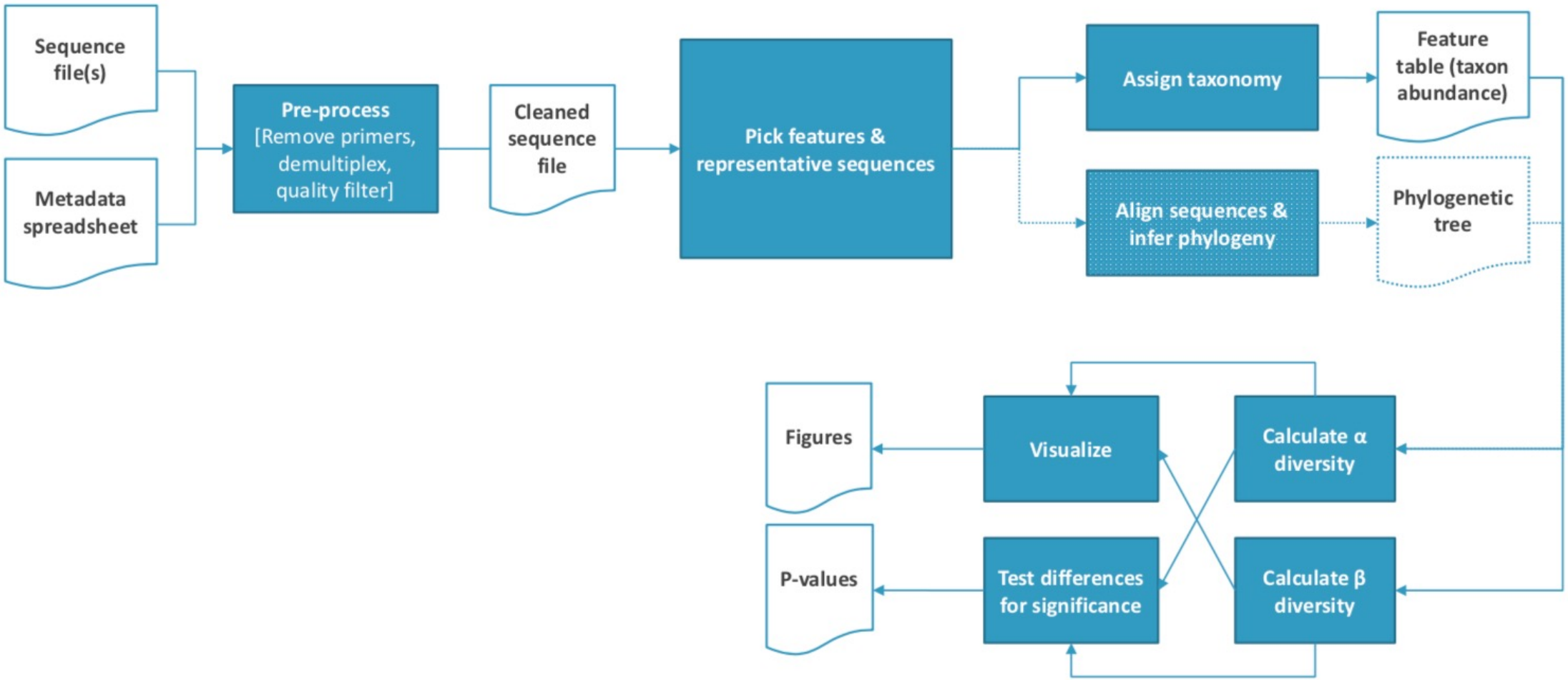
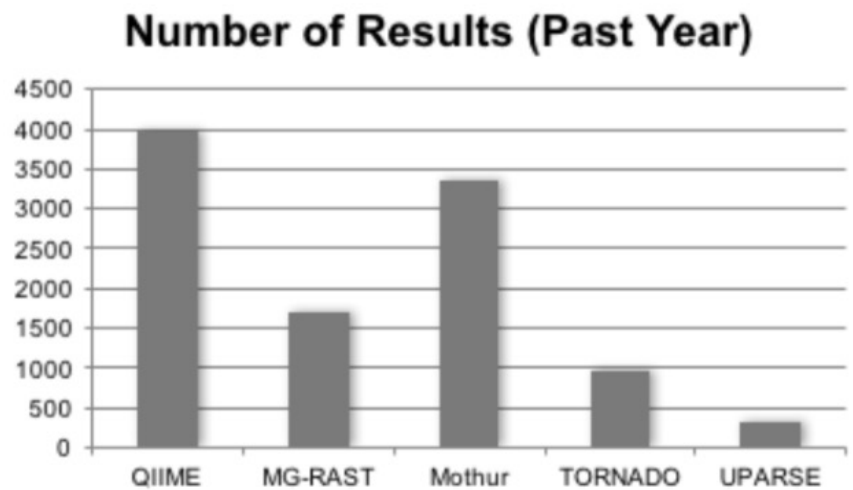
Microbiome (culture-independent) analysis

Marker Gene Analysis Workflow



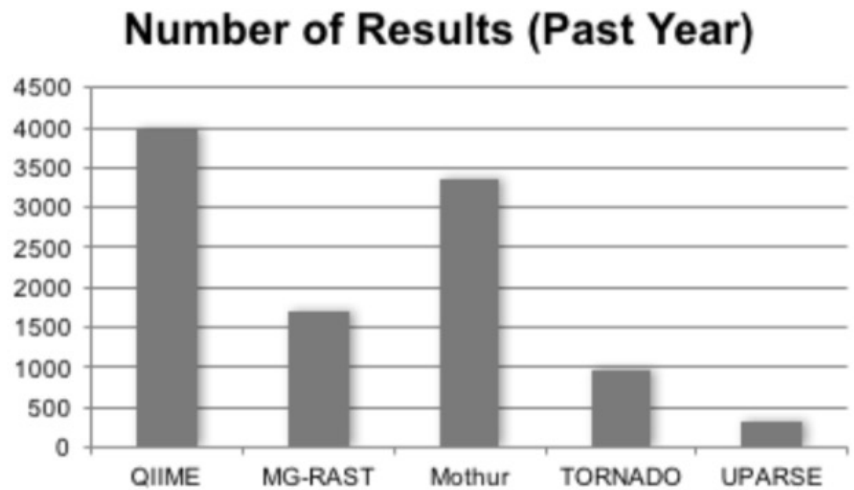
Software Selection

- 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



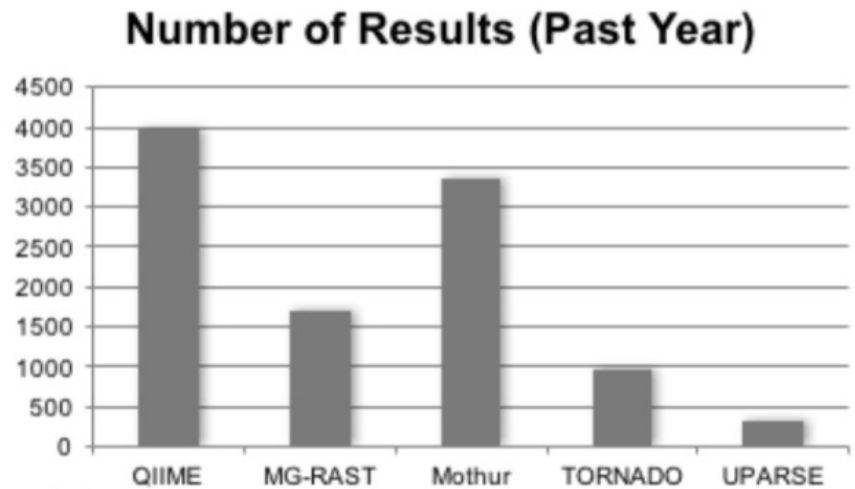
Software Selection

- 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 won’t be supported after end of 2017
  - QIIME 2 not yet feature-complete
    - But already much easier to use!
  - This tutorial uses QIIME 2 **only**
- **I’m not a QIIME 2 developer**
  - I’m not taking credit for this tool, just demonstrating it!



