Microbiome 16S Analysis: A Quick-Start Guide

Amanda Birmingham

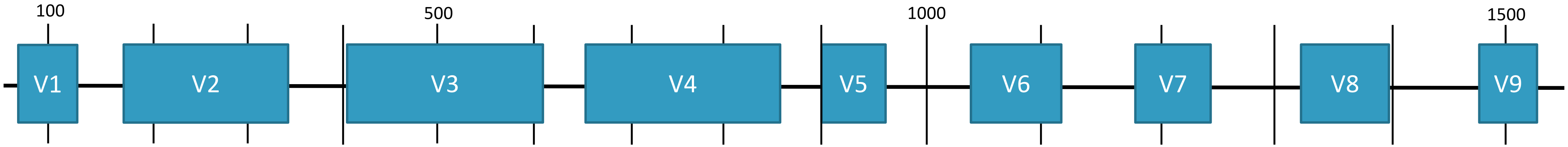
Center for Computational Biology & Bioinformatics
University of California at San Diego

Agenda:

- Rapid introduction to 16S microbiome studies
- Summary of analysis steps and software tools
- Minimal instruction on compute environment
- Practicum on 16S analysis with QIIME 2
  - Alternating lecture and tutorial
    - Goal: Any topic I've lectured about, you will get to test live (even if we don't finish all topics)
- Notice an emphasis on speed ☺
  - Red dot on slide means I won't be covering it in depth
  - Considerable additional material is described in the Supplemental Slides

Marker Gene Metagenomics Basics

- Approach: PCR amplicons of a conserved constitutive gene (a "marker gene") to determine identity and abundance of microbes present
  - Usually the “conserved constitutive gene” of choice is an rRNA
    - For bacteria/archea, usually the 16S—small sub-unit (SSU) of ribosome
      - Excludes eukaryotic DNA as eukaryotes’ SSU is 18S



- 16S rRNA widely conserved across bacteria/archaea (so shared primer sites)
  - But also has 9 hypervariable regions
    - Can be used to id different “species” and build phylogenetic trees
- Can’t study fungi with 16S (they don’t have it) nor 18S (evolves too slowly)
  - Internal transcribed spacer (ITS) is standard fungi marker gene; 28S also used

When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification
- When you're cheap ☺

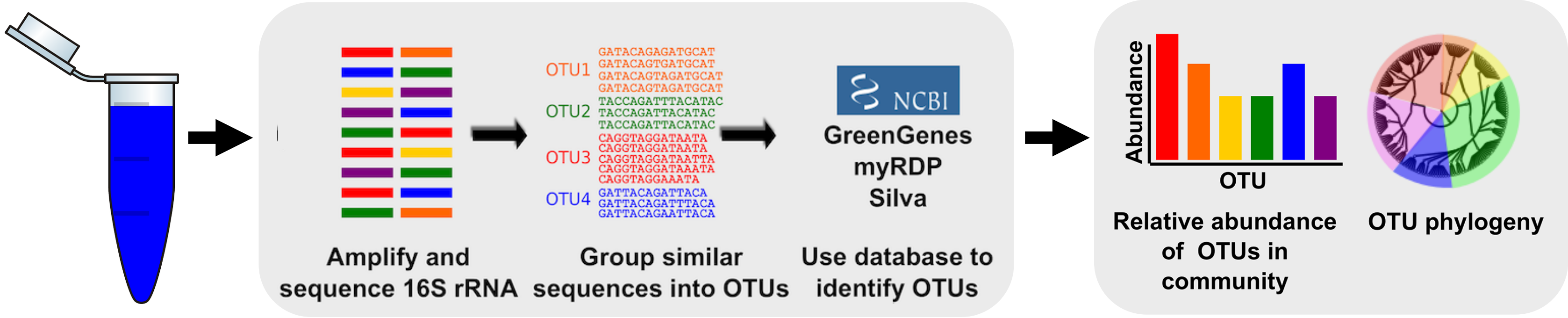


Image modified from Morgan & Huttenhower (2012). PLoS Comput Biol 8(12): e1002808.

● When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification
- When you're cheap ☺
- The good news:
  - Target gene studies are slightly cheaper to prep and sequence than shotgun ones
  - Analysis software is mature, and many studies can be analyzed on a laptop
  - Known taxa can be detected with very low (100s of reads) sequence depth

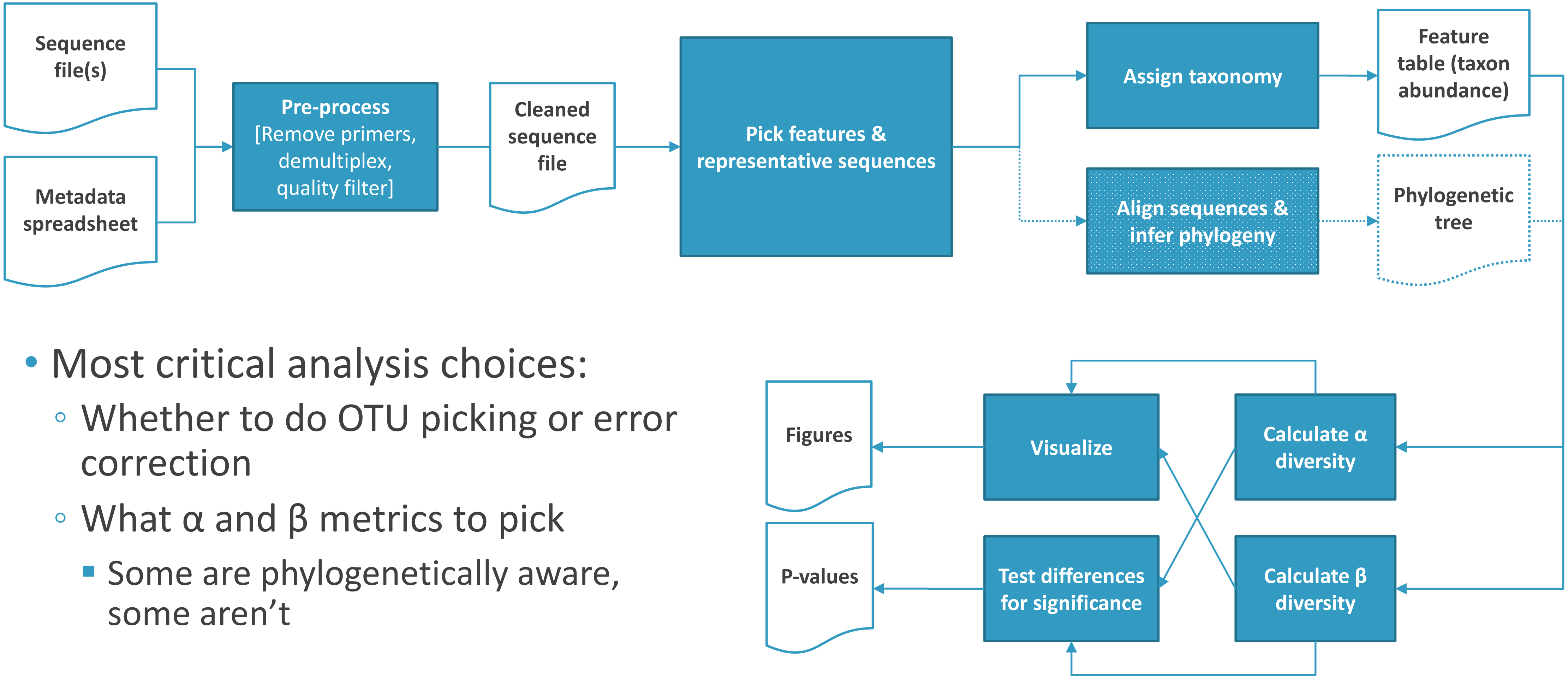
When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification
- When you're cheap ☺
- The good news:
  - Target gene studies are slightly cheaper to prep and sequence than shotgun ones
  - Analysis software is mature, and many studies can be analyzed on a laptop
  - Known taxa can be detected with very low (100s of reads) sequence depth
- The bad news
  - No target gene distinguishes all microbes well
    - And, for a given gene, no primer pair distinguishes all microbes well
  - No other genome information (outside target gene) is captured

● 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

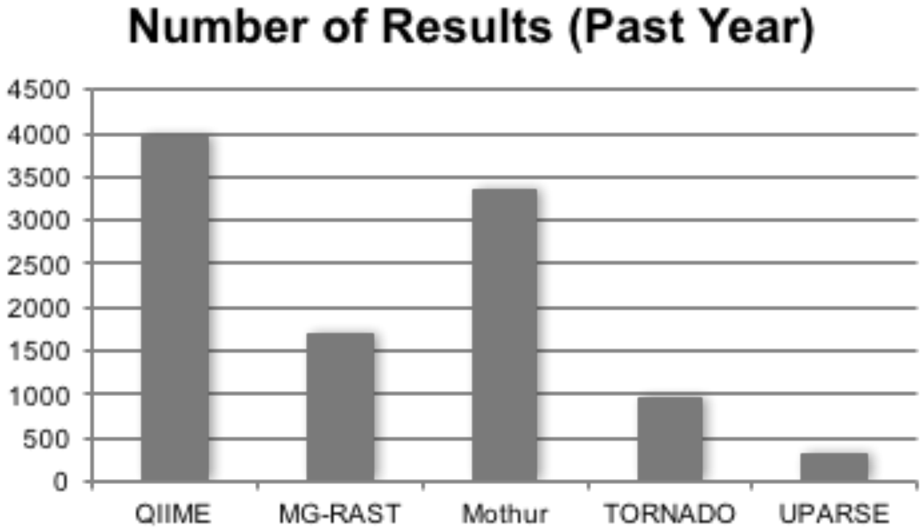
Marker Gene Analysis Workflow


Sequence file(s)
Metadata spreadsheet
Pre-process [Remove primers, demultiplex, quality filter]
Cleaned sequence file
Pick features & representative sequences
Assign taxonomy
Feature table (taxon abundance)
Align sequences & infer phylogeny
Phylogenetic tree
Visualize
Calculate α diversity
Figures
P-values
Test differences for significance
Calculate β diversity

- Most critical analysis choices:
  - Whether to do OTU picking or error correction
  - What α and β metrics to pick
    - Some are phylogenetically aware, some aren't

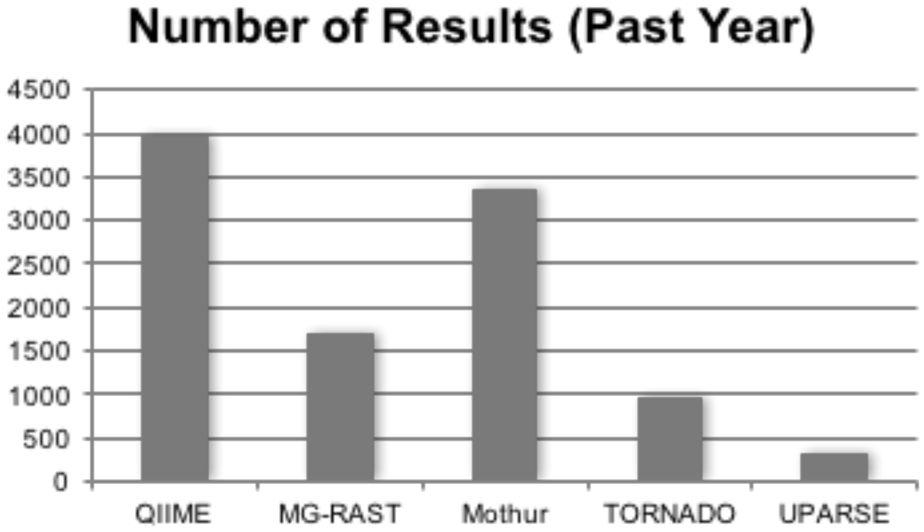
Software Selection

- 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



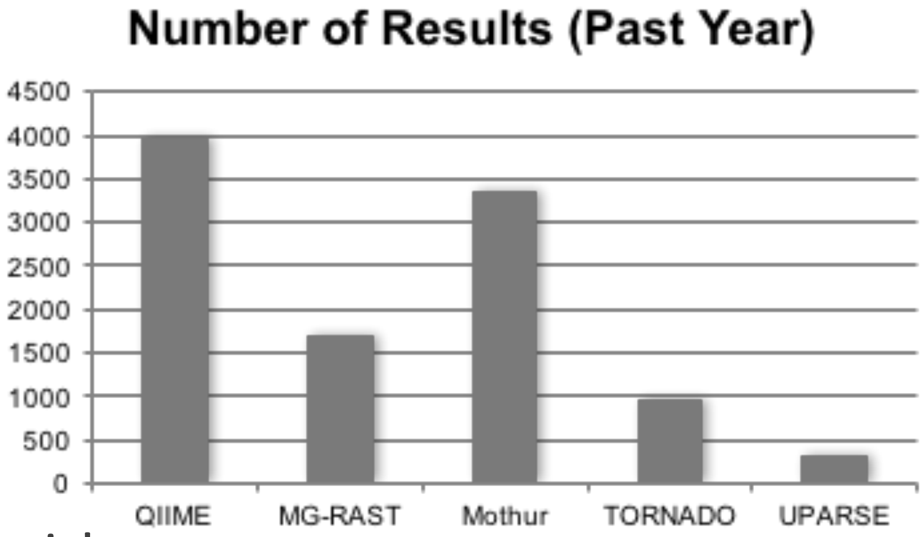
Software Selection

- 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!



Getting the Software & Data

- Not covered in this tutorial, for sake of time
- QIIME 2 is very easy to install with the Conda environment- and package-manager
  - Conda is also very easy to install—either Miniconda or Anaconda versions
  - Once Conda installed, QIIME 2 install is one line, e.g. for linux 64-bit,

```
conda create -n qiime2-2017.6 --file
 https://data.qiime2.org/distro/core/qiime2-2017.6-conda-linux-64.txt
```

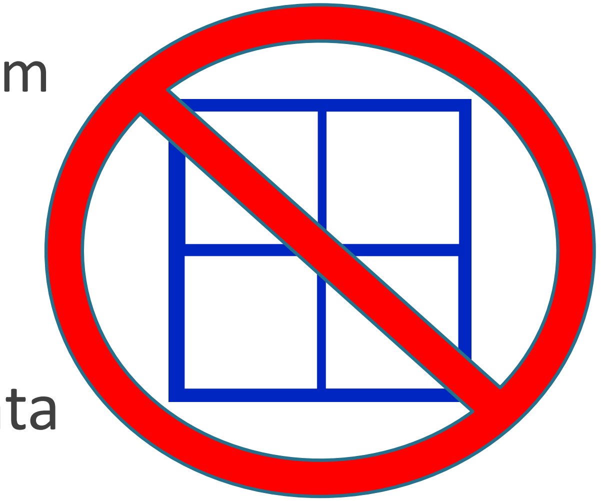
- Data acquisition method is project-specific
  - Public data can often be pulled down from internet with `wget` or `curl` commands
  - Sequencing data from a core usually available by ftp
  - If all else fails, use a flash drive ☺

Get Ready To Practice!

