

16S community sequencing pipeline using dada2 in R

Daniel Padfield

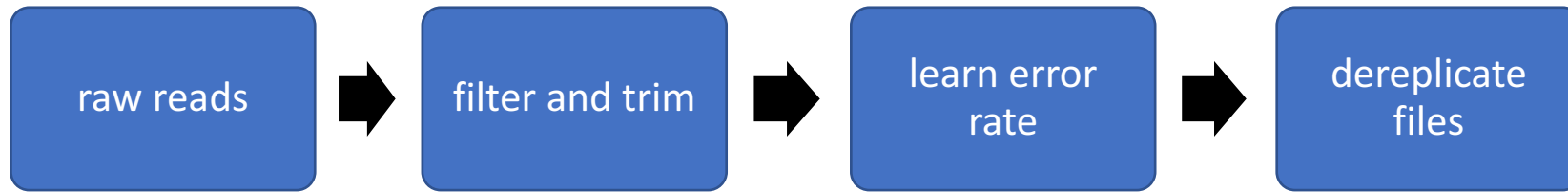
16S community sequencing analysis

- 16S rRNA variable regions
- Most methods cluster sequences that occur with 97% similarity and assign these to “OTUs” from references trees
- Do not incorporate quality scores of sequences

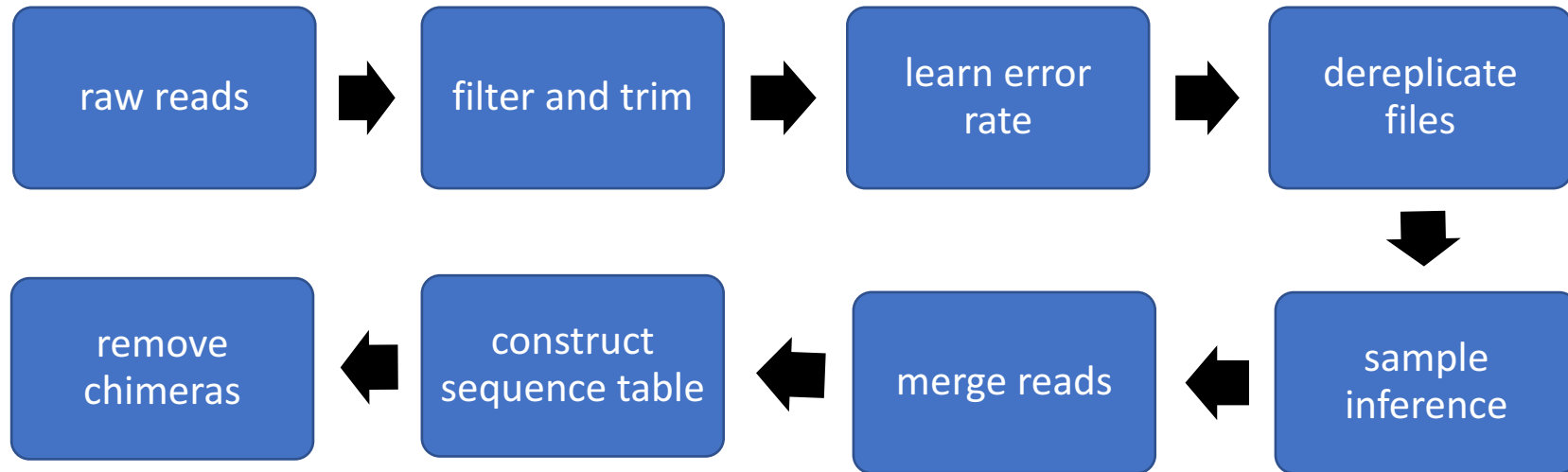
16S community sequencing analysis

- 16S rRNA variable regions
- Most methods cluster sequences that occur with 97% similarity and assign these to “OTUs” from references trees
- Do not incorporate quality scores of sequences
- **Enter dada2**
- Incorporates quality score information
- Distinguishes sequencing errors from real biological variation
- Finds Amplicon Sequence Variants (ASVs)

