

Microbiome research and amplicon sequencing

Learning objectives



Describe pros and cons of different microbiome data types



Understand what makes a good amplicon marker



Describe how 16S rRNA gene amplicon data are generated

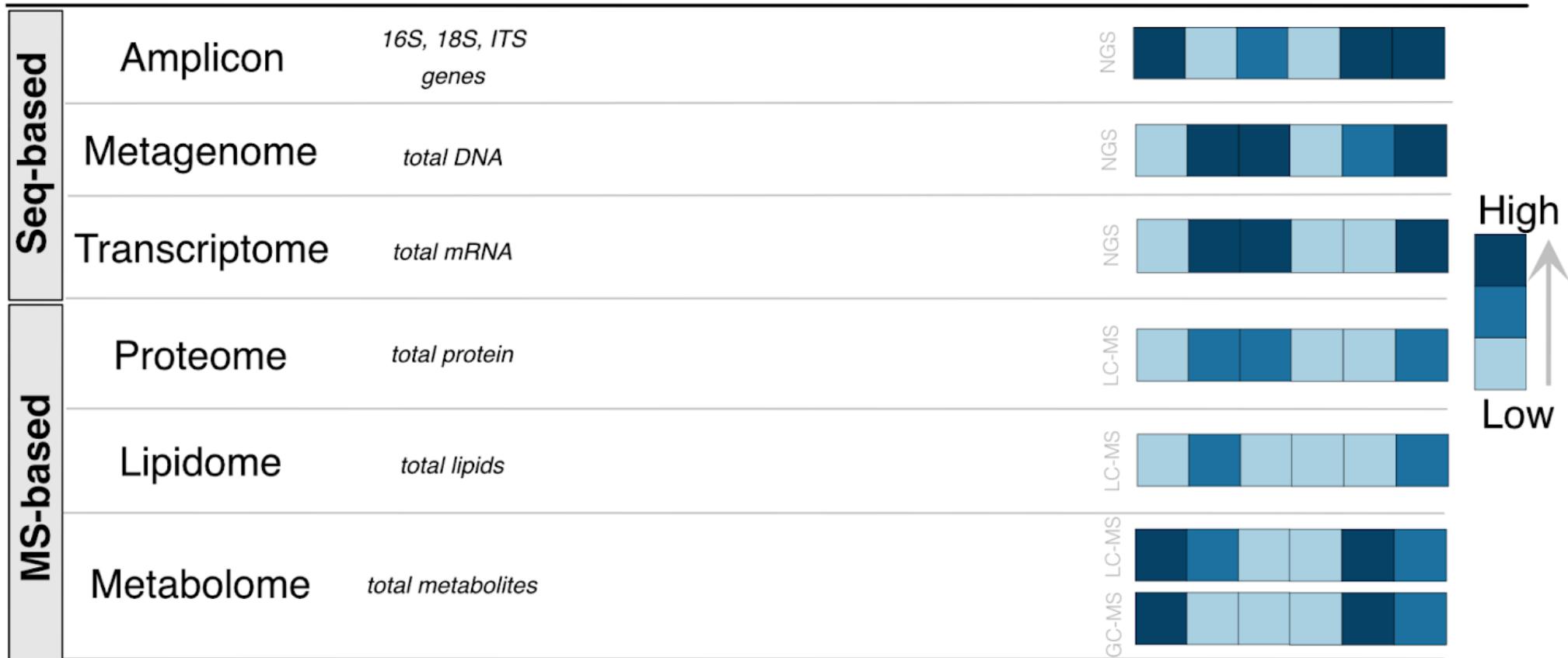
We study ...

structure,
diversity,
function,
communication



We sample

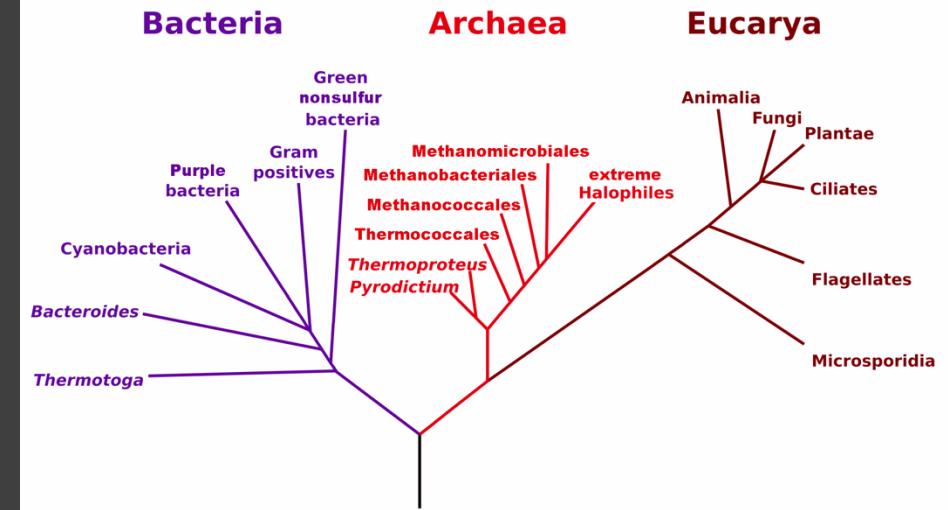
DNA,
RNA,
proteins,
metabolites



Throughput
Cost
Comp. power
Database infrastructure
Accessibility
Reproducibility

Some ancient history about 16S rRNA

Phylogenetic Tree of Life



Carl Woese
Three domains of life
using ribosomal RNA

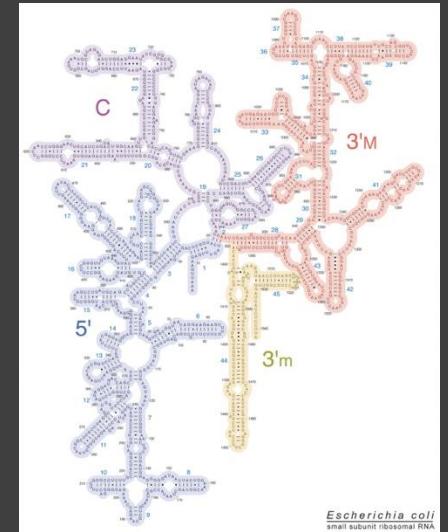
Norm Pace
Culture-independent 16S
rRNA

<https://www.theatlantic.com/science/archive/2017/07/the-man-who-blew-the-door-off-the-microbial-world/534246/>

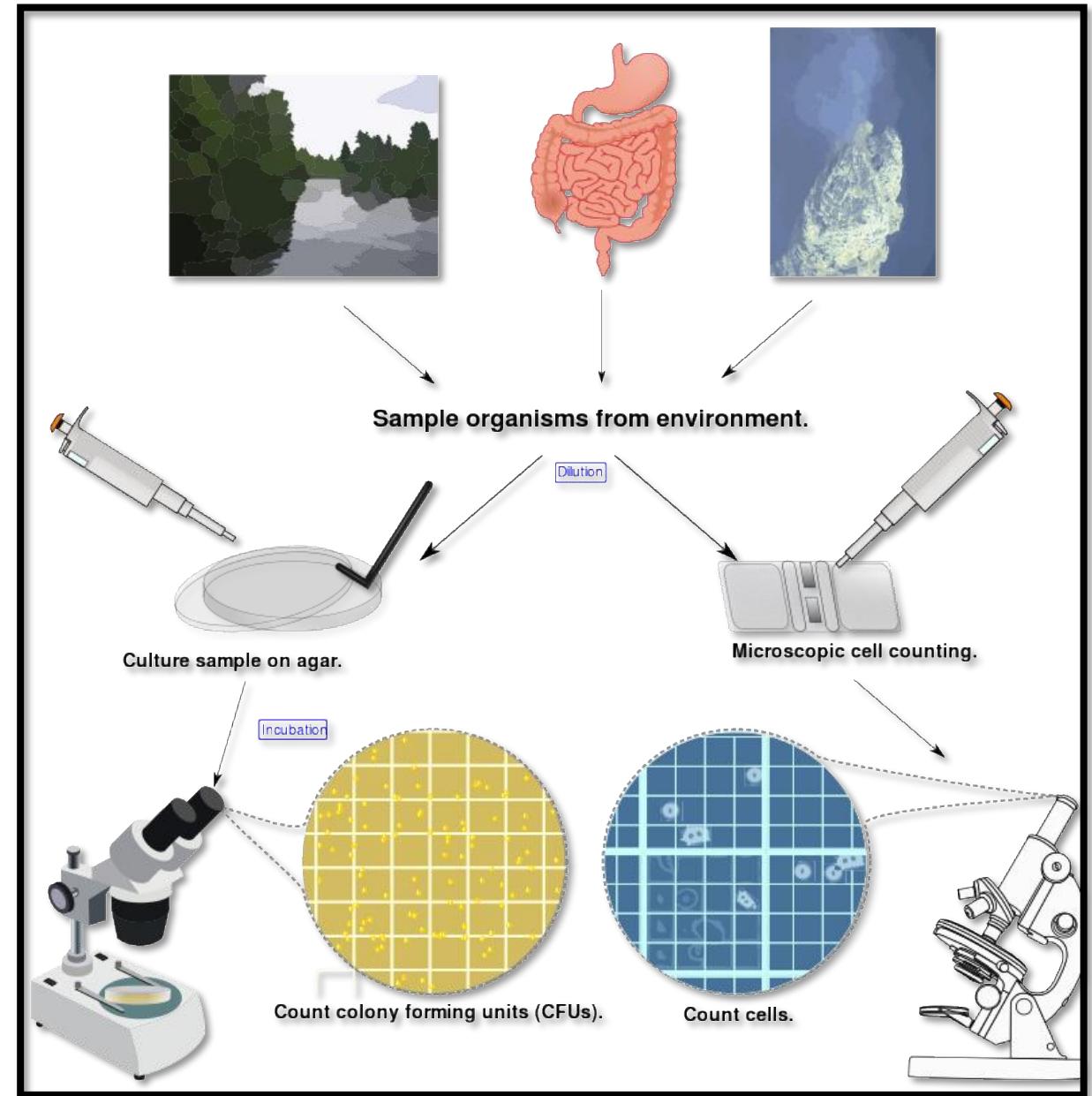


A Molecular View of Microbial Diversity and the Biosphere

Norman R. Pace
See all authors and affiliations
Science 02 May 1997;
Vol. 276, Issue 5313, pp. 734-740
DOI: 10.1126/science.276.5313.734

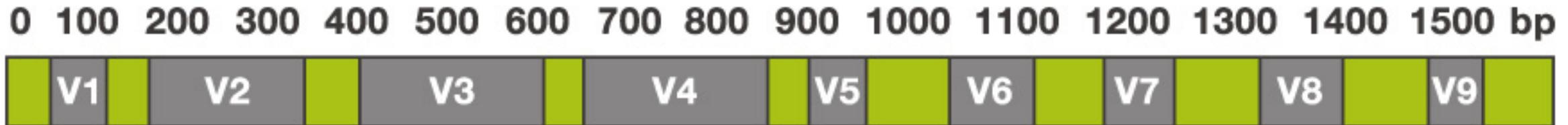


The great plate count anomaly



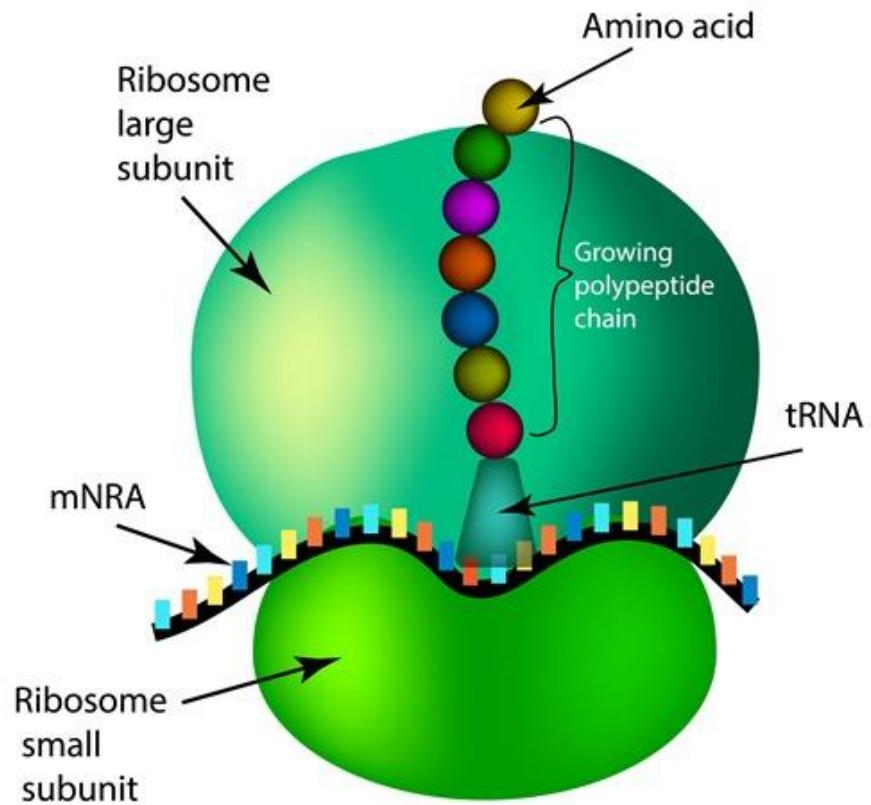
So what characteristics of a gene make it a good marker?

- Genes that are ubiquitous (e.g. important to the function of all living organisms)
- Genes that contains both:
 - **Conserved region** – common between all microbes of interest e.g. a gene region present in all bacteria and archaea (so universal primers can find it)
 - **Variable region** – different between taxa contained within your microbial group of interest e.g. a region within a bacterial marker gene that differentiates *E. coli* or *P. aeruginosa*



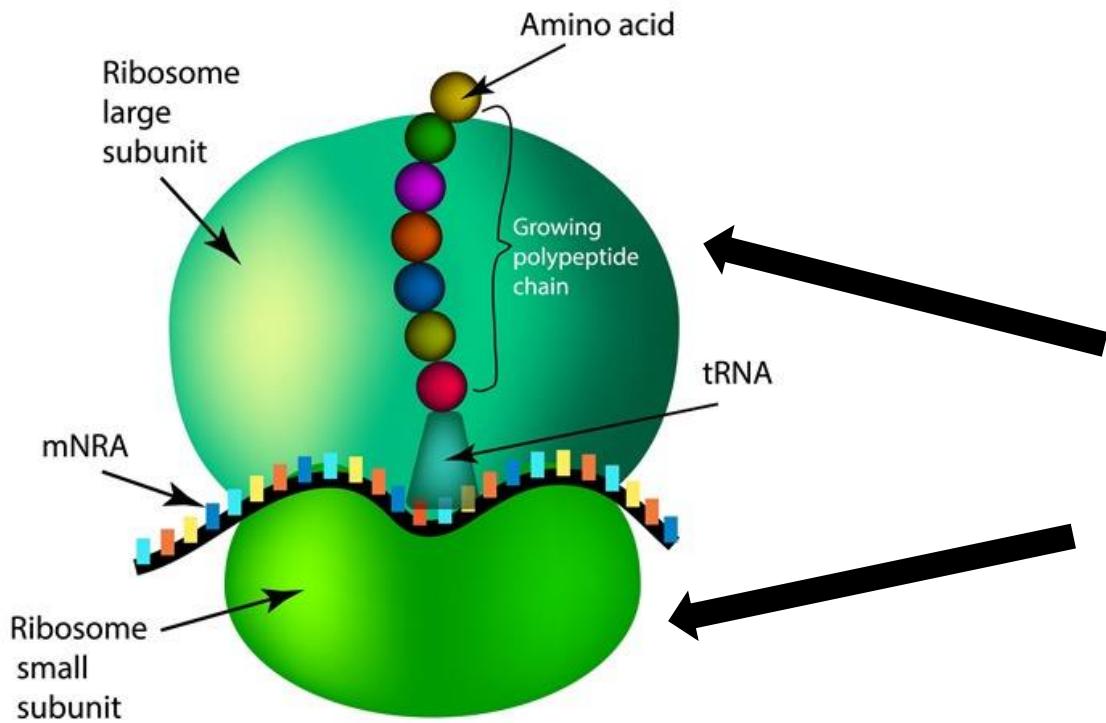
What is a ribosome?

Ribosome



Why would its genes be conserved?

