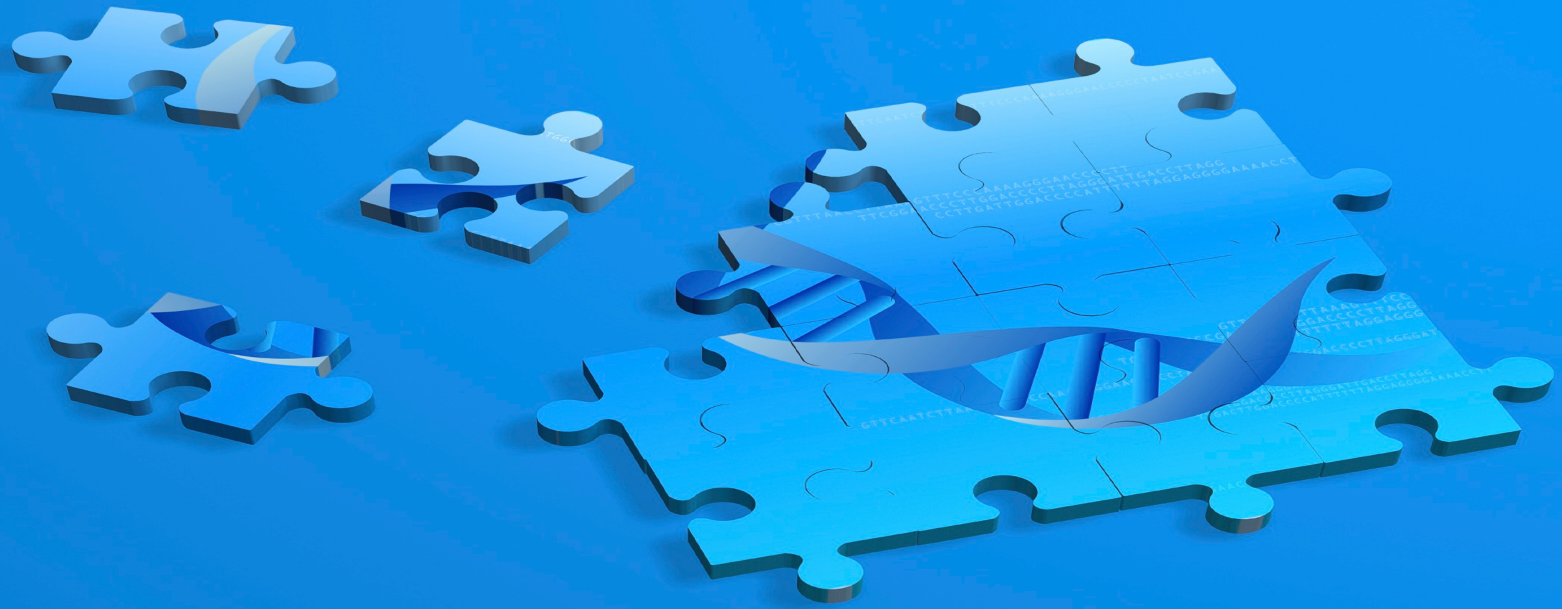
qb3
ucb·ucsc·ucsf



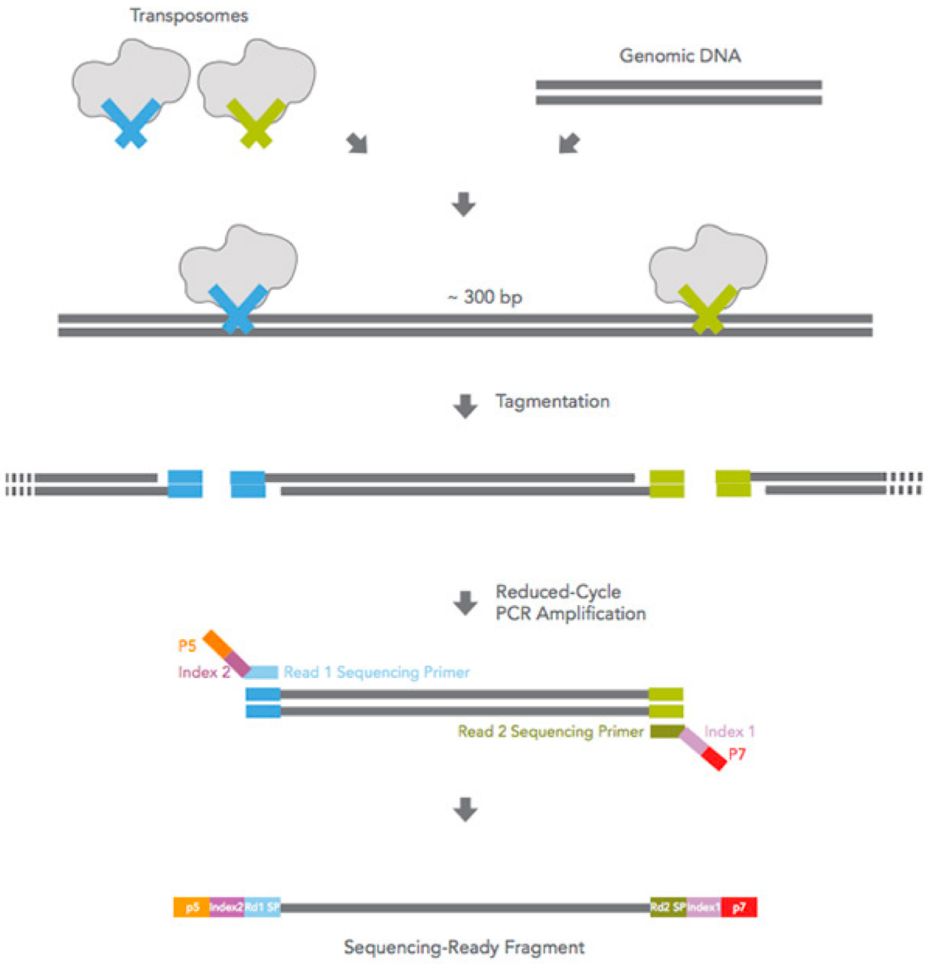
California Institute for Quantitative Biosciences

Community Amplicon Sequencing (iTag)

QB3 Functional Genomics & Genomics Sequencing Labs Services and Technologies

Dylan Smith
Facility Manager, Genomics Sequencing Laboratory

NGS on Illumina sequencers



Amplicon sequencing for community analysis

- Amplicons are generated via PCR with highly specific primers
- PCR primers carry unique built-in barcodes/tags/indices, adapters
- Tagged amplicons are pooled, sequenced in parallel
- Barcodes used to “demultiplex” data pool back to original samples

Amplicon sequencing for community analysis

- <u>Amplicons are generated via PCR with highly specific primers</u>

- PCR primers carry unique built-in barcodes/tags/indices, adapters

- Tagged amplicons are pooled, sequenced in parallel

- Barcodes used to “demultiplex” data pool back to original samples

Amplicon sequencing for community analysis

- Amplicons are generated via PCR with highly specific primers
- PCR primers carry unique built-in barcodes/tags/indices, adapters
- Tagged amplicons are pooled, sequenced in parallel
- Barcodes used to "demultiplex" data pool back to original samples

What is iTag?

- Tagged (barcoded) amplicon sequencing
  - Could theoretically be any locus from any organism

What is iTag?

- Tagged (barcoded) amplicon sequencing
  - Could theoretically be any locus from any organism
  - “iTag” experiments most typically sequence a **barcoding gene**
  - Microbial communities lend themselves well to community amplicon sequencing because they are
    - **Small**, hard to see and find
    - **Superabundant**
    - Often **cryptically defined**
    - **Distribution is poorly understood**

Amplicon sequencing for community analysis

- Filtered reads are assigned taxonomy by aligning against a database of known, taxonomically assigned reads

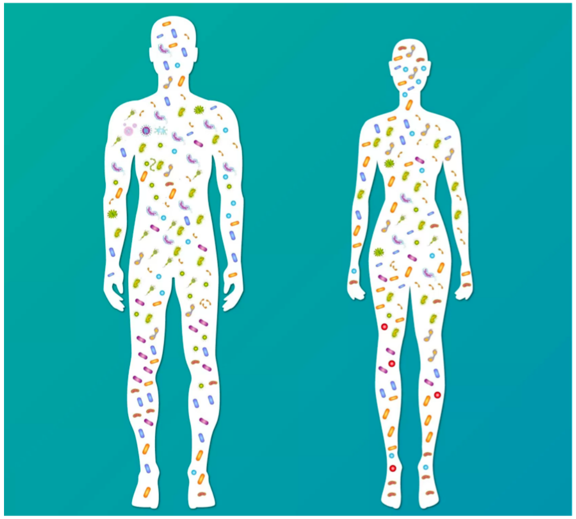
Amplicon sequencing for community analysis

- Filtered reads are assigned taxonomy by aligning against a database of known, taxonomically assigned reads

The point of community amplicon sequencing is to identify, or barcode, organisms from complex communities via specific DNA markers