Making a Mapping File

#SampleID	LinkerPrimerSequence	BarcodeSequence	ReportedAntibioticUsage	DaysSinceExperimentStart	SampleType
L1S140	GTGCCAGCMGCCGCGGTAA	ATGGCAGCTCTA	Yes	0	gut
L2S155	GTGCCAGCMGCCGCGGTAA	ACGATGCGACCA	No	84	left palm

- "Mapping file" contains metadata for study
  - Must contain info needed to process sequences and test YOUR hypotheses
- QIIME 1 required certain columns in certain order, but QIIME 2 is more flexible
  - Tab-separated text file with column labels in first line + at least one data line
    - Column label values must be unique (i.e. no duplicate values)
  - First column is the "identifier" column (sample ID)
    - All values in the first column must be unique (i.e. no duplicate values)
  - See https://docs.qiime2.org/2017.6/tutorials/metadata/
- The easiest way to make a mapping file is with a spreadsheet
  - But **Excel is not your friend!**
    - Routinely corrupts gene symbols, anything interpreted as a dates, etc, & isn't reversible

Practicum: Viewing A Mapping File

- Open Terminal
  - For below, remember to try tab completion!

```
nano sample-metadata.tsv
```

- Stretch the window so you can look at the contents; then, to close, type

```
Ctrl + x
```

- Mapping file errors can lead to QIIME 2 errors—or worse, garbage results!
  - Keemei (pronounced ‘key may’) tool checks for errors in **Google Sheets**
    - **Chrome only,** and must have Google account to use
    - See http://keemei.qiime.org/

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`

> **Note**
>
> It has been brought to our attention that the term *artifact* may be confusing, as it is frequently used by biologists to refer to an experimental error. We use the term artifact here to mean an object that is made by some process (similar, for example, to an archaeological artifact). Throughout our documentation and other educational materials, we try to clarify that we are talking about *QIIME 2 artifacts* as they are defined in this section.

https://docs.qiime2.org/2017.6/concepts/#data-files-qiime-2-artifacts

- Different input commands for
  - Different kinds of input data (e.g., single-end vs paired-end)
  - Different formats of input data (e.g., sequences & barcodes in same or different file)

Practicum: Importing Data

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

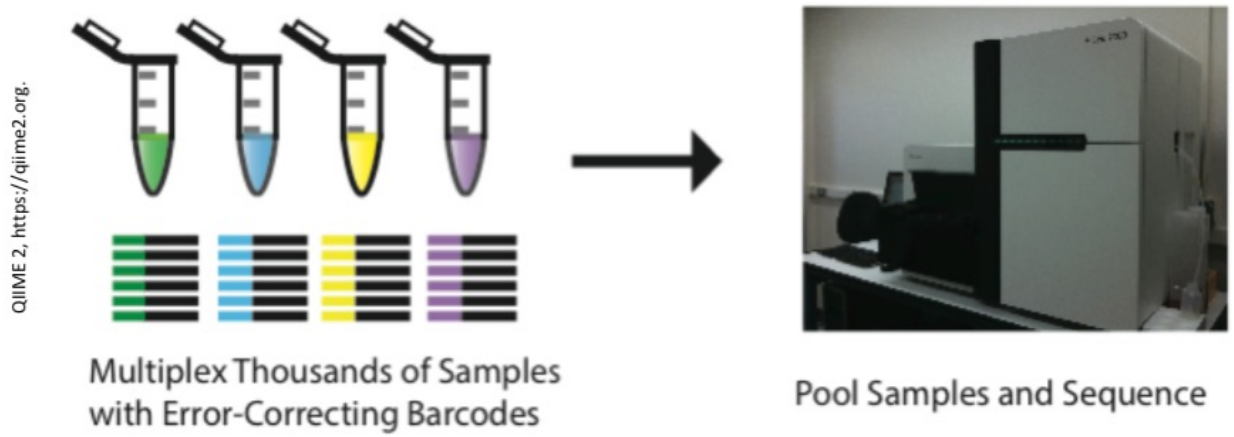
- A backslash \ is used to break up a command onto multiple lines
  - If you prefer to type the whole command onto one run-on line, you can leave it out

Practicum: Importing Data

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

- What does this command actually do?
  - Tells qiime to look into the folder `emp-single-end-sequences` …
  - For the kind of sequence files expected for `EMPSingleEndSequences` …
  - And load them into a new qiime artifact named `emp-single-end-sequences.qza`
- 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-category 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
- QIIME 2 artifact files can't be viewed directly (e.g., in nano)
- New concept: QIIME 2 visualization file
  - Has `.qzv` extension
  - Is intended for human (rather than computer) use
  - Generally provide info via a web browser

Practicum: Demultiplexing (cont.)

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

- Now view the visualization, locally

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

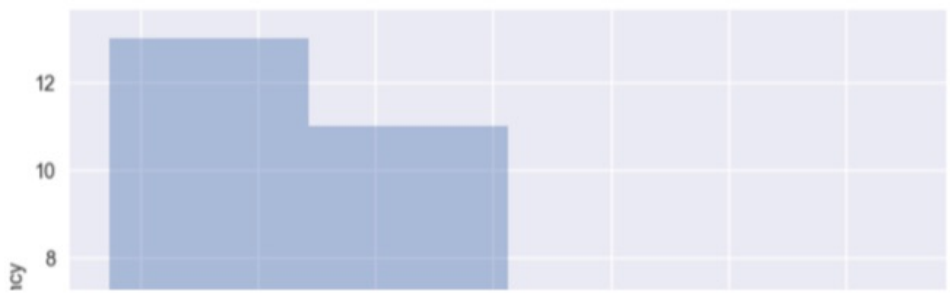
Practicum: Demultiplexing (cont.)

qiime2

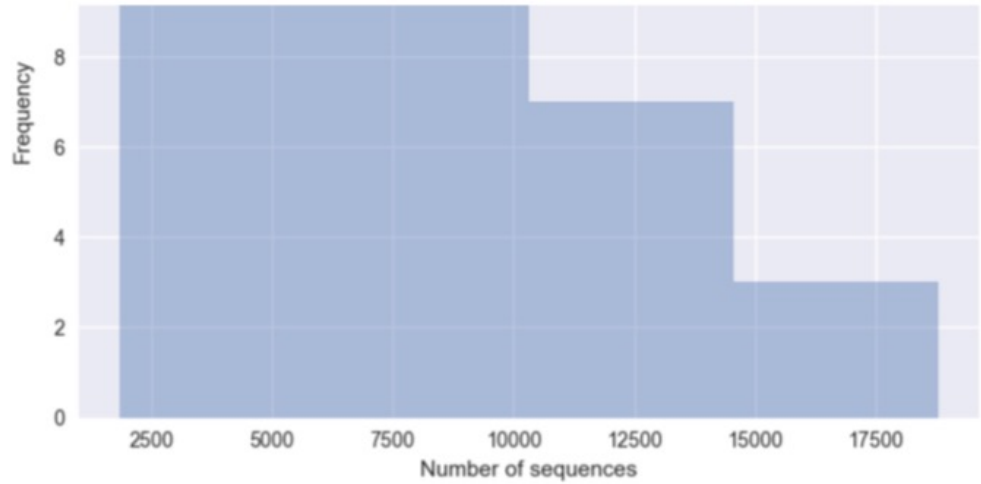
Overview | Interactive Quality Plot

Demultiplexed sequence counts summary

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



Practicum: Demultiplexing (cont.)

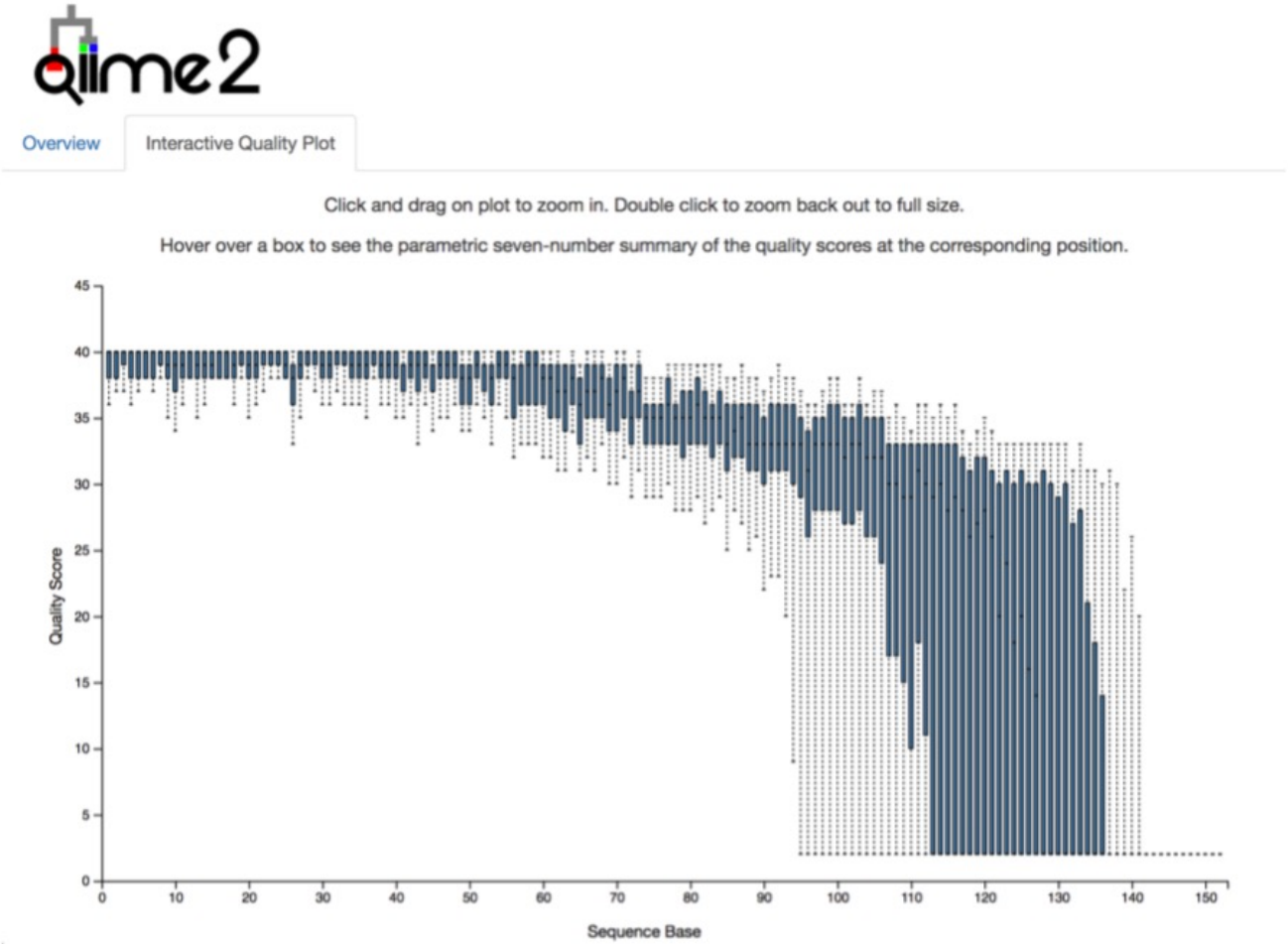


Download as PDF

Per-sample sequence counts

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

Practicum: Demultiplexing (cont.)



