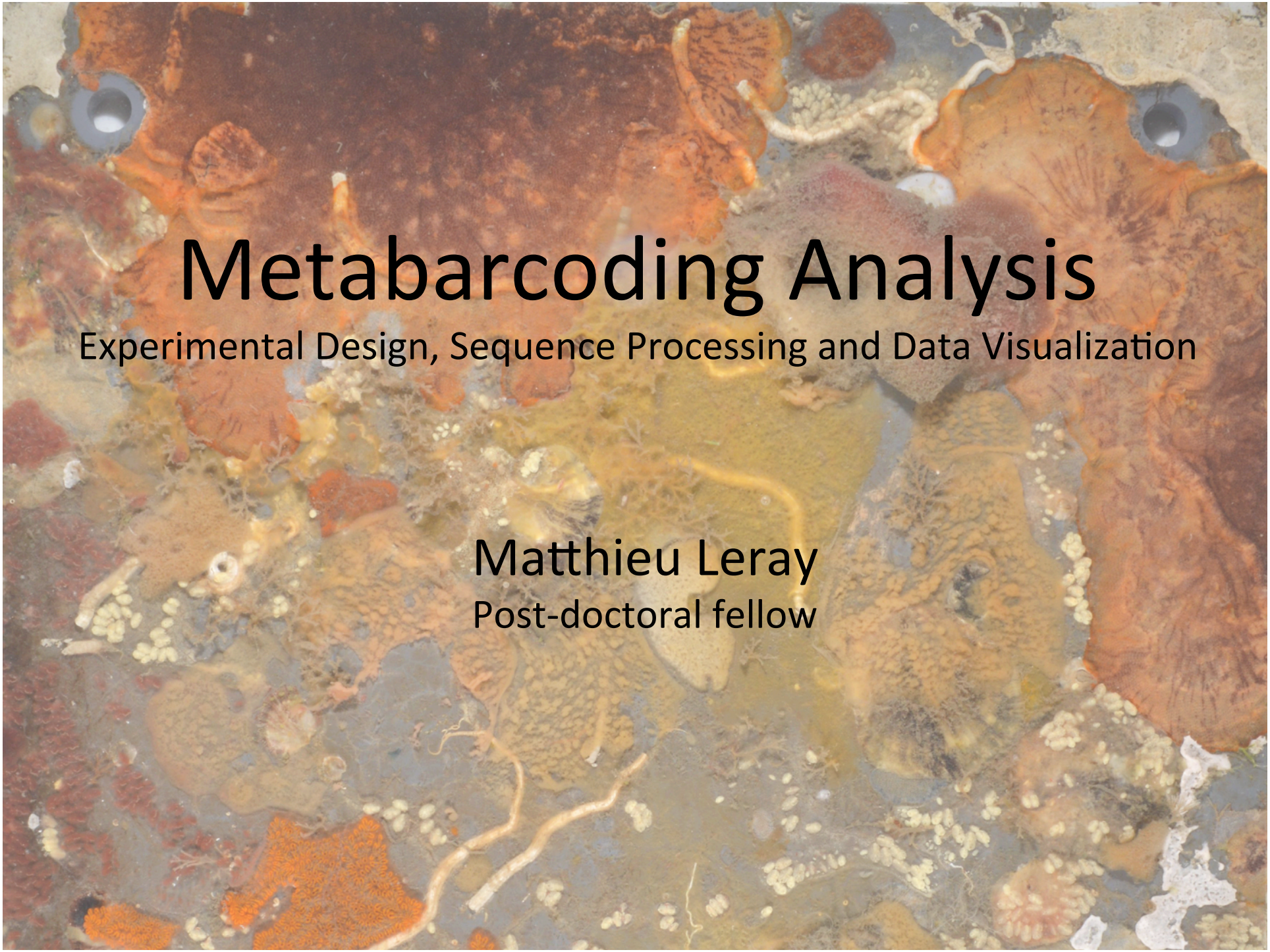


Metabarcoding Analysis

Experimental Design, Sequence Processing and Data Visualization

Matthieu Leray
Post-doctoral fellow

Molecular Approaches for biodiversity studies

DNA barcoding

One organism

at a time
1 sequence

Monocorophium sp.

DNA metabarcoding

Many organisms

Simultaneously
10,000+ sequences

Spionida
Amphipoda
Tubificoides wasselli
Limapontiidae
Tubificidae
…

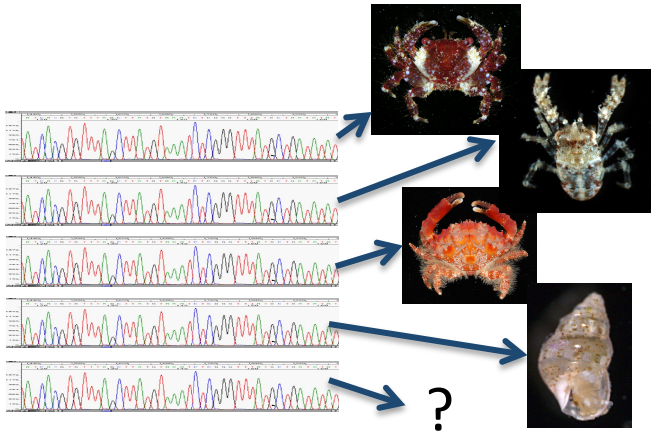
DNA metabarcoding

Community
(environmental sample)

DNA extraction

DNA amplification

High-throughput sequencing



Sequence processing & data analysis
How many taxa? and what are they?

Illumina MiSeq

- Illumina V3 kit sequences 2x300bp paired-end

- Markers of up to 550bp with 25bp overlap

- 25 million reads per run (~12 million contigs)

Number of samples	Number of contigs / sample
5	2,400,000
50	240,000
100	120,000
500	24,000

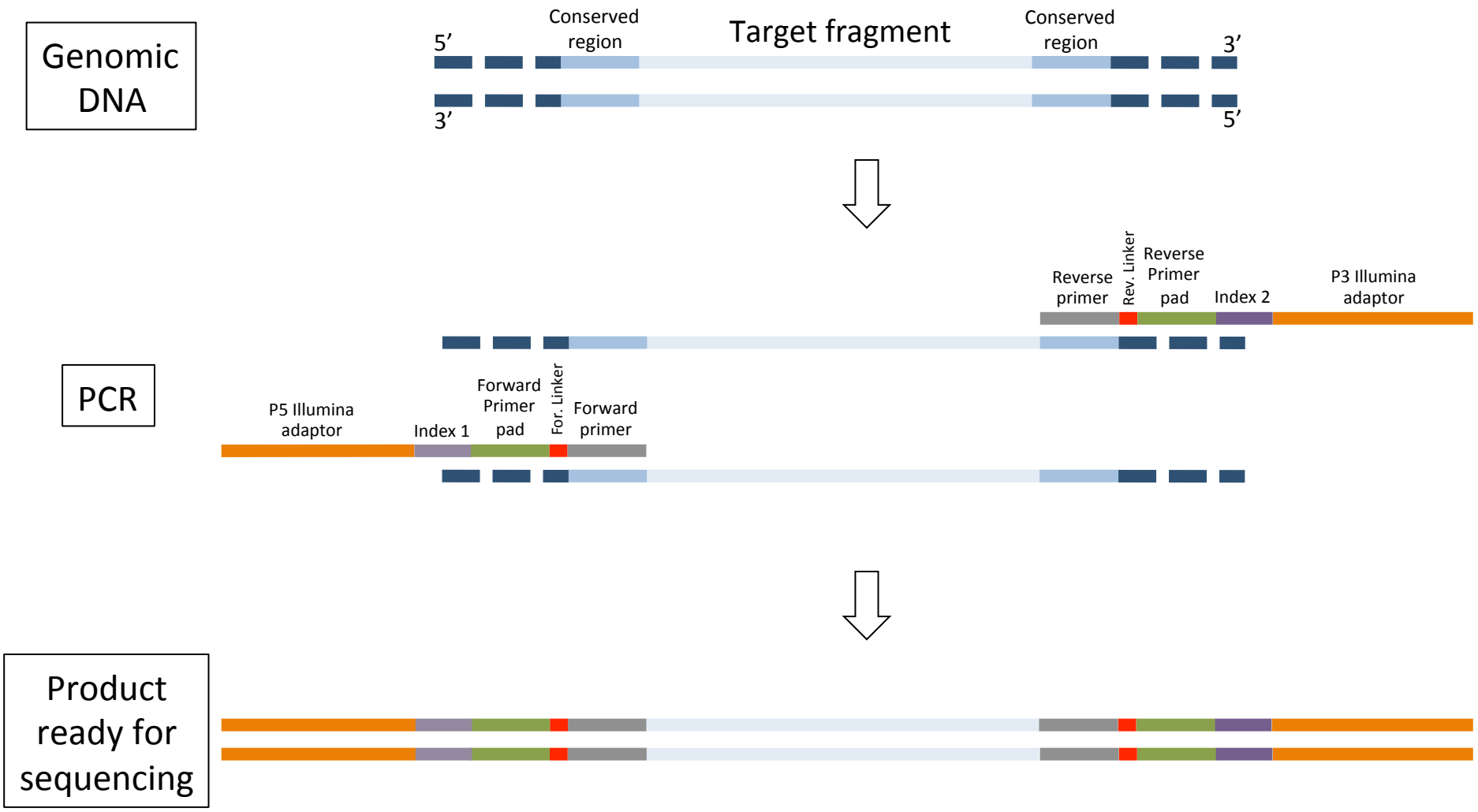
1) Library prep for MiSeq

- Fusion primers
- 2-step PCR (Nextera)
- Ligation of Y-adaptor (TruSeq)
- Ligation of Y-adaptor (TruSeq) w/ tagged primers

2) Sequence processing

3) Data visualization/analysis

Fusion primers



Fadrosh et al. 2014 Microb.
Caporaso et al. 2011 PNAS
Kozich et al. 2013 Appl Environ Microbiol

Fusion primers

2-nt linker sequence
should have minimum homology with potential targets

10-nt pad sequences
- used to increase annealing temperature

pad + linker + primer **>** 60°C

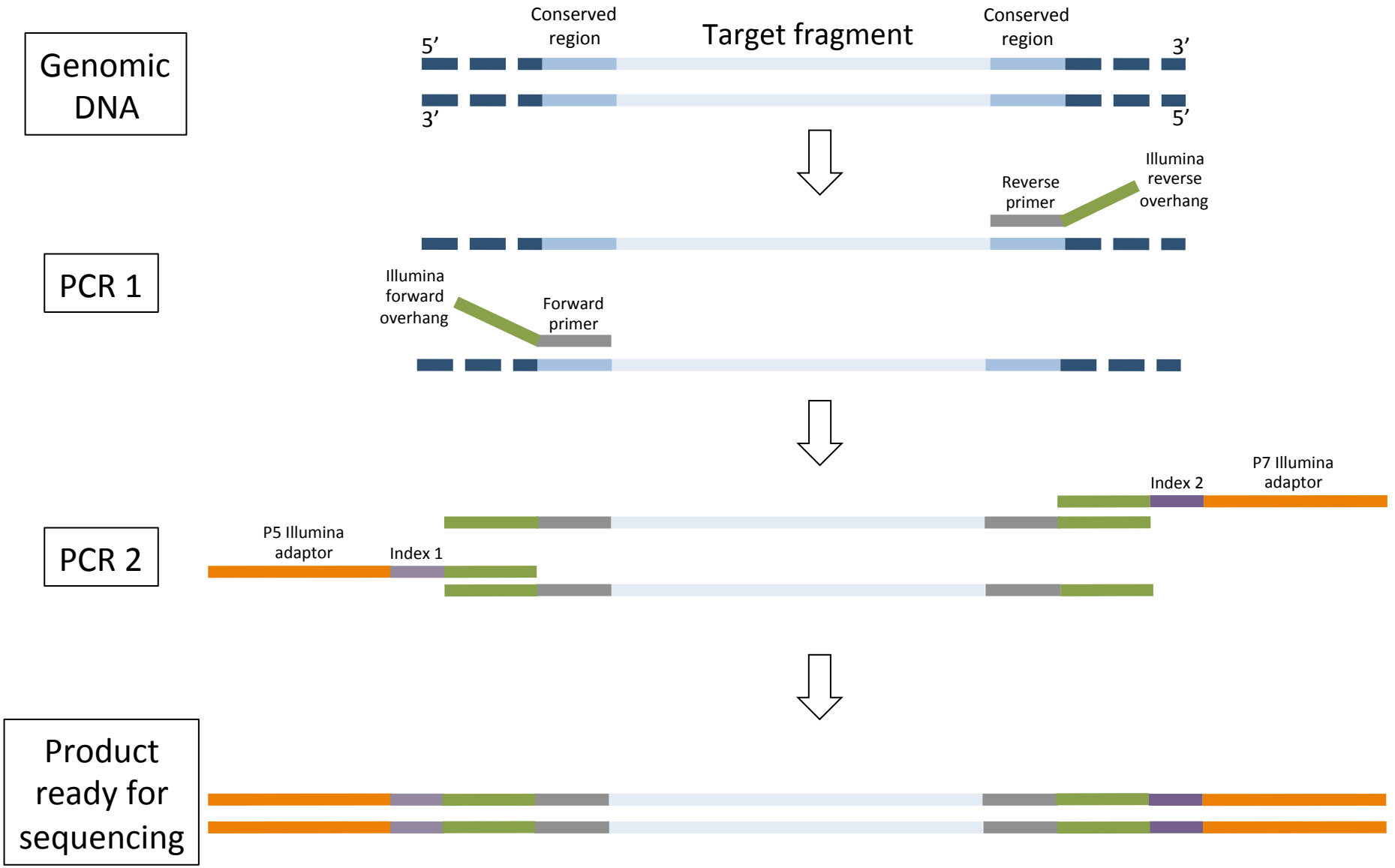
- Carefully designed to avoid hairpins

8-nt barcode sequence
- should differ by at least 2-nt
- should provide an equal intensity into the two light channels used by the sequencer: even representation of all 4 bases at each position

