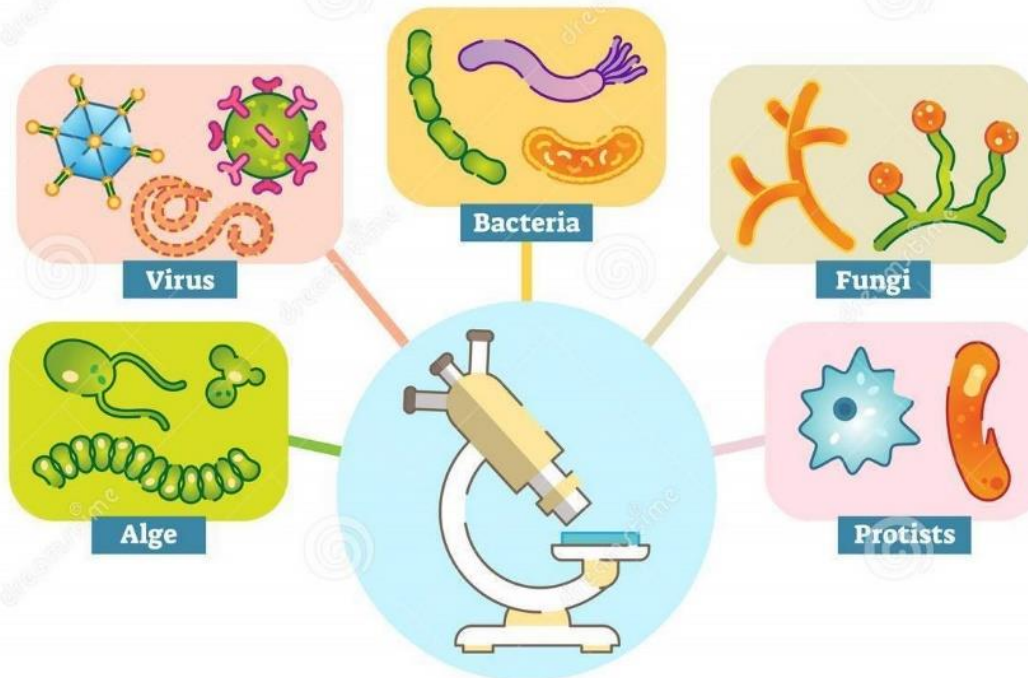


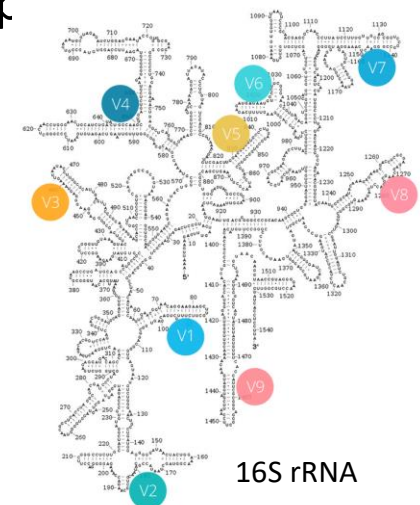
Why do we study metagenomics?

- Metagenomics is the study of genetics of microorganisms such as bacteria, virus, protozoa, fungi, and algae
- Essential components of life on earth
- Bacteria: benefits and adversaries



Classification of bacteria genomes

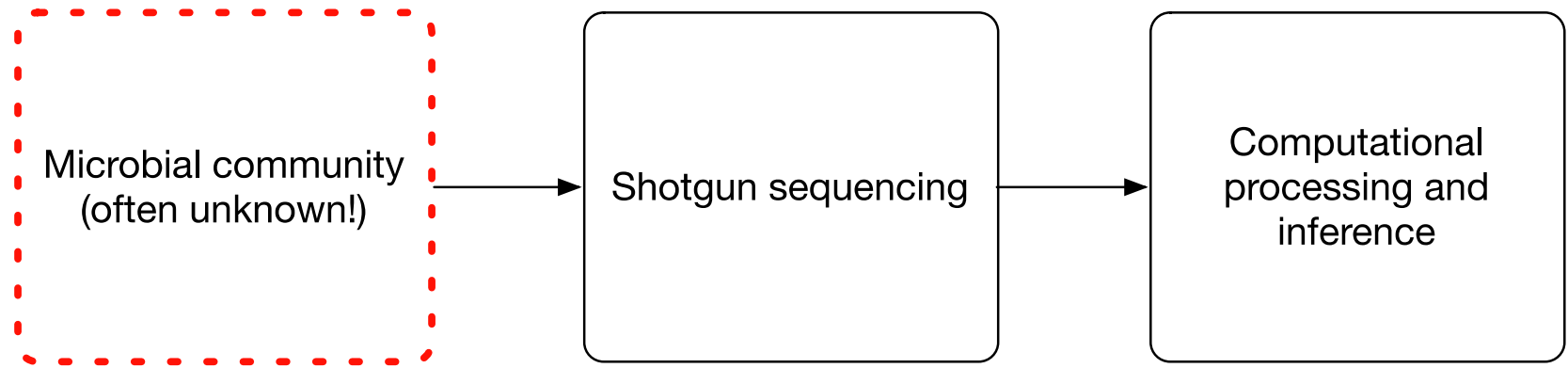
- A common approach for profiling the bacterial communities is to conduct comparative analysis of ribosomal RNA sequences (rRNA), specifically the **16S ribosomal RNA**.
- 16S rRNA gene is considered mostly highly conserved between different species of bacteria, and it's a standard marker gene for bacterial classification.
- 16S rRNA gene does also contain several hypervariable regions (V1-V9), which can provide maximum discriminating power to identify different bacterial groups.



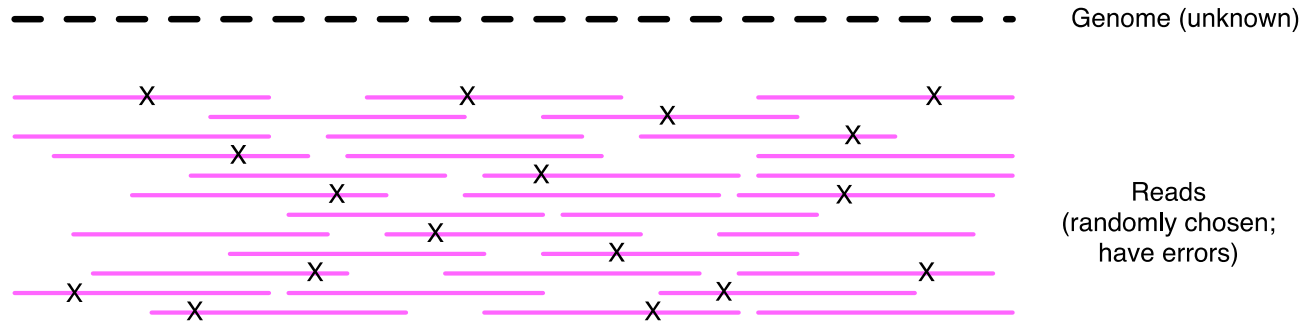
Classification of bacteria genomes

- Next Generation Sequencing (NGS) technologies are used for 16S rRNA sequencing.
- (1) Whole Genome Shotgun (or **16S shotgun (SG)**) is used to obtain the full-length 16S rRNA genes.
- (2) **Amplicon (AMP)** is used to obtain only the specific hypervariable region (V3-V4) of the 16S rRNA genes.

As usual...



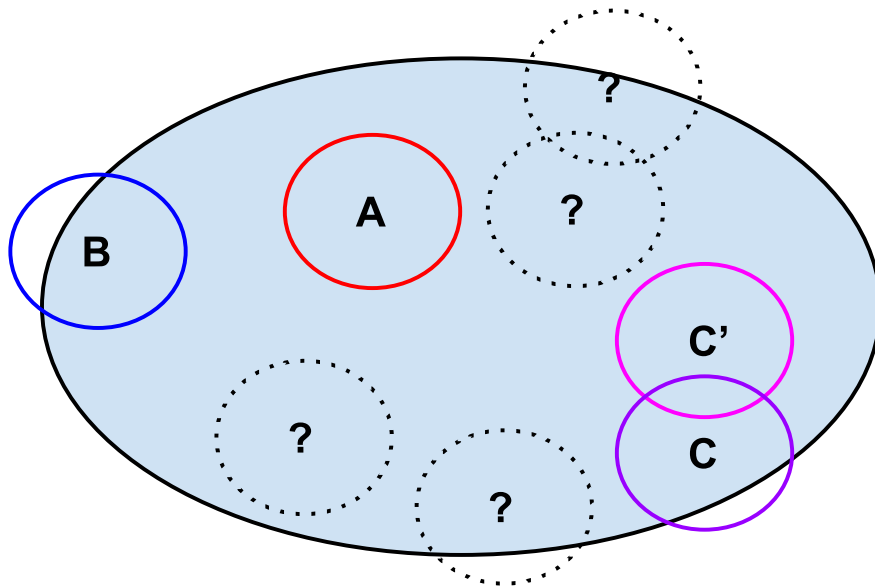
Shotgun sequencing



“Coverage” is the average number of reads that overlap each true base in (meta)genome.

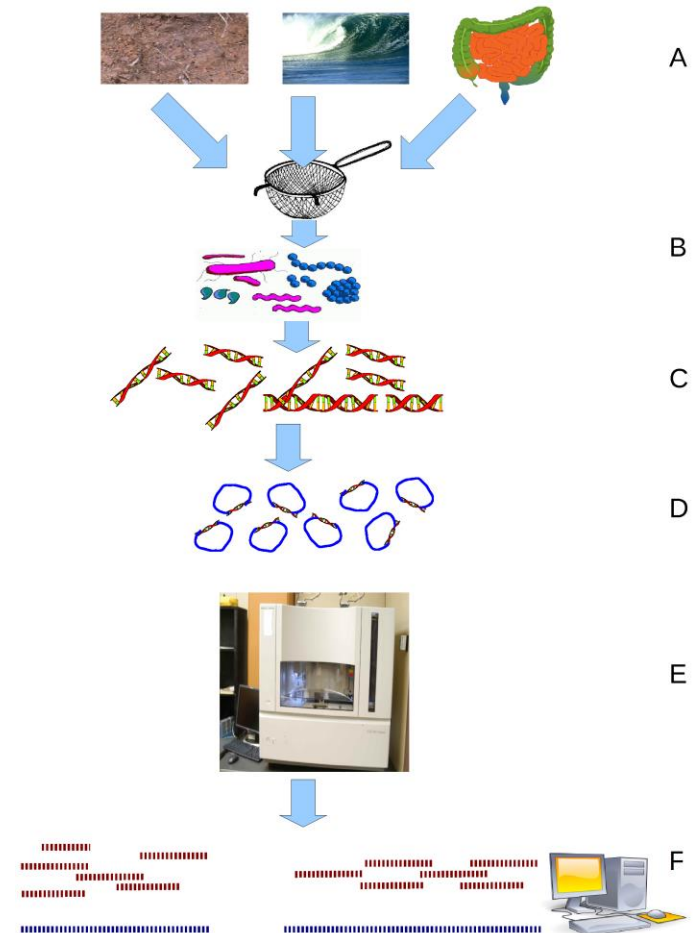
Here, the coverage is ~10 – just draw a line straight down from the top through all of the reads.

Shotgun *metagenomics*: sequencing *communities*.



Shotgun metagenomics

- Collect samples;
- Extract DNA;
- Feed into sequencer;
- Computationally analyze.



Wikipedia: Environmental shotgun
sequencing.png

Goals of shotgun metagenomics

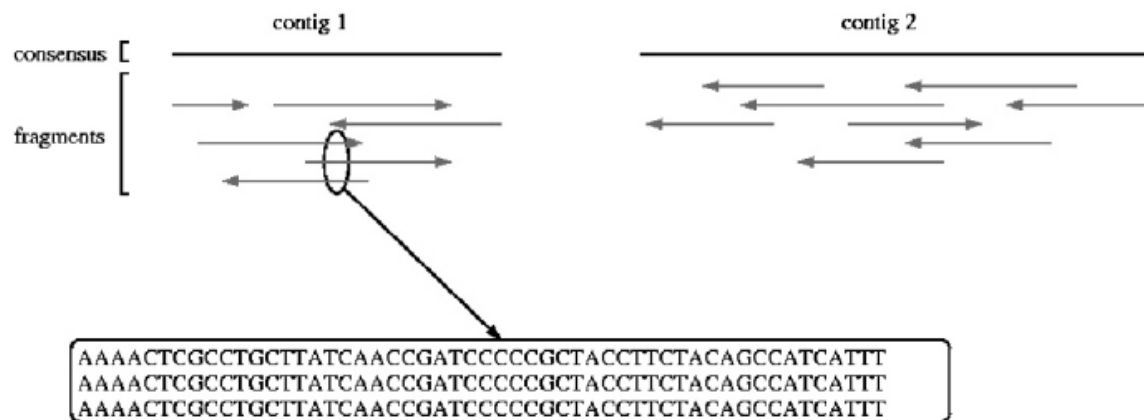
- Expand beyond taxonomic/community structure characterization possible with 16s;
- Analyze virus, plasmid, strain-level content;
- Evaluate metabolic capacity (e.g. “is nirK present?”)
- Reconstruct **genomes** from metagenomes, if possible.

16s? Or Shotgun metagenomics?

- Cons vs amplicon:
 - lower coverage / more expensive (good? bad? what are the tradeoffs?)
 - much more computationally challenging to analyze
- Pros vs amplicon:
 - different bias (no primers)
 - virus/phage can be detected
 - function can be more directly detected
 - recover (putative) genomes

Shotgun metagenome assembly: reconstruct original genome by finding overlaps in data

Randomly sequencing DNA, then finding overlaps and inferring true sequence:



UMD assembly primer (cbcb.umd.edu)

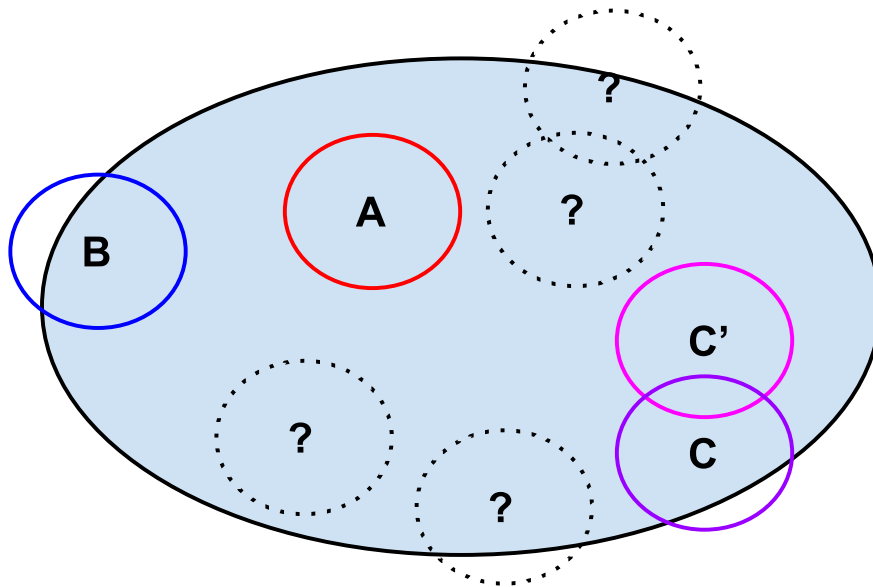
Shotgun sequencing & *de novo* assembly:

It was the Gest of times, it was the wor
, it was the worst of timZs, it was the
isdom, it was the age of foolisXness
, it was the worVt of times, it was the
mes, it was Ahe age of wisdom, it was th
It was the best of times, it Gas the wor
mes, it was the age of witdom, it was th
isdom, it was tle of foolishness



It was the best of times, it was the worst of times, it was the age of wisdom, it was
the age of foolishness

Note: Shotgun metagenome data is *always* incomplete.



Smaller circles represent (some of) your actual community.

Blue circle represents what's in your sequencing data set.

Shotgun metagenome data may not contain everything in your community; may contain strain variants; may contain “unknown” microbes.

What can we do with shotgun metagenomics?

- do taxonomic analysis directly on the reads - *(Tuesday!)*
- search the reads for genes of interest (function, taxonomy)
- assemble the reads into **contigs**, longer stretches of DNA - *(next few days)*

Important notes on assembly:

- assembly squashes abundance 8:
- assembly ignores complicated regions :(
- assembly is **surprisingly accurate** and (when using megahit, at least) **computationally tractable** :)

Some open *computational* research questions:

- right now assembly based approaches simply ...*discard* some proportion of the data. we should figure out a better way.
- what is the value of long reads in shotgun metagenomics?

