UCDAVIS Bioinformatics Core

Microbial Community Profiling

Matt Settles, PhD
University of California, Davis
settles@ucdavis.edu

Bioinformatics Core

Genome Center

UC Davis

bioinformatics.ucdavis.edu

Disclaimer

- This talk/workshop is full of opinion, there are as many different way to perform analysis as there are Bioinformaticians.
- My opinion is based on over a decade of experience and spending a considerable amount of time to understand the data and how it relates to the biological question.
- Each experiment is unique, this workshop is a starting place and should be adapted to the specific characteristics of your experiment.

Outline

This workshop is intended for those who are interested in and are in the planning stages of conducting an experiment to study microbial communities using amplicon based methods (16S, ITS, etc.). Topics to be discussed will include:

- Bioinformatics IS a Data Science
- The difference between Microbial Community Analysis and Metagenomics
- Experimental Design
- Sample preparation, best practices
- High throughput sequencing basics and choices
- Cost estimation
- Microbial Community Analysis
  - Data cleanup and quality assurance
  - Assigning reads to taxa and counting
  - Analysis of the output
- Downstream analysis/visualizations and tables

Bioinformatics IS a Data Science

Section

The data deluge

- Plucking the biology from the Noise

Reality

- Its much more difficult than we may first think

Data Science

Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.

Five Fundamental Concepts of Data Science
statisticsviews.com November 11, 2013 by Kirk Borne

HTS Experiment Philosophy

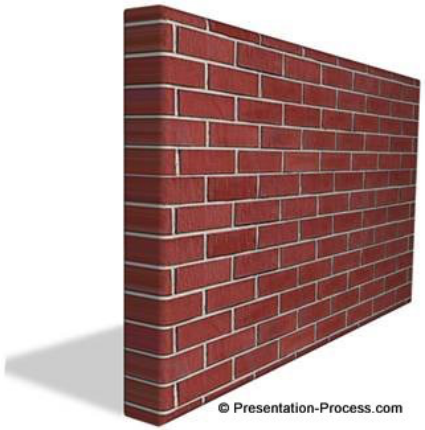
Design Experiment

Perform Experiment

Sample and Extract RNA/DNA

Prepare Libraries

Sequence



Analyze

Interpret

7 Stages to Data Science

1. Define the question of interest

2. Get the data
3. Clean the data
4. Explore the data

5. Fit statistical models

6. Communicate the results
7. Make your analysis reproducible

Bioinformatics done well

Data science done well looks easy – and that's a big problem for data scientists

simplystatistics.org
March 3, 2015 by Jeff Leek

Substrate

Cloud Computing

Cluster Computing

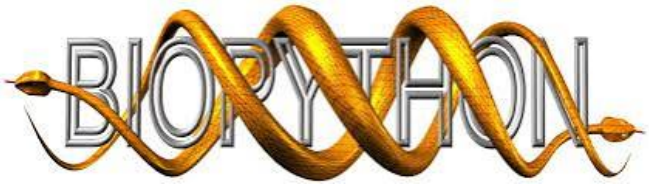
BAS™

LINUX Laptop & Desktop

Environment

“Command Line” and “Programming Languages”

Open Source



VS

Bioinformatics Software Suite

Prerequisites

- Access to a multi-core (24 cpu or greater), ‘high’ memory 64Gb or greater Linux server.
- Familiarity with the ’command line’.
- Basic knowledge of how to install software
- Basic knowledge of R and statistical programming
- Basic knowledge of Statistics and model building
- At least one programming language (eg python)

Bioinformatics

- Know and Understand the experiment
  - "The Question of Interest"
- Develop a deep familiarity and understanding of the techniques/tools/methods
- Build a set of assumptions/expectations
  - Mix of technical and biological
  - Spend your time testing your assumptions/expectations
  - Don't spend your time finding the "best" software
- Don't under-estimate the time Bioinformatics may take
- Be prepared to accept 'failed' experiments

THE Bottom Line

The Bottom Line:

- Spend the time (and money) planning and producing **good quality, accurate and sufficient data** for your experiment.
- Get to know to your data, develop and test expectations
- Result, you'll **spend much less time** (and less money) extracting biological significance and results during analysis.

Amplicons vs. Metagenomics

- Metagenomics
  - Shotgun libraries intended to sequence random genomic sequences from the entire bacterial community.
  - Can be costly per sample ($500 to multi thousands per sample)
  - Better resolution and sensitivity to characterize the sample
  - Due to cost, can only do relatively few samples
- Amplicon community profiling
  - Sequence only one regions of one gene (e.g. 16s, ITS, LSU)
  - Cheap per sample (at scale, down to $20/sample)
  - Due to cost, can do many hundreds of samples make more global inferences

Metagenomic Sequencing Designs

- Taxonomic Identification
  - Amplicon based (e.g. 16s variable regions)
  - Shotgun Metagenomics
- Functional Characterization
  - Shotgun Metagenomics
  - Shotgun Metatranscriptomics (active)
- Genome Assembly, Function and Variation
  - Shotgun Metagenomics
  - Shotgun Metatranscriptomics

UCDAVIS Bioinformatics Core

Experimental Design: Community profiling

Section

General rules for preparing samples

- Prepare more samples then you are going to need, i.e. expect some will be of poor quality, or outright fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you can, reliably
- Quality should be established using Fragment analysis traces (pseudo-gel images, FOR RNA RIN > 7.0)
- DNA/RNA should not be degraded
  - 260/280 ratios for RNA should be approximately 2.0 and 260/230 should be between 2.0 and 2.2. Values over 1.8 are acceptable
- Quantity should be determined with a Fluorometer, such as a Qubit.

BE CONSISTENT ACROSS ALL SAMPLES!!!

Proposal Stage – Reads per sample

Considerations

- Number of reads being sequenced
- Proportion that is diversity sample (e.g. PhiX)
- Number of samples being pooled in the run

The back of the envelope calculation

$$\frac{reads}{sample} = \frac{reads\_sequenced * (1 - diversity\_sample)}{num\_samples}$$

example

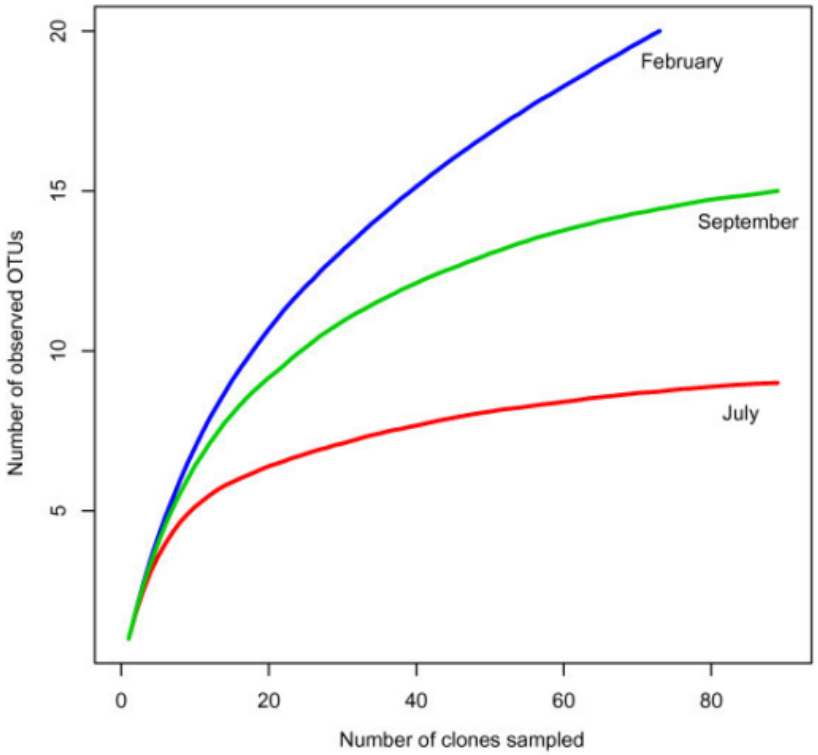
$$\frac{102{,}000}{sample} = \frac{18e6 * (1 - 0.15)}{150}$$

Recommendations

- Illumina 'recommends' 100K per sample
- I've used 30K per sample historically, others are fine with 3K per sample
- Really should have as many reads as your experiment needs

Rarefaction curves

- 'Deep' sequence a number of test samples (~1M+ reads).
- Plot rarefactions curves of Amplicons, to determine if saturation is achieved



However:

"Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible"

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531

Take Homes

- Experience and/or literature searches (other peoples experiences) will provide the best justification for estimates on needed depth.
- ‘Longer’ reads are better than short reads.
- Libraries can be sequenced again, so do a pilot, estimate ‘contamination’, compute rarefaction/16s coverage then sequence more accordingly.

