UCDAVIS Bioinformatics Core

dbcAmplicons pipeline Amplicons

Matthew L. Settles
Genome Center Bioinformatics Core
University of California, Davis
settles@ucdavis.edu; bioinformatics.core@ucdavis.edu

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

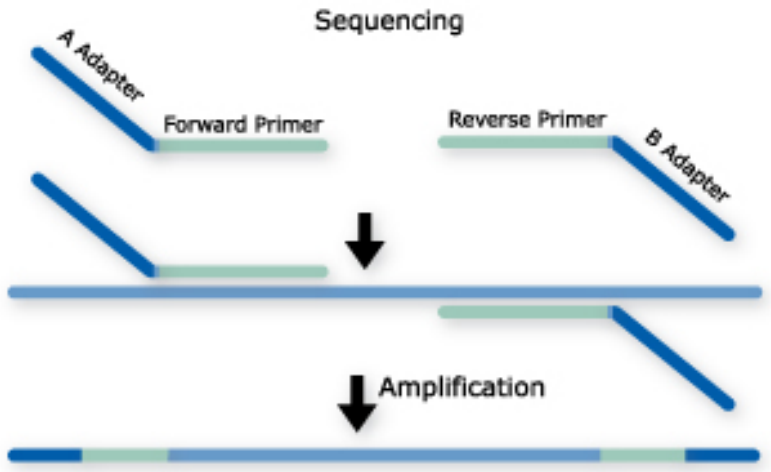
High-throughput sequencing technologies (such as Illumina) can sequence millions of amplicons, across thousands of samples in a single run, and are today our best approach to deeply assess the environmental or clinical diversity of complex microbial assemblages of archaea, bacteria, and eukaryotes.

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 (among many)

- Single PCR
- Long primer sequences (~75bp) that contain barcodes and sequencing adapters
- Single or dual barcodes
  (dual barcode often within read 2)