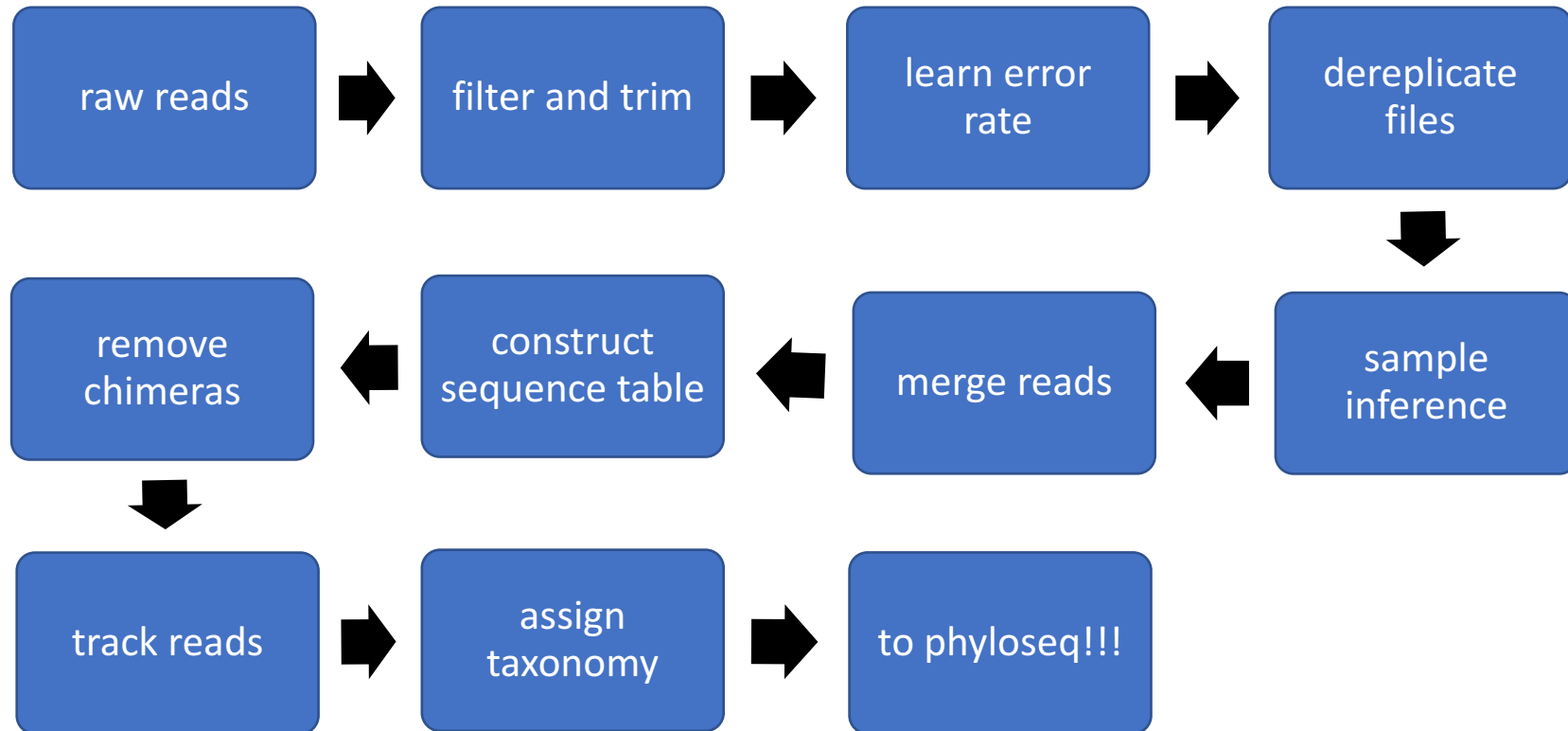
dada2 pipeline



dada2 pipeline



dada2 pipeline



The AB sequencing pipeline

padpadpadpad / AB_dada2_pipeline_R

Unwatch

1

Star

0

Fork

0

<> Code

Issues 3

Pull requests 0

Projects 0

Wiki

Insights

Settings

Pipeline for running sequencing analyses using dada2

Edit

Add topics

13 commits

1 branch

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

padpadpadpad update pipeline

Latest commit 6c0f734 2 days ago

data	update pipeline	2 days ago
plots	initial commit	11 days ago
scripts	update pipeline	2 days ago
workshop	Add prep info	2 days ago
.gitignore	update .gitignore	2 days ago
AB_dada2_pipeline_R.Rproj	initial commit	11 days ago
README.md	Update dummy analysis	10 days ago

The AB sequencing pipeline

padpadpadpad / AB_dada2_pipeline_R

Unwatch 1 Star 0 Fork 0

Code Issues 3 Pull requests 0 Projects 0 Wiki Insights Settings

Pipeline for running sequencing analyses using dada2 [Edit](#)

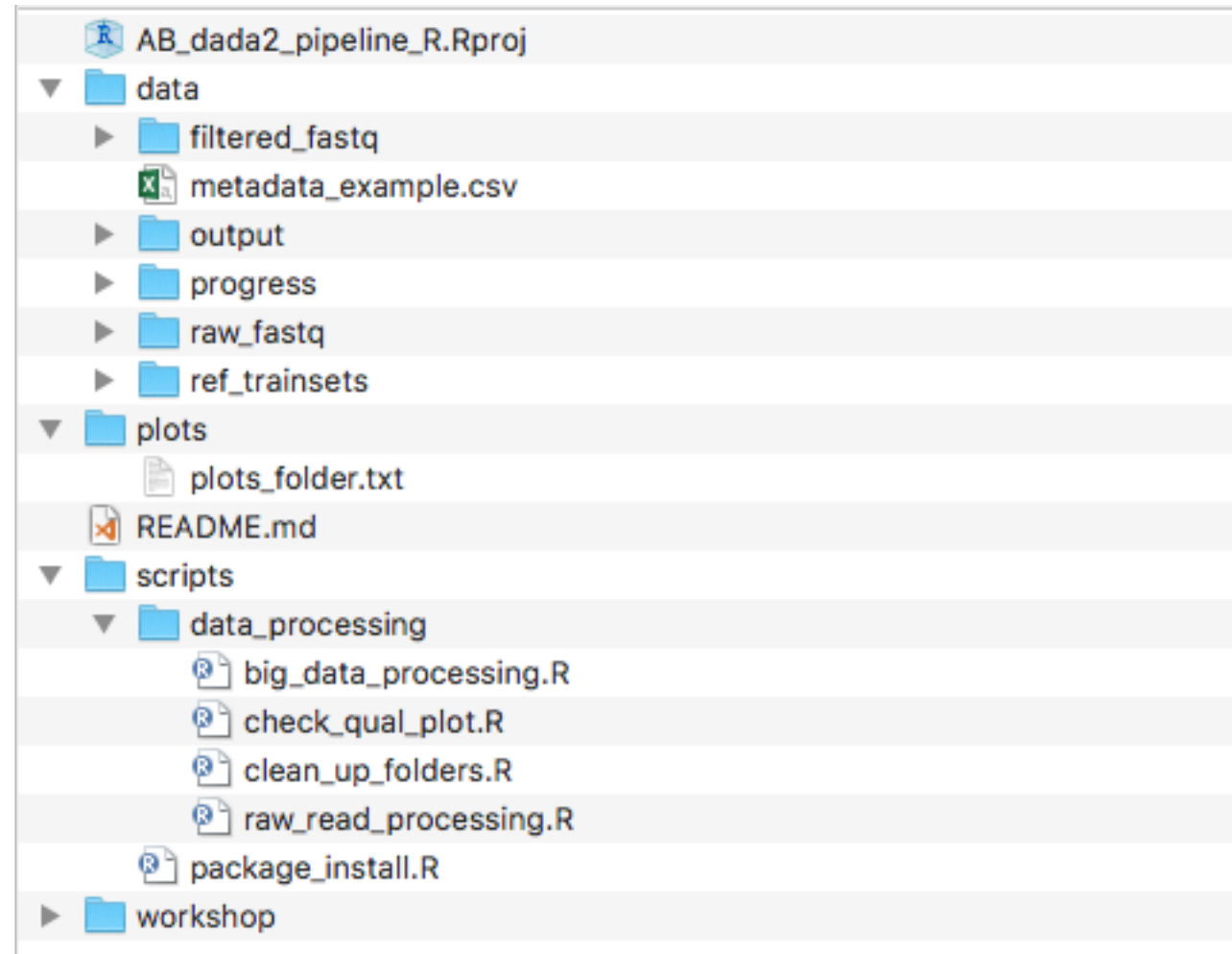
[Add topics](#)

13 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file **Clone or download**

padpadpadpad update pipeline		Latest commit 6c0f734 2 days ago
data	update pipeline	2 days ago
plots	initial commit	11 days ago
scripts	update pipeline	2 days ago
workshop	Add prep info	2 days ago
.gitignore	update .gitignore	2 days ago
AB_dada2_pipeline_R.Rproj	initial commit	11 days ago
README.md	Update dummy analysis	10 days ago

The AB sequencing pipeline



raw_read_processing vs big_data_processing

- Pooling data allows information to be shared across samples. However is much much slower for big data sets
- dada2 resolves sequence variants exactly, and as such DNA sequences are consistent across samples and can be recombined separately
- big_data_processing speeds things up if needs be!