Microbial Community Analysis

- ***Example***
  - ***What is the total microbial diversity in a given environment?***
    - ***How many species are there? (Alpha diversity)***
    - ***What is the community structure of that diversity? (Beta diversity)***
  - ***How does that diversity change between environments?***



iTag ≠ Metagenomics

- ***Metagenomics***
  - ***Shotgun sequencing of randomly sheared and size-selected gDNA fragments from a microbial community***
  - ***Can gain functional information from communities***
  - ***Often requires extensive sequencing, more expensive***

iTag ≠ Metagenomics

- ***Metagenomics***
  - ***Shotgun sequencing of randomly sheared and size-selected gDNA fragments from a microbial community***
  - ***Can gain functional information from communities***
  - ***Often requires extensive sequencing, more expensive***

- ***iTag***
  - ***Sequencing of highly specific loci determined by choice of PCR primers and experimental design***
  - ***Usually does not contain a functional aspect***
  - ***Loci chosen for taxonomic resolution and phylogenetic relevance***

iTag ≠ Metagenomics

- ***Metagenomics***
  - ***Can paint a broad picture of a microbial community***
  - ***Can include:***
    - ***Taxonomic ID***
    - ***Function***
    - ***Activity (transcriptomics)***

iTag ≠ Metagenomics

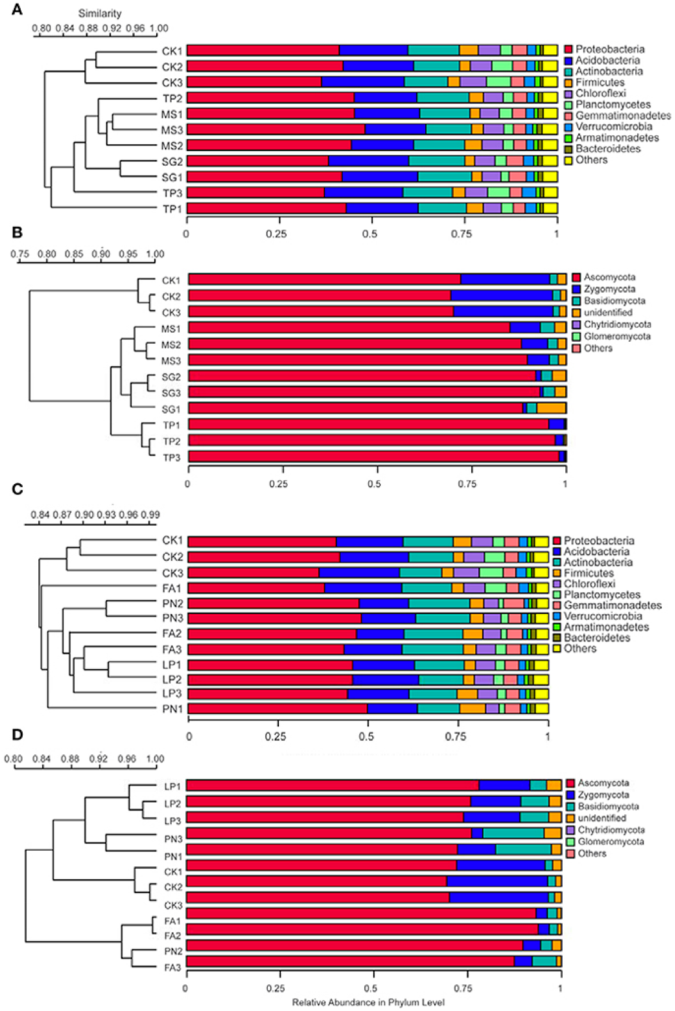
- ***Metagenomics***
  - ***Can paint a broad picture of a microbial community***
  - ***Can include:***
    - ***Taxonomic ID***
    - ***Function***
    - ***Activity (transcriptomics)***

- ***iTag***
- ***Who's there?***
- ***How many are there?***
- ***Who is dominant/rare?***
- ***How do all these observations change between environment and why?***

iTag is a ”counting-based” analysis

- ***Coverage* is typically a metagenomics word**
- ***Reads per sample* is more appropriate and specific to an iTag experiment**
- ***Reads recovered from your sequencing run are treated as analogous to biological occurrence of an organism***
  - ***Cannot be used as an absolute measure***
  - ***Between sample comparison is valid***

iTag – Amplicon sequencing for community analysis



Typical iTag Workflow

- ***Experimental design*** *(this is the* ***most*** *important part)*
- ***Library prep***
  - *gDNA extraction*
  - *PCR amplification (can be one-step or two-step PCR design)*
  - *Library quantification/qualification, QC*
  - *Normalization and pooling*
- ***Sequencing***
- ***Filtering/processing*** *of raw reads (read pairing)*
- ***OTU/ASV table x sample***
- ***Statistical analysis***

iTag Library Prep: Choice of locus

- ***Which organism(s) do I want to sequence?***
- ***What is the goal of my study?***
  - ***Taxonomic ID***
  - ***Phylogenetics***
  - ***Function***
- ***What is the predicted diversity within my environment?***

iTag Library Prep: Choice of locus

- ***Factors to consider***
  - ***Mutation rate***
    - ***Am I likely to be comparing species or larger taxonomic guilds?***
  - ***Length***
    - ***Which sequencing platform will I eventually run my samples on? Do I want overlap? (yes)***
  - ***Utility of taxonomic information***
    - ***Are there good, reliable databases for my locus?***

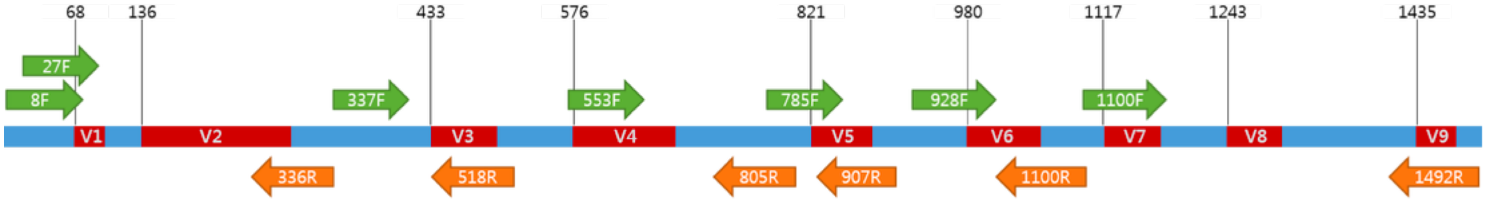
iTag Library Prep: Choice of locus

- ***Factors to consider***
  - ***Mutation rate***
    - ***Am I likely to be comparing species or larger taxonomic guilds?***
  - ***Length***
    - ***Which sequencing platform will I eventually run my samples on? Do I want overlap? (yes)***
  - ***Utility of taxonomic information***
    - ***Are there good, reliable databases for my locus?***

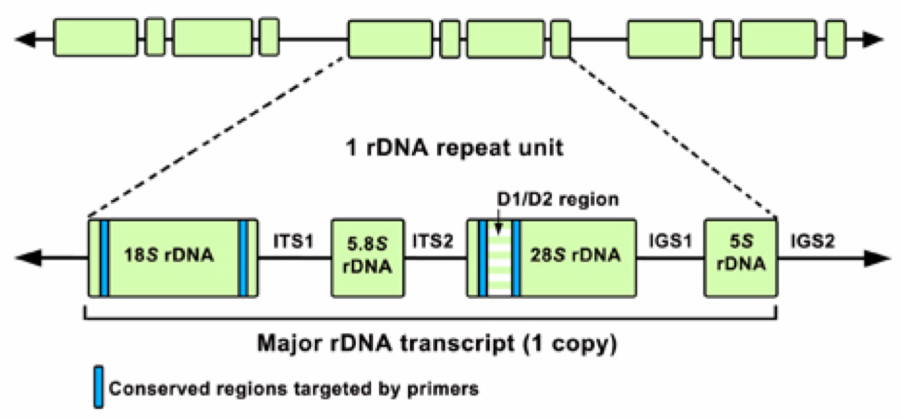
When in doubt, a literature search is usually the best course for determining the right locus/primer choice for your study.