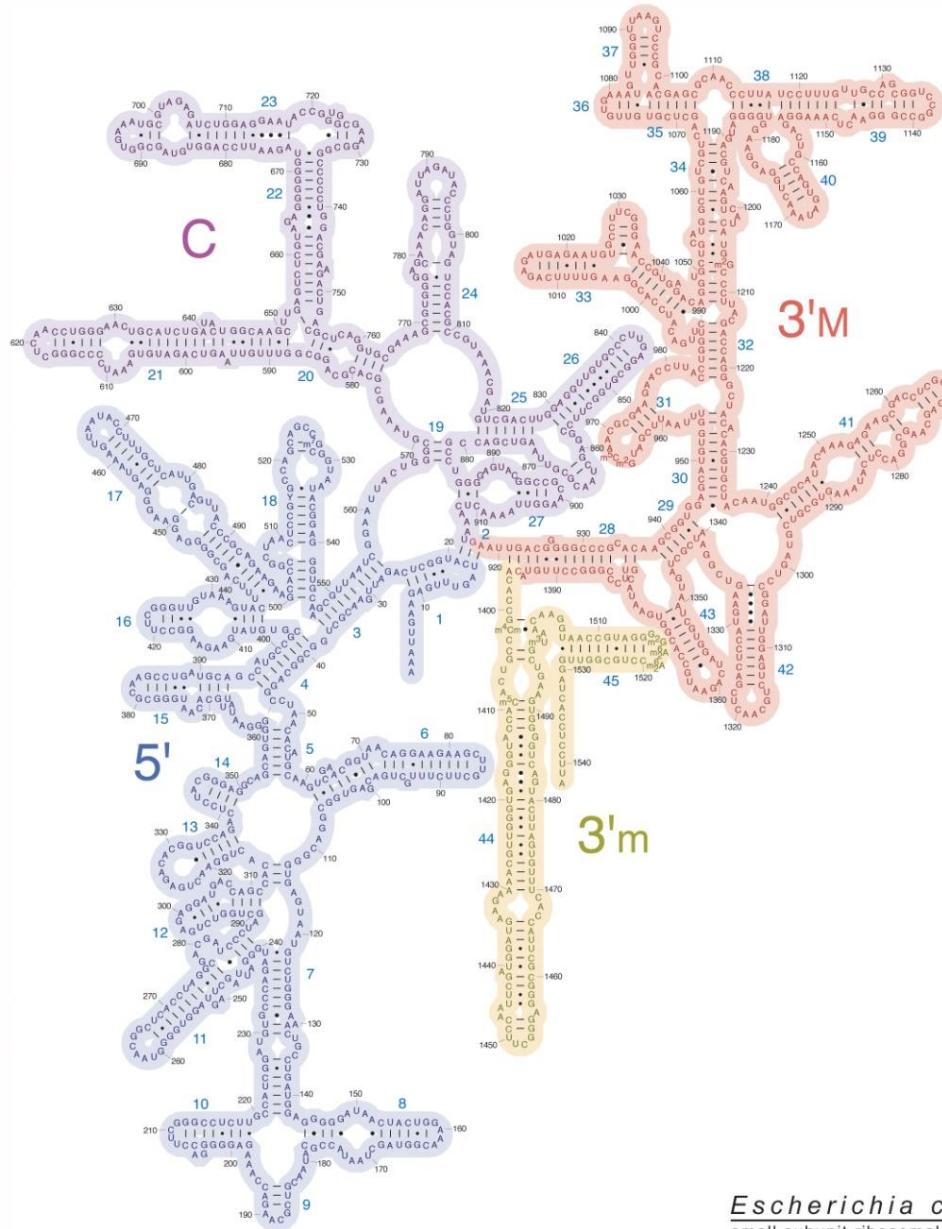
Ribosome



each subunit contains r-proteins
+ rRNA

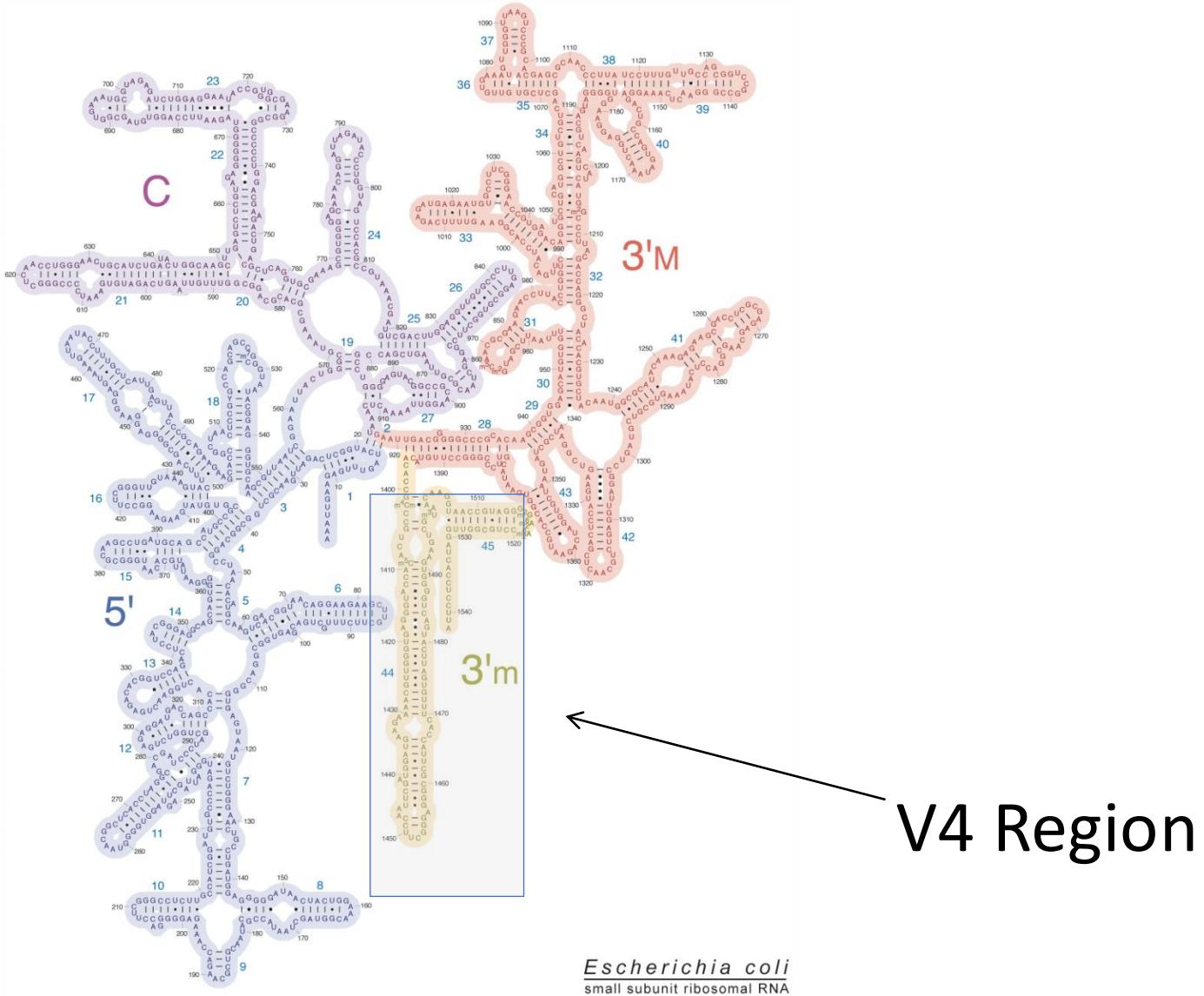
16S ribosomal RNA

- 16S rRNA – part of the small subunit in prokaryotic ribosomes
 - Has multiple functions:
 - Binds to the Shine-Dalgarno sequence, a ribosomal binding site in bacterial and archaeal mRNA that is involved in recruiting the ribosome to initiate translation
 - Acts as a scaffold for ribosomal proteins
 - Helps to stabilize correct protein synthesis



Escherichia coli
small subunit ribosomal RNA

16S rRNA amplicon sequencing



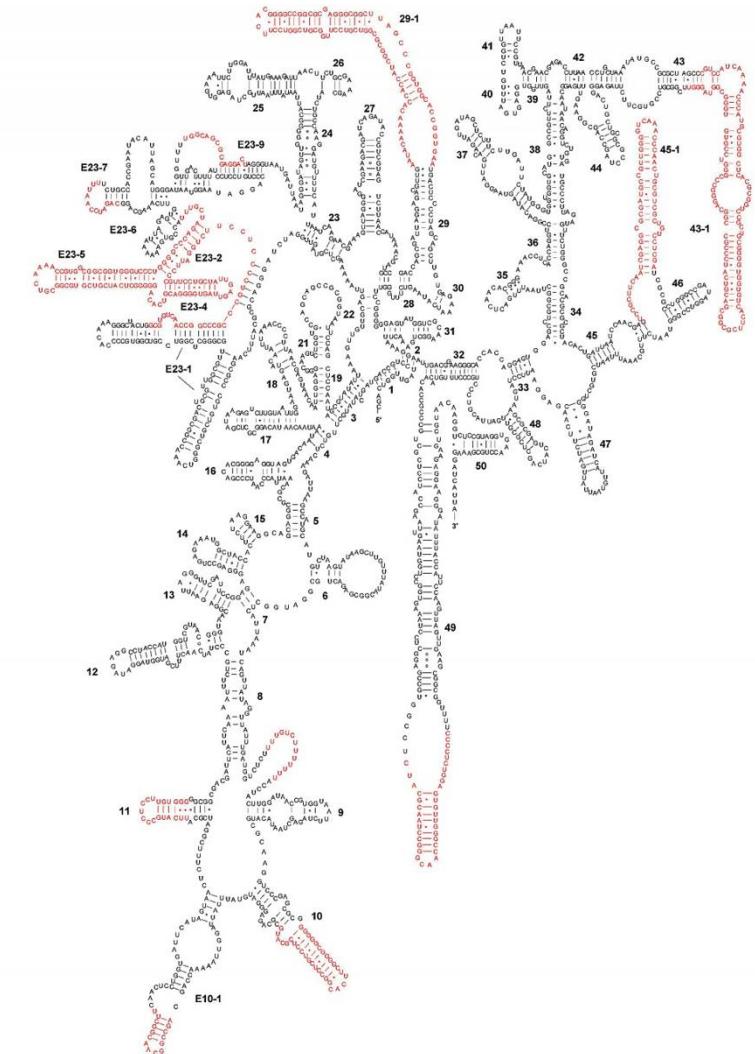


CONSERVED REGIONS: unspecific applications

VARIABLE REGIONS: group or species-specific applications

18S ribosomal RNA

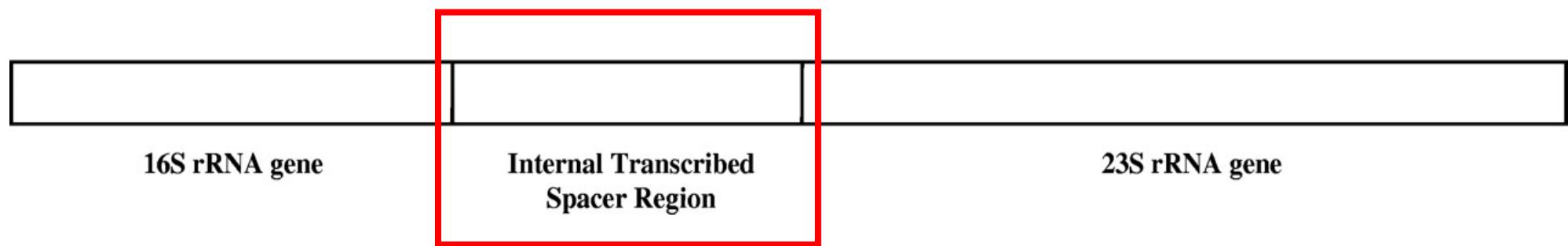
- Part of the small subunit in eukaryotic ribosomes
 - Important for maintaining structure of the small subunit
 - Active center of protein synthesis



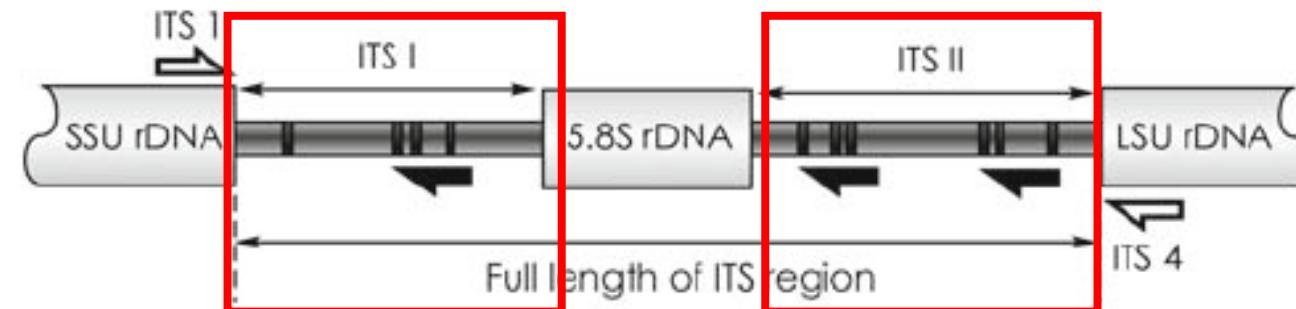
Internal transcribed spacer (ITS)

- Spacer DNA located between the small and large rRNA subunits genes (in the transcribed region)
- Most often used to identify fungi

Prokaryotes



Eukaryotes



- Amplicon data - 16S rRNA, 18S rRNA, ITS markers
 - Provides a snapshot of the taxonomic diversity
 - Inexpensive, can process a lot of samples cheaply
 - Works well with low biomass samples and samples with high amounts of host DNA (e.g. lung samples) because of PCR step
 - Not good for species level identification
 - Can be biased based on primer choice, sample preservation methods, and other technical artifacts
- Metagenomic data
 - Can also generate taxonomic profiles (using multiple target genes)
 - Can provide potential functional capacity of genome
 - Can provide species and maybe strain level taxonomy information
 - Expensive, requires a lot of DNA compared to amplicon methods
 - Aren't great methods for samples with high host DNA content
 - But, high-throughput, shallow coverage metagenomic sequencing is where the field is going