Bits to run supervised

- Quality profiles
- Setting trimming parameters
- Check sample names = metadata\$SampleID
- Make sure there is a column called SampleID!

Correctly format metadata

SampleID == sample_names !!!

	A	B	C
1	SampleID	treatment	ancestral
2	sample_1-1	A	warm
3	sample_10-10	A	amb
4	sample_11-11	A	amb
5	sample_12-12	A	warm
6	sample_14-14	A	warm
7	sample_15-15	A	warm
8	sample_16-16	A	amb
9	sample_17-17	A	warm
10	sample_18-18	A	amb
11	sample_19-19	A	warm
12	sample_2-2	A	amb
13	sample_20-20	A	amb
14	sample_21-C1	A	comb
15	sample_22-C2	A	comb
16	sample_23-C3	A	comb
17	sample_24-C4	A	comb
18	sample_25-C5	A	comb
19	sample_26-C6	A	comb
20	sample_27-C7	A	comb

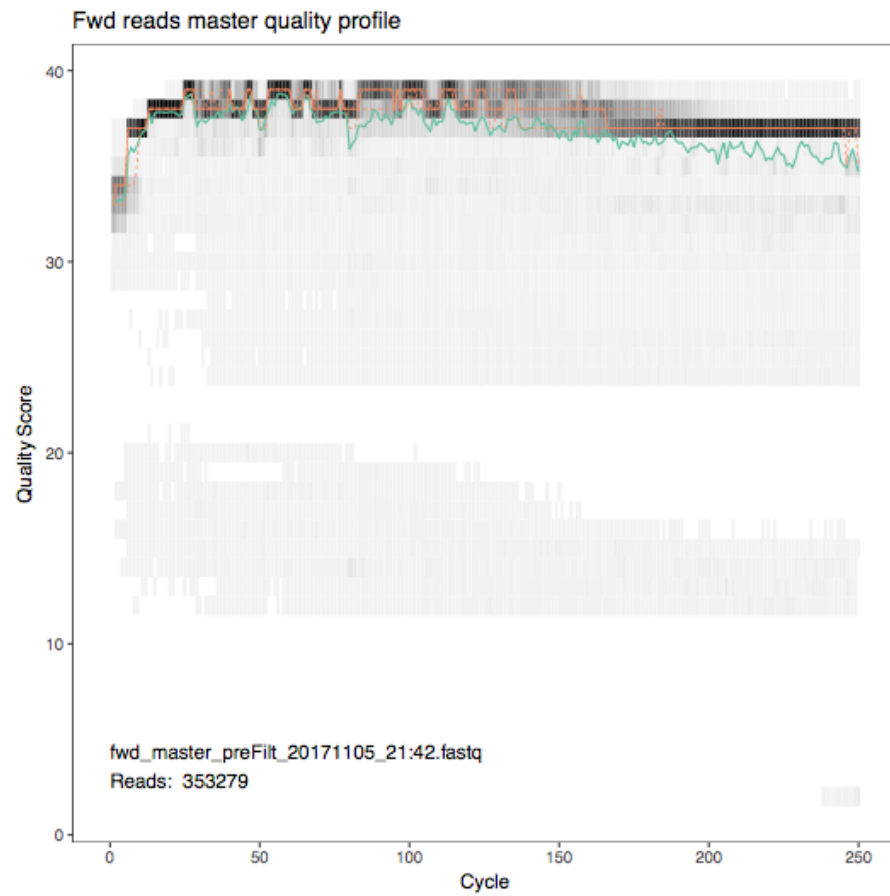
Progress file

```
Run started at 2017-11-03 23:22:55

This run is done using raw_read_processing.R
Filtering completed at 2017-11-03 23:11
Forward error rates completed at 2017-11-03 23:42:09
Forward error rate: 68.6218264955491
Forward error rate: 1.4268482212032
Forward error rate: 0.0806254374037106
Forward error rate: 0.0045528251682451
Forward error rate: 0.000171459734376565
Forward error rate: 0
Reverse error rates completed at 2017-11-04 00:01:03
Reverse error rates: 65.4204274041772
Reverse error rates: 1.28380362213372
Reverse error rates: 0.0450086135629872
Reverse error rates: 0.00562287641467286
Reverse error rates: 0.00151911546430804
Reverse error rates: 0
Dereplicated forward and reverse files 2017-11-04 00:01:19
Forward sequences inferred 2017-11-04 00:05:25
Reverse sequences inferred 2017-11-04 00:09:43
Inferred forward and reverse sequences merged 2017-11-04 00:10:00
Sequence table constructed 2017-11-04 00:10:00
Chimeric sequences removed 2017-11-04 00:10:56
Taxonomy assigned 2017-11-04 00:17:09
Assigning species at 2017-11-04 00:17:09
80 out of 1546 were assigned to the species level. Of which 74 had genera consistent with the input table.
Species assigned 2017-11-04 00:18:48
End of raw read processing without construction of phylogeny 2017-11-04 00:18:55
This run (without phylogeny estimation) took: 0.933248837457763 hours
Sequences aligned 2017-11-04 00:20:39
Constructed phylogenetic tree 2017-11-04 01:35:28
End of raw read processing 2017-11-04 01:35:29
This run took: 2.20936129470666 hours
```