Practicum: Demultiplexing (cont.)

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

- 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)

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
```

- This runs the command with default values for all the tuneable parameters
  - To see the optional parameters, their descriptions, and their defaults, run just

```
qiime dada2 denoise-single \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left 0 \
  --p-trunc-len 120 \
  --o-representative-sequences rep-seqs-dada2.qza \
  --o-table table-dada2.qza \
  --o-denoising-stats stats-dada2.qza

qiime metadata tabulate \
  --m-input-file stats-dada2.qza \
  --o-visualization stats-dada2.qzv

mv rep-seqs-dada2.qza rep-seqs.qza
mv table-dada2.qza table.qza
```

Practicum: Quality Control

Per-sample sequence counts

	total-input-reads	total-retained-reads	fraction-retained	reads-truncated	reads-too-short-after-truncation	reads-exceeding-maximum-ambiguous-bases
sample-id						
Totals	263878	186324	0.706099	245862	76489	1065
L4S137	18787	11642	0.619684	17454	7123	22
L4S63	17167	11505	0.670181	15160	5634	28
L4S112	16265	10012	0.615555	15054	6232	21
L1S8	12386	8433	0.680849	12035	3916	37
L2S240	11986	7110	0.593192	11454	4845	31
L1S57	11750	10000	0.851064	11000	1716	34
L1S105	11340	9232	0.814109	10782	2066	42
L1S208	11335	10148	0.895280	10667	1161	26
L6S93	11270	8580	0.761313	10282	2680	10
L2S175	10691	5574	0.521373	10216	5092	25

Practicum: Feature Table Creation

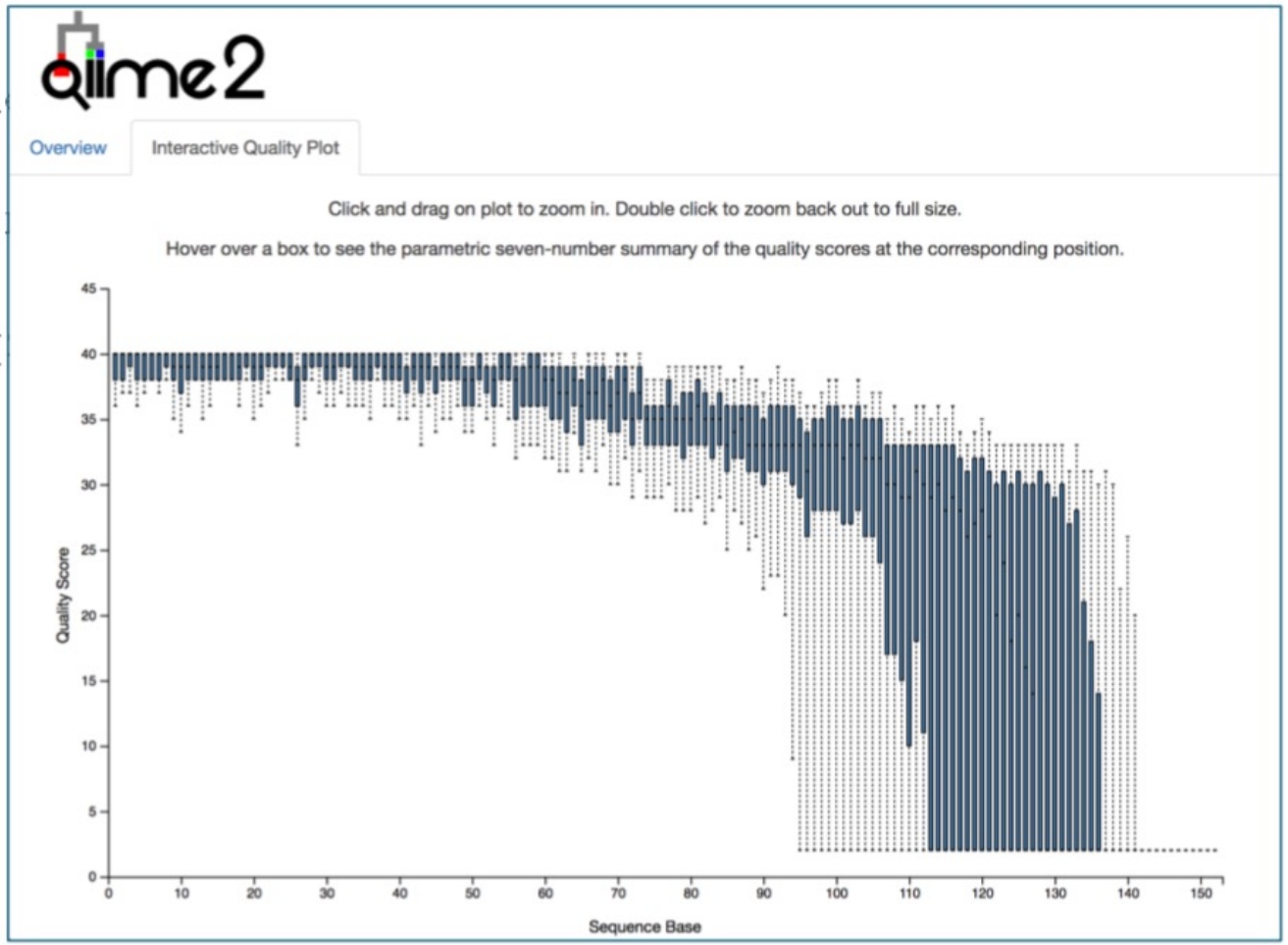
```
qiime dada2 denoise-single \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left 0 \
  --p-trunc-len 120 \
  --o-representative-sequences rep-seqs-dada2.qza \
  --o-table table-dada2.qza \
  --o-denoising-stats stats-dada2.qza
```

- This command can take a few minutes to run
  - So don’t worry if the command prompt doesn’t immediately return after you hit enter
- Where do you guess the number 120 came from?

Practicum: Feature Table Creation

