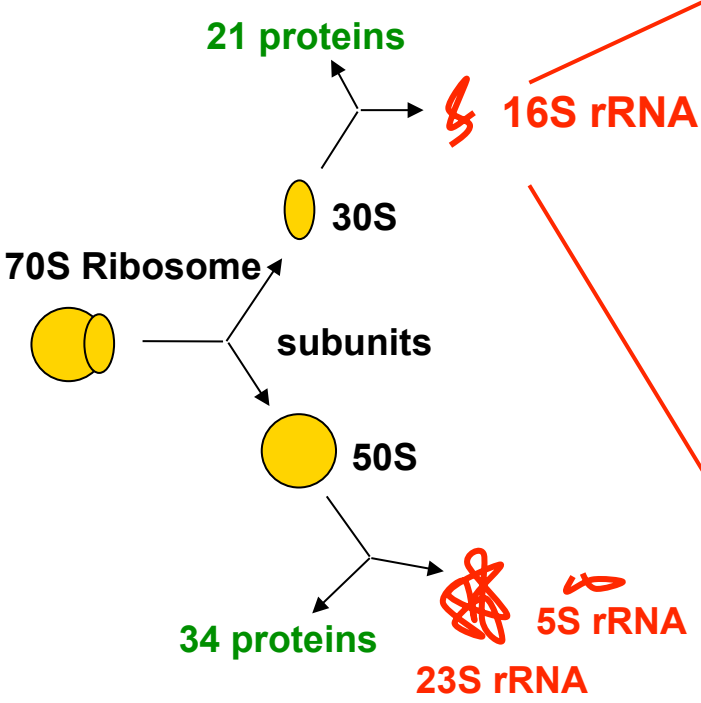
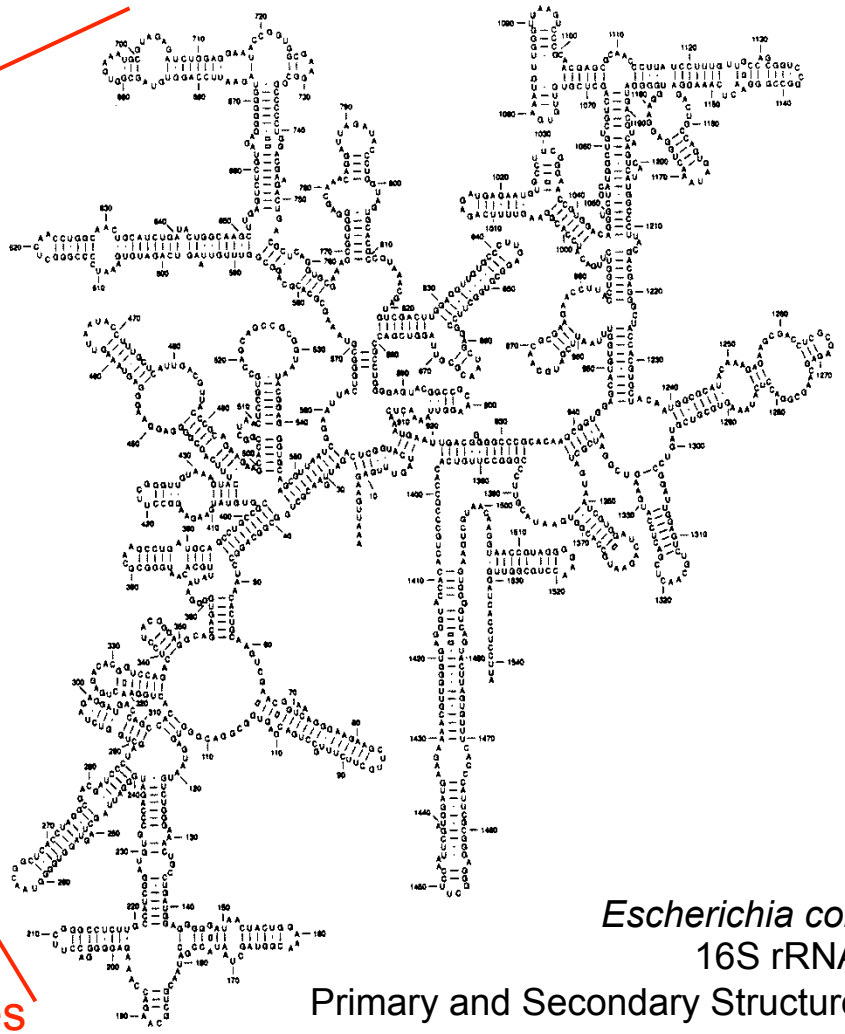
Evaluation of multiplexed 16S rRNA microbial population surveys using Illumina MiSeq platform

Julien Tremblay, PhD
SFAF, Santa Fe, NM – June 7th 2012

16S rRNA as phylogenetic marker gene



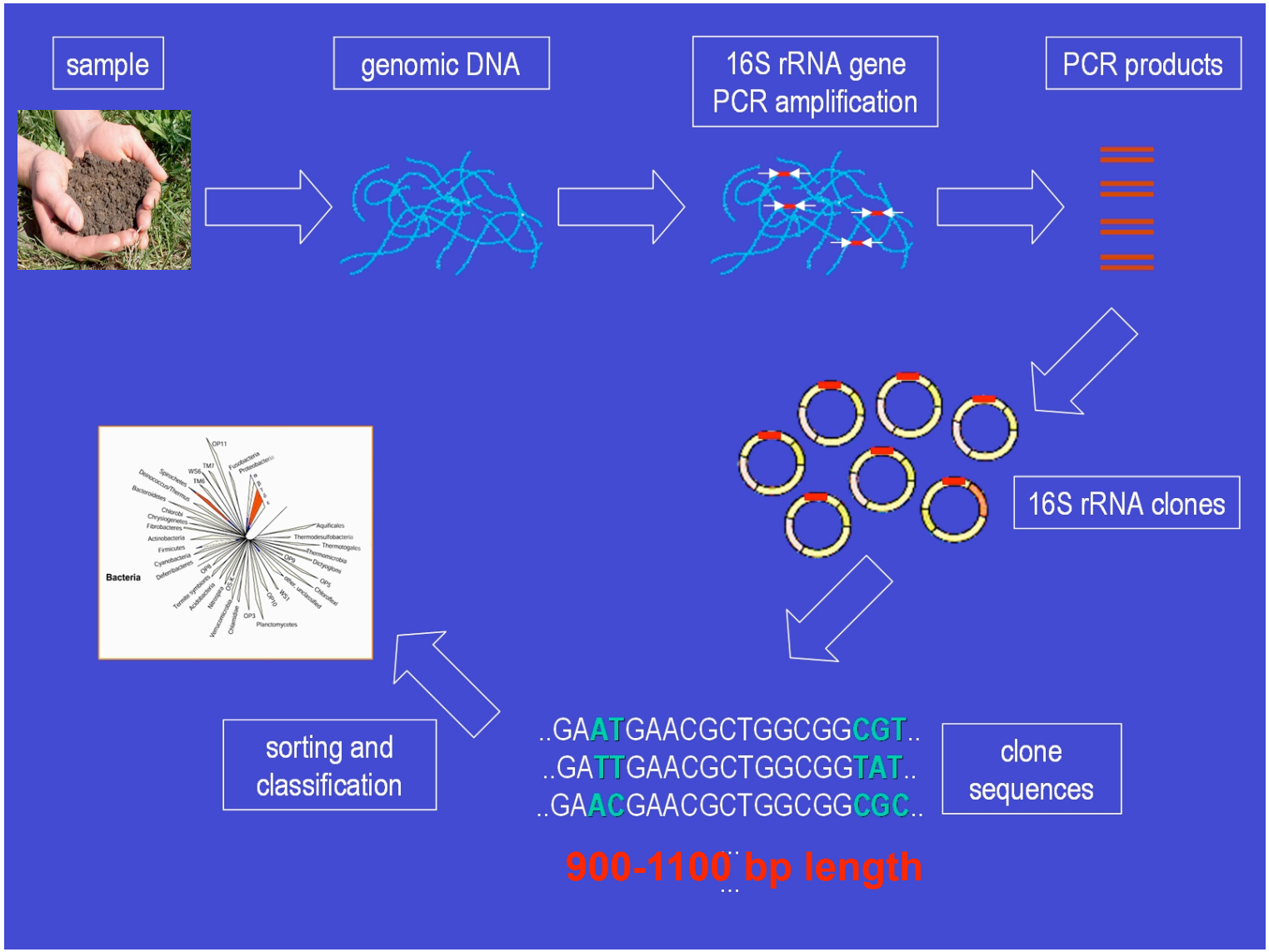


Escherichia coli
16S rRNA
Primary and Secondary Structure

highly conserved between different species of bacteria and archaea

Falk Warnecke

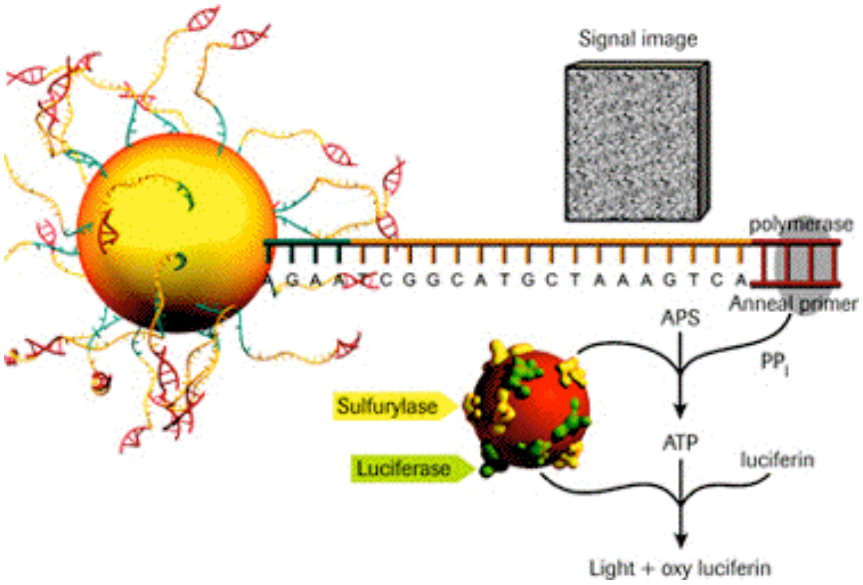
16S rRNA in environmental microbiology (Sanger clone libraries)