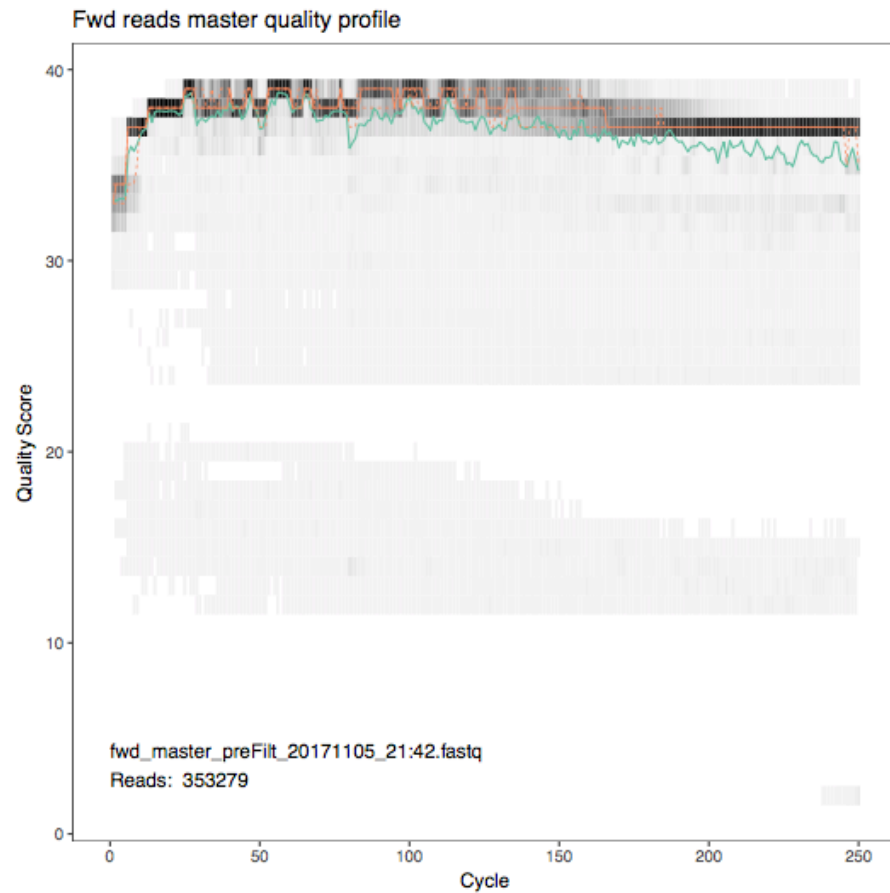
Plots

Before trimming and filtering

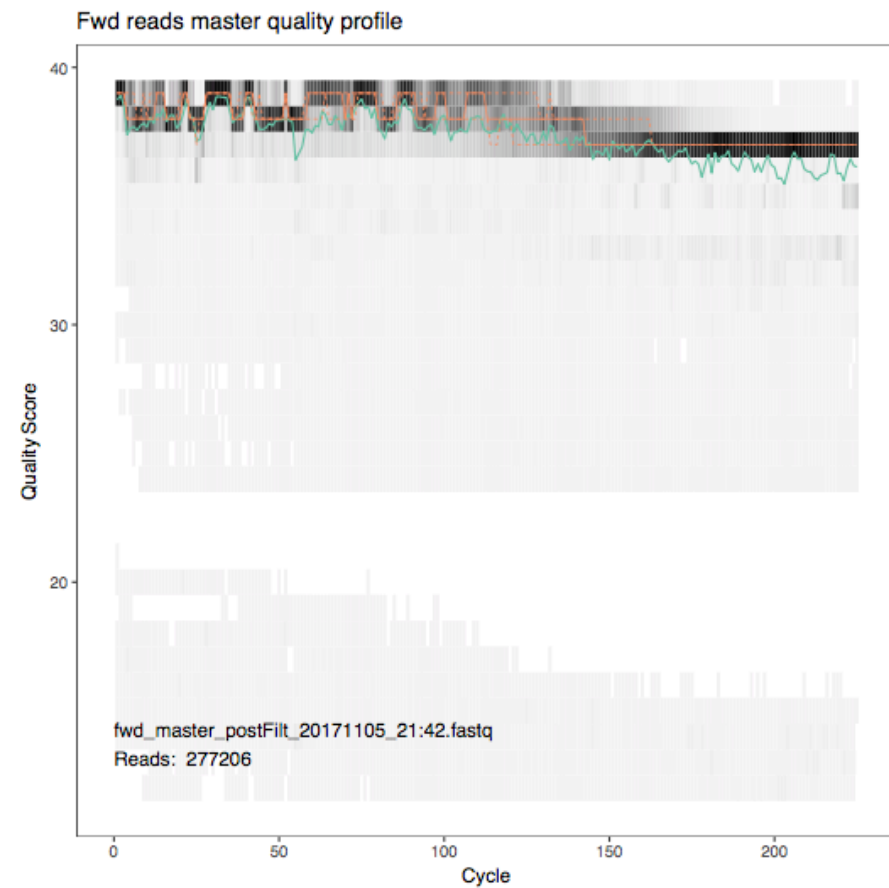


Plots

Before trimming and filtering

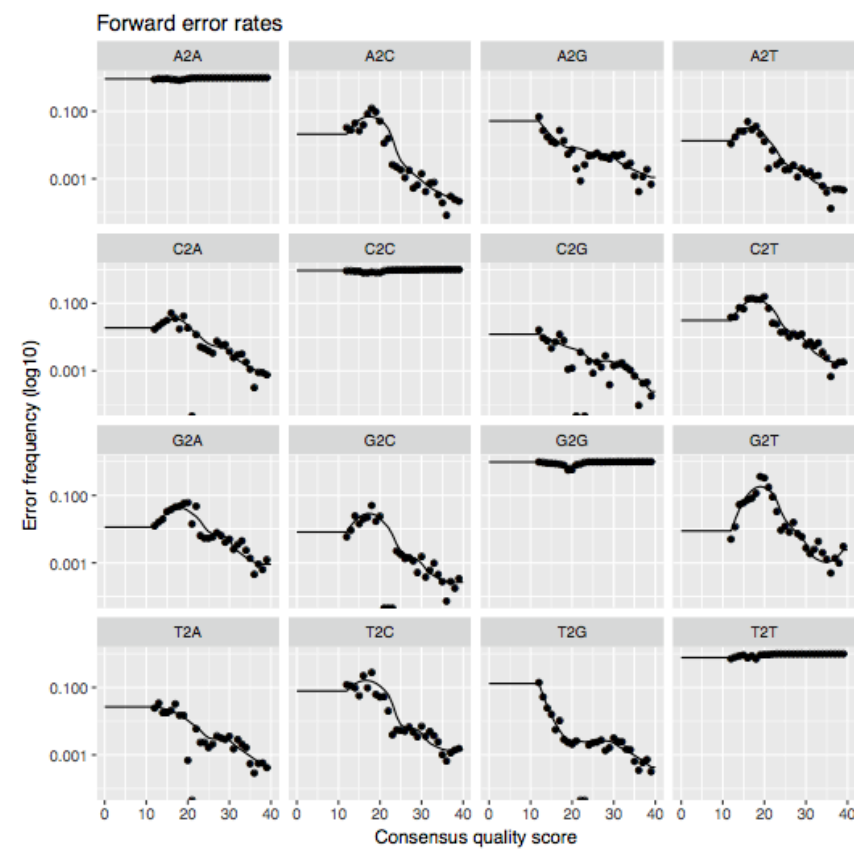