Assembly results

```
% megahit -1 R1.fq.gz -2 R2.fq.gz
```

```
...
```

```
--- [STAT] 7774 contigs, total 4987609 bp, min 200 bp, max 8658 bp,  
avg 642 bp, N50 1049 bp
```

```
% head final.contigs.fa
```

```
>k141_3 flag=0 multi=20.8090 len=230
```

```
TGATCCTGTAGTGACTAACGGACGGATGTAAGCATCTTTTAGACCATTACGTT  
CTAACAATTCGTAAGTAATTTGAGTTAGTTGTTCTTCAGAATAATTCAATTTAAT  
ATTCATGACTTCTGCTCCATACTTAAGCCTTTTGTAATGCTCATATGACTTAAAA  
ATTTTAGTTTCTTCGTTGGTATTGTATGCTCTTATACCTTCGAAAACCCCATTTTC  
CATAGTGTAATA
```


Quast results

# contigs (≥ 0 bp)	7774
# contigs (≥ 1000 bp)	1308
# contigs (≥ 5000 bp)	23
# contigs (≥ 10000 bp)	0

Total length (≥ 0 bp)	4987609
Total length (≥ 1000 bp)	2560563
Total length (≥ 5000 bp)	133727
Total length (≥ 10000 bp)	0

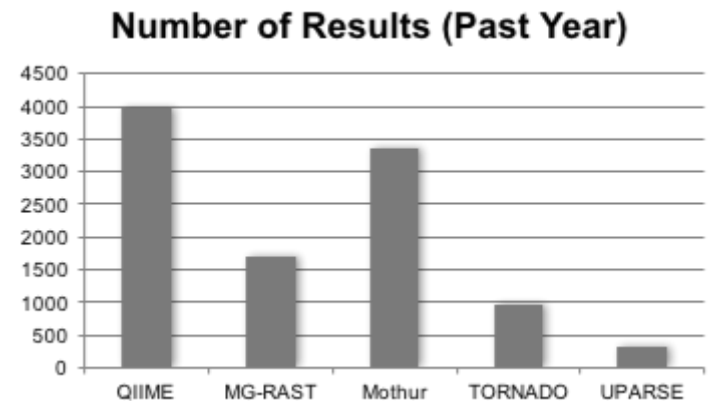
# contigs	2398
Largest contig	8658
Total length	3331977
GC (%)	31.60
N50	1684

Open question: How much should I
sequence?

Open question: How accurate/effective is functional classification on shotgun metagenome data?

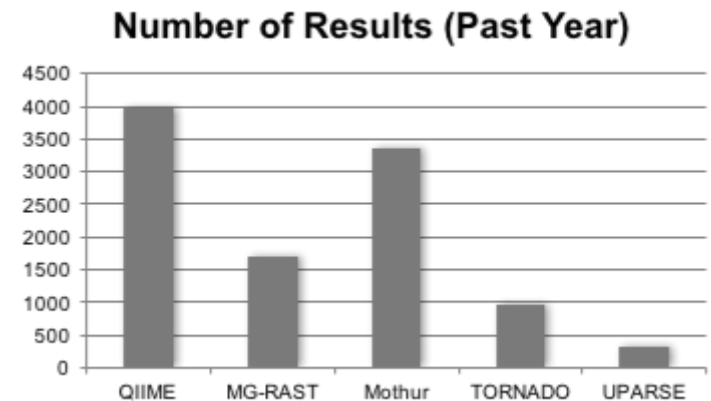
Software Selection

- Google “16S analysis <program name>”; main contenders are
- Mothur
 - Name: not an acronym (play on DOTUR, SONS)
 - Philosophy: single piece of re-implemented software
 - Top pro: easy to install
 - Top con: re-implementations could be buggy
 - Language: C++
 - Model: open-source
 - License: GPL
 - Published: 2009
 - Developed: at Umichigan



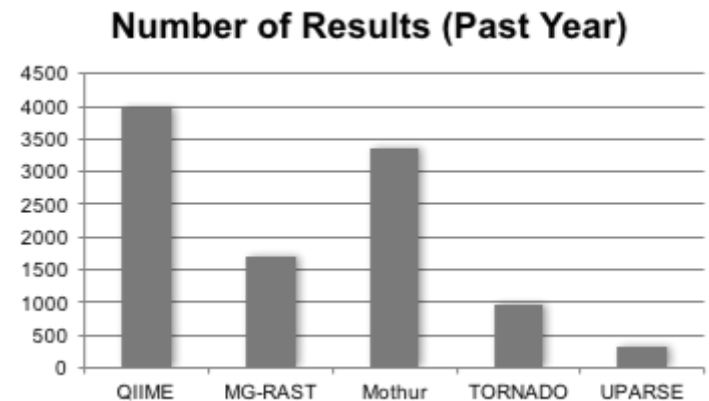
Software Selection

- Google “16S analysis <program name>”; main contenders are
- QIIME
 - Name: Quantitative Insights Into Microbial Ecology
 - Philosophy: wrapper of best-in-class software
 - Top pro: extremely flexible
 - Top con: QIIME 2 not yet feature-complete
 - Language: python (wrapper)
 - Model: open-source
 - License: mixed
 - Published: 2010
 - Developed: At UCSD, NAU



Software Selection

- Google “16S analysis <program name>”
 - Main contenders are Mothur and QIIME
 - Both widely used
 - Both pride themselves on quality of support
- Will discuss only QIIME in this tutorial
- QIIME 1 vs QIIME 2
 - QIIME 1 is no longer supported (since end of 2017)
 - This tutorial uses QIIME 2 **only**
- **I’m not a QIIME 2 developer**
 - I’m not taking credit for this tool, just demonstrating it!



Common Issues in Marker Gene Studies

- Neglecting metadata
 - Analysis can not test for effects of, or discard bias from, categories you didn't record!
- Picking novel 16S primers—not all created equal
 - Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes
- Not taking precautions to support amplicon sequencing
 - Some Illumina machines require high PhiX, low cluster density
- Selecting an inappropriate reference database
 - E.g., Greengenes (16S) reference database when sequencing ITS
- Expecting species-level taxonomy calls
 - Most sequence variants only specify to family or genus level
- Using inappropriate statistical tests
 - Taxa abundance requires a compositionality-aware test like ANCOM
 - Differences in β diversity distances across groups requires test like PERMANOVA, not ANOVA



Importing Data

- After sequence data is on your machine, must be imported to a QIIME 2 “artifact”
 - Artifact = data + metadata
 - QIIME 2 artifacts have extension `.qza`
 - Different kinds of input data (e.g., single-end vs paired-end) and different formats of input data (e.g., sequences & barcodes in same or different file) need different imports
 - See “Importing data” tutorial at <https://docs.qiime2.org/>



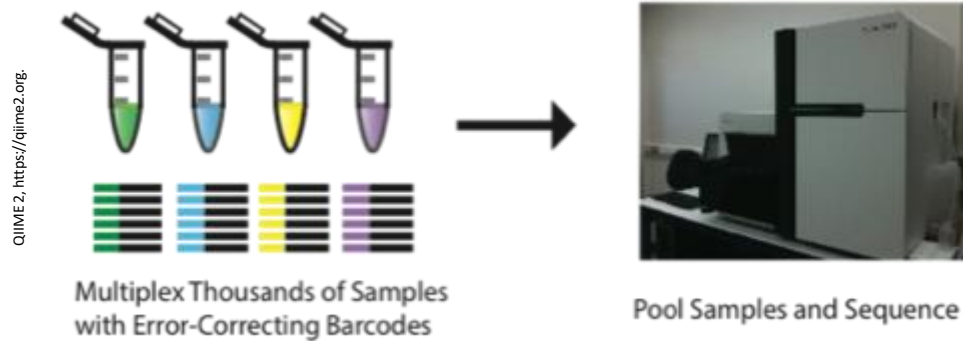
Practicum: Importing Data

```
qiime tools import \  
  --type EMPSingleEndSequences \  
  --input-path emp-single-end-sequences \  
  --output-path emp-single-end-sequences.qza
```

- Backslash is line continuation
 - Could leave out and just type whole command as one run-on line 😊
- Note structure of arguments to `qiime` command
 - Plugin name then method name then arguments
 - Order matters



Demultiplexing



- Must assign resulting sequences to samples to analyze
- **You may not need to do this!**
 - If sequencing done by a core, results may be demultiplexed before returned to you



Practicum: Demultiplexing

```
qiime demux emp-single \  
  --i-seqs emp-single-end-sequences.qza \  
  --m-barcodes-file sample-metadata.tsv \  
  --m-barcodes-column BarcodeSequence \  
  --o-per-sample-sequences demux.qza
```

- Arguments have a naming convention
 - Inputs (--i-<whatever>), metadata (--m-<whatever>), parameter (--p-<whatever>), output (--o-<whatever>)
 - Order *doesn't* matter



Practicum: Demultiplexing (cont.)

- Presumably you'd like to know how your demultiplexing worked
- But the artifact doesn't show you that info, so create a *visualization*

```
qiime demux summarize \  
  --i-data demux.qza \  
  --o-visualization demux.qzv
```

- Note that visualizations have the extension `.qzv` instead of `.qza`

- Now view the visualization, locally

```
qiime tools view demux.qzv
```

- When done examining, in Terminal, type **JUST** `q`
 - Don't need to hit Enter afterwards
 - Beware: quitting visualization doesn't close web page (but page becomes unreliable)



Demultiplexing Summary View

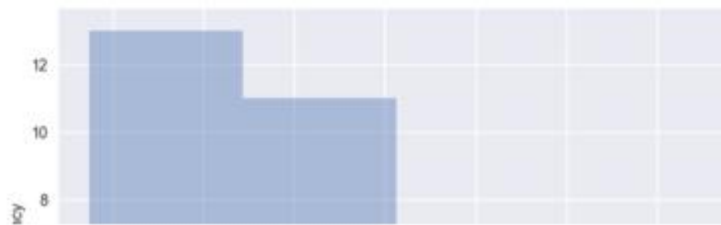


Overview

Interactive Quality Plot

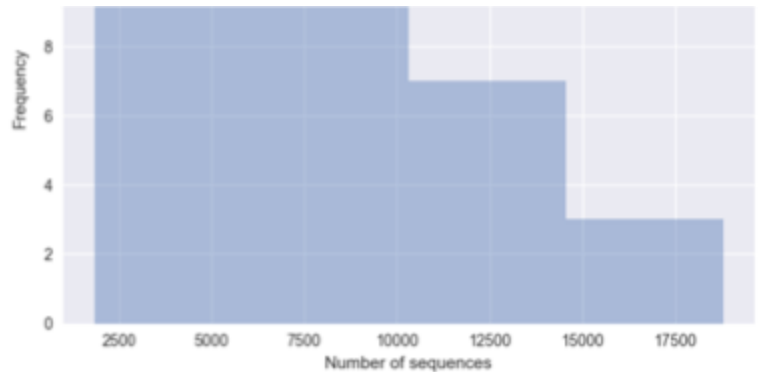
Demultiplexed sequence counts summary

Minimum:	1853
Median:	8645.0
Mean:	7761.11764706
Maximum:	18787
Total:	263878



Demultiplexing Summary View

(cont.)

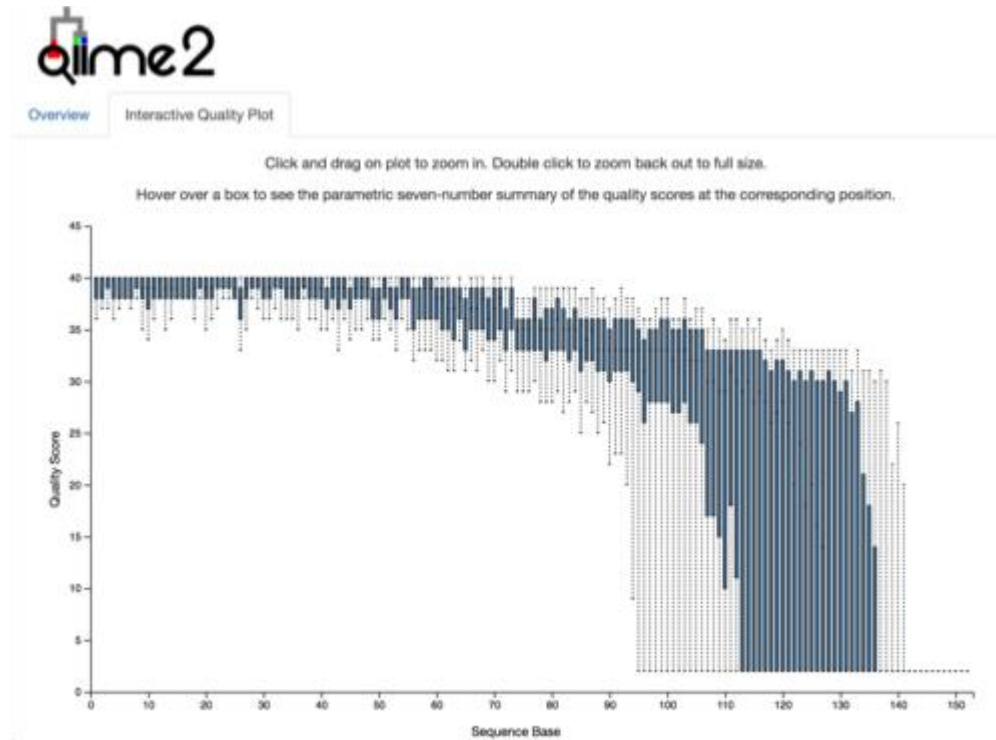


[Download as PDF](#)

Per-sample sequence counts

Sequence count	
Sample name	
L4S137	18787
L4S63	17167
L4S112	16265
L1S8	12386

Demultiplexing Summary View (cont.)



Practicum: Peeking At An Artifact

- What happens if you type

```
qiime tools view  
demux.qza
```



Practicum: Peeking At An Artifact (cont.)

- What happens if you type

```
qiime tools view demux.qza
```

- You get

```
Usage: qiime tools view [OPTIONS] VISUALIZATION_PATH
```

```
Error: Invalid value: demux-filtered.qza is not a QIIME  
2 Visualization. Only QIIME 2 Visualizations can be  
viewed
```

- Instead, run

```
qiime tools view demux.qza
```



Practicum: Peeking At An Artifact (cont.)

- What happens if you type

```
qiime tools view demux.qza
```

- You get

```
Usage: qiime tools view [OPTIONS] VISUALIZATION_PATH
```

```
Error: Invalid value: demux-filtered.qza is not a QIIME  
2 Visualization. Only QIIME 2 Visualizations can be  
viewed
```

- Instead, run

```
qiime tools view demux.qza
```

- See something like

```
UUID:          cce55836-0f04-42de-8476-83224254b419
```

```
Type:          SampleData[SequencesWithQuality]
```

```
Data format: SingleLanePerSampleSingleEndFastqDirFmt
```

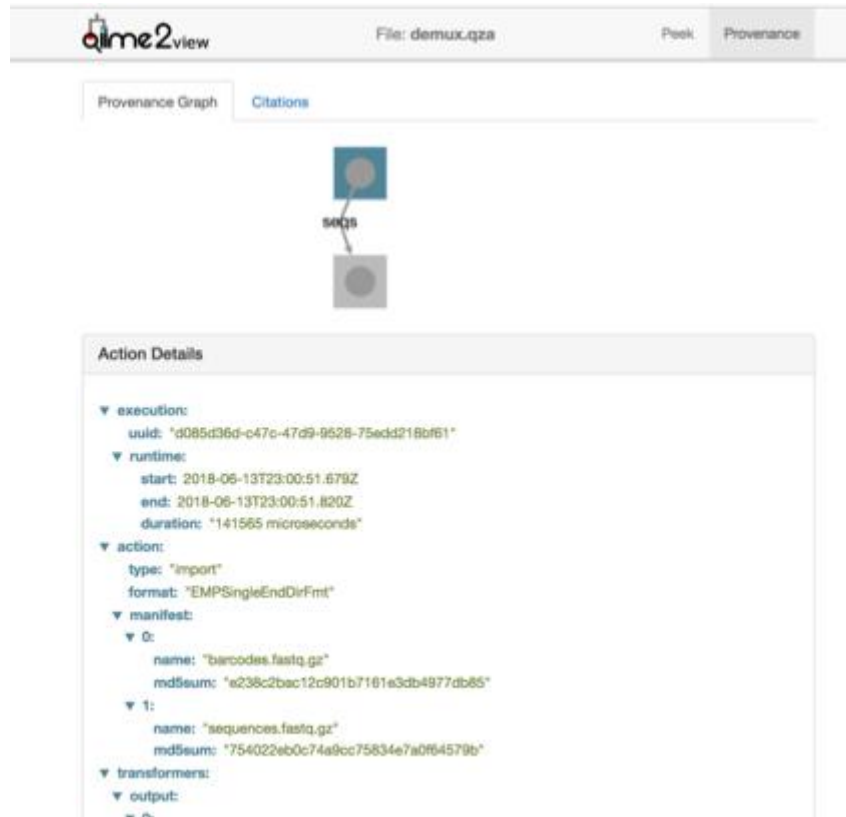


Aside: Viewing Artifact Provenance

- Provenance tracking is **absolutely critical** to reproducible analyses
 - Almost no tool actually tracks it **for** you—really a fantastic new QIIME 2 feature
- Provenance can be viewed through the QIIME2 View website
 - Open Chrome and go to <https://view.qiime2.org>
 - Drag and drop file `demux.qza`
 - Click on “Provenance” tab



Aside: Viewing Artifact Provenance (cont.)