Falk Warnecke

Next generation sequencing (NGS)

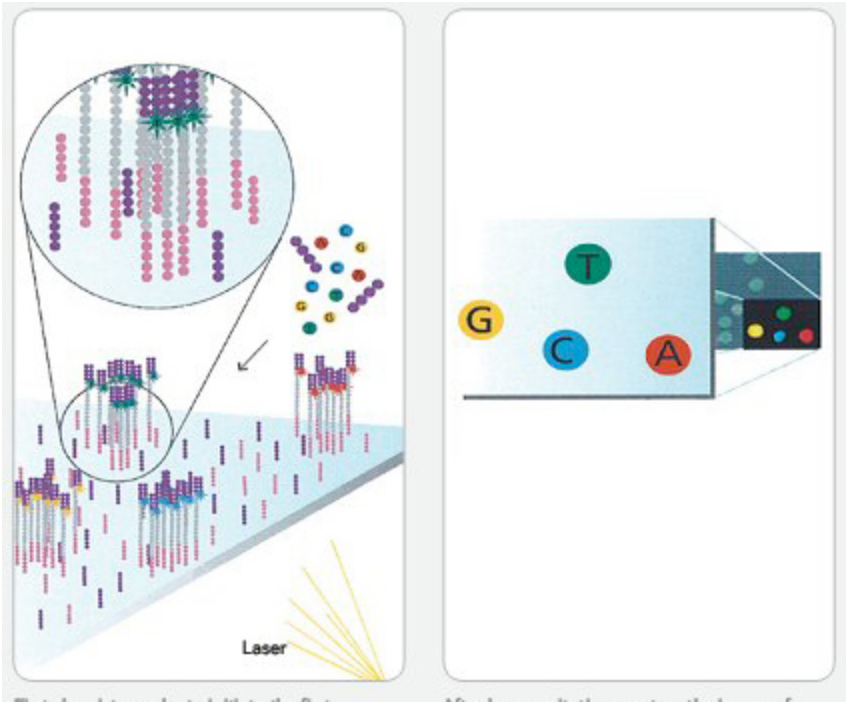
454



0.5M 450bp reads

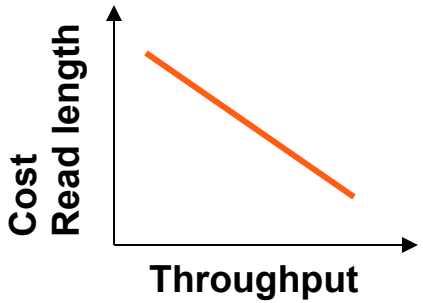
$$

Illumina

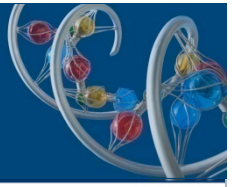


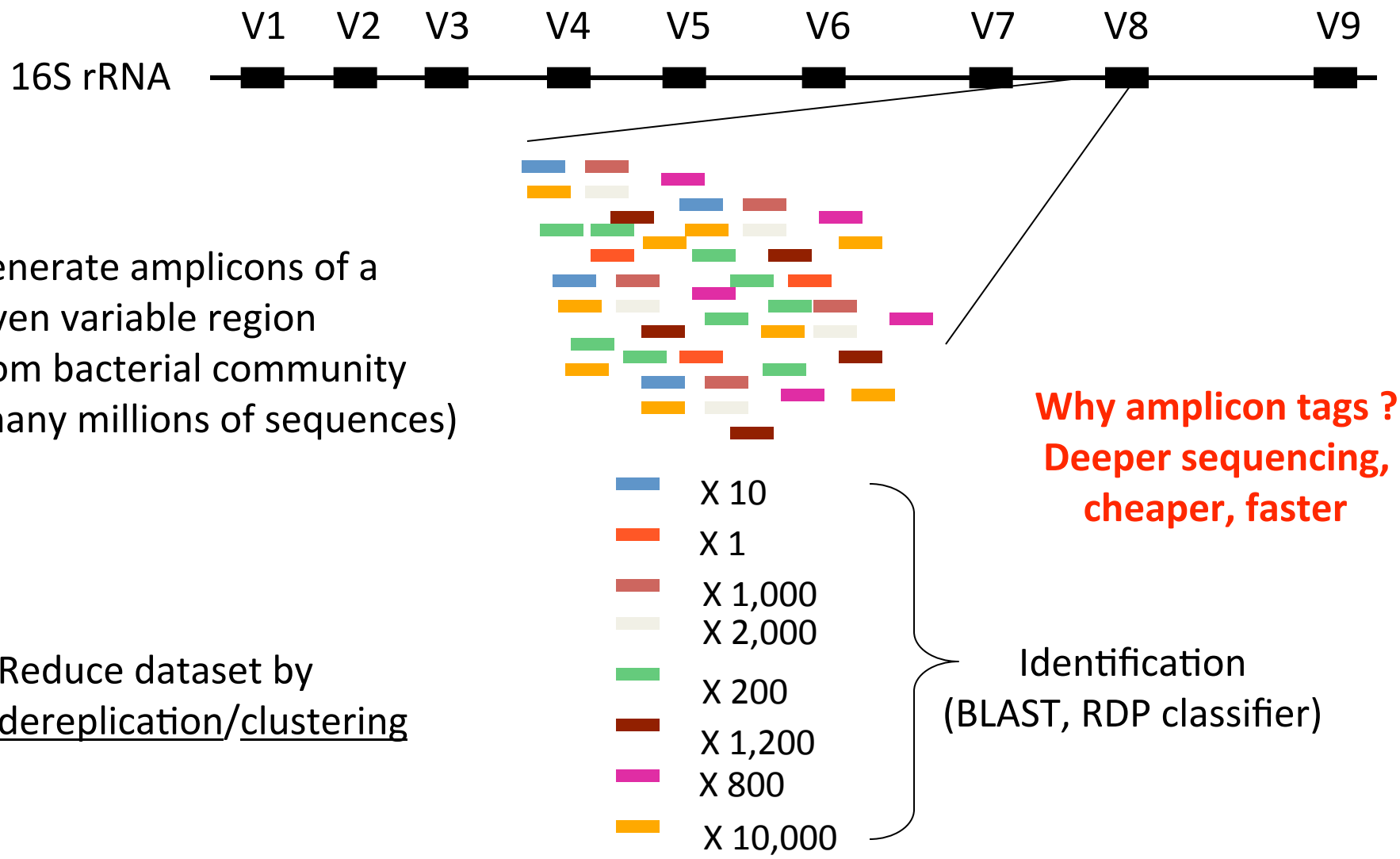
10-350M 150bp reads/lane

$



Surveying microbial diversity with short 16S amplicons