- **“Why are you making me type?!”**
  - QIIME 2 has a GUI—but still very under development
  - QIIME 2 command-line interface is easy to install and ready to run
  - Emphasize typing rather than copy/pasting commands because in your real analyses, you will need to type in the appropriate commands for your data
    - Need to make realistic typing mistakes now so you know how to correct them later!

Get Ready To Practice!

- **“Why are you making me type?!”**
  - QIIME 2 has a GUI—but still very under development
  - QIIME 2 command-line interface is easy to install and ready to run
  - Typing rather than copy/pasting commands because in your real analyses, you will need to type in the appropriate commands for your data
    - Need to make realistic typing mistakes now so you know how to correct them later!
- Will be working in shell on the Ubuntu linux operating system
  - Terminal on Mac OSX is very similar, windows ISN’T
    - Most bioinformatics software doesn’t support Windows
    - Use virtual machine or Cygwin



- Note that you will be training on unusually tractable data
  - Beautiful, clear clustering, significant p-values, etc.
  - If your own data don’t give such clear results, that doesn't mean the analysis is wrong

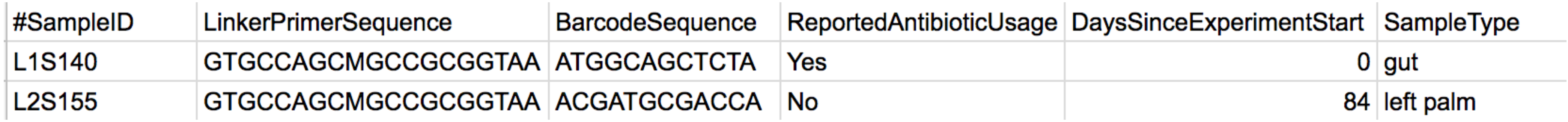
Tips to Help

- When typing a file name or directory path, you can use **tab completion**
  - Start typing file/directory path, then hit tab—if only one file/directory matches what you already typed, shell fills that in
    - Very helpful for correctly entering long file names
    - If >1 matches, shell fills in as much as it can
- Press **up arrow** to get back previous commands you typed
- If you type a command, press enter, and “nothing happens”, **don’t just run it again**
  - Many unix commands produce no visible output to shell—just get back command prompt
  - That doesn’t mean they do nothing, so running them *again* can screw up results
  - **Do not store commands in a word processing program** (or PowerPoint, etc)
    - E.g., MS Word changes hyphens to “m dash”—which command line can’t understand
  - Shell commands are **case-sensitive**

Pre-Acknowledgment

- Please join me in thanking **Pedro Fernandes**
- Without his impeccably managed training computers, resources, and room, these tutorials would not be possible!

Making a Mapping File

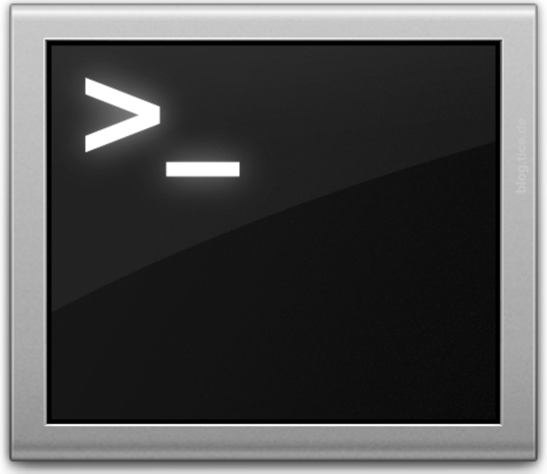


#SampleID	LinkerPrimerSequence	BarcodeSequence	ReportedAntibioticUsage	DaysSinceExperimentStart	SampleType
L1S140	GTGCCAGCMGCCGCGGTAA	ATGGCAGCTCTA	Yes	0	gut
L2S155	GTGCCAGCMGCCGCGGTAA	ACGATGCGACCA	No	84	left palm

- “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

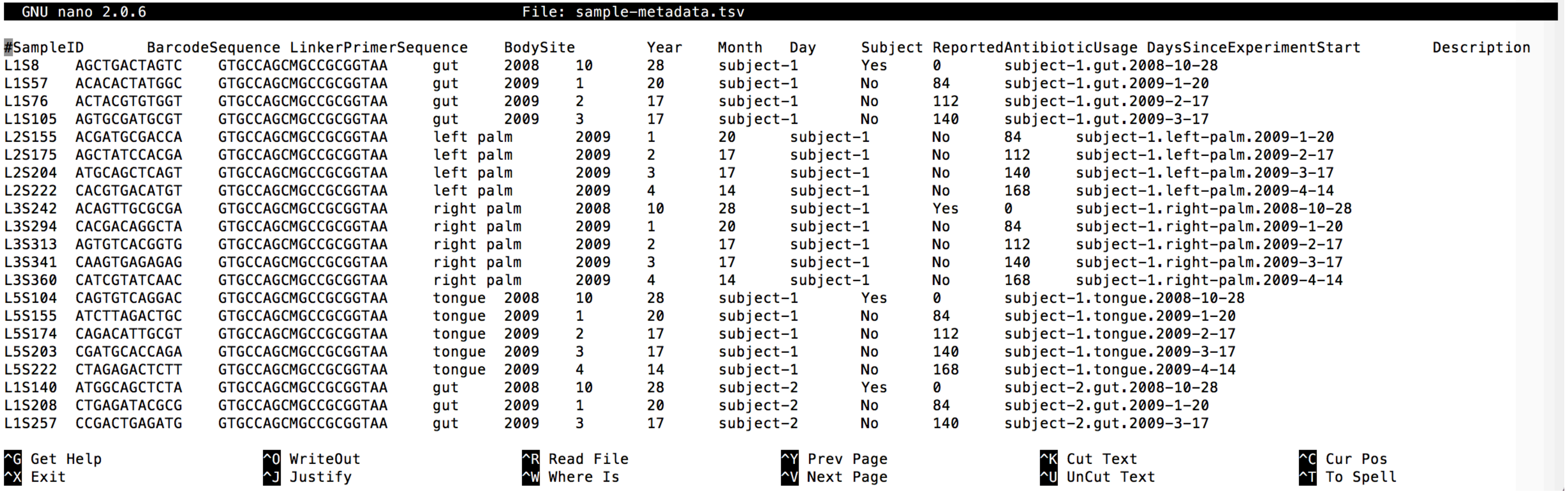
- Open Terminal
  - For below, remember to try tab completion!



```
source activate qiime2-2017.6

cd tutorial-qiime2

ls

nano sample-metadata.tsv
```

- Turn on your green light when the mapping file opens for you

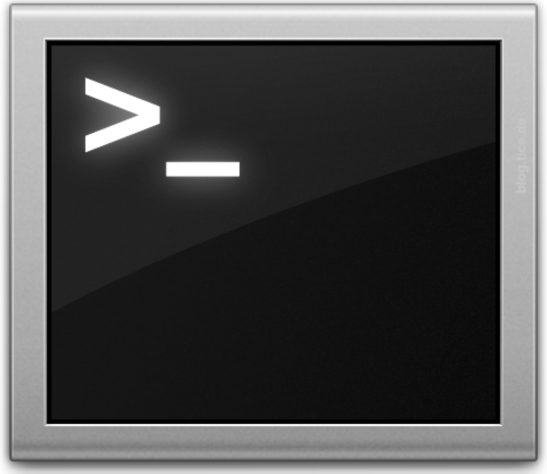
Mapping File View



```
  GNU nano 2.0.6                                         File: sample-metadata.tsv

#SampleID       BarcodeSequence LinkerPrimerSequence    BodySite        Year    Month   Day     Subject ReportedAntibioticUsage DaysSinceExperimentStart        Description
L1S8    AGCTGACTAGTC    GTGCCAGCMGCCGCGGTAA     gut     2008    10      28      subject-1       Yes     0       subject-1.gut.2008-10-28
L1S57   ACACACTATGGC    GTGCCAGCMGCCGCGGTAA     gut     2009    1       20      subject-1       No      84      subject-1.gut.2009-1-20
L1S76   ACTACGTGTGGT    GTGCCAGCMGCCGCGGTAA     gut     2009    2       17      subject-1       No      112     subject-1.gut.2009-2-17
L1S105  AGTGCGATGCGT    GTGCCAGCMGCCGCGGTAA     gut     2009    3       17      subject-1       No      140     subject-1.gut.2009-3-17
L2S155  ACGATGCGACCA    GTGCCAGCMGCCGCGGTAA     left palm       2009    1       20      subject-1       No      84      subject-1.left-palm.2009-1-20
L2S175  AGCTATCCACGA    GTGCCAGCMGCCGCGGTAA     left palm       2009    2       17      subject-1       No      112     subject-1.left-palm.2009-2-17
L2S204  ATGCAGCTCAGT    GTGCCAGCMGCCGCGGTAA     left palm       2009    3       17      subject-1       No      140     subject-1.left-palm.2009-3-17
L2S222  CACGTGACATGT    GTGCCAGCMGCCGCGGTAA     left palm       2009    4       14      subject-1       No      168     subject-1.left-palm.2009-4-14
L3S242  ACAGTTGCGCGA    GTGCCAGCMGCCGCGGTAA     right palm      2008    10      28      subject-1       Yes     0       subject-1.right-palm.2008-10-28
L3S294  CACGACAGGCTA    GTGCCAGCMGCCGCGGTAA     right palm      2009    1       20      subject-1       No      84      subject-1.right-palm.2009-1-20
L3S313  AGTGTCACGGTG    GTGCCAGCMGCCGCGGTAA     right palm      2009    2       17      subject-1       No      112     subject-1.right-palm.2009-2-17
L3S341  CAAGTGAGAGAG    GTGCCAGCMGCCGCGGTAA     right palm      2009    3       17      subject-1       No      140     subject-1.right-palm.2009-3-17
L3S360  CATCGTATCAAC    GTGCCAGCMGCCGCGGTAA     right palm      2009    4       14      subject-1       No      168     subject-1.right-palm.2009-4-14
L5S104  CAGTGTCAGGAC    GTGCCAGCMGCCGCGGTAA     tongue  2008    10      28      subject-1       Yes     0       subject-1.tongue.2008-10-28
L5S155  ATCTTAGACTGC    GTGCCAGCMGCCGCGGTAA     tongue  2009    1       20      subject-1       No      84      subject-1.tongue.2009-1-20
L5S174  CAGACATTGCGT    GTGCCAGCMGCCGCGGTAA     tongue  2009    2       17      subject-1       No      112     subject-1.tongue.2009-2-17
L5S203  CGATGCACCAGA    GTGCCAGCMGCCGCGGTAA     tongue  2009    3       17      subject-1       No      140     subject-1.tongue.2009-3-17
L5S222  CTAGAGACTCTT    GTGCCAGCMGCCGCGGTAA     tongue  2009    4       14      subject-1       No      168     subject-1.tongue.2009-4-14
L1S140  ATGGCAGCTCTA    GTGCCAGCMGCCGCGGTAA     gut     2008    10      28      subject-2       Yes     0       subject-2.gut.2008-10-28
L1S208  CTGAGATACGCG    GTGCCAGCMGCCGCGGTAA     gut     2009    1       20      subject-2       No      84      subject-2.gut.2009-1-20
L1S257  CCGACTGAGATG    GTGCCAGCMGCCGCGGTAA     gut     2009    3       17      subject-2       No      140     subject-2.gut.2009-3-17

^G Get Help                    ^O WriteOut                    ^R Read File                   ^Y Prev Page                   ^K Cut Text                    ^C Cur Pos
^X Exit                        ^J Justify                     ^W Where Is                    ^V Next Page                   ^U UnCut Text                  ^T To Spell
```

Practicum: Viewing A Mapping File

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



```
source activate qiime2-2017.6

cd tutorial-qiime2

ls

nano sample-metadata.tsv
```

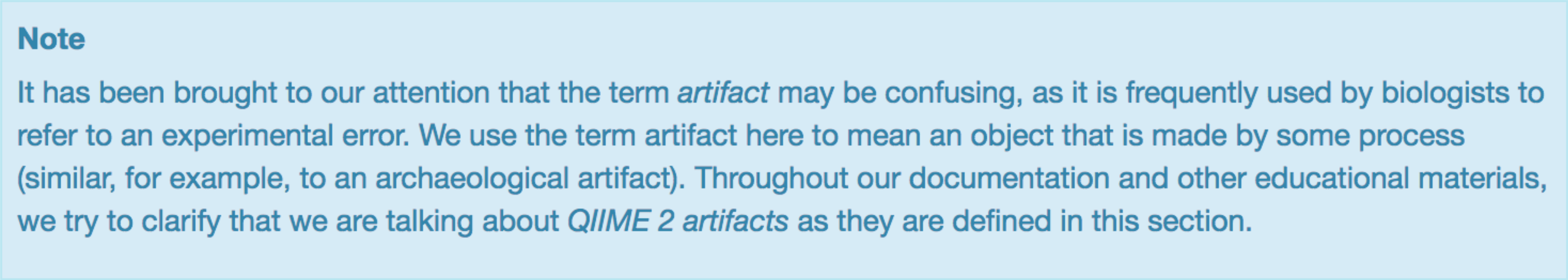
- Stretch the window so you can look at the contents; then, to close, type

```
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)
- 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
```

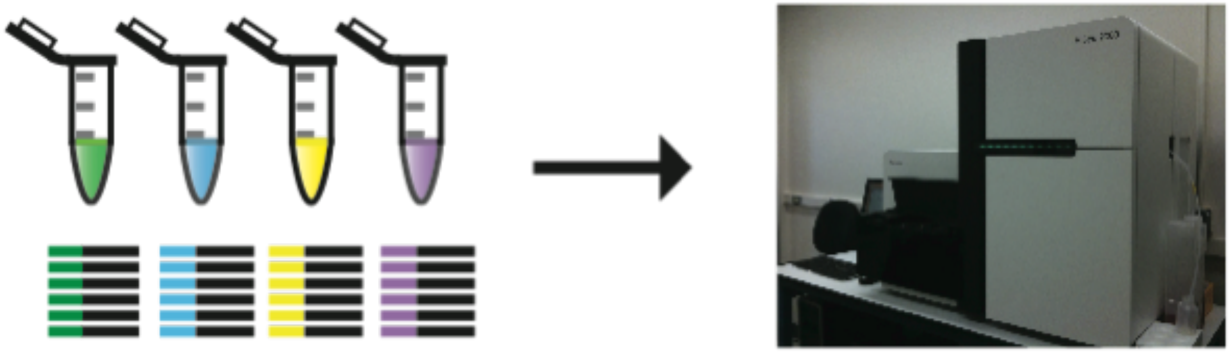
- A backslash \ is used to break up a command onto multiple lines
  - If you prefer to type the whole command onto one run-on line, you can leave it out

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



Multiplex Thousands of Samples with Error-Correcting Barcodes

Pool Samples and Sequence

QIIME 2, https://qiime2.org.