Fusion primers

- To multiplex 9 samples, you need 6 primers
- To multiplex 96 samples, you need 20 primers
- All primers are locus specific

2-step PCR (Nextera)



http://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

2-step PCR (Nextera)

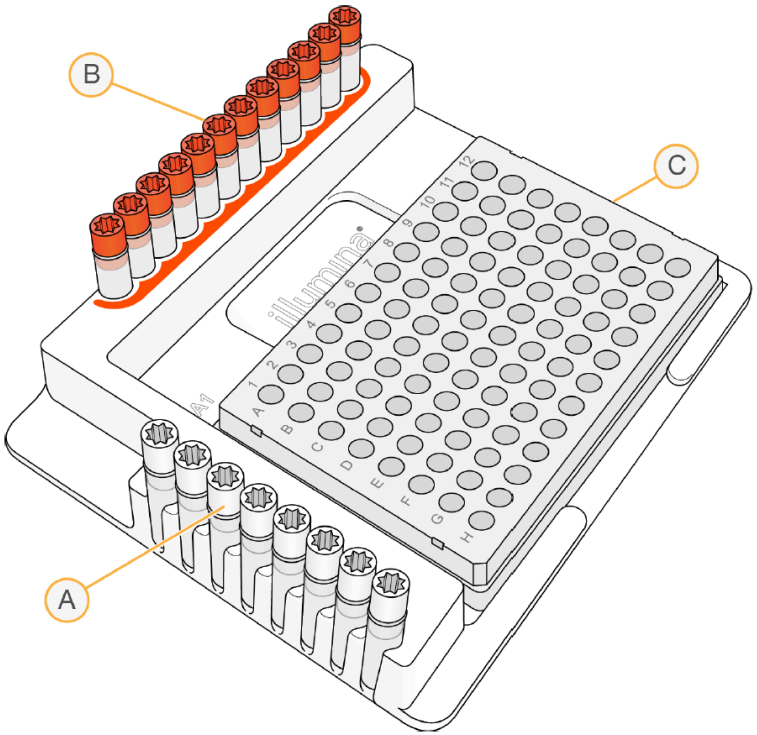
- To multiplex 9 samples, you need 2 primers, 2 P7 and 3 P5 indices

  (provided in Nextera index XT kit - FC-131-1001)

- To multiplex 96 samples, you need 2 primers, 8 P7 and 12 P5 indices

  (provided in Nextera index XT kit - FC-131-1002)

2-step PCR (Nextera)



A Index 2 primers (white caps)
B Index 1 primers (orange caps)
C 96-well plate

- To multiplex 9 samples, you need 2 primers, 2 P7 and 3 P5 indices

  (provided in Nextera index XT kit - FC-131-1001)

- To multiplex 96 samples, you need 2 primers, 8 P7 and 12 P5 indices

  (provided in Nextera index XT kit - FC-131-1002)

Ligation of Y-adaptors (TruSeq)

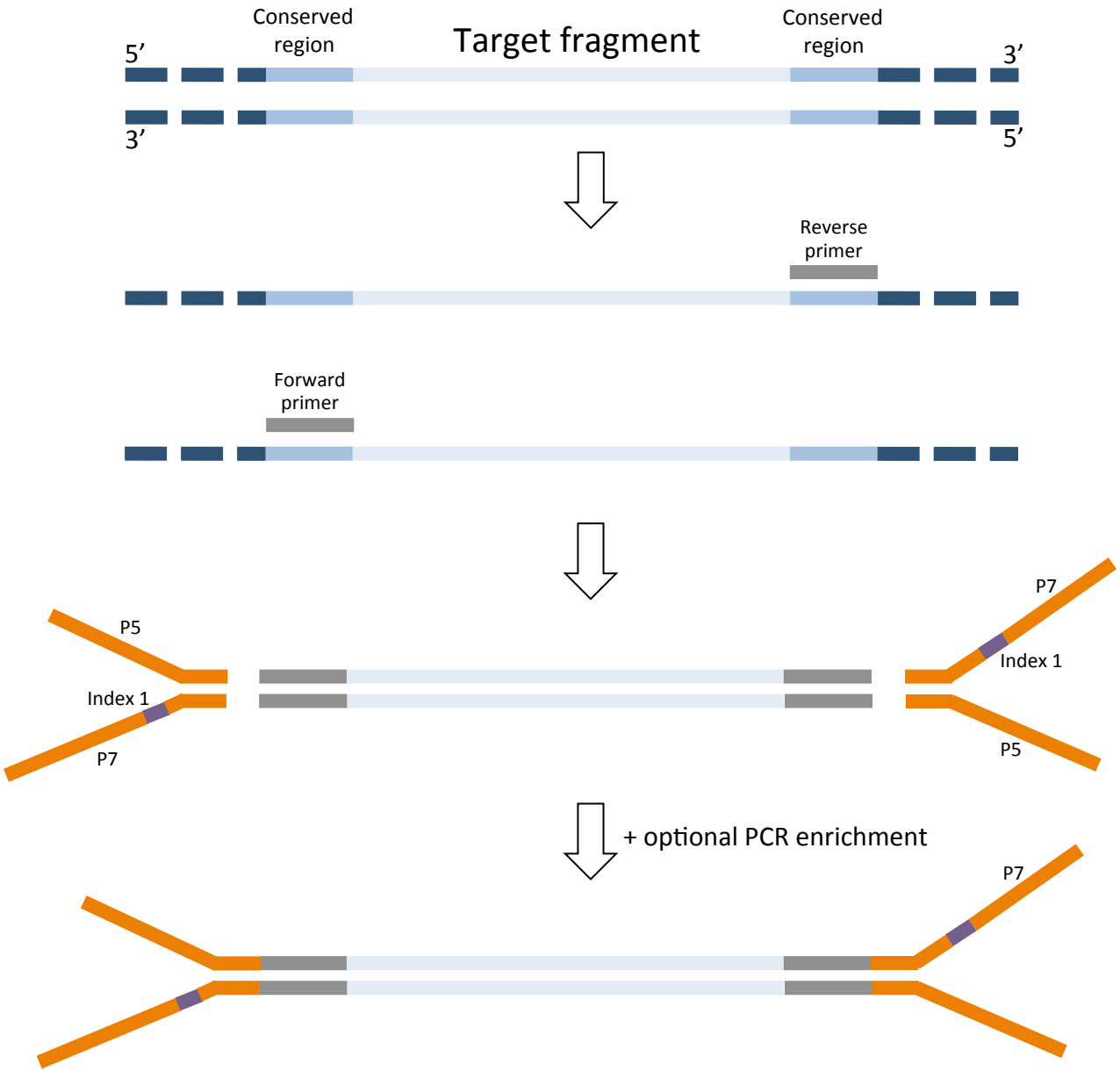
Genomic DNA

PCR

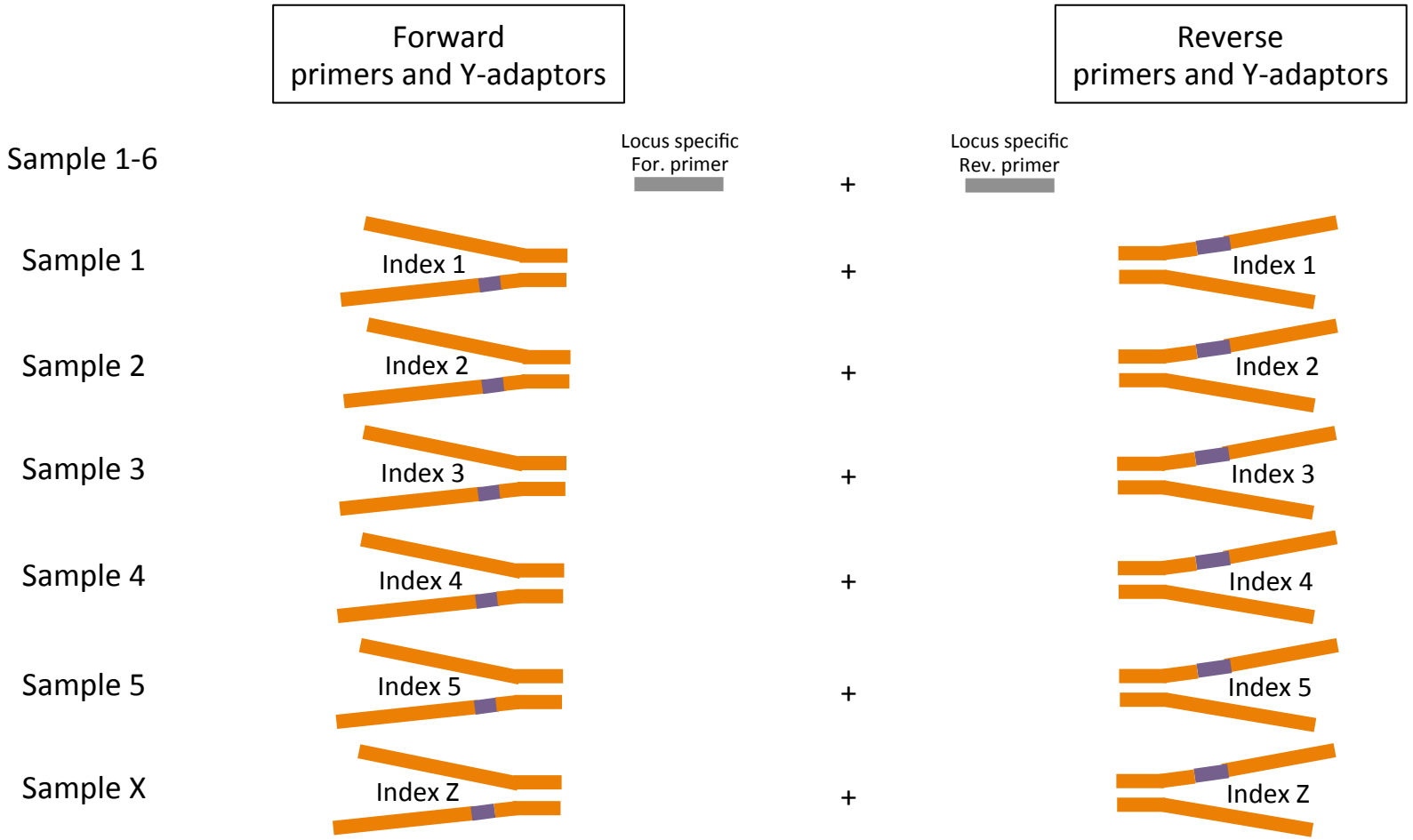
Ligation of Y adaptors

Here single index is illustrated

Product ready for sequencing



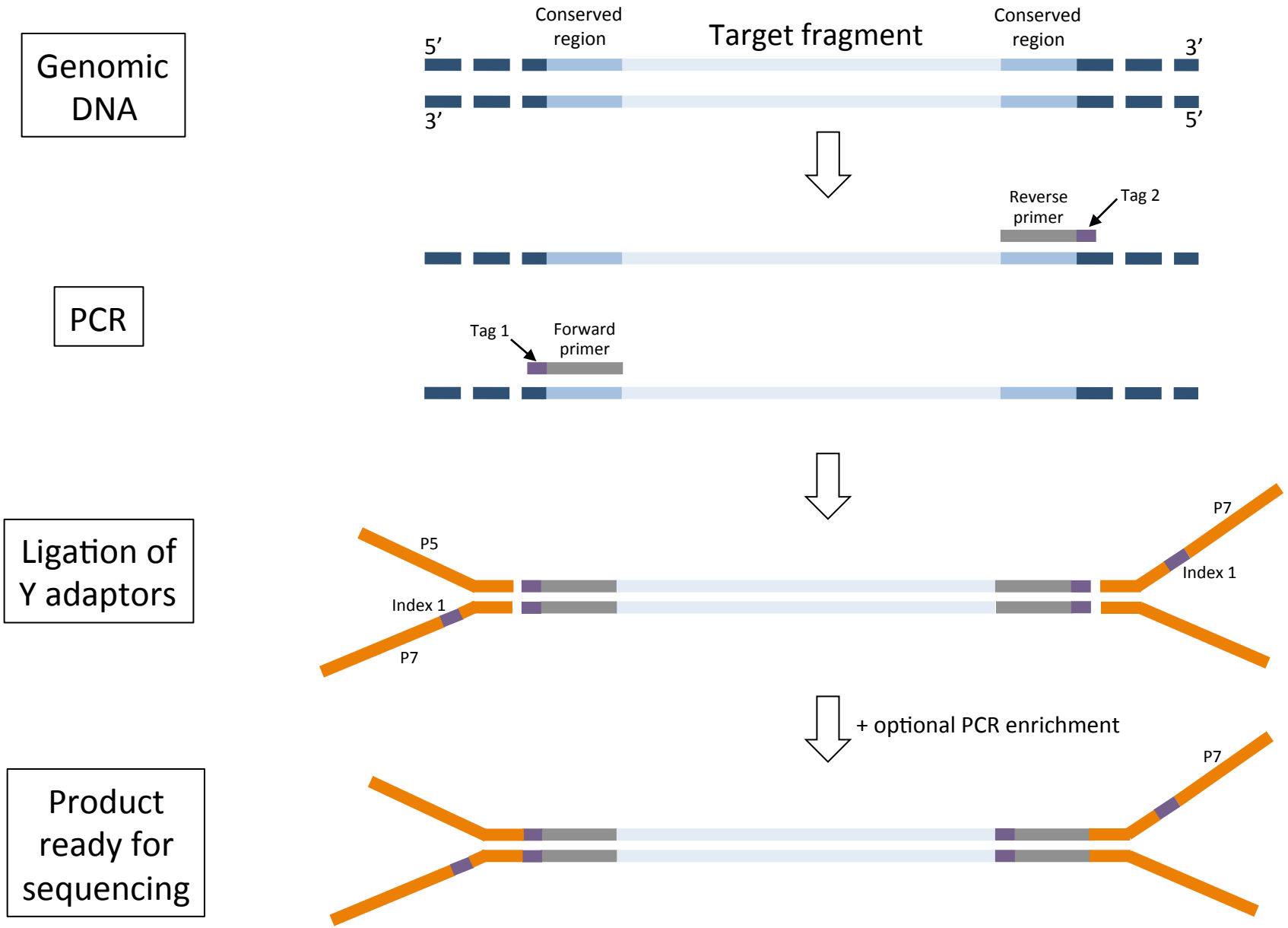
Ligation of Y-adaptors (TruSeq)