```
qiime dada2 denoise-single \
  --i-demultiplexed-seqs demux.q
  --p-trim-left 0 \
  --p-trunc-len 120 \
  --o-representative-sequences r
  --o-table table-dada2.qza \
  --o-denoising-stats stats-dada
```

- Where do you guess the number 120 came from?
  - It is the length to which all sequences will be trimmed
  - It was chosen by viewing the Interactive Quality Plot in `demux.qzv`
  - You might even choose a more conservative length, like 110

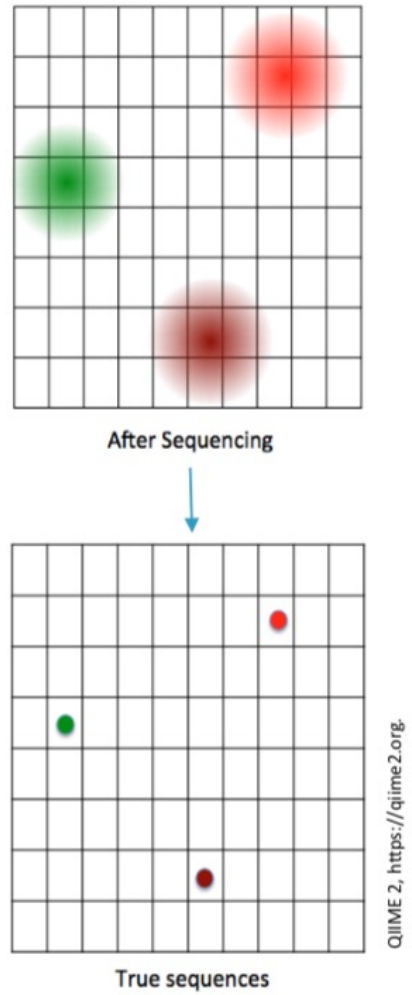


Feature Table Creation—The Past

- 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

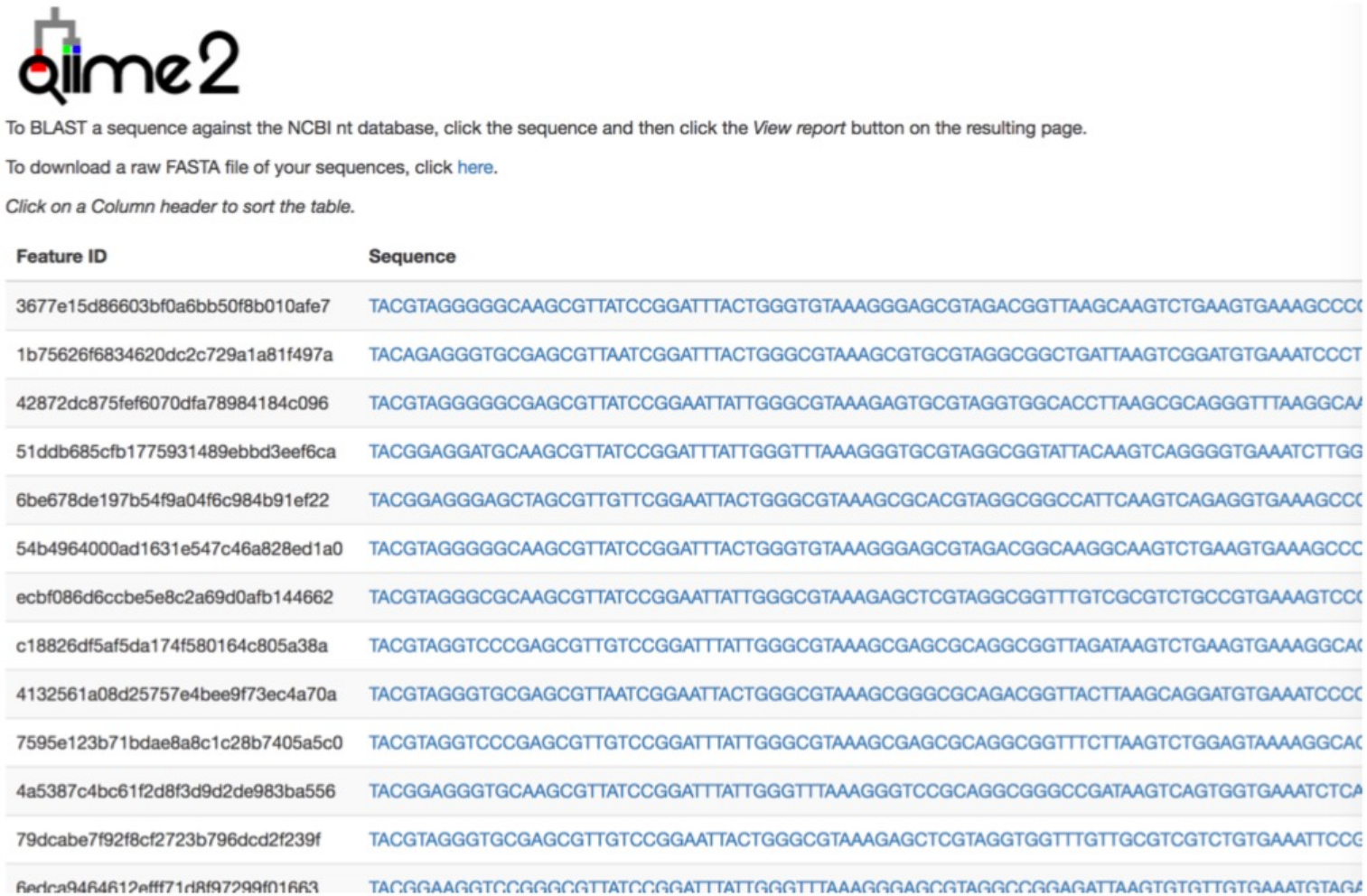
- This year: sOTU (sub-OTU) methods
  - Use error modeling to *in silco* correct sequencing mistakes
    - Sounds impossible but is actually quite accurate, with right error model
      - Error model is specific to the sequencing type (e.g., 454, Illumina Hi/MiSeq)
  - Result: only sequences likely to 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



Practicum: Feature Table Creation (cont.)

```
#FeatureTable and FeatureData summaries
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



To BLAST a sequence against the NCBI nt database, click the sequence and then click the *View report* button on the resulting page.

To download a raw FASTA file of your sequences, click here.

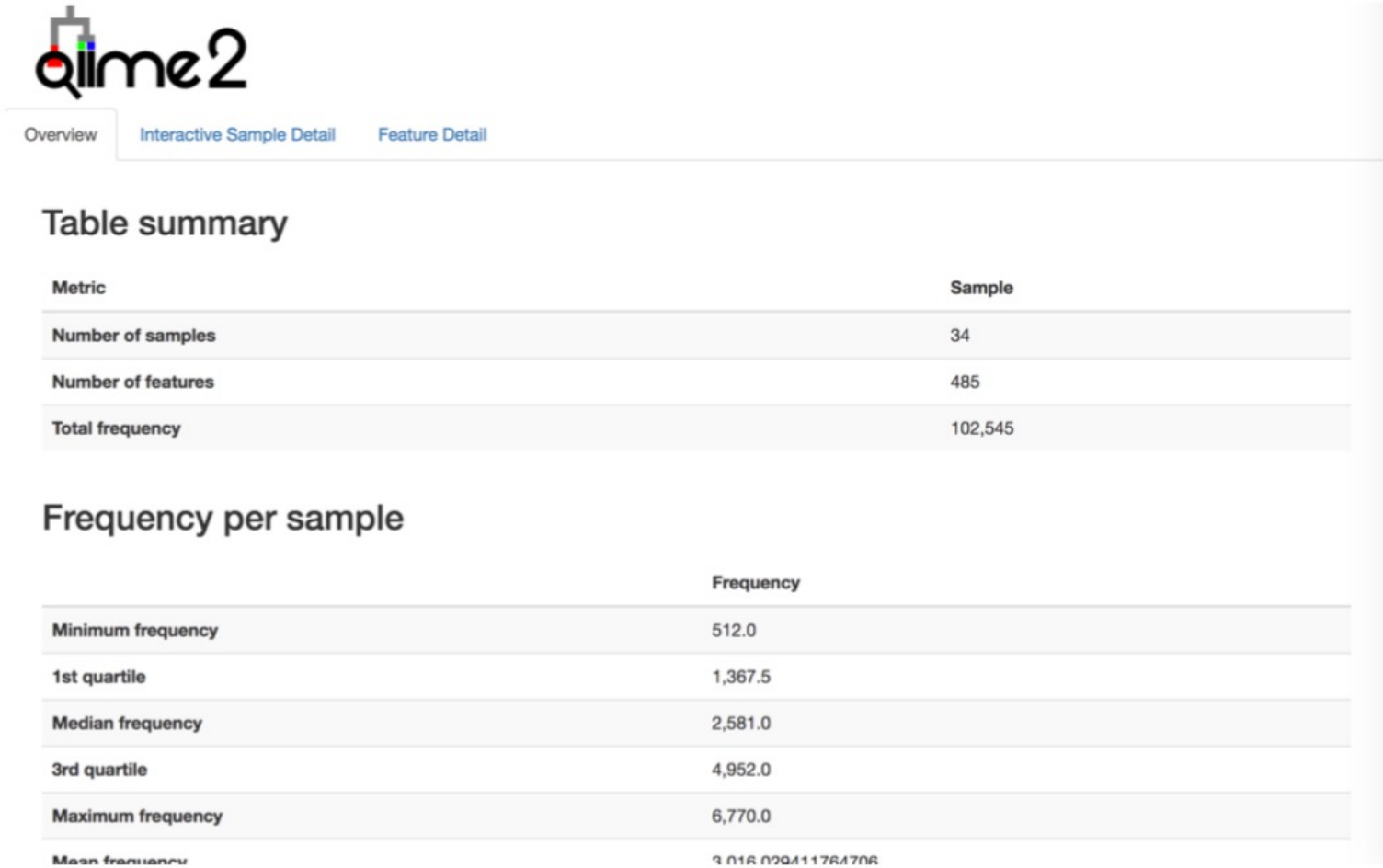
Click on a Column header to sort the table.