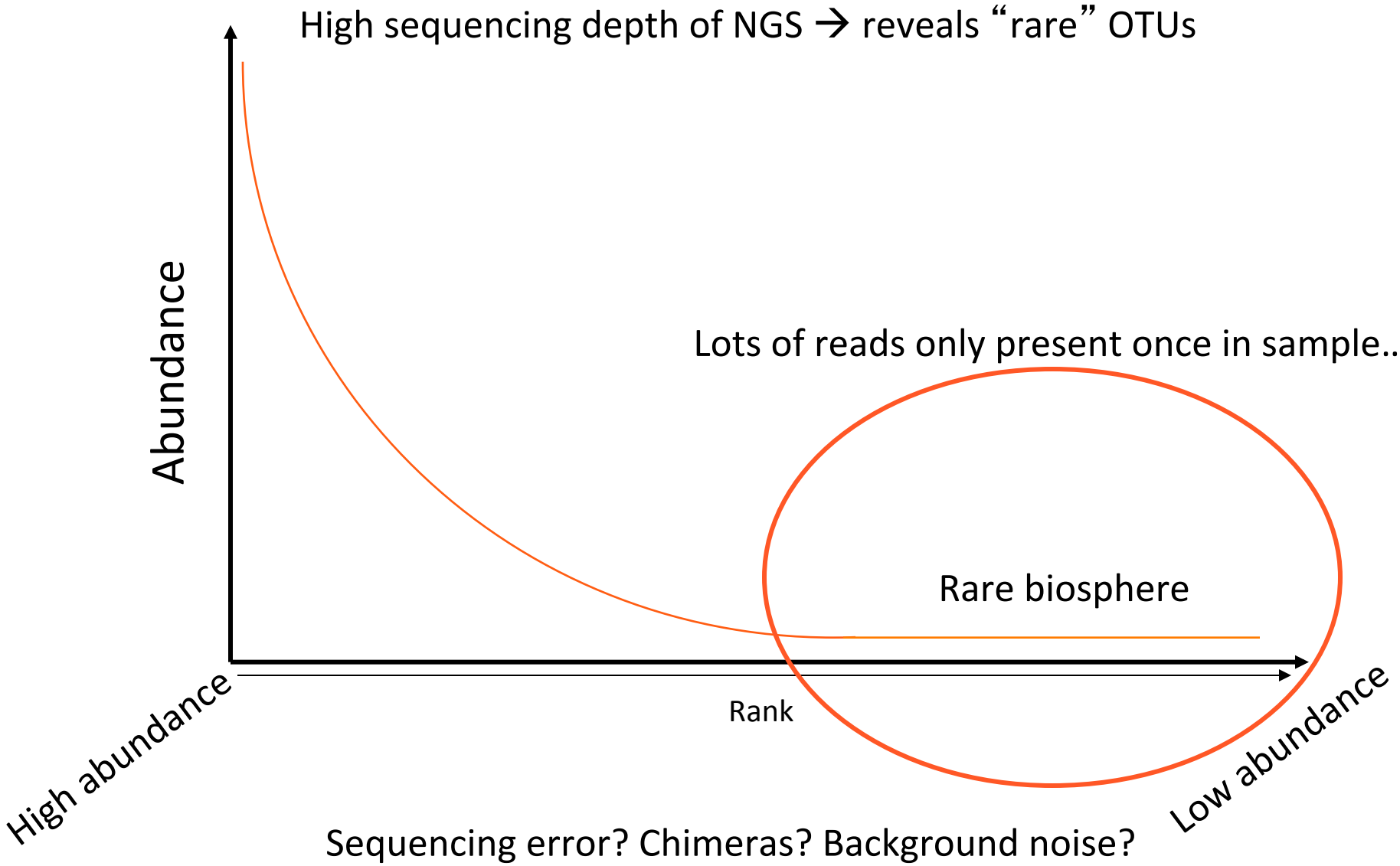
Generate amplicons of a given variable region from bacterial community (many millions of sequences)

Why amplicon tags ? Deeper sequencing, cheaper, faster

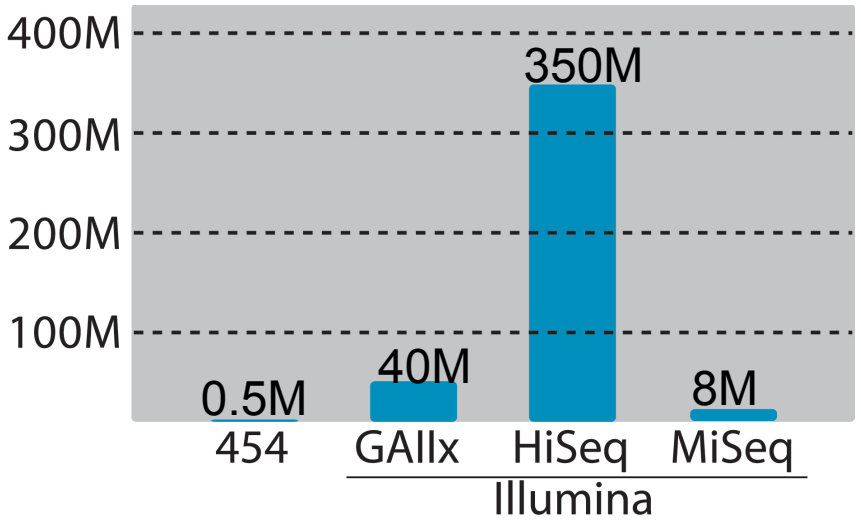
Reduce dataset by dereplication/clustering

Identification (BLAST, RDP classifier)