Metagenomics – sequencing DNA from environmental samples



Jo Handelsman



Jill Banfield

Cost per Human Genome

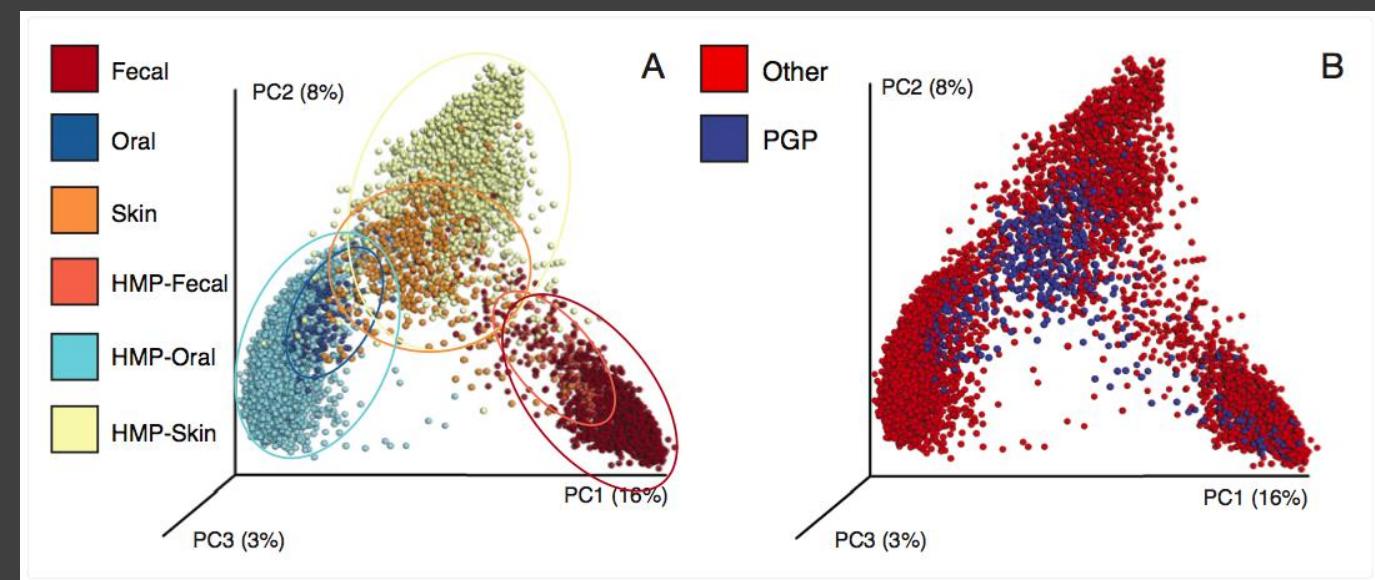


Computational approaches for analyzing 16S rRNA data

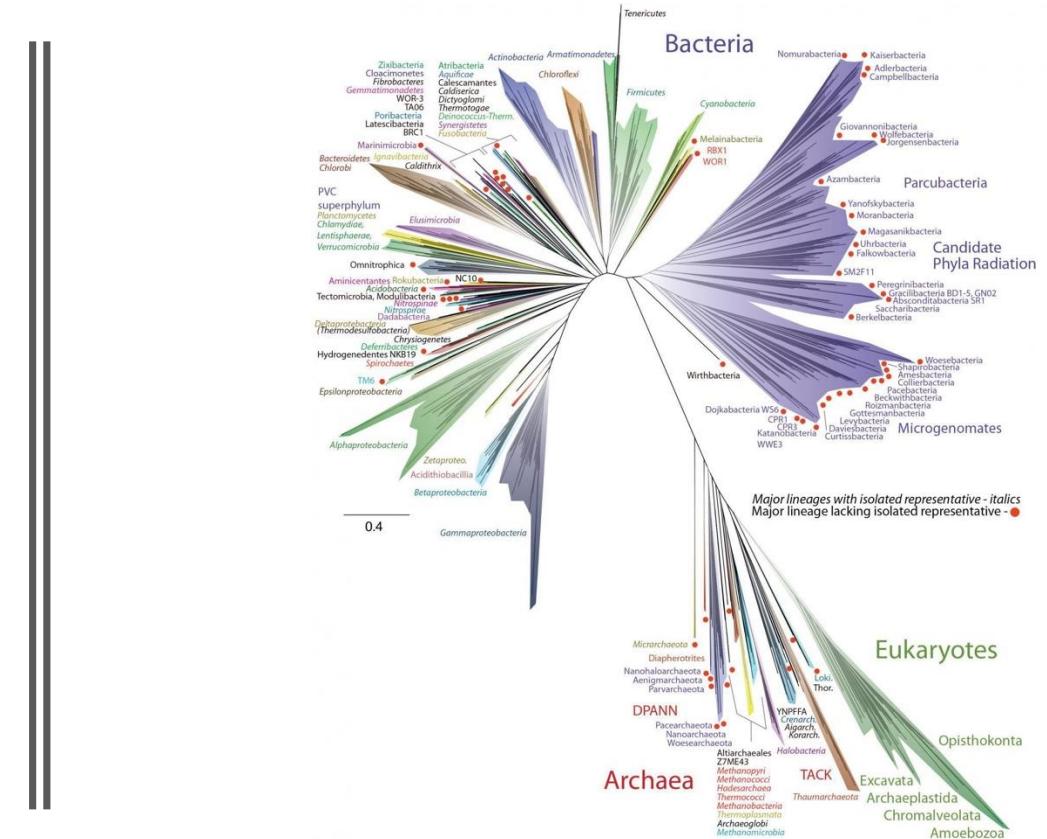
Rob Knight

Cathy Lozupone

Greg Caporaso



Laura Hug

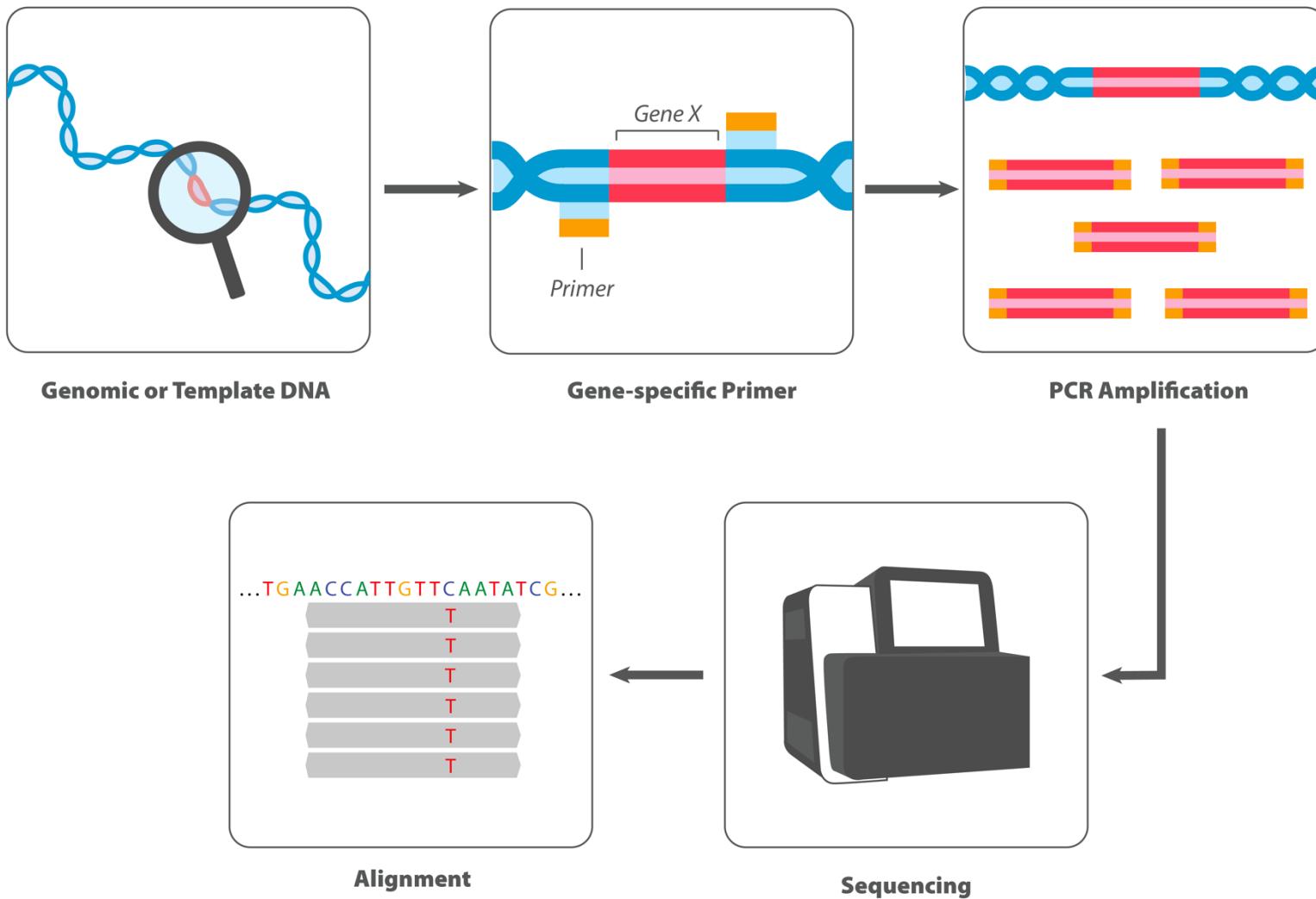


Human Microbiome Project

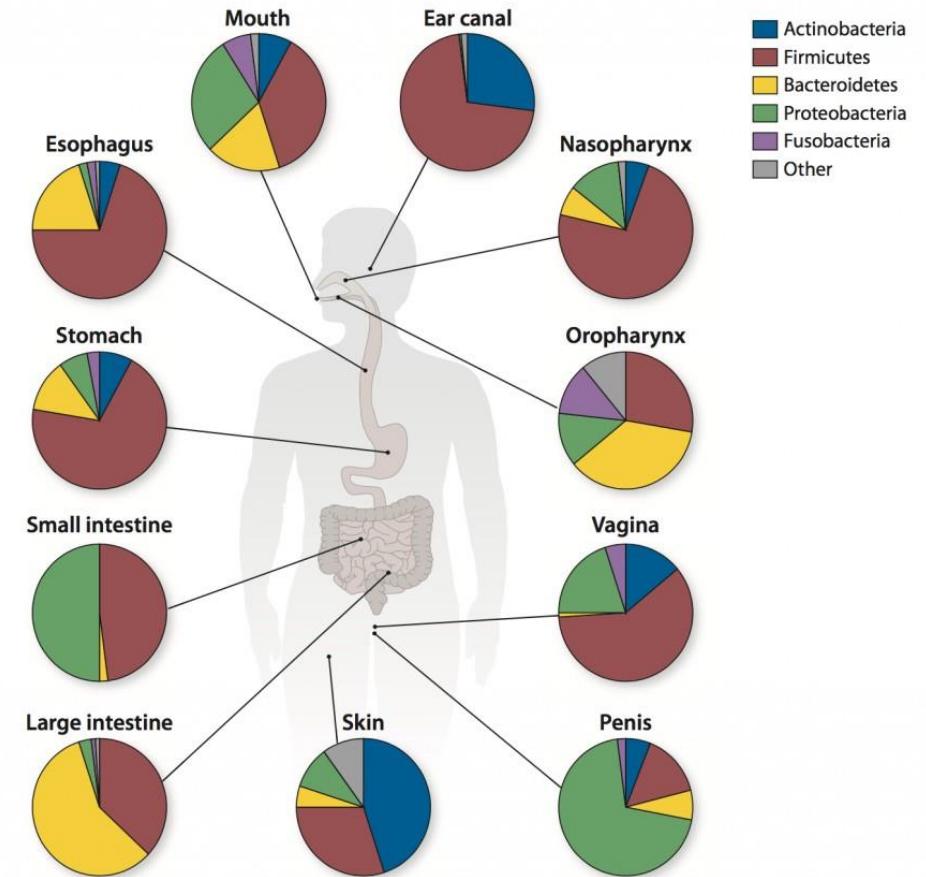
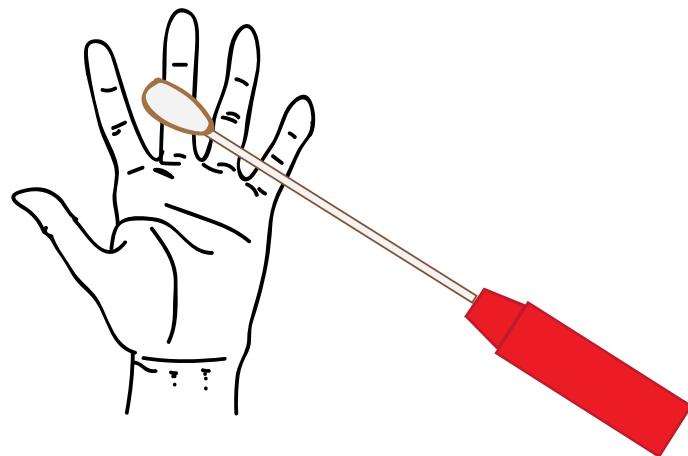


How do we sequence
DNA?

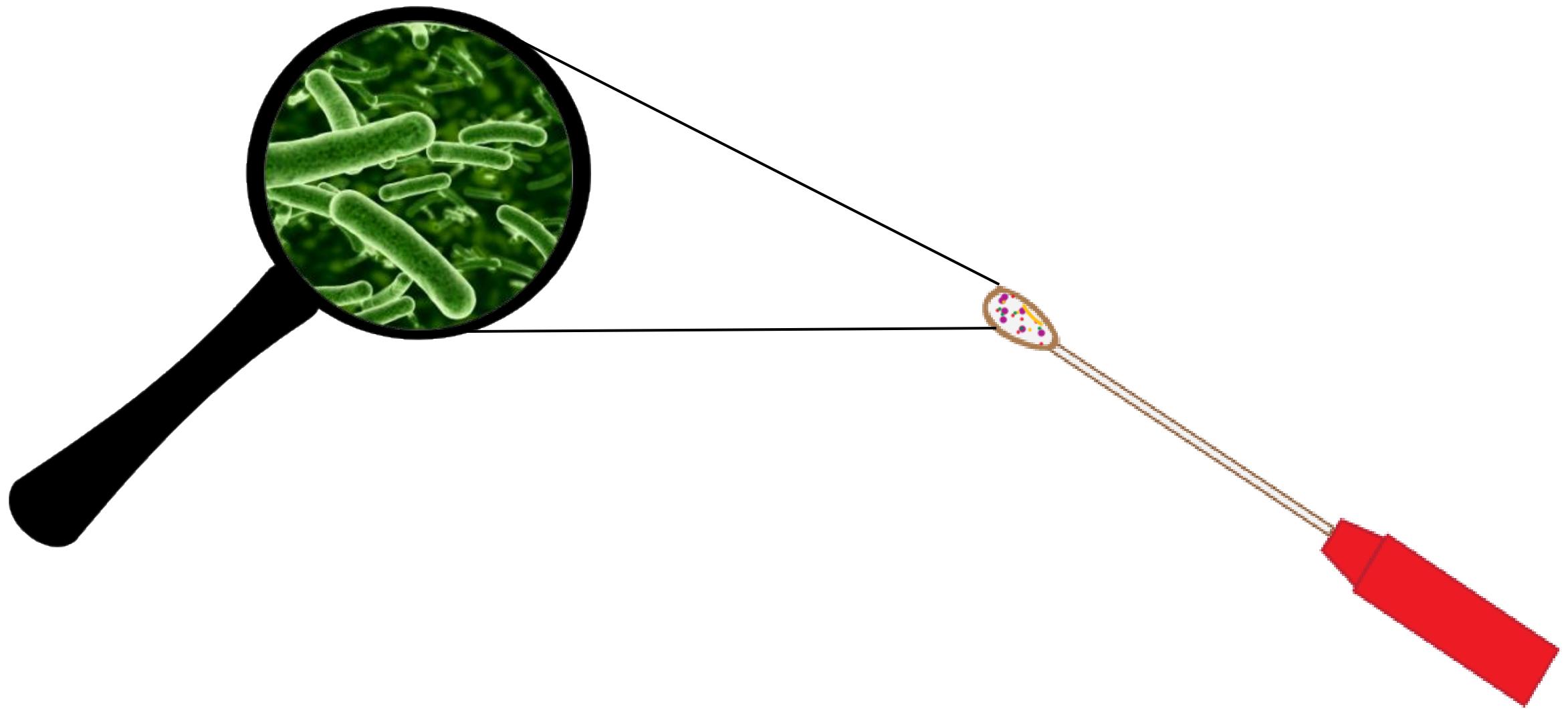
Amplicon sequencing



Sampling a microbiome

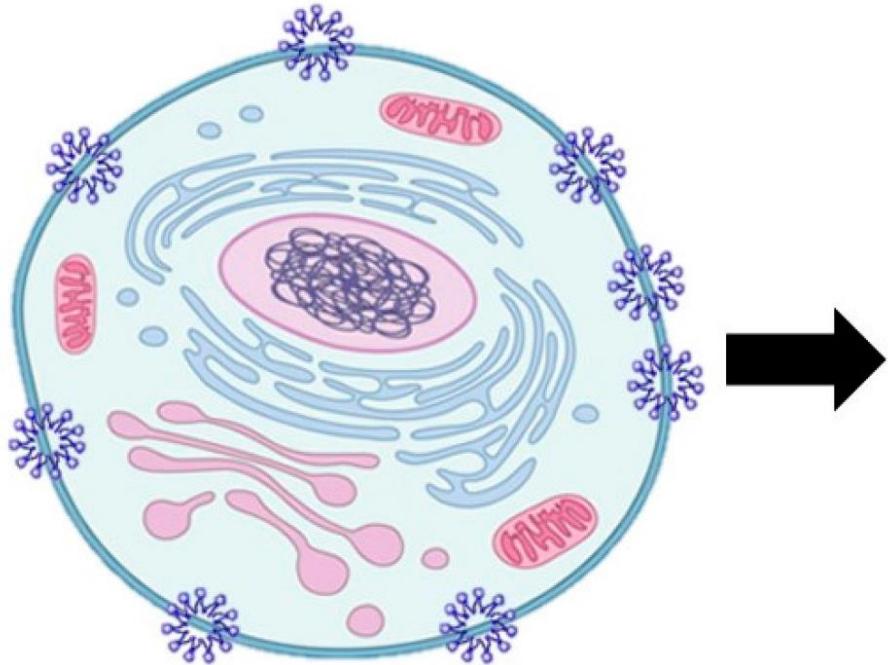


What kind of microbial environments are you interested in?