- 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

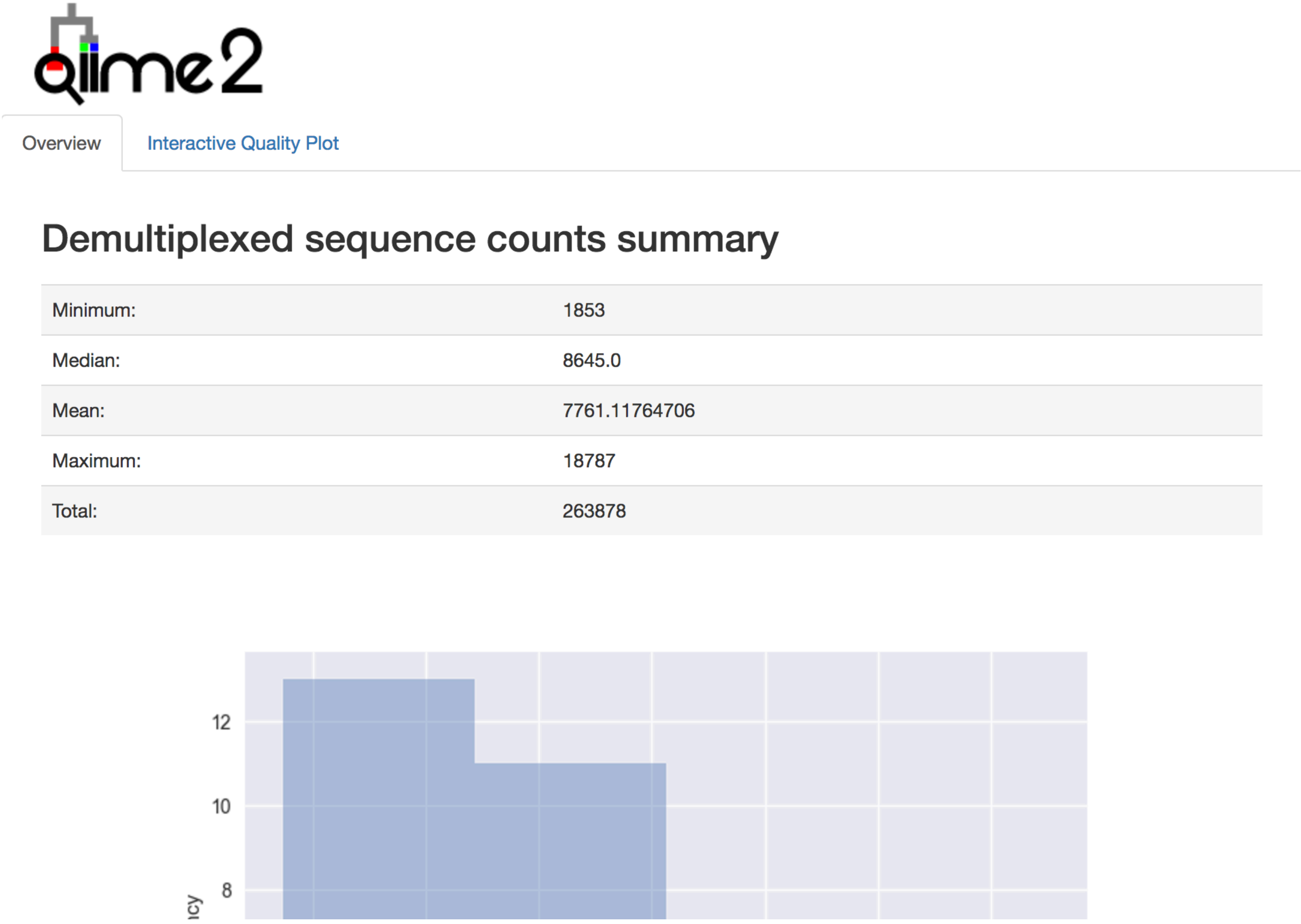
Practicum: Demultiplexing (cont.)

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

- Now view the visualization, locally

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

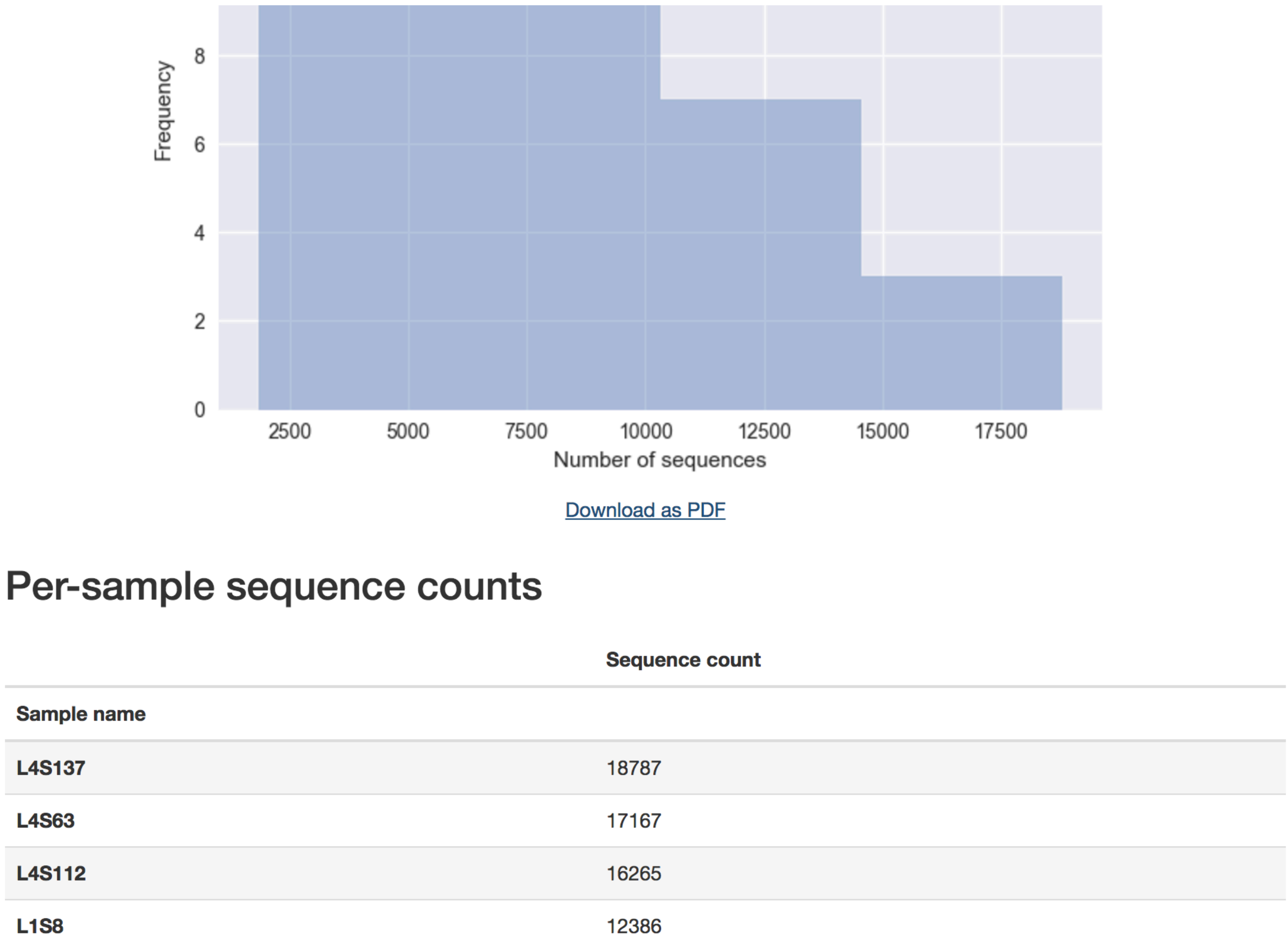
Practicum: Demultiplexing (cont.)



Demultiplexed sequence counts summary

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

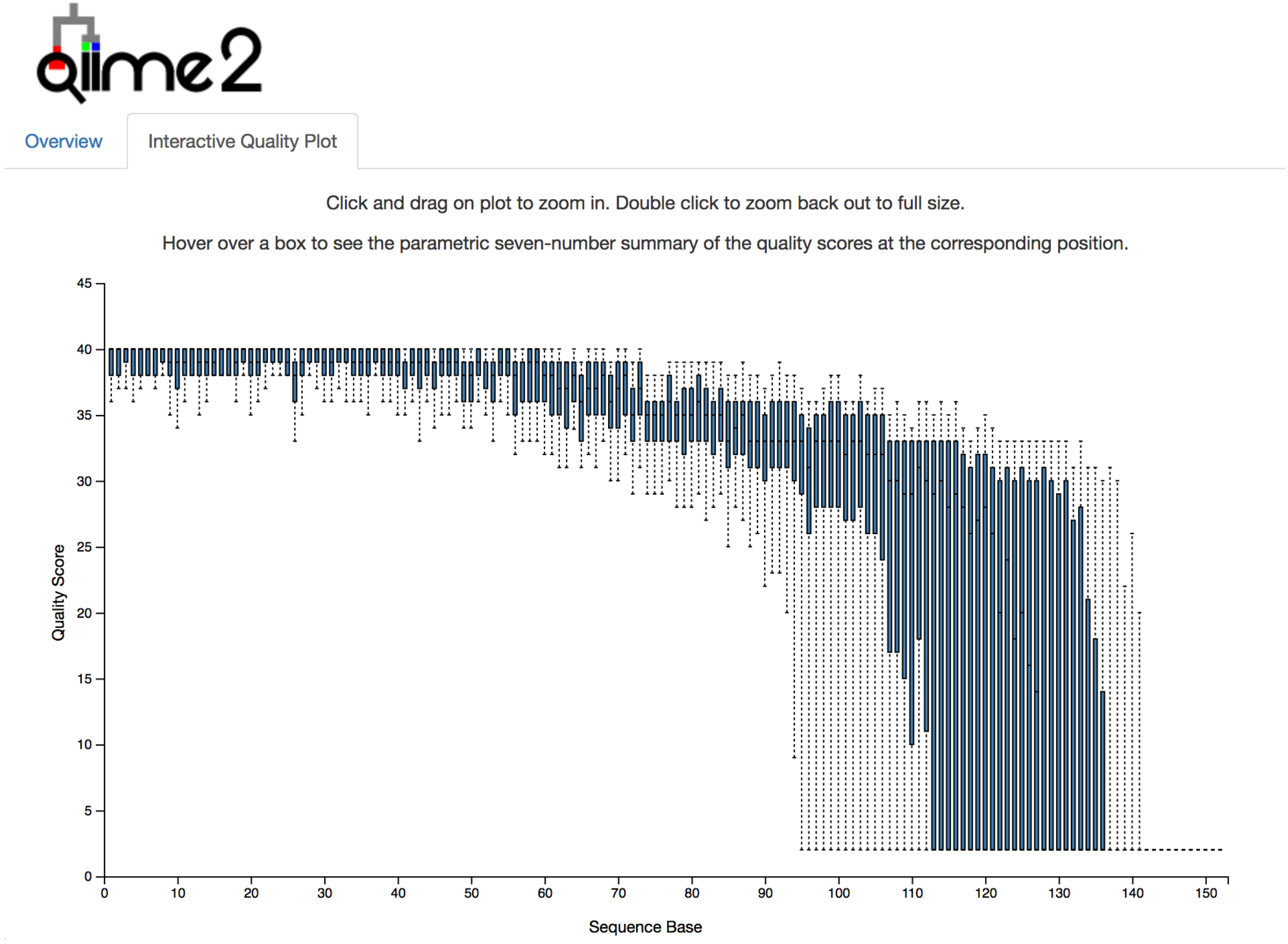
Practicum: Demultiplexing (cont.)



Per-sample sequence counts

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

Practicum: Demultiplexing (cont.)



Click and drag on plot to zoom in. Double click to zoom back out to full size.

Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.

Practicum: Demultiplexing (cont.)

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

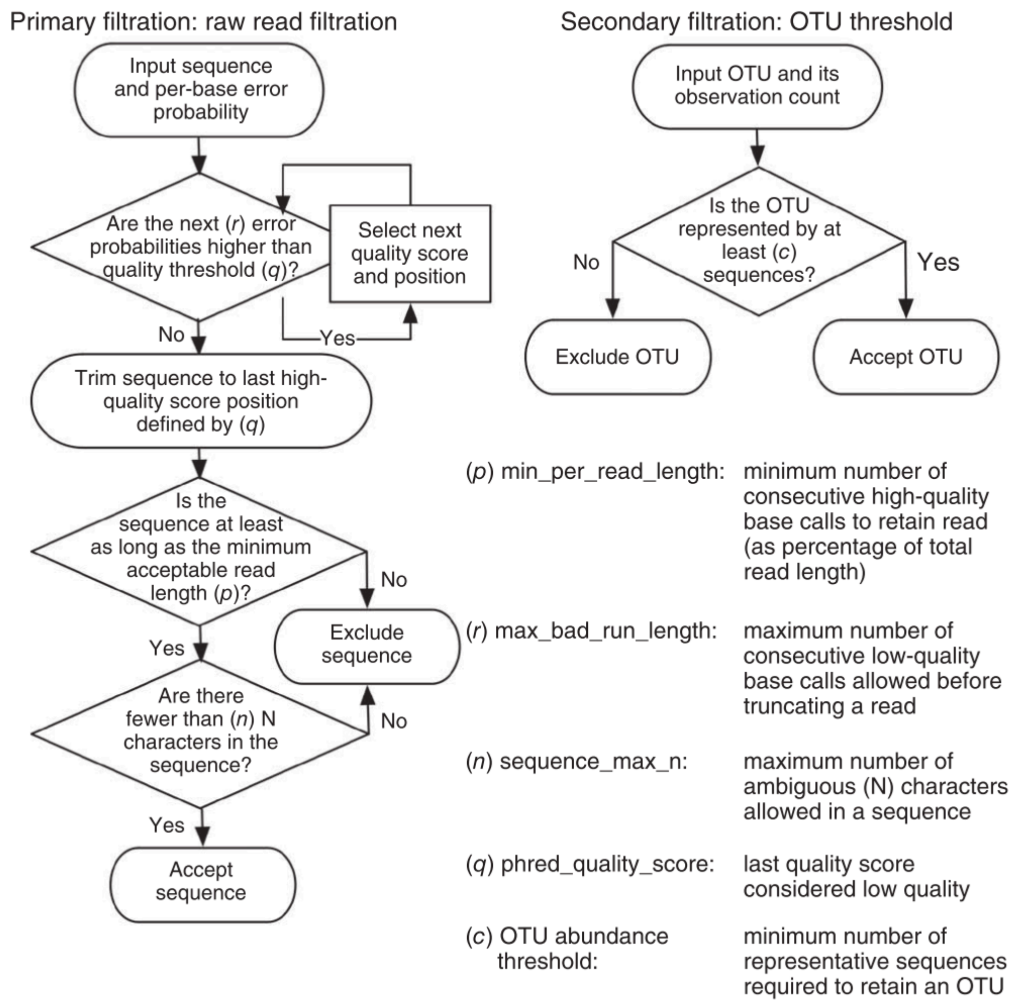
- Now view the visualization, locally

```
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)

Quality Control

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



Primary filtration: raw read filtration
Input sequence and per-base error probability
Are the next (*r*) error probabilities higher than quality threshold (*q*)?
Select next quality score and position
Yes
No
Trim sequence to last high-quality score position defined by (*q*)
Is the sequence at least as long as the minimum acceptable read length (*p*)?
No
Exclude sequence
Yes
Are there fewer than (*n*) N characters in the sequence?
No
Yes
Accept sequence
Secondary filtration: OTU threshold
Input OTU and its observation count
Is the OTU represented by at least (*c*) sequences?
No
Yes
Exclude OTU
Accept OTU

(*p*) min_per_read_length: minimum number of consecutive high-quality base calls to retain read (as percentage of total read length)

(*r*) max_bad_run_length: maximum number of consecutive low-quality base calls allowed before truncating a read

(*n*) sequence_max_n: maximum number of ambiguous (N) characters allowed in a sequence

(*q*) phred_quality_score: last quality score considered low quality

(*c*) OTU abundance threshold: minimum number of representative sequences required to retain an OTU

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
```