Feature ID	Sequence
3677e15d86603bf0a6bb50f8b010afe7	TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGTTAAGCAAGTCTGAAGTGAAAGCCC
1b75626f6834620dc2c729a1a81f497a	TACAGAGGGTGCGAGCGTTAATCGGATTTACTGGGCGTAAAGCGTGCGTAGGCGGCTGATTAAGTCGGATGTGAAATCCCT
42872dc875fef6070dfa78984184c096	TACGTAGGGGGCGAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGCACCTTAAGCGCAGGGTTTAAGGCAA
51ddb685cfb1775931489ebbd3eef6ca	TACGGAGGATGCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTATTACAAGTCAGGGGTGAAATCTTGG
6be678de197b54f9a04f6c984b91ef22	TACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCCATTCAAGTCAGAGGTGAAAGCCC
54b4964000ad1631e547c46a828ed1a0	TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGCAAGGCAAGTCTGAAGTGAAAGCCC
ecbf086d6ccbe5e8c2a69d0afb144662	TACGTAGGGCGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGCCGTGAAAGTCCC
c18826df5af5da174f580164c805a38a	TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTGAAAGGCAC
4132561a08d25757e4bee9f73ec4a70a	TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGGGCGCAGACGGTTACTTAAGCAGGATGTGAAATCCCC
7595e123b71bdae8a8c1c28b7405a5c0	TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTTCTTAAGTCTGGAGTAAAAGGCAC
4a5387c4bc61f2d8f3d9d2de983ba556	TACGGAGGGTGCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTCCGCAGGCGGGCCGATAAGTCAGTGGTGAAATCTCA
79dcabe7f92f8cf2723b796dcd2f239f	TACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTTGCGTCGTCTGTGAAATTCCG
6edca9464612efff71d8f97299f01663	TACGGAAGGTCCGGGGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCCGGAGATTAAGTGTGTTGTGAAATGTAGA

Practicum: Feature Table Creation (cont.)

```
#FeatureTable and FeatureData summaries
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 Summary View



Overview | Interactive Sample Detail | Feature Detail

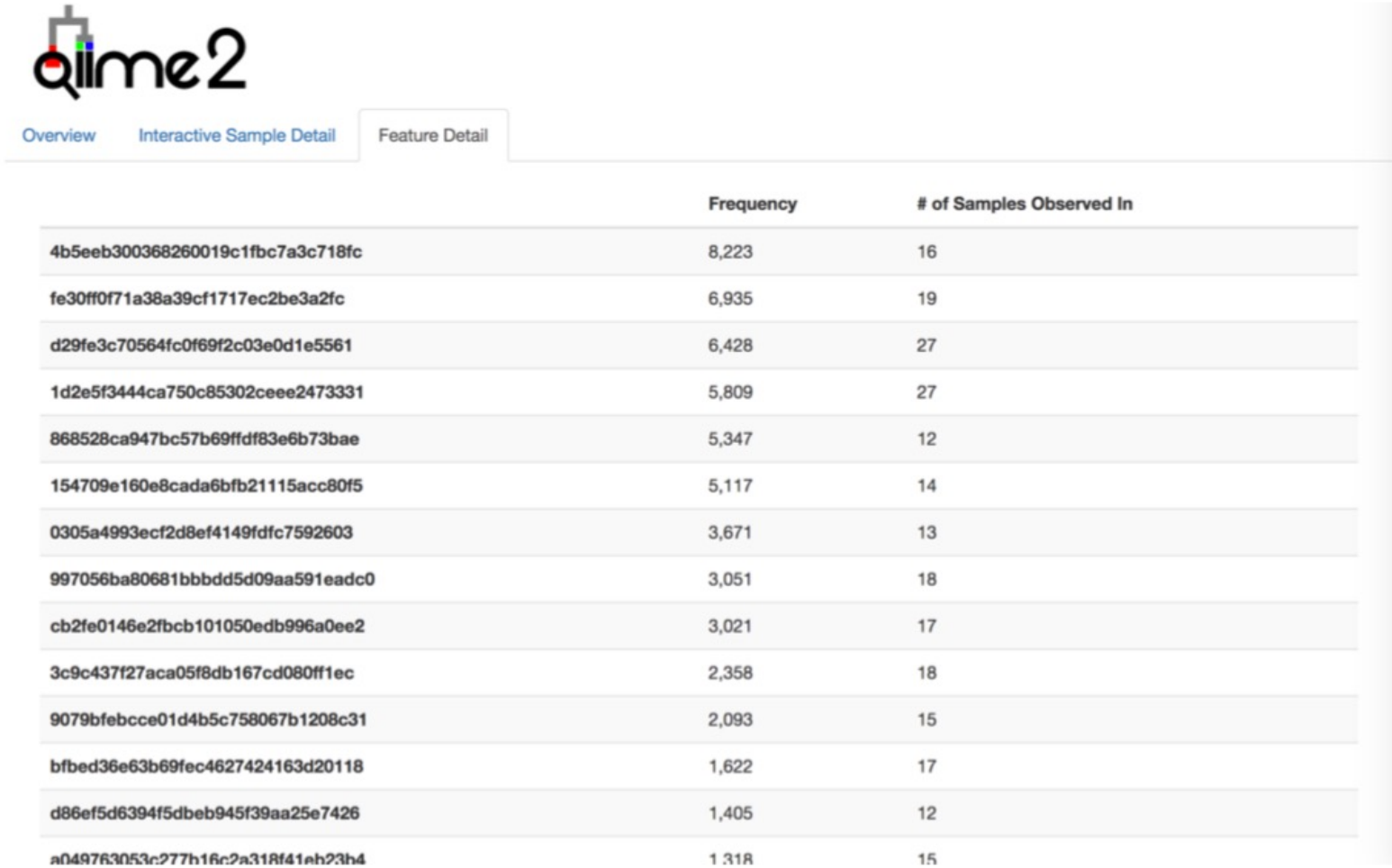
Table summary

Metric	Sample
Number of samples	34
Number of features	485
Total frequency	102,545

Frequency per sample

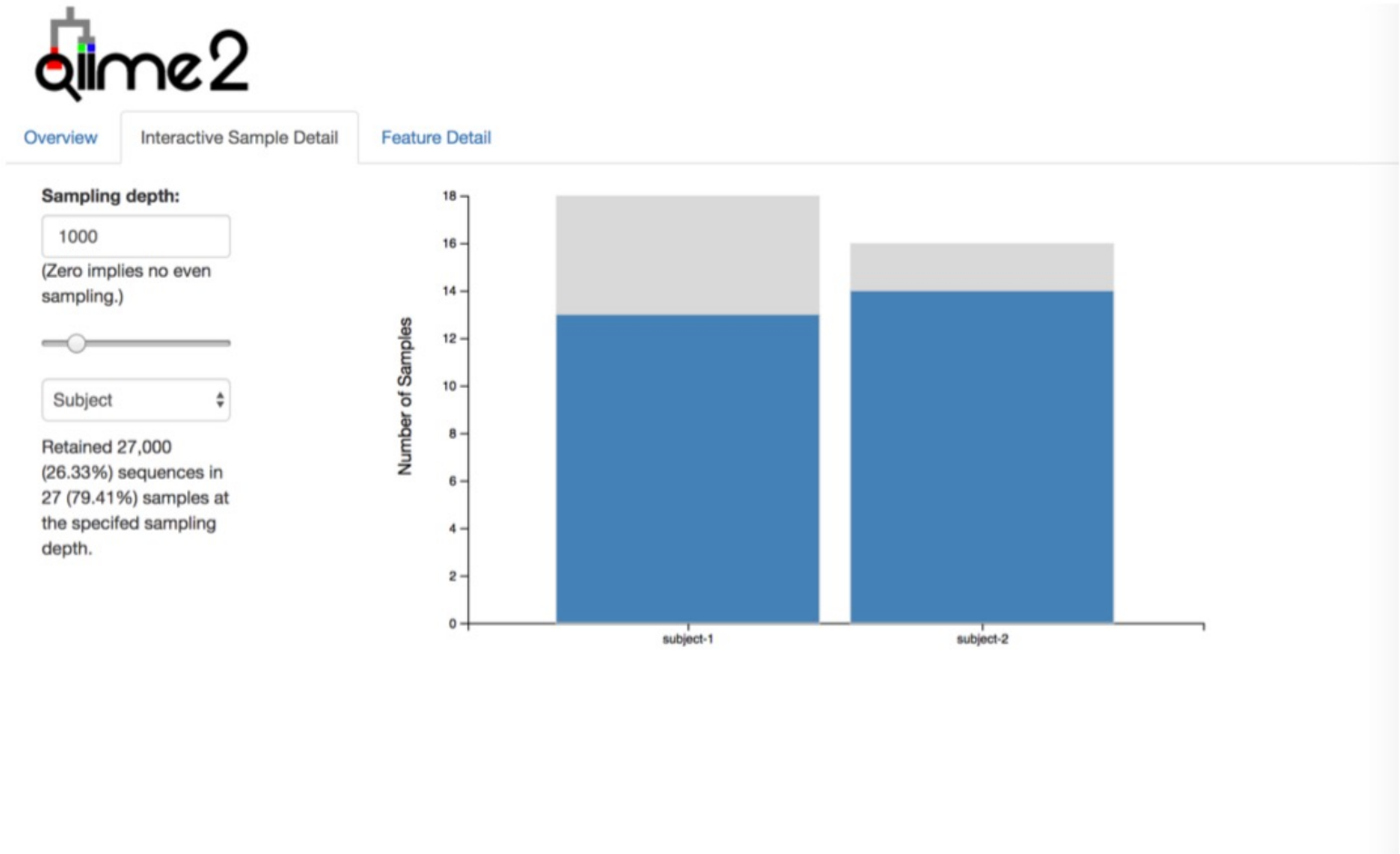
	Frequency
Minimum frequency	512.0
1st quartile	1,367.5
Median frequency	2,581.0
3rd quartile	4,952.0
Maximum frequency	6,770.0
Mean frequency	3,016.029411764706

Feature Table Summary View



	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
0305a4993ecf2d8ef4149fdfc7592603	3,671	13
997056ba80681bbbdd5d09aa591eadc0	3,051	18
cb2fe0146e2fbcb101050edb996a0ee2	3,021	17
3c9c437f27aca05f8db167cd080ff1ec	2,358	18
9079bfebcce01d4b5c758067b1208c31	2,093	15
bfbed36e63b69fec4627424163d20118	1,622	17
d86ef5d6394f5dbeb945f39aa25e7426	1,405	12
a049763053c277b16c2a318f41eb23b4	1,318	15

Feature Table Summary View



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 qualitiative 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)

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

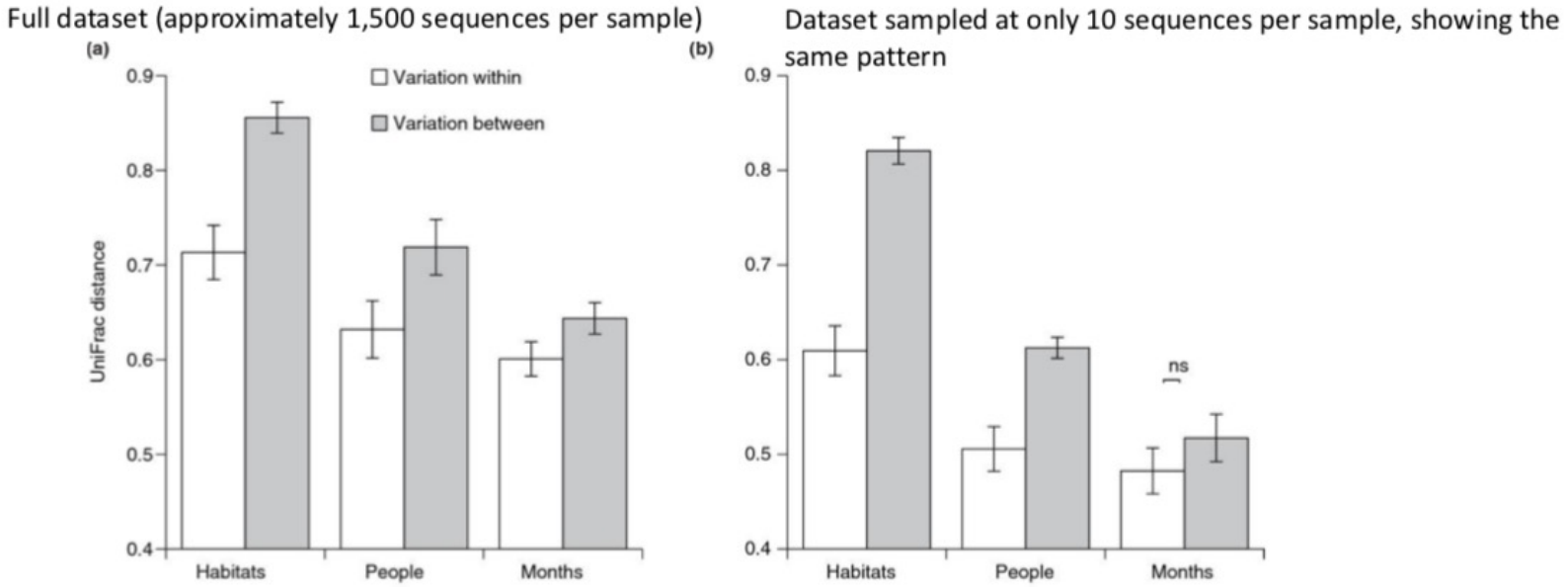


Fig. 2, Kuczynski, J. et al., "Direct sequencing of the human microbiome readily reveals community differences", Genome Biology, 2010

  - 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

Exercise: Core Metrics Sampling Depth

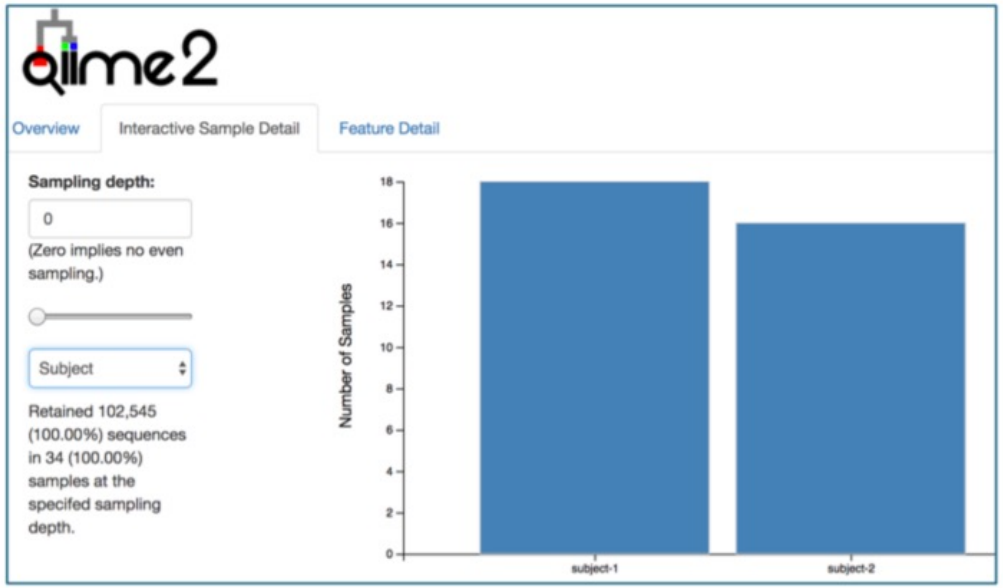
```
qiime diversity core-metrics \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth ??? \
  --output-dir metrics
```

- Note that the core metrics command requires a sampling depth

Exercise: Core Metrics Sampling Depth

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

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



  - 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?

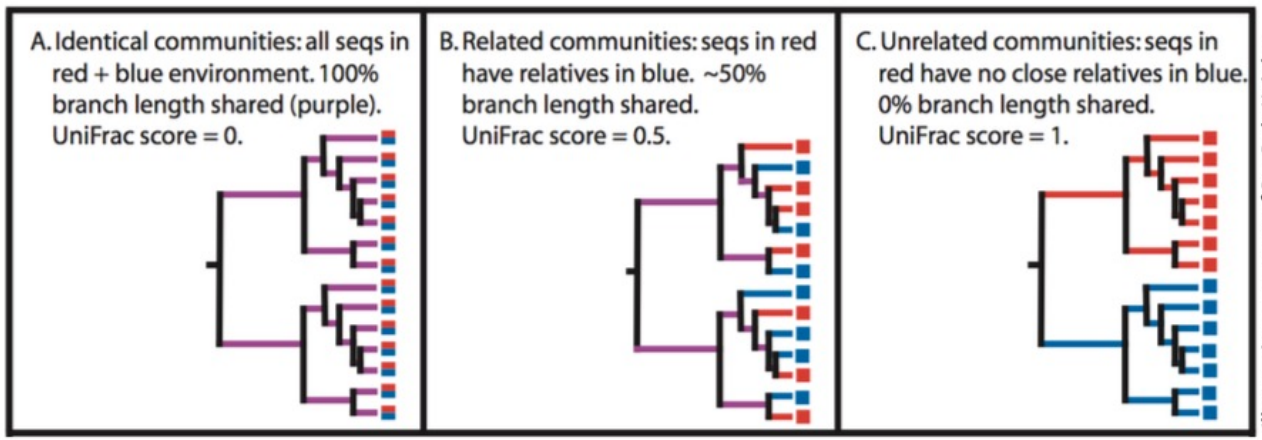
Answers: Core Metrics