The screenshot shows the QIME2 web interface. At the top, there's a header with the QIME2 logo, the file name 'demux.qza', and tabs for 'Peek' and 'Provenance'. Below the header, there are two tabs: 'Provenance Graph' and 'Citations'. The 'Provenance Graph' tab is active, showing a graph with two nodes: a blue square at the top and a grey square at the bottom, connected by a downward arrow labeled 'seqs'. Below the graph, there's an 'Action Details' panel. This panel contains a tree view of the action's metadata:

- ▼ execution:
 - uuid: "d085d36d-c47c-47d9-9528-75ecd218bf61"
- ▼ runtime:
 - start: 2018-06-13T23:00:51.679Z
 - end: 2018-06-13T23:00:51.820Z
 - duration: "141565 microseconds"
- ▼ action:
 - type: "import"
 - format: "EMPSingleEndDirFmt"
- ▼ manifest:
 - ▼ 0:
 - name: "barcodes.fastq.gz"
 - md5sum: "a238c2bac12c901b7161a3db4977db85"
 - ▼ 1:
 - name: "sequences.fastq.gz"
 - md5sum: "754022eb0c74a9cc75634e7a0f64579b"
- ▼ transformers:
 - ▼ output:
 - ▼ 0:

- Click on square to see action details
- Click on circle+arrow to see file passed between actions
- Note that citations are also provided!



Quality Control

Bokulich, N. et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*, 10(1), 57–59.

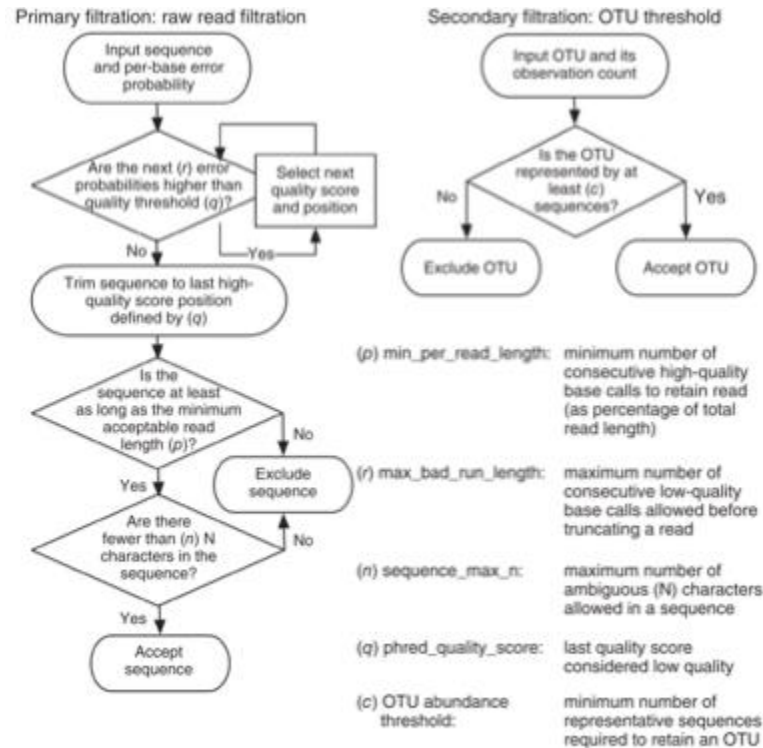


Figure 1 | Quality-filtration process flow in QIIME v1.5.0.

QIIME defaults:

- $r = 3$
- $q = 3$
- $p = 0.75$
- $n = 0$
- $c = 0.005\%$ or 2



Practicum: Quality Control

```
qiime quality-filter q-score \  
  --i-demux demux.qza \  
  --o-filtered-sequences demux-filtered.qza \  
  --o-filter-stats demux-filter-stats.qza
```



Practicum: Quality Control (cont.)

- qiime quality-filter q-score \
 - --i-demux demux.qza \
 - --o-filtered-sequences demux-filtered.qza \
 - --o-filter-stats demux-filter-stats.qza
- qiime metadata tabulate \
 - --m-input-file demux-filter-stats.qza \
 - --o-visualization demux-filter-stats.qzv



Quality Control Summary View



Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

sample-id	total-input-reads	total-retained-reads	reads-truncated	reads-too-short-after-truncation	reads-exceeding-maximum-ambiguous-bases
#q2types	numeric	numeric	numeric	numeric	numeric
L1S105	11340	9232	10762	2066	42
L1S140	9736	8584	9457	1113	39
L1S208	11335	10148	10667	1161	26
L1S257	8216	7302	7672	876	38
L1S281	8904	7763	8343	1117	24
L1S57	11750	10000	11000	1716	34
L1S76	10100	8984	9678	1092	24
L1S8	12386	8433	12035	3916	37
L2S155	9261	5066	8932	4167	28
L2S175	10691	5574	10216	5092	25

Practicum: Feature Table Creation

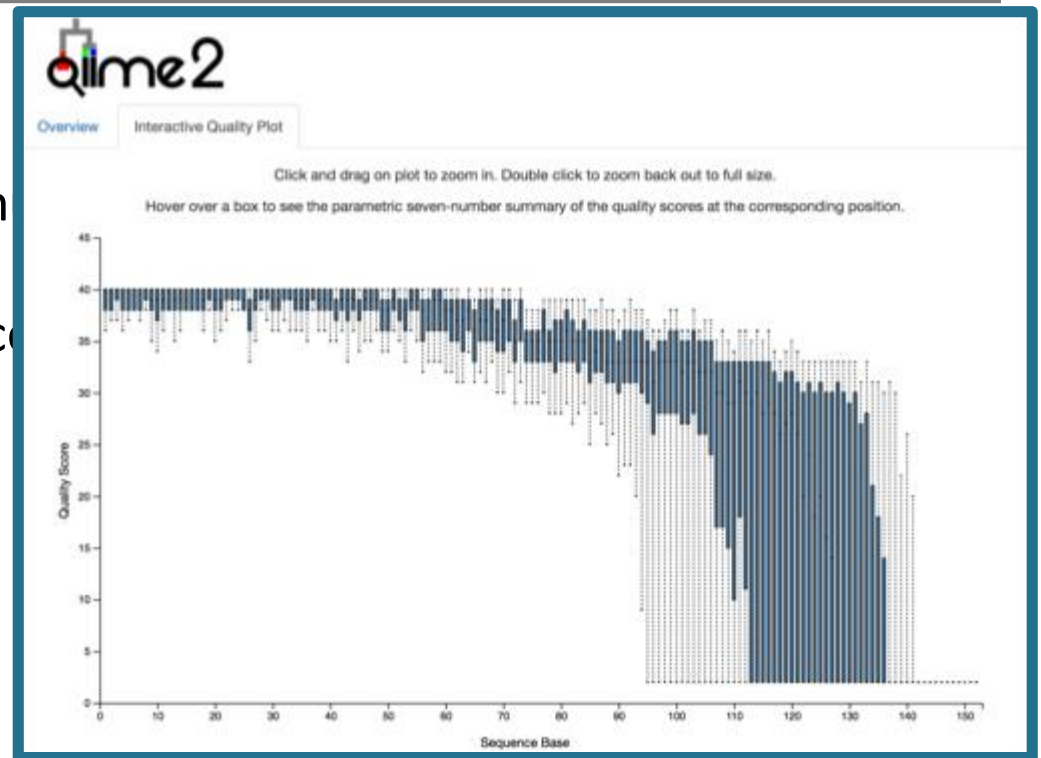
```
qiime deblur denoise-16S \  
  --i-demultiplexed-seqs demux-filtered.qza \  
  --p-trim-length 120 \  
  --o-representative-sequences rep-seqs.qza \  
  --o-table table.qza \  
  --p-sample-stats \  
  --o-stats deblur-stats.qza
```

- This can take up to 10 minutes to run, so while we wait ...
 - Where do you guess the number 120 came from?



Practicum: Feature Table Creation

- qiime deblur denoise-16S \
 - --i-demultiplexed-seqs dem
 - --p-trim-length 120 \
 - --o-representative-sequences
 - --o-table table.qza \
 - --p-sample-stats \
 - --o-stats deblur-stats.qza



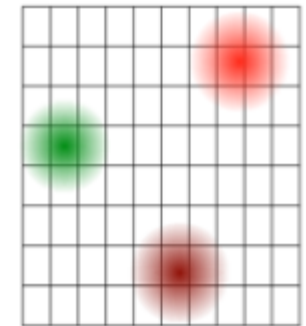
Feature Table Creation—The Past

- Last year: OTU (Operational Taxonomic Unit)
 - “an operational definition of a species used when only DNA sequence data is available”
 - Sequences at/above a given similarity threshold considered part of the same OTU
 - 97% is the usual “species-level” threshold
 - Similarity determined using alignment (time-consuming)
 - Purpose is to minimize impact of sequencing errors
 - But also masks fine (sub-OTU) variation in real biological sequences
 - Results very difficult to compare across studies if done *de novo*
 - “Closed reference”, “open reference” methods increase comparability require reference database
- Output is a “feature table”:
 - Rows are samples
 - Columns are OTUs (arbitrary identifiers if **de novo**, from reference database if closed reference)
 - Values are frequency of reads from that OTU in that sample

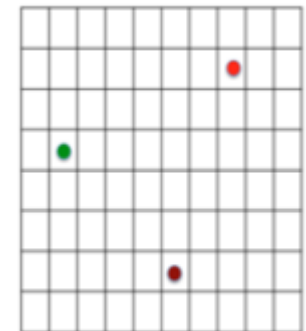


Feature Table Creation—The Present

- This year: sOTU (sub-OTU) methods
 - Use error modeling to *in silico* correct sequencing mistakes
 - Sounds impossible but is actually quite accurate, with right error model
 - Error model is specific to the sequencing type (e.g., 454, Illumina Hi/MiSeq)
 - Result: only sequences likely to have been input to the sequencer
 - Options include (NOT a complete list):
 - DADA2 (2016)
 - Deblur (2017)
- Output is STILL a feature table:
 - Rows are samples
 - Columns are SEQUENCES
 - Values are frequency of reads from that SEQUENCE in that sample



After Sequencing



True sequences

QIIME 2, <https://qiime2.org>.



Practicum: Feature Table Creation (cont.)

```
qiime deblur visualize-stats \  
  --i-deblur-stats deblur-stats.qza \  
  --o-visualization deblur-stats.qzv
```



Deblur Statistics

View



Per-sample Deblur stats

Click on a Column header to sort the table.

Mouse over a Column header to get a description.

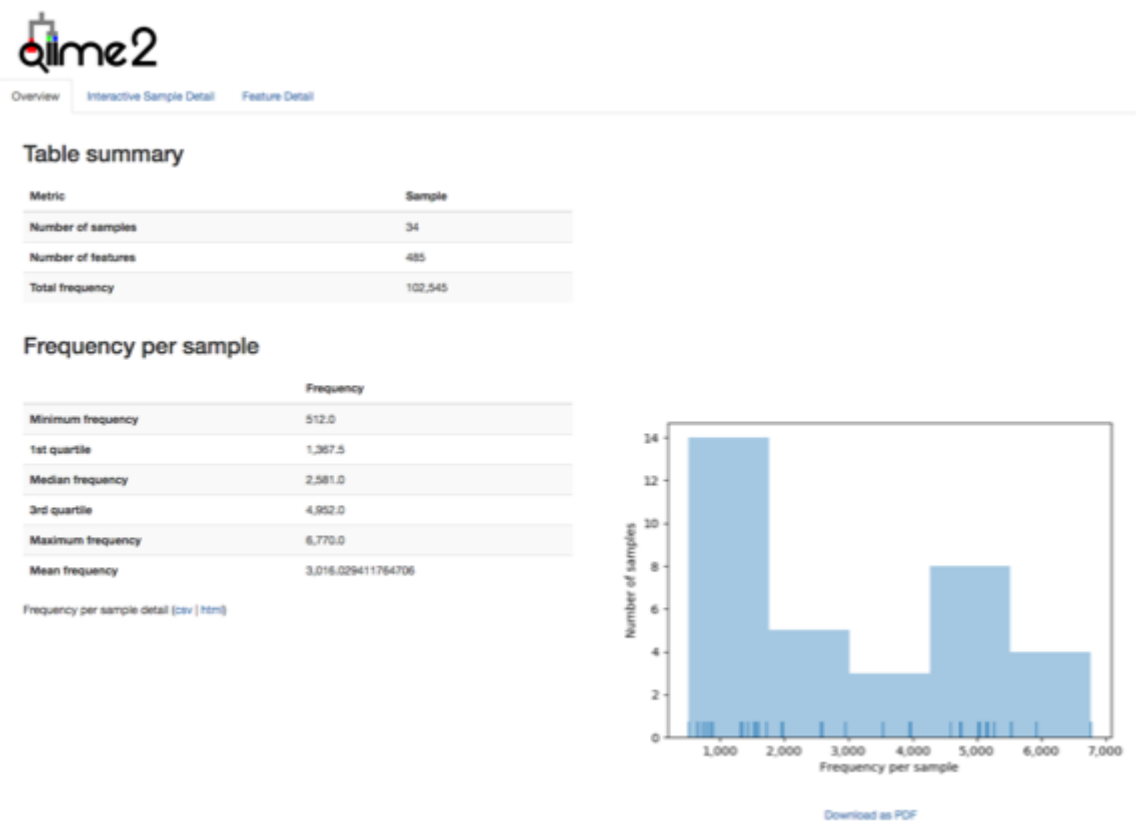
	sample-id	reads-raw	fraction-artifact-with-minsize	fraction-artifact	fraction-missed-reference	unique-reads-derep	reads-derep	unique-reads-deblur	reads-deblur	unique-reads-hit-artifact	reads-hit-artifact	unique-reads-chimeric	reads-chimeric	unique-reads-hit-reference	reads-hit-reference	unique-reads-missed-reference	reads-missed-reference
0	L3S360	1341	0.518270	0.0	0.011785	118	646	111	600	0	0	1	6	73	512	3	7
1	L2S222	4459	0.498094	0.0	0.019086	327	2238	287	1999	0	0	4	8	147	1603	3	38
2	L2S309	1904	0.457458	0.0	0.003185	124	1033	99	942	0	0	0	0	76	895	1	3
3	L3S341	1293	0.438515	0.0	0.000000	95	726	86	675	0	0	0	0	78	653	0	0
4	L2S204	4349	0.429064	0.0	0.012851	236	2483	152	2102	0	0	1	1	106	1968	2	27
5	L2S357	3100	0.419032	0.0	0.000000	149	1801	87	1554	0	0	0	0	75	1533	0	0
6	L3S294	1523	0.401182	0.0	0.002398	82	912	65	836	0	0	1	2	52	800	1	2
7	L3S313	1340	0.391045	0.0	0.000000	85	816	69	747	0	0	1	1	66	741	0	0
8	L2S240	7110	0.390717	0.0	0.000000	253	4332	78	3578	0	0	9	17	59	3535	0	0
9	L2S155	5066	0.388077	0.0	0.006227	260	3100	178	2730	0	0	0	0	119	2579	3	17
10	L2S175	5574	0.371726	0.0	0.001308	281	3502	177	3059	0	0	1	1	132	2954	1	4

Practicum: Feature Table Creation (cont.)

```
qiime feature-table summarize \  
  --i-table table.qza \  
  --o-visualization table.qzv \  
  --m-sample-metadata-file sample-  
  metadata.tsv
```



Feature Table Summary View

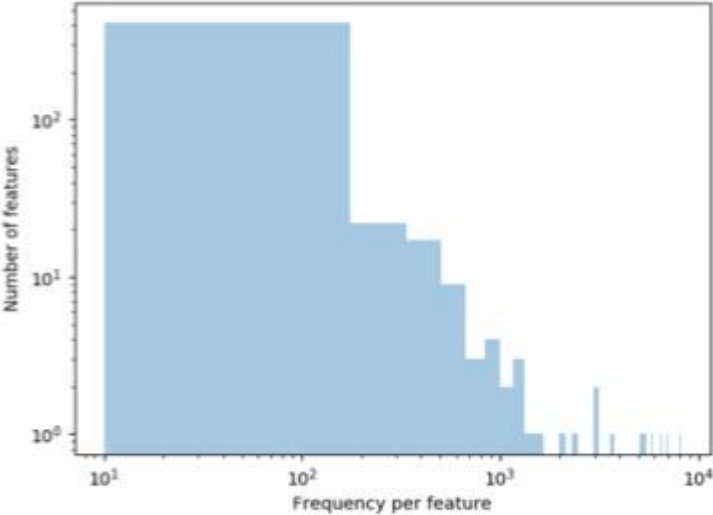


Feature Table Summary View (cont.)