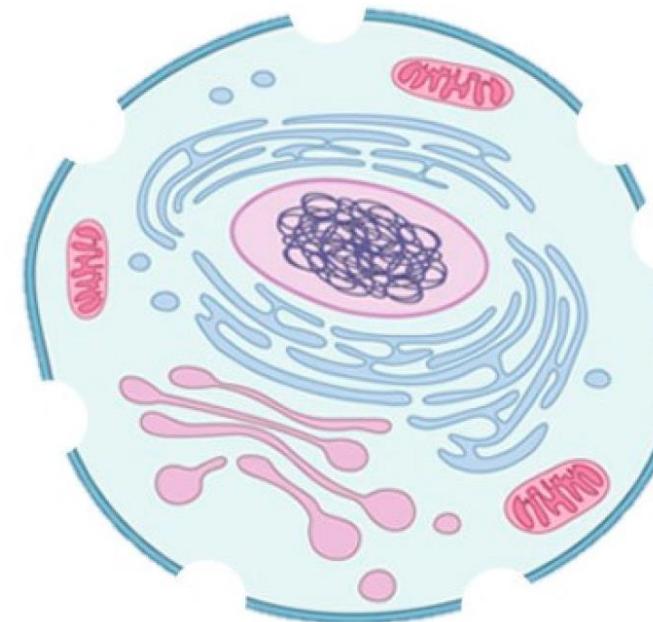




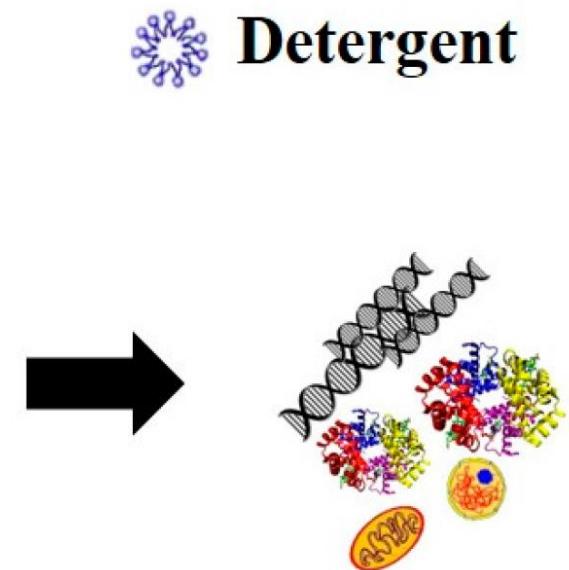
beads →



**Detergent reacts
with cell membrane**

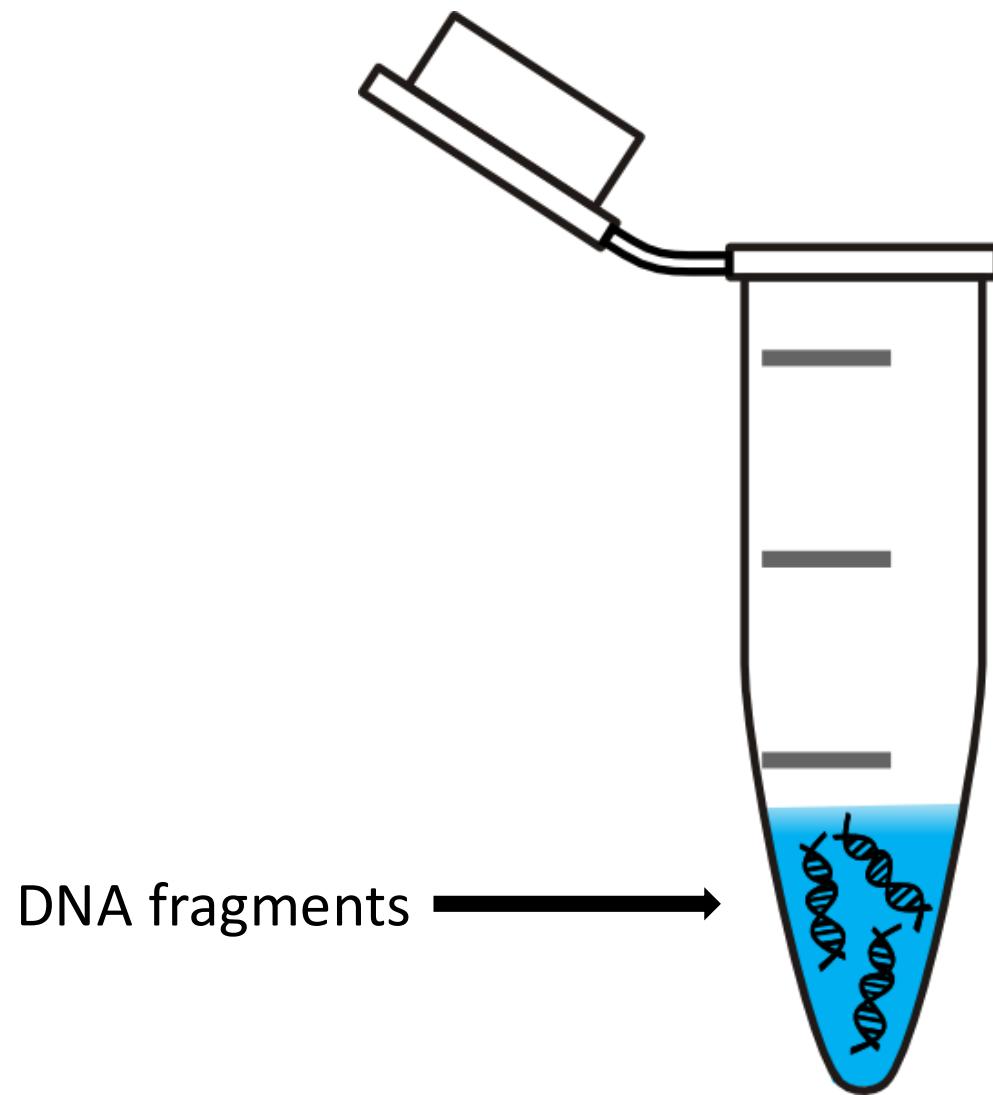


**Detergent destroys
the cell membrane**

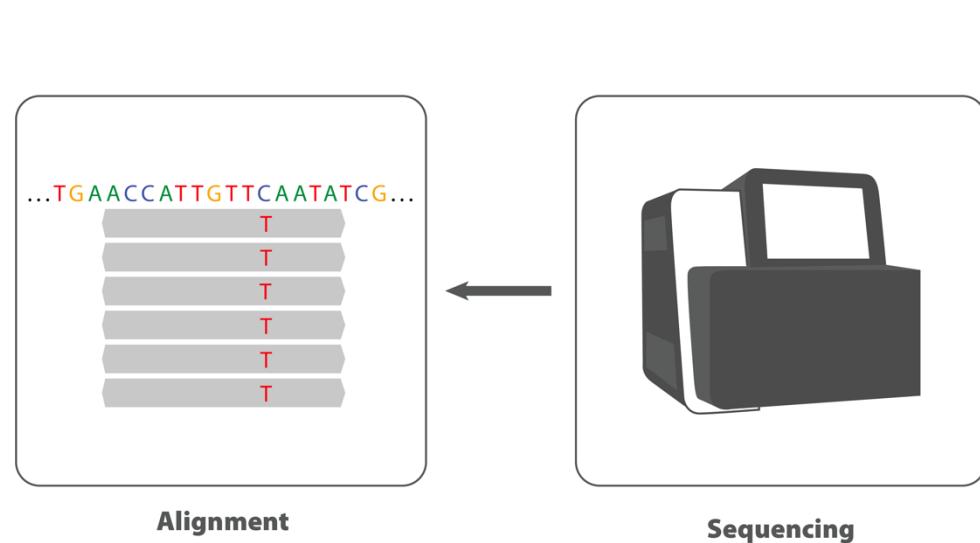
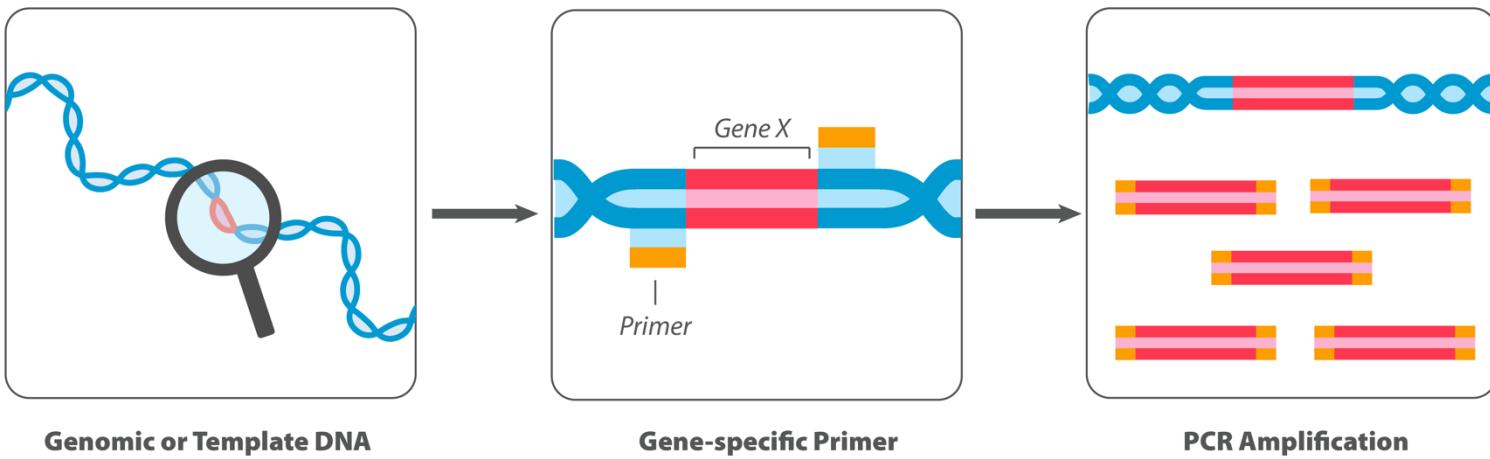


Detergent

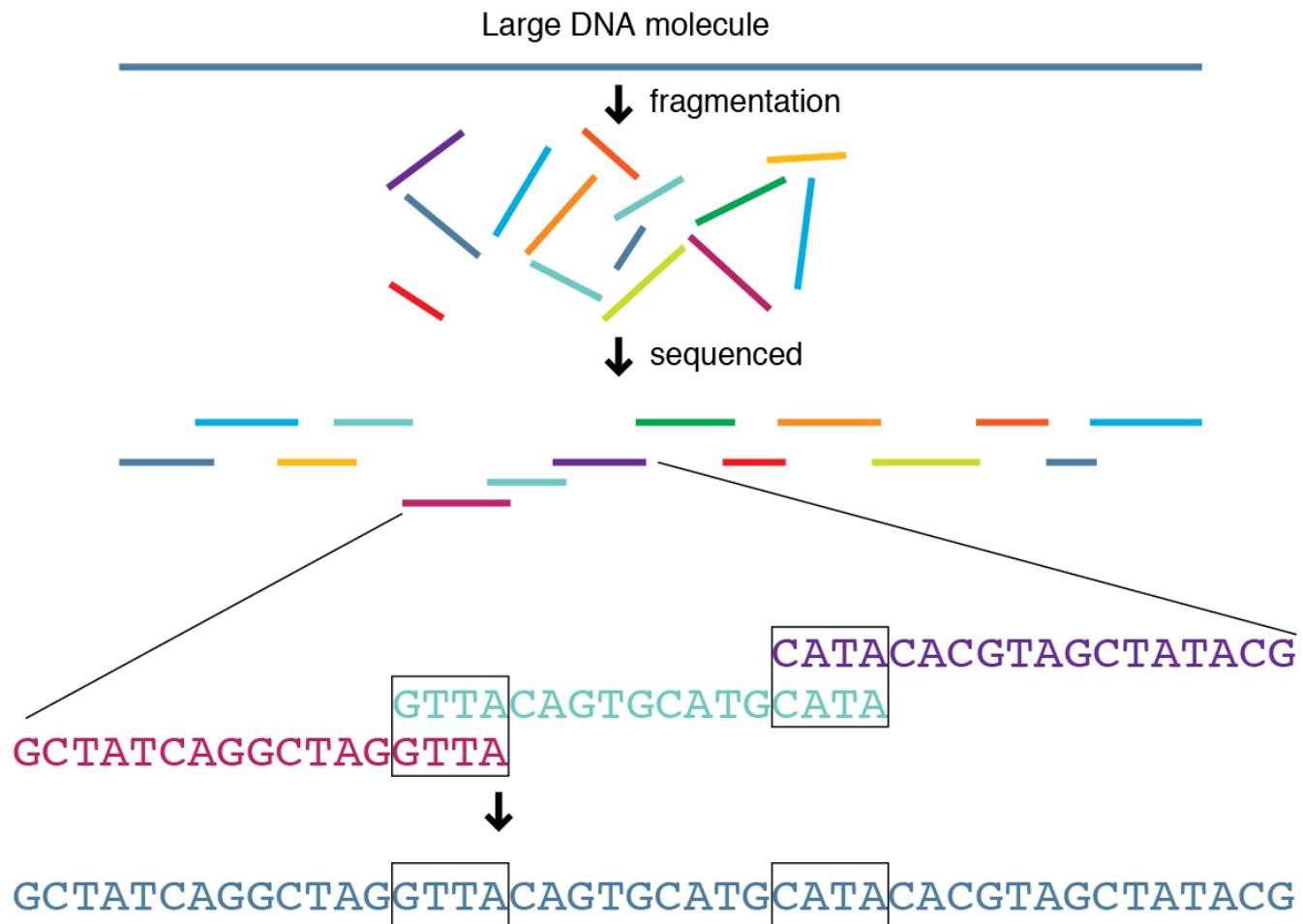
**Intracellular
components are
released**



Amplicon sequencing



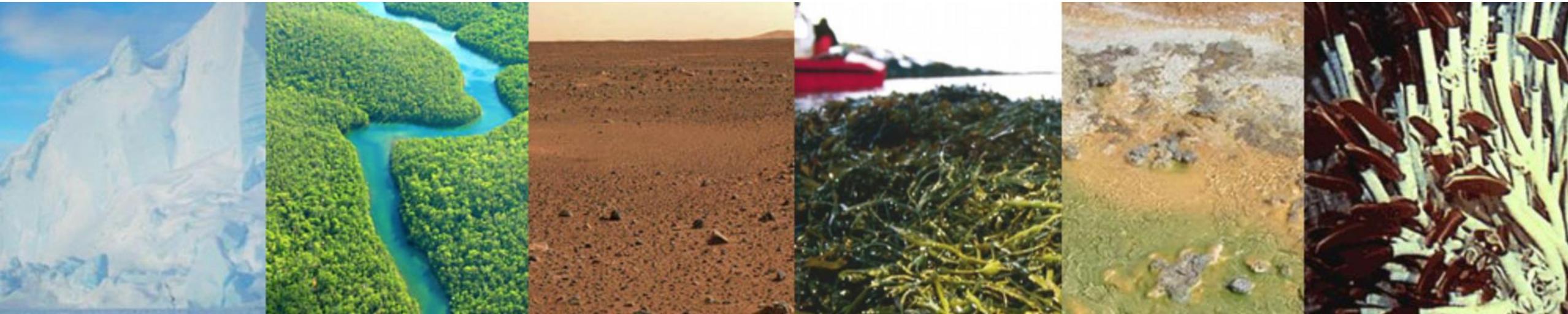
Metagenomic sequencing



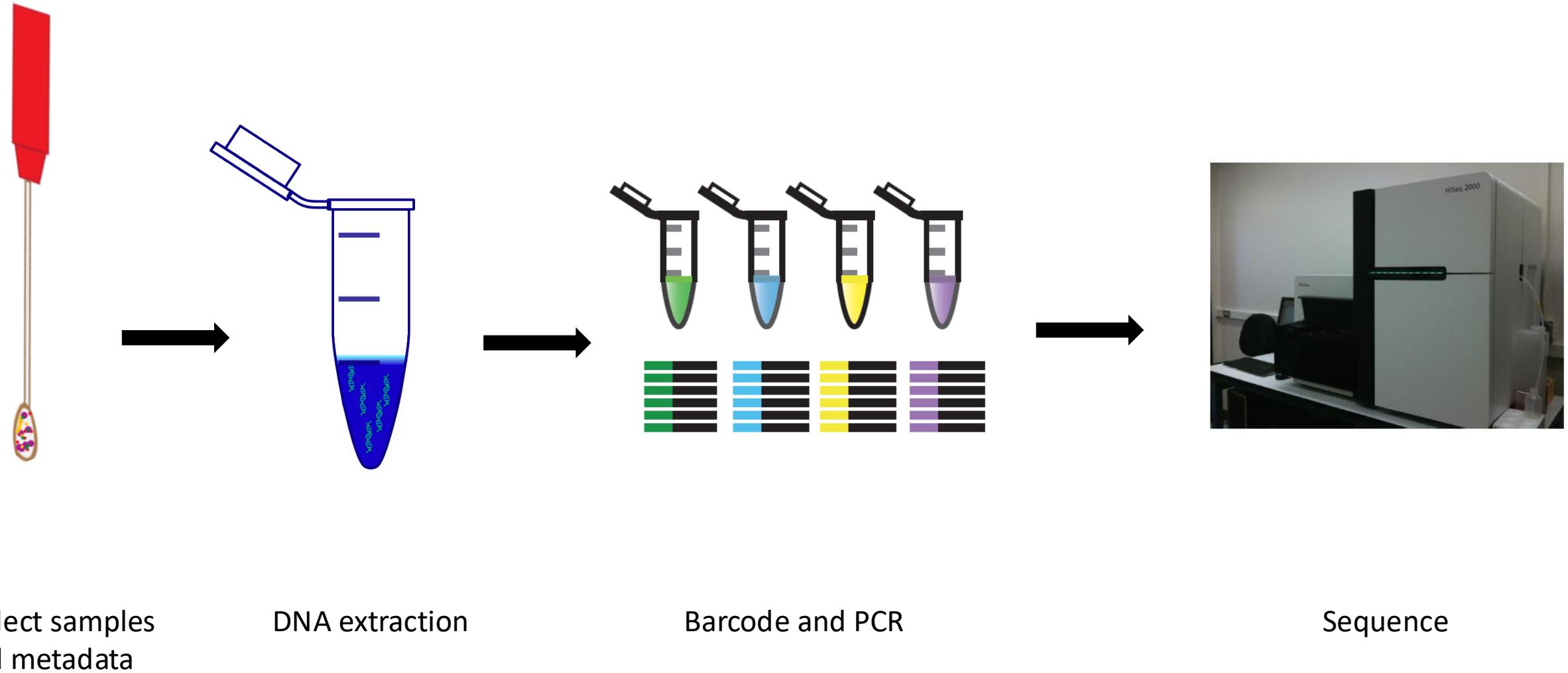
Amplicon sequencing deep dive



- <http://www.earthmicrobiome.org/protocols-and-standards/>

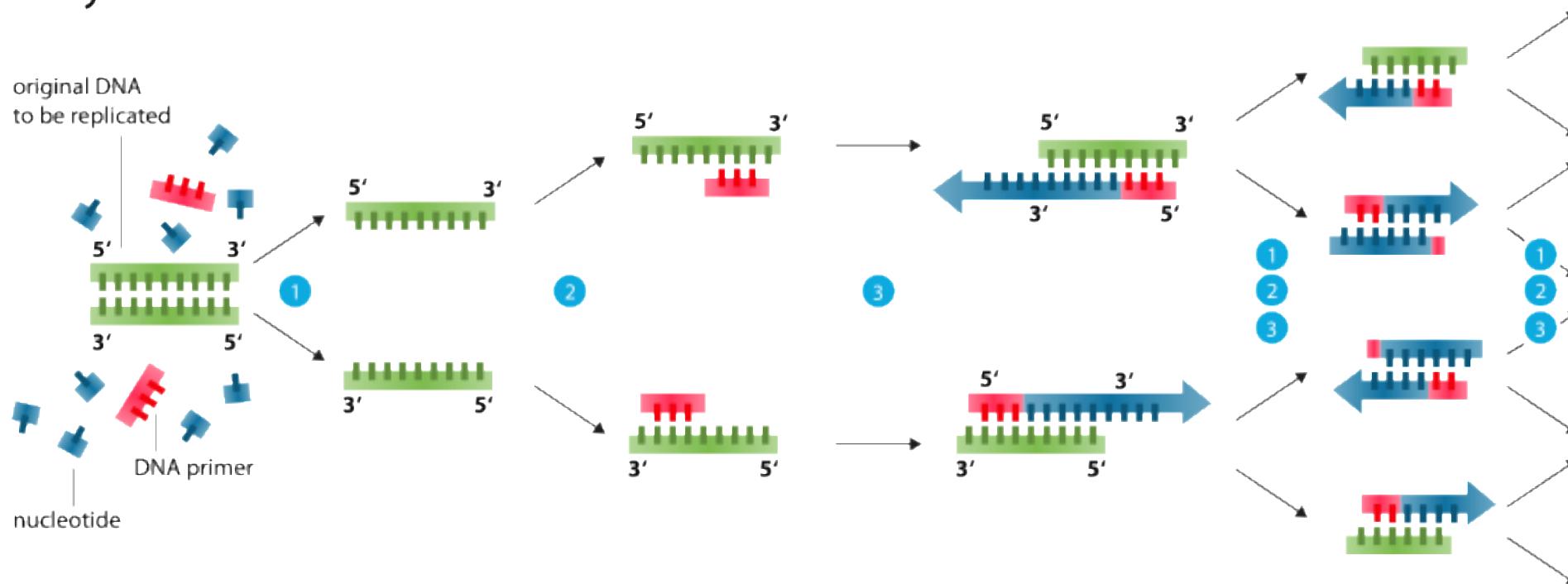


Generating amplicon data



Making DNA copies

Polymerase chain reaction - PCR



- 1 **Denaturation** at 94-96°C
- 2 **Annealing** at ~68°C
- 3 **Elongation** at ca. 72 °C