Cost Estimation

- PCR reactions
  - Library QA/QC (Bioanalyzer and Qubit/microplate reader)
  - Pooling
- Sequencing (Number of Lanes / runs)
  - Because sequencing is an absolute cost, community profiling gets cheaper per sample the more you pool together. I have pooled as much as 1500 amplicons on one MiSeq V3 sequencing run.
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

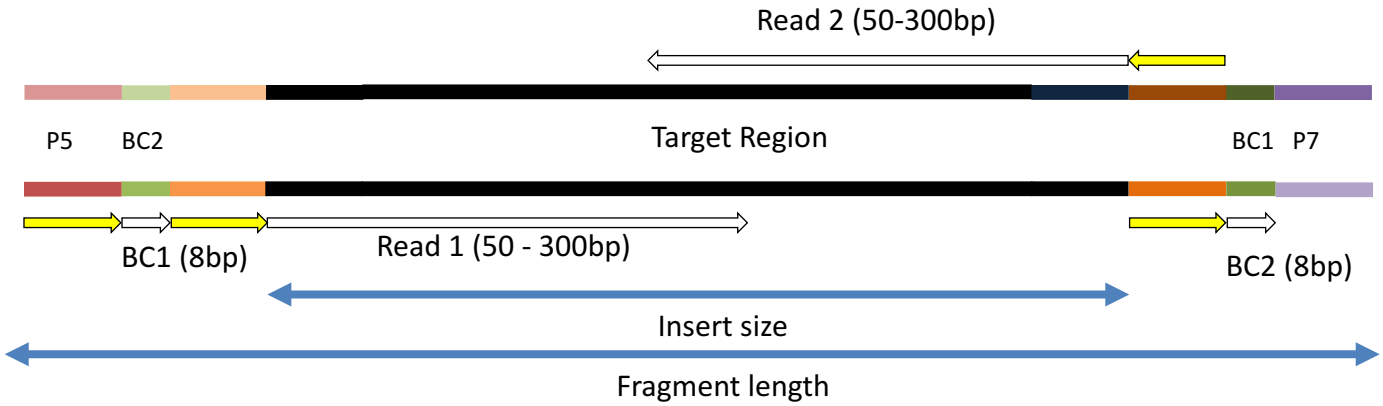
http://dnatech.genomecenter.ucdavis.edu/prices/

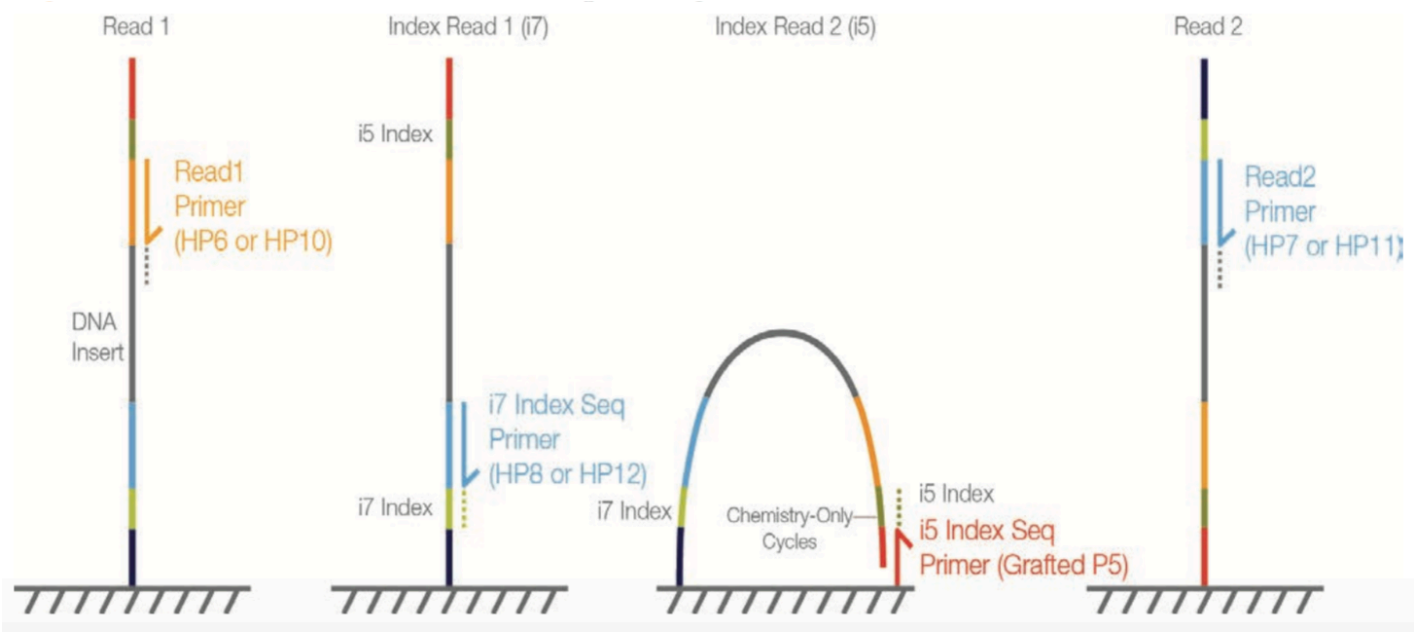
Sequencing

Illumina

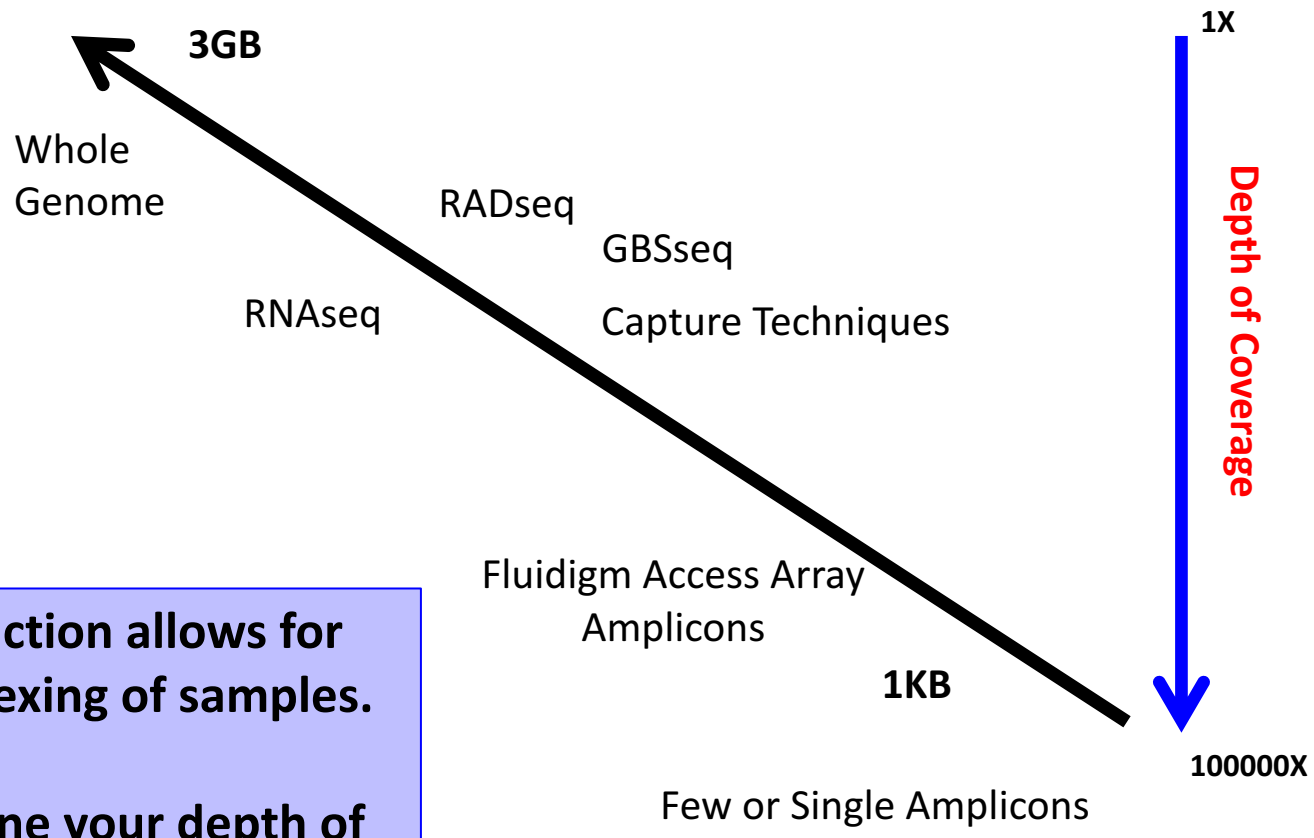
Illumina sequencing

- Illumina SBS





Genomic Reduction



Genomic reduction allows for greater multiplexing of samples.