- 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 quality-filter q-score
```

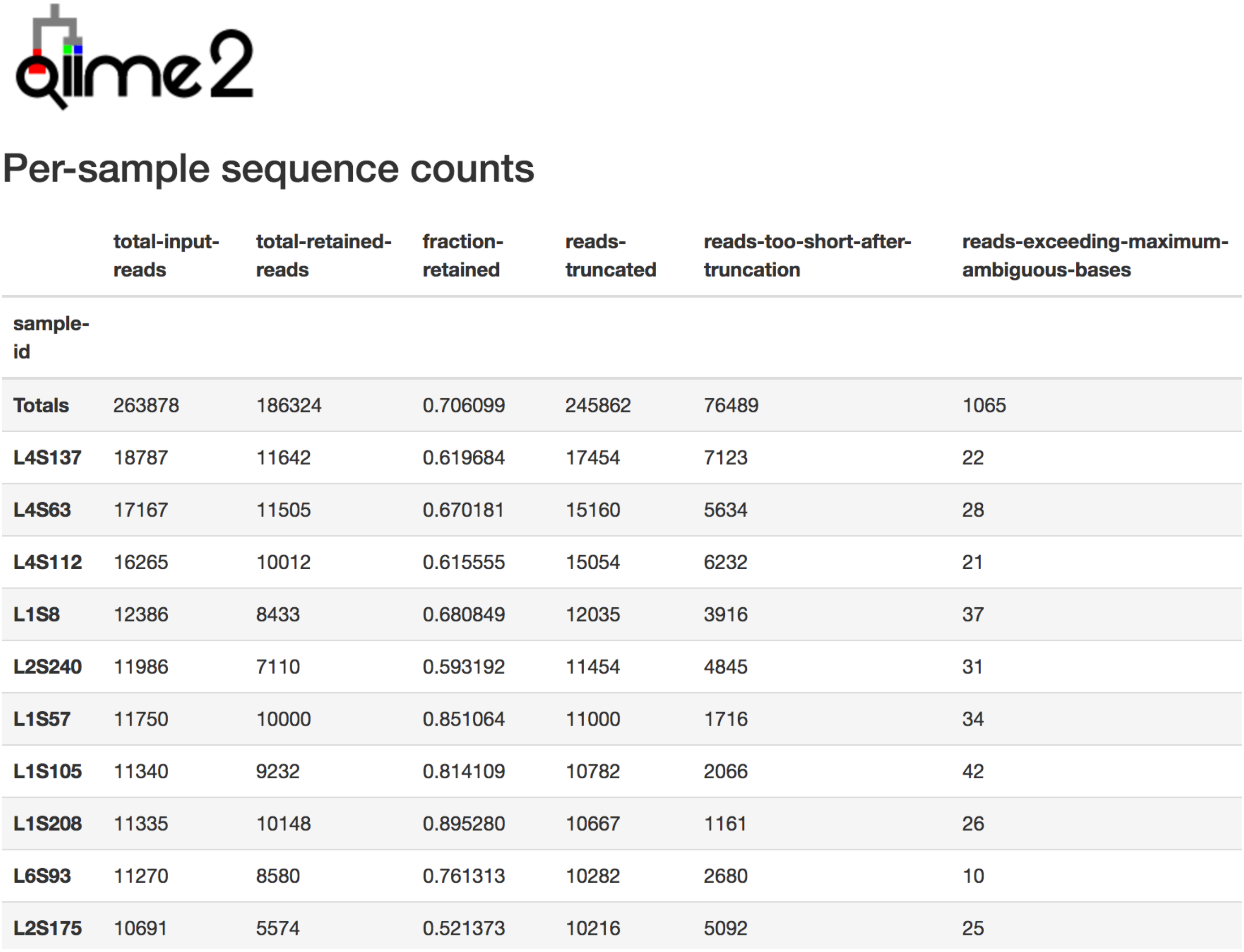
Practicum: Quality Control

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

qiime quality-filter visualize-stats \
 --i-filter-stats demux-filter-stats.qza \
 --o-visualization demux-filter-stats.qzv
```

- Remember: for the remainder of this tutorial, any time you create a visualization file, you will need to run an additional command to view it!

```
qiime tools view yourvisualizationfilename.qzv
```

Practicum: Quality Control



Per-sample sequence counts

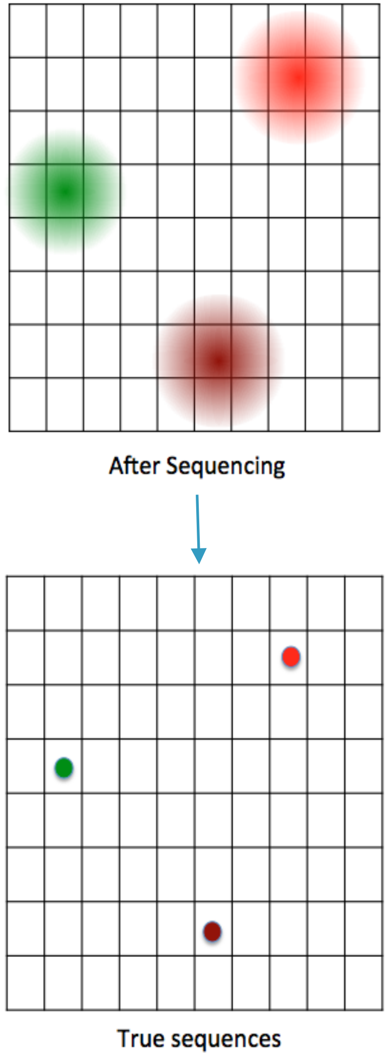
	total-input-reads	total-retained-reads	fraction-retained	reads-truncated	reads-too-short-after-truncation	reads-exceeding-maximum-ambiguous-bases
sample-id						
Totals	263878	186324	0.706099	245862	76489	1065
L4S137	18787	11642	0.619684	17454	7123	22
L4S63	17167	11505	0.670181	15160	5634	28
L4S112	16265	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

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



QIIME 2, https://qiime2.org.

Practicum: Feature Table Creation

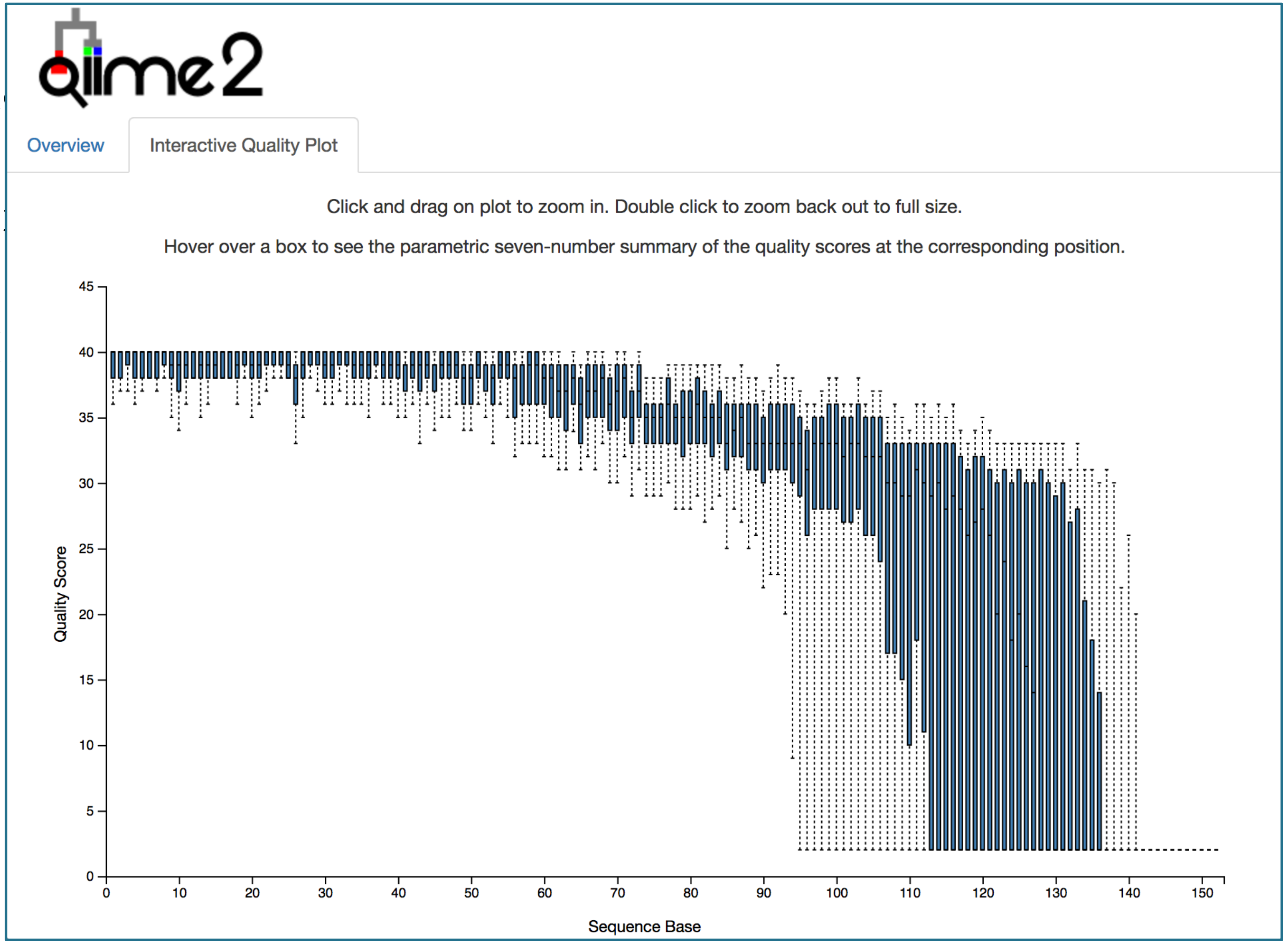
```
qiime deblur denoise-16S \
  --i-demultiplexed-seqs demux-filtered.qza \
  --p-trim-length 120 \
  --o-representative-sequences rep-seqs.qza \
  --o-table table.qza \
  --o-stats deblur-stats.qza
```

- 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 deblur denoise-16S \
  --i-demultiplexed-seqs d
  --p-trim-length 120 \
  --o-representative-seque
  --o-table table.qza \
  --o-stats deblur-stats.q
```

- 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



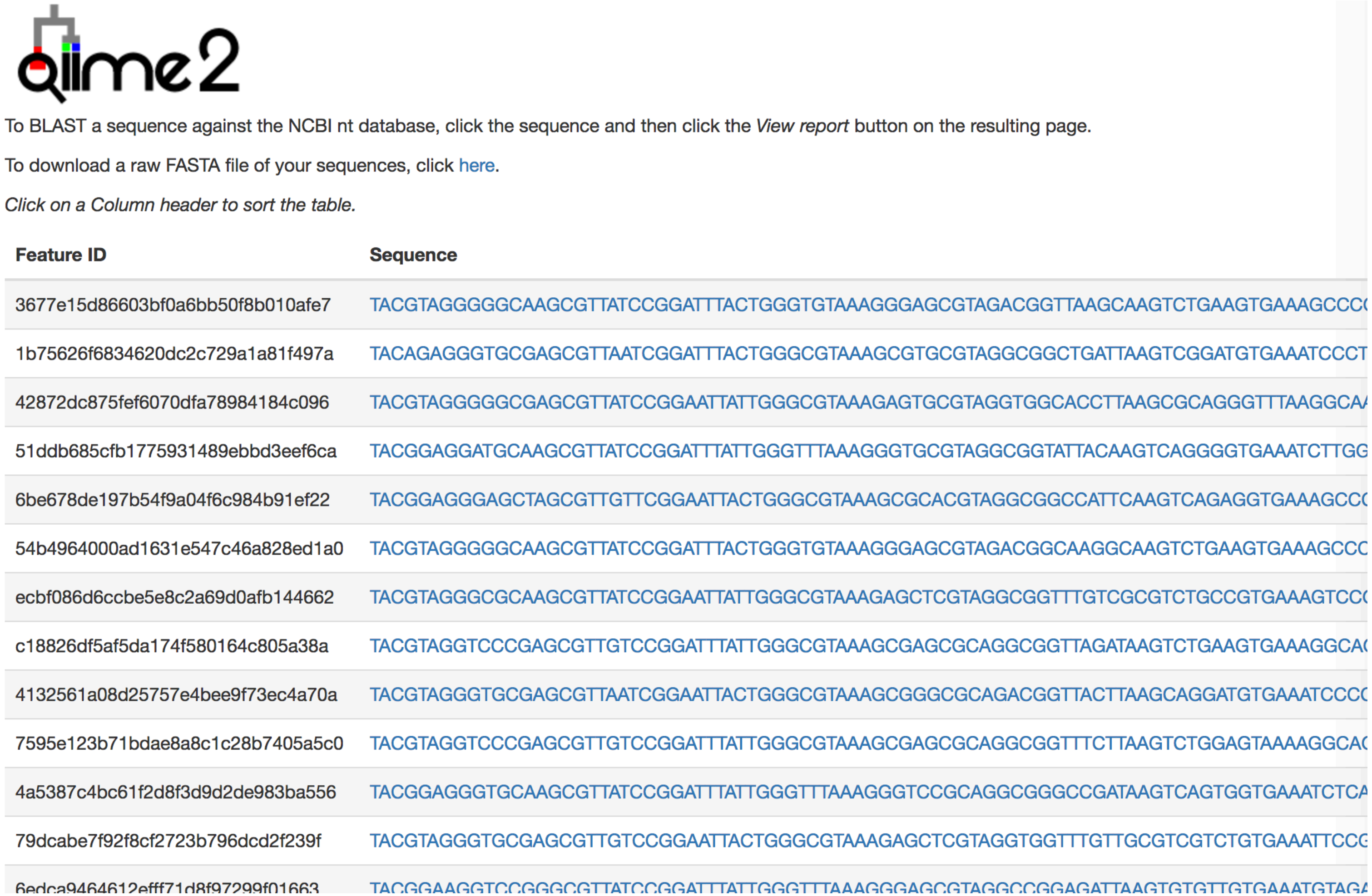
Click and drag on plot to zoom in. Double click to zoom back out to full size.

Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.

Practicum: Feature Table Creation (cont.)

