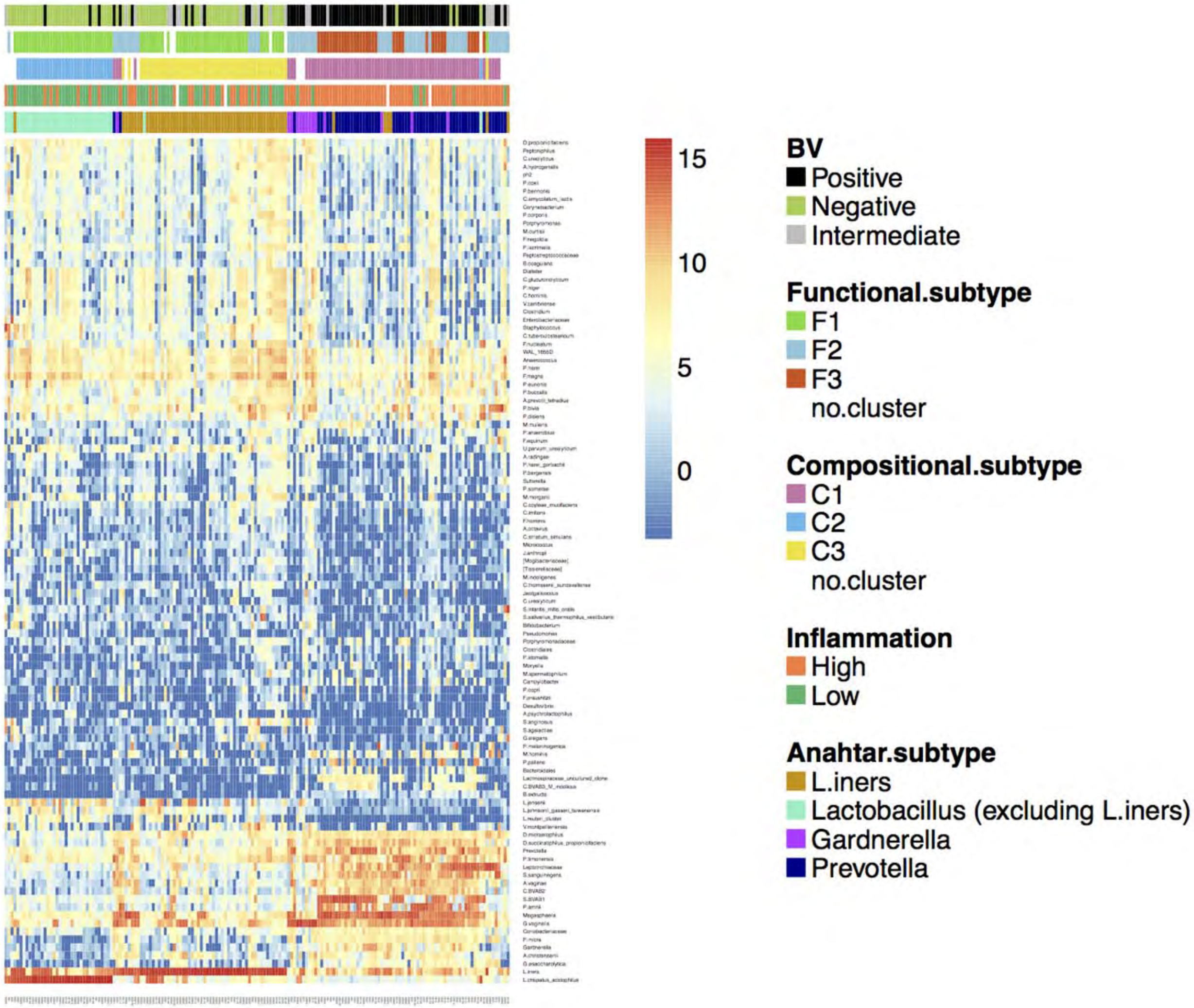
https://aem.asm.org/content/aem/71/12/8228.full.pdf