After trimming and filtering



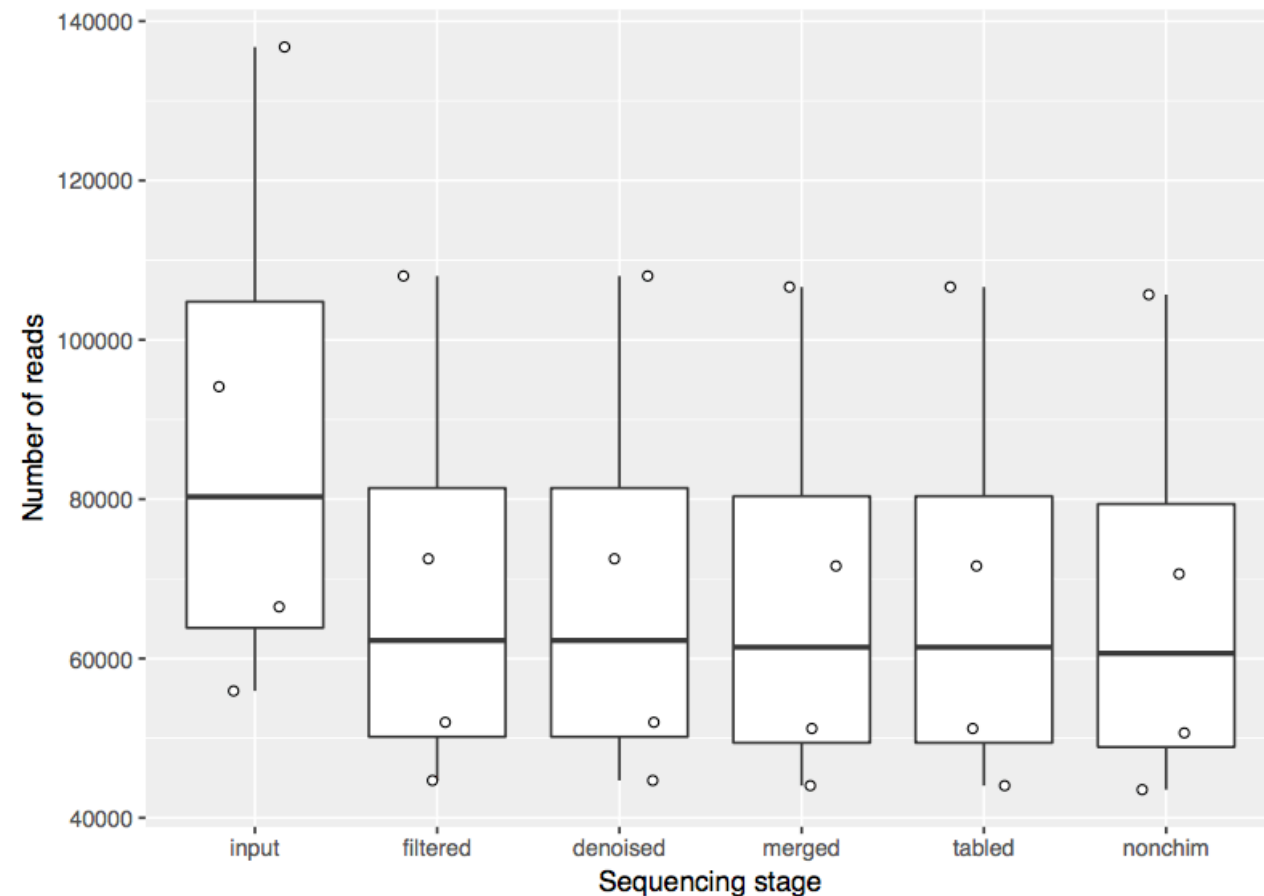
Plots

Plot of error rates



Plots

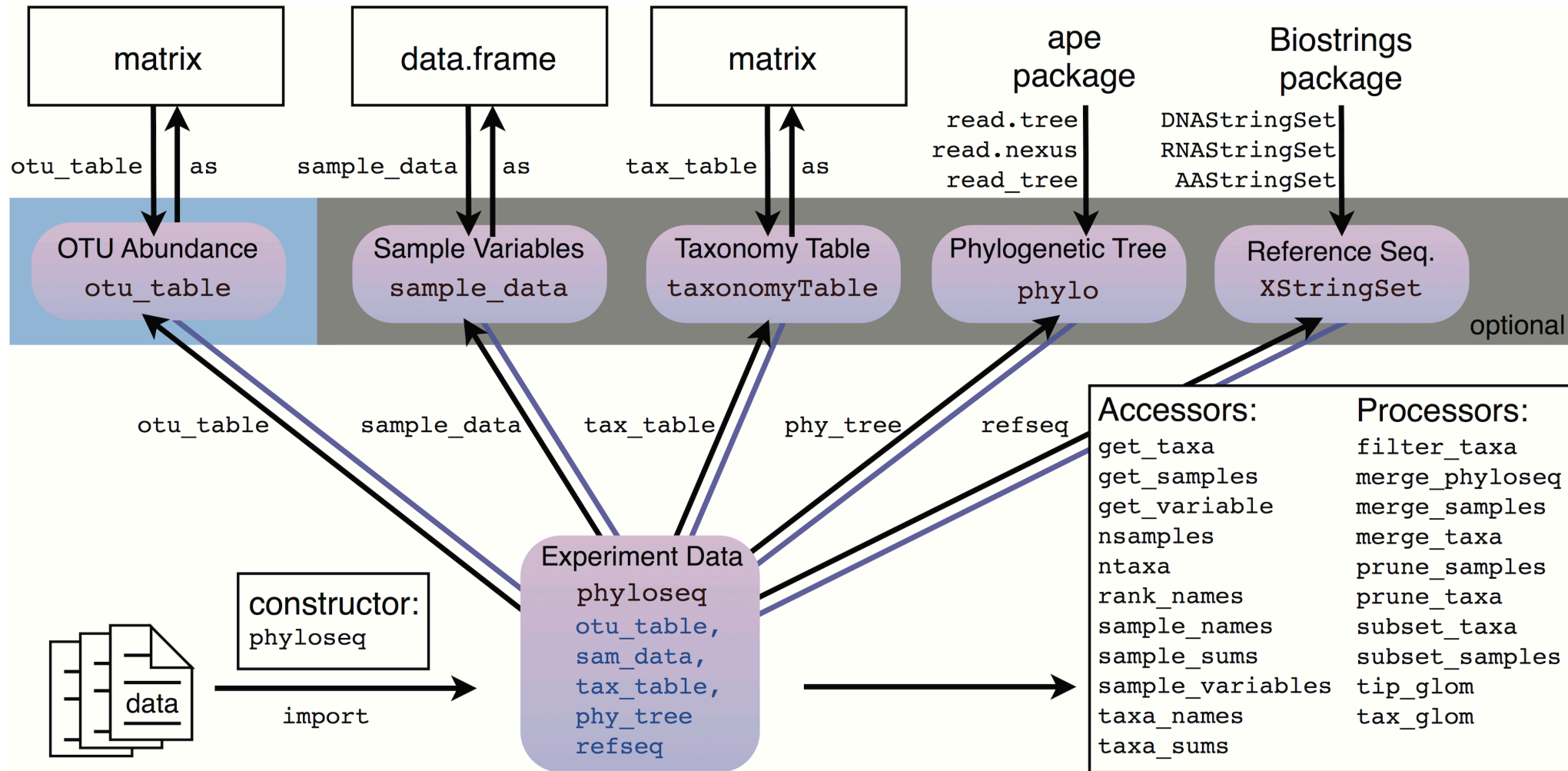
Track the number of reads through each stage