Common approach



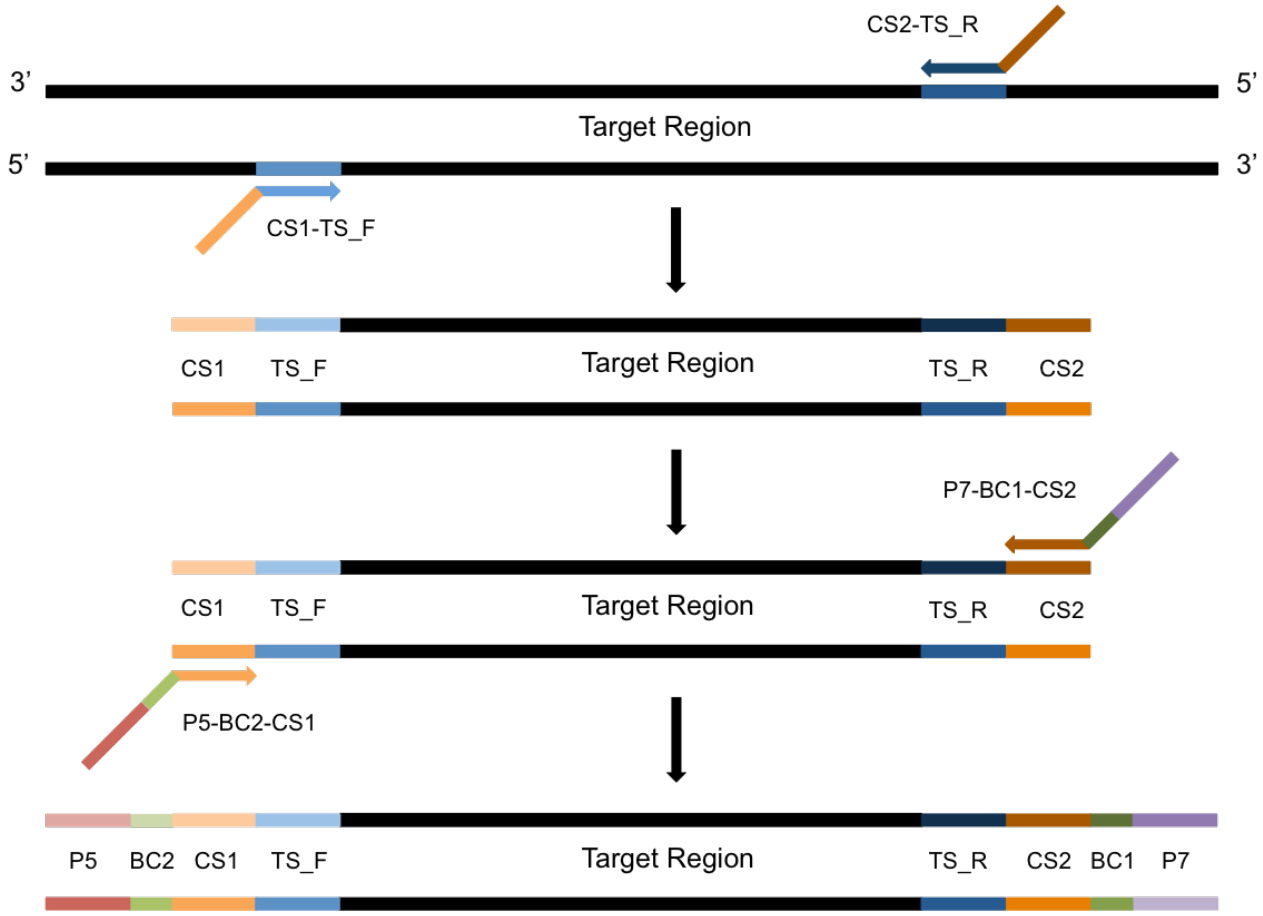
UCDAVIS Bioinformatics Core

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.

UCDAVIS Bioinformatics Core

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

Amplicon sequencing with dual barcoding

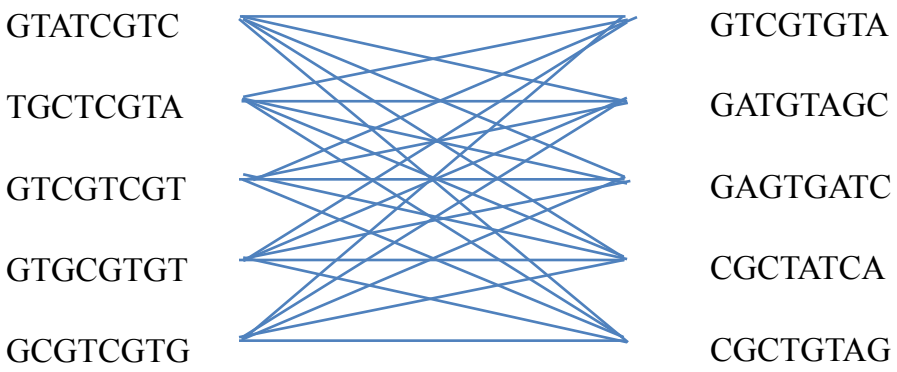
- 2-step PCR, where the first PCR extracts out the target specific region and the second PCR add on adapters and barcodes. Target specific primers include universal sequences CS1 and CS2, the second PCR extends the universal sequences with adapters and barcodes.
- Adapters and barcodes are not included in the target specific primers which allows for maximum flexibility in target specific primer usage and the ability to swap out targets, or include multiple targets in the same sequencing reaction without needing to purchase a large number of barcoded, target specific primers.
- Barcodes are included in both adapters, therefor a pair of barcodes are used to uniquely identify a samples. This allows for 32 barcode pairs to be able to uniquely identify 1024 samples.