Rare biosphere

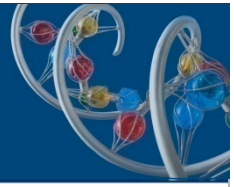


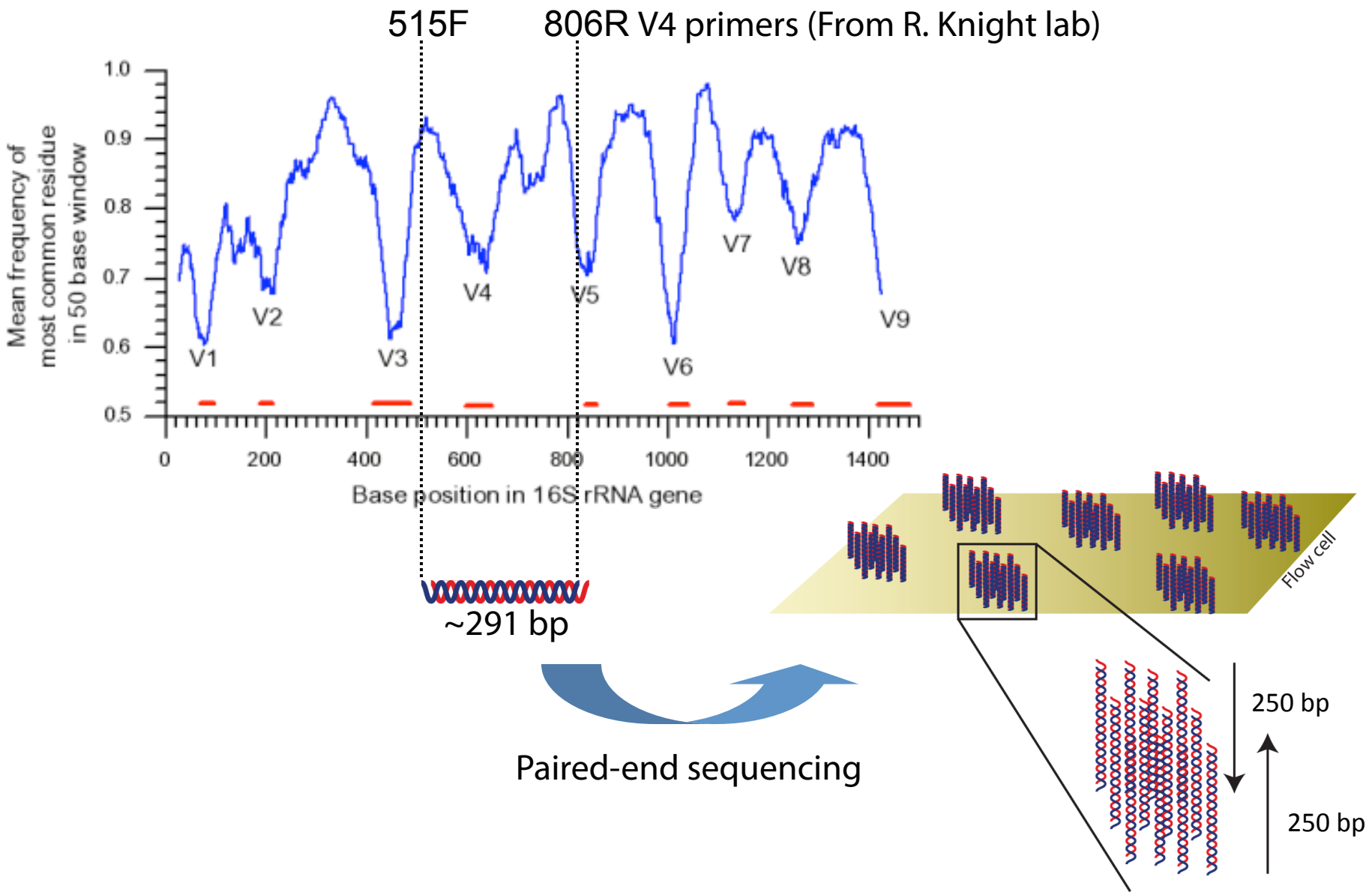
Platforms throughput



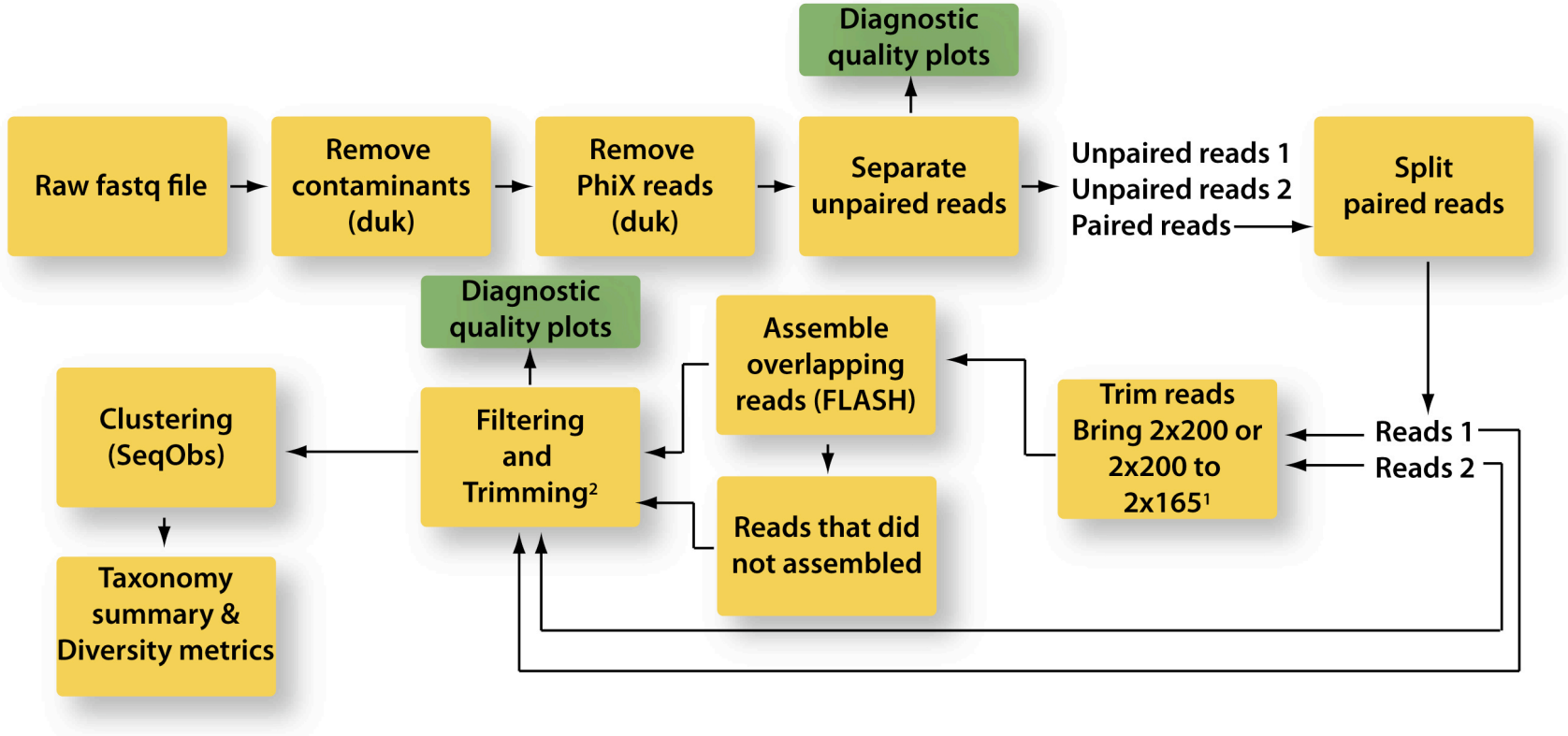
- **Move 16S tags sequencing from 454 to Illumina MiSeq platform**
  - **$ and higher quality reads with Illumina**
  - **HiSeq = Slow (18 days) and huge output compared to 454 (suitable for big projects 1000+ indexes(barcodes)/libraries**
  - **MiSeq = moderately high throughput (More suitable for standard projects) and fast (~30 hrs run)**
  - **Longer reads with MiSeq (250 bp) compared to HiSeq (150 bp)**
- **↑throughput → Clustering algorithm development.**

Sequencing design for 16S tags on MiSeq





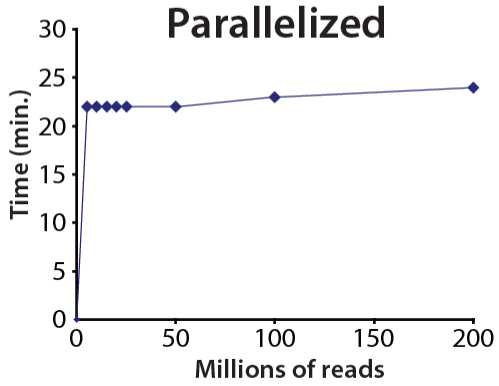
Itags pipeline



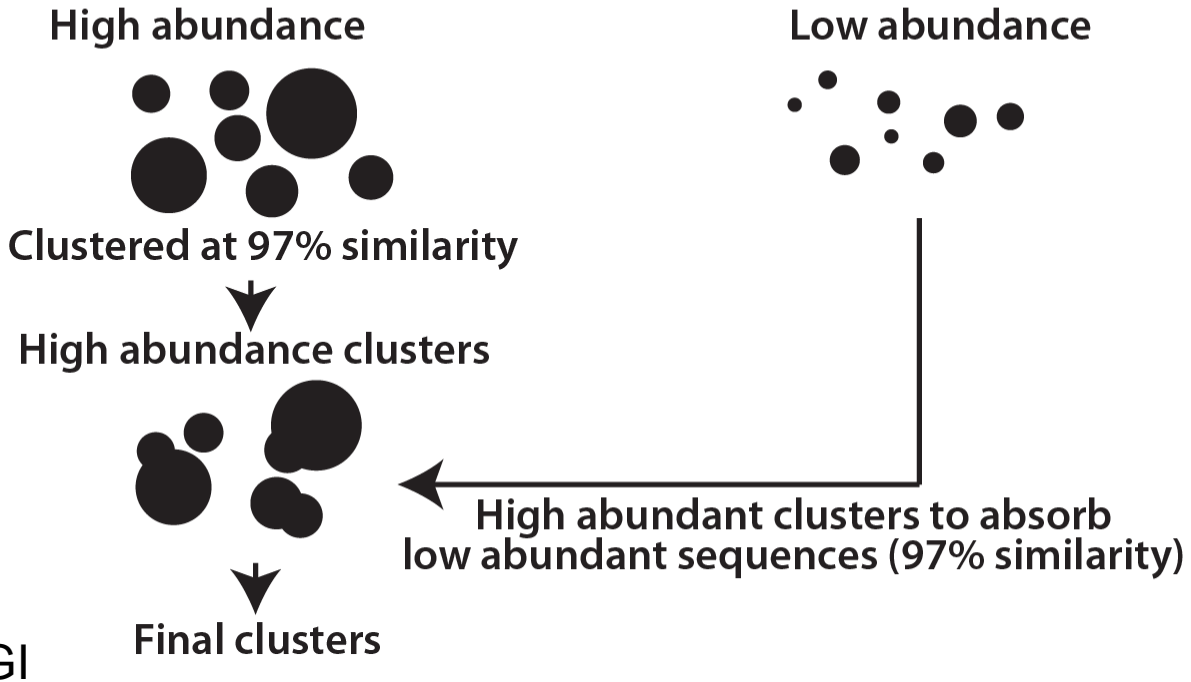
- Remove low quality reads
- Reduce dataset
- Provide diversity metrics and taxonomic classification

16S tags clustering

Sequences sorted by alphabetical order
Counts of each unique sequences (100% dereplication)

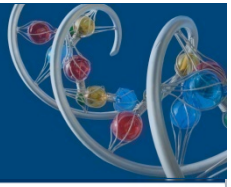


Dereplicated sequences binned by abundance:



Edward Kirton, JGI

Reads count summary

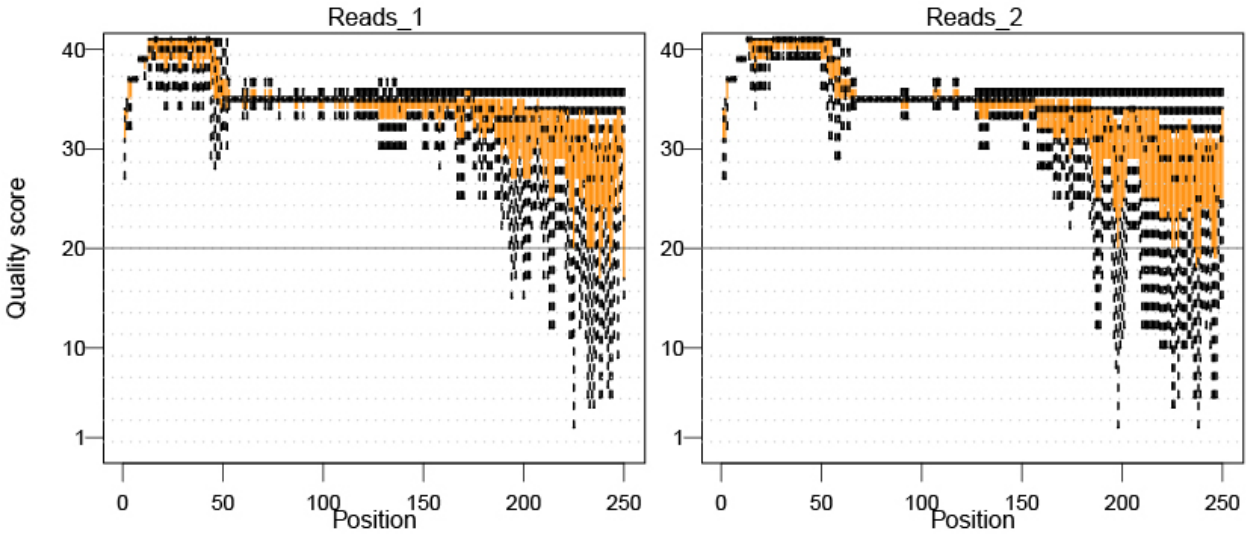


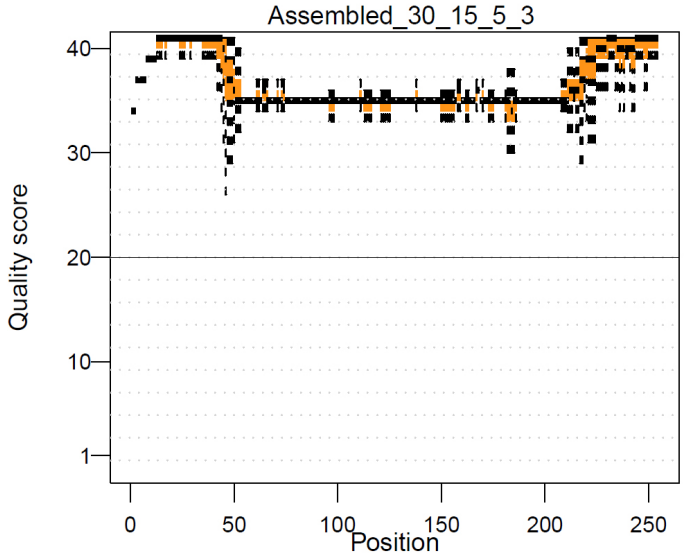
- **MiSeq 96 barcodes typical run**

Total reads	7.6M
Contaminants reads	0.12M
PhiX reads	2M
Paired reads 1 (non-PhiX)	2.7M
Paired reads 2 (non-PhiX)	2.7M
Assembled reads	2.5M
Assembled reads QC passed	2.1M
Assembled reads / barcodes	22,574 ± 8,244

How is quality?

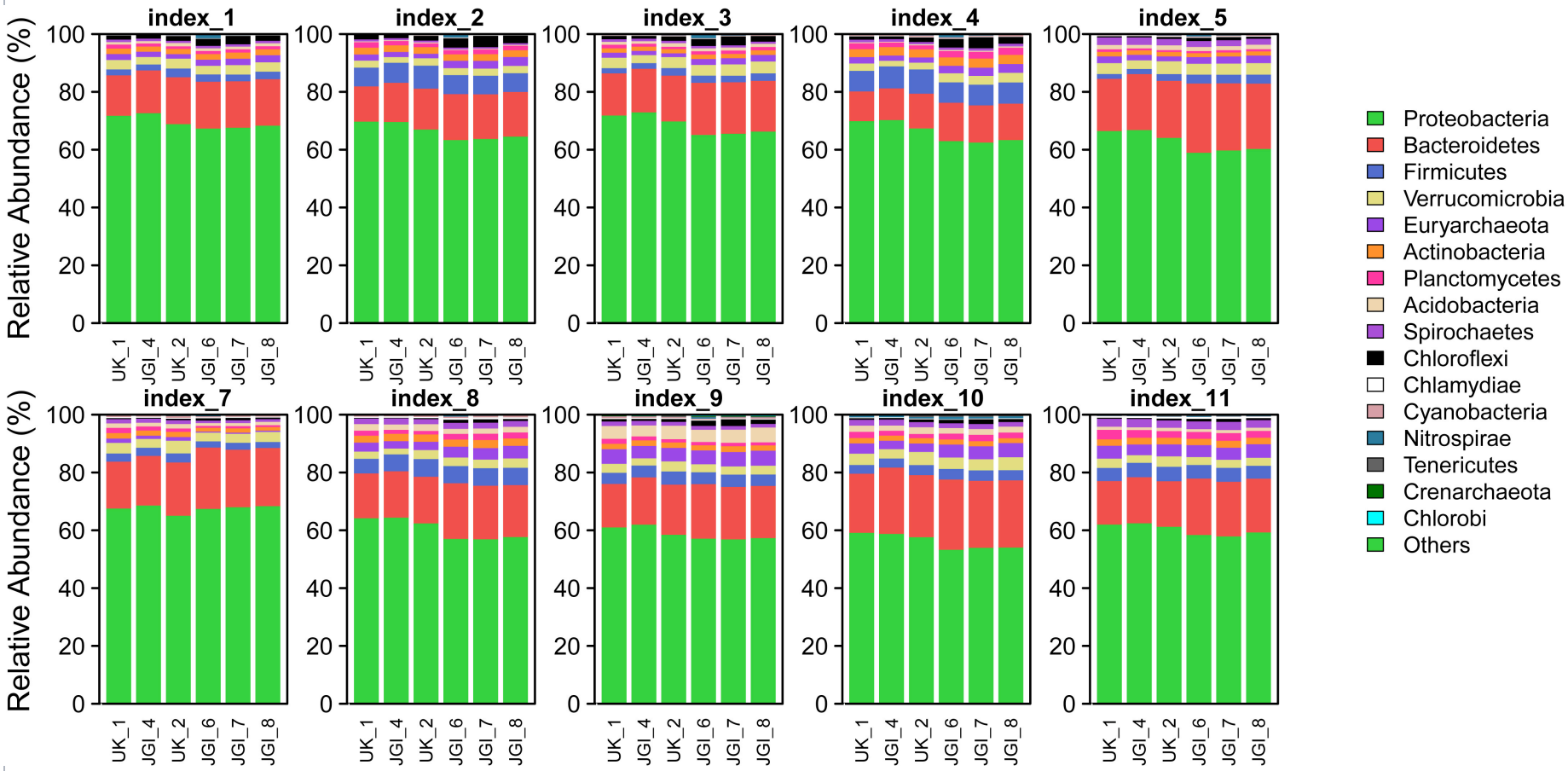
- **2x250 runs**





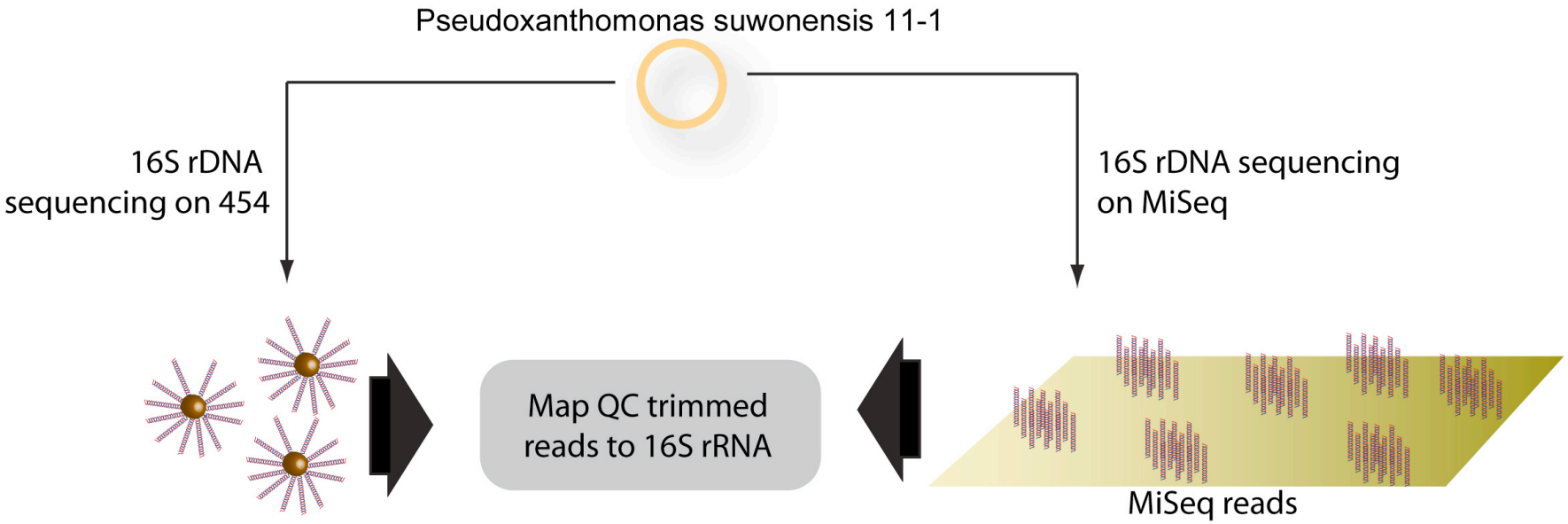
MiSeq 16S tags sequencing reproducibility