Frequency per feature

	Frequency
Minimum frequency	10.0
1st quartile	16.0
Median frequency	33.0
3rd quartile	88.0
Maximum frequency	8,223.0
Mean frequency	211.43298969072166

[Frequency per feature detail \(csv | html\)](#)



[Download as PDF](#)

Feature Table Summary View (cont.)

[Overview](#)[Interactive Sample Detail](#)[Feature Detail](#)

	Frequency	# of Samples Observed in
4b5eeb300368260019c1fbc7a3c718fc	8,223	16
fe30ff0f71a38a39cf1717ec2be3a2fc	6,935	19
d29fe3c70564fc0f69f2c03e0d1e5561	6,428	27
1d2e5f3444ca750c85302ceee2473331	5,809	27
868528ca947bc57b69ffdf83e6b73bae	5,347	12
154709e160e8cada6bfb21115acc80f5	5,117	14
0305a4993ecf2d8ef4149fdcf7582603	3,671	13
997056ba80681bbbdd5d09aa591eadc0	3,051	18
cb2fe0146e2fbc101050edb996a0ee2	3,021	17
3c9c437f27aca05f8db167cd080ff1ec	2,358	18
9079bfbcce01d4b5c758067b1208c31	2,093	15
bfbcd36e63b69fec4627424163d20118	1,622	17
d86ef5d6394f5dbeb945f39aa25e7426	1,405	12
a048763053c2777h1fc2a31f841ah23hd	1,318	15

Feature Table Summary View (cont.)



Overview

Interactive Sample Detail

Feature Detail

Sampling depth:

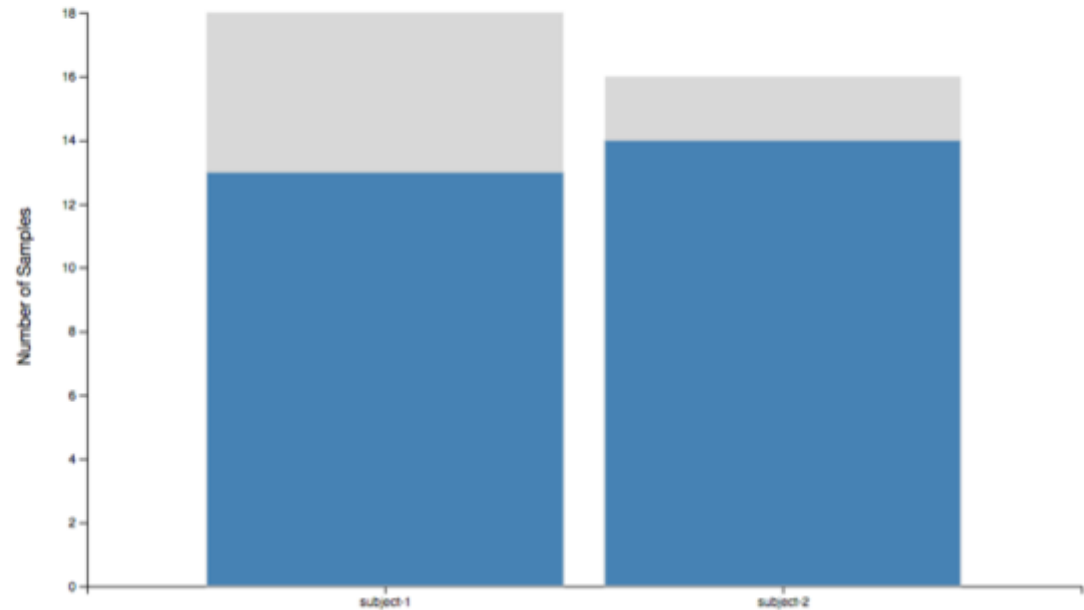
1000

(Zero implies no even sampling.)



Subject

Retained 27,000 (26.33%)
sequences in 27 (79.41%)
samples at the specified
sampling depth.



Feature Table Summary View (cont.)

Sample ID	Sequence Count
L4S137	6,770
L4S63	5,912
L1S57	5,525
L4S112	5,523
L6S93	5,261
L6S20	5,170
L1S208	5,136
L1S76	5,037
L4S102	5,020
...	...
L5S155	1,347
L5S240	1,329
L2S309	895
L3S378	849
L3S294	800
L3S313	741
L3S242	660
L3S341	653
L3S360	512




Practicum: Feature Table Creation (cont.)

- qiime feature-table summarize \
 - --i-table table.qza \
 - --o-visualization table.qzv \
 - --m-sample-metadata-file sample-metadata.tsv
- qiime feature-table tabulate-seqs \
 - --i-data rep-seqs.qza \
 - --o-visualization rep-seqs.qzv



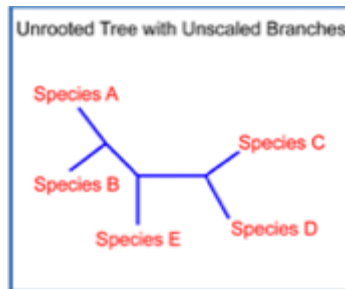
Feature Table Tabulation View

 <p>To BLAST a sequence against the NCBI nt database, click the sequence and then click the View report button on the resulting page.</p> <p>To download a raw FASTA file of your sequences, click here.</p> <p>Click on a Column header to sort the table.</p>	
Feature ID	Sequence
19bc6716f02b0a22f03b1a686895aae1	GACGGGGGGGCAAGTGTCTTDCGAA7GACTGGGCGTAAAGGSCACGTAGGCGGTGAATCGGGTTGAAAGTGAAGTCGCCAAAACTGGCGGAATGCTCTCGAAACCAAT
b1e666a5c5f127a8b92de2863503a9e5	TACGTAGGTCCCGAGCGTTGTCCGATTATTGGGCGTAAAGCGAGCGCAGGCGGTTTGATAGTCTGAAGTAAAGGCTGTGGCTTAACCATAGTATGCTTTGGAAACTGTTAA
5a165119cbd604e6c33d3ee6c62a45e2	TACAGAGGGTGCAGCGTTAATCGGAATTAAGTGGGCGTAAAGCGAGCGGTAGGTGGCTTGATAGTCAAGTCAGATGTGAAAGCCCCGGGCTTAACCTGGGAACGGCATCTGATACTGTT
101968ec709b68fcd964a68f226dcd1	TACGTAGGGGGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGGCGCAGCGGCTGTGATCAAGTCAGCTGTAAAGGATGCGGCTTAACCGTGTTTAGCAGTTGAAACTGGAT
921966b15e9f1508e6650cd053443c385	TACGGAGGGTGCAGCGTTATCCGGAATTATTGGGTTTAAAGGGTCCGTAGGTGGGTAGTAAAGTCAGTGGTGAATCCTGCAGCTTAAGTGTAGAACTGCCATGTATACTGCTAG
79e9e337b10e2d298bb1b3bde946782d	TACGGAGGGTGCAGCGTTAATCGGAATTACTGGGCGTAAAGGCGCAGCGCGGTTTGTAAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACGTGCATCTGATACTGGK
0be23adca941c1335eb04db17dc66e98	TACGGAGGATACGAGCGTTATCCGGAATTATTGGGTTTAAAGGGTGGTGGGTAGTGGCGTATTAAGTCAGTGTGAAAGCTGCAGCTCAACTGTAGTCTTGCCGTTGAAACTGATAT
9da7a39de7e62006cc9533681ad7765a	TACGTATGTCACAGCGTTATCCGGAATTATTGGGCGTAAAGGCGGCTCAGGTGGTTAGTAAAGTCTGATGTGAAATGCGAGGCTCAACTCTGATTGCGTTGGAAACTGTGTAA
71a8f515ee1ac5cb8b529b13e5a89790	TACGGAGGGTGCAGCGTTAATCGGAATTAAGTGGGCGTAAAGGCGCAGCGCGGACTTTTAAAGTGAATGTGAAATCCCCGAGCTTAACCTGGGAACGTGCATTTGAGACTGGK
f023384b8989d014cd3ead77f10db307	TACGTAGGGAGCGAGCGTTGTCCGGAATTACTGGGTTTAAAGGGAGCGTAGGCGGGAAGCAAGTCAGATGTGAAAACTATGGGCTCAACCTGTAGATTGCATTTGAAACTGTTT
c5dfb6a2b481cb89e2602fc20941587	TACGGAGGATACGAGCGTTATCCGGAATTATTGGGTTTAAAGGGTGGTGGGTTGCTTTTAAAGTCAGTGGTGAAGGCTGTGGCTCAACCATAGTCTTGCCGTTGAAACTGAAGG
3b77e15d86603bf0a6be5098b010afe7	TACGTAGGGGGCGAGCGTTATCCGGAATTACTGGGTTTAAAGGGAGCGTAGAGCGTTAAGCAAGTCTGAAGTGAAGGCCCGGGGCTCAACCCGGTACTGCTTTGGAAACTGT
685ea779ee012329ec2e171f1823fa8	TACGTAGGGTGCAGCGTTAATCGGAATTAAGTGGGCGTAAAGCGAGCGCAGCGGTTTATTAAGCAAGTGTGAAATCCCCGAGCTTAACCTGGGAACGTGCGTTTGAAGTGGTAT
8992381f7d3d5a45d162e2f4b38d01b	TACAGAGGGTGCAGCGTTAATCGGAATTAAGTGGGCGTAAAGCGCGGTAGGTGGTTGCTTAAAGTGGATGTGAAATCCCCGGGCTCAACCTGGGAACGTGCATTCAGAACTGGK
4086f6a89c2ead7d91003a0362a00228	TACGTATGTCACAGCGTTATCCGGAATTATTGGGCGTAAAGCGGCTCAGGTGGTTAGTAAAGTCTGAATGTGAAATGCGAGGCTCAACTCTGTATGCGTTGGAAACTGTGTAA
fd4f95c05b686060121f7090850f21	TACGTATGGAGCGAGCGTTGTCCGGAATTATTGGGCGTAAAGGGTACGCGAGGCGGTTTAAAGTGAATGTGAAATGCGAGGCTCAACCCCGTAAAGCATGTGAAACTGATAA
edac7632doad21c328669bef9e2afe	TACGTAGGGGGCGAGCGTTATCCGGAATTACTGGGTTTAAAGGGAGCGTAGGCGGCGAGCAAGTCAGAAAGTGAAGGCCCGGGGCTCAACCCCGGAGCGGCTTTTGAAGCTT
4851a0c85e3c1b3f3cf7e88d8a38960a	TACGGAGGGTGCAGCGTTAATCGGAATTAAGTGGGCGTAAAGGCGCAGCGCGGATTTTAAAGTGAAGTGTGAAATCCTTGGGCTTAACCTGGGAATTCGATTTGAGACTGGG

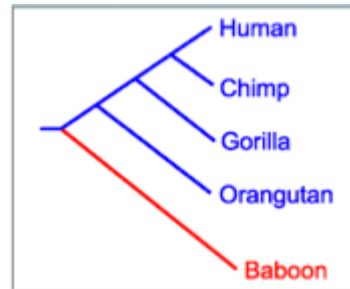
Phylogenetic Tree Creation

- Evolution is the core concept of biology
 - There's only so much you can learn from microbes while ignoring evolution!
- Evolution-aware analyses of a dataset need a phylogenetic tree of its sequences
 - *De novo*: infer tree using only sequences from dataset
 - Reference-based: insert sequences from dataset into an existing phylogenetic tree
 - Not all existing phylogenies are created equal—have strengths and weaknesses based on intended purpose when developed
- Phylogenetically based analyses in QIIME 2 need a **rooted** tree

Unrooted:



Rooted:



Geer, R.C., Messersmith, D.J., Alpi, K., Bhagwat, M., Chattopadhyay, A., Gaedeke, N., Lyon, J., Minie, M.E., Morris, R.C., Ohles, J.A., Osterbur, D.L. & Tennant, M.R. 2002. NCBI Advanced Workshop for Bioinformatics Information Specialists. [Online] <http://www.ncbi.nlm.nih.gov/Class/NAWBIS/>



Practicum: Phylogenetic Tree Creation

```
qiime alignment mafft \  
  --i-sequences rep-seqs.qza \  
  --o-alignment aligned-rep-seqs.qza
```

- Note: here we are doing *de novo* phylogenetic tree creation
 - Not necessarily the BEST approach, but an easy one to show you 😊



Practicum: Phylogenetic Tree Creation (cont.)

```
qiime alignment mafft \  
  --i-sequences rep-seqs.qza \  
  --o-alignment aligned-rep-seqs.qza
```

```
qiime alignment mask \  
  --i-alignment aligned-rep-seqs.qza \  
  --o-masked-alignment masked-aligned-rep-  
seqs.qza
```



Practicum: Phylogenetic Tree Creation (cont.)

```
qiime alignment mafft \  
  --i-sequences rep-seqs.qza \  
  --o-alignment aligned-rep-seqs.qza
```

```
qiime alignment mask \  
  --i-alignment aligned-rep-seqs.qza \  
  --o-masked-alignment masked-aligned-rep-  
seqs.qza
```

```
qiime phylogeny fasttree \  
  --i-alignment masked-aligned-rep-seqs.qza \  
  --o-tree unrooted-tree.qza
```



Practicum: Phylogenetic Tree Creation (cont.)

```
qiime alignment mafft \  
  --i-sequences rep-seqs.qza \  
  --o-alignment aligned-rep-seqs.qza
```

```
qiime alignment mask \  
  --i-alignment aligned-rep-seqs.qza \  
  --o-masked-alignment masked-aligned-rep-  
seqs.qza
```

```
qiime phylogeny fasttree \  
  --i-alignment masked-aligned-rep-seqs.qza \  
  --o-tree unrooted-tree.qza
```

```
qiime phylogeny midpoint-root \  
  --i-tree unrooted-tree.qza \  
  --o-rooted-tree rooted-tree.qza
```

- No visualizations provided for these artifacts



Core Metrics

- So how do you actually compare microbial communities?
 - Can't just eyeball the (gigantic, sparse) feature tables and look for differences
 - Instead, calculate metrics that compress a lot of info into a single number
 - Then do statistical tests on metrics to look for significant differences
 - **BE CAREFUL**—microbiome data is sparse, compositional, etc, so requires unusual tests
 - QIIME 2 uses appropriate tests; if doing your own, **MUST** check the literature first
- These metrics are lossy!
 - No metric exposes all the information in the full feature table
 - If it did, it would BE the feature table
 - Different metrics capture different aspects of the communities
- **Thus ...**
 - **Don't ask, "Which metric should I use?" UNTIL you know what you're looking for!**



Core Metrics (cont.)

- QIIME 2 calculates a smorgasbord of metrics for you with one command
- Alpha diversity
 - Shannon's diversity index (a quantitative measure of community richness)
 - Observed OTUs (a qualitative measure of community richness)
 - Faith's Phylogenetic Diversity (a qualitative measure of community richness that incorporates phylogenetic relationships between the features)
 - Evenness (or Pielou's Evenness; a measure of community evenness)
- Beta diversity
 - Jaccard distance (a qualitative measure of community dissimilarity)
 - Bray-Curtis distance (a quantitative measure of community dissimilarity)
 - unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
 - weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)



Normalization for Core Metrics

#Full OTU Counts

#OTU ID	PC.354	PC.355	PC.356	PC.481	PC.593
wf_otu_0	0	0	0	0	0
wf_otu_1	0	0	0	1	0
wf_otu_10	0	1	0	0	0
wf_otu_100	0	0	1	0	0
wf_otu_101	0	0	3	0	0
wf_otu_102	0	1	0	0	0
wf_otu_103	0	1	0	0	1
wf_otu_104	0	0	0	1	0
wf_otu_105	0	1	0	0	0
wf_otu_106	0	0	0	1	0
wf_otu_107	0	0	0	1	0
wf_otu_108	0	0	0	1	0
wf_otu_109	0	0	1	0	2
wf_otu_11	0	0	0	0	1
wf_otu_110	0	0	0	2	0
wf_otu_111	0	0	0	0	1
wf_otu_112	0	0	0	0	1
wf_otu_113	0	0	0	1	0

- Calculated metric values depend on sampling depth
- Ex: circled column has more non-zero counts than others
 - Is its community really more diverse—or do we just SEE more?
 - Samples with more sequences (greater sampling depth) show more diversity

- Normalization is necessary for valid comparisons of abundance/diversity
 - “But how?!”
 - Longstanding approach: rarefaction (reduce all samples to uniform sampling depth)
 - Recent publication caused concern
 - *Waste not, want not: why rarefying microbiome data is inadmissible*. McMurdie PJ, Holmes S. PLoS Comput Biol. 2014;10(4).
 - Further work demonstrated concern is excessive
 - *Normalization and microbial differential abundance strategies depend upon data characteristics*. Weiss S, et al. Microbiome. 2017 Mar 3;5(1):27. (Note: I’m an author, so not objective)



Rarefaction

- What is rarefaction?
 - randomly subsampling the same number of sequences from each sample
 - NB: samples without that number of sequences are discarded
- Concerns:
 - Too low: ignore a lot of samples' information
 - Too high: ignore a lot of samples
 - *Still* a good choice for normalization (Weiss S, et al. Microbiome. 2017):
 - “Rarefying more clearly clusters samples according to biological origin than other normalization techniques do for ordination metrics based on presence or absence”
 - “Alternate normalization measures are potentially vulnerable to artifacts due to library size”
- Researcher must choose sampling depth—but how?