H3ABioNet
Pan African Bioinformatics Network for H3Africa

Ordination methods

- Multidimensional scaling (MDS) = Principal coordinates analysis (PCoA)
  - Principal components analysis (PCA): simplest case of MDS where the dissimilarity metric is Euclidean distance (which is not appropriate for ecological data)
  - Axes measure of importance (% variation explained)
- Non-metric MDS (NMDS)
  - Iterative method (non-metric) converting raw dissimilarity values into ranks
    - NB: iterative so random start seed should be recorded
  - Locates samples in N-dimensional space so that Euclidean distance (in ordination) between samples correspond to compositional dissimilarity (calculated by Bray-Curtis, Unifrac etc.)
    - Stress value: measure of fit between ordination and input dissimilarity (lower is better, with 0.1 as ballpark minimum acceptable)
- Should I use MDS or NMDS?
  - NMDS is better than MDS if MDS requires 3 or more dimensions to represent the main distance relationship among sites. NMDS is able to 'squeeze' (distort) the ordination into two dimensions (which is useful for publication purposes)

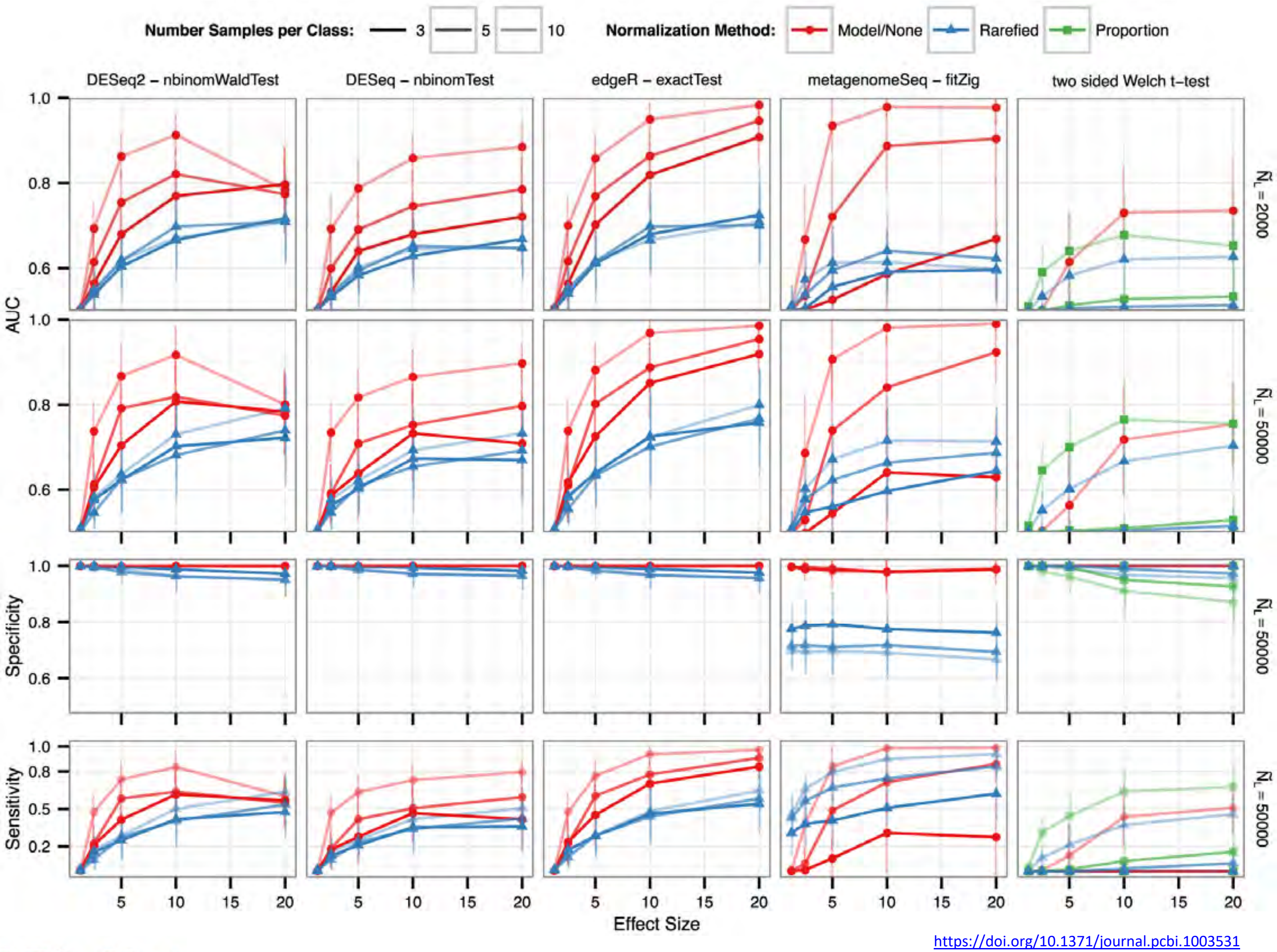
Annotated heatmaps



Differential abundance testing

- Suitable methods should consider:
  - Zero-inflated nature of 16S data
  - Differences in sampling depth between samples
- Non-parametric tests like Wilcoxon rank-sum or Kruskal-Wallis?
- More powerful parametric alternatives:
  - DESeq2