- **Multiple runs with same sample**
  - **Highly similar taxonomy**



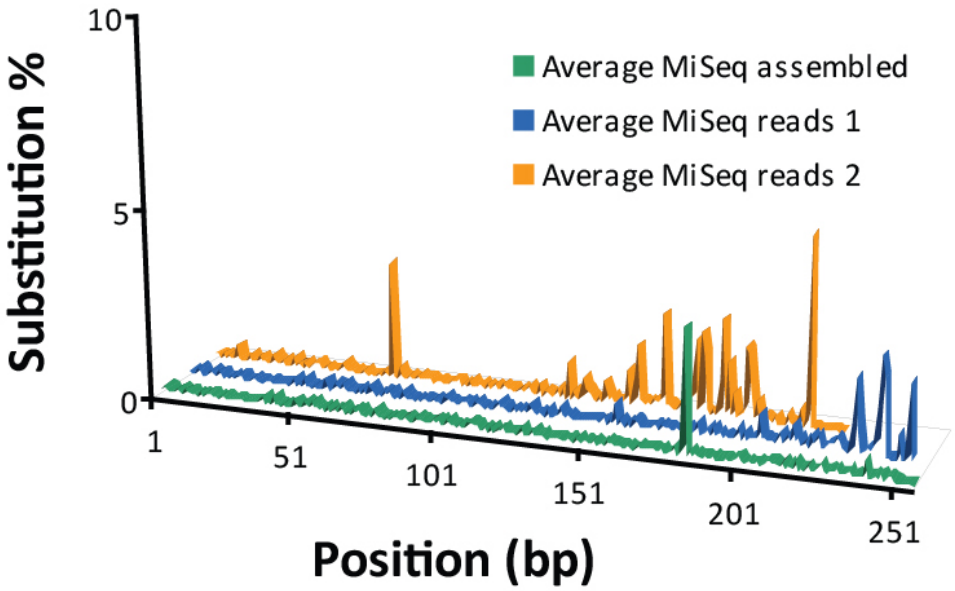
How does 16S tags on MiSeq compares with 454?

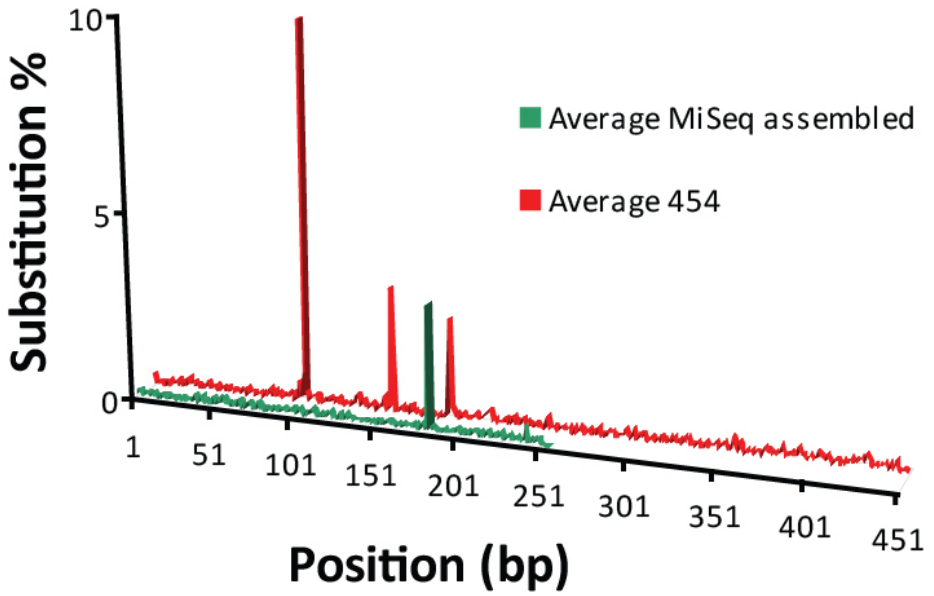
- **Is quality better on MiSeq?**
  - **Sequence 16S rDNA amplicons of a single microbial genome on MiSeq and 454**
  - **Map reads to 16S rDNA and estimate error**



Each experiment done in triplicate

Quality of MiSeq reads > 454

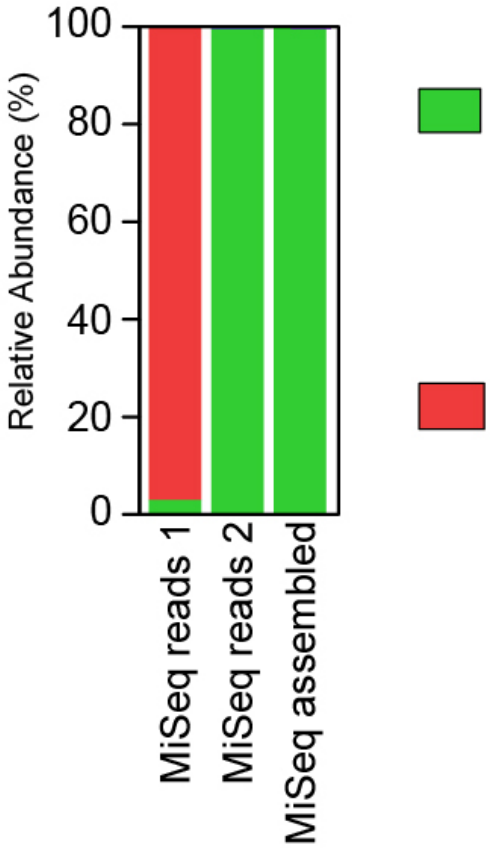




Error type (error/ 1,000,000 sequenced bases)	MiSeq reads 1	MiSeq reads 2	MiSeq assembled	454
Insertion rate	0.49	0.14	0.39	2,365
Deletion rate	82	62	13	2,816
Substitution rate	1,338	1,827	1,163	4,392

Taxonomic classification

- **Advantage of using assembled reads?**



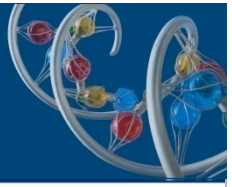
p__Proteobacteria
c__Gammaproteobacteria
o__Xanthomonadales;
f__Xanthomonadaceae
g__Pseudoxanthomonas
s__Pseudoxanthomonas taiwanensis

p__Proteobacteria
c__Gammaproteobacteria
o__Xanthomonadales
f__Xanthomonadaceae
g__Xanthomonas
s__Xanthomonas axonopodi

So why bother assembling reads?

	Count	Average length
assembled	63,352	250 ± 1.3
reads 1	64,070	188 ± 36.2
reads 2	57,656	153 ± 24.3

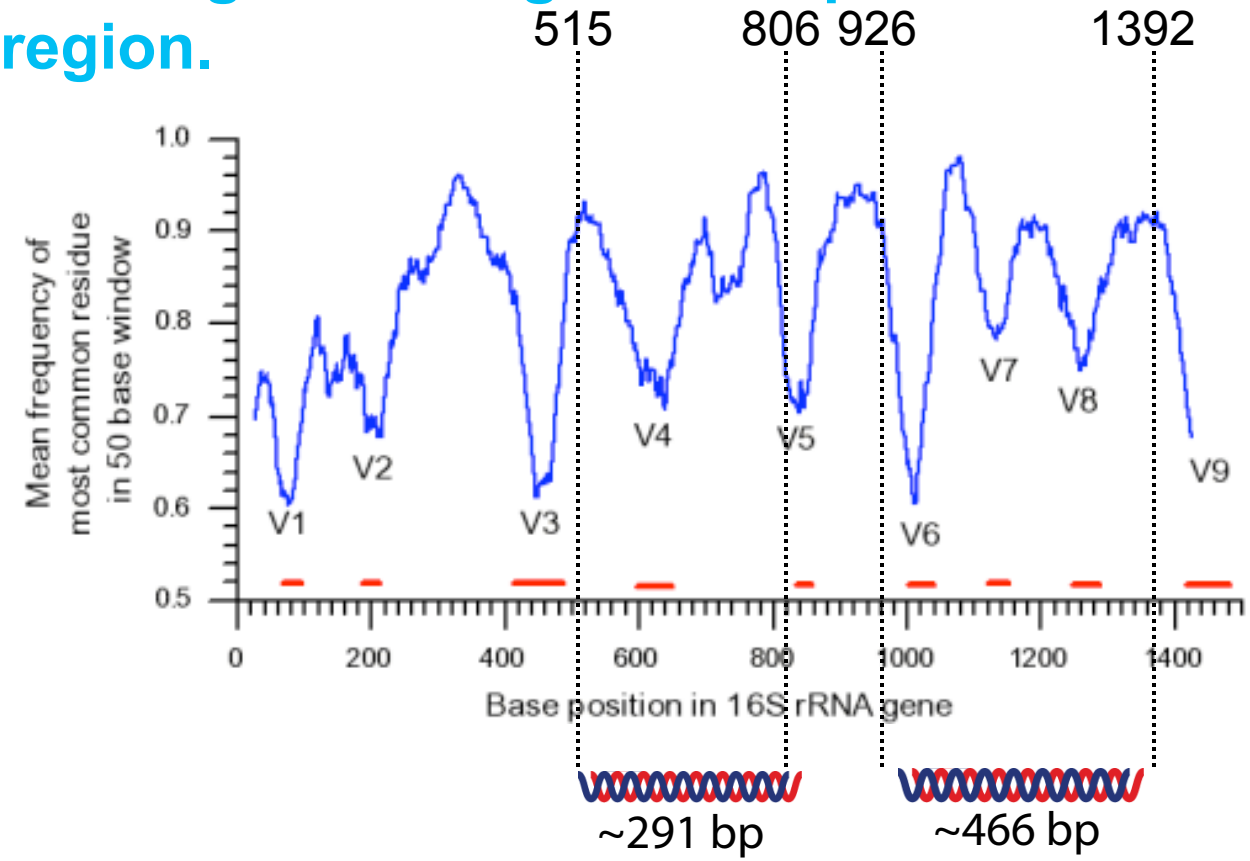
Conclusion



- **MiSeq offers deeper sequencing than 454**
  - Better (assembled) reads quality on MiSeq
  - Lower error rate
  - Higher throughput
  - lower cost
  - Higher multiplexing
- **Cons**
  - Lower read length (Compensated by high quality of reads?)
  - Longer primers (more expensive)
  - Sequencing run is fast, but library preparation time is long.