Multiplex Samples

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

Pairing of BC1 and BC2 uniquely identifies sample

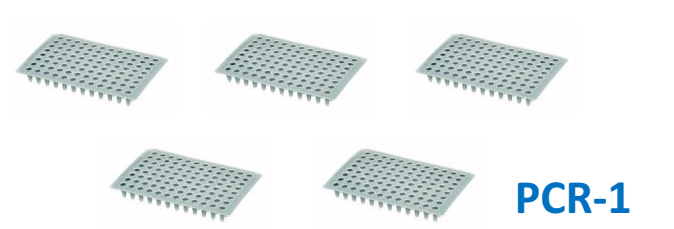
8bp [Error-Correcting] Barcodes



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

UCDAVIS Bioinformatics Core

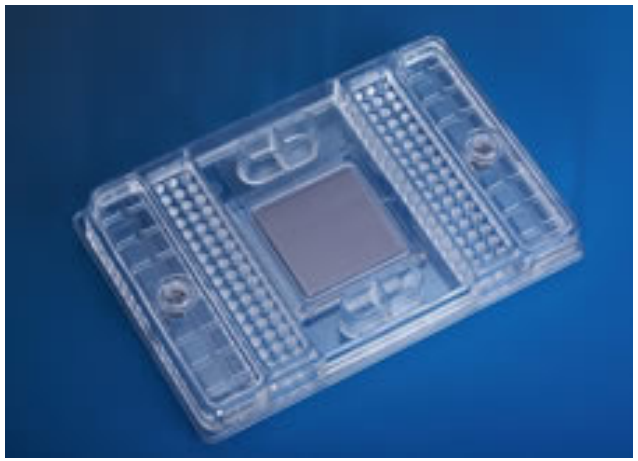
Multiplex Amplicon Targets



PCR-1 product

PCR-2 product

Fluidigm Access Array System

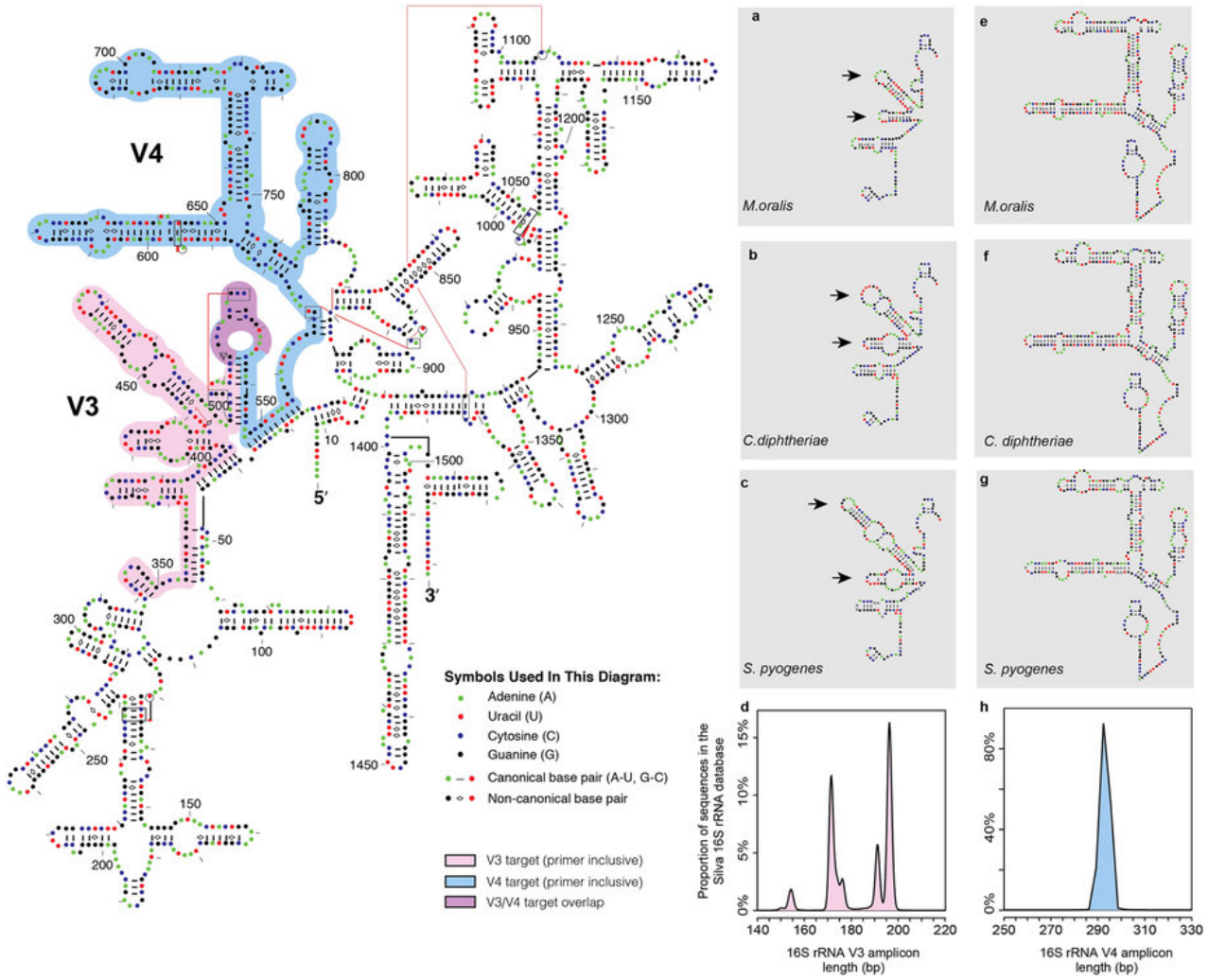


48 samples X 48 amplicons
2304 Two-step PCR reactions

Pool many samples

UCDAVIS Bioinformatics Core

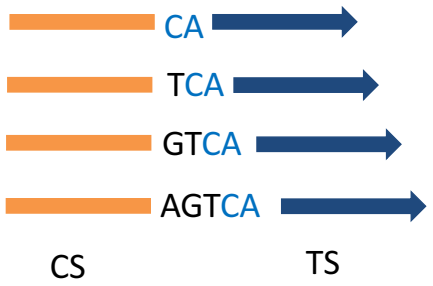
Primer Design



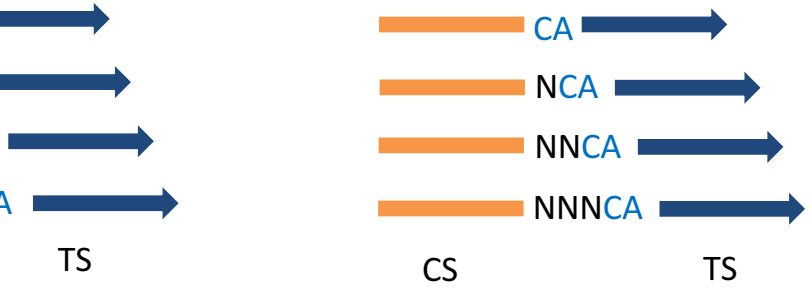
Template Specific Primer Design

- 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 and should only ever appear once

Components of the target specific primer

- CS1, or CS2 sequence
  - Provides the sequence necessary for priming of PCR-2, also serves as the sequencing primer site
- Phase-shifting bases
  - Generates diversity in the sequencing reaction
- Linker sequence
  - Buffers the target specific primer sequence from the rest of the primer, preventing some taxa (longer priming) from being more efficient than others.
- Target specific primer sequence

Example target specific primers