```
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	TACGTAGGGGGCGAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGCACCTTAAGCGCAGGGTTTAAGGCA
51ddb685cfb1775931489ebbd3eef6ca	TACGGAGGATGCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTATTACAAGTCAGGGGTGAAATCTTGG
6be678de197b54f9a04f6c984b91ef22	TACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCCATTCAAGTCAGAGGTGAAAGCCC
54b4964000ad1631e547c46a828ed1a0	TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGCAAGGCAAGTCTGAAGTGAAAGCCC
ecbf086d6ccbe5e8c2a69d0afb144662	TACGTAGGGCGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGCCGTGAAAGTCC
c18826df5af5da174f580164c805a38a	TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTGAAAGGCA
4132561a08d25757e4bee9f73ec4a70a	TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGGGCGCAGACGGTTACTTAAGCAGGATGTGAAATCCC
7595e123b71bdae8a8c1c28b7405a5c0	TACGTAGGTCCCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTTCTTAAGTCTGGAGTAAAAGGCA
4a5387c4bc61f2d8f3d9d2de983ba556	TACGGAGGGTGCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGTCCGCAGGCGGGCCGATAAGTCAGTGGTGAAATCTC
79dcabe7f92f8cf2723b796dcd2f239f	TACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTTGCGTCGTCTGTGAAATTCC
6edca9464612efff71d8f97299f01663	TACGGAAGGTCCGGGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCCGGAGATTAAGTGTGTTGTGAAATGTAG

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

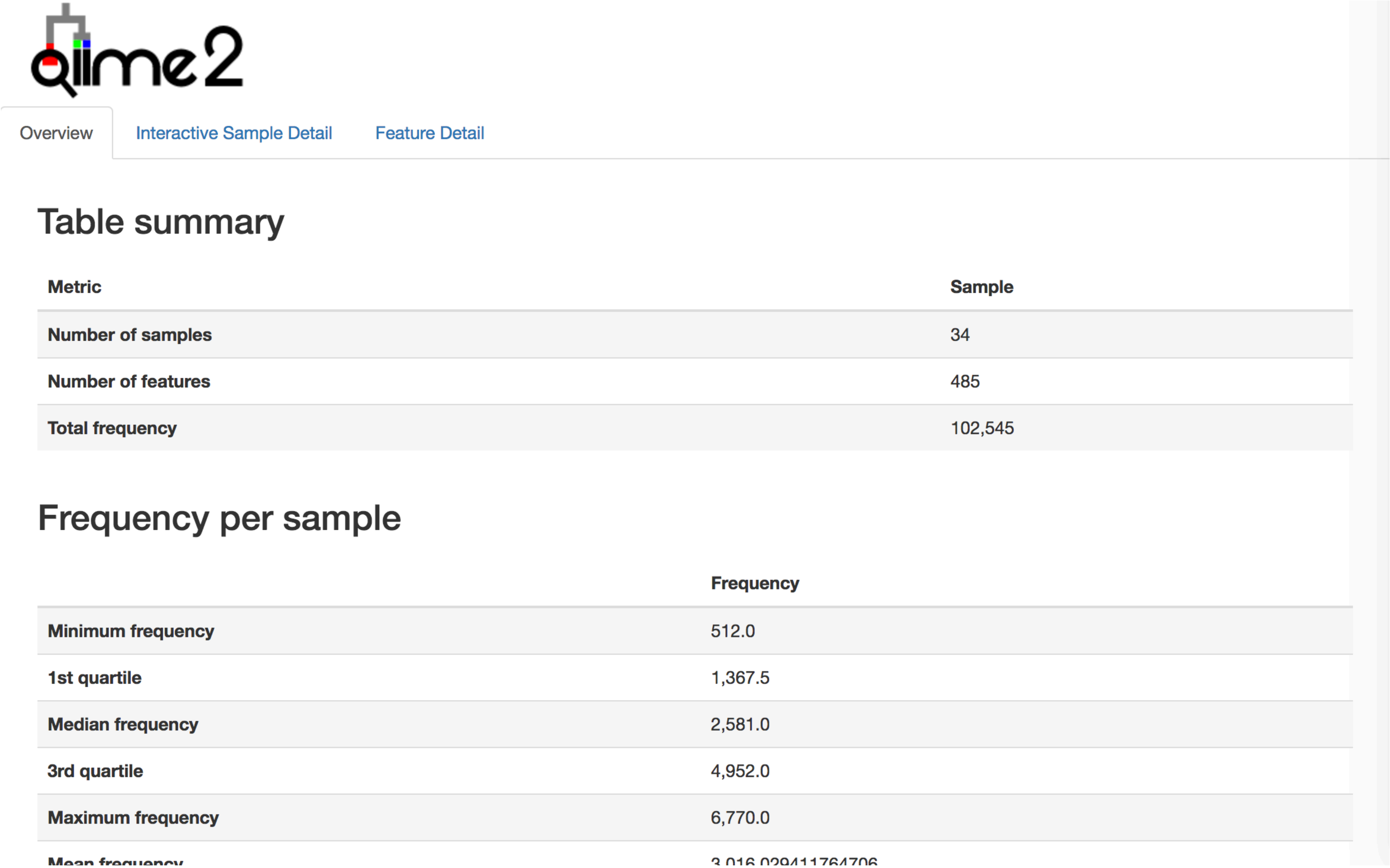


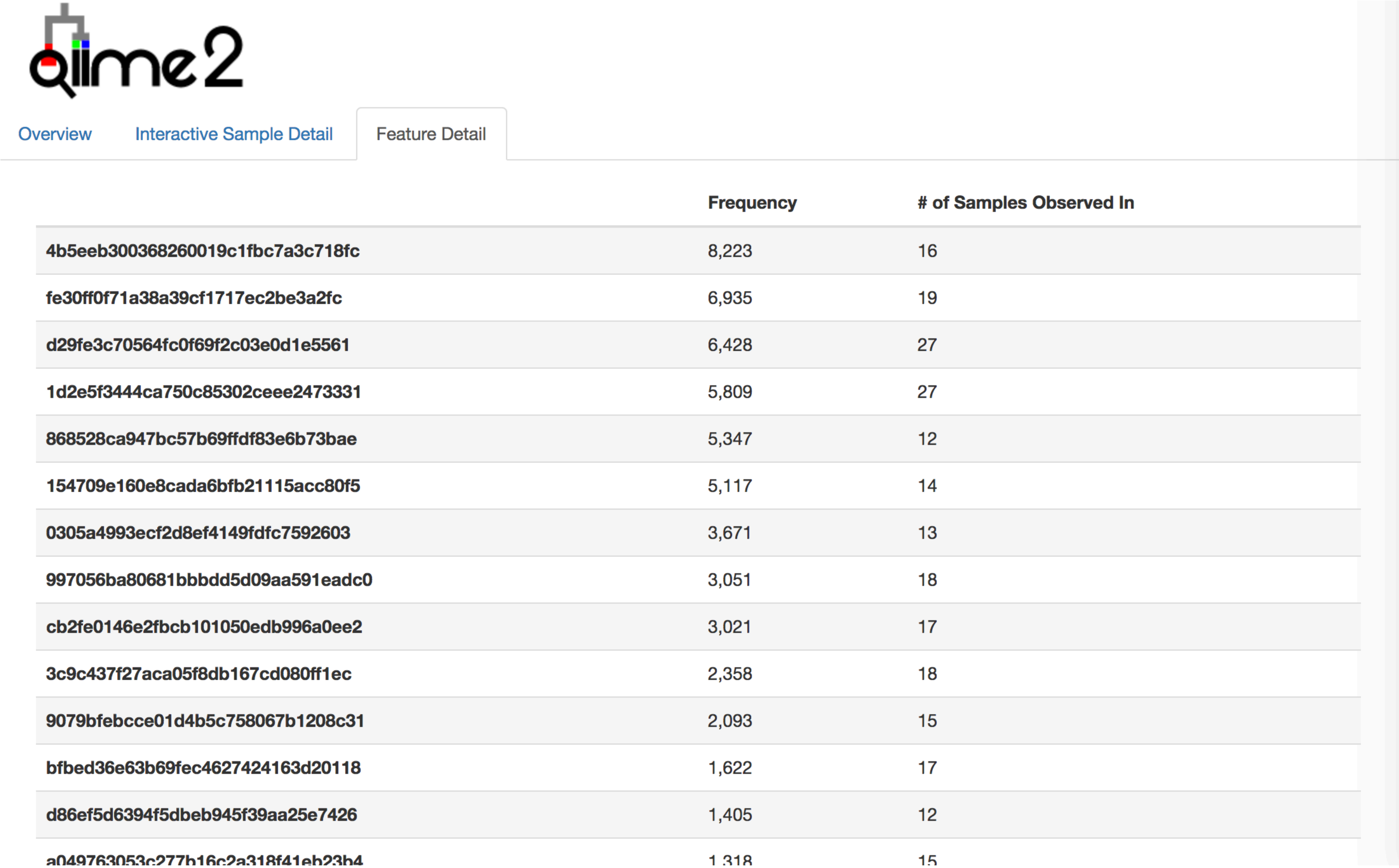
Table summary

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

Frequency per sample

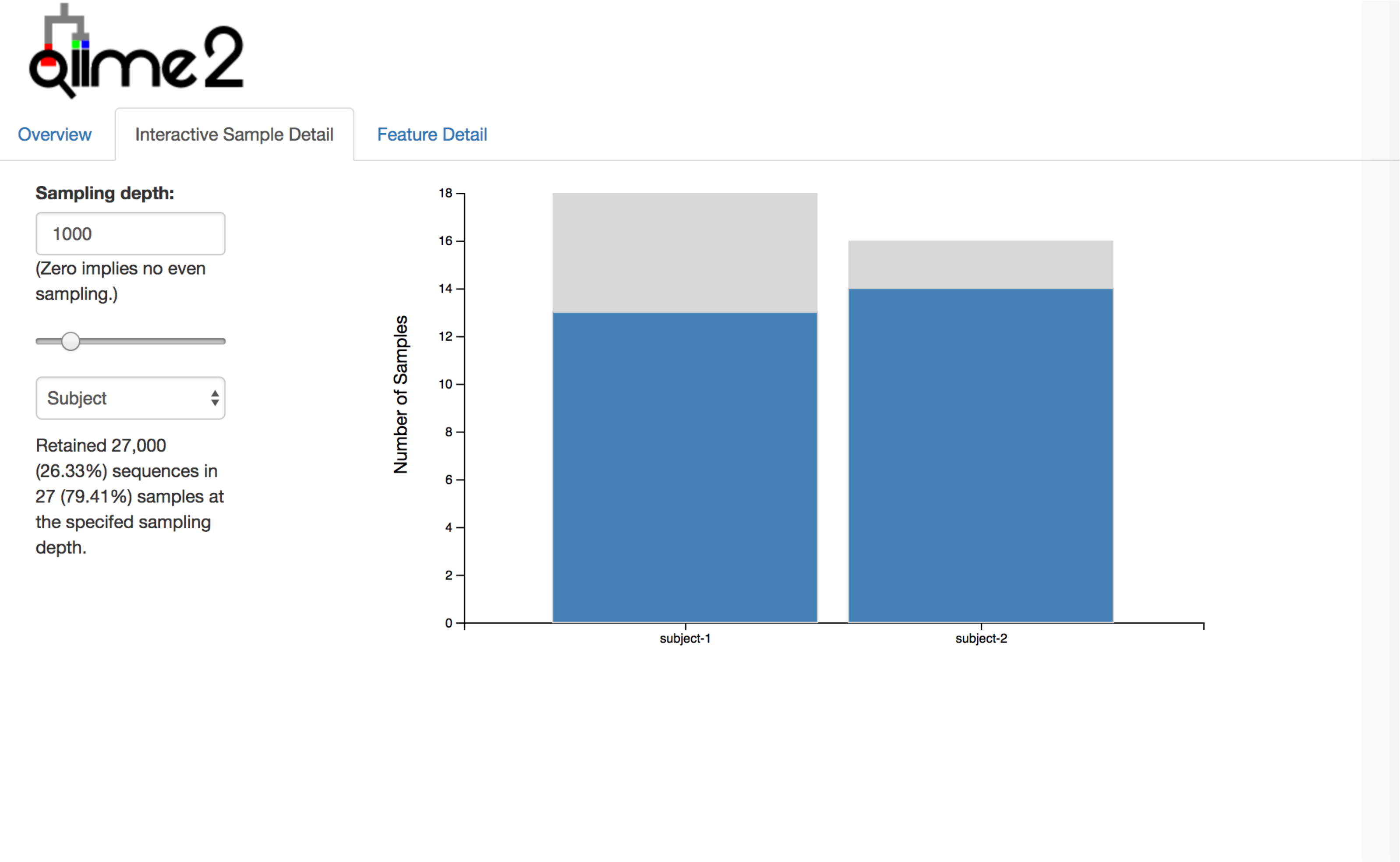
	Frequency
Minimum frequency	512.0
1st quartile	1,367.5
Median frequency	2,581.0
3rd quartile	4,952.0
Maximum frequency	6,770.0
Mean frequency	3,016.029411764706

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



Sampling depth:

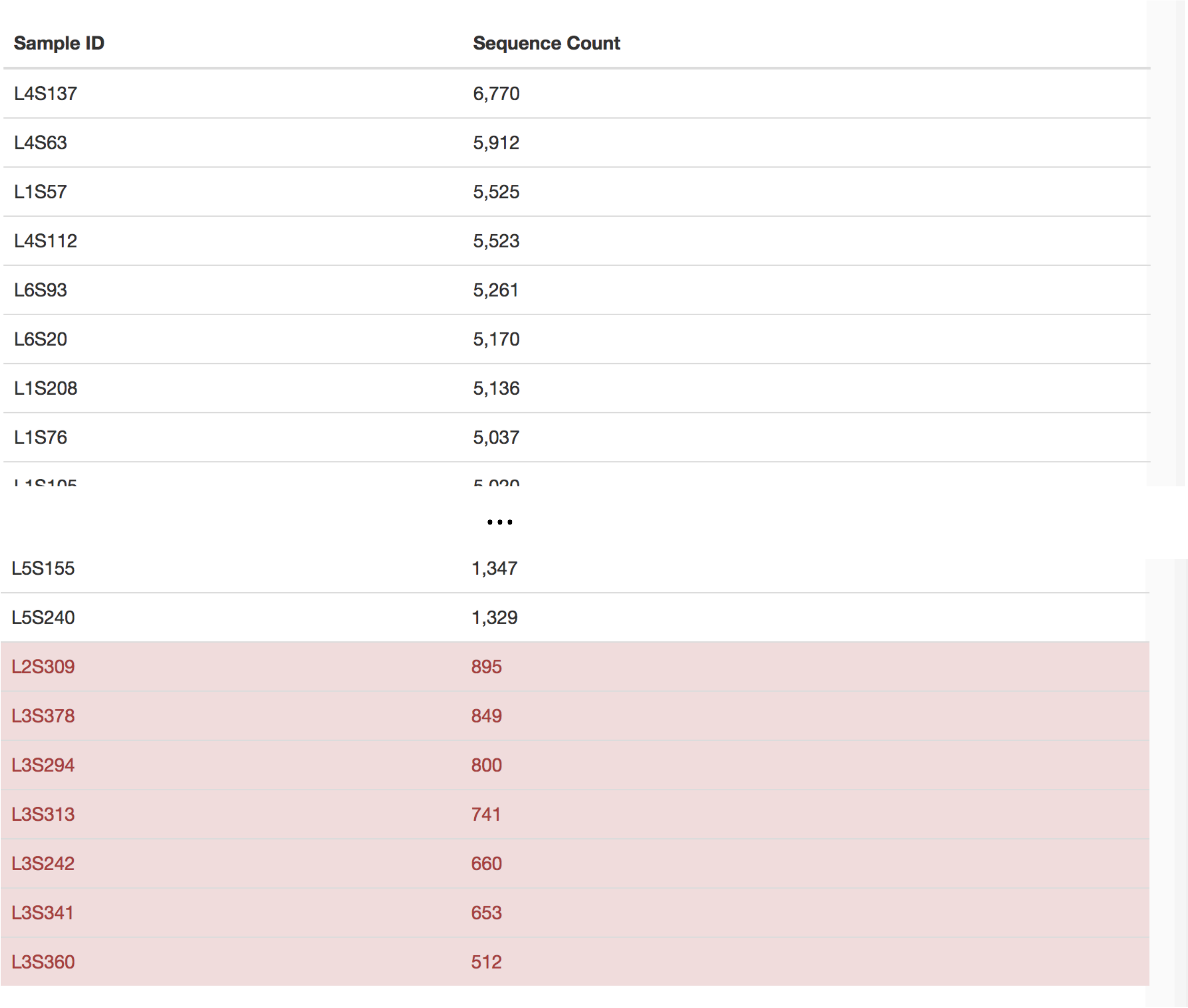
1000

(Zero implies no even sampling.)