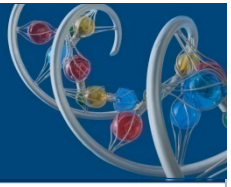
Future directions

- **Compare runs of 454 and MiSeq of same sample**
  - **Although challenge to compare V4 with V6-8 region.**



JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

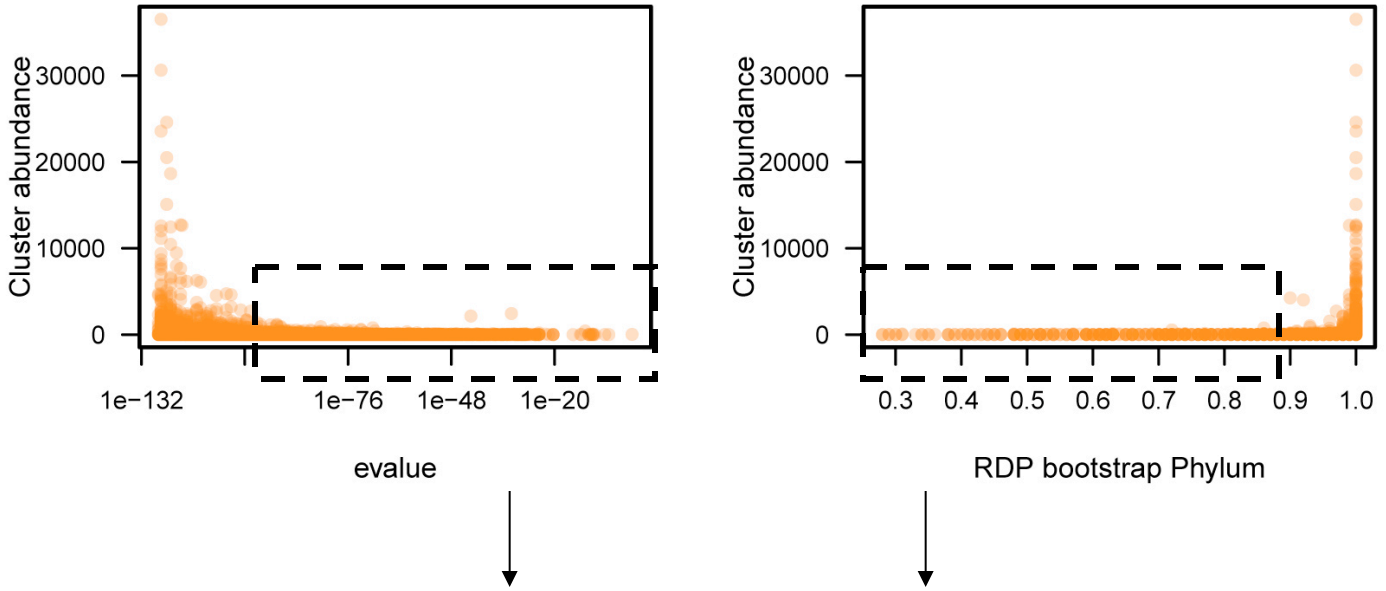
Acknowledgements



- Susannah Tringe
- Feng Chen
- Kanwar Singh
- Ed Kirton
- Alison Fern
- Chris Daum
- Christine Naca
- And more!

itags – rare OTUs

MiSeq wetlands samples



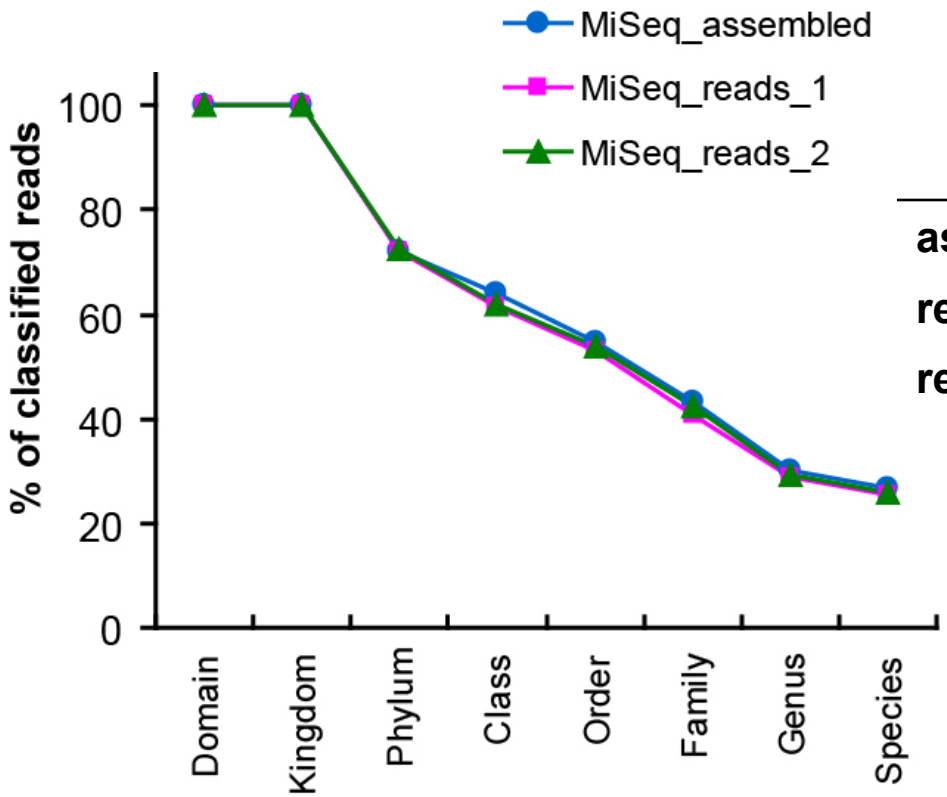
Low abundant reads consistently shows low confidence in Classification.

Low abundant reads = errors, artifacts?

Low abundant reads are underrepresented in databases?

Taxonomic classification

- **Advantages of using assembled reads?**
  - **Test samples from a wetlands site (highly complex environment)**

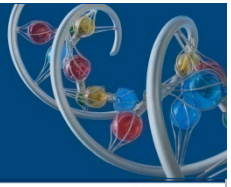


	Count	Average length
assembled	960,125	251 ± 0.6
reads 1	1,016,129	219 ± 27.4
reads 2	1,029,569	185 ± 22.1

	Average RDP bootstrap value at class level
assembled	0.68
reads 1	0.66
reads 2	0.64