Common Barcoding Regions for Microbes

- ***Bacteria/Archaea: 16S rRNA gene***



- ***Fungi: Internal Transcribed Spacer rRNA***



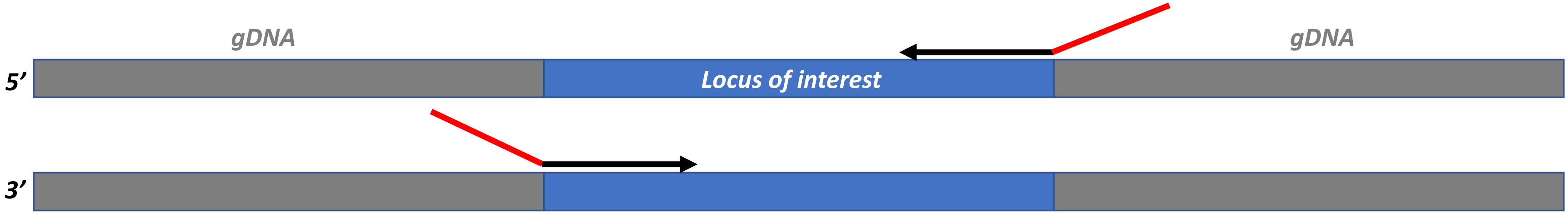
http://help.bioiplug.com/16s-rrna-and-16s-rrna-gene/ https://mycology.adelaide.edu.au/descriptions/

iTag Library Prep: Primer design

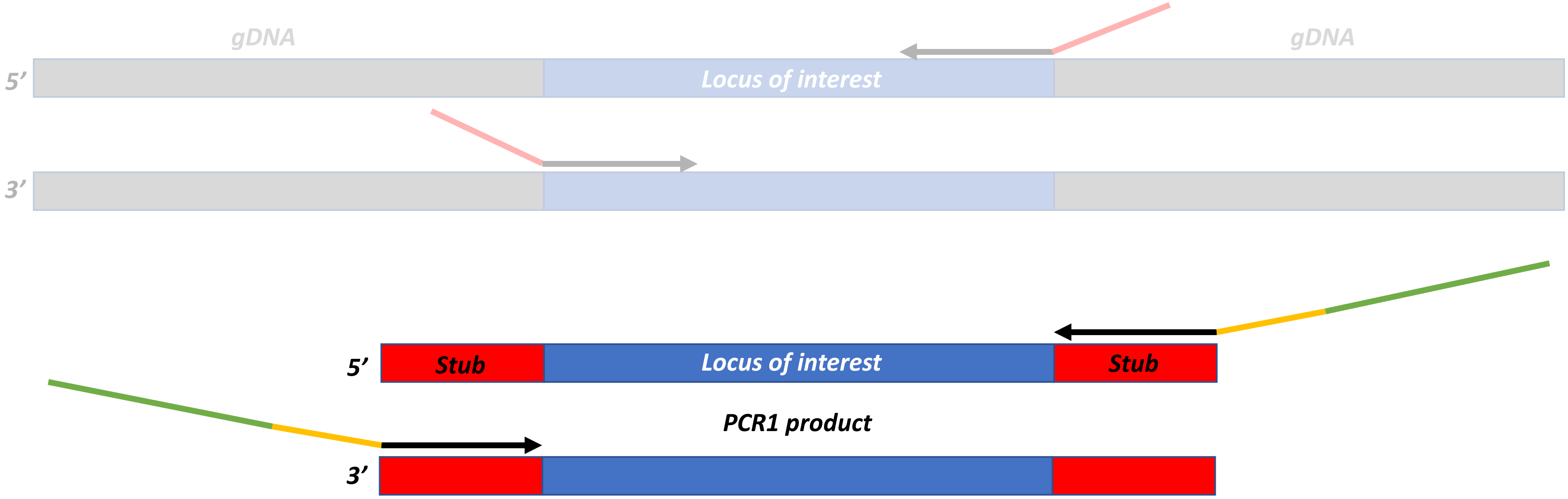
- ***Dimerization and secondary structures***
  - ***Self dimerization***
  - ***Cross dimerization***
  - ***Self complimentary***

- ***Melting temperatures***
  - ***Will my libraries be run with PhiX?***
  - ***PhiX is a common diversity and loading concentration control run with Illumina libraries***

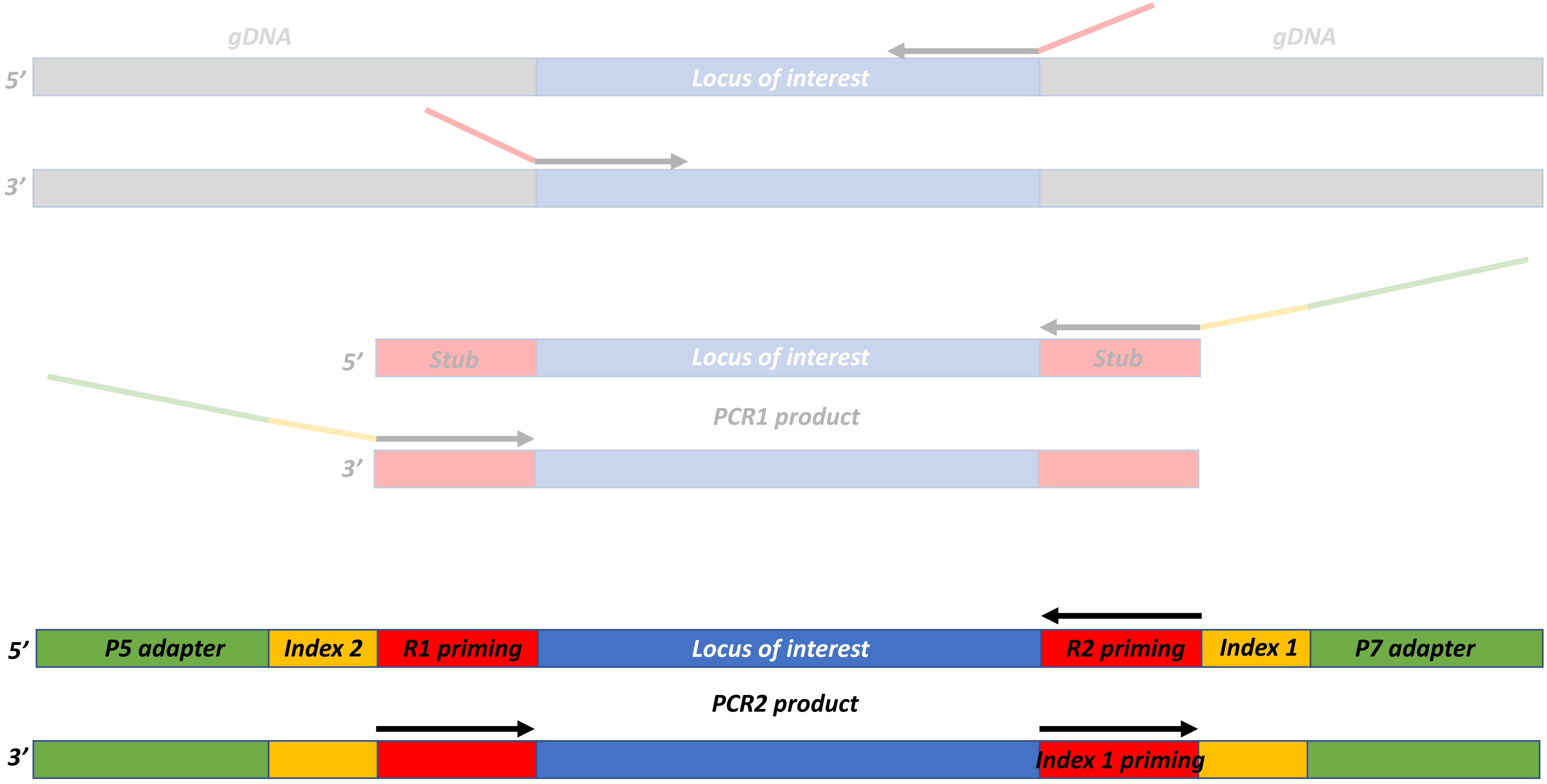
iTag Library Prep: Primer design



iTag Library Prep: Primer design

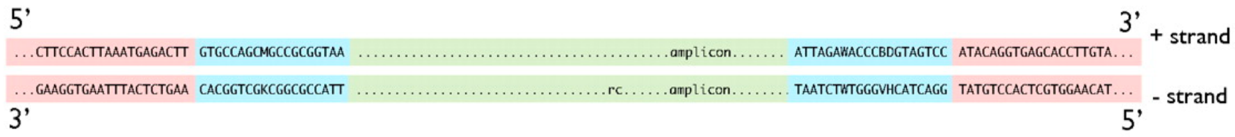


iTag Library Prep: Primer design



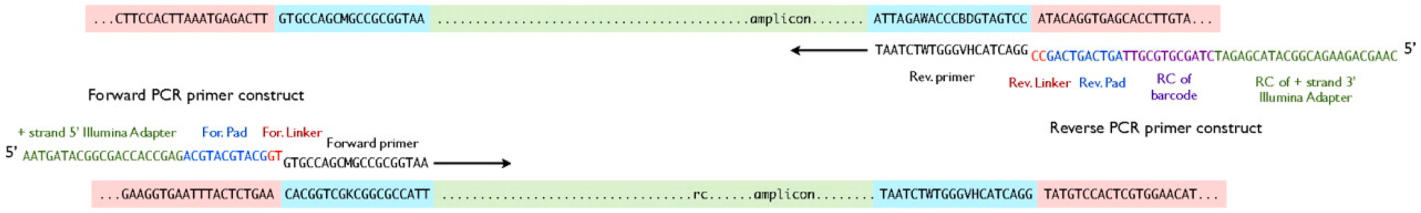
iTag Library Prep: Primer design

Target gene:



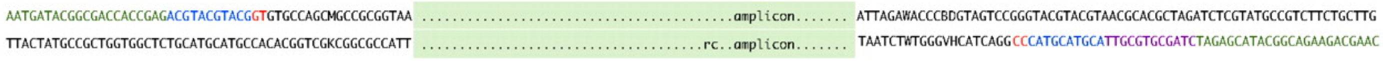
```
5'                                                                                                                   3'
...CTTCCACTTAAATGAGACTT GTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCC ATACAGGTGAGCACCTTGTA...   + strand
...GAAGGTGAATTTACTCTGAA CACGGTCGKCGGCGCCATT ................................rc......amplicon....... TAATCTWTGGGVHCATCAGG TATGTCCACTCGTGGAACAT...   - strand
3'                                                                                                                   5'
```