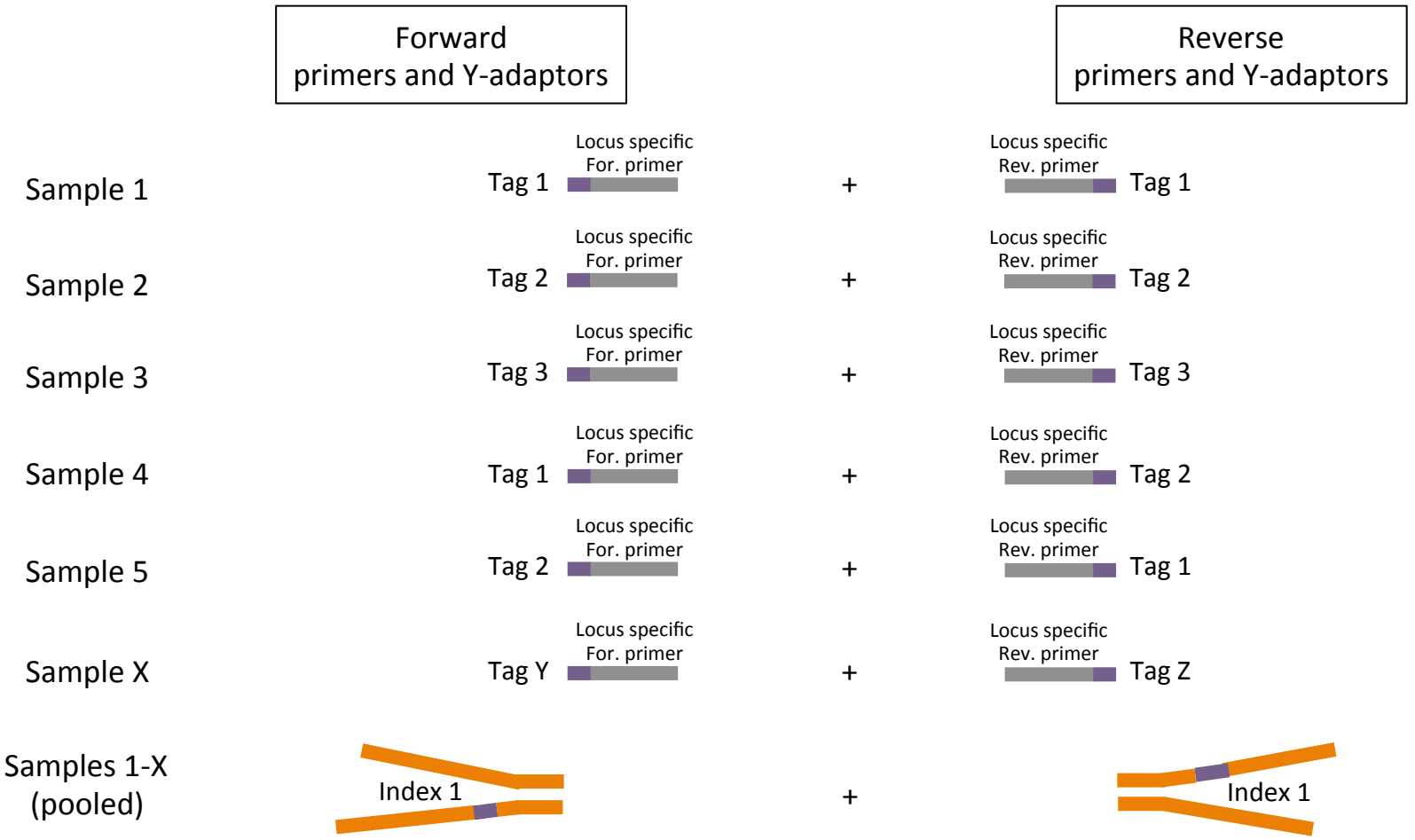
- To multiplex 6 samples, you need 2 primers and 6 unique Y-adaptors

- To multiplex 96 samples, you need 2 primers and 96 unique Y-adaptors (dual index)

Ligation of Y-adaptors (TruSeq) w/ tagged primers



Ligation of Y-adaptors (TruSeq) w/ tagged primers



- To multiplex 9 samples, you need 6 tagged primers and 1 Y-adaptor
- To multiplex 90 samples with 6 tagged primers, you need 10 unique Y-adaptors

Which library prep?

	Additional amplification bias	Hands-on time	Max number of sample per run
Fusion primers	Very likely ~70 bp primers	1 hour	No maximum
2-step PCR (Nextera)	Very likely 2 PCRs & long primers	2 hours	96
Ligation of Y-adaptors (TruSeq)	None	4 hours	96
Ligation of Y-adaptors (TruSeq) w/ tagged primers	None	4 hours	No maximum

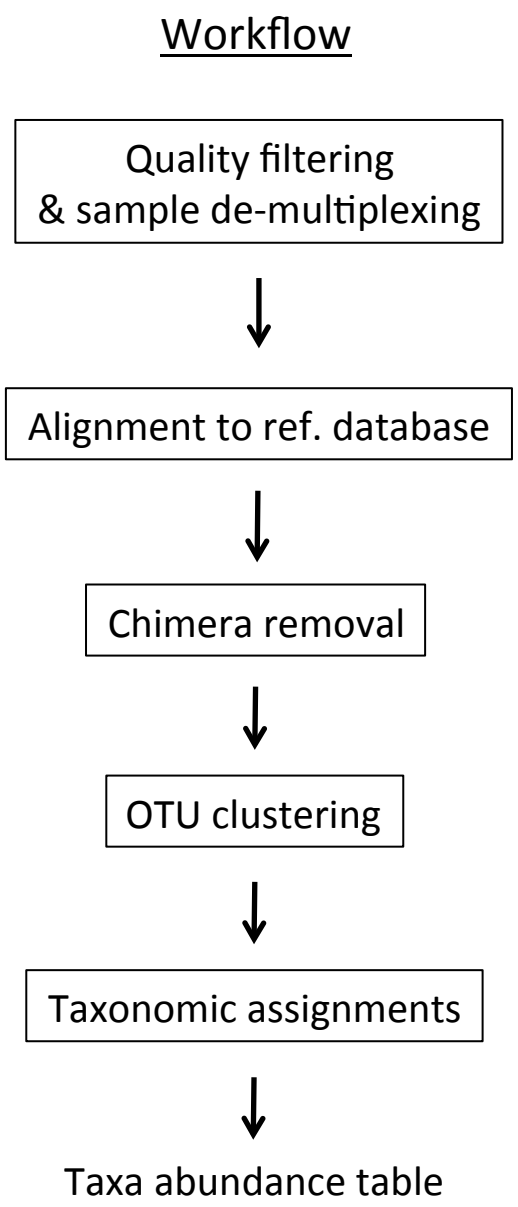
Which library prep?

	Library Prep US$ / sample	Library Prep + Sequencing* US$ / 96 samples
Fusion primers	$35 for primer set (2x70nt - 25 nmol) $6.5 for PCR + clean-up = **$41.5**	Library prep : $1124 Sequencing* : $1485 **= $2609**
2-step PCR (Nextera)	$30 for primer set (2x60nt - 25 nmol) $5.2 for Nextera indices $10 for PCR + clean-up = **$45.2**	Library prep : $290 Sequencing* : $1485 **= $1775**
Ligation of Y-adaptors (TruSeq)	$12.5 for primer set (2x25nt - 25 nmol) $6.5 for PCR + clean-up $15 for ligation (TruSeq LT kit) = **$34**	Library prep** : $2880 Sequencing* : $1485 **= $4265**
Ligation of Y-adaptors (TruSeq) w/ tagged primers	$12.5 for primer set (2x25nt - 25 nmol) $6.5 for PCR + clean-up $15 for ligation = **$34**	Library prep*** : $826 Sequencing* : $1485 **= $2311**

* MiSeq v3 kit 600-cycle
** TruSeq HT kit (dual index)
*** 6 tagged primers & 11 Y-adapters TruSeq LT kit (single index)

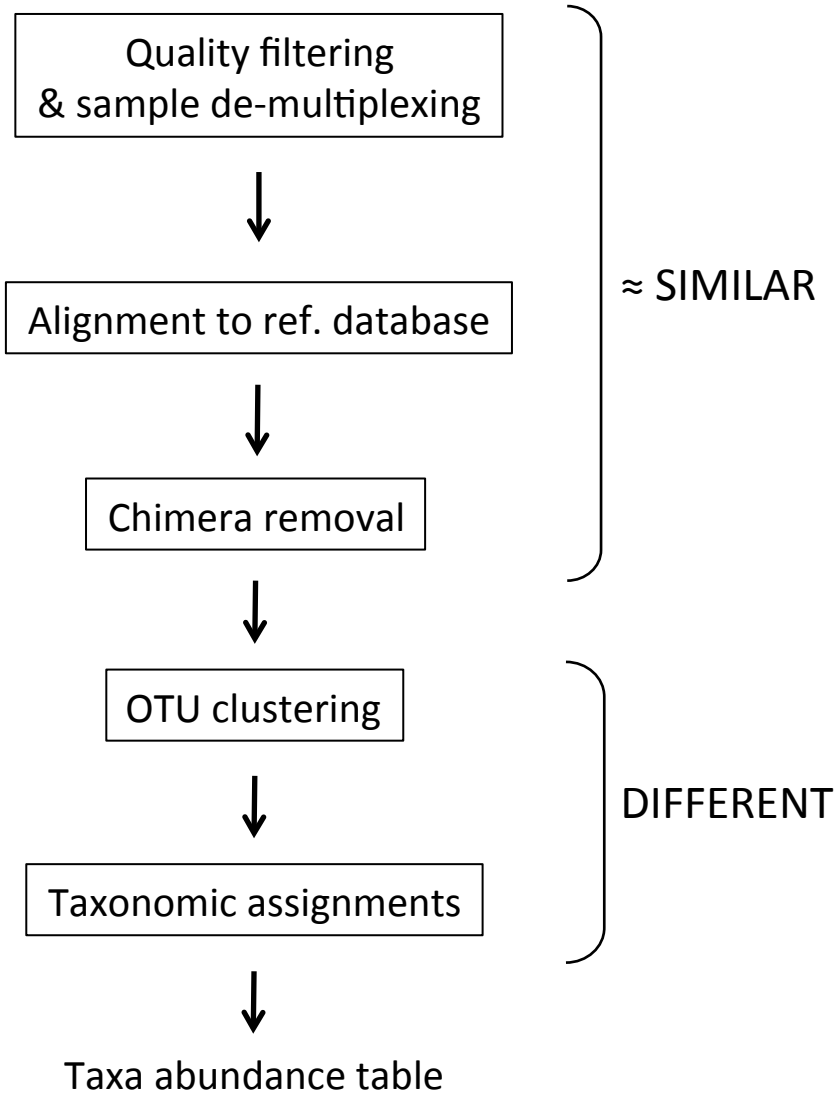
PCR w/ proofreading taq: $2/PCR
PCR primers: $0.25/nt
Ligation Truseq Y-adaptors (FC-121-3001-3): $15/ligation
Nextera indices (FC-131-1001-2): $2.6/sample
Ampure clean-up: $0.5/clean-up