You can fine tune your depth of coverage needs and sample size with the reduction technique

Illumina HISEQ sequencing

- http://www.illumina.com/systems/hiseq-3000-4000/specifications.html

2500

MiSeq

	HISEQ 3000 SYSTEM	HISEQ 4000 SYSTEM
No. of Flow Cells per Run	1	1 or 2
Data Yield: 2 × 150 bp 2 × 75 bp 1 × 50 bp	 650-750 Gb 325-375 Gb 105-125 Gb	 1300-1500 Gb 650-750 Gb 210-250 Gb
Clusters Passing Filter (Single Reads) (8 lanes per flow cell)	2.1-2.5 billion	4.3-5 billion
Quality Scores: 2 × 50 bp 2 × 75 bp 2 × 150 bp	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1-3.5 days	< 1-3.5 days
Human Genomes pe	up to 6	up to 12
Exomes per Run**	up to 48	up to 96
Transcriptomes per F	up to 50	up to 100

UCDAVIS Bioinformatics Core

Illumina MISEQ SEQUENCING

MISEQ REAGENT KIT V2		
READ LENGTH	TOTAL TIME*	OUTPUT
1 × 36 bp	~4 hrs	540-610 Mb
2 × 25 bp	~5.5 hrs	750-850 Mb
2 × 150 bp	~24 hrs	4.5-5.1 Gb
2 × 250 bp	~39 hrs	7.5-8.5 Gb

MISEQ REAGENT KIT V3		
READ LENGTH	TOTAL TIME*	OUTPUT
2 × 75 bp	~21 hrs	3.3-3.8 Gb
2 × 300 bp	~56 hrs	13.2-15 Gb

Reads Passing Filter†

MISEQ REAGENT KIT V2	
Single Reads	12-15 M
Paired-End Reads	24-30 M

MISEQ REAGENT KIT V3	
Single Reads	22-25 M
Paired-End Reads	44-50 M

Quality Scores††

MISEQ REAGENT KIT V2
> 90% bases higher than Q30 at 1 x 36 bp
> 90% bases higher than Q30 at 2 x 25 bp
> 80% bases higher than Q30 at 2 x 150 bp
> 75% bases higher than Q30 at 2 x 250 bp

MISEQ REAGENT KIT V3
> 85% bases higher than Q30 at 2 x 75 bp
> 70% bases higher than Q30 at 2 x 300 bp

Double barcoded primer design

Section

Goal: A culture independent method for profiling the diversity of a community.

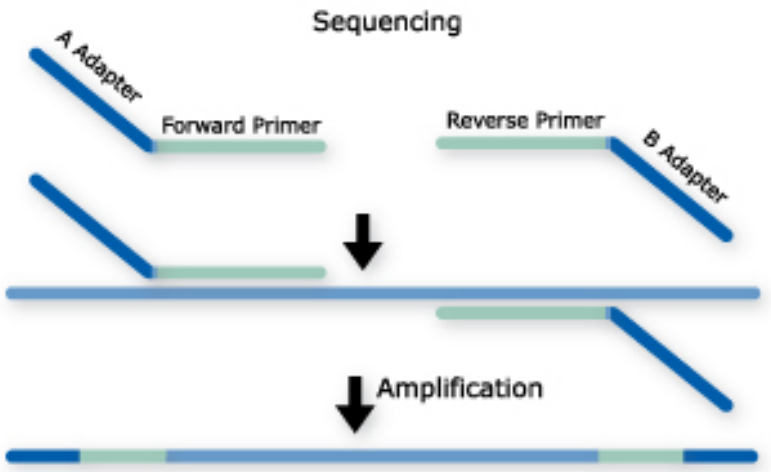
High-throughput sequencing technologies (such as Illumina) can generate millions of amplicons, across thousands of samples in a single run, and are today our best approach to deeply assess the environmental or clinical diversity of complex microbial assemblages of archaea, bacteria, and eukaryotes.

Using sequence variation within a common gene (e.g. 16s) to assign and count community members rather than counting individual cells. Assume each sequence variant is one community member.

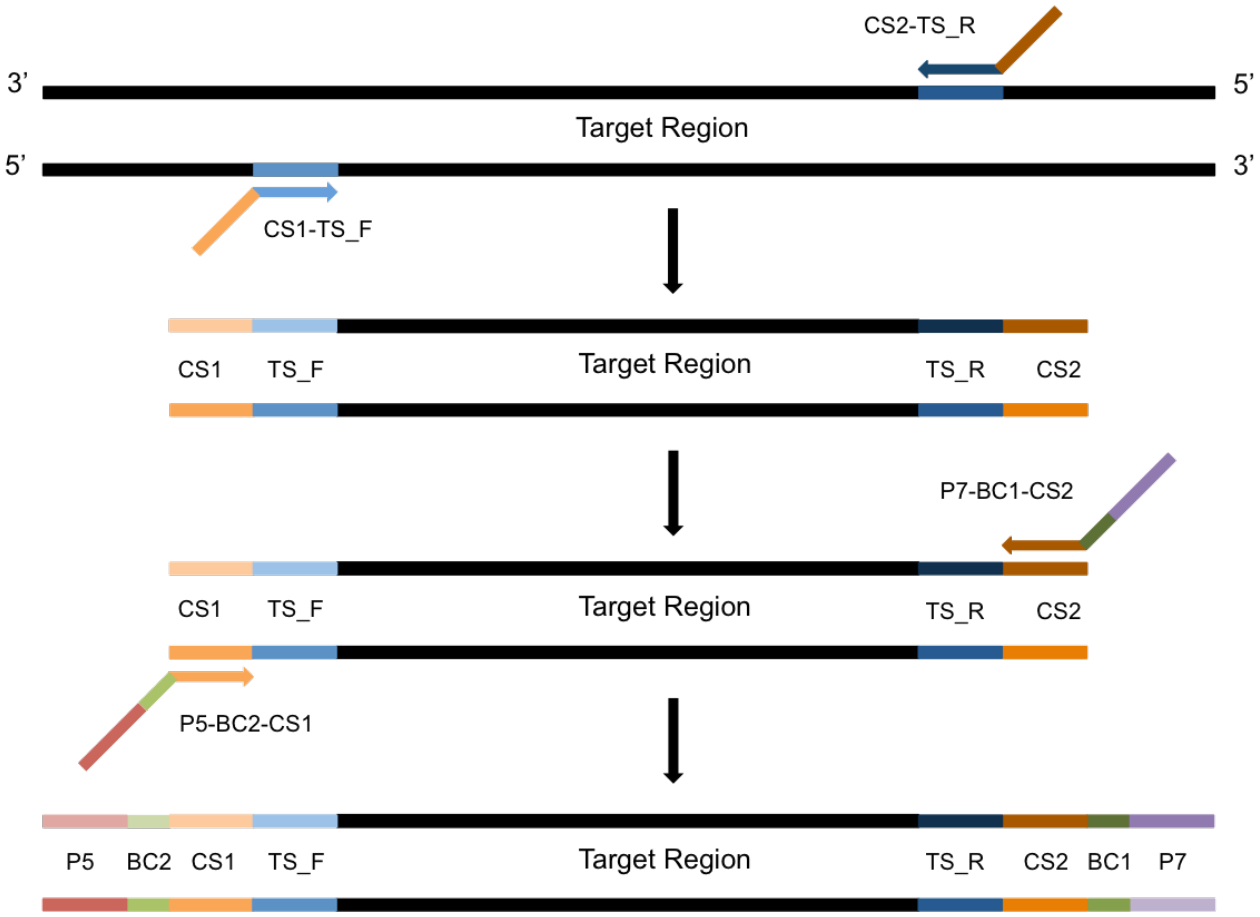
Amplicons: Common Approach

- Single PCR
- Long primer sequences (up to ~75bp) that contain barcodes and sequencing adapters
- Custom sequencing primers amplify from target primer
- Single or dual barcodes
  (dual barcode are sometimes within read)

Common approach



Amplicons: Two Step PCR Approach



PCR-1
Template with overhang

PCR-2
Add barcodes and adapters

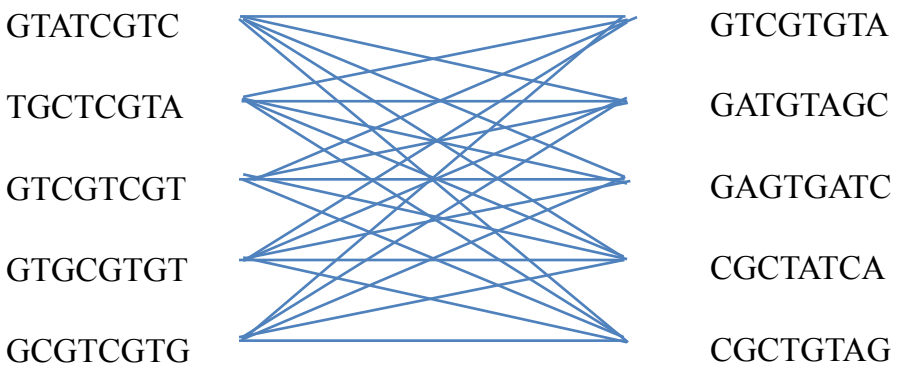
Barcodes and adapters are added in the second round of PCR