Amplification primers with annealing sites:



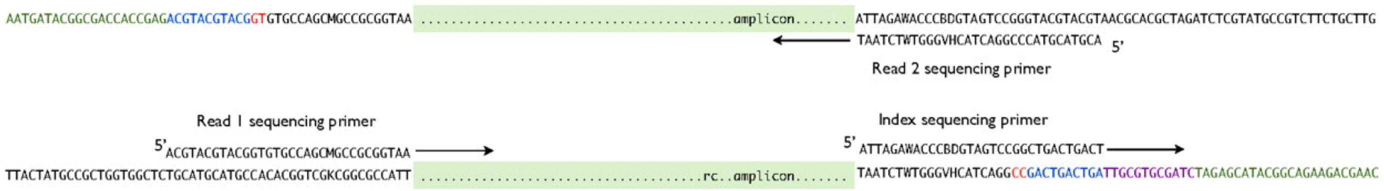
```
...CTTCCACTTAAATGAGACTT GTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCC ATACAGGTGAGCACCTTGTA...
                                                                          <------ TAATCTWTGGGVHCATCAGGCCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC 5'
                                                                                   Rev. primer       Rev. Linker  Rev. Pad  RC of     RC of + strand 3'
                                                                                                                            barcode   Illumina Adapter
Forward PCR primer construct
                                                                                                     Reverse PCR primer construct
+ strand 5' Illumina Adapter   For. Pad   For. Linker
                                                     Forward primer
5' AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA ------>
...GAAGGTGAATTTACTCTGAA CACGGTCGKCGGCGCCATT ................................rc......amplicon....... TAATCTWTGGGVHCATCAGG TATGTCCACTCGTGGAACAT...
```

Amplification products:



```
AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTGCATGCATGCCACACGGTCGKCGGCGCCATT ....................................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCCATGCATGCATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

Sequencing primers with annealing sites:



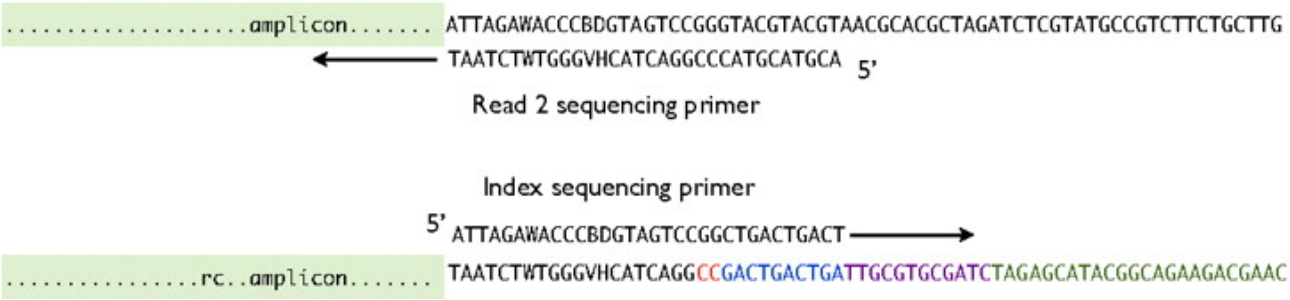
```
AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
                                                                                               <------ TAATCTWTGGGVHCATCAGGCCCATGCATGCA 5'
                                                                                                        Read 2 sequencing primer

                    Read 1 sequencing primer                                                            Index sequencing primer
                  5'ACGTACGTACGGTGTGCCAGCMGCCGCGGTAA ------>                                          5' ATTAGAWACCCBDGTAGTCCGGCTGACTGACT ------>
TTACTATGCCGCTGGTGGCTCTGCATGCATGCCACACGGTCGKCGGCGCCATT ....................................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

iTag Library Prep: Primer design

- ***"Linker"***
  - ***Short (typically 2-3bp) intentional mismatch to your organism's genome***
  - ***Designed to have the 5' end of the primer physically hang off the genomic template***
  - ***Thought to decrease overall PCR primer bias of certain taxa over others***