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

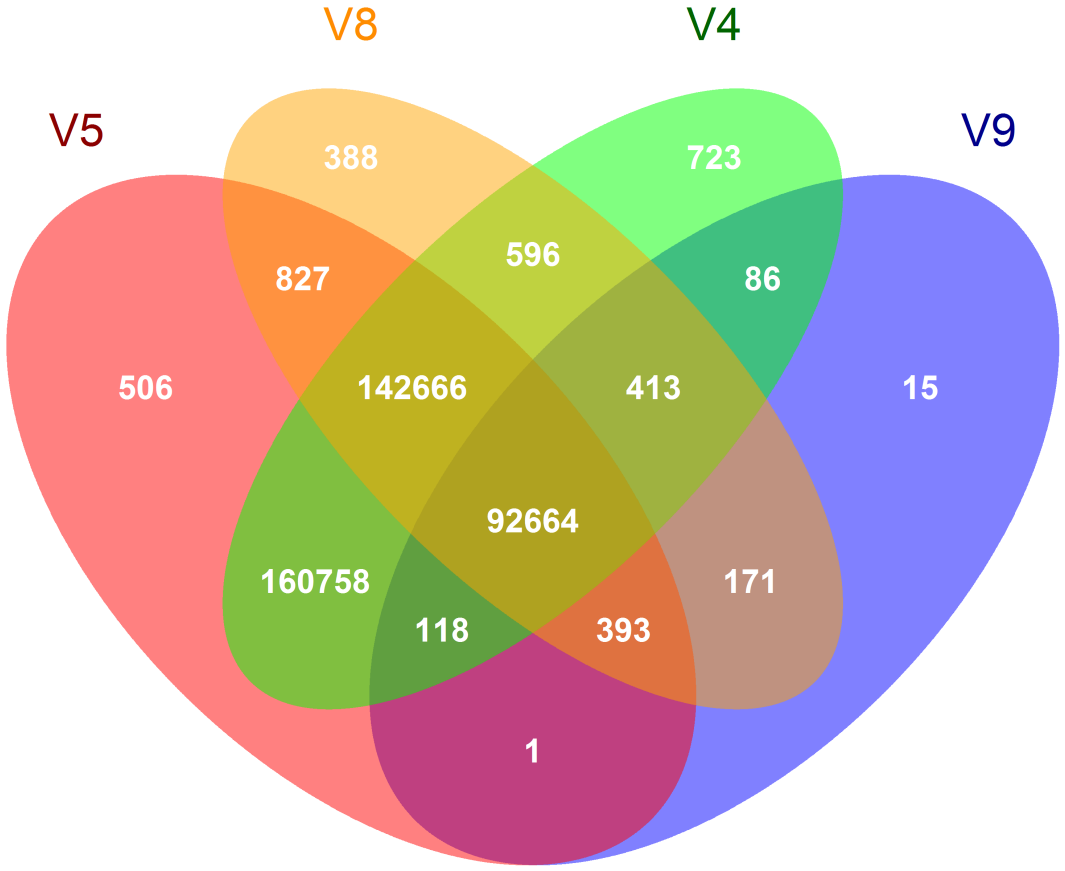
Future directions



- Short size of amplicon
- What filtering parameters to use (stringency level)?
  - → balance between stringency filter and keeping as much data as we can
- Whole new dimension for rare biosphere?
- Handling large numbers of sample (tens of thousand magnitude)
- Sequencing run is fast, but library preparation time is long.
- Cost of barcoded primers (will need lots of barcodes), handling.
- Huge ammount of samples → statistics models…

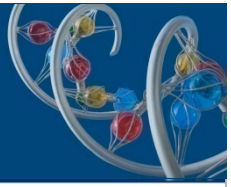
Comparing 454 with illumina

- **Primer pair of variable region is likely to affect outcome of results.**



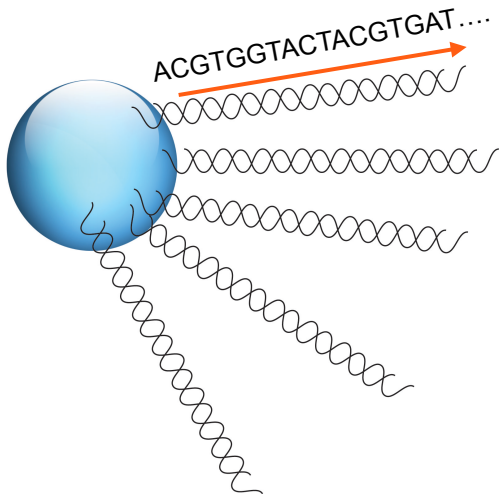
In silico PCR on 16S Greengenes database.

454 and Illumina technology



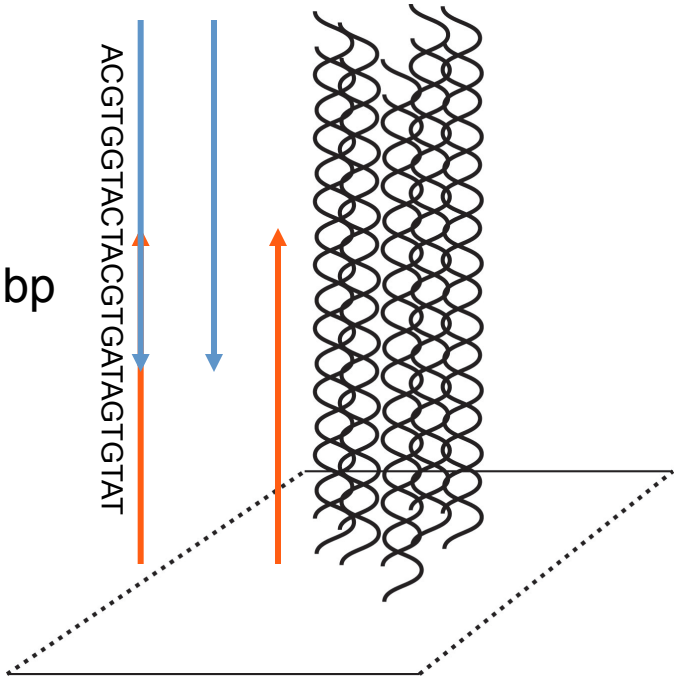
454

~400 bp



Illumina

~250 bp

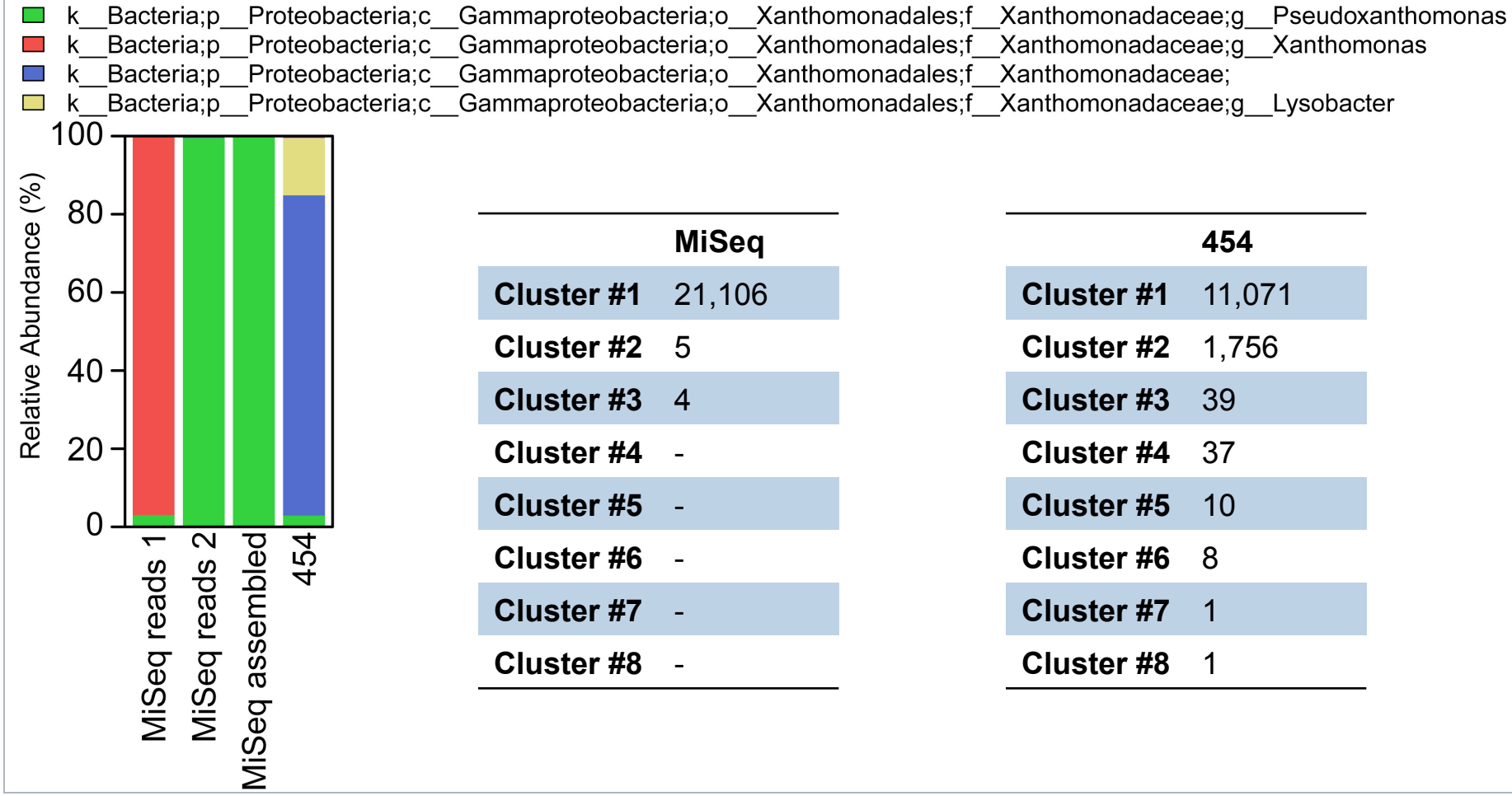


- **454 = "1" read**
- **Illumina = "2" reads => have to be assembled**
  - **Both reads need to be of good quality**

Taxonomic classification

- **Advantage of using assembled reads?**

Raw reads → assembled → QC → clustering → classification with RDP classifier



	MiSeq
Cluster #1	21,106
Cluster #2	5
Cluster #3	4
Cluster #4	-
Cluster #5	-
Cluster #6	-
Cluster #7	-
Cluster #8	-

	454
Cluster #1	11,071
Cluster #2	1,756
Cluster #3	39
Cluster #4	37
Cluster #5	10
Cluster #6	8
Cluster #7	1
Cluster #8	1