Sampling Depth Selection

- Don't sweat it too much
 - “Low” depths (10-1000 sequences per sample) capture all but very subtle variations

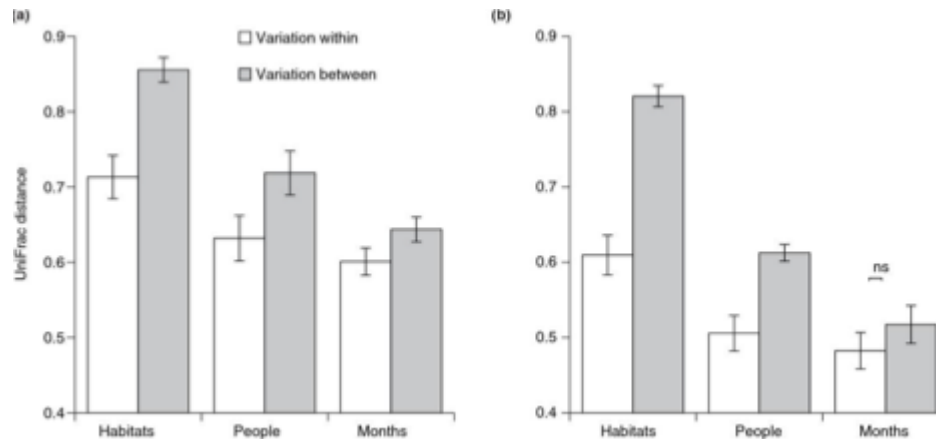


Fig. 2, Kuczynski, J. et al., "Direct sequencing of the human microbiome readily reveals community differences", Genome Biology, 2010

- Retaining samples is usually more important than retaining sequences
 - May care not just how many samples are left out but WHICH samples are left out



Practicum: Core Metrics

```
qiime diversity core-metrics-phylogenetic \  
  --i-phylogeny rooted-tree.qza \  
  --i-table table.qza \  
  --p-sampling-depth ??? \  
  --m-metadata-file sample-metadata.tsv \  
  --output-dir metrics
```

- Which sampling depth should we use?
 - How can we decide?



Exercise: Core Metrics

```
qiime diversity core-metrics-phylogeny \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth ??? \
  --m-metadata-file sample-metadata.tsv \
  --output-dir metrics
```

- Which sampling depth should we use?

- How can we decide?

```
qiime tools view table.qzv
```

- Work with your partner to choose a sampling depth, then answer:
 - Why did you choose this value?
 - How many samples will be excluded from your analysis based on this choice?
 - How many total sequences will you be analyzing in the core metrics command?



Answers: Core Metrics

```
qiime diversity core-metrics-phylogenetic \  
  --i-phylogeny rooted-tree.qza \  
  --i-table table.qza \  
  --p-sampling-depth 800 \  
  --m-metadata-file sample-metadata.tsv \  
  --output-dir metrics
```

- My answers:
 - Why did you choose this value?
 - Anything higher excludes \geq half of right palm samples
 - How many samples will be excluded from your analysis based on this choice?
 - 4, all from right palm of subject 1
 - How many total sequences will you be analyzing in the core metrics command?
 - 24,000 (23.40%)
- Note: there is no single visualization for core metrics

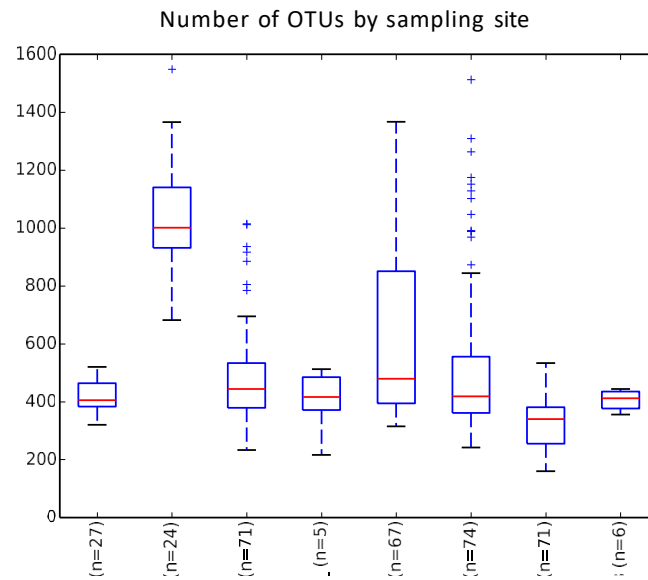


AlphaDiversity

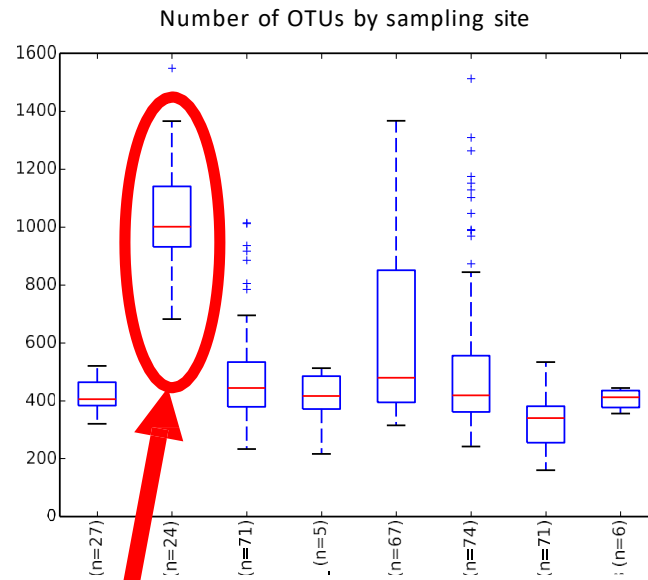
- “Within-sample” diversity
 - Many different metrics exist
 - Taxonomy-based (e.g., number of observed OTUs)
 - Assume everything is equally dissimilar
 - More likely to see differences based on close relatives
 - Phylogeny-based (e.g., phylogenetic diversity over whole tree)
 - Treat less related items as more dissimilar
 - Better at scaling the observed differences
 - The “correct” metric(s) are those relevant to your hypothesis
 - Please do HAVE a hypothesis!
- Testing approach:
 - Examine alpha diversity metric by metadata values
 - Test whether differences in metric distribution is different between groups (if metadata is categorical) or correlated with metadata (if metadata is continuous)



Alpha Diversity



Alpha Diversity



High within-sample diversity—why?



Practicum: Alpha Diversity Group Significance

```
qiime diversity alpha-group-significance \  
  --i-alpha-diversity metrics/faith_pd_vector.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/faith-pd-group-significance.qzv
```

- Note: only showing you the group significance visualization of ONE alpha diversity metric
 - Remember that 3 others are calculated by `core-metrics-phylogenetic` alone
 - *The one I am showing is not “the correct one”—pick the one that fits your hypothesis*
- To check the group significance of a different metric, just input a different vector file
 - To find them:

```
cd metrics/  
ls *_vector.qza
```



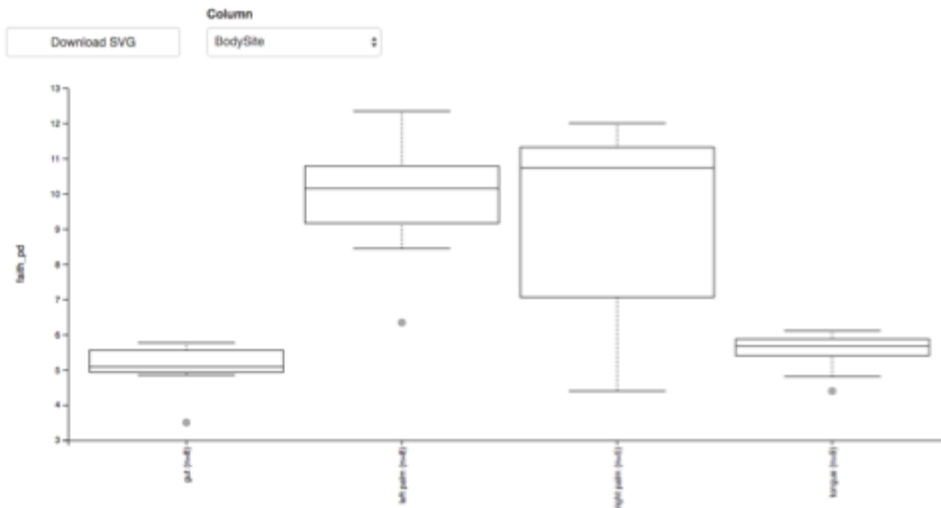
Alpha Diversity Group Significance View



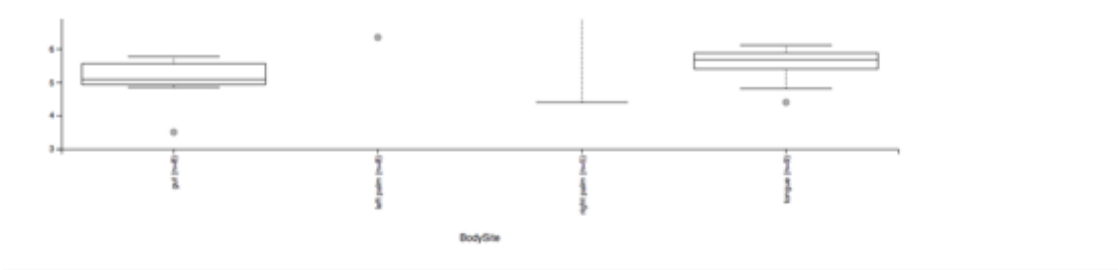
The following metadata columns have been omitted because they didn't contain categorical data: Day, DaysSinceExperimentStart, Month, Year

The following categorical metadata columns have been omitted because the number of groups was equal to the number of samples, there was only a single group, or the column consisted only of missing data: BarcodeSequence, Description, LinkerPrimerSequence

Alpha Diversity Boxplots



Alpha Diversity Group Significance View



Kruskal-Wallis (all groups)	
Result	
H	16.60616487455198
p-value	0.0008515500394924999

Kruskal-Wallis (pairwise)		H	p-value	q-value
Group 1	Group 2			
gut (n=8)	left palm (n=8)	11.294118	0.000778	0.002333
	right palm (n=5)	3.621429	0.057040	0.107791
	tongue (n=9)	1.564815	0.210962	0.253154
left palm (n=8)	right palm (n=5)	0.000000	1.000000	1.000000
	tongue (n=9)	12.000000	0.000532	0.002333
right palm (n=5)	tongue (n=9)	3.240000	0.071861	0.107791

Exercise: Alpha Diversity Group Significance

```
qiime diversity alpha-group-significance \  
  --i-alpha-diversity metrics/faith_pd_vector.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/faith-pd-group-significance.qzv
```

- **Work with your partner to answer these questions:**
 - Is BodySite value associated with significant differences in phylogenetic diversity?
 - Which two sites have the most significant difference in phylogenetic diversity distributions?
 - Note difference between p-value and q-value
 - Is Subject value associated with significant differences in phylogenetic diversity?



Answers: Alpha Diversity Group Significance

```
qiime diversity alpha-group-significance \  
  --i-alpha-diversity metrics/faith_pd_vector.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/faith-pd-group-significance.qzv
```

- My answers:

- Is BodySite value associated with significant differences in phylogenetic diversity?
 - Yes, with $p < 1 \text{ E-}3$
- Which two sites have the most significantly difference in phylogenetic diversity distributions?
 - Left palm is (equally) most significantly different from gut and tongue
 - Consider: any idea why perhaps left palm but not right?
- Is Subject value associated with significant differences in phylogenetic diversity?
 - No



Practicum: Alpha Diversity Correlation

```
qiime diversity alpha-correlation \  
  --i-alpha-diversity metrics/evenness_vector.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/evenness-alpha-  
correlation.qzv
```

- Same caveat as before:
 - Only showing the correlation visualization of ONE alpha diversity metric
 - Not necessarily “the correct one”!

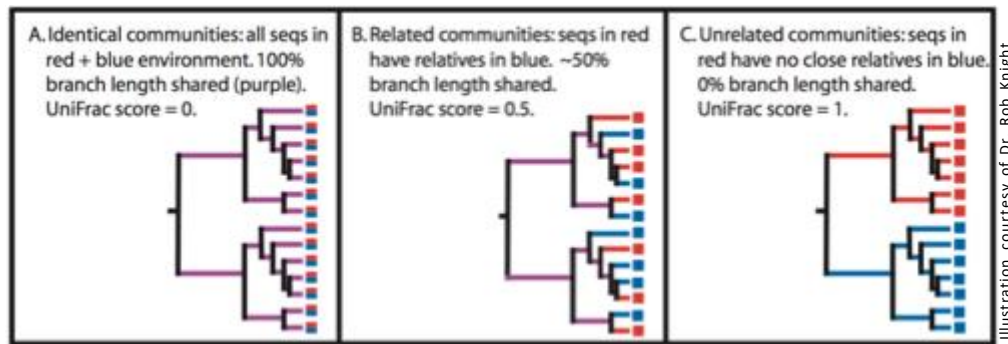


Alpha Diversity Correlation View



Beta Diversity

- “Between-sample” diversity
 - Has similar categories, caveats as α diversity
- A popular phylogenetic option is 'UniFrac':
 - Measures how different two samples' component sequences are



- Weighted UniFrac: takes abundance each sequence into account



Beta Diversity Ordination

- **Ordination:** multivariate techniques that arrange samples along axes on the basis of composition
- **Principal Coordinates Analysis:** a way to map non-Euclidean distances into a Euclidean space to enable further investigation
 - Abbreviated as PCoA, not to be confused with PCA (Principal Component Analysis)
 - Starting point is distance matrix
 - NOT the full set of independent variables for each sample
 - n pairwise distances are projected into n-1 dimensions
 - PCA performed to reduce the dimensionality back down
- PCoA axes can't be decomposed into independent variable contributions
 - But results can be compared to metadata to identify patterns

	A	B	C
A	0	.3	.7
B	.3	0	.6
C	.7	.6	0

Distance Matrix



Practicum: Beta Diversity Ordination

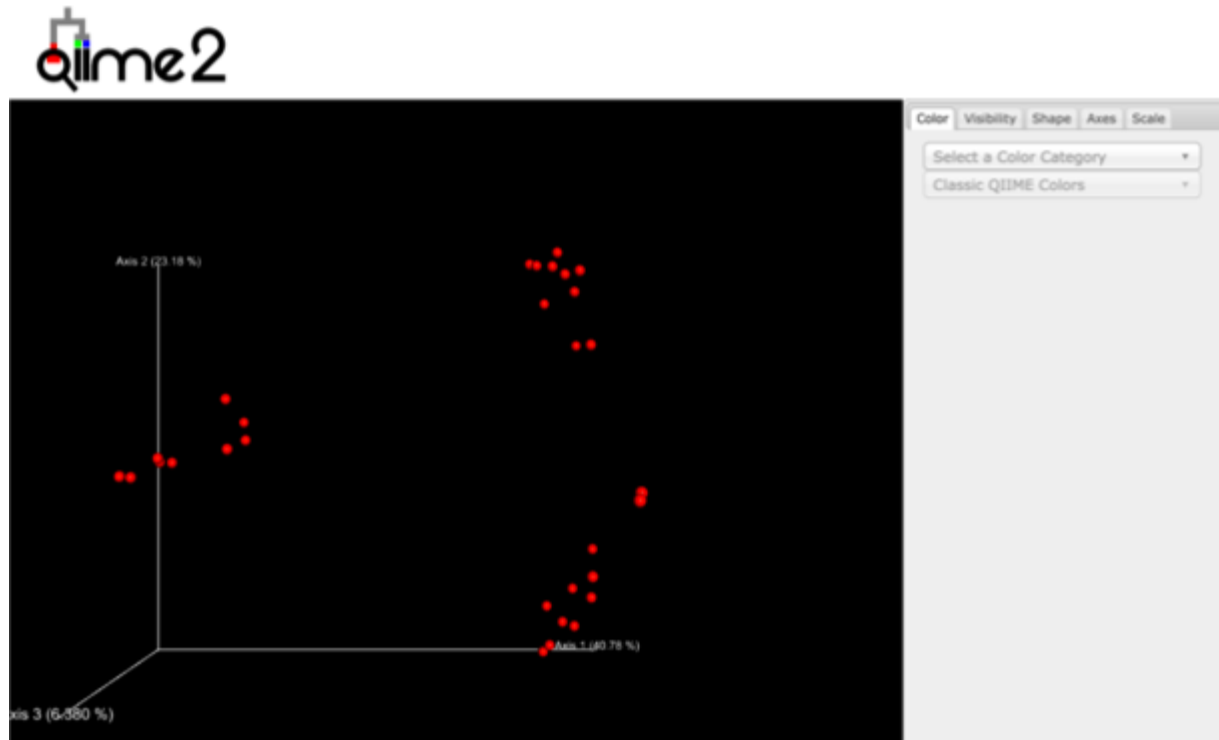
```
qiime emperor plot \  
  --i-pcoa metrics/unweighted_unifrac_pcoa_results.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/unweighted-unifrac-emperor.qzv
```