iTag Library Prep: Primer design

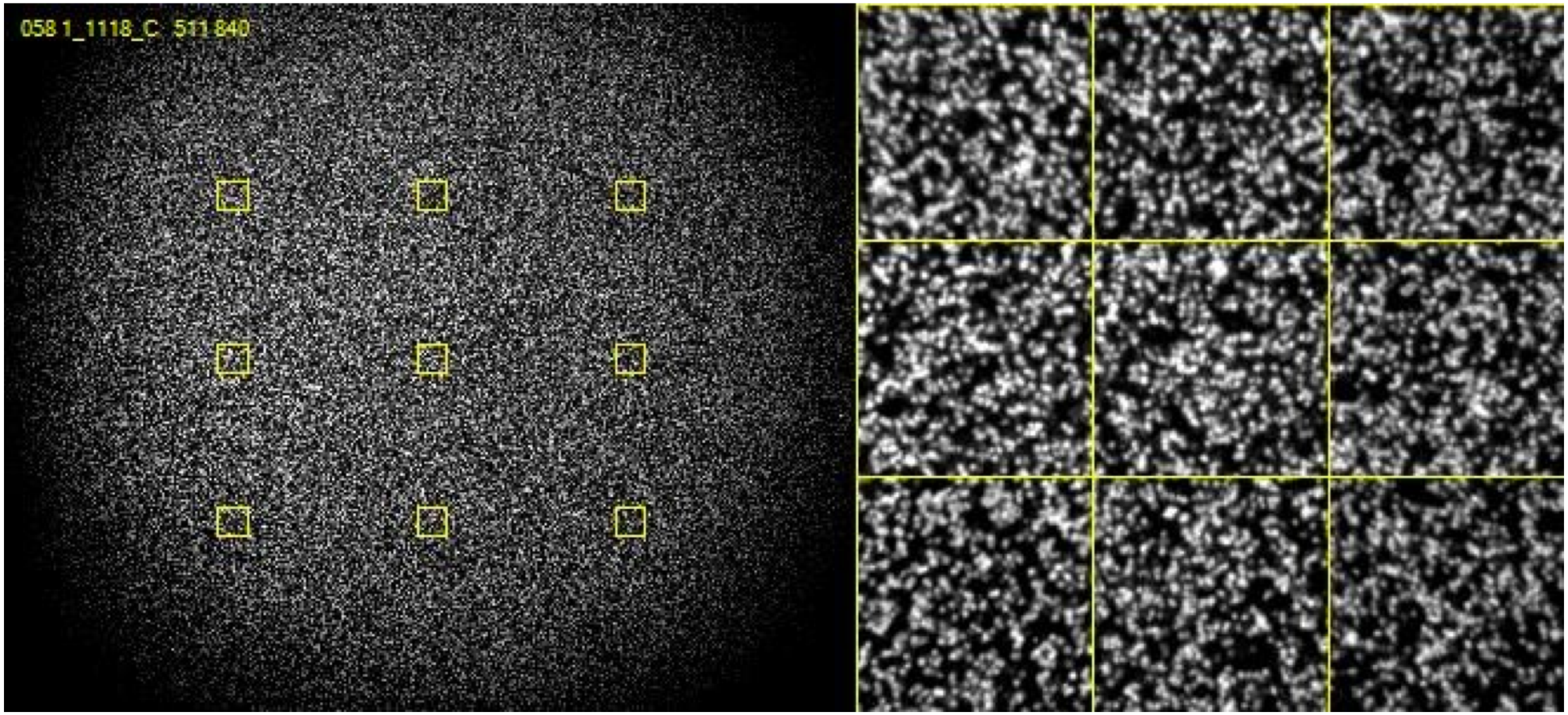
- ***"Pad"***
  - ***Stretch of random bases designed into PCR primers that is NOT designed to match any actual genomic sequence***
  - ***Bases provide space (length) on which sequencing primers sit***
  - ***Pad may provide additional chemistry advantages (think dimerization and melting temperatures) for the PCR primer***



```
....................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
                                    TAATCTWTGGGVHCATCAGGCCCATGCATGCA  5'
                                      Read 2 sequencing primer

                                      Index sequencing primer
                                 5' ATTAGAWACCCBDGTAGTCCGGCTGACTGACT
................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

iTag Sequencing

- ***Sequence diversity at every base position matters!***



iTag Sequencing

Sequence diversity can be added into primer design

- ***Sequence multiple loci at once***
- ***"N stagger"***

…TGAGACTT**N**GTGCCAGCMGCC…
…TGAGACTT**NN**GTGCCAGCMGCC…
…TGAGACTT**NNN**GTGCCAGCMGCC…
…TGAGACTT**NNNN**GTGCCAGCMGCC…
…TGAGACTT**NNNNN**GTGCCAGCMGCC…

- ***"N shuffle"***

…TGAGACTT**NNNNN**GTGCCAGCMGCC…
…TGAGACTT**NNNNN**GTGCCAGCMGCC…
…TGAGACTT**NNNNN**GTGCCAGCMGCC…
…TGAGACTT**NNNNN**GTGCCAGCMGCC…

iTag Library Prep/Sequencing at Berkeley

- ***16S – V3/V4 hypervariable regions***
  - ***Fwd 515Fb, Rev 806Rb***
- ***ITS – Smith/Peay ITS1***
  - ***Fwd Smith/Peay ITS1f, Rev Smith/Peay ITS2***
- ***~ 50% of reads are "flipped" in orientation***
- ***Currently running single-locus amplicon pools with 20M-25M read return, 0% PhiX***

Common iTag sequencing questions

- ***Which sequencer should I use?***
- ***How many bases should I run?***
- ***Should I do paired end reads?***
- ***How many reads do I need?***
- ***What are the proper controls for my study?***

Common iTag sequencing questions

- ***Which sequencer should I use?***
- ***How many bases should I run?***
- ***Should I do paired end reads?***
- ***How many reads do I need?***
- ***What are the proper controls for my study?***

When in doubt, try a literature search

Basic Illumina Stats

MiSeq

HiSeq 2500 Rapid

15-25 million reads (v3 chemistry

300PE run ~4 days

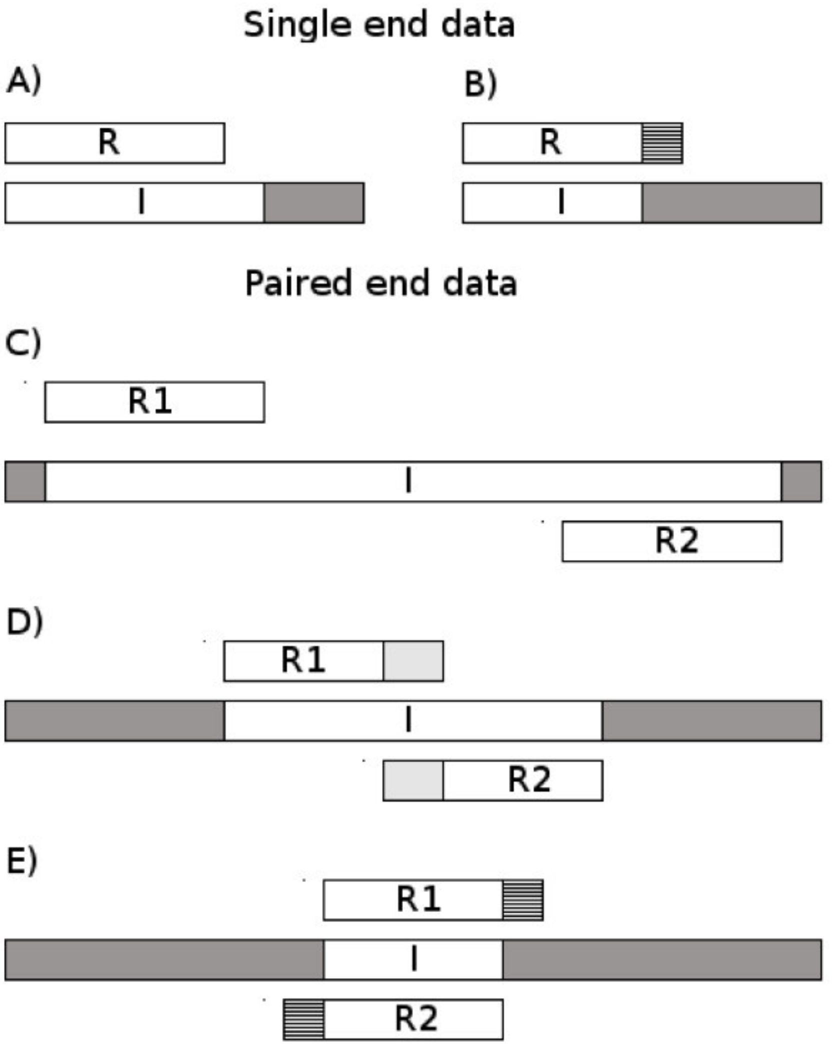
Single lane

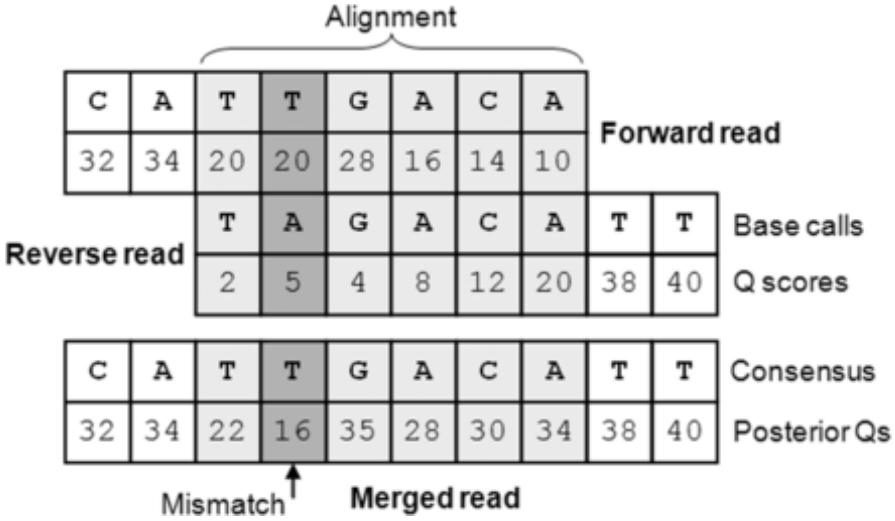
100-150 million reads per lane*

Run type (# of cycles) more flexible, faster

*Usually must run two lanes

Benefits of paired end reads