- Example 27F and 534R (red bases are the inserted bases)

CS1-27F	ACACTGACGACATGGTTCTACAAGAGTTTGATCCTGGCTCAG
CS1-27F_2	ACACTGACGACATGGTTCTACACAGAGTTTGATCCTGGCTCAG
CS1-27F_3	ACACTGACGACATGGTTCTACATCAGAGTTTGATCCTGGCTCAG
CS1-27F_4	ACACTGACGACATGGTTCTACAGTCAGAGTTTGATCCTGGCTCAG

CS2-534R	TACGGTAGCAGAGACTTGGTCTATTACCGCGGCTGCTGG
CS2-534R_2	TACGGTAGCAGAGACTTGGTCTCATTACCGCGGCTGCTGG
CS2-534R_3	TACGGTAGCAGAGACTTGGTCTTGATTACCGCGGCTGCTGG
CS2-534R_4	TACGGTAGCAGAGACTTGGTCTGCGATTACCGCGGCTGCTGG

CS1/CS2 sequence, Phase-shifting bases, Linker sequence, target specific primer

Components of the Barcoded adapter primers sequence

- P5, or P7 sequence
  - Primers to the Illumina flow cell, Sequence on the P5 strand typically constitutes R1, those on P7 strand typically constitutes R2.
- Barcode sequence
  - Uniquely identifies sample
- CS1, or CS2 sequence
  - Necessary for extending PCR-1

Example barcoded adapter primers

- Example 27F and 534R (red bases are the inserted bases)

P5-Index1-CS1	AATGATACGGCGACCACCGAGATCTACACTAGATCGCACACTGACGACATGGTTCTACA
P5-Index2-CS1	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTGACGACATGGTTCTACA
P5-Index3-CS1	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTGACGACATGGTTCTACA
P5-Index4-CS1	AATGATACGGCGACCACCGAGATCTACACAGAGTAGAACACTGACGACATGGTTCTACA