- Same caveat as before:
 - Only showing the PCoA visualization of ONE beta diversity metric
 - Not necessarily “the correct one”!
 - Remember that 3 others are calculated by `core-metrics-phylogenetic` alone
- To check the group significance of a different metric, just input a different vector file
 - To find them:

```
cd metrics/  
ls *_pcoa_results.qza
```



Beta Diversity Ordination View



Exercise: Beta Diversity Ordination

```
qiime emperor plot \  
  --i-pcoa metrics/unweighted_unifrac_pcoa_results.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/unweighted-unifrac-emperor.qzv
```

- Work with your partner to answer the following question:
 - Can you find a metadata category that appears associated with the observed clusters?
 - [Hint: Experiment with coloring points by different metadata](#)



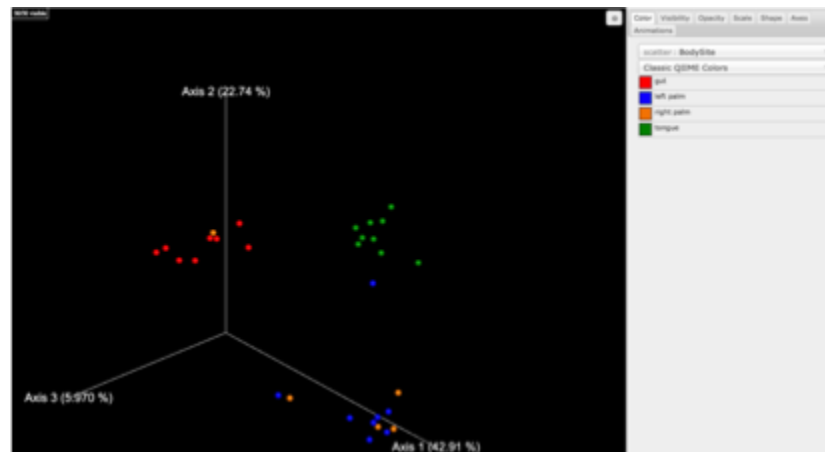
Answers: Beta Diversity Ordination

```
qiime emperor plot \  
  --i-pcoa metrics/unweighted_unifrac_pcoa_results.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization metrics/unweighted-unifrac-emperor.qzv
```

- My answer:

- Can you find a metadata category that appears associated with the observed clusters?

- Yep: BodySite



Practicum: Beta Diversity Ordination (cont.)

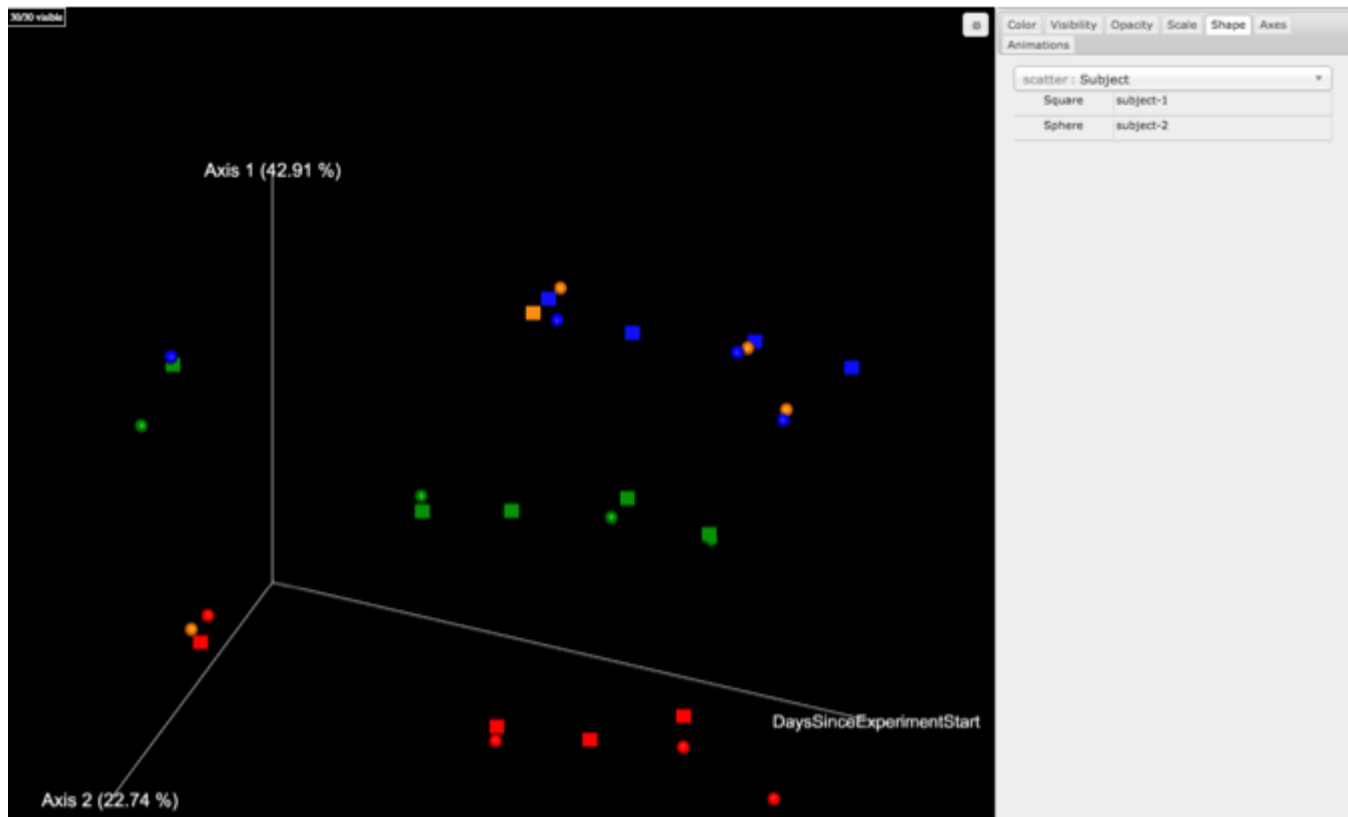
- But wait, this is time-series data!

```
qiime emperor plot \  
  --i-pcoa metrics/unweighted_unifrac_pcoa_results.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --p-custom-axes DaysSinceExperimentStart \  
  --o-visualization metrics/unweighted-unifrac-emperor-  
    bydayssince.qzv
```

- Standard caveats apply



Beta Diversity Ordination View (cont.)



Practicum: Beta Diversity Group Significance

- qiime diversity beta-group-significance \
 - --i-distance-matrix
metrics/unweighted_unifrac_distance_matrix.qza \
 - --m-metadata-file sample-metadata.tsv \
 - --m-metadata-column BodySite \
 - --p-pairwise \
 - --o-visualization metrics/unweighted-unifrac-bodysite-significance.qzv
- Standard caveats apply

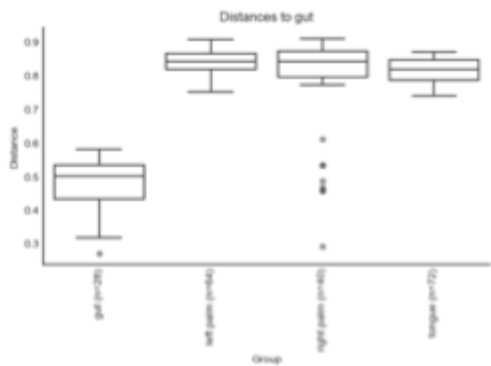


Beta Diversity Group Significance

View



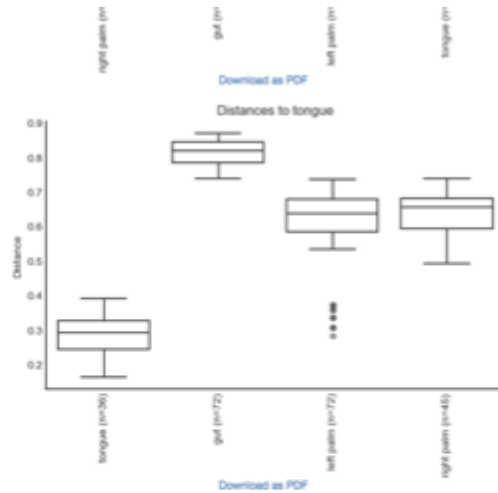
PERMANOVA results	
method name	PERMANOVA
test statistic name	pseudo-F
sample size	30
number of groups	4
test statistic	11.3089
p-value	0.001
number of permutations	999



[Download as PDF](#)



Beta Diversity Group Significance View



Pairwise permanova results

[Download CSV](#)

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
gut	left palm	16	999	16.532145	0.002	0.0024
	right palm	13	999	7.442433	0.001	0.0020
	tongue	17	999	28.727307	0.001	0.0020
left palm	right palm	13	999	0.495504	0.974	0.9740
	tongue	17	999	12.465506	0.001	0.0020
right palm	tongue	14	999	6.999375	0.002	0.0024

Exercise: Beta Diversity Group Significance

- qiime diversity beta-group-significance \
 - --i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \
 - --m-metadata-file sample-metadata.tsv \
 - --m-metadata-category BodySite \
 - --p-pairwise \
 - --o-visualization metrics/unweighted-unifrac-bodysite-significance.qzv
- Work with your partner to answer these questions:
 - Does the group significance analysis bear out your intuition from the ordination?
 - If so, are the differences statistically significant?
 - Are there specific pairs of BodySite values that are significantly different from each other?
 - How about Subject?
 - Hint: you will need to run a new command!



Answers: Beta Diversity Group Significance

- qiime diversity beta-group-significance \
 - --i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \
 - --m-metadata-file sample-metadata.tsv \
 - --m-metadata-category BodySite \
 - --p-pairwise \
 - --o-visualization metrics/unweighted-unifrac-bodysite-significance.qzv
- My answers:
 - Does the group significance analysis bear out your intuition from the ordination?
 - Yes
 - If so, are the differences statistically significant?
 - –Yes, with $p \leq 0.001$ (bonus: why do I say “**less than** or equal to”?)
 - Are there specific pairs of BodySite values that are significantly different from each other?
 - –Yes, all of the pairs except left palm/right palm



Taxonomic Assignment

- Sequence features or OTUs have limited utility
 - At some point, you'll want to link your findings to published work
 - That requires identifying the taxonomy of each sequence feature
- Steps:
 - Pick reference database
 - I hear you cry, "Which one should I use?"
 - Train a classifier algorithm to assign taxonomies to sequences
 - Use the reference database as the training set
 - Run the classifier algorithm on your sequence features



Taxonomic Assignment

- Sequence features or OTUs have limited utility
 - At some point, you'll want to link your findings to published work
 - That requires identifying the taxonomy of each sequence feature
- Steps:
 - Pick reference database
 - I hear you cry, **"Which one should I use?"**
 - Train a classifier algorithm to assign taxonomies to sequences
 - Use the reference database as the training set
 - Run the classifier algorithm on your sequence features



Common Issues in Marker Gene Studies

- Neglecting metadata
 - Analysis can not test for effects of, or discard bias from, categories you didn't record!
- Picking novel 16S primers—not all created equal
 - Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes
- Not taking precautions to support amplicon sequencing
 - Some Illumina machines require high PhiX, low cluster density
- Selecting an inappropriate reference database
 - E.g., Greengenes (16S) reference database when sequencing ITS
- Expecting species level taxonomy calls
 - Most O sequence variants only specify to family or genus level
- Using inappropriate statistical tests
 - Taxa abundance requires a compositionality-aware test like ANCOM
 - Differences in β diversity distances across groups requires test like PERMANOVA, not ANOVA



Marker Gene Reference Databases

- NOT a complete list:
 - Greengenes: 16S
 - Silva: 16S/18S
 - RDP: 16S/18S/28S
 - UNITE: ITS
- Another not complete list at eukref.org/databases (not just eukaryotic)
- At the very least, choose a database that includes your marker gene!
 - Beyond that, formal guidance is hard to find
 - But off the record you might get some informal guidance 😊



Taxonomic Assignment

- Sequence features or OTUs have limited utility
 - At some point, you'll want to link your findings to published work
 - That requires identifying the taxonomy of each sequence feature
- Steps:
 - Pick reference database
 - I hear you cry, "Which one should I use?"
 - Train a classifier algorithm to assign taxonomies to sequences
 - Use the reference database as the training set
 - Run the classifier algorithm on your sequence features



Common Issues in Marker Gene Studies

- Neglecting metadata
 - Analysis can not test for effects of, or discard bias from, categories you didn't record!
- Picking novel 16S primers—not all created equal
 - Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes
- Not taking precautions to support amplicon sequencing
 - Some Illumina machines require high PhiX, low cluster density
- Selecting an inappropriate reference database
 - E.g., Greengenes (16S) reference database when sequencing ITS
- Expecting species-level taxonomy calls
 - Most sequence variants only specify to family or genus level
- Using inappropriate statistical tests
 - Taxa abundance requires a compositionality-aware test like ANCOM
 - Differences in β diversity distances across groups requires test like PERMANOVA, not ANOVA



Taxonomy: Expectation Vs Reality

	Ideal Result	• Real Result
Kingdom	Bacteria	• Bacteria
Phylum	Proteobacteria	• Proteobacteria
Class	Gammaproteobacteria	• Gammaproteo
Order	Enterobacteriales	bacteria
Family	Enterobacteriaceae	Enterobacteriale
Genus	<i>Eschericia</i>	s
Species	<i>coli</i>	• Enterobacteria
Strain	O157:H7	ceae
		• ---
		• OTU 2445338
		• --

Practicum: Taxonomic Assignment

```
qiime feature-classifier classify-sklearn \  
  --i-classifier gg-13-8-99-515-806-nb-classifier.qza \  
  --i-reads rep-seqs.qza \  
  --o-classification taxonomy.qza
```

- Note: the classifier has already been trained for you
 - trained on the Greengenes 13_8 99% OTUs
 - sequences trimmed to only include 250 bases from the region of the 16S that was sequenced in this analysis
 - the V4 region, bound by the 515F/806R primer pair
- Other pre-trained classifiers available in Data Resources page on docs.qiime2.org



Practicum: Taxonomic Assignment

- qiime feature-classifier classify-sklearn \
 - --i-classifier gg-13-8-99-515-806-nb-classifier.qza \
 - --i-reads rep-seqs.qza \
 - --o-classification taxonomy.qza
- qiime metadata tabulate \
 - --m-input-file taxonomy.qza \
 - --o-visualization taxonomy.qzv



Taxonomic Assignment Tabulation View



Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

Feature ID #q2types	Taxon categorical	Confidence categorical
0160e14a78b18b903618f11bc732746e	k__Bacteria; p__Verrucomicrobia; c__Verrucomicrobiae; o__Verrucomicrobiales; f__Verrucomicrobiaceae; g__Akkermansia; s__muciniphila	0.9999999999838138
01b99cb344ed2530f7d80897fe257a9	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Burkholderiales; f__Comamonadaceae	0.9934747834222519
01ce91fd8dbecf637eb5e67cdab5c5aa	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__[Mogilibacteriaceae]; g__;	0.9992651687229891
01e0b7ac306895be84179f2715af269b	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Lachnospira; s__	0.9999998747932687
02ef9a59d6da8b642271166d3ffd1b52	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__;	0.9782760999752006
0305a4993ecf2d8ef4149dfc7592603	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__uniformis	0.9967281012905461

Practicum: Taxonomic Assignment

- qiime taxa barplot \
- --i-table table.qza \
- --i-taxonomy taxonomy.qza \
- --m-metadata-file sample-metadata.tsv \
- --o-visualization taxa-bar-plots.qzv



Taxonomic Assignment Bar Plot View



Download

SVG (bars)

SVG (legend)