```
qiime diversity core-metrics \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth 1103  \
  --output-dir metrics
```

- My answers:
  - Why did you choose this value?
    - Anything higher excludes >= half of right palm samples
  - How many samples will be excluded from your analysis based on this choice?
    - 3
  - How many total sequences will you be analyzing in the core-metrics command?
    - 34,193 (22.22%)

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



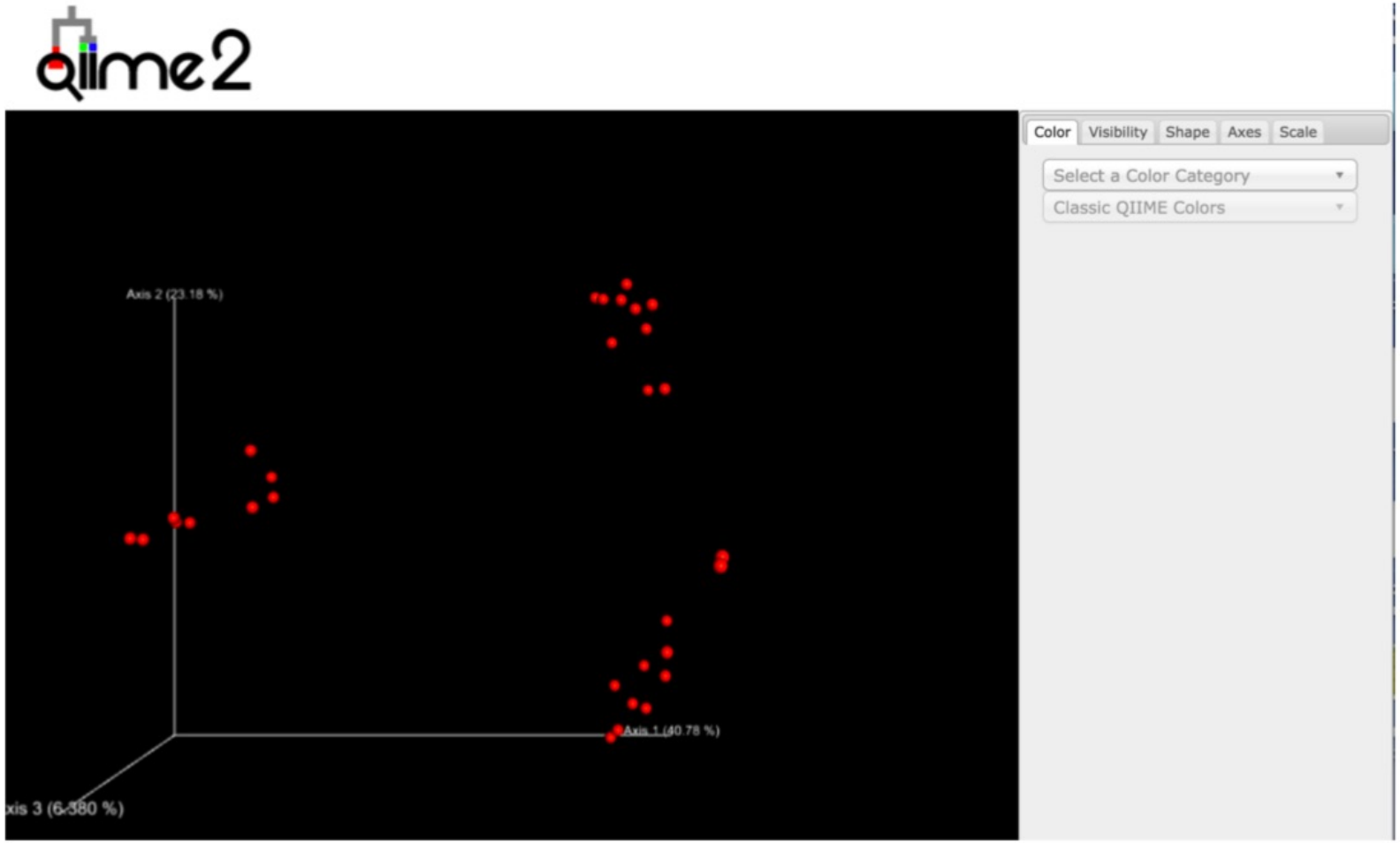
  - Weighted UniFrac: takes abundance each sequence into account

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
```