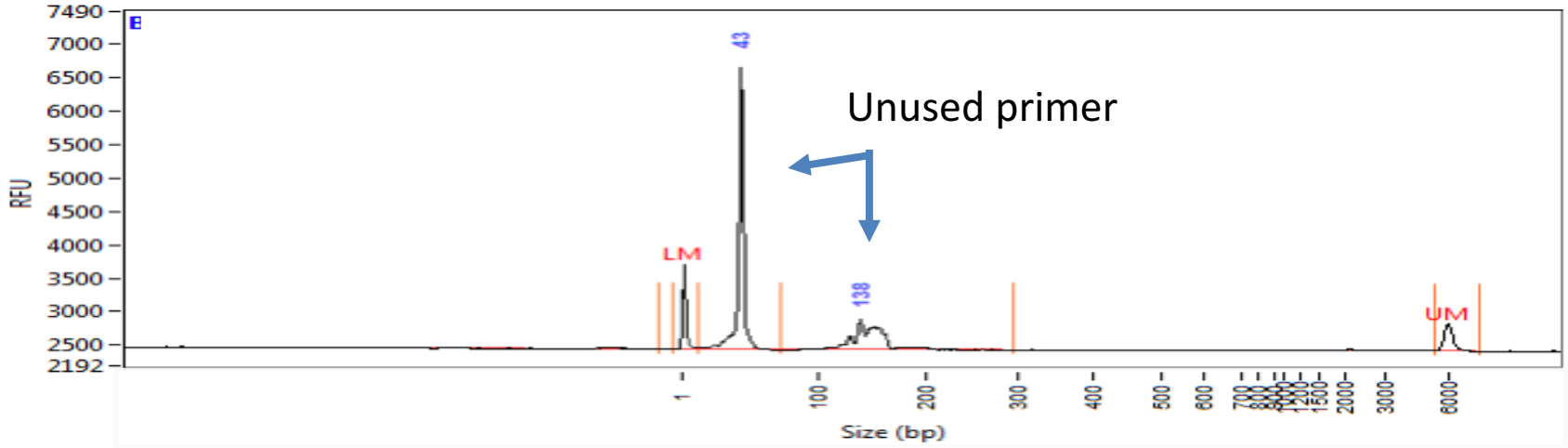
P7-Index1-CS2	CAAGCAGAAGACGGCATACGAGATTAAGGCGATACGGTAGCAGAGACTTGGTCT
P7-Index2-CS2	CAAGCAGAAGACGGCATACGAGATCGTACTAGTACGGTAGCAGAGACTTGGTCT
P7-Index3-CS2	CAAGCAGAAGACGGCATACGAGATAGGCAGAATACGGTAGCAGAGACTTGGTCT
P7-Index4-CS2	CAAGCAGAAGACGGCATACGAGATTCCTGAGCTACGGTAGCAGAGACTTGGTCT

P5/P7 sequence, Barcode, CS1/CS2 sequence

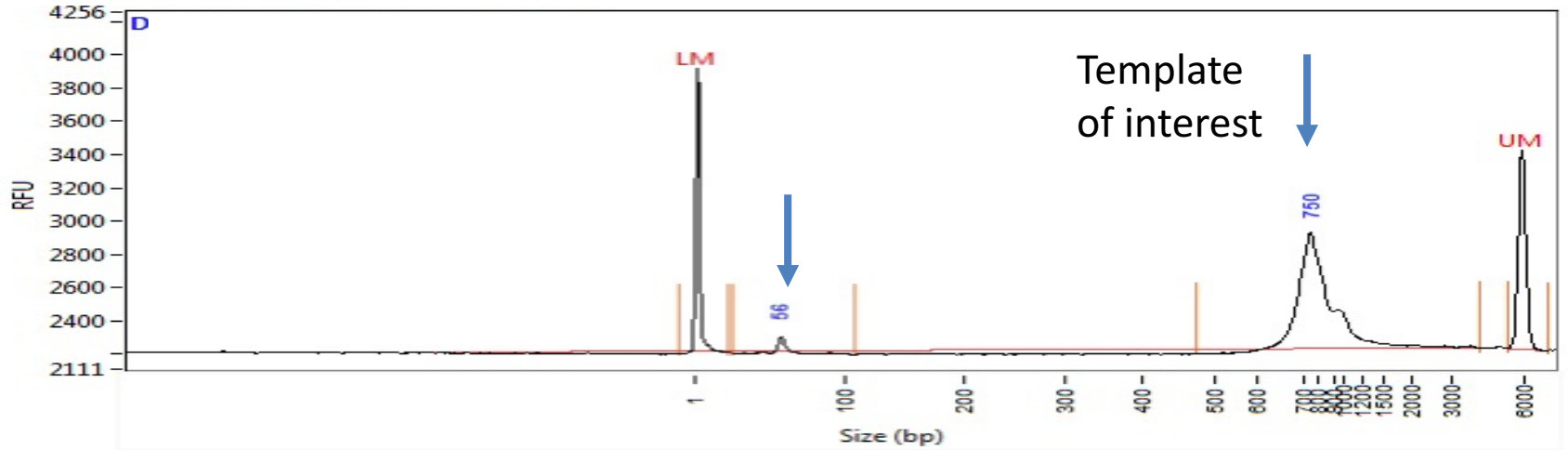
UCDAVIS Bioinformatics Core

QC: what is a “good” 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.