  - edgeR
    - Originally designed for RNASeq data
    - Both fit a generalized linear models and assume that read counts follow a Negative Binomial distribution.
  - **MetagenomeSeq**
    - Specifically designed for zero-inflated count data with variable coverage between samples
    - Generalized linear model with zero-inflated Gaussian distribution (abundance testing) + presence/absence statistic (Fisher's exact)

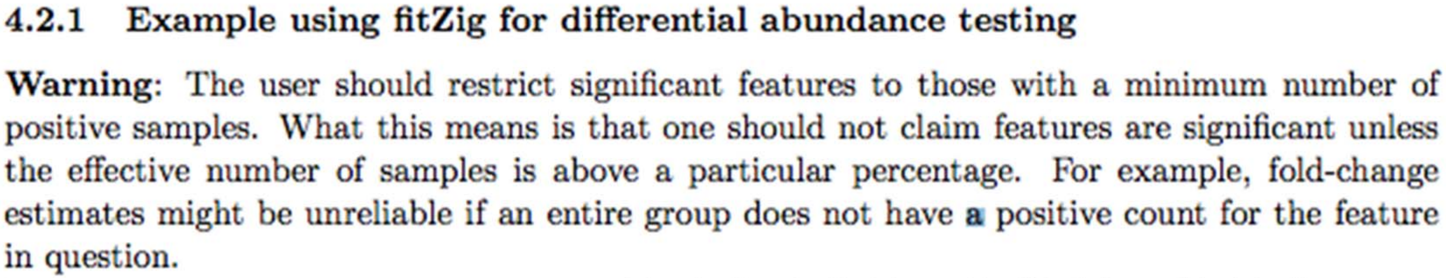
Differential abundance testing: method comparison



https://doi.org/10.1371/journal.pcbi.1003531

metagenomeSeq

- Normalization method accounts for differences in sampling depth (cumulative sum scaling)
- Zero-inflated GLM computes probability of zero due to:
  - low sampling depth (present but not observed) vs.
  - sparsity (biological zero, truly absent)
- Can include covariates (prevent confounding)
- Results require filtering based on ASV presence to avoid false positives, particularly with small sample sizes
- Further reading



4.2.1 Example using fitZig for differential abundance testing

Warning: The user should restrict significant features to those with a minimum number of positive samples. What this means is that one should not claim features are significant unless the effective number of samples is above a particular percentage. For example, fold-change estimates might be unreliable if an entire group does not have a positive count for the feature in question.