- This is only showing the PCoA visualization of ONE beta diversity metric
  - Not necessarily “the correct one”!
  - Remember that 3 others are calculated by `core-metrics` alone

Beta Diversity Ordination View

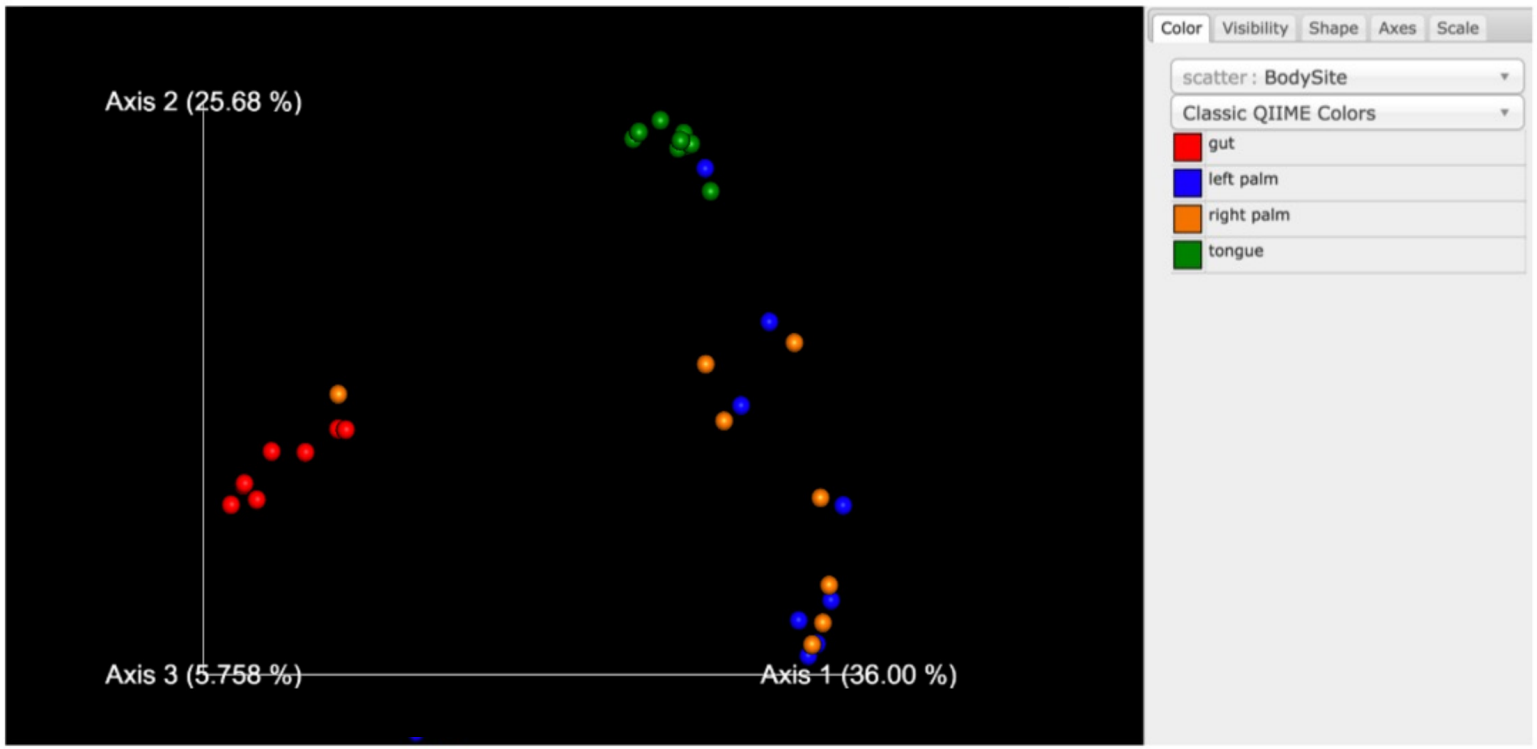


Exercise: Beta Diversity Ordination

- Initial PCoA view (see previous slide) is completely **independent** of metadata
  - Clusters/gradients/etc seen in PCoA are produced by unsupervised learning, based on the feature table information without any awareness of metadata
- It's great to see clear, distinct clusters as in this dataset–but even greater if they can be explained by a known metadata category

Answers: Beta Diversity Ordination

- My answer:
  - Can you find a metadata category that appears associated with the observed clusters?
    - Yep: BodySite. Note left and right palm aren't distinct from each other, unsurprisingly



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

Common Issues in Marker Gene Studies

- Neglecting metadata
  - Analysis can not test for effects of, or discard bias from, features 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

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 ☺

Common Issues in Marker Gene Studies

- Neglecting metadata
  - Analysis can not test for effects of, or discard bias from, features 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 OTUs/features only specified to family or genus level

Taxonomy: Expectation Vs Reality

	Ideal Result	Real Result
Kingdom	Bacteria	Bacteria
Phylum	Proteobacteria	Proteobacteria
Class	Gammaproteobacteria	Gammaproteobacteria
Order	Enterobacteriales	Enterobacteriales
Family	Enterobacteriaceae	Enterobacteriaceae
Genus	*Eschericia*	---
Species	*coli*	OTU 2445338
Strain	O157:H7	--

Practicum: Taxonomic Assignment

```
qiime metadata tabulate \
  --m-input-file taxonomy.qza \
  --o-visualization taxonomy.qzv
```