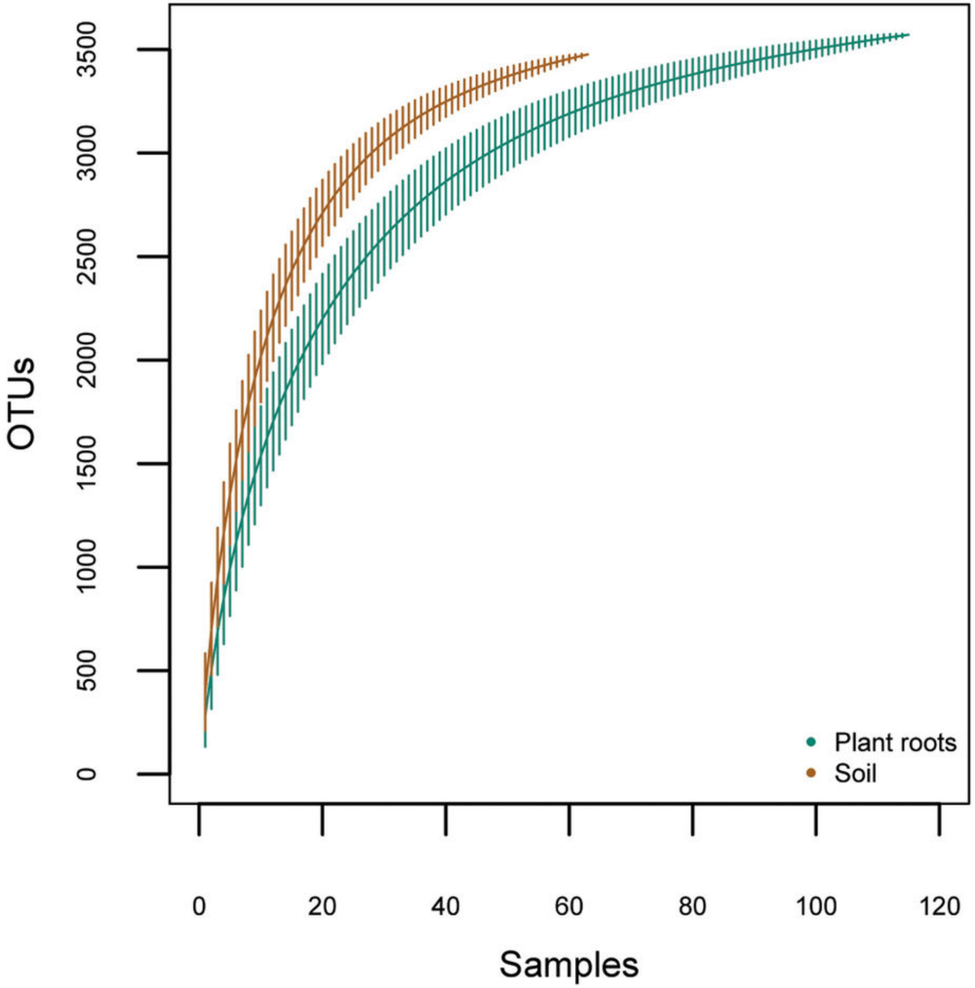


Forward read	C	A	T	T	G	A	C	A		
	32	34	20	20	28	16	14	10		
Reverse read – Base calls			T	A	G	A	C	A	T	T
Q scores			2	5	4	8	12	20	38	40
Consensus	C	A	T	T	G	A	C	A	T	T
Posterior Qs	32	34	22	16	35	28	30	34	38	40

Lindgren 2012, *BMC research notes*

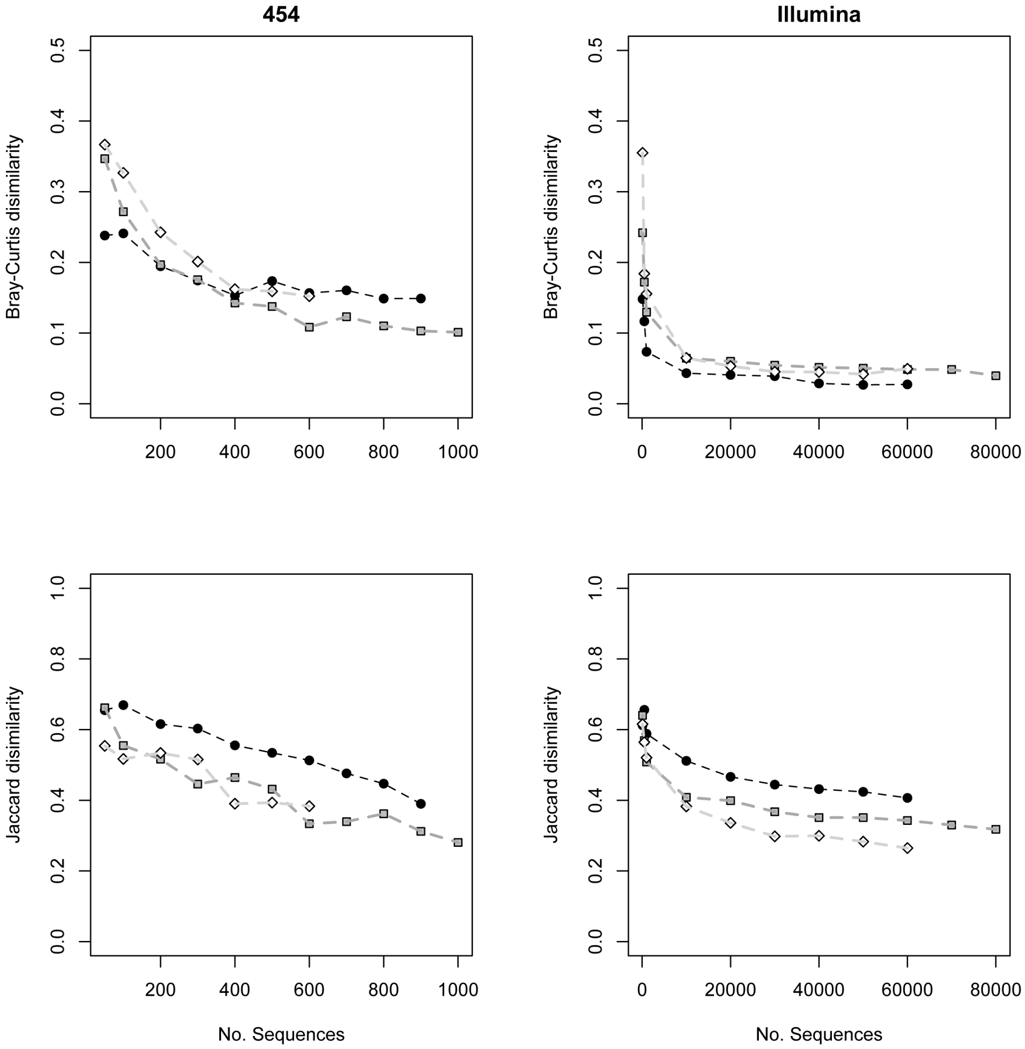
Edgar & Flyvbjerg 2015, *Bioinformatics*

How many reads do I need?



Vic et al. 2013, *Scientific Reports*

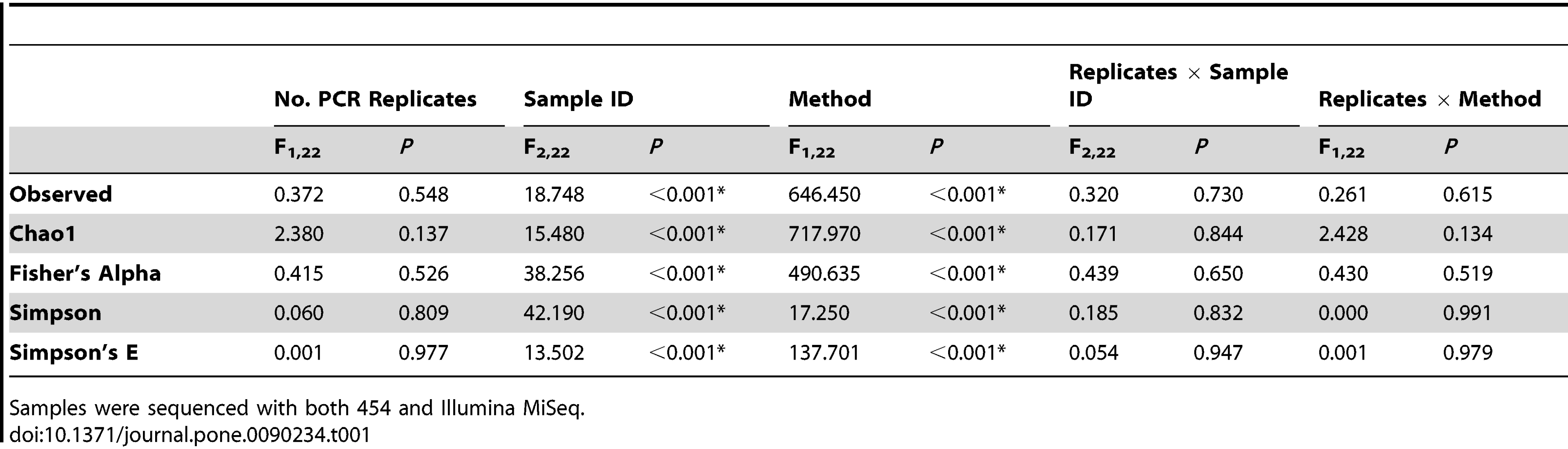
How many reads do I need?



Smith & Peay 2014, *PLoS ONE*

Controls – PCR replication

- ***PCR replication – a common practice not rooted in scientific benefit!***

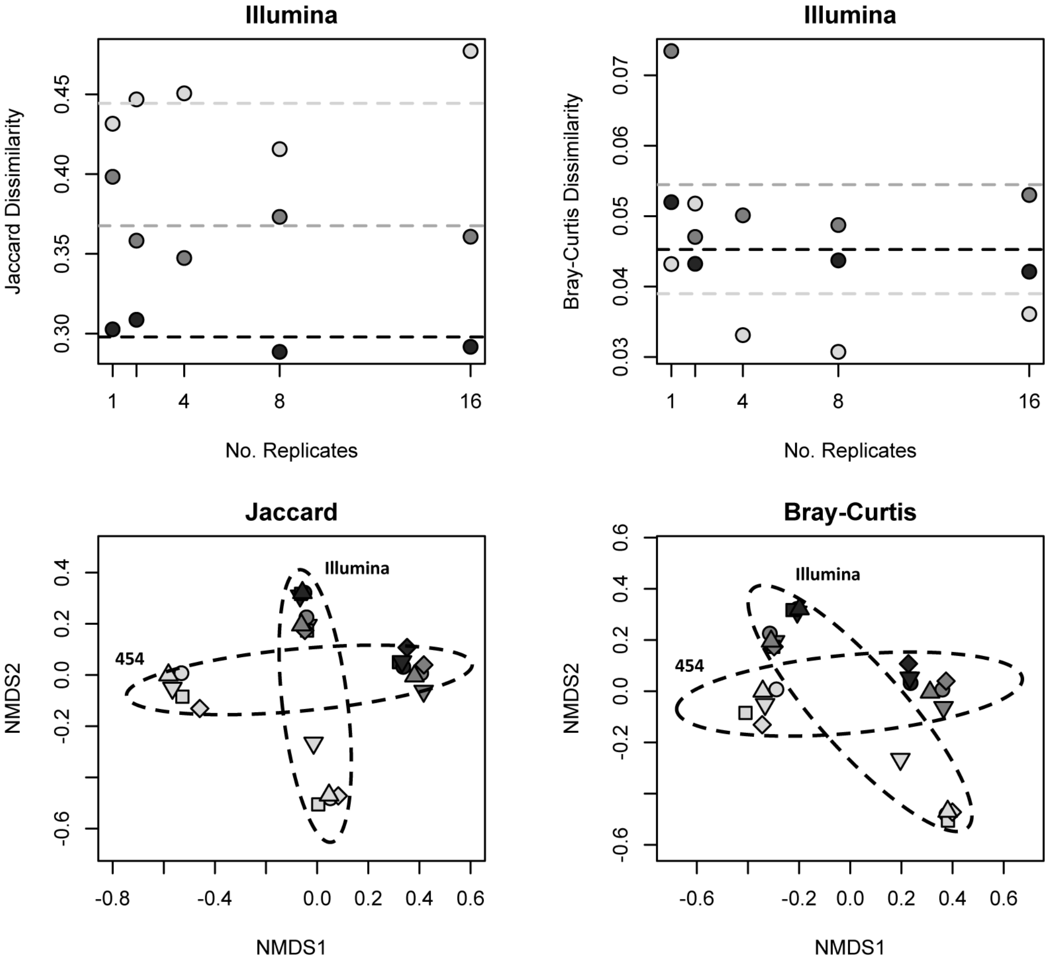


	No. PCR Replicates		Sample ID		Method		Replicates × Sample ID		Replicates × Method	
	$F_{1,22}$	*P*	$F_{2,22}$	*P*	$F_{1,22}$	*P*	$F_{2,22}$	*P*	$F_{1,22}$	*P*
Observed	0.372	0.548	18.748	<0.001*	646.450	<0.001*	0.320	0.730	0.261	0.615
Chao1	2.380	0.137	15.480	<0.001*	717.970	<0.001*	0.171	0.844	2.428	0.134
Fisher's Alpha	0.415	0.526	38.256	<0.001*	490.635	<0.001*	0.439	0.650	0.430	0.519
Simpson	0.060	0.809	42.190	<0.001*	17.250	<0.001*	0.185	0.832	0.000	0.991
Simpson's E	0.001	0.977	13.502	<0.001*	137.701	<0.001*	0.054	0.947	0.001	0.979

Samples were sequenced with both 454 and Illumina MiSeq.
doi:10.1371/journal.pone.0090234.t001

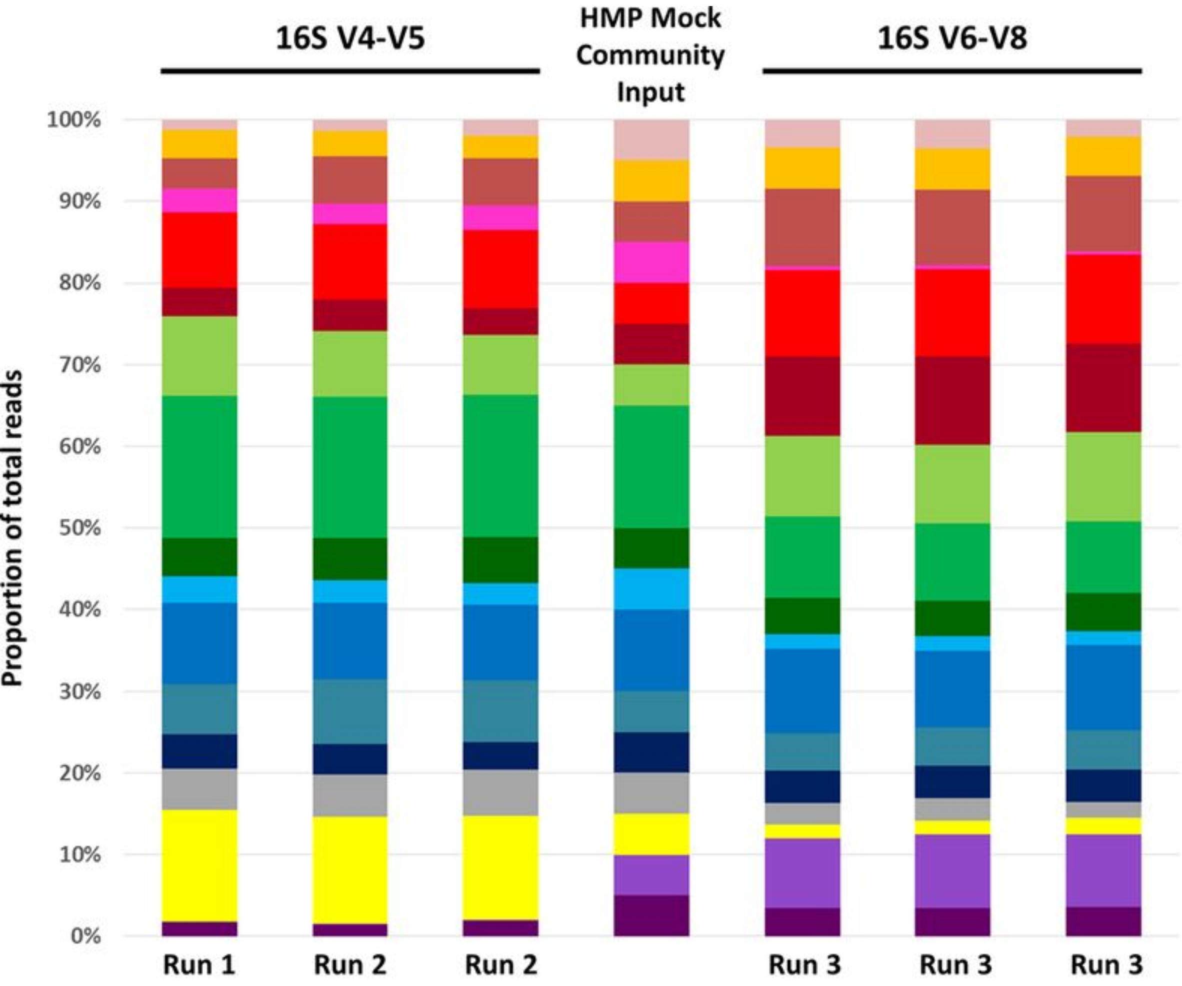
Smith & Peay 2014, *PLoS ONE*

Controls – PCR replication



Smith & Peay 2014, *PLoS ONE*

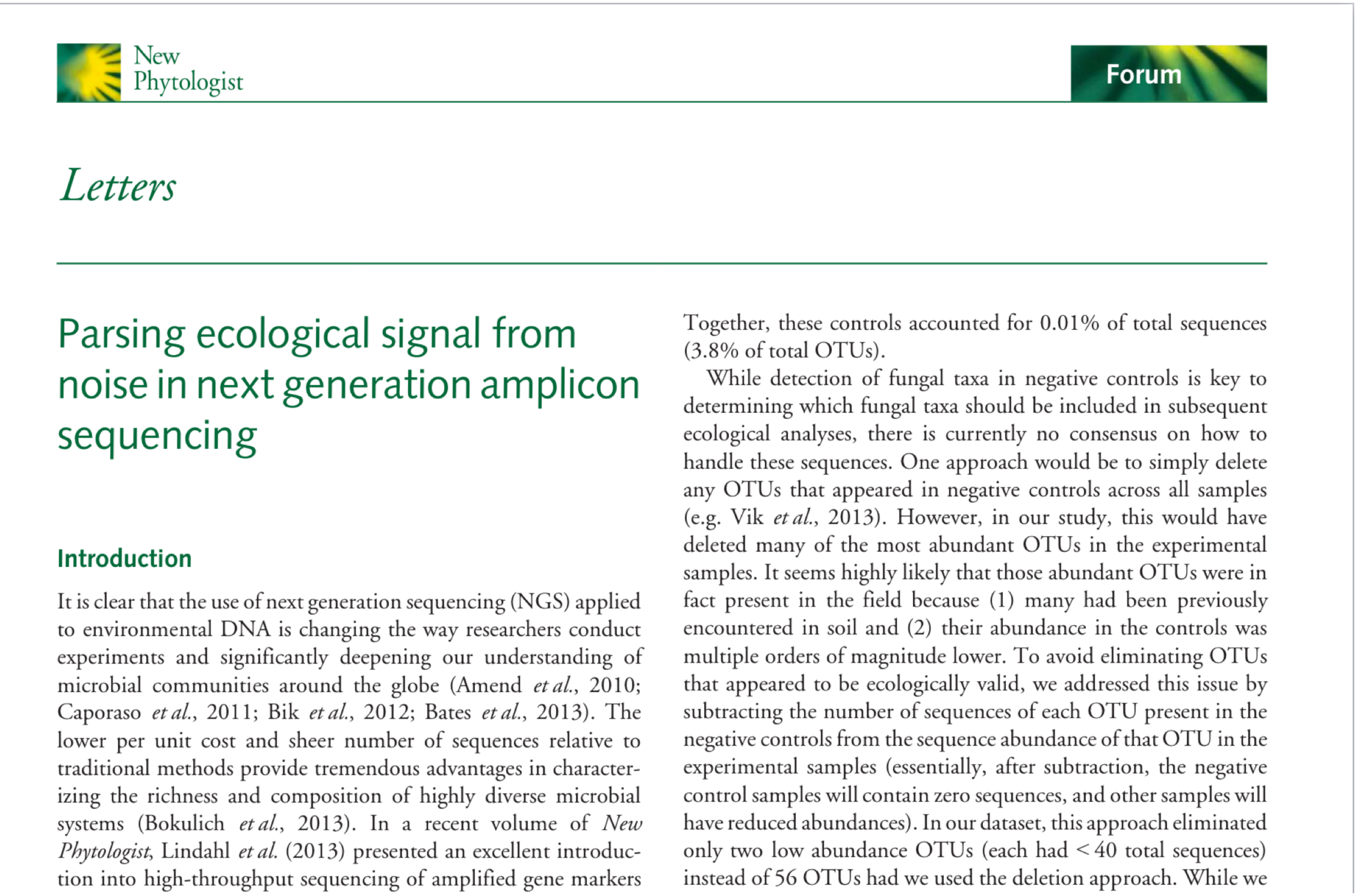
Controls – mock community



		# of OTUs		
		input	V4-V5	V6-V8
Proteo-bacteria	■ *Pseudomonas*	1	1	2
	■ *Acinetobacter*	1	2	2
	■ *Escherichia/Shigella*	1	2	3
	■ *Helicobacter*	1	1	1
	■ *Neisseria*	1	2	2
	■ *Rhodobacter*	1	4	3
Firmicutes	■ *Clostridium*	1	1	3
	■ *Streptococcus*	3	6	4
	■ *Lactobacillus*	1	1	4
	■ *Enterococcus*	1	1	1
	■ *Staphylococcus*	2	4	11
	■ *Listeria*	1	2	3
	■ *Bacillus*	1	3	3
	■ *Deinococcus*	1	1	1
	□ *Bacteroides*	1	3	5
Actino-bacteria	■ *Propionibacterium*	1	1	4
	■ *Actinomyces*	1	2	2
	Total =	20	37	54

Comaeu et al. 2017, *mSystems*

Controls & other concerns



New Phytologist

Forum

Letters

Parsing ecological signal from noise in next generation amplicon sequencing

Introduction