Multiplex Samples

Dual barcoding allows for massively multiplexing of samples using only a relatively few primers

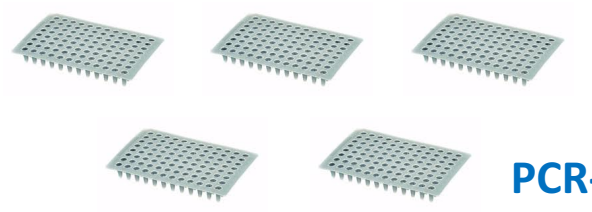
Pairing of BC1 and BC2 uniquely identifies sample

8bp [Error-Correcting] Barcodes



5 Pairs of Barcodes allows for multiplexing of **25 samples**.
32 Pairs can multiplex **1024 samples** in the same sequencing reaction

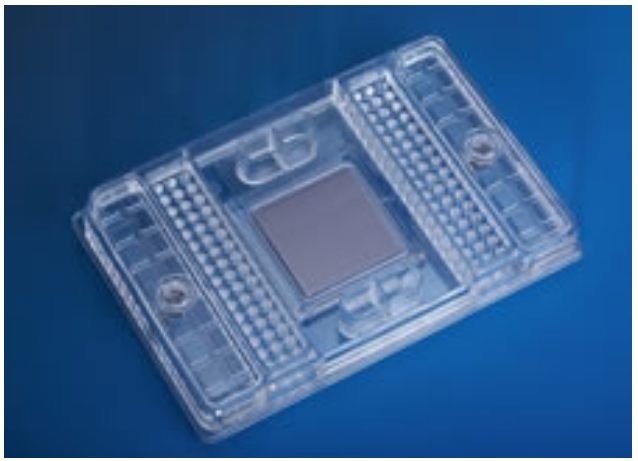
Multiplex Amplicon Targets



PCR-1 product

PCR-2 product

Fluidigm Access Array System



48 samples X 48 amplicons
2304 Two-step PCR reactions

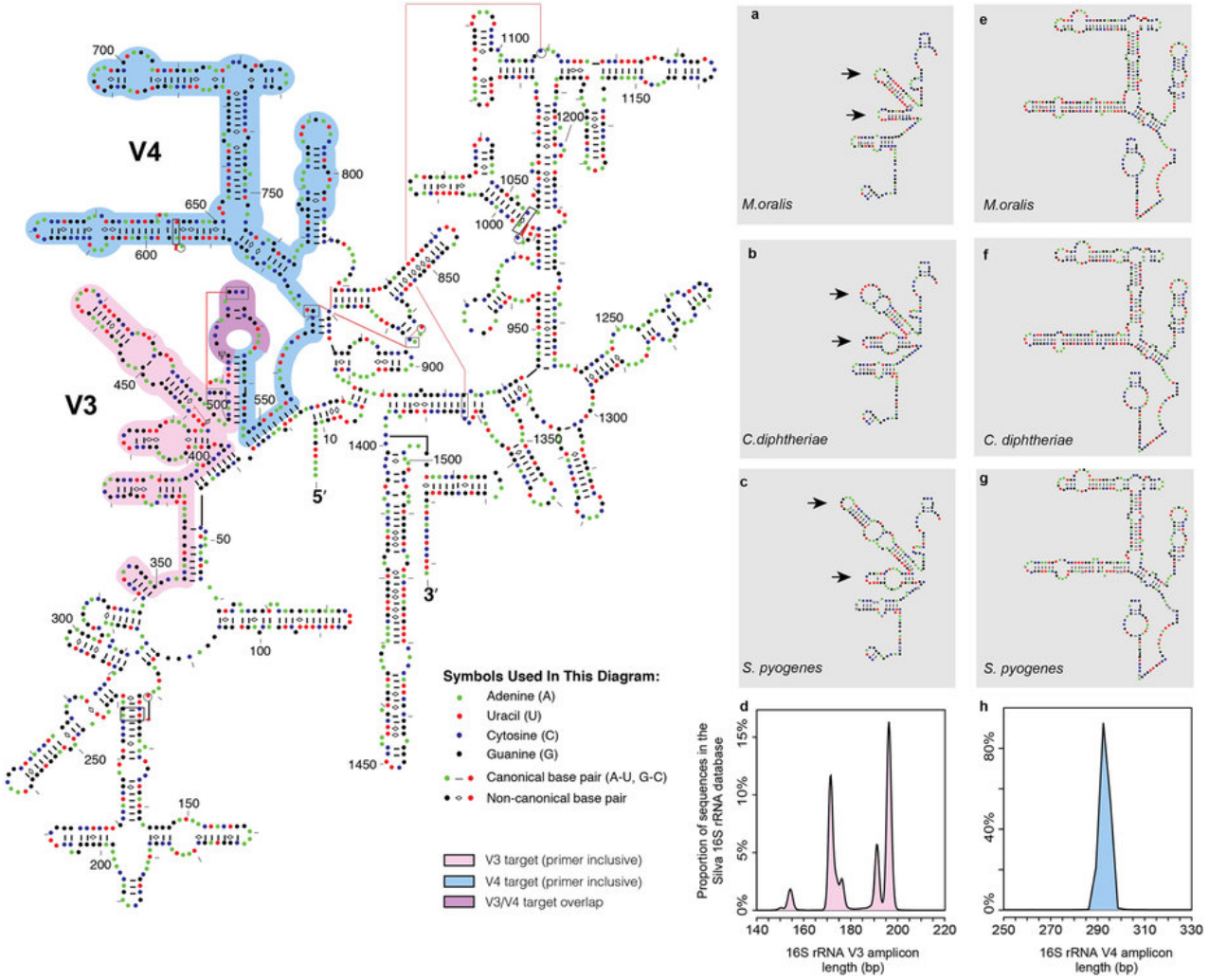
Pool many samples

dbcAmplicons

- Originally conceived in late 2012 to lower per sample costs on relatively short targeted (PCR) regions
  - 16S, ITS, LSU, 18S, etc. Community profiling
  - Extraction of mitochondria, virae, chloroplast regions, plasmids by PCR
  - Genotyping of samples for phylogenomics, genome to phenotype interactions
- Uses the Illumina platform, capably of pooling thousands, or even tens of thousands of barcoded samples/targets per sequencing run.
- Core Facility friendly, facilitates interactions between and across individual labs, standardizing workflows.
- Consistency across labs leads to predictability in the DNA Technologies Core, which leads to consistency in result.

UCDAVIS Bioinformatics Core

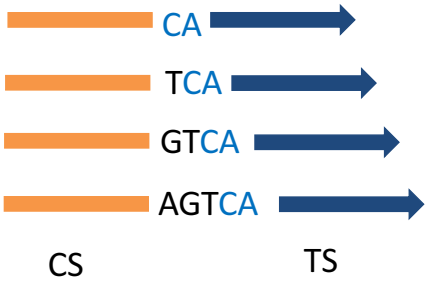
Primer Design



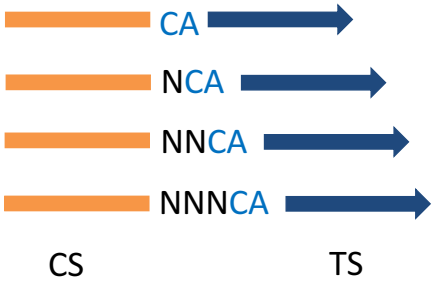
Template Specific Primer Design

- Each primer pair contains the following parts
  - CS1 or CS2 to attach second adapter/barcode primer
  - Phase-shifting bases [see below]
  - Linker sequence
  - Template specific primer sequence

Phase shifting primer



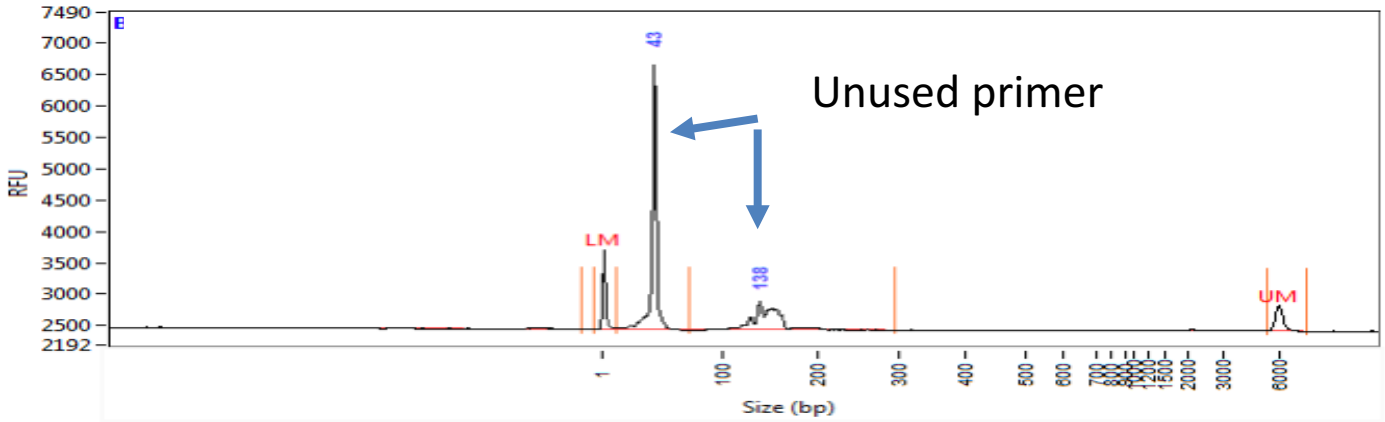
Phase shifting primers with PCR duplicate detection