Subject

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

Feature Table Summary View

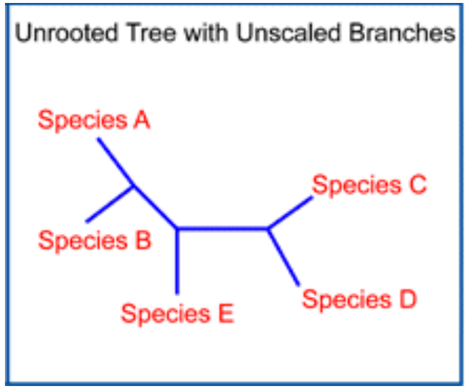


Sample ID	Sequence Count
L4S137	6,770
L4S63	5,912
L1S57	5,525
L4S112	5,523
L6S93	5,261
L6S20	5,170
L1S208	5,136
L1S76	5,037
L1S105	5,020
…	
L5S155	1,347
L5S240	1,329
L2S309	895
L3S378	849
L3S294	800
L3S313	741
L3S242	660
L3S341	653
L3S360	512

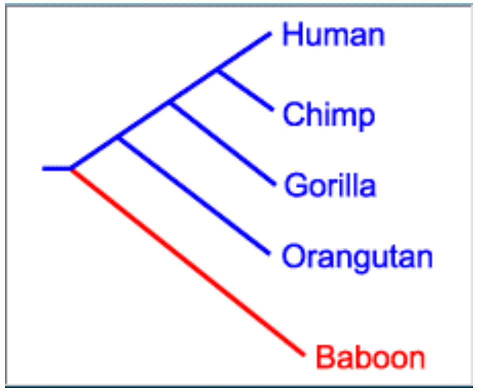
Phylogenetic Tree Creation

- 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

- Note: here we are doing *de novo* phylogenetic tree creation
  - Not necessarily the BEST approach, but an easy to show you ☺
- No visualizations will be produced

Practicum: Phylogenetic Tree Creation

```
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
```

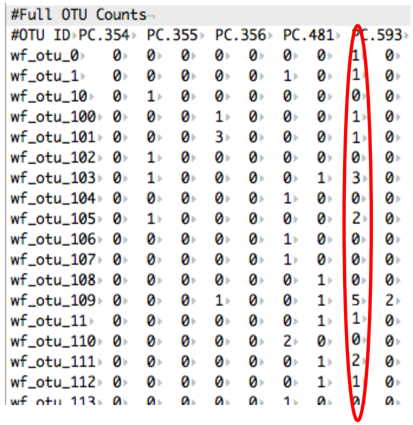
Core Metrics

- 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)

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   0   0   1   0
wf_otu_1     0   0   0   0   0   1   0   1   0
wf_otu_10    0   1   0   0   0   0   0   0   0
wf_otu_100   0   0   0   1   0   0   0   1   0
wf_otu_101   0   0   0   3   0   0   0   1   0
wf_otu_102   0   1   0   0   0   0   0   0   0
wf_otu_103   0   1   0   0   0   0   1   3   0
wf_otu_104   0   0   0   0   0   1   0   0   0
wf_otu_105   0   1   0   0   0   0   0   2   0
wf_otu_106   0   0   0   0   0   1   0   0   0
wf_otu_107   0   0   0   0   0   1   0   0   0
wf_otu_108   0   0   0   0   0   0   1   0   0
wf_otu_109   0   0   0   1   0   0   1   5   2
wf_otu_11    0   0   0   0   0   0   1   1   0
wf_otu_110   0   0   0   0   0   2   0   0   0
wf_otu_111   0   0   0   0   0   0   1   2   0
wf_otu_112   0   0   0   0   0   0   1   1   0
wf_otu_113   0   0   0   0   0   1   0   0   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

Full dataset (approximately 1,500 sequences per sample)

Dataset sampled at only 10 sequences per sample, showing the same pattern

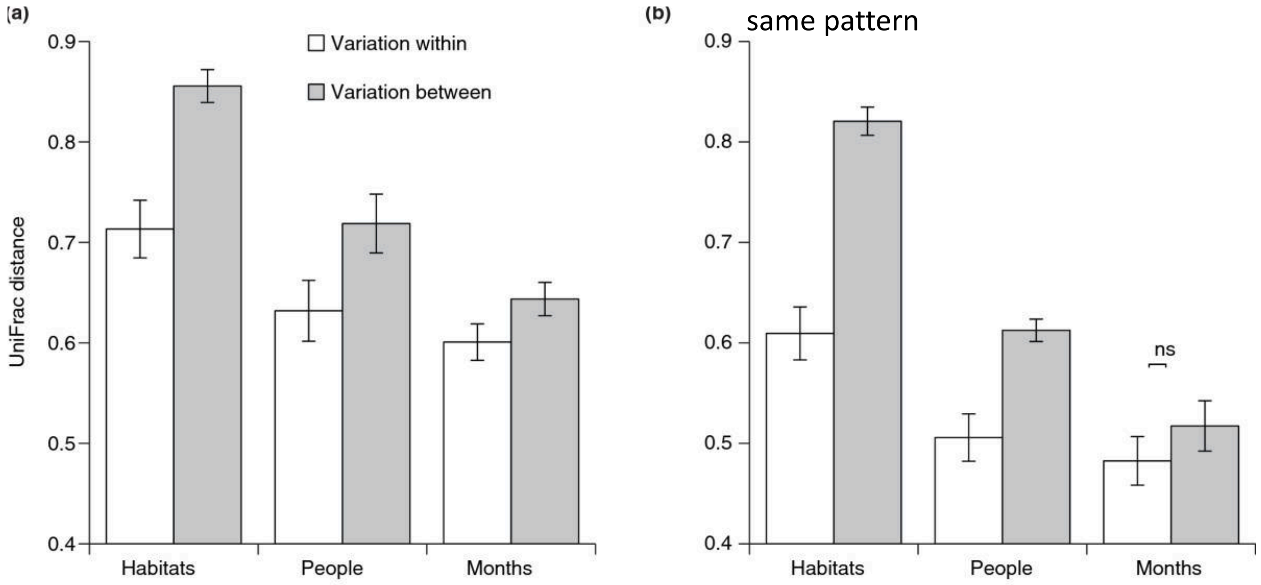


Fig. 2, Kuczynski, J. et al., "Direct sequencing of the human microbiome readily reveals community differences", Genome Biology, 2010

  - 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

- **Do NOT start typing yet!**

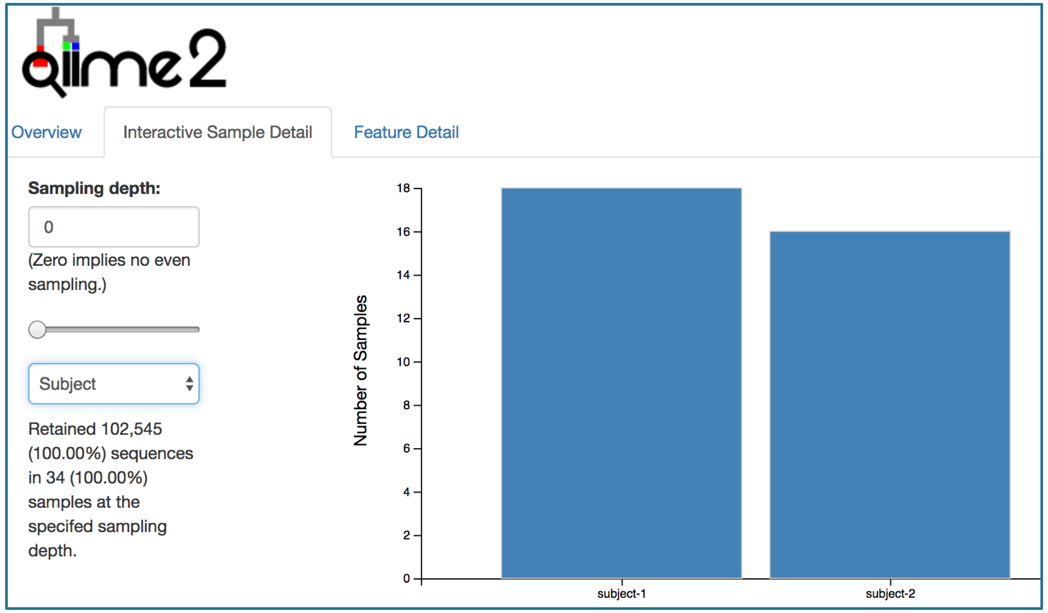
```
qiime diversity core-metrics \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth ??? \
  --output-dir metrics
```

- 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
```



Sampling depth:

0

(Zero implies no even sampling.)

Subject

Retained 102,545 (100.00%) sequences in 34 (100.00%) samples at the specifed sampling depth.

  - 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 \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth 800 \
  --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?
    - 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%)

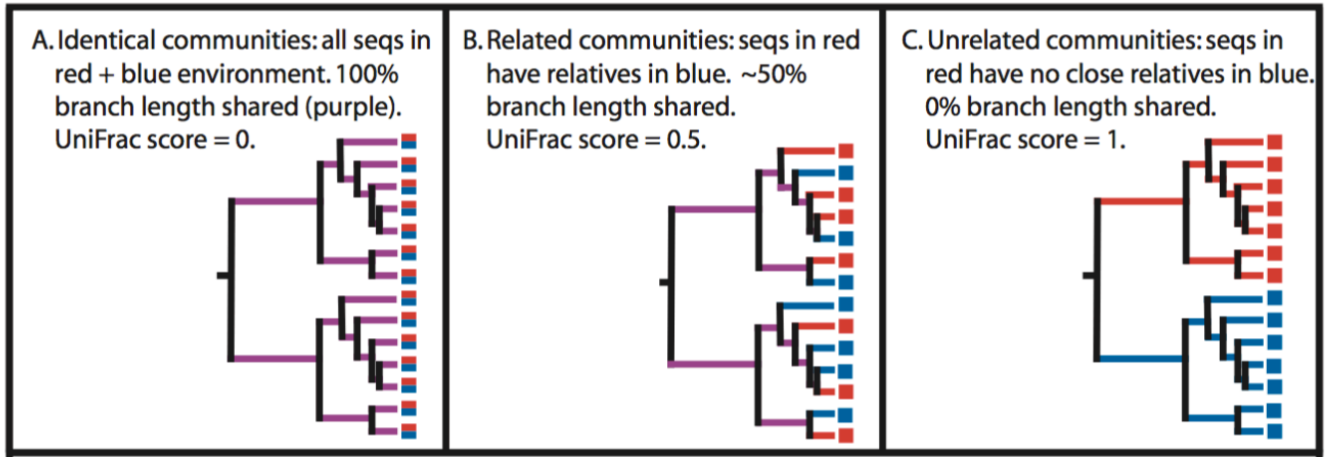
Practicum: Core Metrics

```
qiime diversity core-metrics \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth 800 \
  --output-dir metrics
```

- Note: there is no single visualization for core metrics
  - We will examine a few different visualizations later

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



A. Identical communities: all seqs in red + blue environment. 100% branch length shared (purple). UniFrac score = 0.

B. Related communities: seqs in red have relatives in blue. ~50% branch length shared. UniFrac score = 0.5.

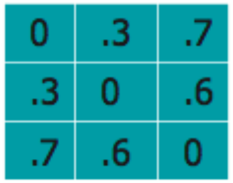
C. Unrelated communities: seqs in red have no close relatives in blue. 0% branch length shared. UniFrac score = 1.

Illustration courtesy of Dr. Rob Knight

  - 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



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

Distance Matrix

- PCoA axes can't be decomposed into independent variable contributions
  - But results can be compared to metadata to identify patterns

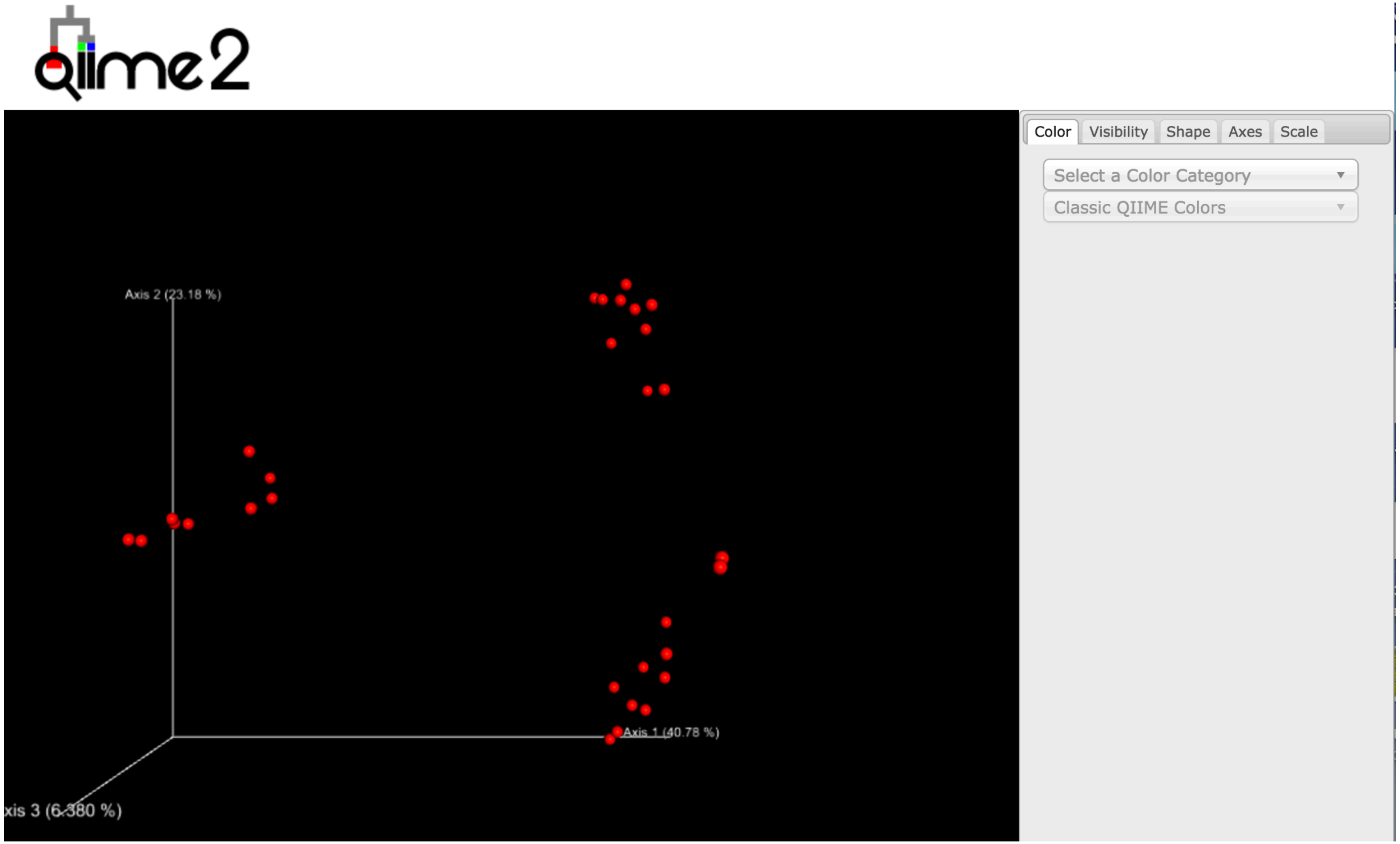
Practicum: Beta Diversity Ordination