Output

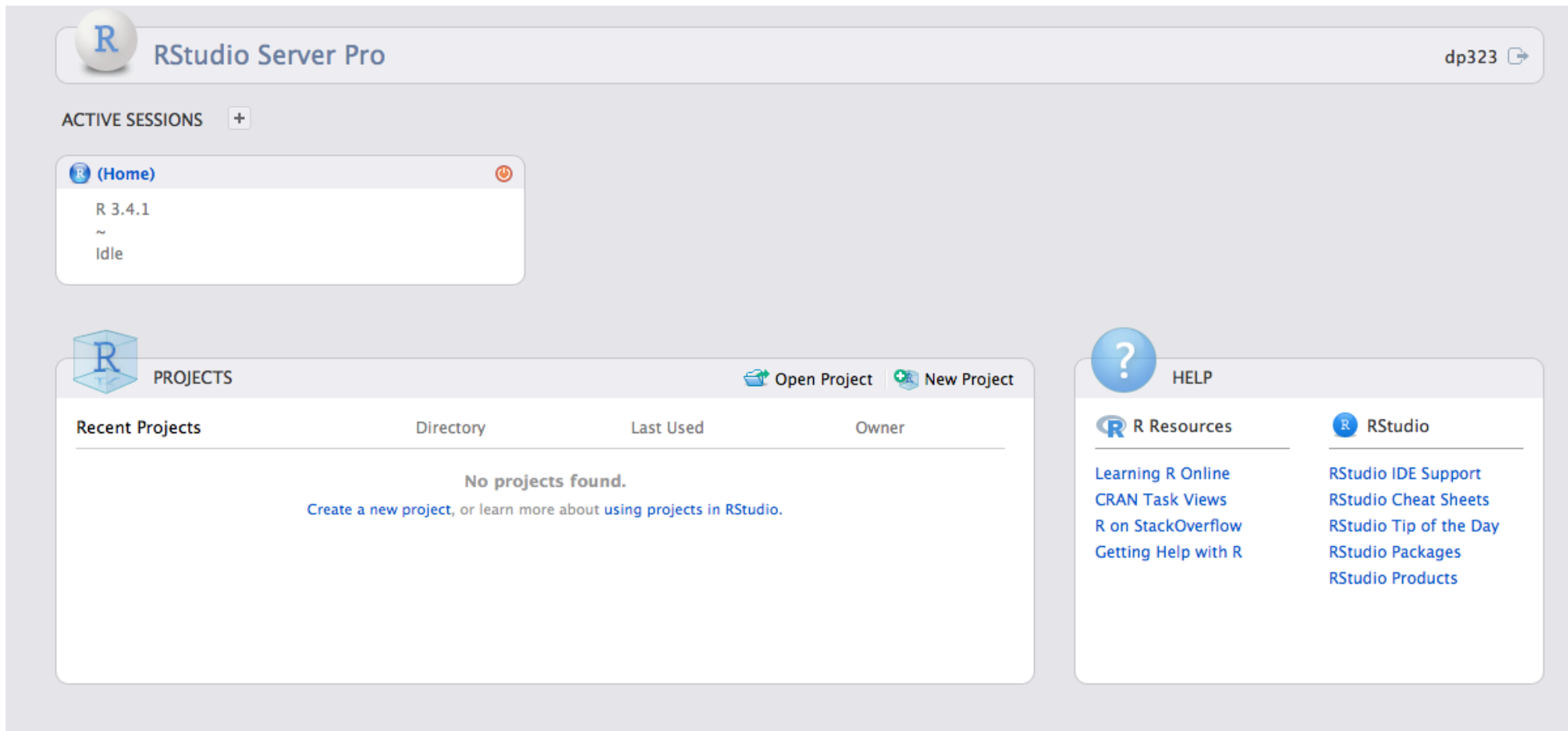
▼	20171107_13h43m	Today, 15:11	--	Folder
	20171107_13h43m_dadaFs.rds	Today, 14:02	6.6 MB	R Data File
	20171107_13h43m_dadaRs.rds	Today, 14:02	6.6 MB	R Data File
	20171107_13h43m_fwd_error.rds	Today, 13:58	22 KB	R Data File
	20171107_13h43m_phytree.rds	Today, 15:11	216 KB	R Data File
	20171107_13h43m_ps.Rdata	Today, 15:11	254 KB	R Data File
	20171107_13h43m_ps.rds	Today, 15:11	254 KB	R Data File
	20171107_13h43m_rev_error.rds	Today, 13:58	24 KB	R Data File
	20171107_13h43m_seqtab.rds	Today, 15:11	72 KB	R Data File
	20171107_13h43m_taxtab.rds	Today, 15:11	78 KB	R Data File
	20171107_13h43m_track_reads_through_stages.rds	Today, 14:03	230 bytes	R Data File
	fwd_master_postFilt_20171107_13h43m.fastq	Today, 13:44	24.2 MB	Document
	fwd_master_preFilt_20171107_13h43m.fastq	Today, 13:43	33.9 MB	Document
	rev_master_postFilt_20171107_13h43m.fastq	Today, 13:44	27.7 MB	Document
	rev_master_preFilt_20171107_13h43m.fastq	Today, 13:43	38.1 MB	Document

Bonus – integration with phyloseq



RStudio server



<https://rstudio01.cles.ex.ac.uk> (can be 01, 02, 03 or 04)




The screenshot displays the RStudio Server Pro web interface. At the top, the header shows the RStudio logo and the text "RStudio Server Pro" on the left, and the user identifier "dp323" with a refresh icon on the right. Below the header, the interface is divided into three main sections. The "ACTIVE SESSIONS" section on the left shows a single session for user "(Home)" using R 3.4.1, which is currently idle. The "PROJECTS" section in the center features a table with columns for "Recent Projects", "Directory", "Last Used", and "Owner". It indicates that no projects are currently found and provides a link to create a new project or learn more about using projects. The "HELP" section on the right, marked with a question mark icon, lists various resources under two categories: "R Resources" (including Learning R Online, CRAN Task Views, R on StackOverflow, and Getting Help with R) and "RStudio" (including RStudio IDE Support, RStudio Cheat Sheets, RStudio Tip of the Day, RStudio Packages, and RStudio Products).

RStudio Server Pro dp323


ACTIVE SESSIONS +



 (Home) 

R 3.4.1
~
Idle

 PROJECTS Open Project New Project

Recent Projects	Directory	Last Used	Owner
No projects found. Create a new project , or learn more about using projects in RStudio .			