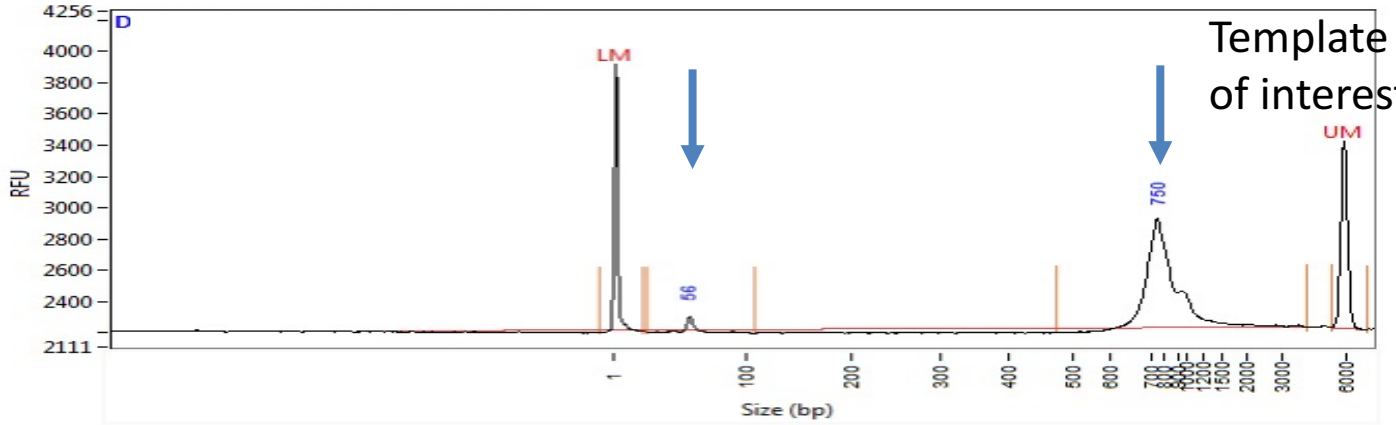
Ns, resolve to be PCR duplicate keys
(Unique Molecular Indexes, UMIs)
and should only ever appear once

QC: what is a “good” PCR library?

Unused primers and adapters



BAD!



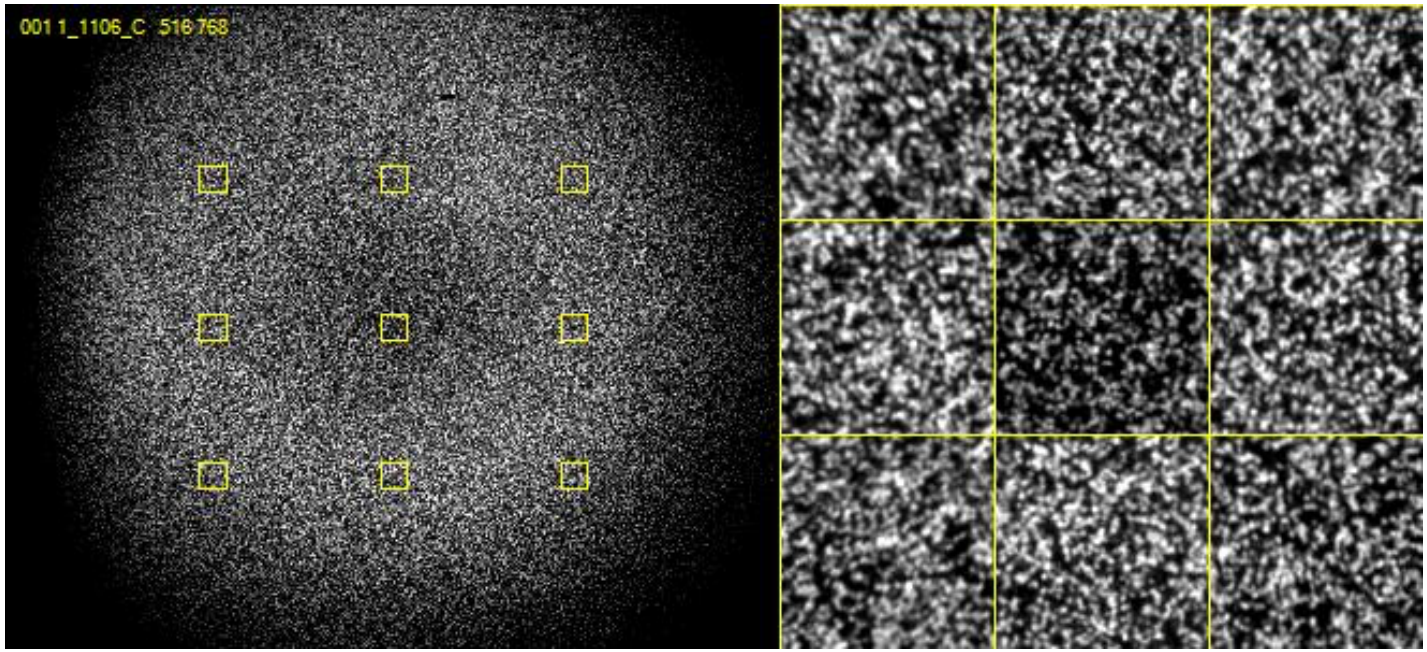
Good!

Pooling Samples/Amplicons

- Even amplicon representation is important and difficult to achieve. Amplicon counts can vary from sample to sample by 100x
- Each amplicon should be evaluated by quality (ideally by trace) and quantity (fluourometery). Both qualities will effect final counts.
- Best practices
  - First group amplicons by quality/quantity profiles
  - Pool each group separately
  - If a small number of groups consider qPCR on each group for final pooling concentrations
  - If a large number re-quantify and pool to final pool.

Nucleotide diversity

Critically important for imaging clusters



Amplicon runs are Low Complexity

- Sequencing with lower cluster density
- Adding in highly diverse libraries (shotgun libraries ~15%)
- Phase-shifting template specific primers:

What is nucleotide diversity?

The term "nucleotide diversity" describes the proportion of each nucleotide (A, T, C, and G) at each position in a sequencing library. A "balanced" or "diverse" library has *equal proportions* of A, C, G, and T nucleotides at each base position in a sequencing library. Figure 9 illustrates cluster images and SAV data by cycle plots from diverse/ balanced, low-diversity, and unbalanced libraries.

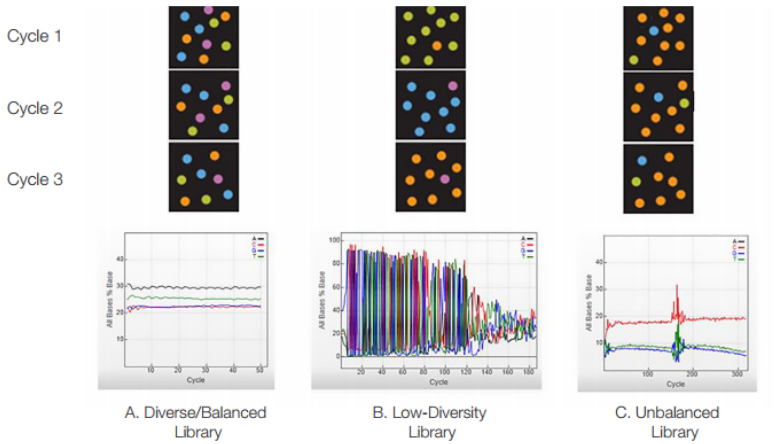


Figure 9: Nucleotide Diversity and Data by Cycle: % Base Plots. A) Diverse/balanced libraries contains equal proportions of A, T, C, and G nucleotides. The Data by Cycle: % Base plot shows even, horizontal curves centered around 25%. B) Low-diversity libraries, such as amplicon, enriched/targeted, or ChIP libraries, have an uneven proportion of nucleotides across the flow cell from one cycle to the next. The Data by Cycle: % Bases plot shows large intensity spikes at each cycle. C) Unbalanced libraries, such as bisulfite converted or Tetrahymena libraries, have one base at a much lower percentage than the others. This example shows a library with a low percentage of the "A" nucleotide.