```
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
- To check the group significance of a different metric, just input a different file
  - To find them:

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

Beta Diversity Ordination View

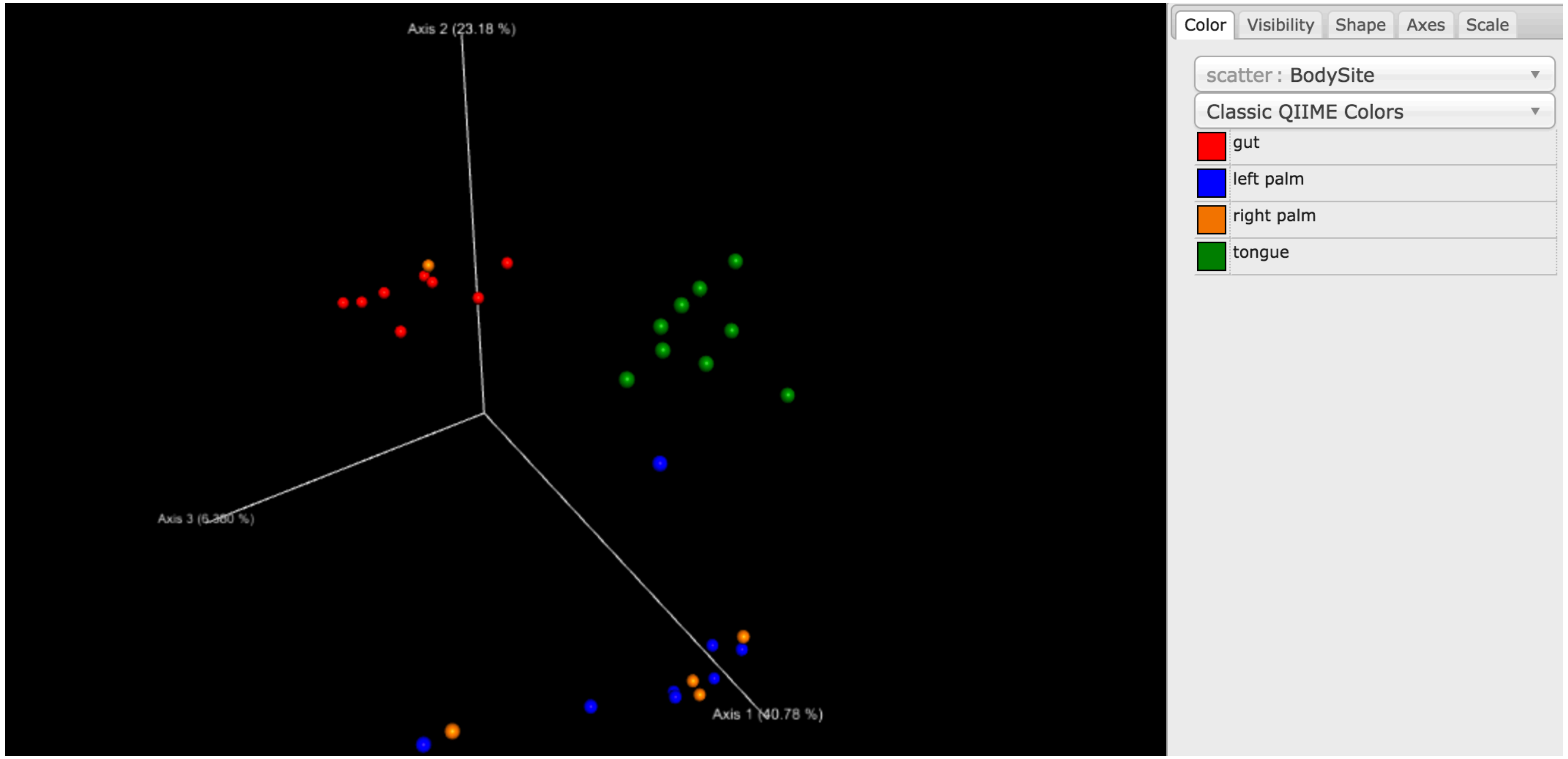


Exercise: Beta Diversity Ordination

- 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
- 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

- 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



scatter : BodySite

Classic QIIME Colors

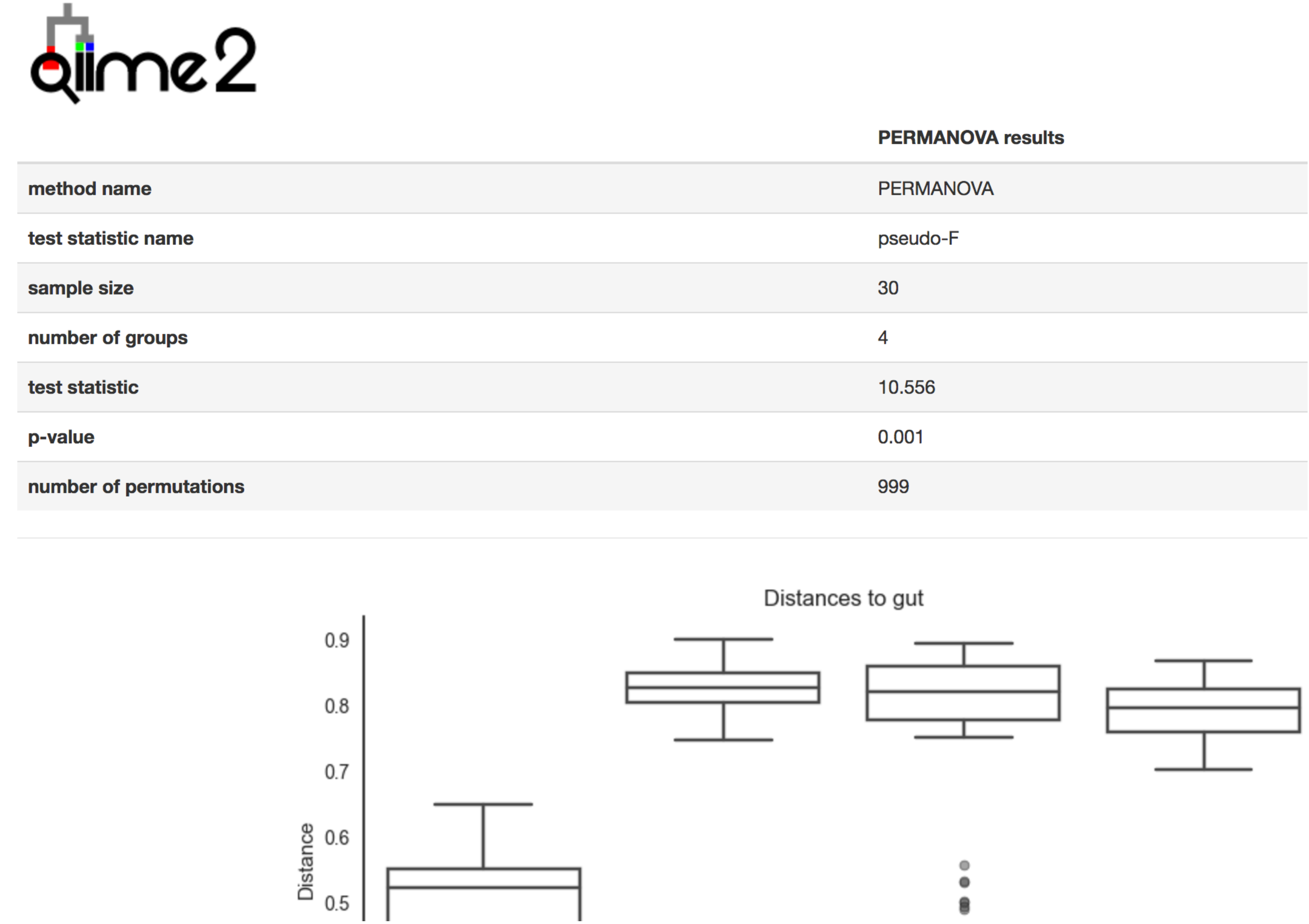
- 🟥 gut
- 🟦 left palm
- 🟧 right palm
- 🟩 tongue

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-category BodySite \
  --p-pairwise  \
  --o-visualization \
   metrics/unweighted-unifrac-bodysite-significance.qzv
```

- Standard caveats apply—not the “one true metric”, etc

Beta Diversity Group Significance View



	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	30
number of groups	4
test statistic	10.556
p-value	0.001
number of permutations	999

Distances to gut

Exercise: Beta Diversity Group Significance

- 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

- 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 <= 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
  - How about Subject?

```
qiime diversity beta-group-significance \
  --i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \
  --m-metadata-file sample-metadata.tsv \
  --m-metadata-category Subject\
  --p-pairwise  \
  --o-visualization metrics/unweighted-unifrac-subject-significance.qzv
```

- Nope: significance of difference of distributions of unweighted unifrac metric grouped by subject has p =0.442

Acknowledgements

- Center for Computational Biology & Bioinformatics, University of California at San Diego
- 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

Practicum: Beta Diversity Ordination

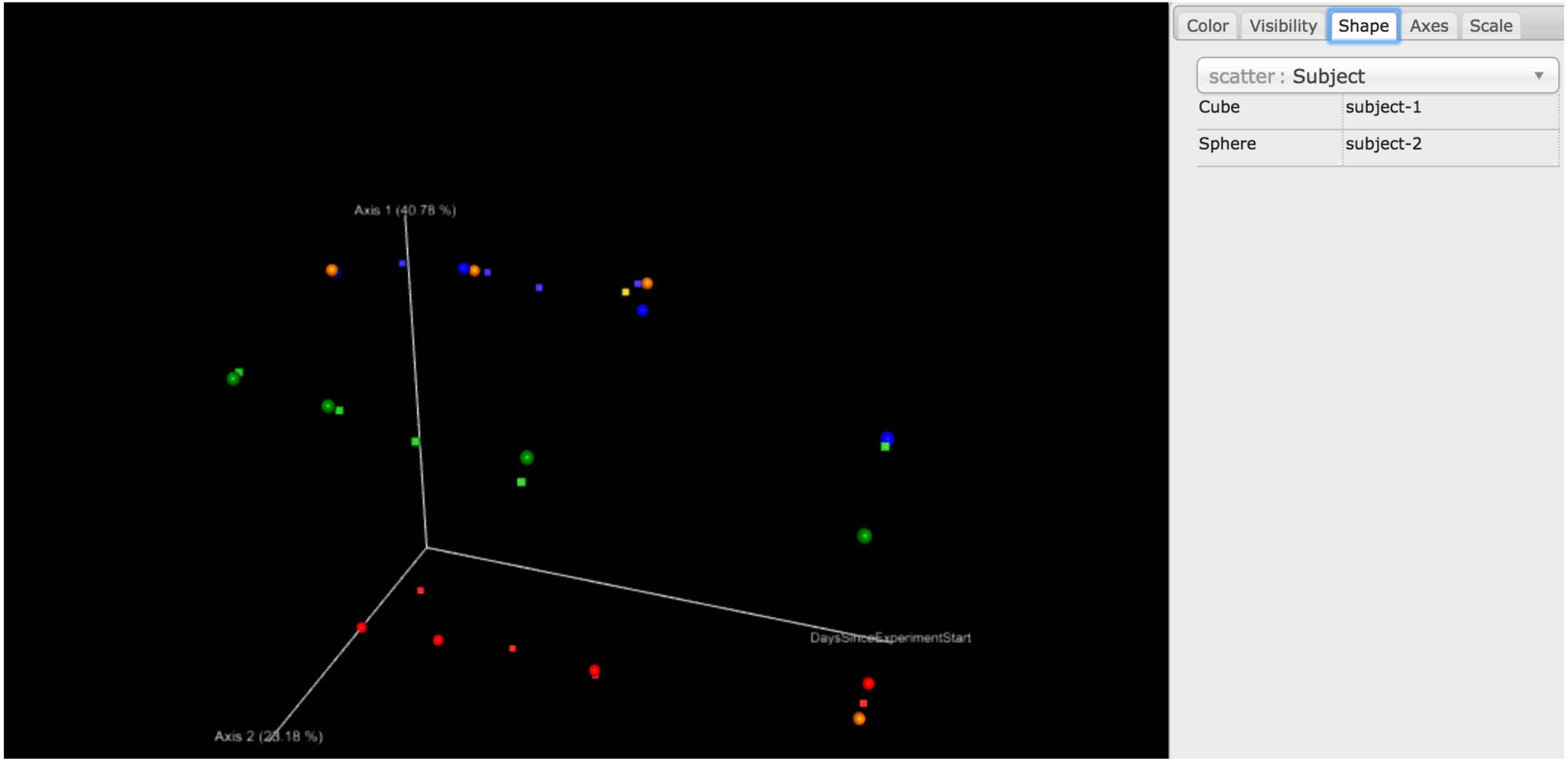
- But wait, this is time-series data! Maybe we'd like to view it on a time axis:

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

- Standard caveats apply

Supplemental Slides

Beta Diversity Ordination View (cont.)



scatter : Subject ▾

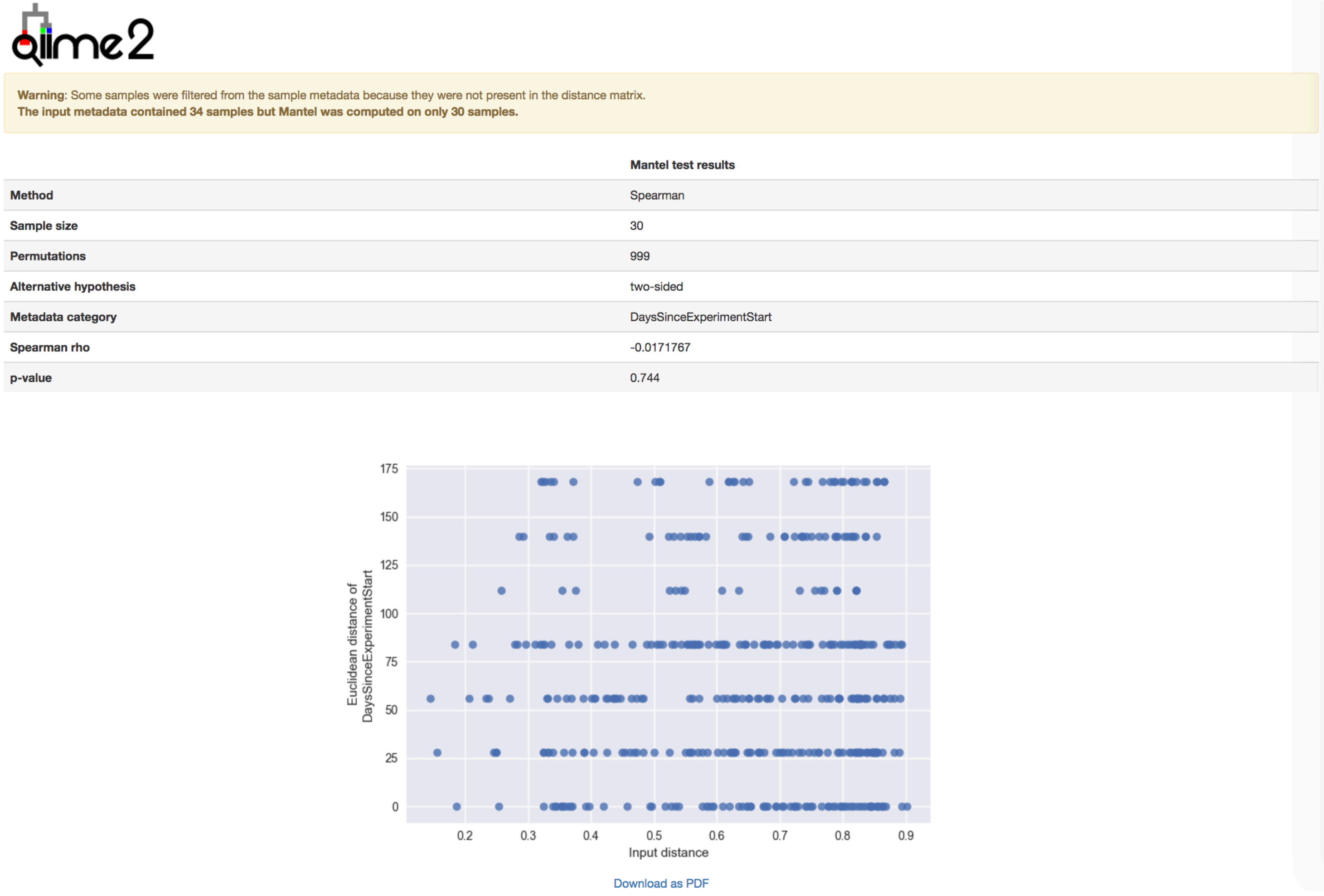
Cube	subject-1
Sphere	subject-2

Practicum: Beta Diversity Correlation