Adapter sequence

Barcode

Primer pad

Linker

515f

AATGATACGGCGACCACCGAGATCTACACGCT XXXXXXXXXXXX TATGGTAATT GT GTGYCAGCMGCCGCGGTAA

Plate Name(s)	Plate Number	Well Position	Sequence	Barcode
IL_515fBC_Jed_Arch_1	Plate 1	A1	AATGATACGGCGACCACCGAGATCTACACGCTAGCCTCGTCGC TATGGTAATTGTGTGYCAGCMGCCGCGGTAA	AGCCTTCGTCGC
IL_515fBC_Jed_Arch_1	Plate 1	A2	AATGATACGGCGACCACCGAGATCTACACGCTTCCATACCGGAATATGTAATTGTGTGYCAGCMGCCGCGGTAA	TCCATACCGGAA
IL_515fBC_Jed_Arch_1	Plate 1	A3	AATGATACGGCGACCACCGAGATCTACACGCTAGCCCTGCTACATATGTAATTGTGTGYCAGCMGCCGCGGTAA	AGCCCTGCTACA
IL_515fBC_Jed_Arch_1	Plate 1	A4	AATGATACGGCGACCACCGAGATCTACACGCTCTAACGGTCCATATGTAATTGTGTGYCAGCMGCCGCGGTAA	CCTAACGGTCCA
IL_515fBC_Jed_Arch_1	Plate 1	A5	AATGATACGGCGACCACCGAGATCTACACGCTCGCGCCTAAACTATGTAATTGTGTGYCAGCMGCCGCGGTAA	CGCGCCTAAC
IL_515fBC_Jed_Arch_1	Plate 1	A6	AATGATACGGCGACCACCGAGATCTACACGCTTATGGTACCCAGTATGTAATTGTGTGYCAGCMGCCGCGGTAA	TATGGTACCCAG
IL_515fBC_Jed_Arch_1	Plate 1	A7	AATGATACGGCGACCACCGAGATCTACACGCTTACAATATCTGTTATGTAATTGTGTGYCAGCMGCCGCGGTAA	TACAATATCTGT
IL_515fBC_Jed_Arch_1	Plate 1	A8	AATGATACGGCGACCACCGAGATCTACACGCTAATTAGGTAGGTATGTAATTGTGTGYCAGCMGCCGCGGTAA	AATTAGGTAGG
IL_515fBC_Jed_Arch_1	Plate 1	A9	AATGATACGGCGACCACCGAGATCTACACGCTGACTCAACCAGTTATGTAATTGTGTGYCAGCMGCCGCGGTAA	GACTCAACCACT
IL_515fBC_Jed_Arch_1	Plate 1	A10	AATGATACGGCGACCACCGAGATCTACACGCTGCCCTACGTCGTATGTAATTGTGTGYCAGCMGCCGCGGTAA	GCCTCTACGTCG

Adapter primer

Barcode

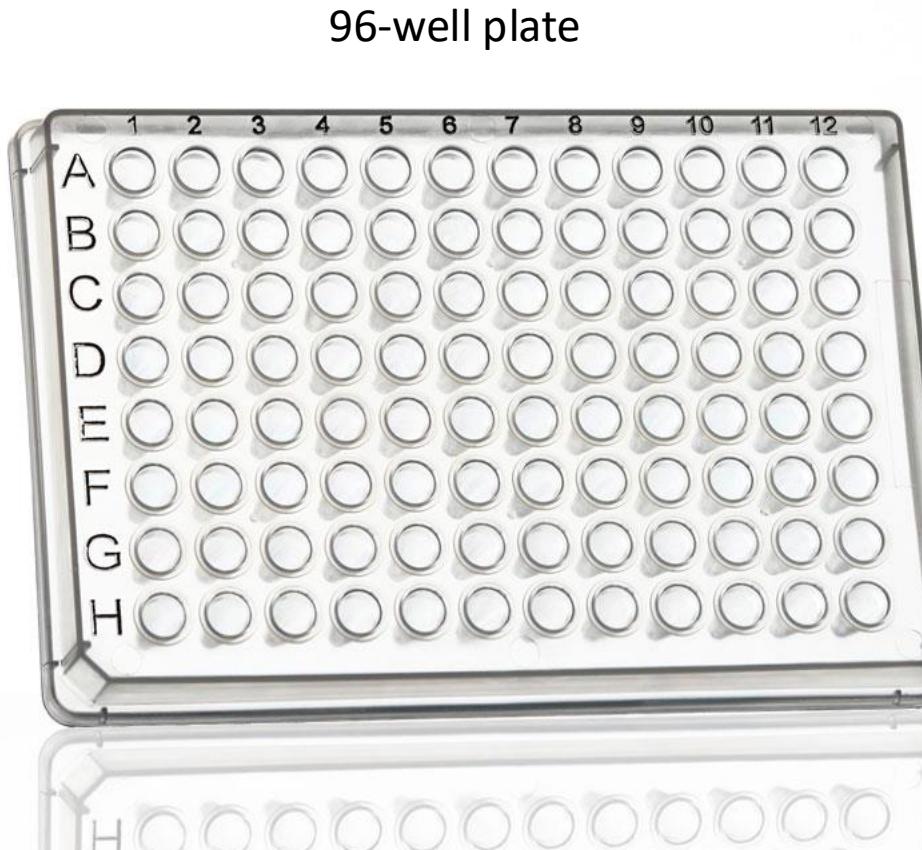
Primer pad

Linker

515f

AATGATAACGGCGACCACCGAGATCTACACGCT TATGGTAATT GT GTGCAGCMGCCGCGTAA

Plate Name(s)	Plate Number	Well Position	Sequence	Barcode
IL_515fBC_Jed_Arch_1	Plate 1	A1	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A2	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A3	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A4	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A5	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A6	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A7	AATGATA	CCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A8	AATGATA	GCCGCGGTAA
IL_515fBC_Jed_Arch_1	Plate 1	A9	AATGATAACGGCGACCACCGAGATCTACACGCT TATGGTAATT GT GTGCAGCMGCCGCGTAA	GACTCAACCACT
IL_515fBC_Jed_Arch_1	Plate 1	A10	AATGATAACGGCGACCACCGAGATCTACACGCTGCCTCTACGTCGTATGGTAATT GT GTGCAGCMGCCGCGTAA	GCCTCTACGTCG



Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex

Micah Hamady¹, Jeffrey J Walker², J Kirk Harris³,
Nicholas J Gold² & Rob Knight⁴

We constructed error-correcting DNA barcodes that allow one run of a massively parallel pyrosequencer to process up to 1,544 samples simultaneously. Using these barcodes we processed bacterial 16S rRNA gene sequences representing microbial communities in 286 environmental samples, corrected 92% of sample assignment errors, and thus characterized nearly as many 16S rRNA genes as have been sequenced to date by Sanger sequencing.

Illumina Error Rate = ~0.1%
OR
1 in 1,000 bases

TCCATA**CC**GGAA
TCCATA**CC**GGAA
TCCATA**CC**GGAA
TCCATA**G**CGGAA
TCCATA**CC**GGAA
TCCATA**CC**GGAA
TCCATA**CC**GGAA
 Consensus
TCCATA**CC**GGAA

Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton

Amy Apprill^{1,*}, Sean McNally^{1,2}, Rachel Parsons², Laura Weber¹

¹Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

²Bermuda Institute of Ocean Sciences, Ferry Reach, St. George's GE01, Bermuda

Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys

William Walters,^a Embriette R. Hyde,^b Donna Berg-Lyons,^c Gail Ackermann,^b Greg Humphrey,^b Alma Parada,^d Jack A. Gilbert,^{e,f,g,h,i} Janet K. Jansson,^j J. Gregory Caporaso,^k Jed A. Fuhrman,^d Amy Apprill,^l Rob Knight^{b,m}

Discussion questions

- Based on your specific research questions, which do you think might be better for your project – amplicon or metagenomic sequencing?
- What types (if any) of ‘omics data are you familiar with? (e.g. amplicon sequencing, metagenomics, metatranscriptomics, metabolomics) What types of questions did you answer with them?

What happens when you hand off a library to the sequencing core?

Jessica Henley, Director
CSU NGS core

Library Preparation



For clustering:

Libraries must have P5 and P7 binding regions on either end of a library

For sequencing:

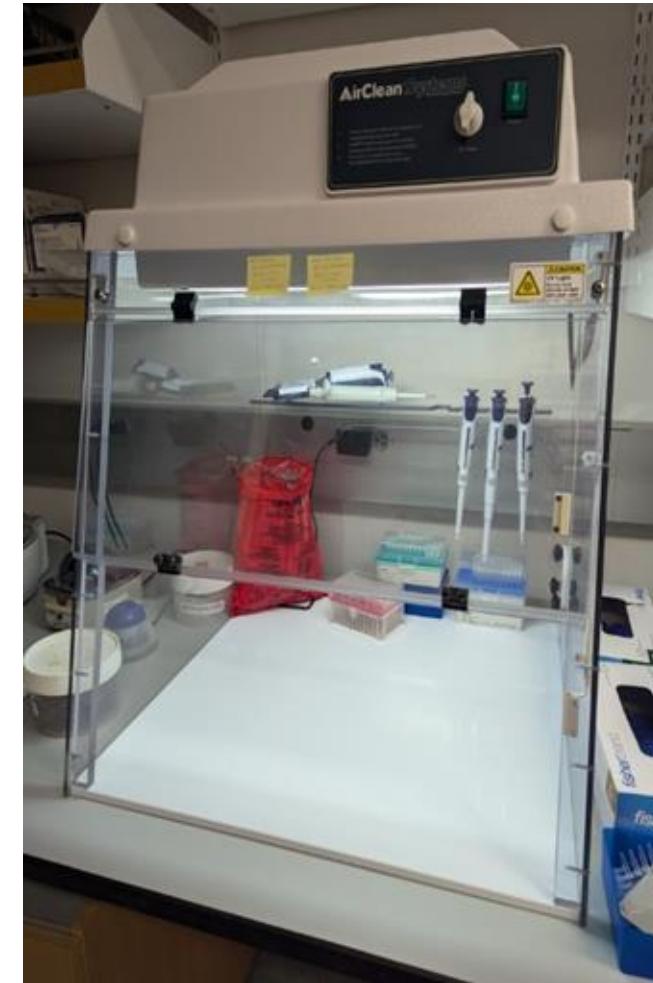
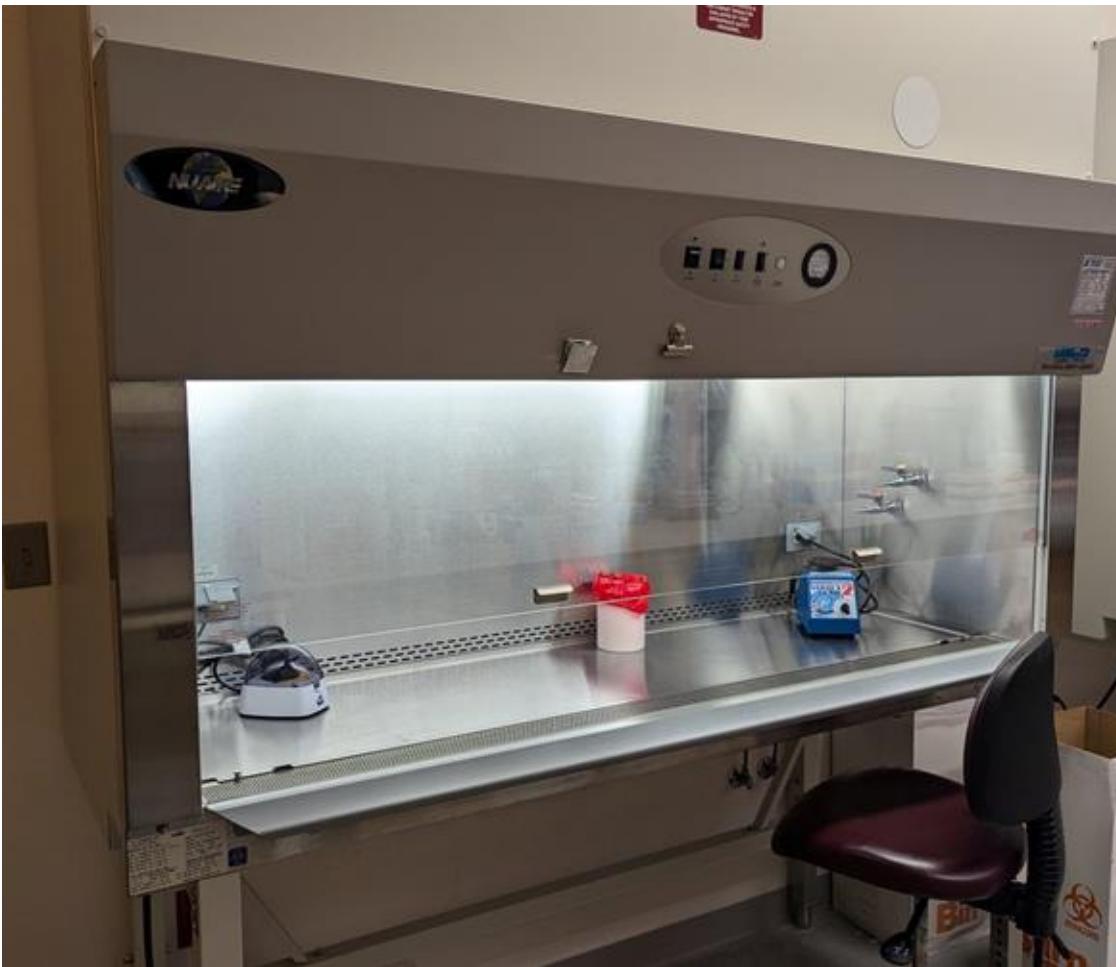
Libraries must have sequencing primer binding regions

For mixing samples:

Libraries must have a unique index or barcodes sequence

Library Preparation Steps

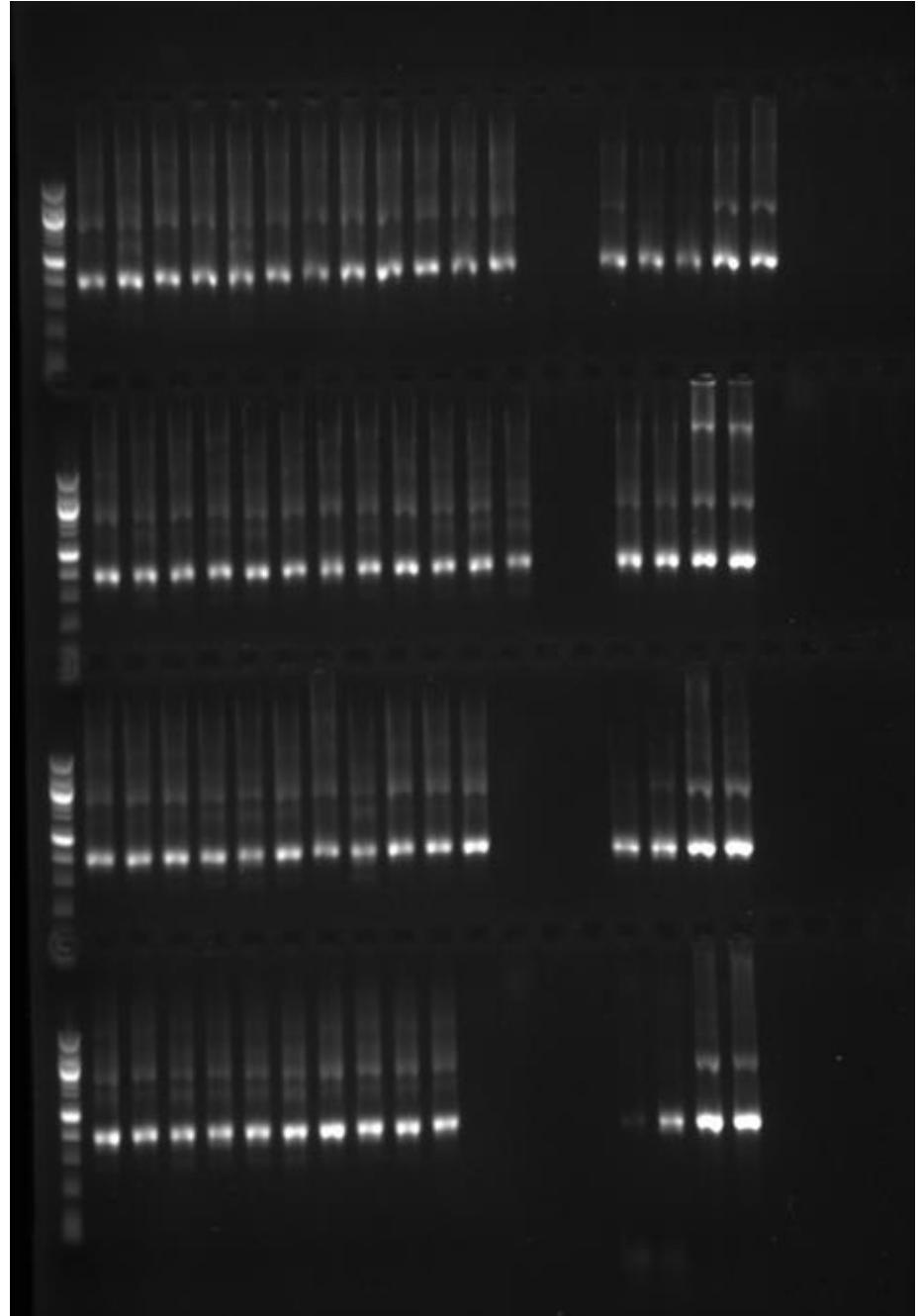
1. PCR with custom barcoded target gene primers



paper: Sourdough starters, Clark et al

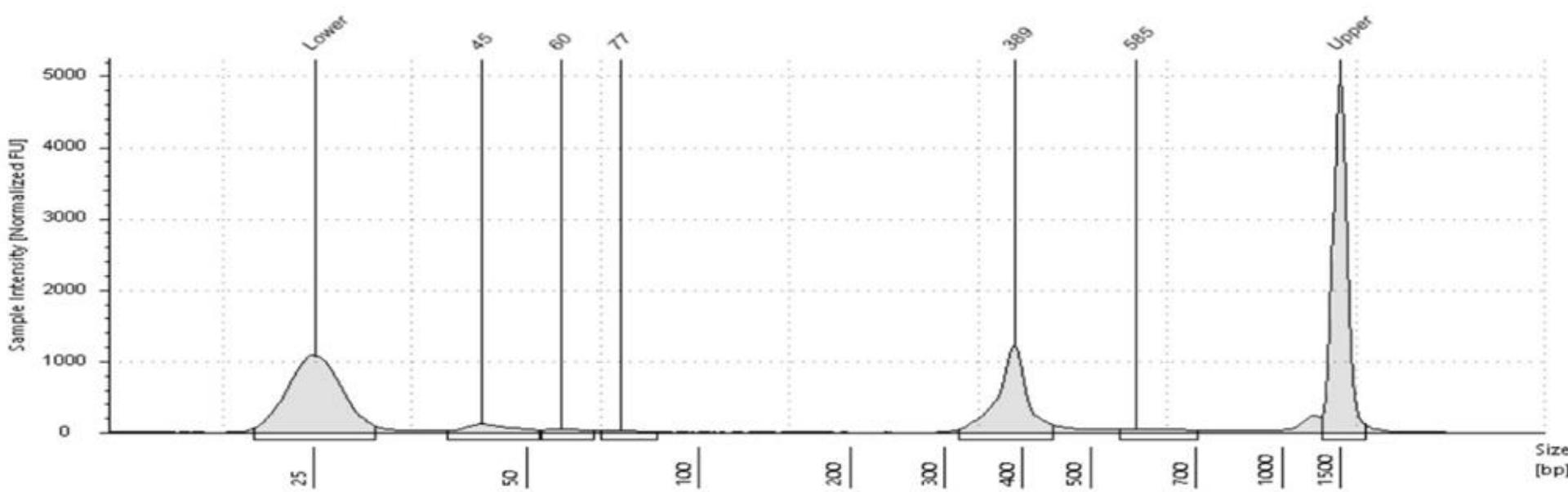
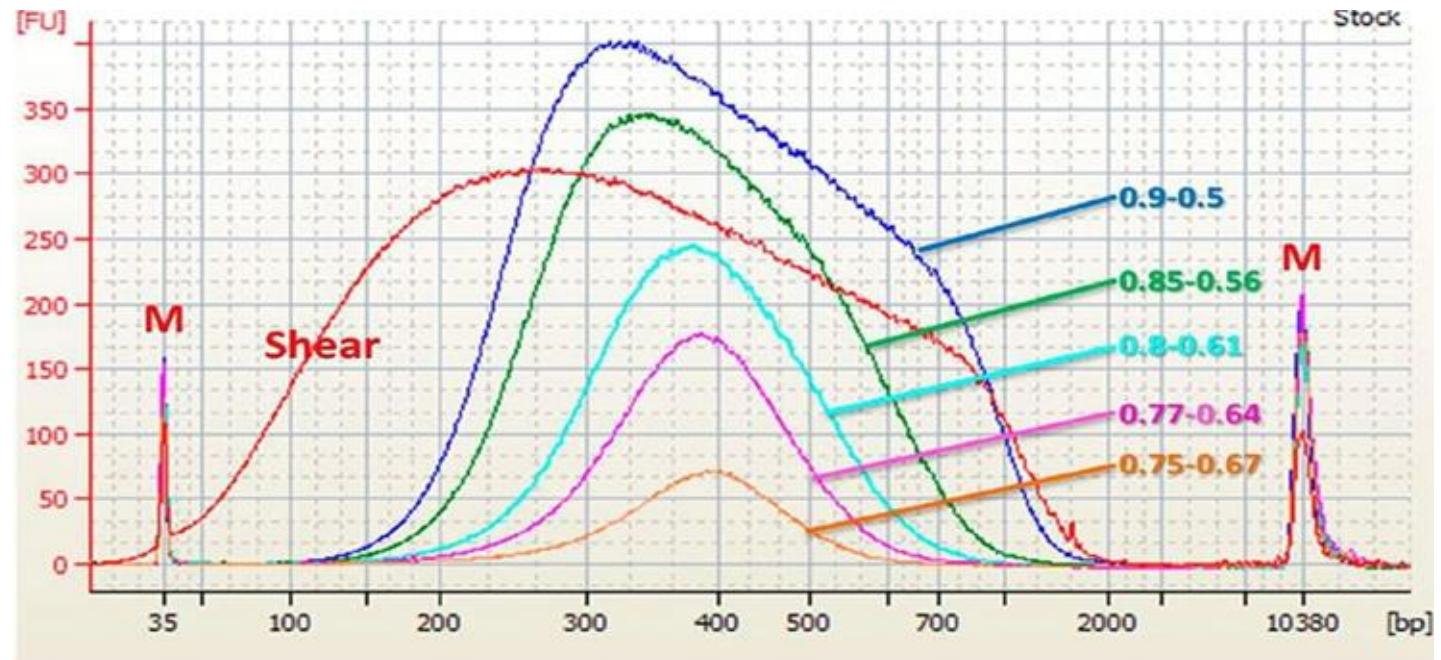
2. Gel

3. Quantify, Normalize & Pool



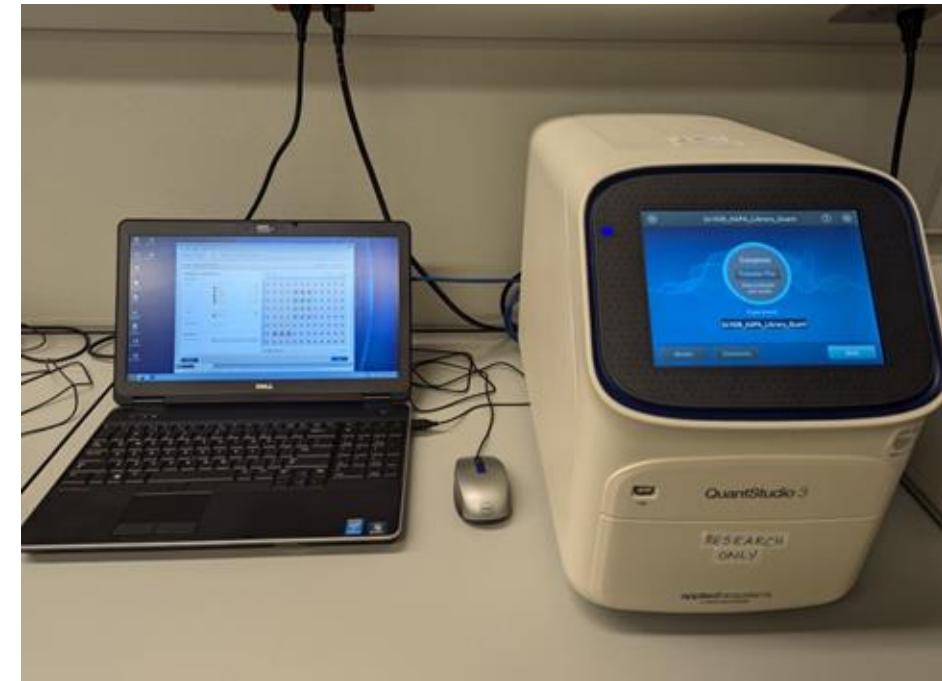
4. Clean up pools

5. Quant & Final Pool

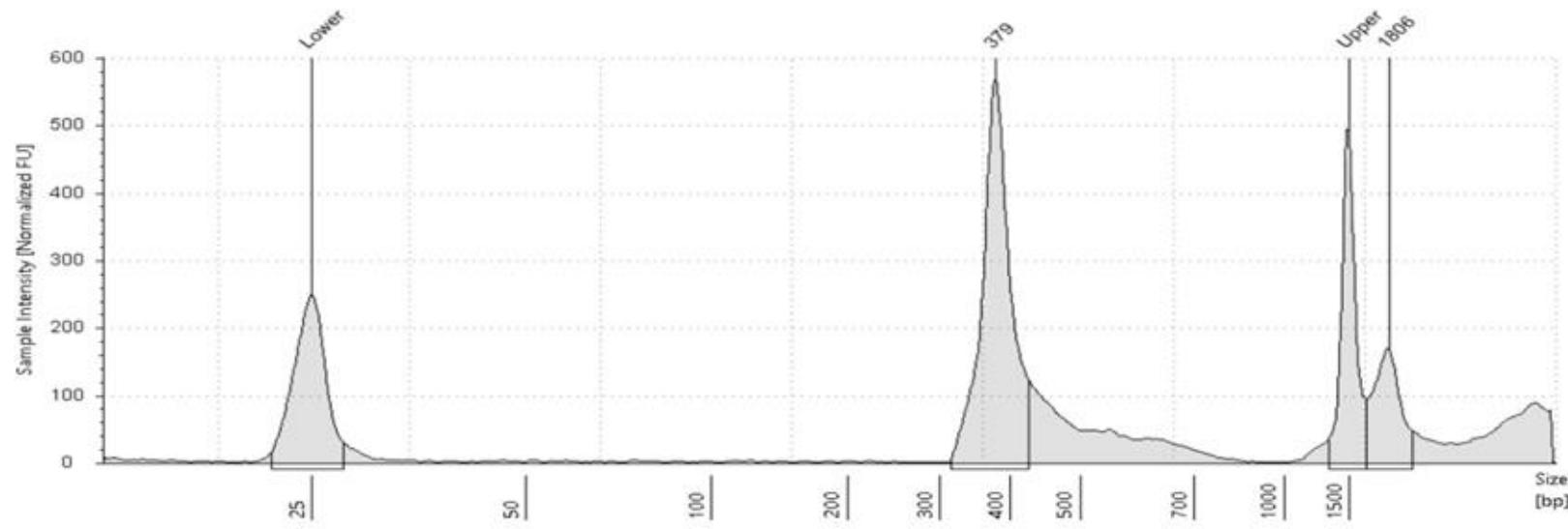


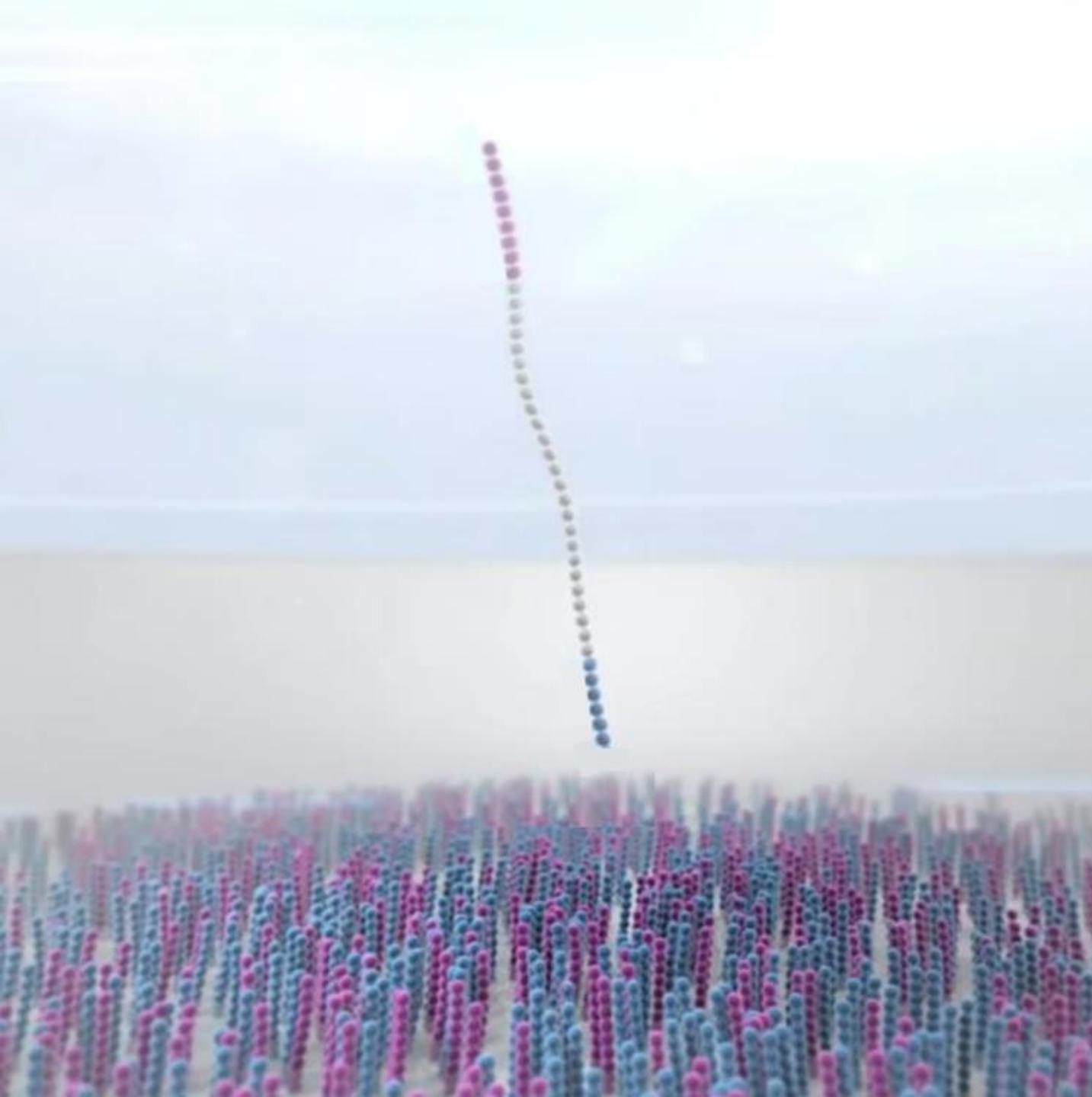


Quality Control at the NGS Core



B1: W1



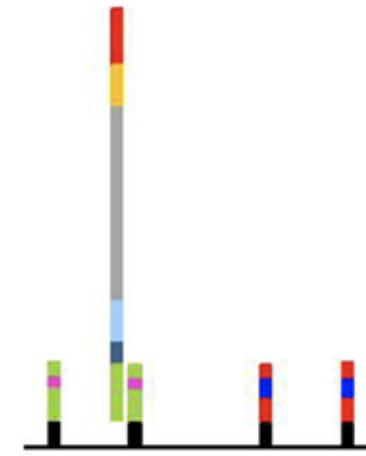
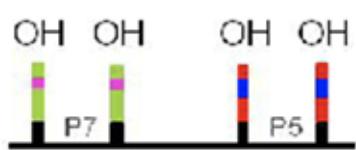


Sequencing By Synthesis

(SBS)