It is clear that the use of next generation sequencing (NGS) applied to environmental DNA is changing the way researchers conduct experiments and significantly deepening our understanding of microbial communities around the globe (Amend *et al.*, 2010; Caporaso *et al.*, 2011; Bik *et al.*, 2012; Bates *et al.*, 2013). The lower per unit cost and sheer number of sequences relative to traditional methods provide tremendous advantages in characterizing the richness and composition of highly diverse microbial systems (Bokulich *et al.*, 2013). In a recent volume of *New Phytologist*, Lindahl *et al.* (2013) presented an excellent introduction into high-throughput sequencing of amplified gene markers

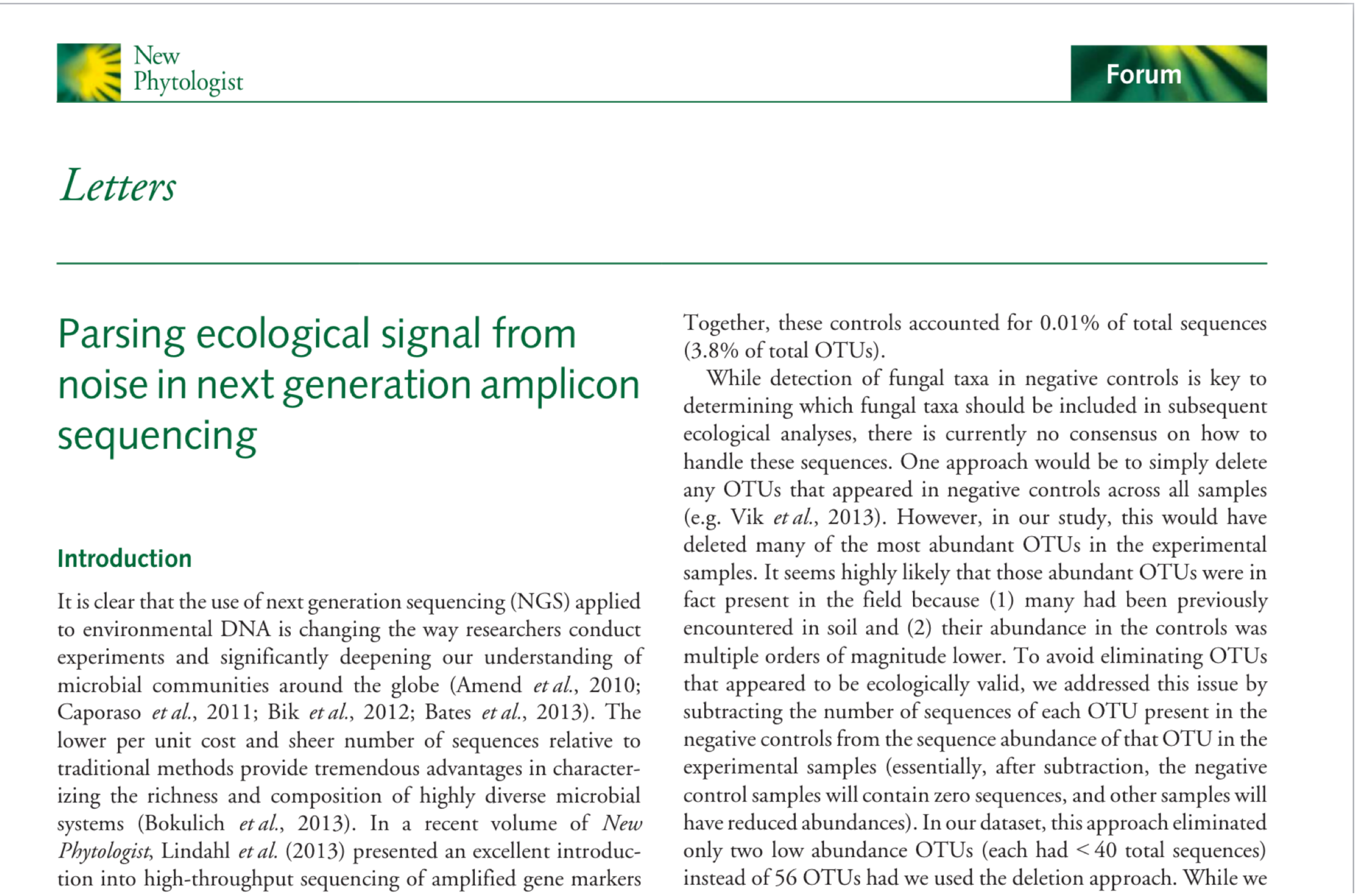
Together, these controls accounted for 0.01% of total sequences (3.8% of total OTUs).

While detection of fungal taxa in negative controls is key to determining which fungal taxa should be included in subsequent ecological analyses, there is currently no consensus on how to handle these sequences. One approach would be to simply delete any OTUs that appeared in negative controls across all samples (e.g. Vik *et al.*, 2013). However, in our study, this would have deleted many of the most abundant OTUs in the experimental samples. It seems highly likely that those abundant OTUs were in fact present in the field because (1) many had been previously encountered in soil and (2) their abundance in the controls was multiple orders of magnitude lower. To avoid eliminating OTUs that appeared to be ecologically valid, we addressed this issue by subtracting the number of sequences of each OTU present in the negative controls from the sequence abundance of that OTU in the experimental samples (essentially, after subtraction, the negative control samples will contain zero sequences, and other samples will have reduced abundances). In our dataset, this approach eliminated only two low abundance OTUs (each had < 40 total sequences) instead of 56 OTUs had we used the deletion approach. While we

- Negative controls
- OTU clustering methods
- Low abundance OTUs
- Singletons

Nguyen et al. 2015, *New Phytologist*

Controls & other concerns



New Phytologist

Forum

Letters

Parsing ecological signal from noise in next generation amplicon sequencing

Introduction

It is clear that the use of next generation sequencing (NGS) applied to environmental DNA is changing the way researchers conduct experiments and significantly deepening our understanding of microbial communities around the globe (Amend *et al.*, 2010; Caporaso *et al.*, 2011; Bik *et al.*, 2012; Bates *et al.*, 2013). The lower per unit cost and sheer number of sequences relative to traditional methods provide tremendous advantages in characterizing the richness and composition of highly diverse microbial systems (Bokulich *et al.*, 2013). In a recent volume of *New Phytologist*, Lindahl *et al.* (2013) presented an excellent introduction into high-throughput sequencing of amplified gene markers