 HELP

 R Resources  RStudio

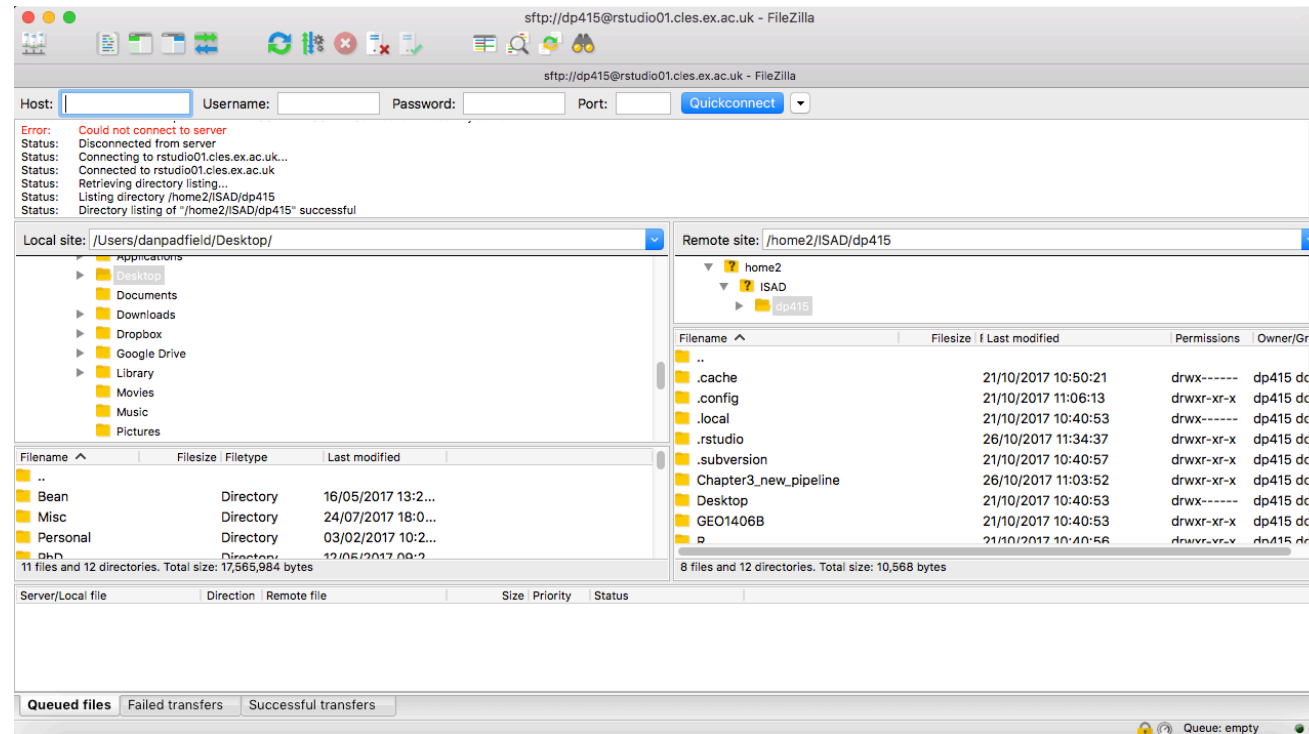
[Learning R Online](#)
[CRAN Task Views](#)
[R on StackOverflow](#)
[Getting Help with R](#)

[RStudio IDE Support](#)
[RStudio Cheat Sheets](#)
[RStudio Tip of the Day](#)
[RStudio Packages](#)
[RStudio Products](#)

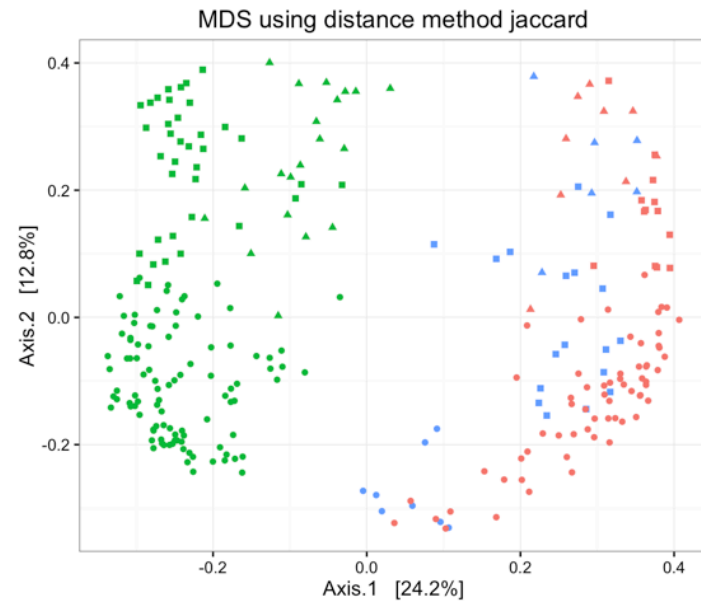
Filezilla

<https://filezilla-project.org>

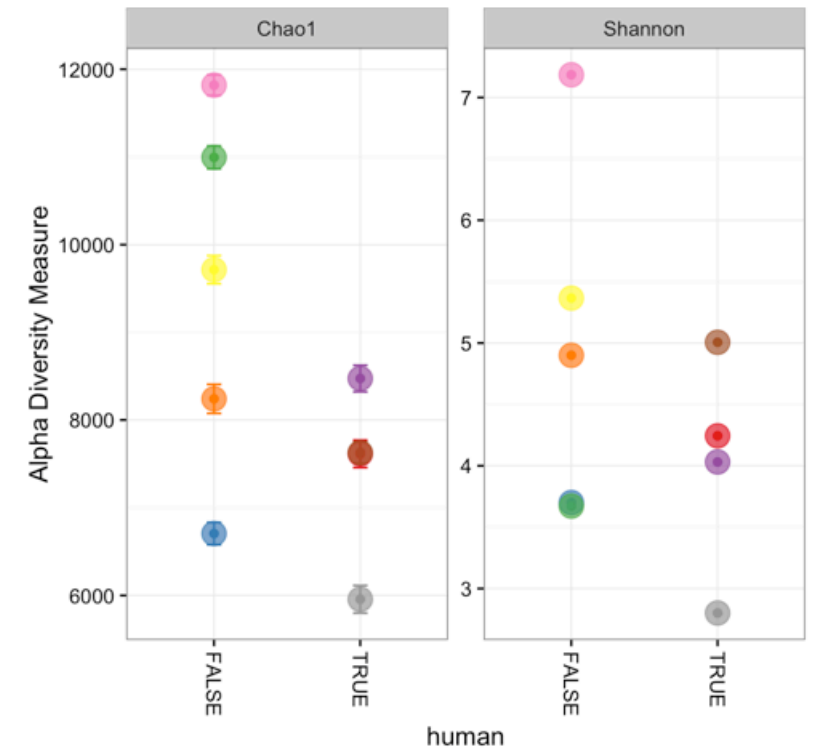
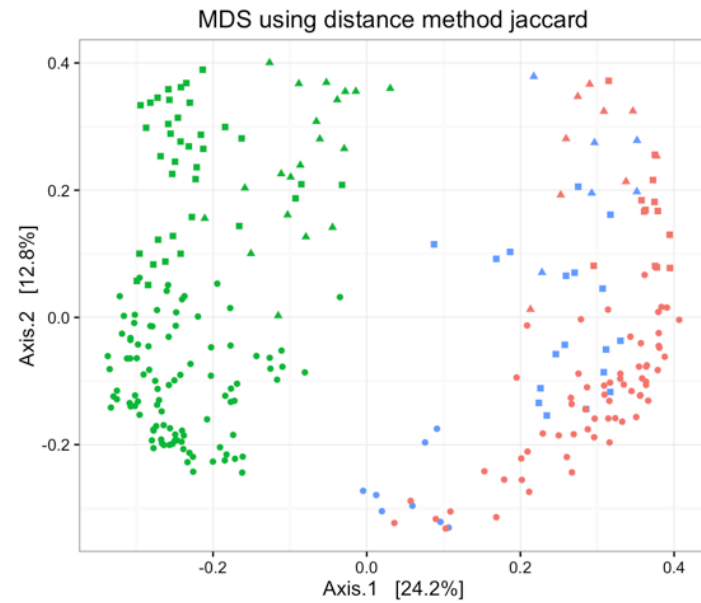
sftp://https://<your username>@rstudio01.cles.ex.ac.uk