Benefits/Drawbacks

Benefits

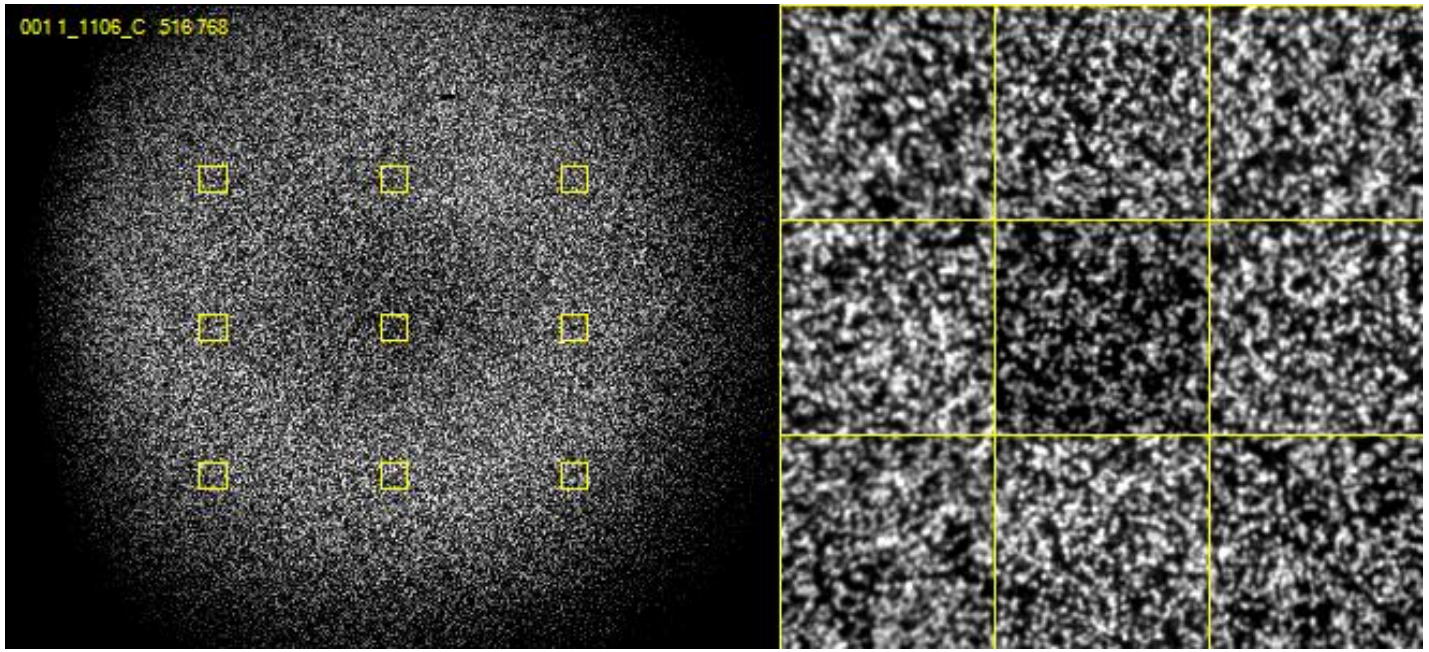
- Maximum Flexibility, fewer target specific primers needed.
- Dual barcoding, allowing for massively multiplexing of samples to occur.
- Pool multiple targets per run
- Software for demultiplexing