Benefits/Drawbacks

Benefits

- Maximum Flexibility, fewer target specific primers needed.
- Dual barcoding, allowing for massively multiplexing of samples to occur.
- Pool multiple targets per run
- Software for demultiplexing

Drawbacks

- Two – step PCR reaction
- Sequence the target specific primer
  - ❖ In my opinion NOT a feature not a drawback

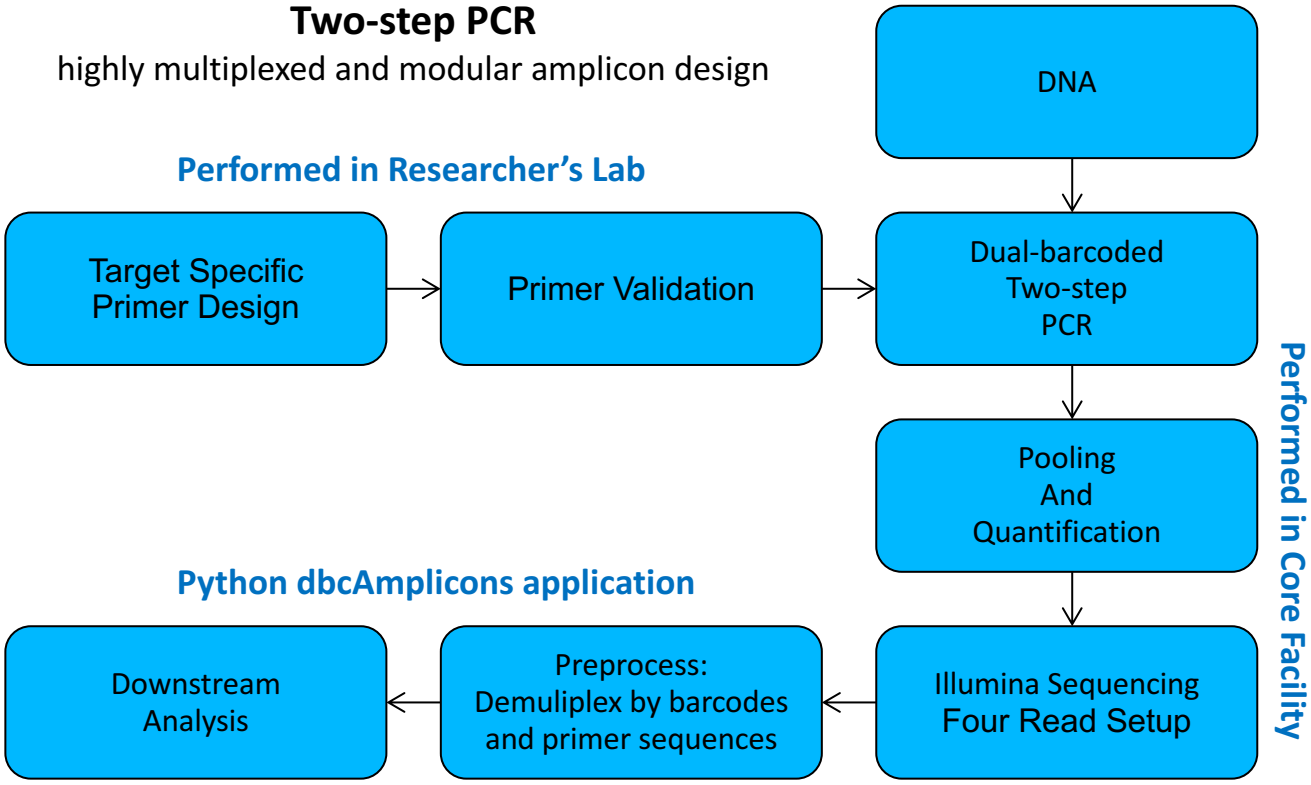
Common Analysis Workflow

1. Identify barcodes
2. Identify primer sequence (if present) and trim
3. Overlap paired end reads to produce single read, full amplified target sequence
4. Generate operational taxonomic unites (“OTUs”), via clustering or classification
5. Assign ”OTUs” to an organism
6. Generate abundance tables
7. Statistical testing

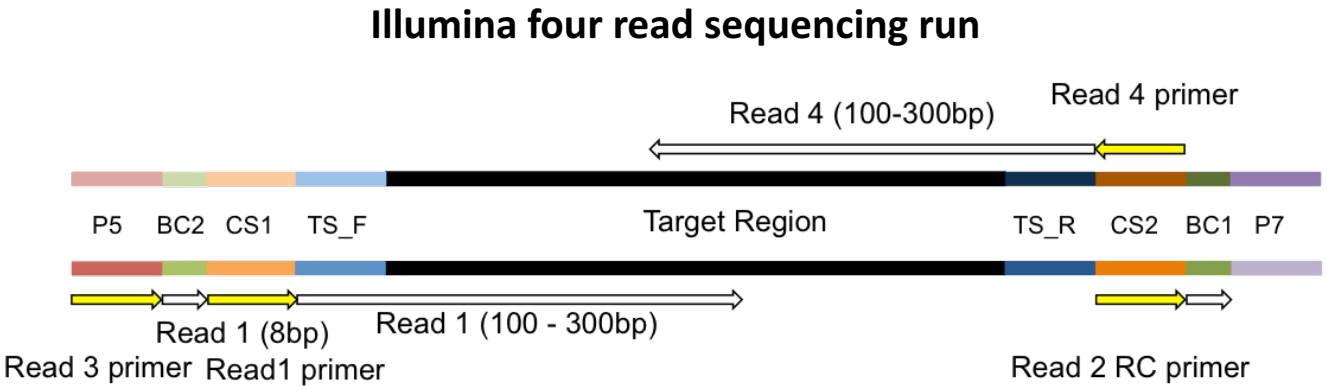
Common workflows

- Qiime
  - Worst piece of software to install ever
- Mothur
- dbcAmplicons (my software)

dbcAmplicons

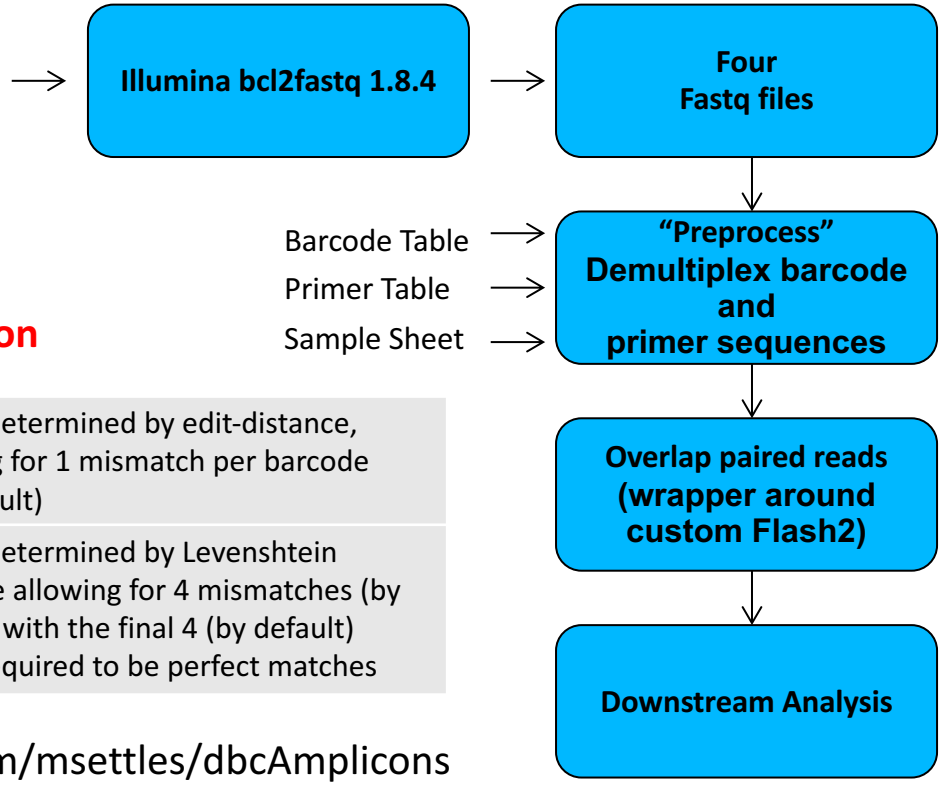


Sequencing



Read	Sequencing Primer
Read1 primer	CS1 - 5' ACACTGACGACATGGTTCTACA
Read2 primer	CS2 - 5' TACGGTAGCAGAGACTTGGTCT
BC1 primer	CS2rc - 5' AGACCAAGTCTCTGCTACCGTA
BC2 primer	Uses the P5 amplification primer

Bioinformatics



dbcAmplicons
Python Application

Barcode	Match determined by edit-distance, allowing for 1 mismatch per barcode (by default)
Primers	Match determined by Levenshtein Distance allowing for 4 mismatches (by default) with the final 4 (by default) bases required to be perfect matches

https://github.com/msettles/dbcAmplicons

Downstream Analysis

Population Community Profiling (i.e. microbial, bacterial, fungal, etc.)

dbcAmplicons Python Application

Screen	Using Bowtie2, screen targets against a reference fasta file, separating reads by those that produce matches and those that do not match sequences in the reference database.
Classify	Wrapper around the MSU Ribosomal Database Project (RDP) Classifier for Bacterial and Archaeal 16S rRNA sequences, Fungal 28S rRNA, fungal ITS regions
Abundance	Reduce RDP classifier results to abundance tables (or biom file format), rows are taxa and columns are samples ready for additional community analysis.