2) Sequence processing



2) Sequence processing

Workflow



16S & 18S pipelines

Open-source, user-friendly packages

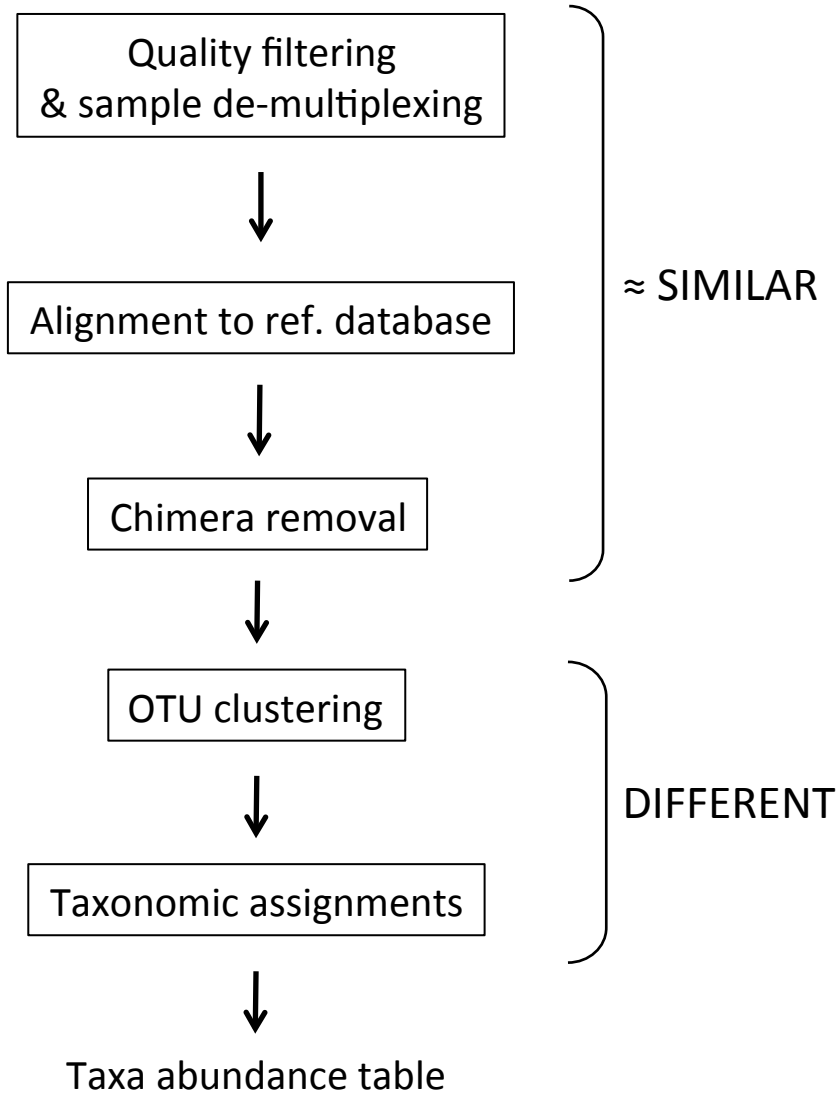
- Mothur http://www.mothur.org
- Qiime http://qiime.org
- CloVR http://clovr.org
- RDP tools https://rdp.cme.msu.edu
- USEARCH http://www.drive5.com/usearch/
- LotuS http://psbweb05.psb.ugent.be/lotus/
- …

16S & 18S pipelines

	OTU clustering algorithm	Taxonomic assignments	Speed
Mothur	- Hierarchical	- Bayesian classifier - kmers	Slow (requires pairwise dist. calculations)
Qiime	- Hierarchical - BLAST - Greedy heuristic - Closed ref. clustering - Open ref. clustering	- Bayesian classifier - kmers - BLAST	Flexible
CloVR	- Hierarchical - BLAST - Greedy heuristic	- Bayesian classifier	Flexible
RDP tools	- Hierarchical	- Bayesian classifier	Slow (requires pairwise dist. calculations)
USEARCH	- Greedy heuristic	- kmers - BLAST	Fast
LotuS	- Greedy heuristic	- Bayesian classifier	Very Fast (30 min on a laptop for a MiSeq dataset)

2) Sequence processing

Workflow



16S & 18S pipelines

Open-source, user-friendly packages

- Mothur http://www.mothur.org
- Qiime http://qiime.org
- CloVR http://clovr.org
- RDP tools https://rdp.cme.msu.edu
- USEARCH http://www.drive5.com/usearch/
- LotuS http://psbweb05.psb.ugent.be/lotus/
- …

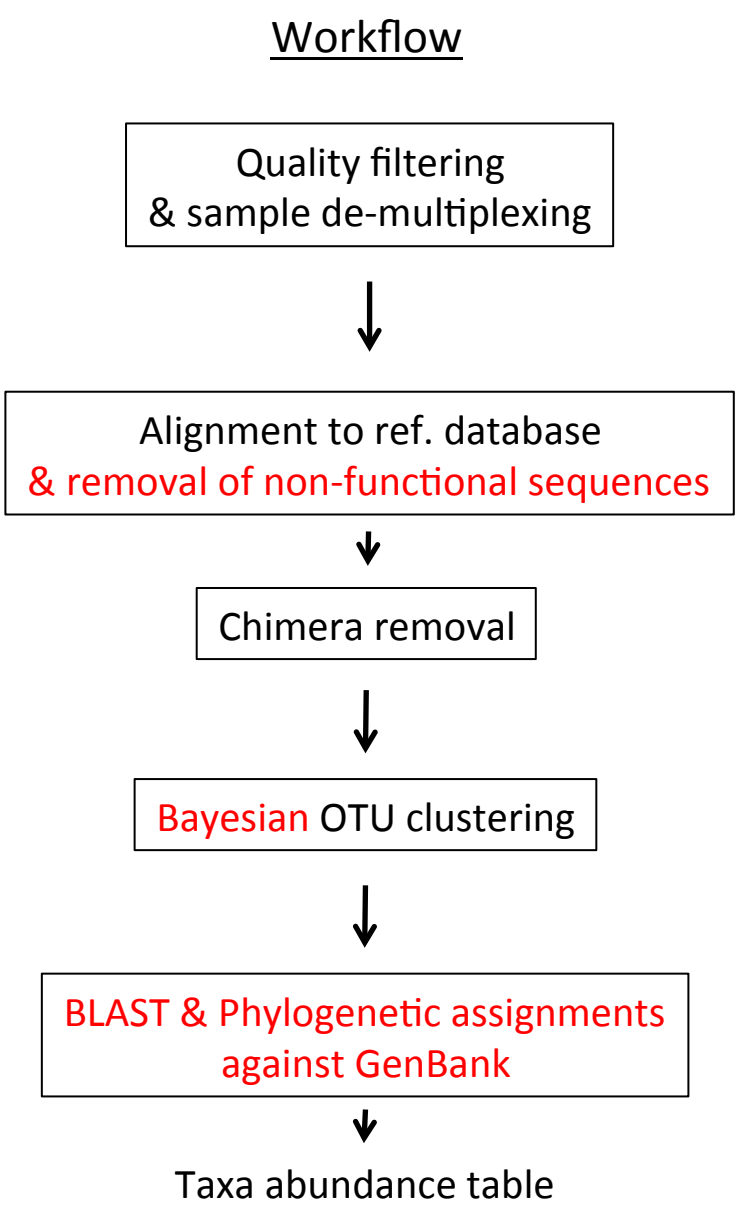
COI & rbcl

- None of the pipelines use amino acid translations for alignment and quality filtering

- All pipelines support only algorithms that use hard-cutoffs for OTU delineation

- None of the pipelines provide curated COI or rbcl datasets for taxonomic assignments

2) Sequence processing



Programs

Mothur (Schloss et al. 2009 Appl Environ Microbiol)

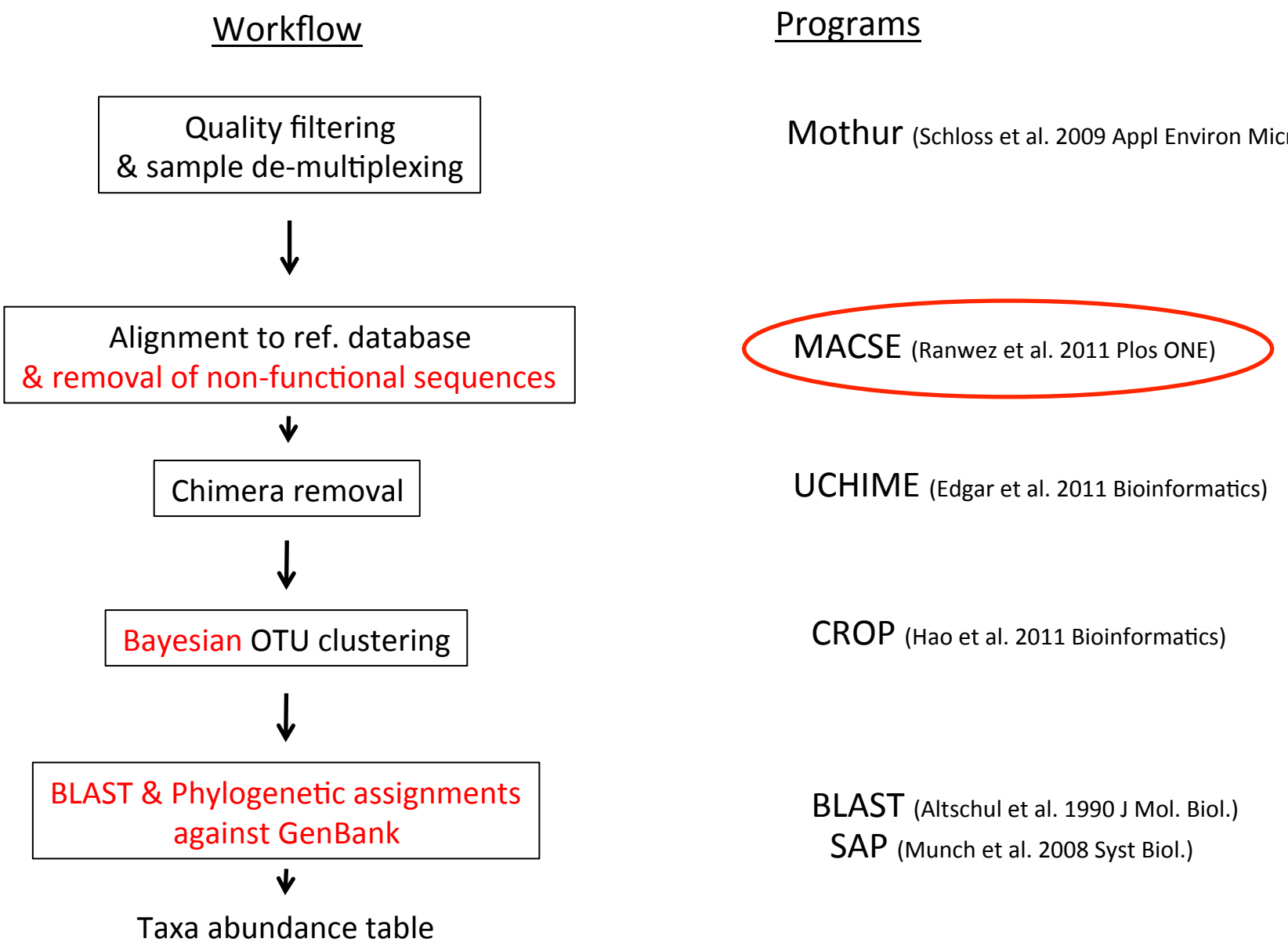
MACSE (Ranwez et al. 2011 Plos ONE)

UCHIME (Edgar et al. 2011 Bioinformatics)

CROP (Hao et al. 2011 Bioinformatics)

BLAST (Altschul et al. 1990 J Mol. Biol.)
SAP (Munch et al. 2008 Syst Biol.)