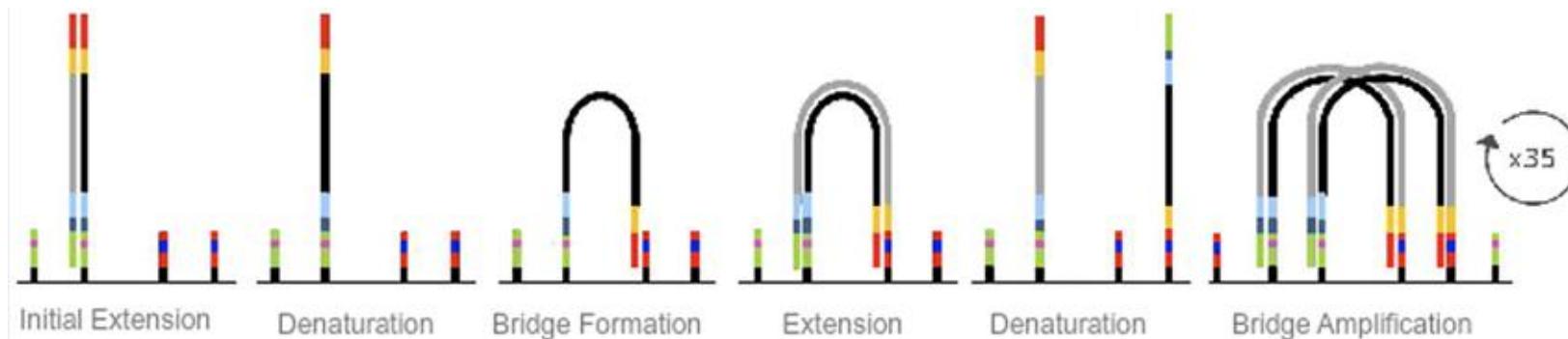


Flow Cell Oligos



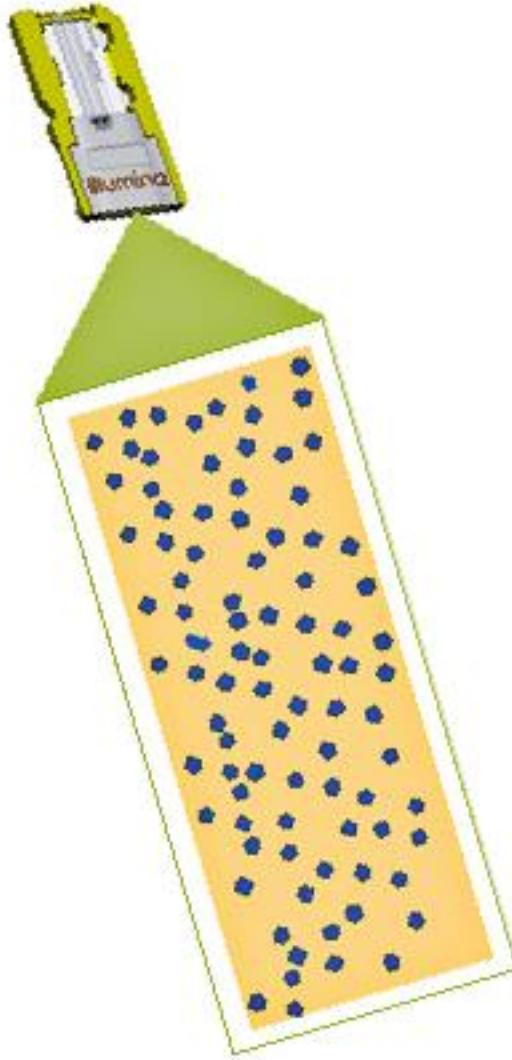
Bridge
Amplification

aka
Cluster
Generation

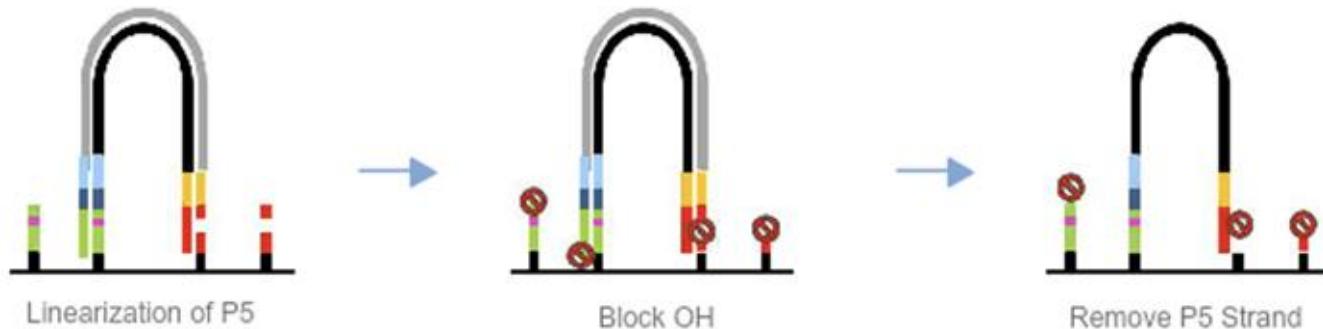


All bridge amplification up to this point has not given us any base calls.

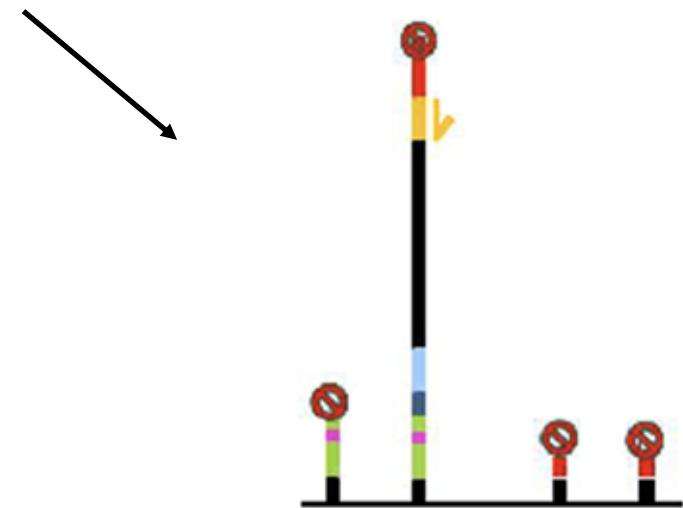
It has generated our clusters, so the fluorescent signals will be readable during base calls.



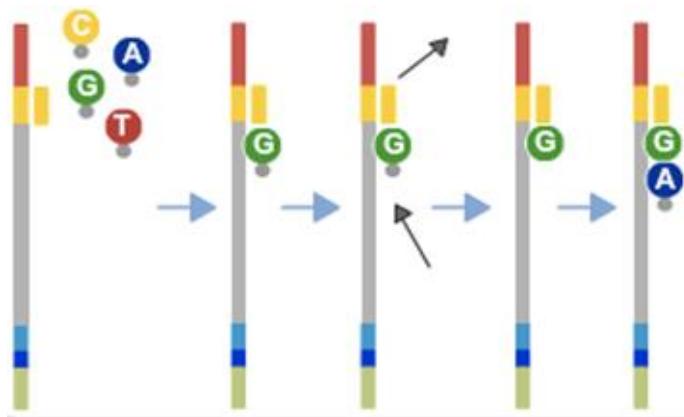
Linearization



The reverse strands are cleaved and washed away (linearization), and the 3' ends and unused oligos are blocked.

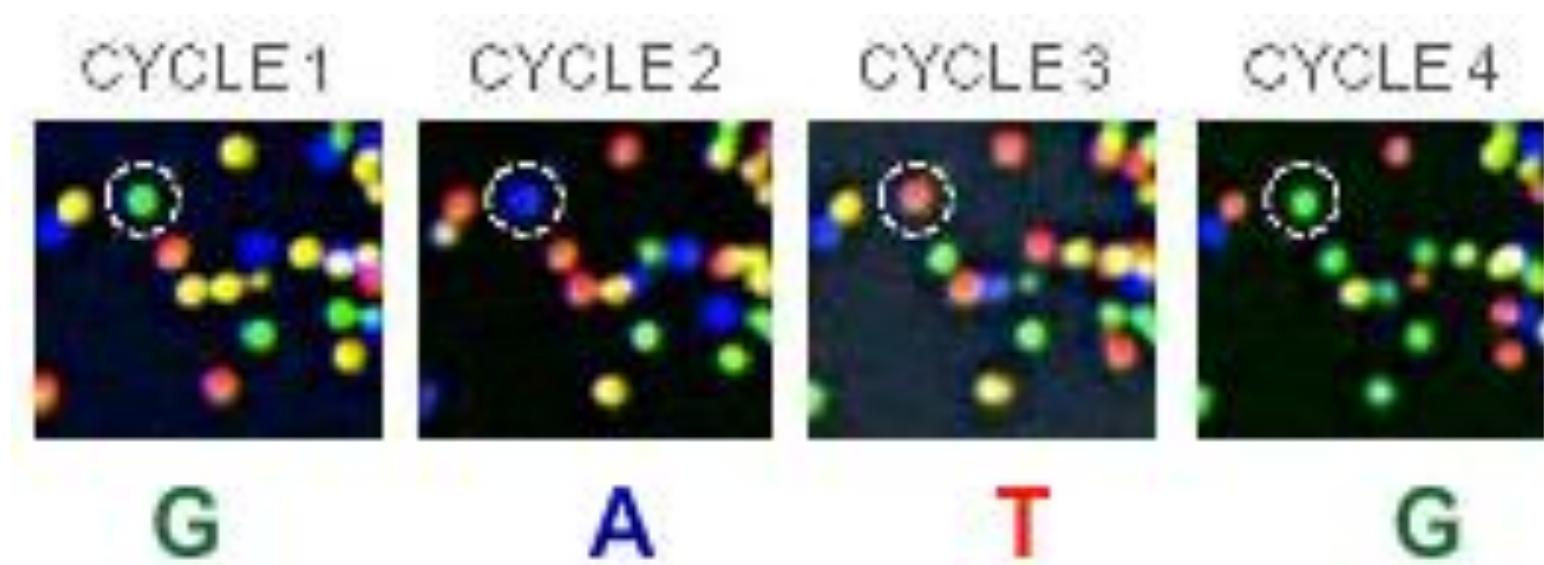


Sequencing primer can anneal and sequencing begins.

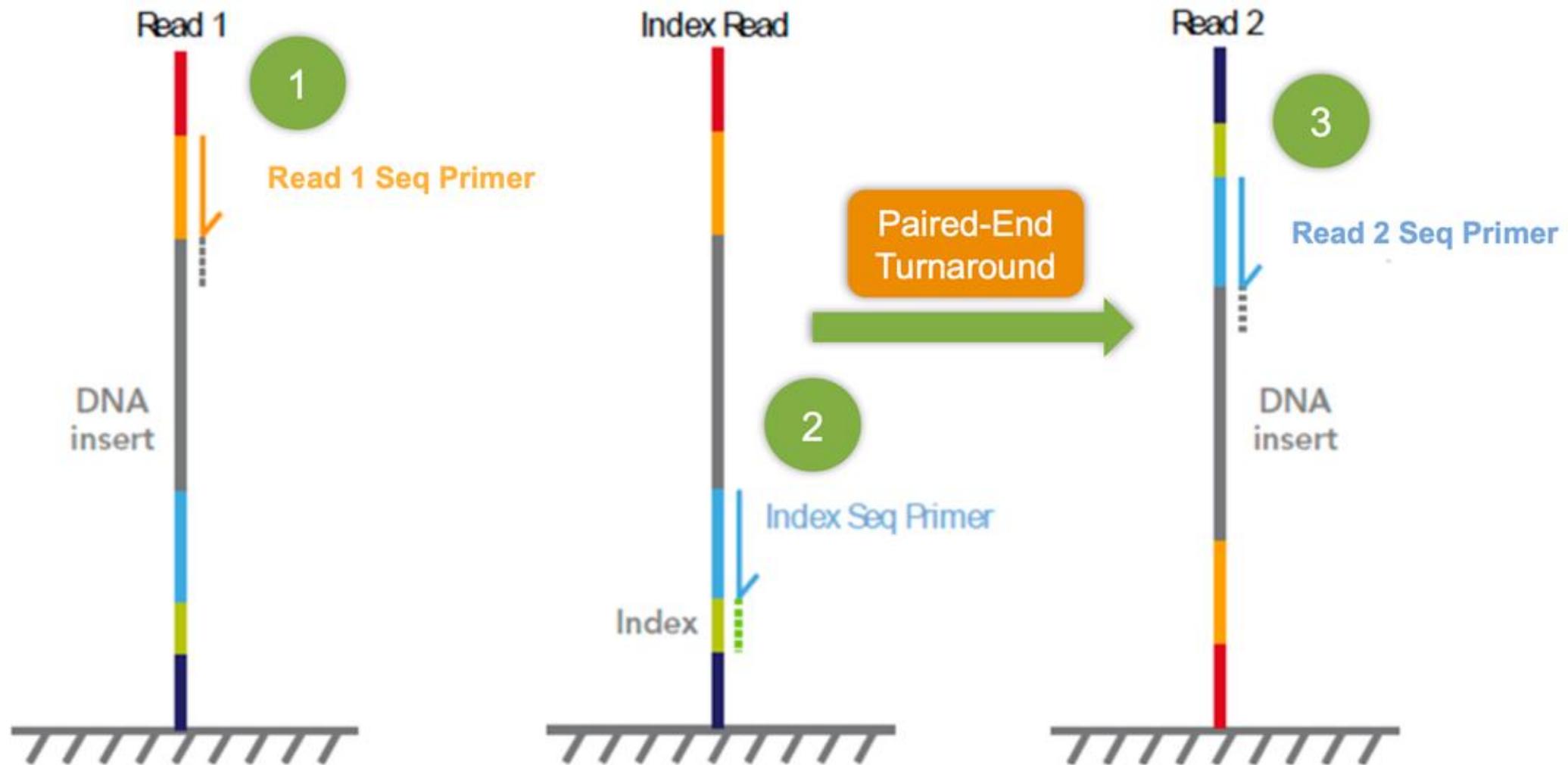


“Sequencing By Synthesis”

Base Calls Four Channel Chemistry

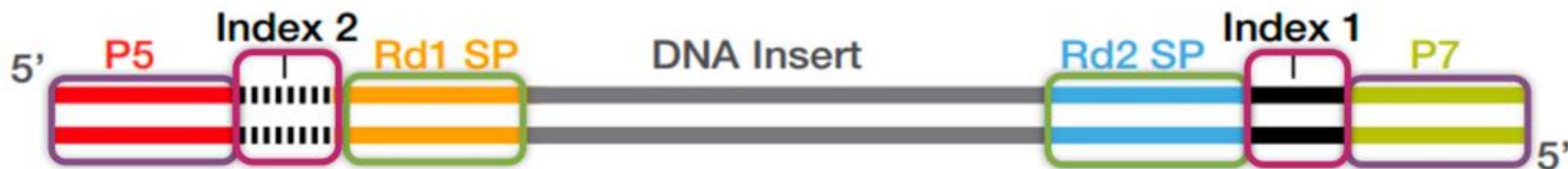


Single Index Read



Turn-Around Chemistry for Read 2

One more Bridge Amplification and then the forward strand is cleaved off, leaving the reverse strand to be sequenced during Read 2.