Drawbacks

- Two – step PCR reaction
- Sequence the target specific primer

Nucleotide diversity

Critically important for imaging clusters

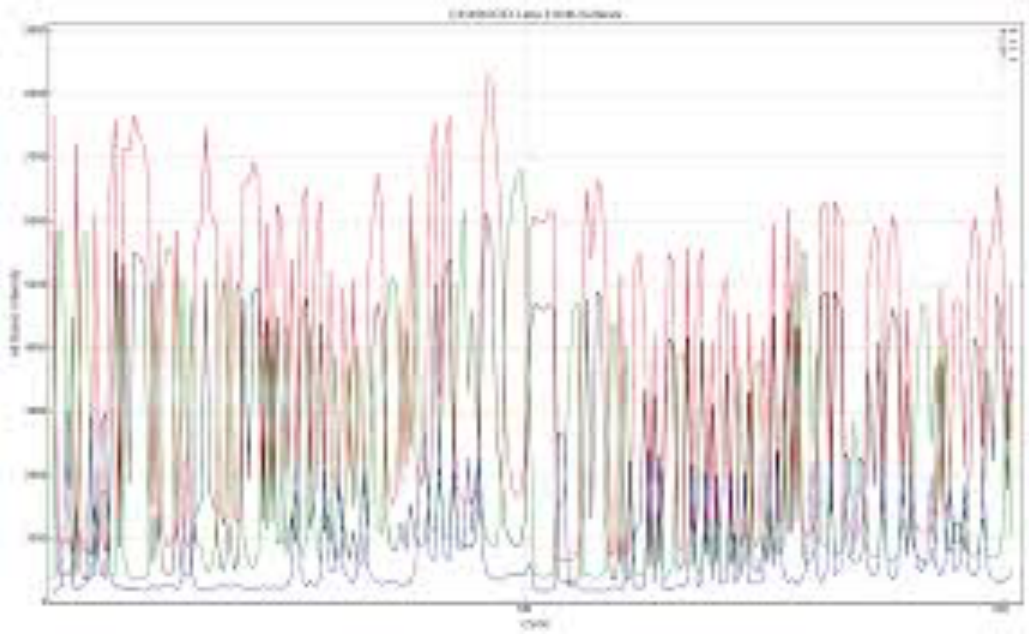


Nucleotide Diversity

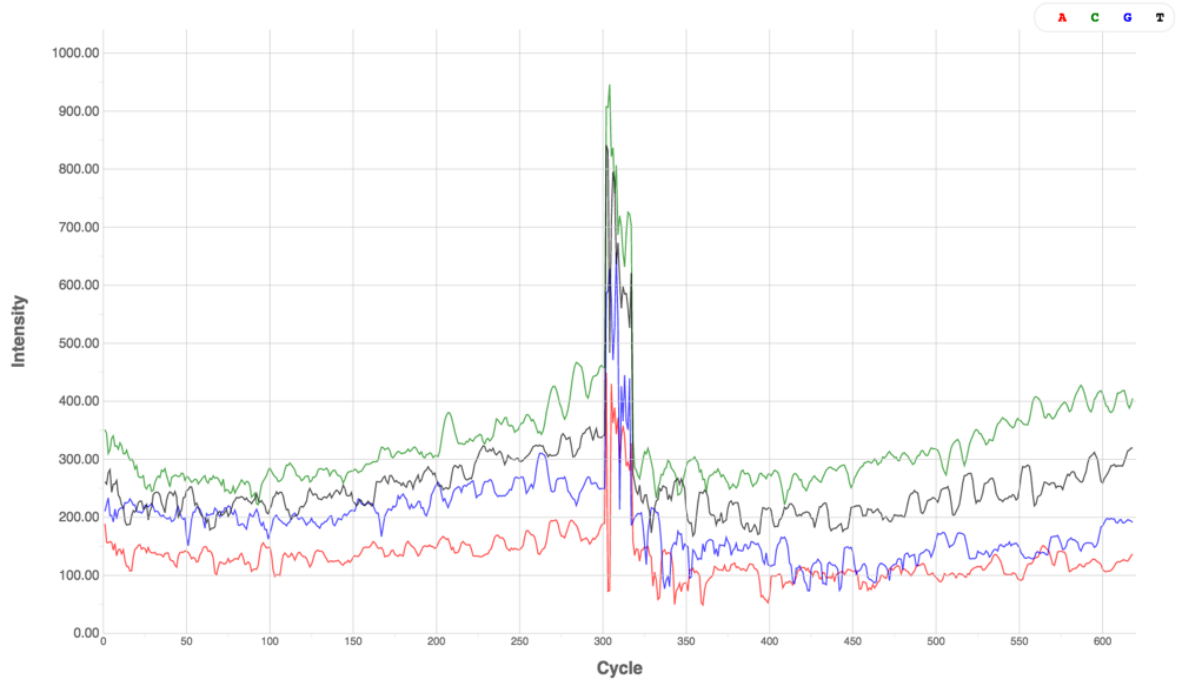
Once a sample library is converted to clusters on a flow cell, “nucleotide diversity” refers to the distribution of nucleotides across the flow cell at any given cycle. From the viewpoint of the instrument software, a high diversity library translates into analyzing images containing an even distribution of spots from 4 different color channels corresponding to the 4 nucleotide bases A, T, C & G. In contrast, an unbalanced nucleotide distribution or “low diversity library” means that for any given image, or to two bases are present at a high percentage