2) Sequence processing



Multiple Alignment of Coding Sequences (MACSE)

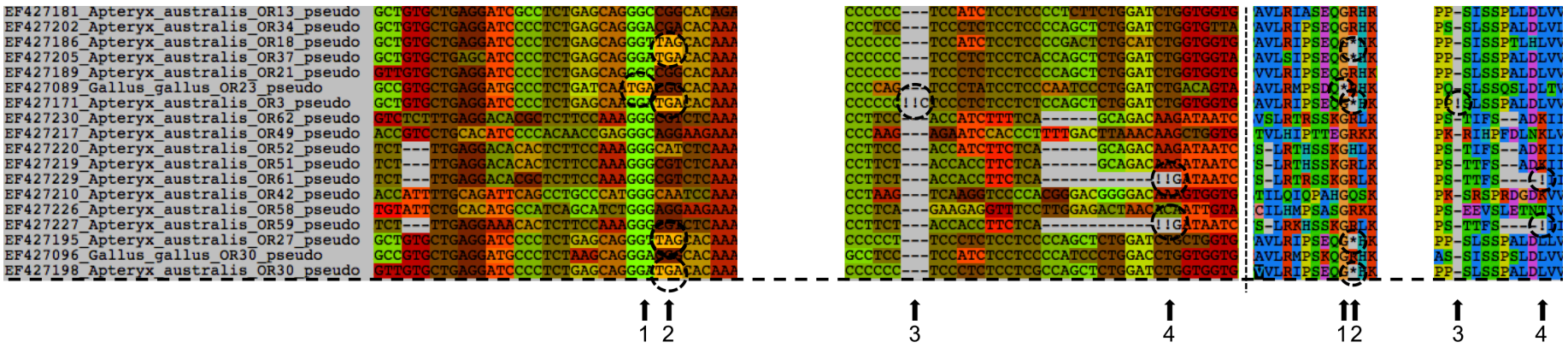
Ranwez et al. 2011 Plos ONE

How it works: aligns sequences one by one to a reference dataset based on amino-acid translations and discards reads that don't meet specified criteria

Command lines:

java -jar macse_v1.0.0i_7.jar **-prog** enrichAlignment **-seq** BIOCODE_MACSE_VR.fasta **-align** BIOCODE_MACSE_VR.fasta **-seq_lr** mls1aa **-maxFS_inSeq** 0 **-maxSTOP_inSeq** 0 **-maxINS_inSeq** 0 -**maxDEL_inSeq** 3 **-gc_def** 5 **-fs_lr** -10 -**stop_lr** -10 **-out_NT** mls1aa_NT **-out_AA** mls1aa_AA -**seqToAdd_logFile** mls1aa_log.csv

java -jar macse_v1.0.0i_7.jar **-prog** exportAlignment **-align** mls1aa_NT **-charForRemainingFS** - -gc_def 5 **-out_AA** mls1aa_AA_macse.fasta **-out_NT** mls1aa_NT_macse.fasta **-statFile** mls1aa.csv



Multiple Alignment of Coding Sequences (MACSE)

Ranwez et al. 2011 Plos ONE

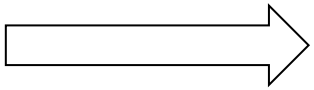
SLOW: Aligns 10,000 sequences in one hour on one laptop…
so it would require 1,000 laptop hours for 10,000,000 reads of a MiSeq run

So we use Hydra and Matt Kweskin's help to make it faster

SCRIPT for batch submission on Hydra
split file

perl split-fasta.pl FILENAME

Login into Hydra and place all files
into destination folder
Submit job
qsub macse-batch.qsub

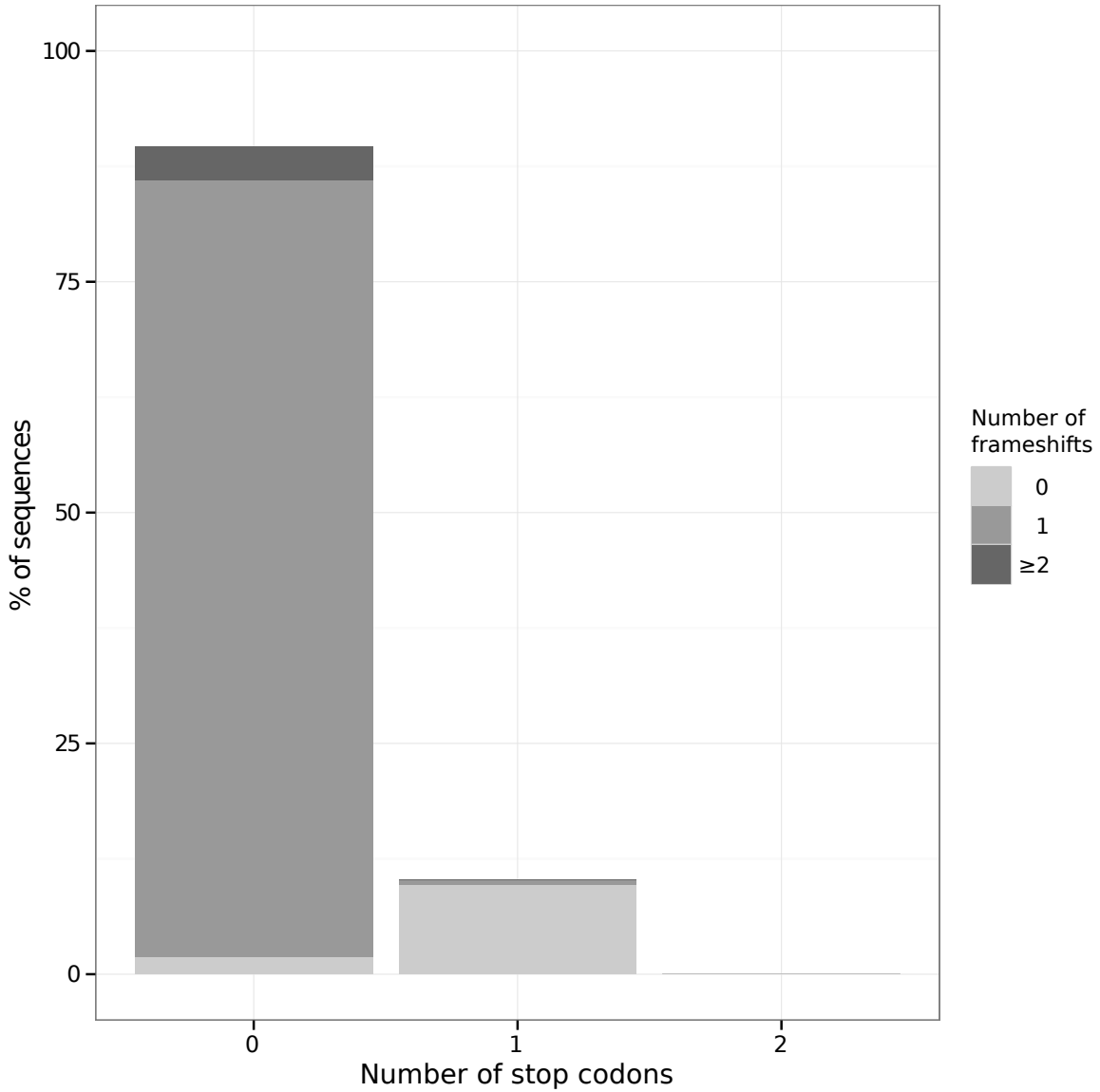


OUTPUTS

- Nucleotide alignments
- Amino acid alignments
- Log files

Multiple Alignment of Coding Sequences (MACSE)

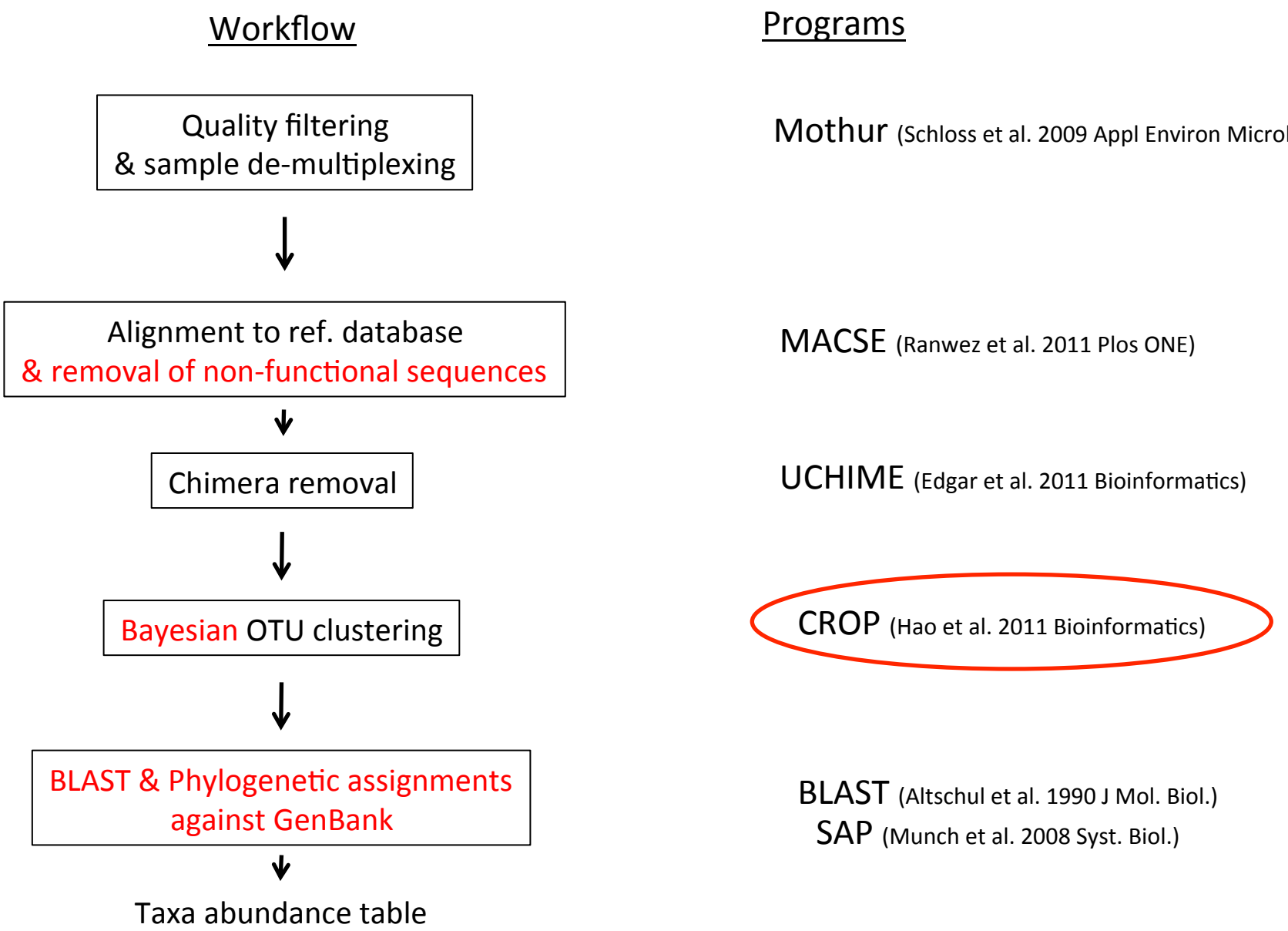
Ranwez et al. 2011 Plos ONE



Example MiSeq run at LAB (COI):

- 87.1% passed
- 12.9% discarded

2) Sequence processing



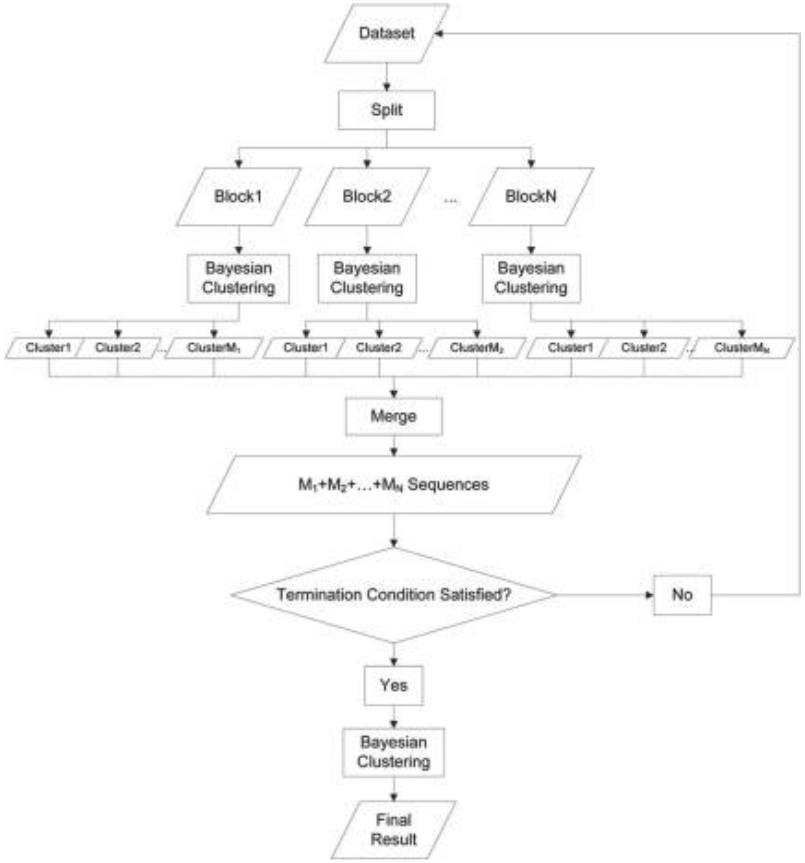
Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering (CROP)

Hao et al. 2011 Bioinformatics

<u>How it works:</u>

clusters based on the natural organization of data without setting a hard cut-off threshold (3%/5%) as required by hierarchical clustering

- Abundant reads carry more weight in the clustering process
- De-duplicates dataset and starts with clustering small blocks of sequences to speed-up the analysis



<u>Command line:</u>

CROP_1_33 CROP **-i** Aligned_file.fasta **-o** Aligned_clustered **-l** 3 **-u** 4 **-b** 500 **-z** 450 (**-r** 0)

-b is the number of blocks in initial round

-r size of the clusters considered “rare”

-l 0.3 **-u** 0.5 ---------- 1%
-l 0.6 **-u** 1.0 ---------- 2%
-l 1.0 **-u** 1.5 ---------- 3%
-l 1.5 **-u** 2.5 ---------- 5%
-l 3.0 **-u** 4.0 ---------- 8%
-l 4.0 **-u** 5.0 ---------- 10%

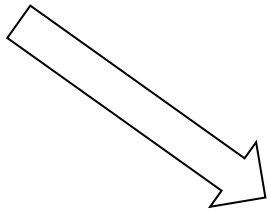
Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering (CROP)

Hao et al. 2011 Bioinformatics

Kind of SLOW: Clusters 10,000,000 sequences in three days on a laptop…
and you need to run it multiple times to choose the best run