For clustering:

Libraries must have P5 and P7 binding regions on either end of a library

For sequencing:

Libraries must have sequencing primer binding regions

For mixing samples:

Libraries must have a unique index or barcodes sequence

Illumina

Focused Power



iSeq™ 100



MiniSeq™



MiSeq™

Flexible Power



NextSeq™ 550



NextSeq™
1000/2000

Population Power

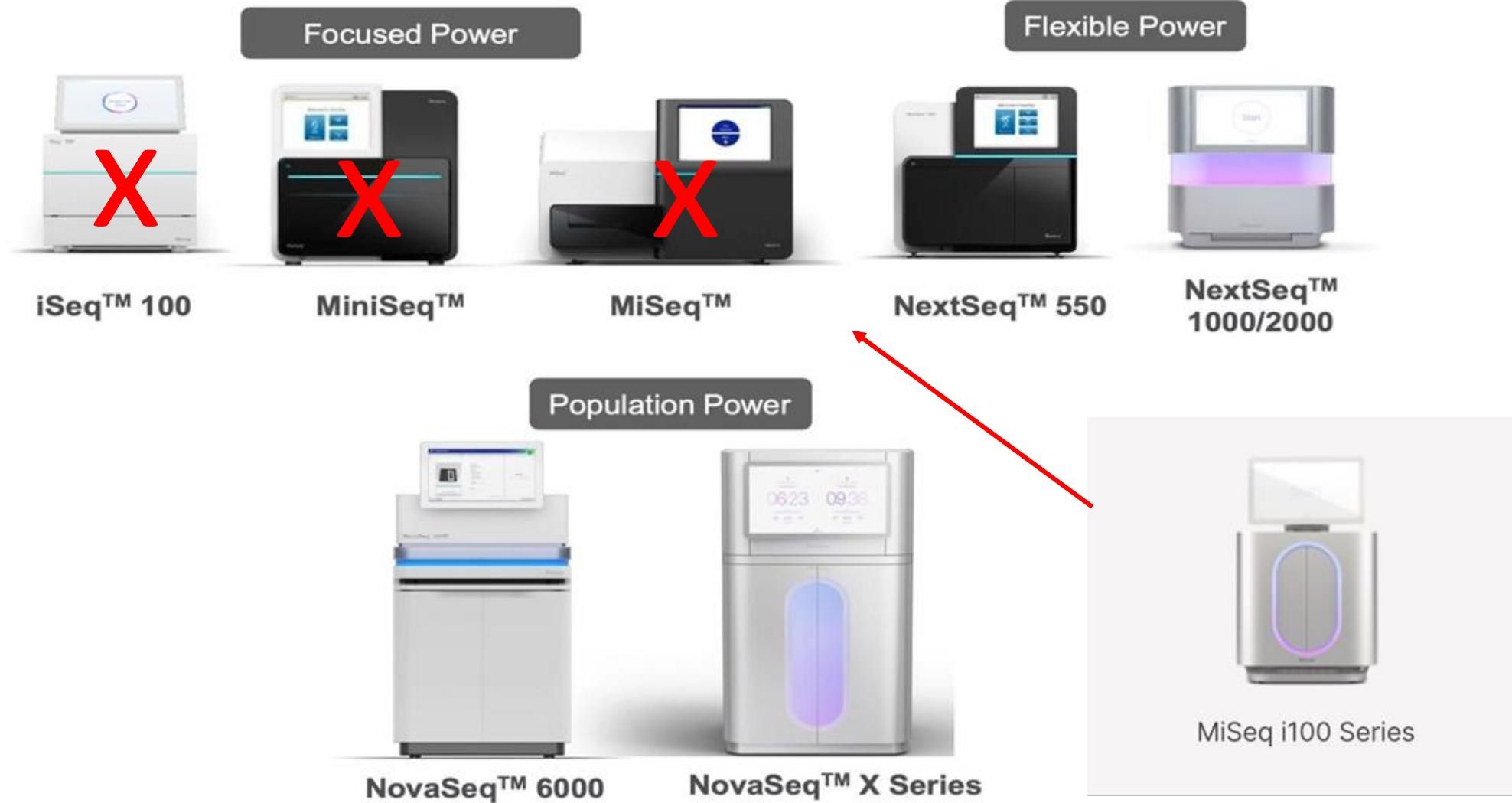


NovaSeq™ 6000



NovaSeq™ X Series

Illumina



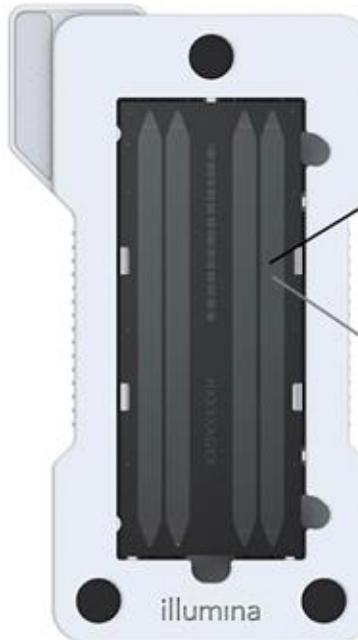


Flow Cell Architecture

Random vs Patterned

Random Flow Cell

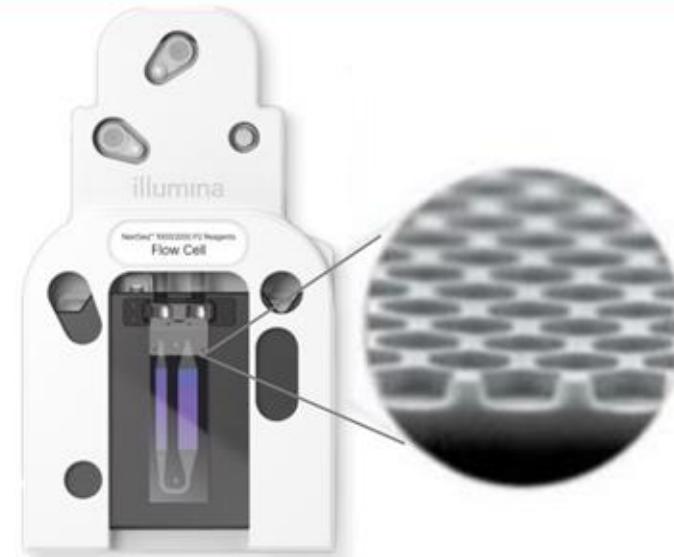
- MiSeq, NextSeq 550, MiniSeq
- Randomly spaced clusters
- More sensitive to overloading



NextSeq 550 Flow Cell

Patterned Flow Cell

- NovaSeqs, NextSeq 1000/2000, iSeq 100
- Defined cluster size and spacing
- Increased Cluster Density
- Simplified imaging

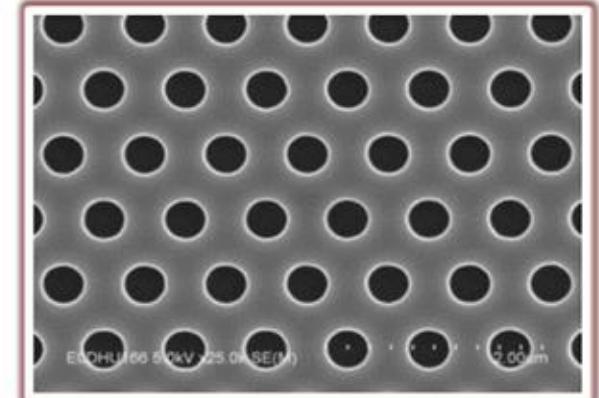
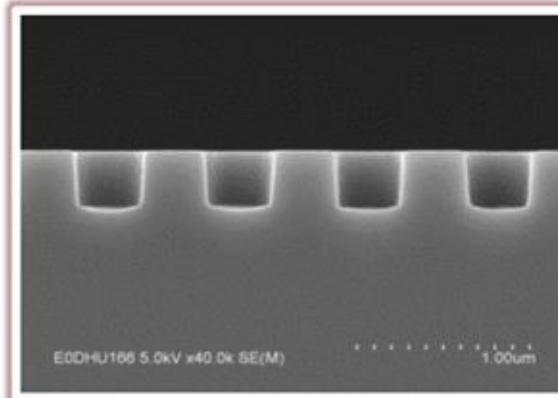
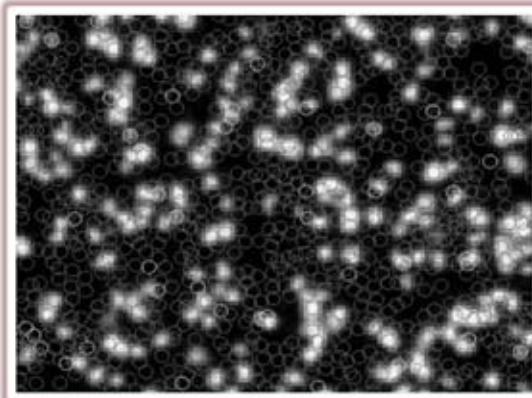
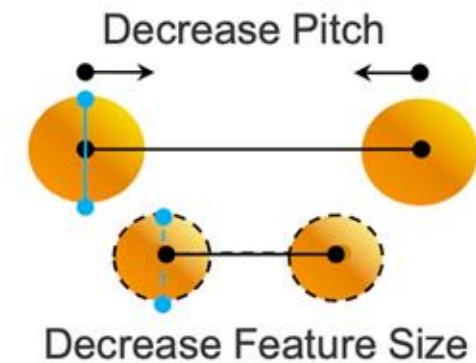
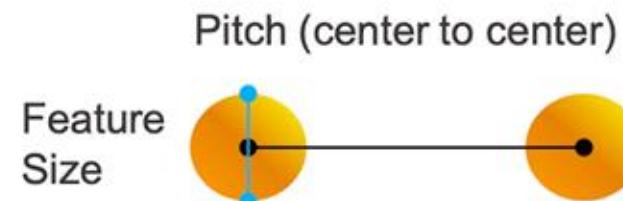
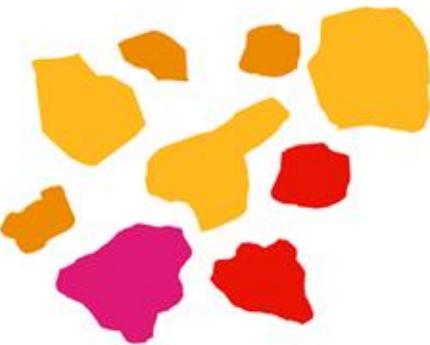


NextSeq 1000/2000 Flow Cell

Patterned Flow Cells

Controlled pitch and feature size

Clusters:



Random Flow Cell

Patterned Flow Cell

Sequencing time is reduced by skipping template generation (finding clusters on flow cell)

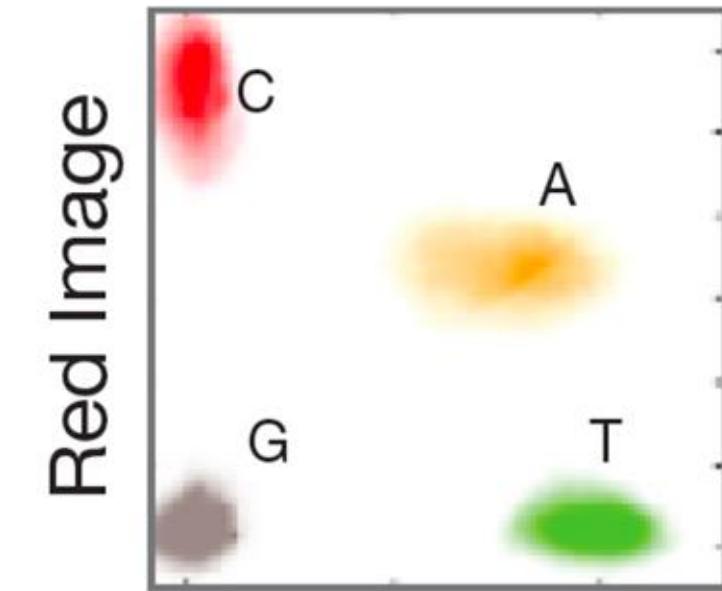
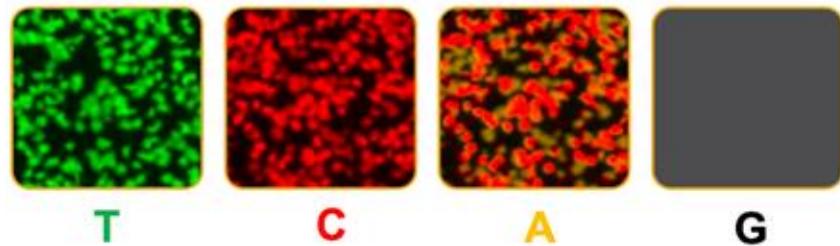
Two Channel SBS Chemistry:

NextSeq Systems, MiniSeq, NovaSeq Systems

- Two channel SBS uses two images:
 - Clusters appearing in the green image only are **T**
 - Clusters appearing in the red*^ image only are **C**
 - Clusters appearing in both^ images are **A**
 - Clusters not present/dark are **G**
- After imaging, cluster intensities are plotted and bases called accordingly

*NextSeq 1000/2000 uses blue for **C** and both blue + green for **A**

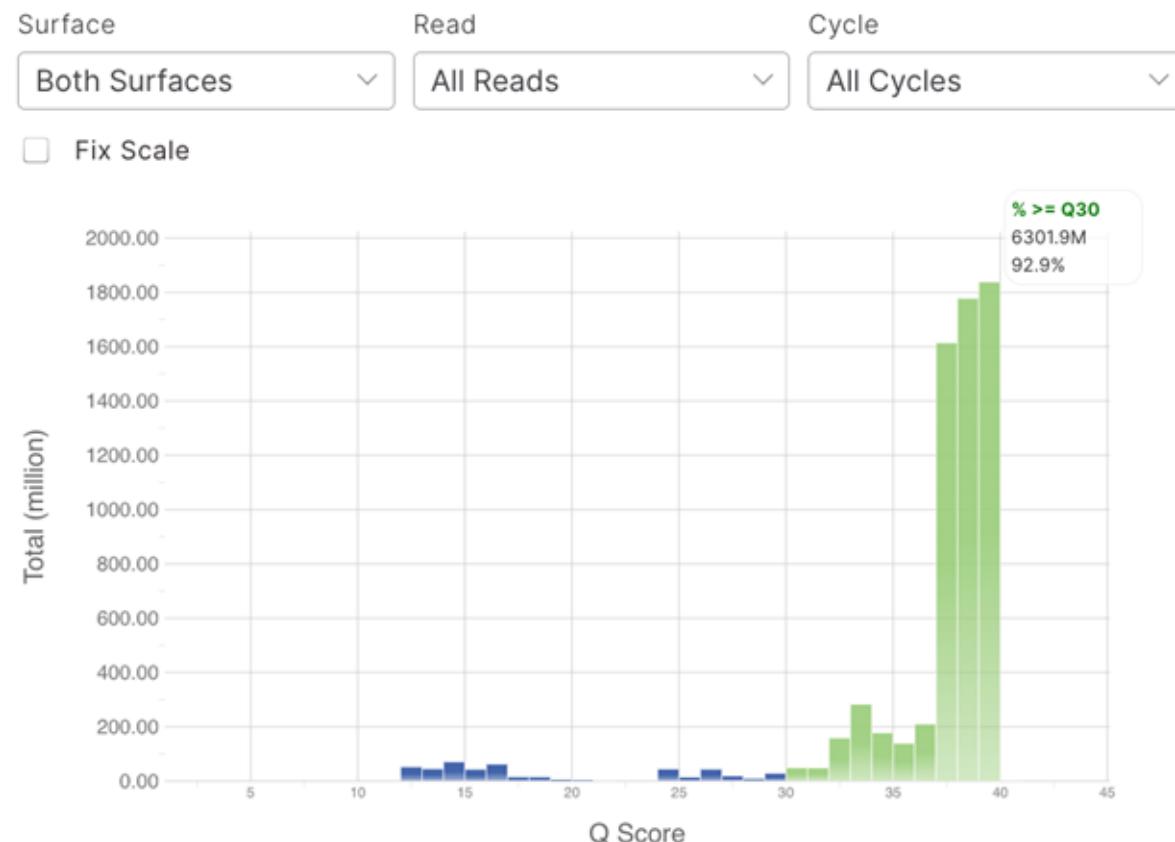
^NovaSeq X Plus uses blue for **A** and both blue + green for **C**



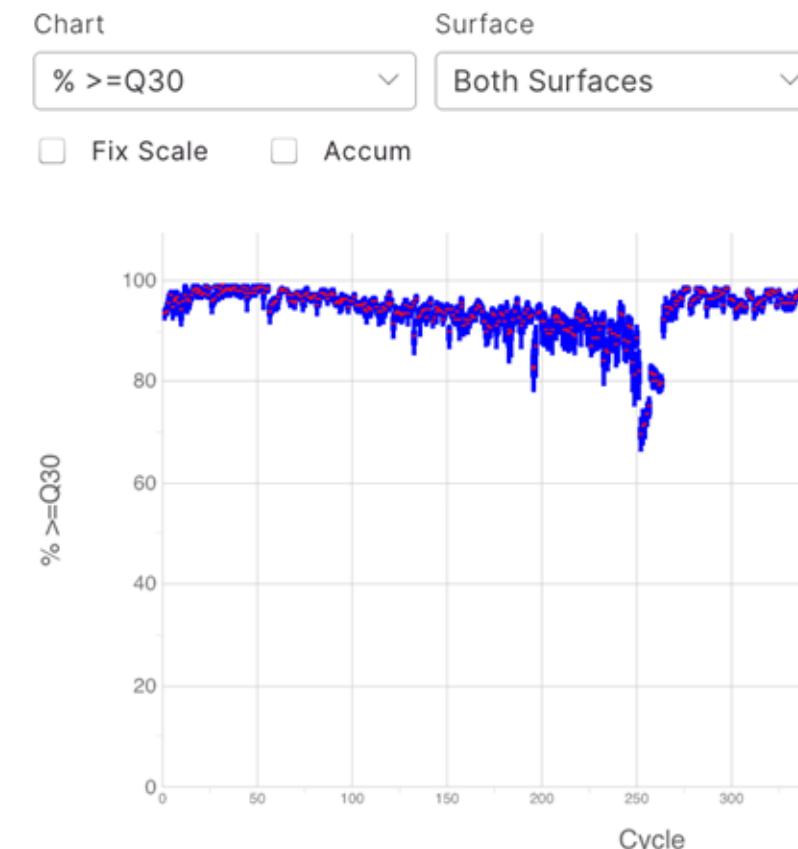
Green Image

Run Results on BaseSpace

Qscore Distribution Chart



Data By Cycle



Areas of Interest

Techniques

Technology

Illumina Research & Innovation

Training

Publications

Data Analysis & Informatics

Illumina Resources & Tools

NGS for Beginners

Genomics Education

Illumina NGS & Array Training

Educational Webinars

Support Webinars & Online Training

Videos

Podcasts

Medical Genetics

All Training



Next-generation sequencing for beginners

Curious about using NGS in your research?
Learn the basics of next-generation sequencing and find tips for getting started.[Learn about NGS](#)

For everyone

Access NGS with us
speed[See All Learning Options](#)[Meet the MiSeq i100 Series](#)

Feedback

<https://knowledge.illumina.com/><https://support.illumina.com/>

Paper discussion: Burcham et al. 2024