Taxonomic Assignment Tabulation View

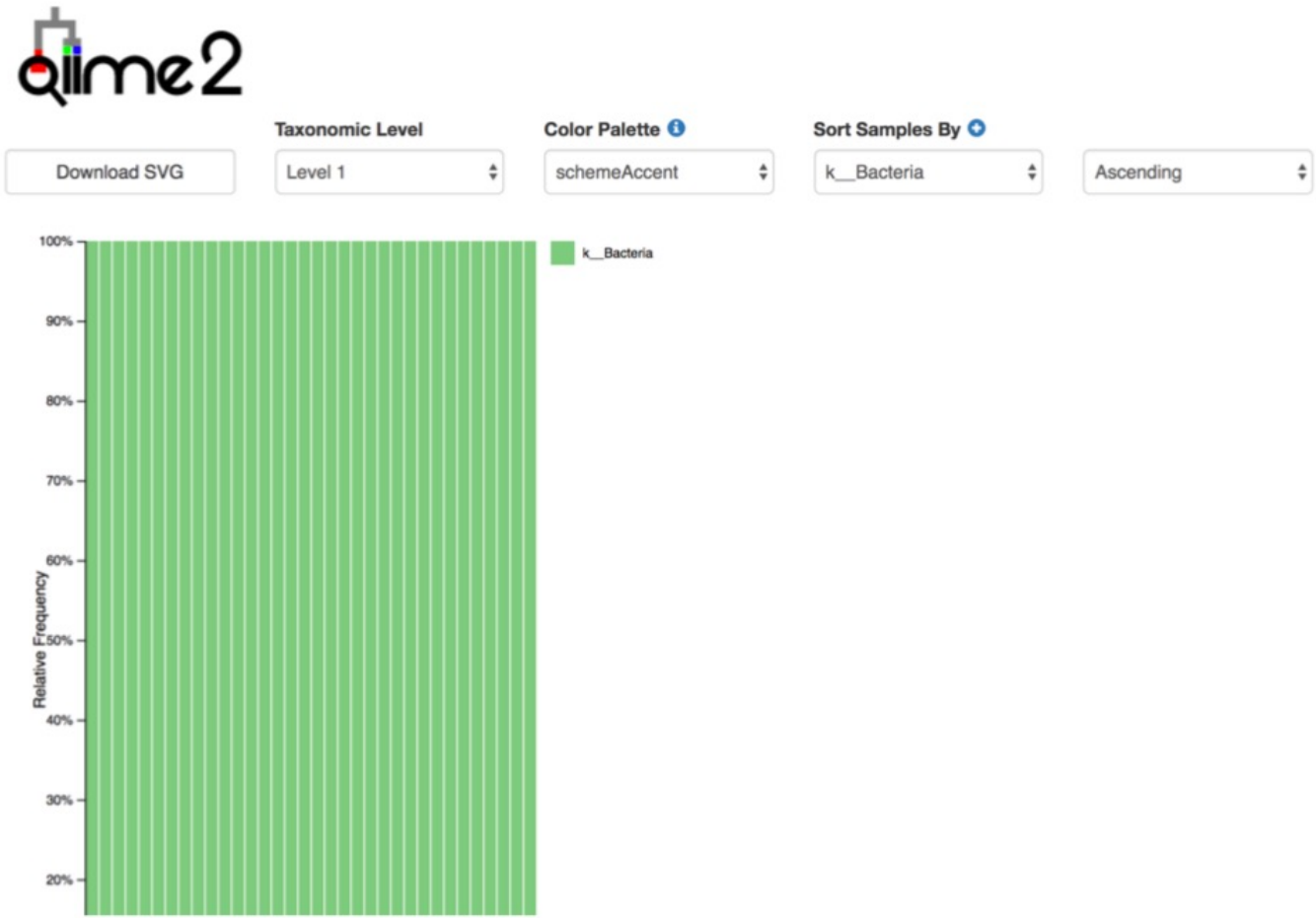
Feature ID	Taxonomy
3677e15d86603bf0a6bb50f8b010afe7	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
1b75626f6834620dc2c729a1a81f497a	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Acinetobacter; s__
42872dc875fef6070dfa78984184c096	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__[Mogibacteriaceae]; g__; s__
51ddb685cfb1775931489ebbd3eef6ca	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Paludibacter; s__
6be678de197b54f9a04f6c984b91ef22	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae
54b4964000ad1631e547c46a828ed1a0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
ecbf086d6ccbe5e8c2a69d0afb144662	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae; g__; s__
c18826df5af5da174f580164c805a38a	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__anginosus
4132561a08d25757e4bee9f73ec4a70a	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Neisseriales; f__Neisseriaceae
7595e123b71bdae8a8c1c28b7405a5c0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae
4a5387c4bc61f2d8f3d9d2de983ba556	k__Bacteria; p__Bacteroidetes; c__Flavobacteriia; o__Flavobacteriales; f__[Weeksellaceae]; g__Chryseobacterium; s__
79dcabe7f92f8cf2723b796dcd2f239f	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Corynebacteriaceae; g__Corynebacterium; s__
6edca9464612efff71d8f97299f01663	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g__Prevotella; s__melaninogenica
fcd4f95c05b868060121ff709085bf21	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__[Tissierellaceae]; g__Finegoldia; s__
f35ce9c514e1398308f5f84ed50b260f	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__[Paraprevotellaceae]; g__[Prevotella]; s__

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

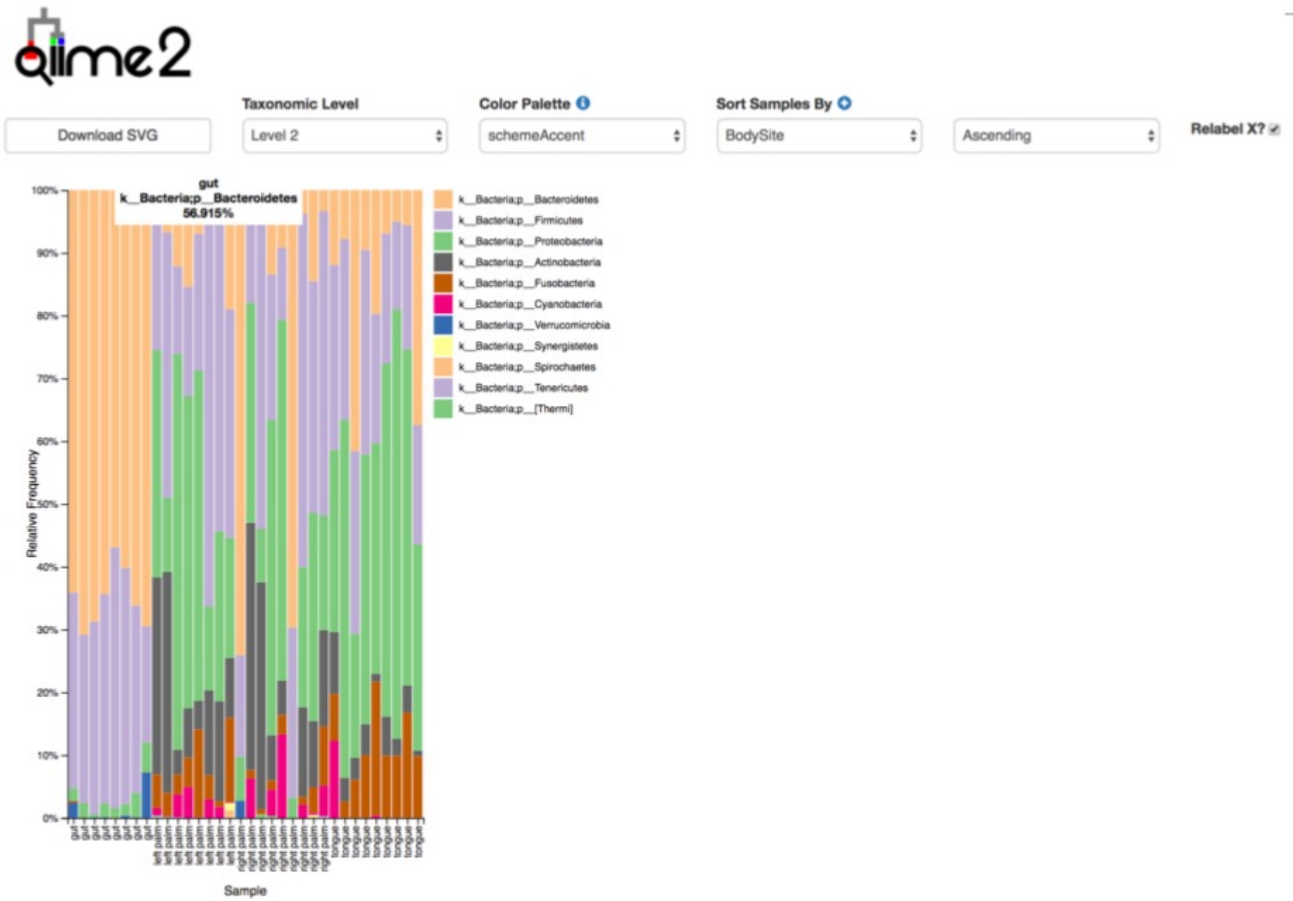
Taxonomic Assignment Bar Plot View



Exercise: Taxonomic Assignment

- “Level 1” = kingdom, “Level 2” = phylum, etc
  - 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

Acknowledgements

- Caporaso lab, Northern Arizona University
- Knight lab, UCSD
- ***QIIME 2 development team!***
  - Especially for the excellent “Moving Pictures” tutorial on which this one is based

- Sophie Weiss,
  Department of Chemical and Biological Engineering, University of Colorado at Boulder