UCDAVIS Bioinformatics Core

Low Diversity Library
vs
High Diversity Library



Low Diversity Library



High Diversity Library

Ways to ensure nucleotide diversity

Appropriate nucleotide diversity and cluster density are important for high quality data. Low nucleotide diversity in combination with high cluster density will most-likely lead to poor data quality and/or low data yield.

Ways to avoid low nucleotide diversity

1. Sequence the sample at a 30-40% lower density
2. Spiking in at a 5-50% a nucleotide balanced library

   (such as PhiX, or better a shotgun library of a sample of interest)
3. Multiplex a high number of amplicon regions 12 or greater)
4. Build phase-shifted primers

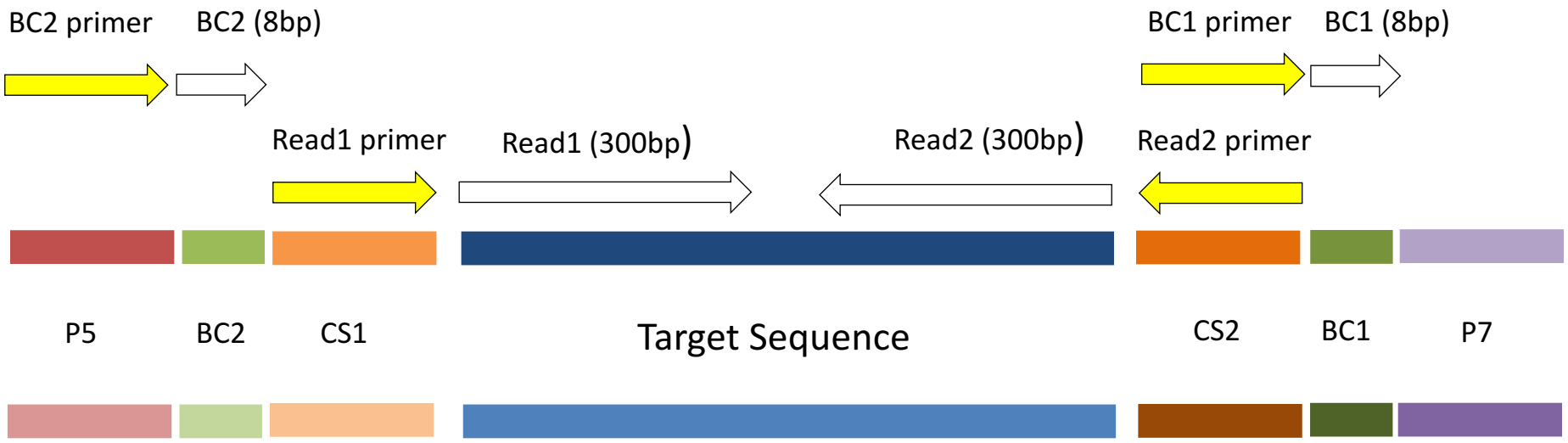
Note: Experience has shown, that 15% shotgun spike-in, plus phase-shifted primers and/or multiple target region typically yields good results.

UCDAVIS Bioinformatics Core

Illumina Sequencing

Requires custom sequencing primers to be added to the reaction

(Typical Illumina sequencing primers remain in the reaction for sequencing of PhiX or other shotgun library)



Read	Sequencing Primers
Read1 primer	CS1 - 5’ ACACTGACGACATGGTTCTACA 3’
Read2 primer	CS2 - 5’ TACGGTAGCAGAGACTTGGTCT 3’
BC1 primer	CS2rc - 5’ AGACCAAGTCTCTGCTACCGTA 3’
BC2 primer	Uses the P5 amplification primer

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
- Dada2
- dbcAmplicons (my software)

UCDAVIS Bioinformatics Core

dbcAmplicons