CSV

Taxonomic Level

Level 1

Color Palette ⓘ

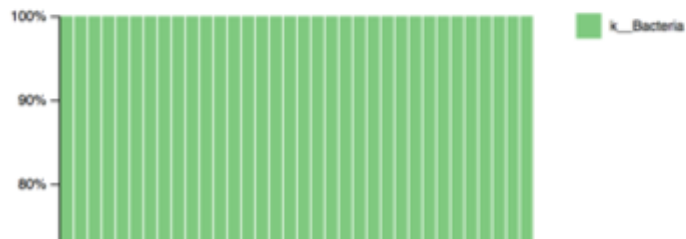
schemeAccent

Sort Samples By ⓘ

k__Bacteria

Ascending

Hover over the plot to learn more



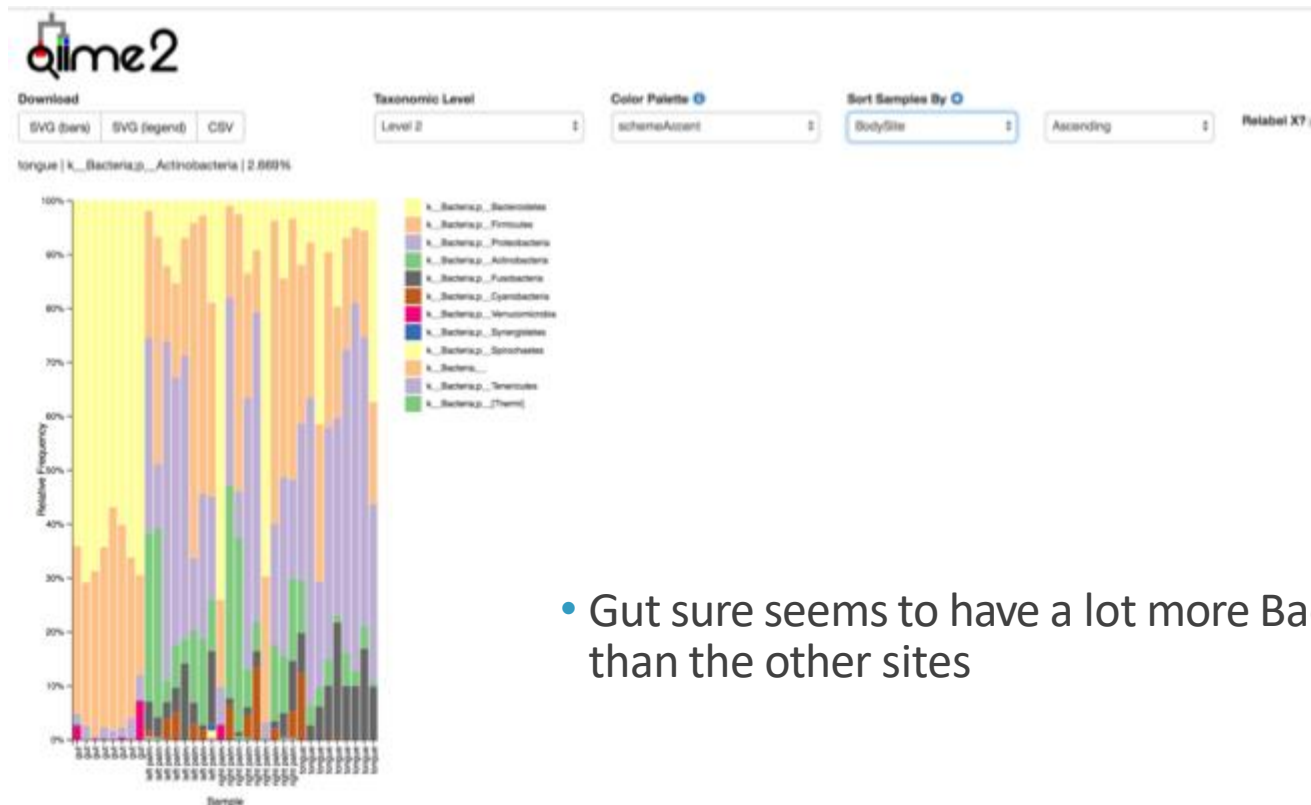
Exercise: Taxonomic Assignment

```
qiime taxa barplot \  
  --i-table table.qza \  
  --i-taxonomy taxonomy.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization taxa-bar-plots.qzv
```

- “Level 1” = kingdom, “Level 2” = phylum, etc
- Work with your partner to:
 - Visualize the taxa at level 2
 - Sort the samples by BodySite
 - Do you see anything suggestive?



Answers: Taxonomic Assignment

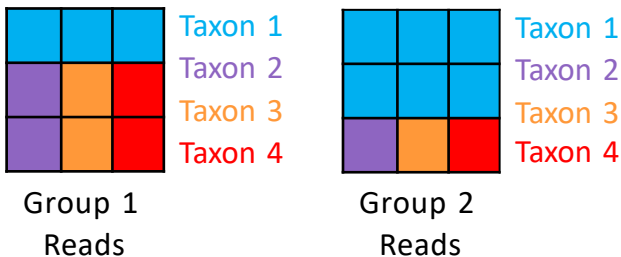


- Gut sure seems to have a lot more Bacteroidetes than the other sites



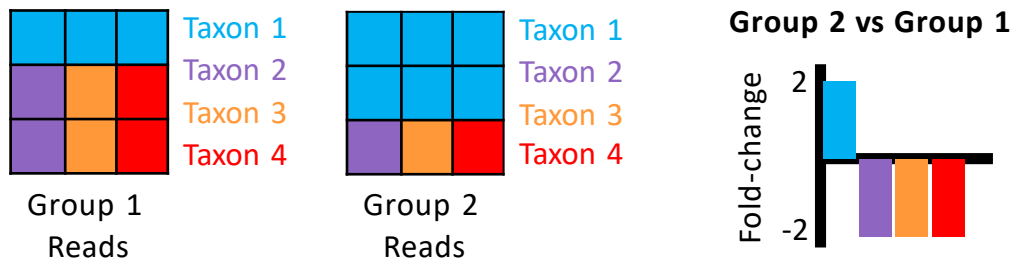
Differential Abundance Analysis

- Why go to the trouble of assigning taxonomies?
 - Probably you want to know whether any particular taxa are differentially abundant
 - In different individuals, environments, time points, etc
- How to test for differential abundance?
 - Microbiome datasets are “compositional” (fixed sum)



Differential Abundance Analysis

- Why go to the trouble of assigning taxonomies?
 - Probably you want to know whether any particular taxa are differentially abundant
 - In different individuals, environments, time points, etc
- How to test for differential abundance?
 - Microbiome datasets are “compositional” (fixed sum)



- **Watch out:** “traditional” statistical methods perform badly for this sort of data
 - e.g., t-test, ANOVA
 - False discovery rate can be as high as 90%!



Differential Abundance Analysis (cont.)

- What to use instead?
 - ANCOM (Analysis of Composition of Microbiomes)
 - Identifies taxa that are present in different abundances across sample groups
 - Compares log ratio of the abundance of each taxon to abundance of all remaining taxa one at a time
 - Assumes that <25% of the features are changing between groups—not a given!
 - ilr (Isometric Log Ratio transforms—a.k.a. balance trees or gneiss)
 - Identifies microbial subcommunities that present in different abundances across sample groups
 - Neither require rarefaction of inputs
 - Both recommend filtering out taxa that don't contain much info, such as
 - Features that have few reads (i.e. less than 10 reads across all samples).
 - Features that are rarely observed (i.e. present in less than 5 samples in a study).
 - Features that have very low variance (i.e. less than $10e-4$)
 - This is left as an exercise for the reader ☺ Check out `feature-table filter-samples`



Practicum: ANCOM Analysis

- We already saw in that different body sites look very different
- Given that, probably many features (sequences) are changing in abundance across body sites
 - **This violates ANCOM's statistical assumption!**
- Therefore, limit ANCOM analysis to samples from a single body site:

```
qiime feature-table filter-samples \  
  --i-table table.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --p-where "BodySite='gut'" \  
  --o-filtered-table gut-table.qza
```



Practicum: ANCOM Analysis (cont.)

- Can run ANCOM on individual “features” (sequences) but isn’t very informative
- Often more useful to collapse features to a chosen taxonomic level before ANCOM
 - Here I chose level 6, e.g., genus

```
qiime taxa collapse \  
  --i-table gut-table.qza \  
  --i-taxonomy taxonomy.qza \  
  --p-level 6 \  
  --o-collapsed-table gut-table-level6.qza
```



Practicum: ANCOM Analysis (cont.)

- Internally, ANCOM takes logs—and log of zero is undefined
 - Common approach: add one count to every value (pseudocount)
 - Not limited to ANCOM—used for any log-based method

```
qiime composition add-pseudocount \  
  --i-table gut-table-level6.qza \  
  --o-composition-table comp-gut-table-level6.qza
```



Practicum: ANCOM Analysis (cont.)

- qiime composition ancom \
 - --i-table comp-gut-table-level6.qza \
 - --m-metadata-file sample-metadata.tsv \
 - --m-metadata-column Subject \
 - --o-visualization ancom-subject-level6.qzv



ANCOM

View

ANCOM statistical results

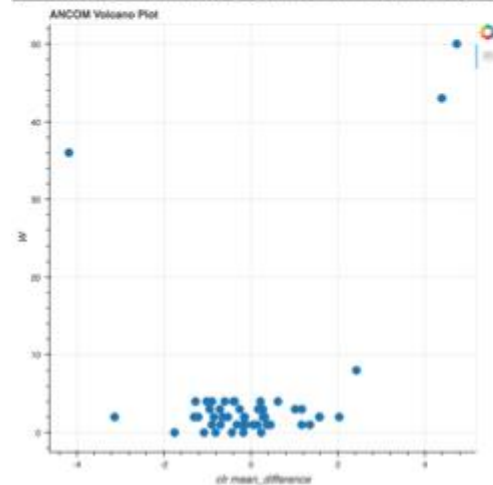
[Download complete table as CSV](#)

	W
k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides	50

Percentile abundances of features by group

[Download complete table as CSV](#)

Percentile	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
Group	subject-1	subject-1	subject-1	subject-1	subject-1	subject-2	subject-2	subject-2	subject-2	subject-2
k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Parabacteroides	1.0	1.0	1.0	1.0	1.0	76.0	100.75	112.0	158.0	287.0



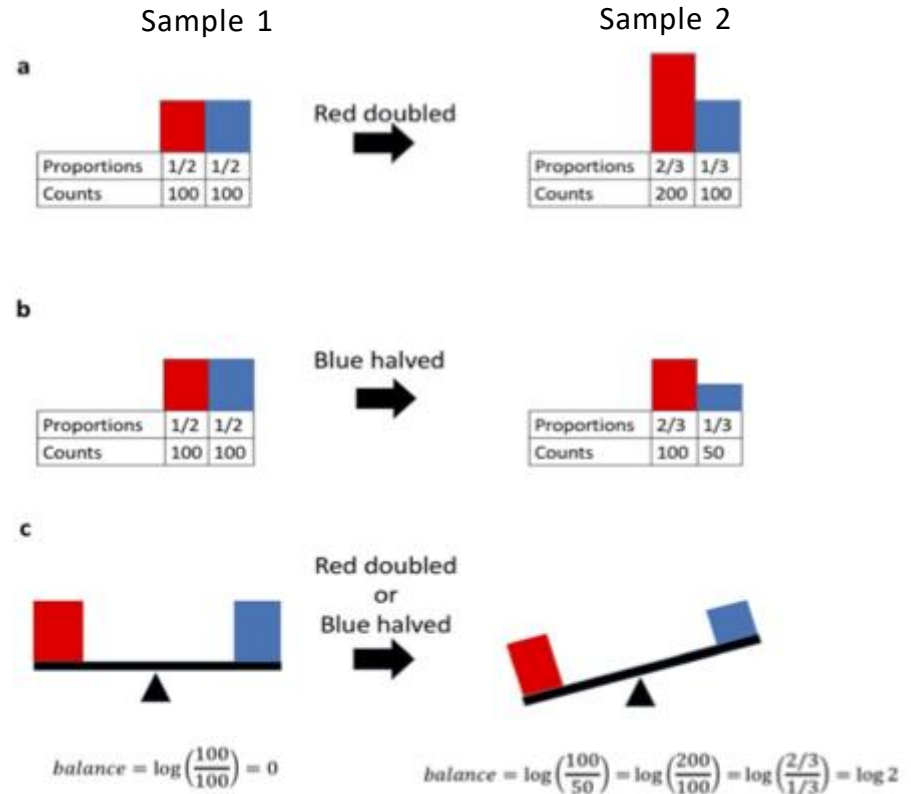
Details: ANCOM View

- W-statistic
 - # of features that a single feature is tested to be significantly different against
 - ANCOM internally decides what W value indicates significance, returns only significant results
 - You aren't going to find a p-value here, no matter how hard you look 😊
- Percentile abundance table:
 - A table of features and their percentile abundances in each group
 - Rows are features or taxa
 - Columns are percentile within a group
 - Values are abundance of reads for given percentile for that group
 - e.g.: The lowest-in-this-taxon 25% of samples in "group" subject-2 had 100.75 or fewer sequences assigned to this taxon



ilr and Balance Trees

- Microbiome sequence data give *proportions* of taxa abundance
 - Because of compositionality
- “[B]ased on proportions alone, it is impossible to determine whether the growth or decline of any individual species has truly occurred”
- Balance trees instead ask an answerable question:
 - Has the balance of sub-communities changed?



Quote and figure from Morton et al, mSystems 2017



ilr and Balance Trees (cont.)

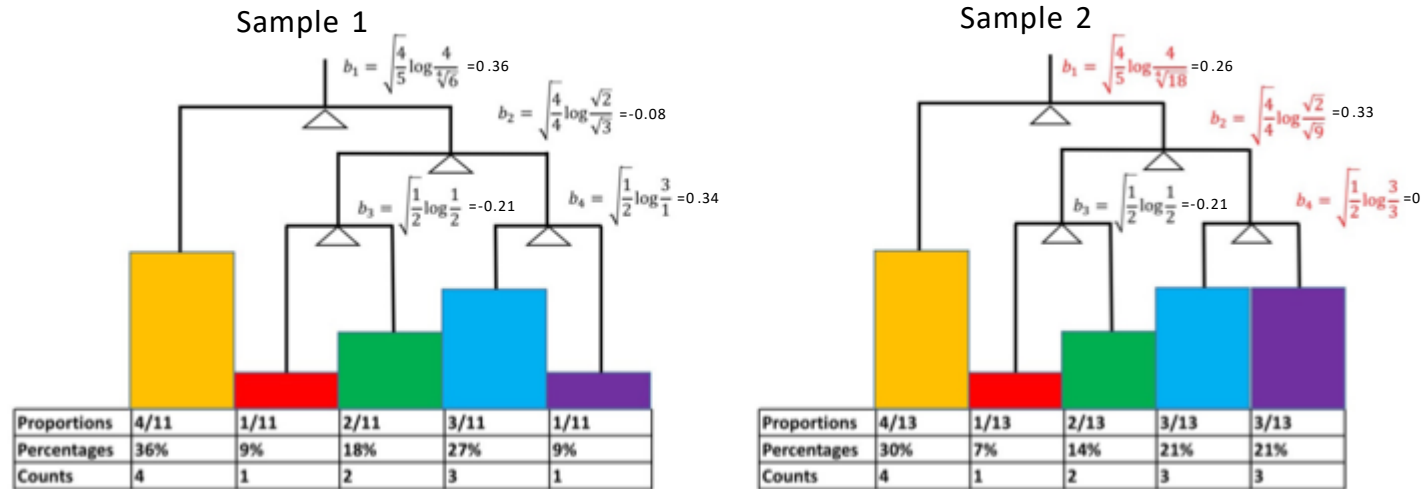


Fig modified from Morton et al, mSystems 2017

- The subcommunities of interest are defined by a tree of all the species
 - Each internal node in the tree is a “balance” between the subcommunities in its left and right children
- For each sample, at each balance, calculate the isometric log ratio transform (ilr)
 - Gives a measure of relative abundance of subcommunities on each side of the balance



ilr and Balance Trees (cont.)

- After ilr, result is a matrix
 - ilr values by sample by balance
- Simple case:
 - Only one balance is of interest
 - Can just do Student's t-test
- More realistic case:
 - All balances are of interest
 - Want to characterize whole community
 - Do multivariate response linear regression
 - Fit a linear regression model for each balance based on all samples for that balance
 - e.g., $Y_{b1} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots \beta_i X_i + \epsilon_{b1}$
 - Coefficients significantly $\neq 0$ = categories associated w/difference in subcommunities of balance

	b1	b2	b3	b4
Sample 1	0.36	-0.08	-0.21	0.34
Sample 2	0.26	0.33	-0.21	0.00

		b1	b2	b3	b4
Group 1	Sample 1	0.36	-0.08	-0.21	0.34
	Sample 3	0.33	-0.11	-0.20	0.29
	Sample 5	0.37	-0.03	-0.18	0.35
Group 2	Sample 2	0.26	0.33	-0.21	0.00
	Sample 4	0.28	0.30	-0.22	0.01
	Sample 6	0.25	0.34	-0.20	0.02



Practicum: gneiss

Analysis

- Balance trees were first developed in geology
 - “gneiss” (say: nice) is a kind of rock
- Like ANCOM, gneiss uses logs, so requires a pseudocount

```
qiime gneiss add-pseudocount \  
  --i-table table.qza \  
  --p-pseudocount 1 \  
  --o-composition-table composition.qza
```



Practicum: gneiss Analysis (cont.)