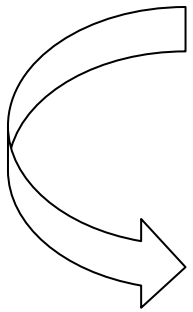
SCRIPT for batch submission on Hydra
Modify name of input file in CROP.job

chmod a+x start-CROP-batch.sh
./start-CROP-batch.sh



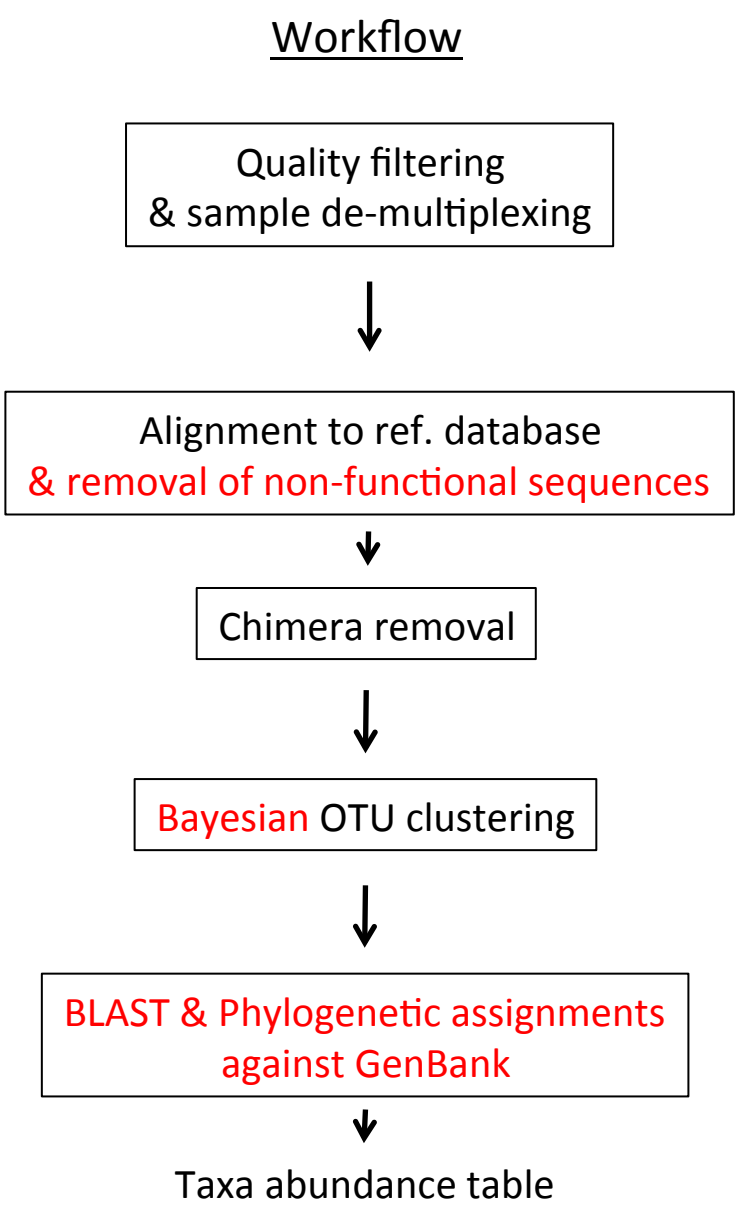
OUTPUTS

- Fasta file with one representative seq. per OTU
- List file containing the label of each sequence clustered within each OTU



Used to build OTU table
& make taxonomic assignments

2) Sequence processing



Programs

Mothur (Schloss et al. 2009 Appl Environ Microbiol)

MACSE (Ranwez et al. 2011 Plos ONE)

UCHIME (Edgar et al. 2011 Bioinformatics)

CROP (Hao et al. 2011 Bioinformatics)

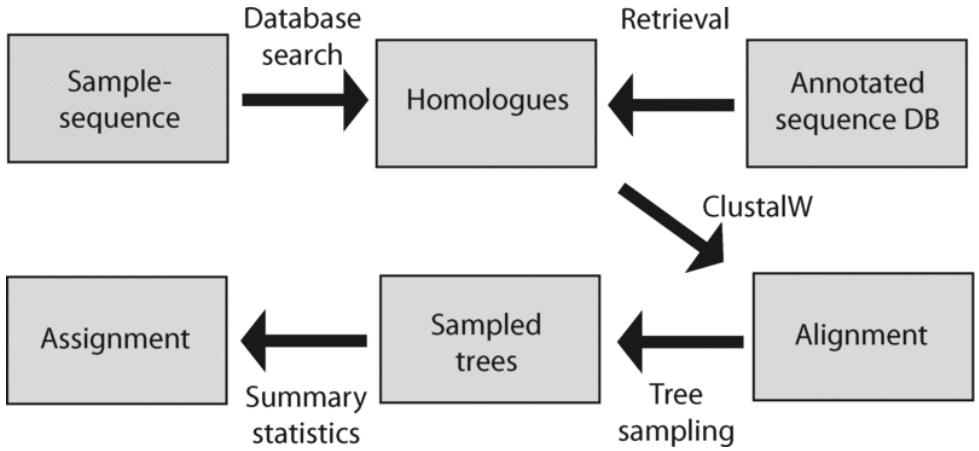
BLAST (Altschul et al. 1990 J Mol. Biol.)

SAP (Munch et al. 2008 Syst Biol.)

Statistical Assignment Package

Munch et al. 2008 Syst. Biol.

How it works: 1) Builds 10,000 unrooted phylogenetic trees from a collection of homologue sequences retrieved from a sequence database (GenBank)
2) Calculates the probability that a query sequence belongs to a monophyletic group within that set of homologues.



Command line:

-x 90 -x 80 --**minidentity** 0.7

OUTPUT

.html file with taxonomic information for each query sequence

Summary

Input query sequences: orignal data: 5
significant number of homologues found for: 3
alignment completed for: 3
mrBayes run completed for: 3
tree-statistics completed for: 3

Input sequences assigned to taxonomic levels: 3

Assignments at 80% level:

phylum		class		order		family		genus		species	
Arthropoda	41_SYFFW_01412_02085: 100%	Maxillopoda	41_SYFFW_01412_02085: 100%	Calanoida	41_SYFFW_01412_02085: 100%	Cyprididae	43_SYFFW_00827_01415: 96%	Eucypris	43_SYFFW_00827_01415: 96%	Eucypris virens	43_SYFFW_00827_01415: 96%
	42_SYFFW_01524_01331: 100%		42_SYFFW_01524_01331: 100%	Podocopida	43_SYFFW_00827_01415: 96%	total	1	total	1	total	1
	43_SYFFW_00827_01415: 97%	Ostracoda	43_SYFFW_00827_01415: 96%	Sessilia	42_SYFFW_01524_01331: 100%	total/all_analyzed	1/3 33.3%	total/all_analyzed	1/3 33.3%	total/all_analyzed	1/3 33.3%
total	3	total	3	total	3	list not assigned		list not assigned		list not assigned	
total/all_analyzed	3/3 100.0%	total/all_analyzed	3/3 100.0%	total/all_analyzed	3/3 100.0%						
list not assigned		list not assigned									

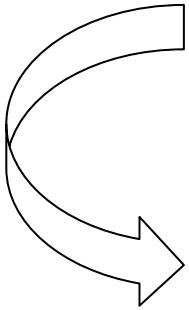
Statistical Assignment Package

Munch et al. 2008 Syst. Biol.

very SLOW: takes one hour to assign each sequence if you query GenBank
but only about 2min/sequence if you query a local curated database

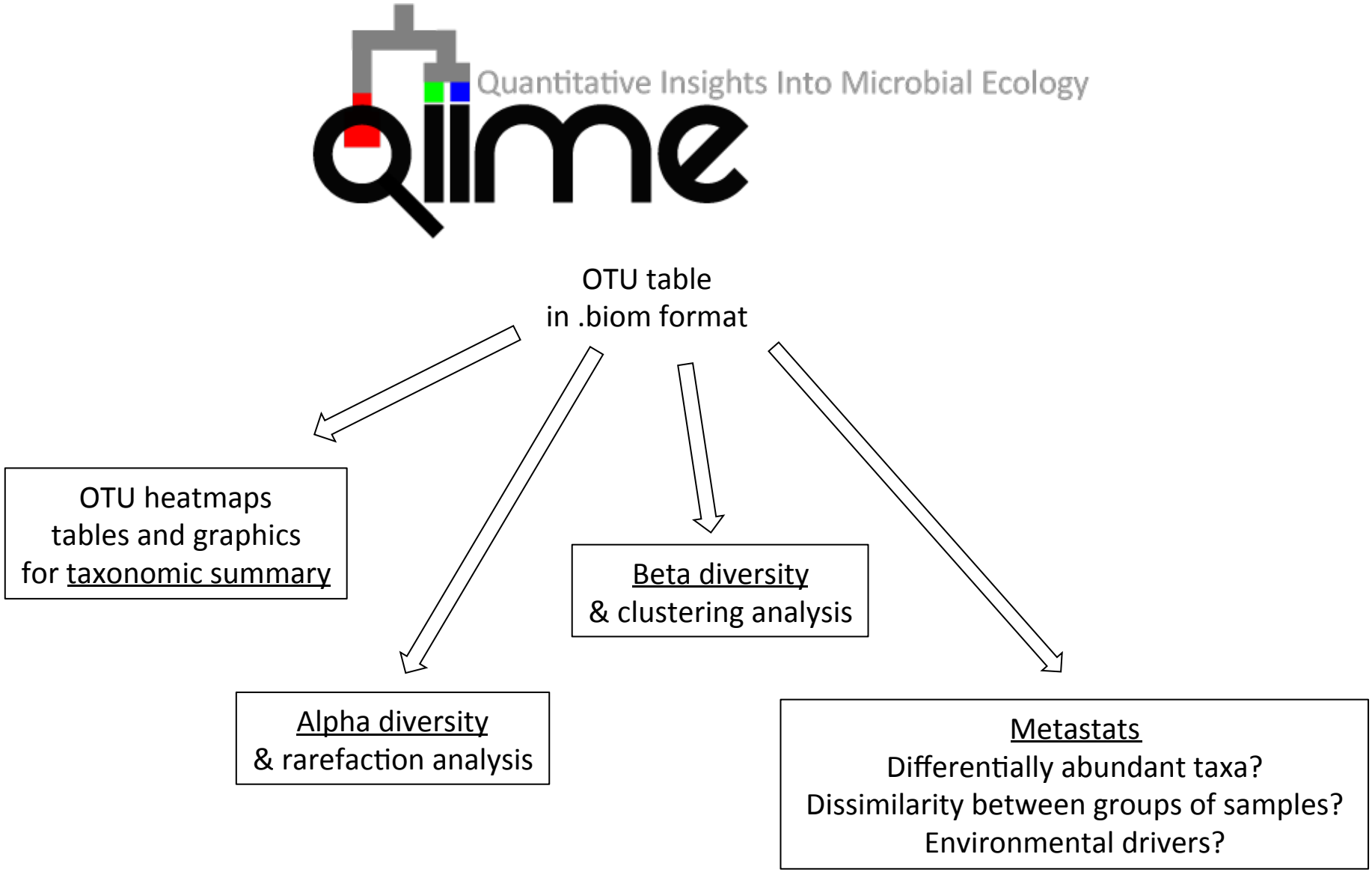
SCRIPT for batch submission on Hydra
it will automatically split the sequence file with query sequences
and run each of the smaller files on Hydra

```
chmod a+x split-and-run.pl
./split-and-run.pl file.fasta
```



Taxonomic information automatically
added to the OTU table

3) Data visualization



Prepare input files for Qiime

1) OTU table

Convert tab-delimited table into a .biom file (http://biom-format.org/documentation/biom_conversion.html)

biom convert **-i** otu_table.taxonomy.txt **-o** otu_table.from_txt.biom **--table-type=**"otu table" **--process-obs-metadata** taxonomy