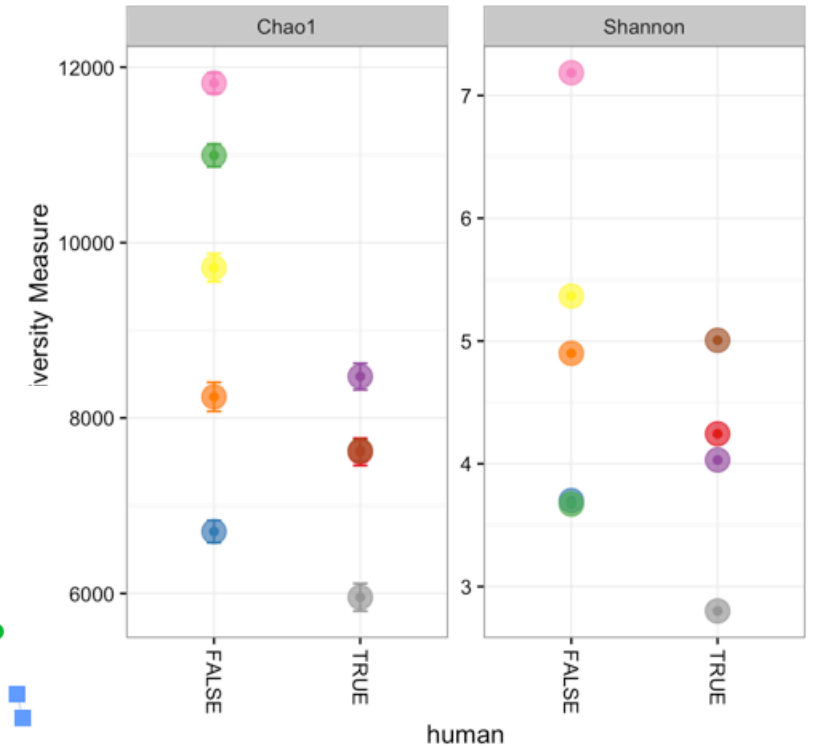
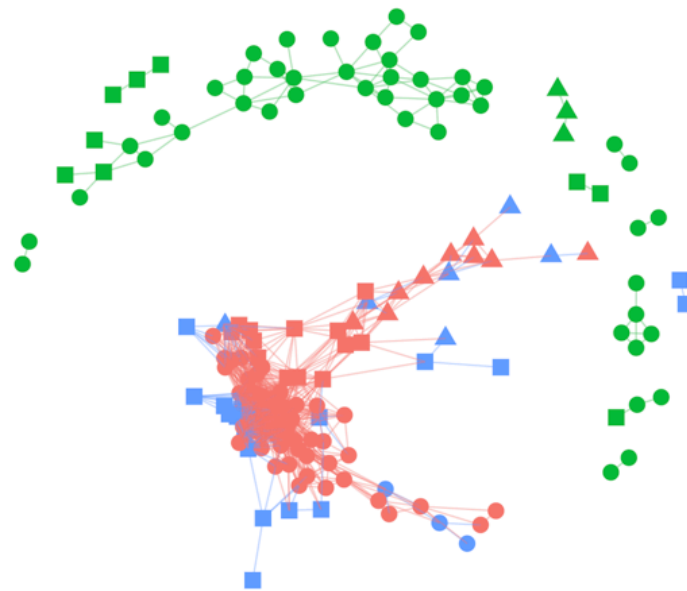
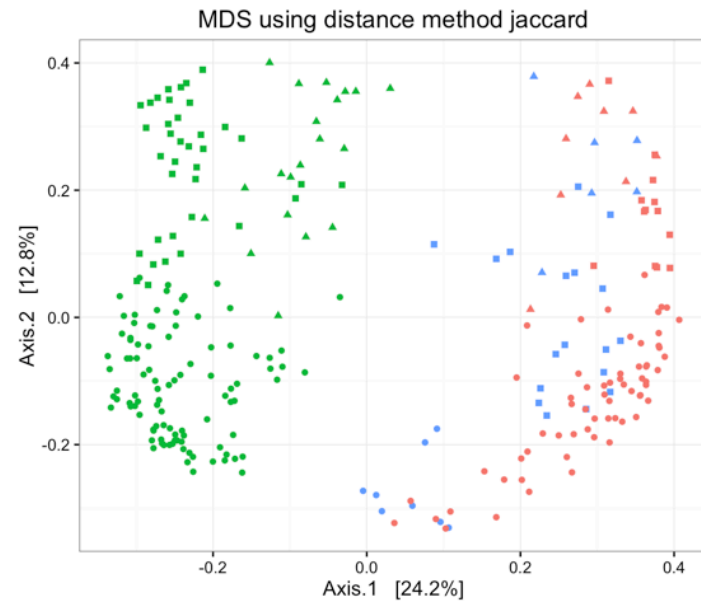
Phyloseq possibilities



Phyloseq possibilities



Phyloseq possibilities



Resources

- [dada2](#)
- [Phyloseq](#)
- [Complete AB sequencing workflow](#)

Lets get our hands diRty!!!