Together, these controls accounted for 0.01% of total sequences (3.8% of total OTUs).

While detection of fungal taxa in negative controls is key to determining which fungal taxa should be included in subsequent ecological analyses, there is currently no consensus on how to handle these sequences. One approach would be to simply delete any OTUs that appeared in negative controls across all samples (e.g. Vik *et al.*, 2013). However, in our study, this would have deleted many of the most abundant OTUs in the experimental samples. It seems highly likely that those abundant OTUs were in fact present in the field because (1) many had been previously encountered in soil and (2) their abundance in the controls was multiple orders of magnitude lower. To avoid eliminating OTUs that appeared to be ecologically valid, we addressed this issue by subtracting the number of sequences of each OTU present in the negative controls from the sequence abundance of that OTU in the experimental samples (essentially, after subtraction, the negative control samples will contain zero sequences, and other samples will have reduced abundances). In our dataset, this approach eliminated only two low abundance OTUs (each had < 40 total sequences) instead of 56 OTUs had we used the deletion approach. While we

- Negative controls
- OTU clustering methods
- Low abundance OTUs
- Singletons

Nguyen et al. 2015, *New Phytologist*

OTUs vs ASVs

- ***OTUs (Operational Taxonomic Units)***
  - ***Several to many sequences ”collapsed” into one reference sequence based on a discrete sequence similarity threshold***
  - ***Functionally equivalent to ”species”***
  - ***Several ways to delineate and pick OTUs***

- ***ASV (Amplicon Sequence Variants)***
  - ***Each sequence treated as a piece of data with a taxonomy assignment***
  - ***No sequence collapsing***
  - ***More***

Callahan et al. 2017, *ISME*

iTag with PacBio

- ***PacBio – Single molecule real time sequencing***
- ***Pros:***
  - ***Potential for much longer read length***
  - ***Better phylogenetic potential***
  - ***Better taxonomic assignments to reads***
- ***Cons:***
  - ***Significantly higher error rate****
  - ***Significantly lower throughput***

Genomics Core Seminar Series

THANK YOU!!!!