2) Mapping file

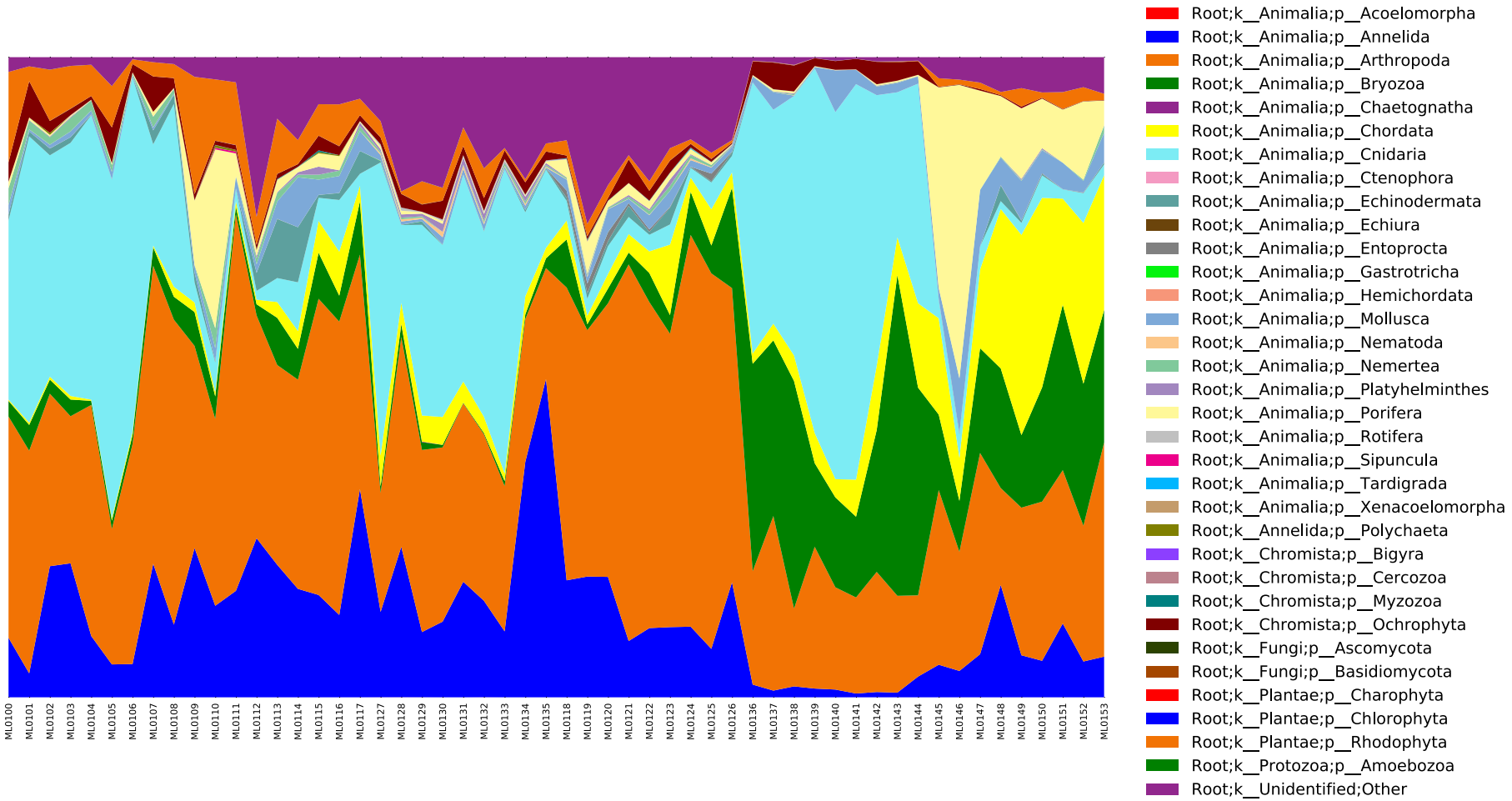
#SampleID	BarcodeSequence	LinkerPrimerSequence	Locality	Site	FractionA	FractionB	Description
ML.0100	WCH A1 500um	WCH_500um	WCH_500umA1	NA	NA		
ML.0101	WCH A1 500um	WCH_500um	WCH_500umA1	NA	NA		
ML.0102	WCH A1 500um	WCH_500um	WCH_500umA1	NA	NA		
ML.0103	WCH A3 500um	WCH_500um	WCH_500umA3	NA	NA		
ML.0104	WCH A3 500um	WCH_500um	WCH_500umA3	NA	NA		
ML.0105	WCH A3 500um	WCH_500um	WCH_500umA3	NA	NA		
ML.0106	WCH A2 500um	WCH_500um	WCH_500umA2	NA	NA		
ML.0107	WCH A2 500um	WCH_500um	WCH_500umA2	NA	NA		
ML.0108	WCH A2 500um	WCH_500um	WCH_500umA2	NA	NA		
ML.0109	FTP B1 500um	FTP_500um	FTP_500umB1	NA	NA		
ML.0110	FTP B1 500um	FTP_500um	FTP_500umB1	NA	NA		

Because we input data into Qiime for downstream analysis only,
no need to specify “BarcodeSequence” and “linkerPrimerSequence”

3) Phylogenetic tree (optional)

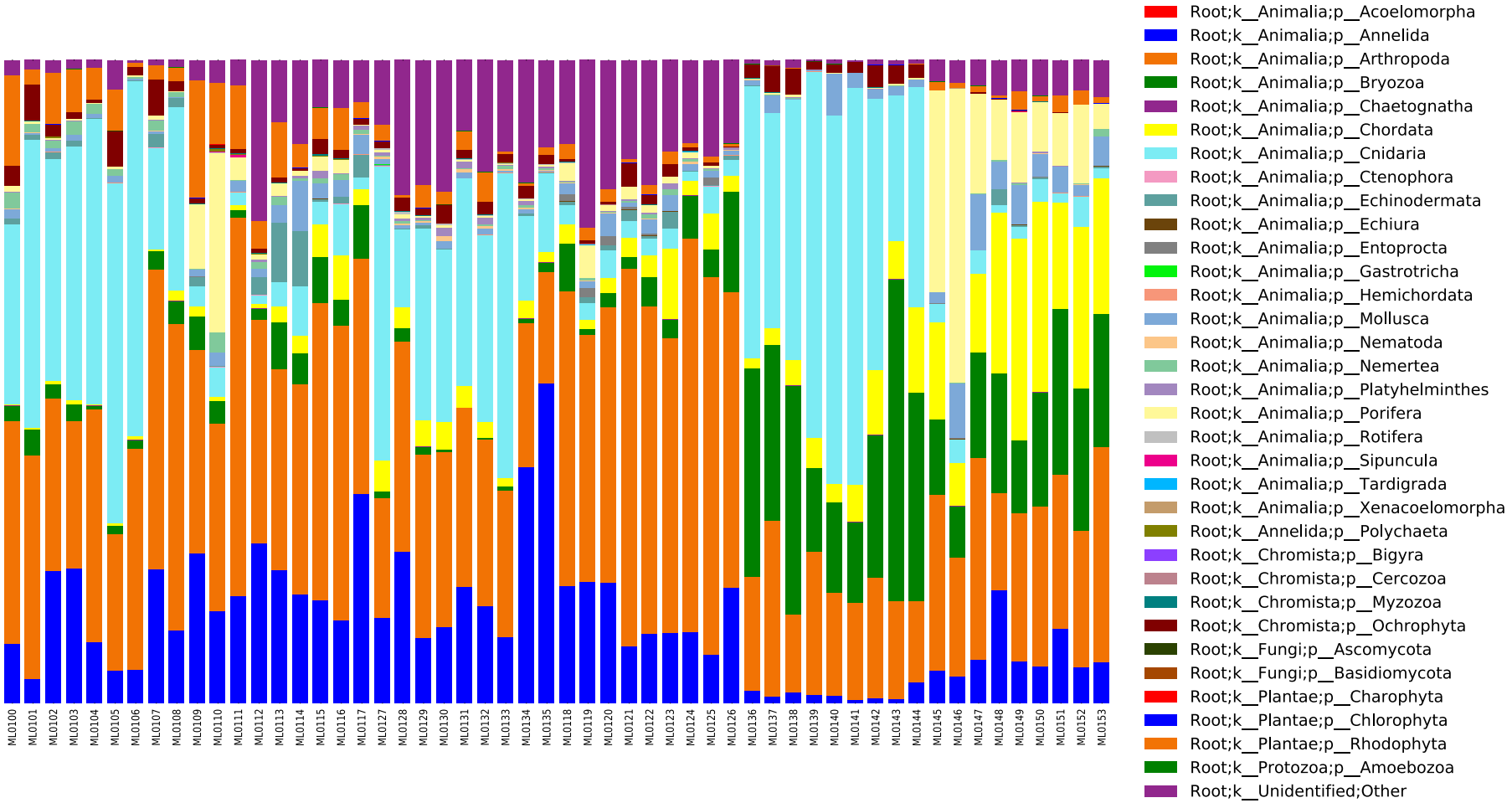
Taxonomic summary

summarize_taxa_through_plots.py -**i** otus/otu_table.biom -**o** taxa_summary
-**m** Fasting_Map.txt



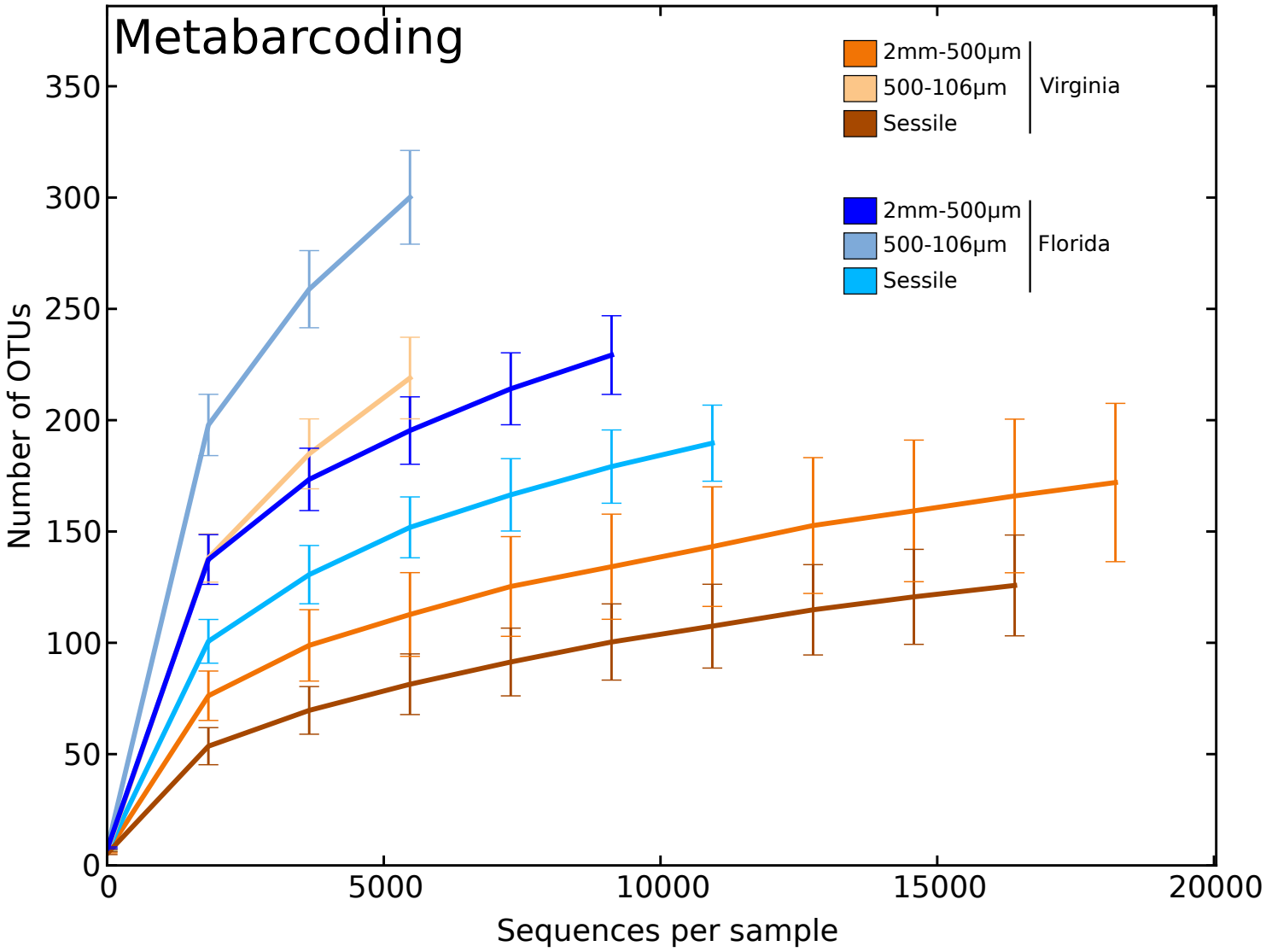
Taxonomic summary

summarize_taxa_through_plots.py -**i** otus/otu_table.biom -**o** taxa_summary
-**m** Fasting_Map.txt