Targeted Re-sequencing

	Consensus - Reduce reads to consensus sequence for each sample and amplicon
R-functions to be added into dbcAmplicons	**Most Common –** Reduce reads to the most commonly occurring read in the sample and amplicon (that is present in at least 5% and 5 reads, by default)
	Haplotypes – Impute the different haplotypes in the sample and amplicon

Supplemental Scripts

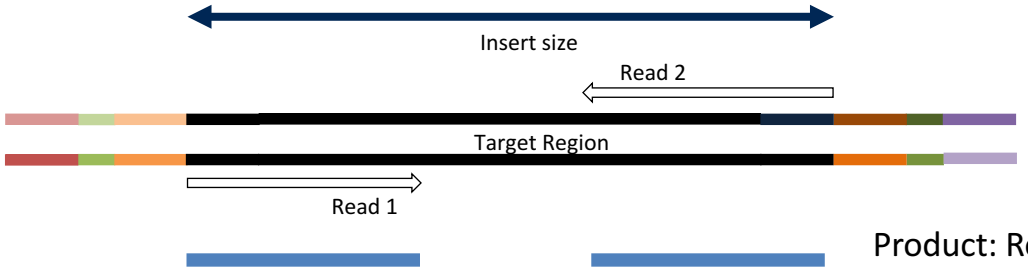
- convert2Readto4Read.py
  - For when samples are processed by someone else
- splitReadsBySample.py
  - To facilitate upload to the SRA
- preprocPair_with_inlineBC.py
  - Cut our inline BC and create 4 reads for standard input processing
  - Will work with ”Mills lab” protocol
- dbcVersionReport.sh
  - Print out version numbers of all tools

dbcAmplicons: dependent software

- dbcAmplicons
  - https://github.com/msettles/dbcAmplicons
- Flash2, for overlapping reads
  - https://github.com/dstreett/FLASH2
- RDP for classification
  - https://github.com/rdpstaff/RDPTools
- Bowtie2 for screening (if relevant)
  - https://github.com/BenLangmead/bowtie2

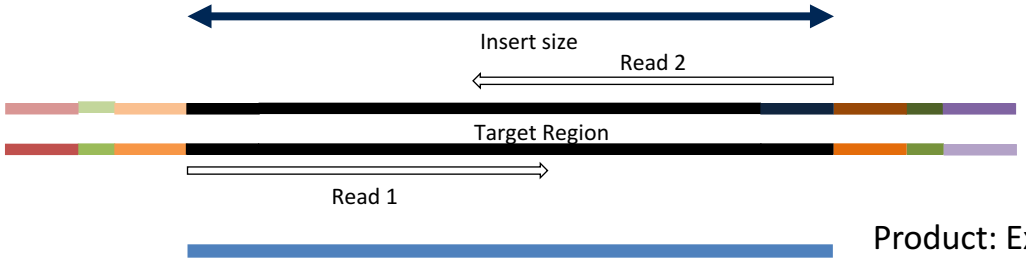
Flash2 – overlapping of reads

Insert size > length of the number of cycles



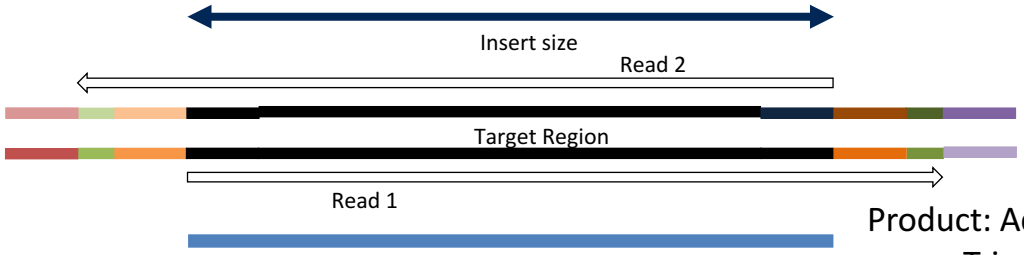
Product: Read Pair

Insert size < length of the number of cycles (10bp min)



Product: Extended Single

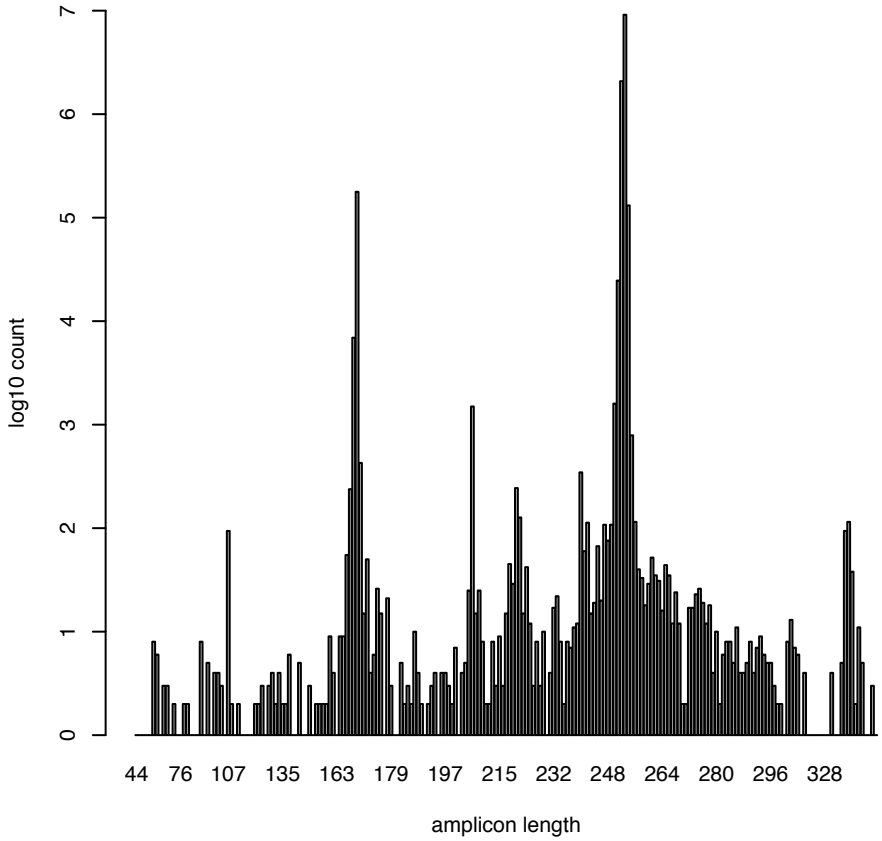
Insert size < length of the read length



Product: Adapter Trimmed Single

Flash2 typically produces tight sizes

241	60
242	113
243	15
244	19
245	67
246	20
247	108
248	76
249	108
250	1598
251	24682
252	2083811
253	9136890
254	131296
255	789
256	115
257	40
258	33
259	18
260	29
261	52
262	35
263	31
264	16



dbcAmplicons: classify

- Uses the RDP (Ribosomal Database Project) classifier for bacterial and archaeal 16S, fungal LSU, ITS warcup/unite databases.
- Classifies sequences to the closest taxonomic reference provides a bootstrap score for reliability
- Concatenates Paired-end reads
  - Can trim off low quality ends, to some value Q

Direct Classification - RDP

- Ribosomal Database Project (RDP) - naïve Bayesian Classifier
  - Compares each read to a database
    - Database is updates periodically
  - Compares by k-mers (15 mers)
  - 100 bootstraps to establish confidence in result
- Order does not matter, no 3% !
- Drawbacks
  - Accepts only fasta (though website implies fastq) files
  - Can be slow
  - Down to genus only (for 16s, species for ITS)
  - Kmer database are based on whole 16s
  - Cannot group together unknown OTUs that represent unique taxa

Direct classification vs clustering

- Clustering – “Because of the increasing sizes of today’s amplicon datasets, fast and greedy *de novo* clustering heuristics are the preferred and only practical approach to produce OTUs”. I DISAGREE

Shared steps in these current algorithms are:

1. An amplicon is drawn out of the amplicon pool and becomes the center of a new OTU (centroid selection)
2. This centroid is then compared to all other amplicons remaining in the pool.
3. Amplicons for which the distance is within a global clustering threshold, *t (e.g. 3%)*, to the centroid are moved from the pool to the OUT
4. The OTU is then closed. These steps are repeated as long as amplicons remain in the pool.