- The tree used defines which balances are assessed
 - i.e., which subcommunities of species are compared
 - How to choose it?
- Use an externally defined tree (e.g., phylogeny)
- Build a tree based on a numeric metric related to your hypothesis
 - e.g., from your metadata, such as pH
 - Using `gradient clustering gneiss` command
- Build a tree with unsupervised clustering across all your data
 - Group together organisms based on how often they co-occur with each other

```
qiime gneiss correlation-clustering \  
  --i-table composition.qza \  
  --o-clustering hierarchy.qza
```



Practicum: gneiss Analysis (cont.)

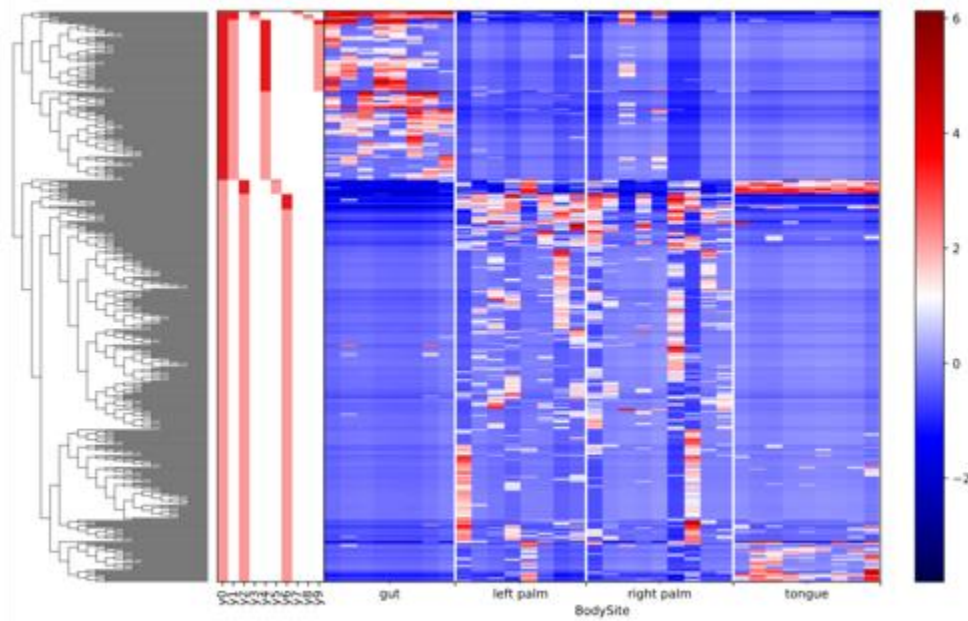
- What does the tree actually look like?

```
qiime gneiss dendrogram-heatmap \  
  --i-table composition.qza \  
  --i-tree hierarchy.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --m-metadata-column BodySite \  
  --p-color-map seismic \  
  --o-visualization  
  tree_heatmap_by_bodysite.qzv
```



gneiss Tree View

Dendrogram heatmap



Numerator Denominator

Practicum: gneiss Analysis (cont.)

- Calculate the ilr transforms

```
qiime gneiss ilr-transform \  
  --i-table composition.qza \  
  --i-tree hierarchy.qza \  
  --o-balances balances.qza
```



Practicum: gneiss Analysis (cont.)

- Fit the linear regression model
 - Requires deciding on your formula, based on your metadata

```
qiime gneiss ols-regression \  
  --p-formula "Subject+BodySite+DaysSinceExperimentStart" \  
  --i-table balances.qza \  
  --i-tree hierarchy.qza \  
  --m-metadata-file sample-metadata.tsv \  
  --o-visualization regression_summary.qzv
```

- As this is a time-course, could instead use `lme-regression` grouped by Subject



gneiss Regression Summary View

Simplicial Linear Regression Summary

No. Observations 34.0000

Model: OLS

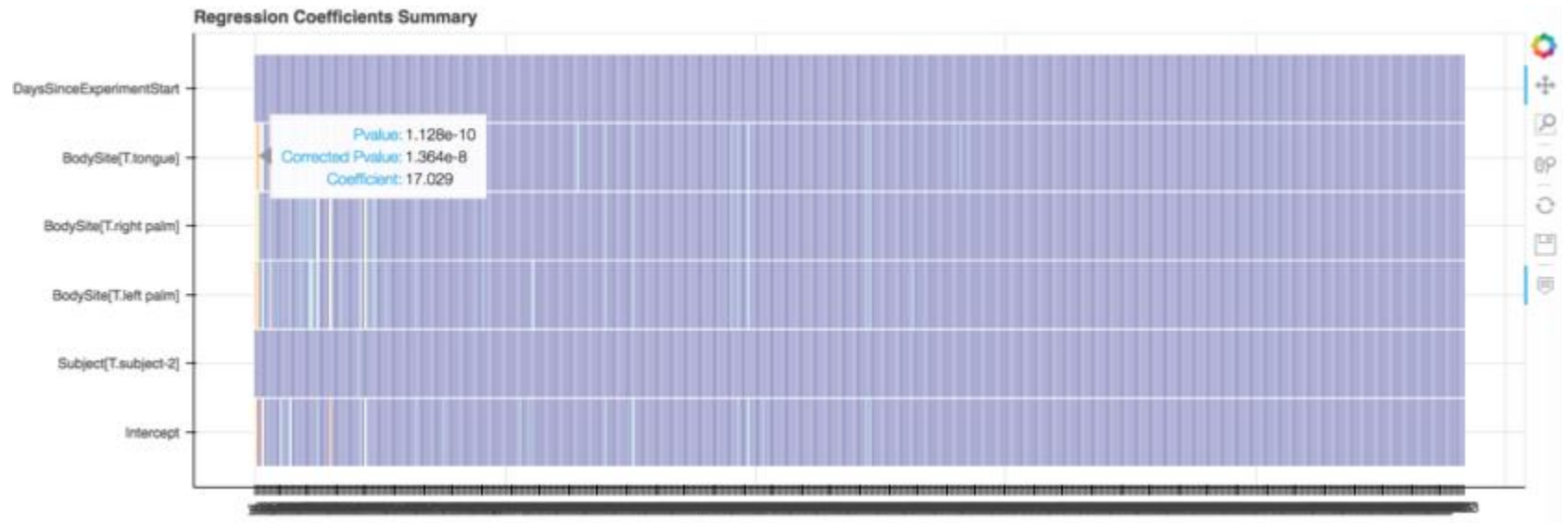
Rsquared: 0.4975

	mse	Rsquared	R2diff
Intercept	18.2693	0.3309	0.1665
Subject[T.subject-2]	15.2408	0.4418	0.0556
BodySite[T.left palm]	19.2142	0.2963	0.2011
BodySite[T.right palm]	18.5499	0.3207	0.1768
BodySite[T.tongue]	19.8765	0.2721	0.2254
DaysSinceExperimentStart	14.5560	0.4669	0.0305

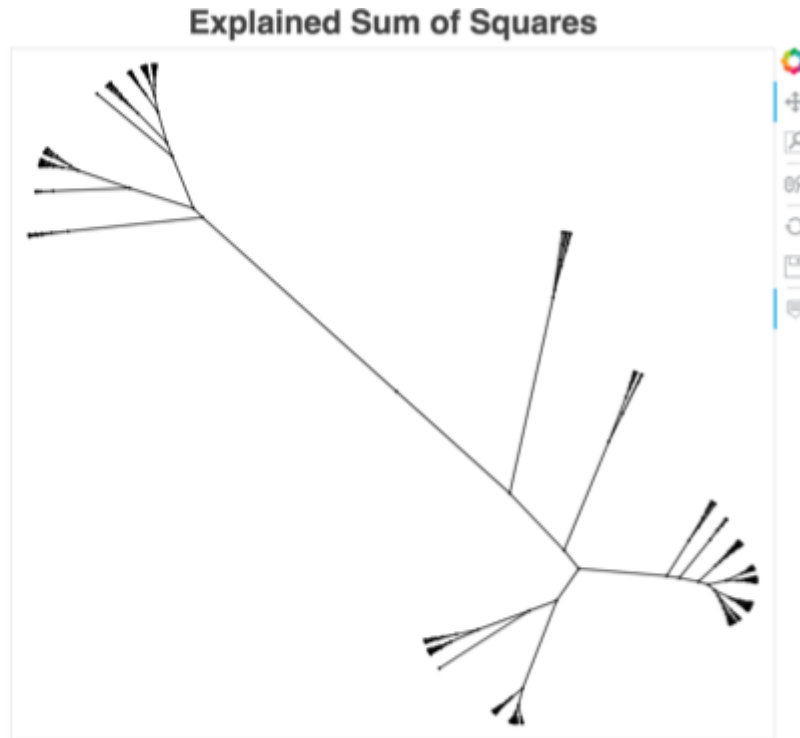
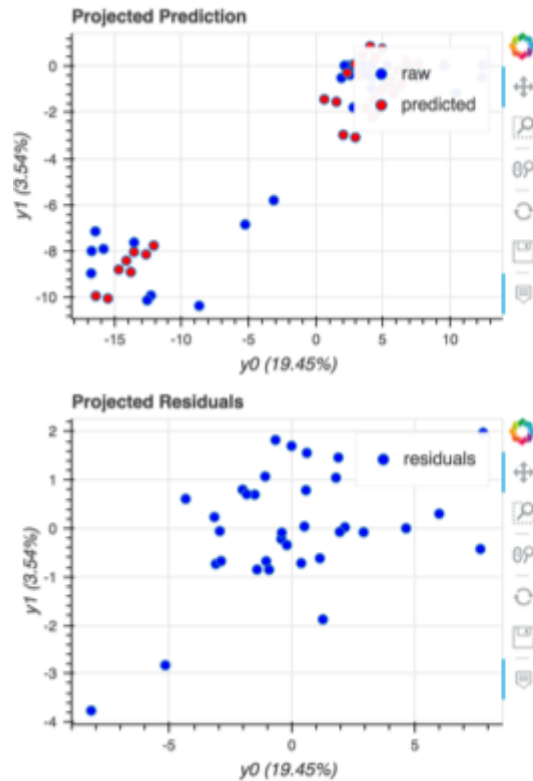
	model_mse	Rsquared	pred_mse
fold_0	11.0272	0.4668	1.8216
fold_1	10.9444	0.4592	2.1930
fold_2	10.8385	0.4979	2.1128
fold_3	10.6139	0.5143	2.5960
fold_4	10.8593	0.5234	2.4676
fold_5	10.6387	0.5243	2.2926
fold_6	11.3915	0.5102	1.4510
fold_7	9.6503	0.5400	3.6218
fold_8	10.6161	0.5135	2.3444
fold_9	12.0860	0.4713	0.7668

Coefficients [Download as CSV](#)

gneiss Regression Summary View (cont.)



gneiss Regression Summary View (cont.)



Practicum: gneiss Analysis

(cont.)

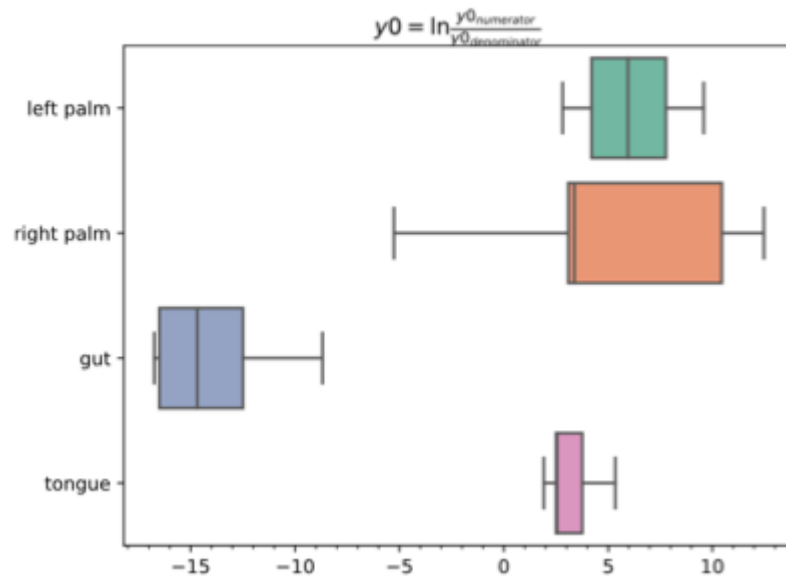
- Well, great, but we seem to have strayed a long way from actual taxa!

```
qiime gneiss balance-taxonomy \  
  --i-table composition.qza \  
  --i-tree hierarchy.qza \  
  --i-taxonomy taxonomy.qza \  
  --p-taxa-level 2 \  
  --p-balance-name 'y0' \  
  --m-metadata-file sample-metadata.tsv \  
  --m-metadata-column BodySite \  
  --o-visualization y0_taxa_summary.qzv
```



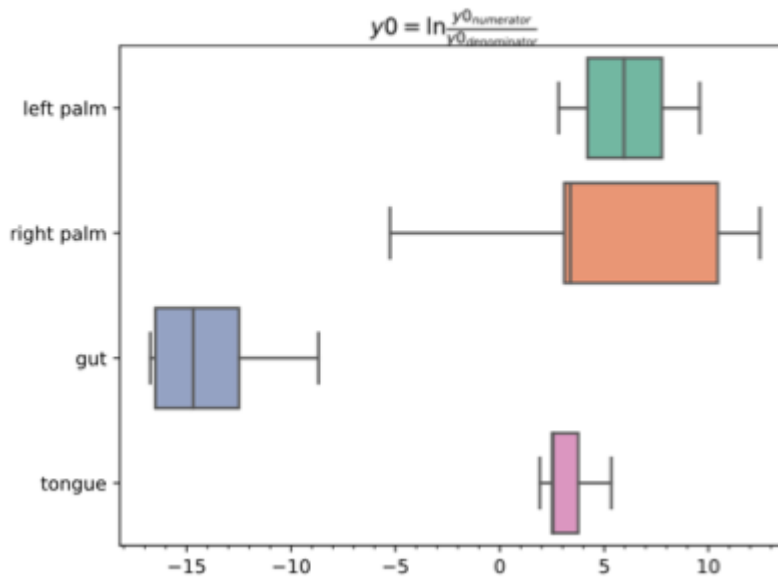
gneiss Balance Taxa Summary View

Balance vs BodySite



gneiss Balance Taxa Summary View

Balance vs BodySite

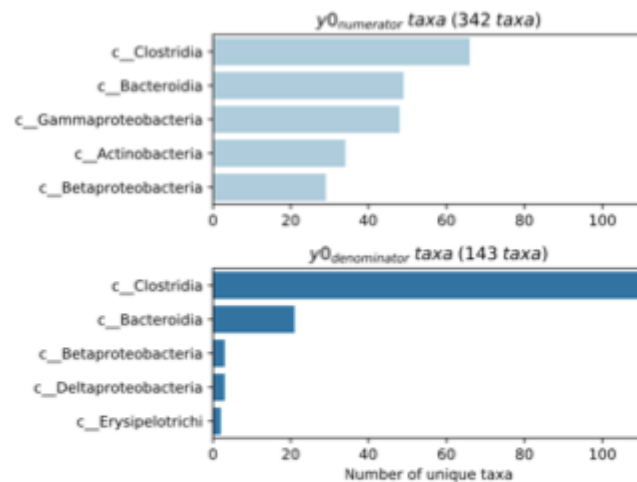


- Possible meanings
 - The taxa in the y_0 numerator on average increase in other sites compared to gut
 - The taxa in the y_0 denominator on average decrease in other sites compared to gut
 - A combination of the above occurs
 - Taxa abundances in both y_0 numerator and y_0 denominator both increase compared to gut, but taxa abundances in numerator increase more compared to denominator
 - Taxa abundances in both y_0 numerator and y_0 denominator both decrease, but taxa abundances in denominator increase more compared to numerator



gneiss Balance Taxa Summary View (cont.)

Balance Taxonomy



Numerator taxa

[Download as CSV](#)

Denominator taxa

[Download as CSV](#)

Practicum

Summary

- Steps practiced
 - Importing data
 - Demultiplexing
 - Running Quality Control
 - Creating a feature table
 - Building a phylogenetic tree
 - Calculating core diversity metrics
 - Testing alpha diversity group significance and correlation
 - Performing beta diversity ordination
 - Testing beta diversity group significance
 - Assigning taxonomies
 - Performing differential abundance analysis with ANCOM and/or gneiss



Acknowledgements

- Center for Computational Biology & Bioinformatics, University of California at San Diego
- Caporaso lab, Northern Arizona University
- Knight lab, UCSD
- ***QIIME 2 development team!***