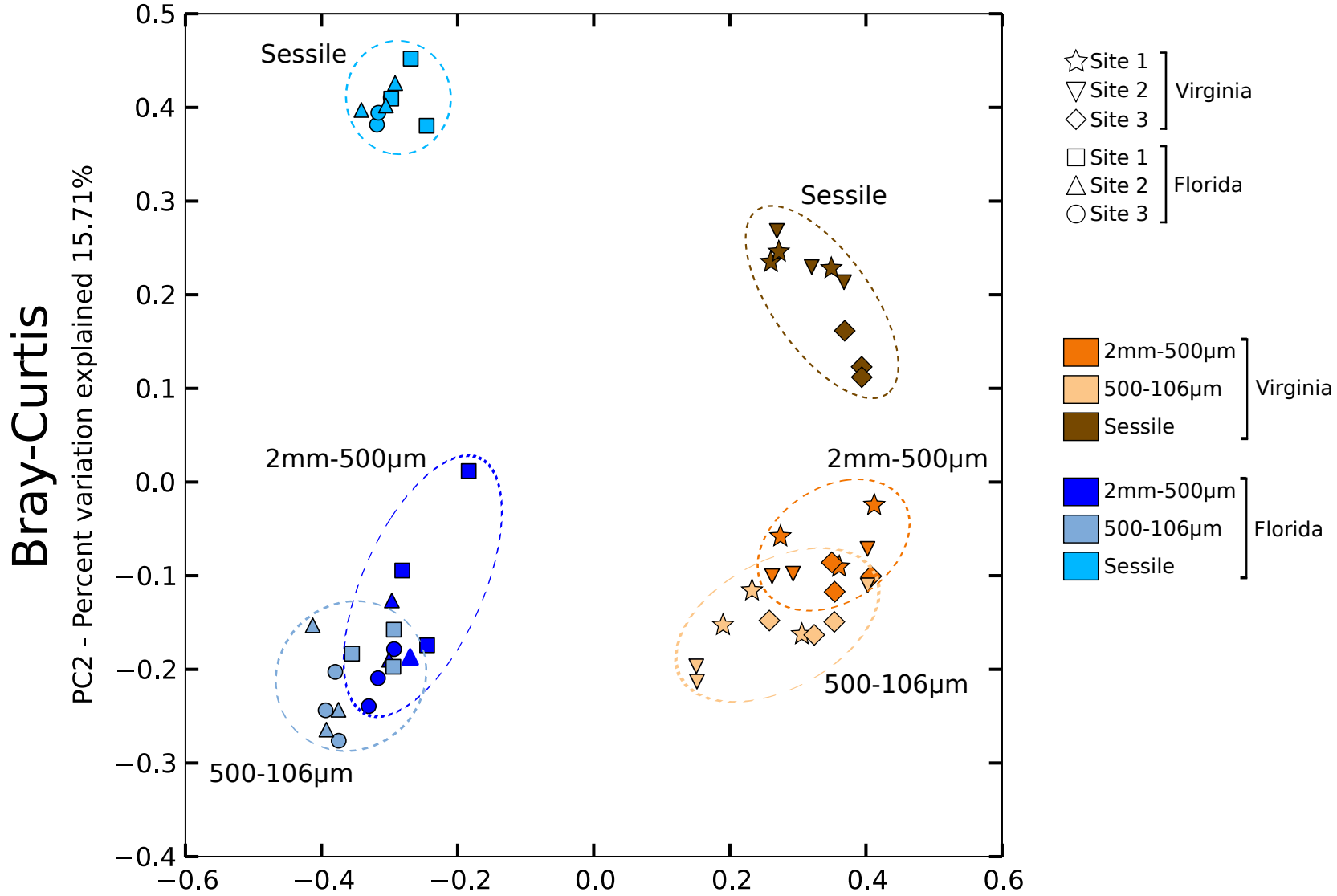
Alpha diversity

alpha_rarefaction.py -**i** otus/otu_table.biom -**m** Fasting_Map.txt -**o** arare -**p** alpha_params.txt



Beta diversity

make_2d_plots.py - **i** bray_curtis_pc.txt -**m** Fasting_Map.txt -**p** beta_params.txt



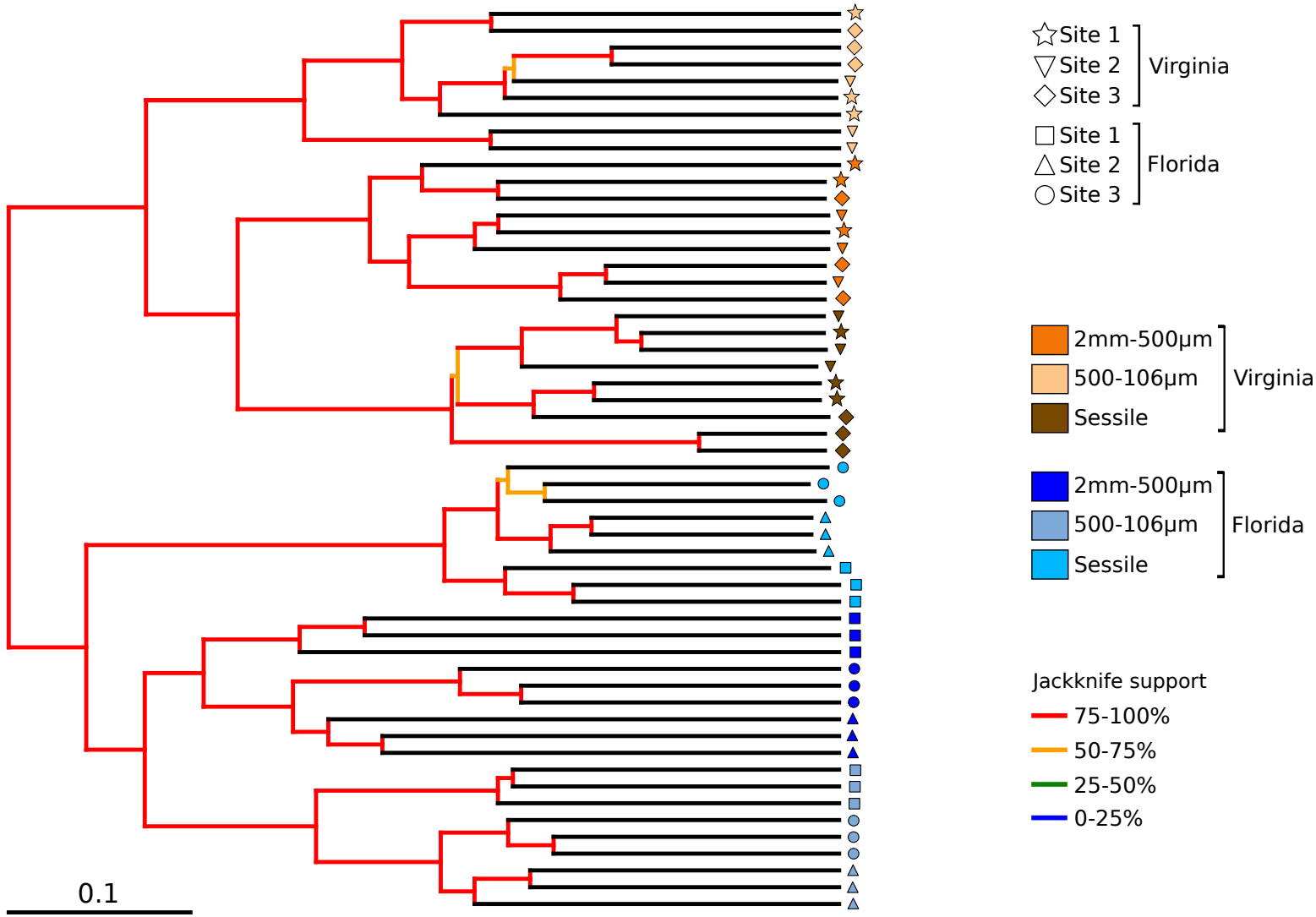
Beta diversity

beta_diversity_through_plots.py **-i** otus/otu_table.biom **-m** Fasting_Map.txt **-o** bdiv_even6114 **-e** 6114

.html file

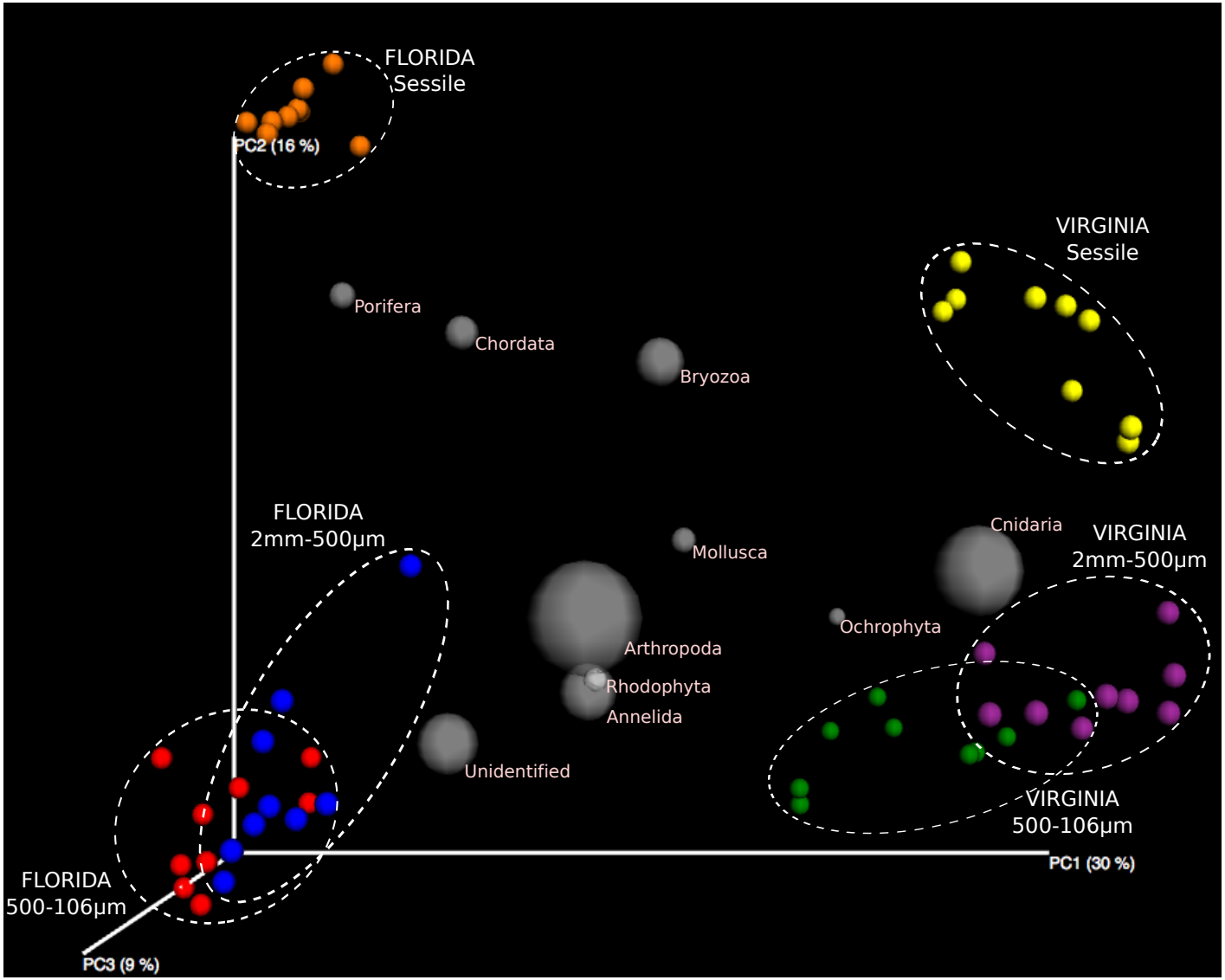
Beta diversity

make_bootstrapped_tree.py -**m** jack/bray_curtis/upgma_cmp/master_tree.tre -**s** jack/bray_curtis/upgma_cmp/jackknife_support.txt -**o** jack/bray_curtis/upgma_cmp/jackknife_named_nodes.pdf



Beta diversity

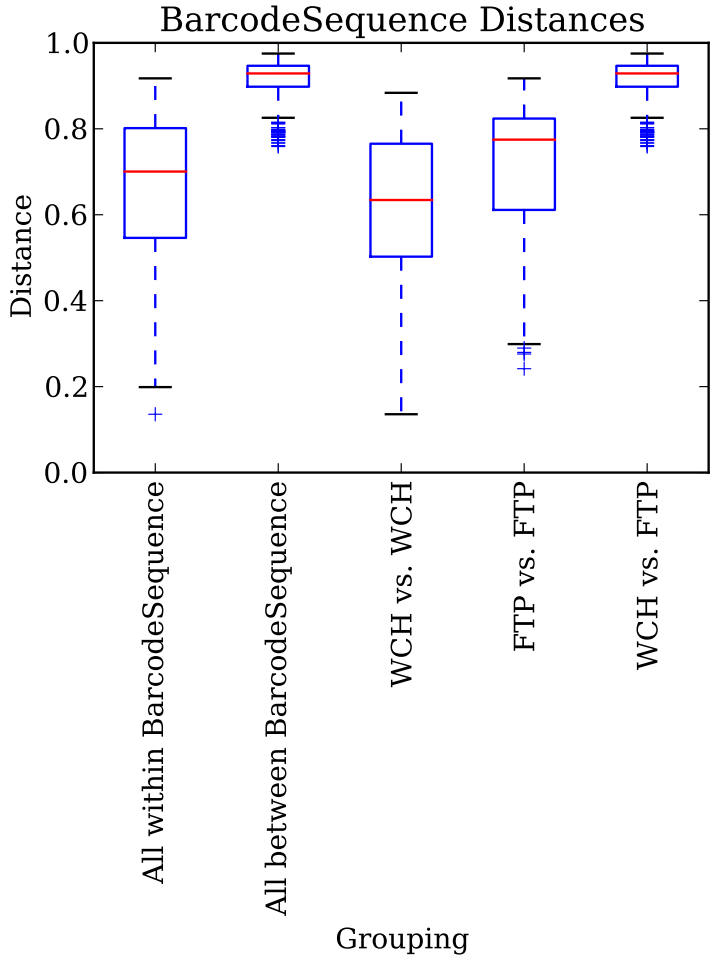
make_emperor.py **-i** bdiv_even6114/bray_curtis_pc.txt **-m** Fasting_Map.txt **--n_taxa_to_keep** 5 **-o** biplots



Metastats

- Plot within and between distances between treatments

```
make_distance_boxplots.py -m Fasting_Map.txt
-d bdiv_even6114/bray_curtis_dm.txt -f
Treatment -o tutorial_output
```



- Test for differences in community composition between groups of samples

```
compare_categories.py --method adonis -i
bray_curtis_dm.txt -m Fasting_Map.txt -c Site -
o adonis_out -n 999
```

http://qiime.org

Questions?