Reasons why I'm not a fan

1. Little to no biological rational to any of the clustering parameters, modify the parameters to get a result you like.
2. Dependent on ordering, reorder our reads you can get different set of OTUs. Often not repeatable from run to run.
3. 3% (or any other cutoff) is BS.
4. Most clustering algorithms do not consider sequencing errors.
5. If you generate more data you have to start the clustering process all over again as population of sequences matters.
6. I'm sure there is more

OTU clustering Comparison

	Clone43			
	Expected OTUs	**Inferred* OTUs(2%)**	**inferred OTUs(3%)**	**inferred OTUs(4%)**
Mothur	43	1882	720	369
Muscle+Mothur		2478	1418	784
ESPRIT		4474	4397	1733
ESPRIT-Tree		2301	1096	279
SLP		286	245	227
Uclust		2177	1883	597
CD-HIT		1473	1464	481
DNAClust		3768	3658	1103
GramCluster		2119	2071	2071
CROP		339	133	62

*: all the listed numbers of OTU are the average numbers over xx simulations.
doi:10.1371/journal.pone.0070837.t002

Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H (2013) A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. PLoS ONE 8(8): e70837. doi:10.1371/journal.pone.0070837
http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0070837

Abundance tables and Biom files

	A	B	C	D	E	F	G	H	I	J	K	L
1	Taxon_Name	Level	MeanBootstraptValue	Sample1	Sample10	Sample11	Sample12	Sample13	Sample14	Sample15	Sample16	Sample17
2	Acidovorax	genus	0.973	0	0	0	0	0	0	0	0	0
3	Actinobaculum	genus	1	0	0	0	0	0	0	0	0	0
4	Actinomyces	genus	0.988	0	0	0	0	0	0	0	0	0
5	Actinomycetaceae	family	0.83	0	0	0	0	0	0	0	0	0
6	Actinomycetales	order	0.923	0	0	0	0	0	0	0	0	0
7	Aeriscardovia	genus	0.68	0	0	0	0	0	0	0	0	0
8	Aerococcus	genus	1	0	0	0	0	0	1	0	0	0
9	Alloscardovia	genus	1	0	3	0	0	0	0	0	0	0
10	Lactobacillus	genus	0.977	41	75	55	96	56	66	12	0	114
11	Anaerococcus	genus	0.946	0	1	0	0	0	1	0	0	0
12	Anaerosphaera	genus	0.582	0	0	0	0	0	0	0	0	0
13	Archaea	domain	0.587	0	0	0	0	0	0	0	0	0
14	Atopobium	genus	1	0	0	0	0	0	0	0	0	0
15	Bacilli	class	0.606	0	1	1	2	1	0	0	0	5
16	Bacillus	genus	0.9	0	0	0	0	0	0	0	0	0

Can open file in spreadsheet; lists lowest taxonomically classified level for each sample by read count

Biom file

The **BIOM file format** (canonically pronounced *biome*) is designed to be a general-use format for representing biological sample by observation contingency tables. BIOM is a recognized standard for the **Earth Microbiome Project** and is a **Genomics Standards Consortium** supported project.

Future Directions

- dbcAmplicons is a data reduction pipeline, produces abundance/biome files, post processing most typically done in R.
- Include “error-correcting barcodes” in demultiplexing
- Identification of PCR duplicates (using UMI)
- Replace RDP classification with another scheme
  - Have ideas (for years) but no time
- Use amplicon length in classification
- Include screening of diversity sample in preprocessing to get an idea of actual proportion in the pool
- Incorporate the R genotyping pipeline into dbcAmplicons
  - Extend to inferring copy number (or ploidy levels)
- Correct for copy number (16s)

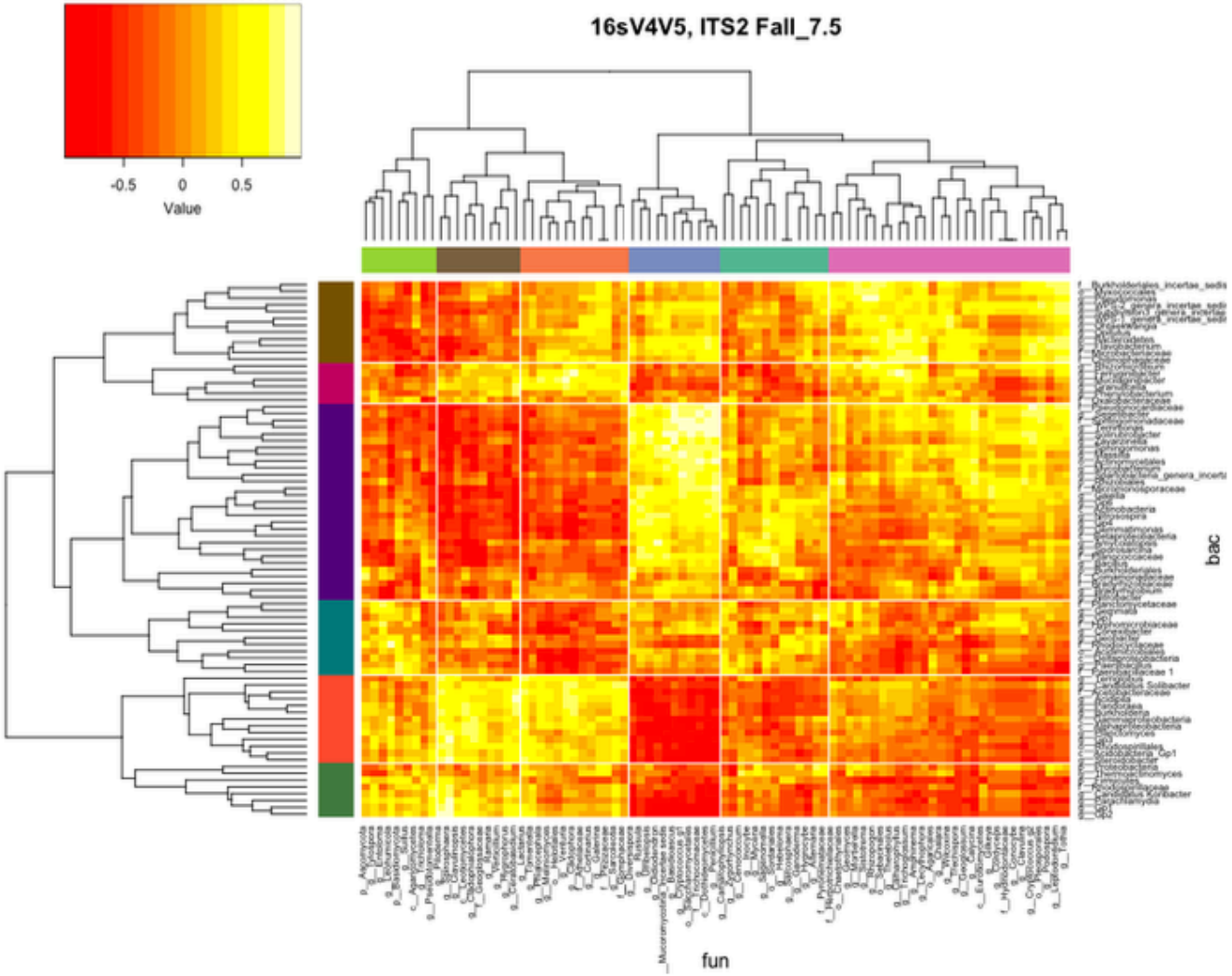
Post Processing

- I pretty much do all of my post analysis (abundance table, Biome) in R
  - Common Packages
    - Vegan
      - ❖ https://github.com/vegandevs/vegan
    - Vegetarian
      - ❖ https://github.com/cran/vegetarian
    - Phyloseq (uses vegan, ade4, ape, picante)
      - ❖ https://joey711.github.io/phyloseq/
- Ecological Diversity Analysis
  - how does the structure of OTU across samples/groups compare
- Ordination Analysis (multivariate analysis)
  - Visualize the relative similarity/dissimilarity across samples, test for taxa/environment relationships
- Differential Abundance Analysis (univariate analysis)
  - Uses tools from RNAseq (limma, edgeR)
- Visualization (temporal, heatmaps, 'trees', more)

Standardization/Normalization

- Relative (proportional) abundances
  - Divide by sum of sample, values 0-100%
- LogCPM from RNAseq
- Hellinger standardization
  - http://biol09.biol.umontreal.ca/PLcourses/Section_7.7_Transformations.pdf
- Others
  - Wisconsin

Multi community analysis



Getting Help

- Host Microbe Systems Biology Core (HMSB)
  - http://www.ucdmc.ucdavis.edu/medmicro/hmsbcore/index.html
  - DNA Extractions
  - Library Construction
  - Data Analysis
- DNA Technologies Core
  - http://dnatech.genomecenter.ucdavis.edu/
  - Illumina Sequencing of Libraries
- Bioinformatics Core
  - http://bioinformatics.ucdavis.edu/
  - Experimental design
  - Data analysis
  - Custom analysis