```
qiime diversity beta-correlation \
  --i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \
  --m-metadata-file sample-metadata.tsv \
  --m-metadata-category DaysSinceExperimentStart \
  --o-visualization metrics/unweighted-unifrac-dayssinceexperimentstart-beta-correlation.qzv
```

- Standard caveats apply

Beta Diversity Correlation View



qiime2

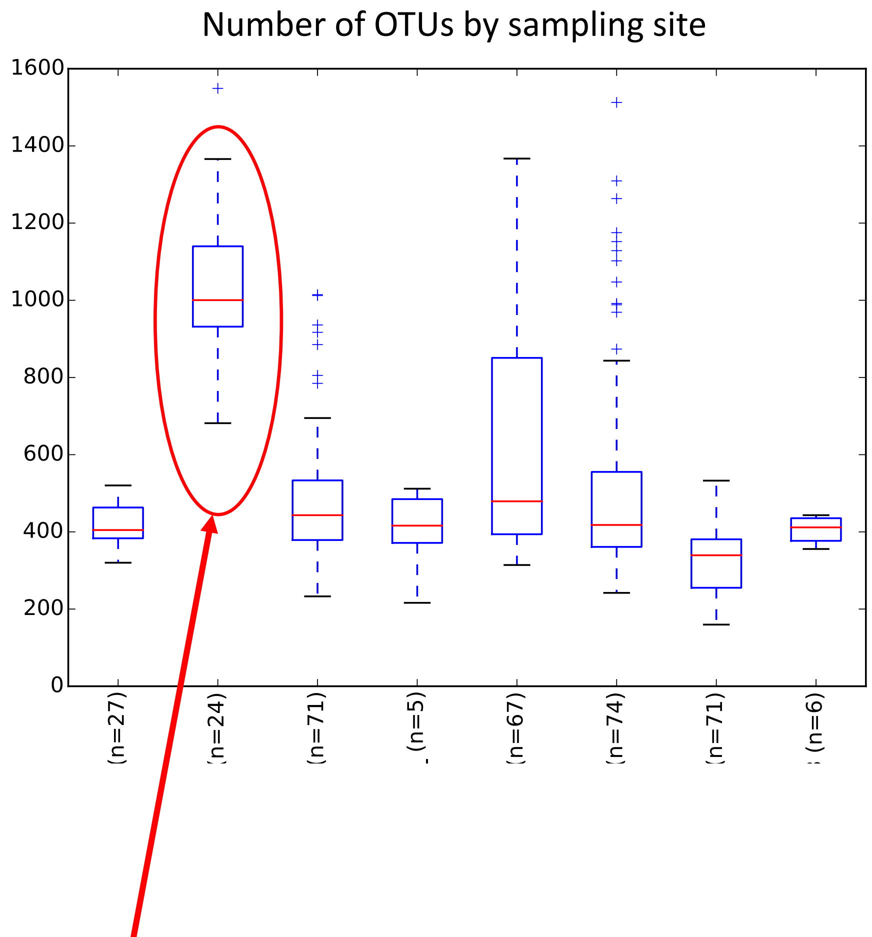
Warning: Some samples were filtered from the sample metadata because they were not present in the distance matrix.
The input metadata contained 34 samples but Mantel was computed on only 30 samples.

	Mantel test results
Method	Spearman
Sample size	30
Permutations	999
Alternative hypothesis	two-sided
Metadata category	DaysSinceExperimentStart
Spearman rho	-0.0171767
p-value	0.744

Alpha Diversity

- “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



Number of OTUs by sampling site

High within-sample diversity—why?

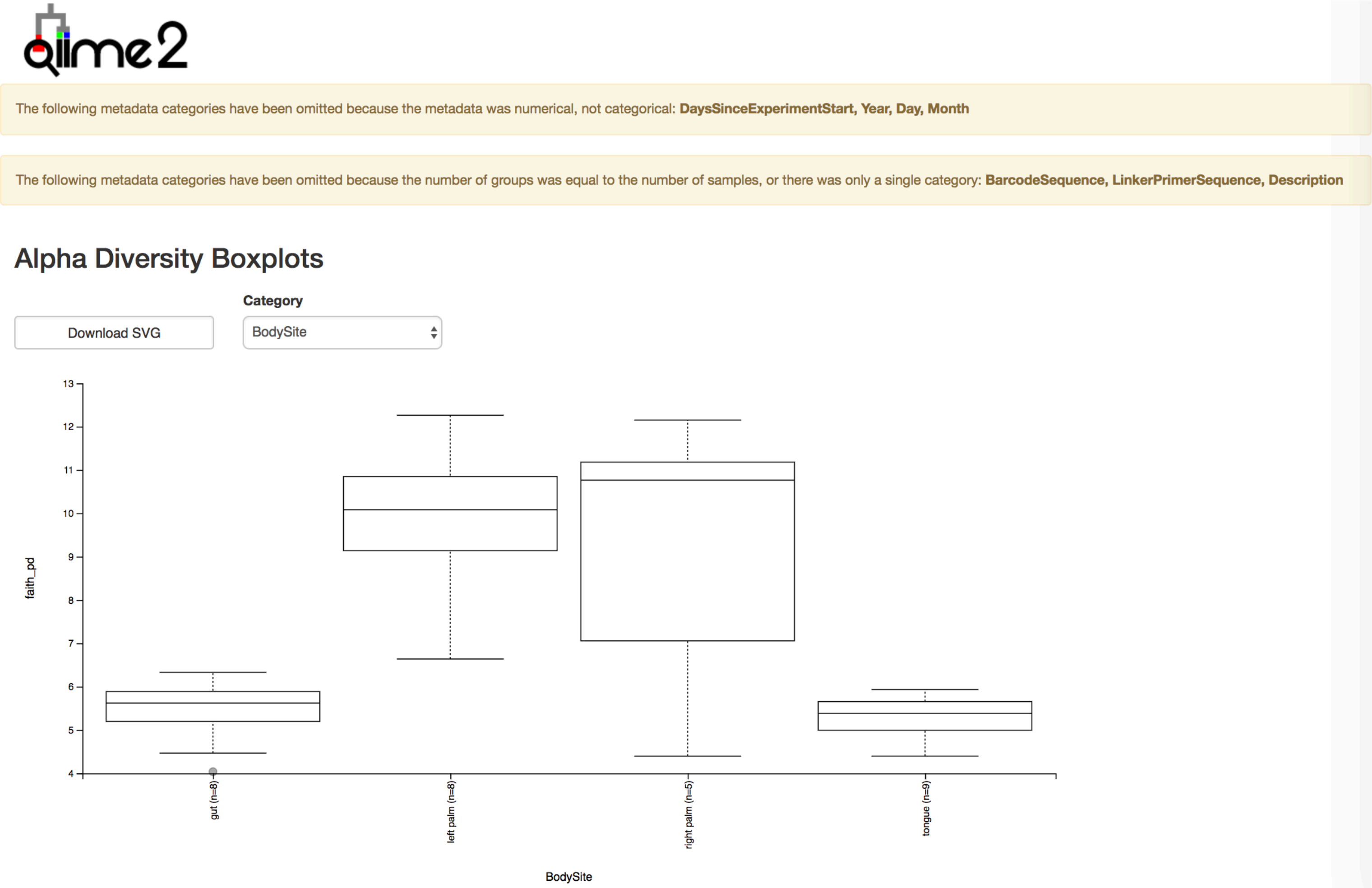
Practicum: Alpha Diversity Group Significance

```
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` 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



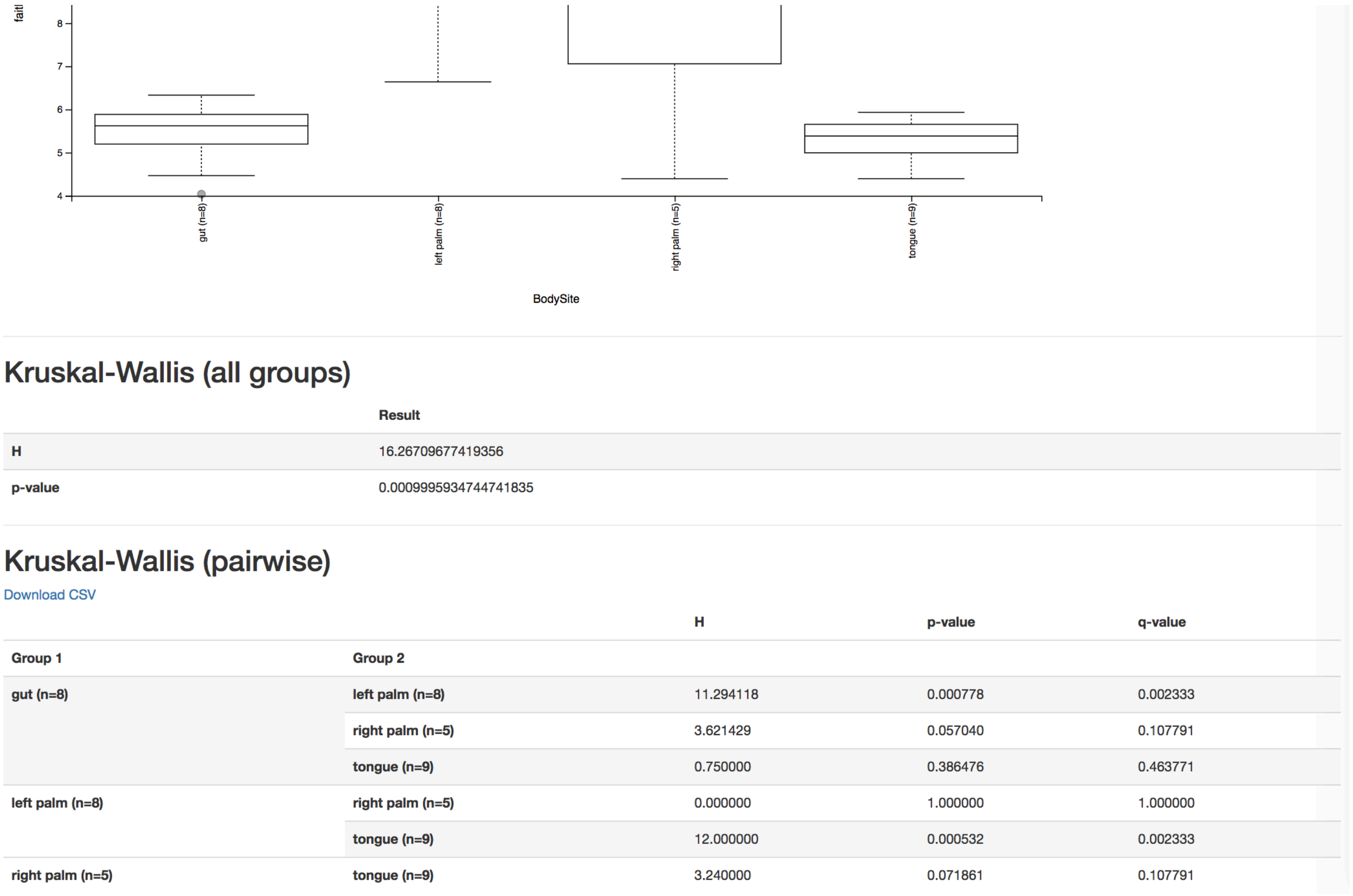
The following metadata categories have been omitted because the metadata was numerical, not categorical: **DaysSinceExperimentStart, Year, Day, Month**

The following metadata categories have been omitted because the number of groups was equal to the number of samples, or there was only a single category: **BarcodeSequence, LinkerPrimerSequence, Description**

Alpha Diversity Boxplots

Category: BodySite

Alpha Diversity Group Significance View



Kruskal-Wallis (all groups)

	Result
H	16.26709677419356
p-value	0.0009995934744741835

Kruskal-Wallis (pairwise)

Download CSV

		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)	0.750000	0.386476	0.463771
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 different between p-value and q-value
  - Is Subject value associated with significant differences in phylogenetic diversity?

Answers: Alpha Diversity Group Significance

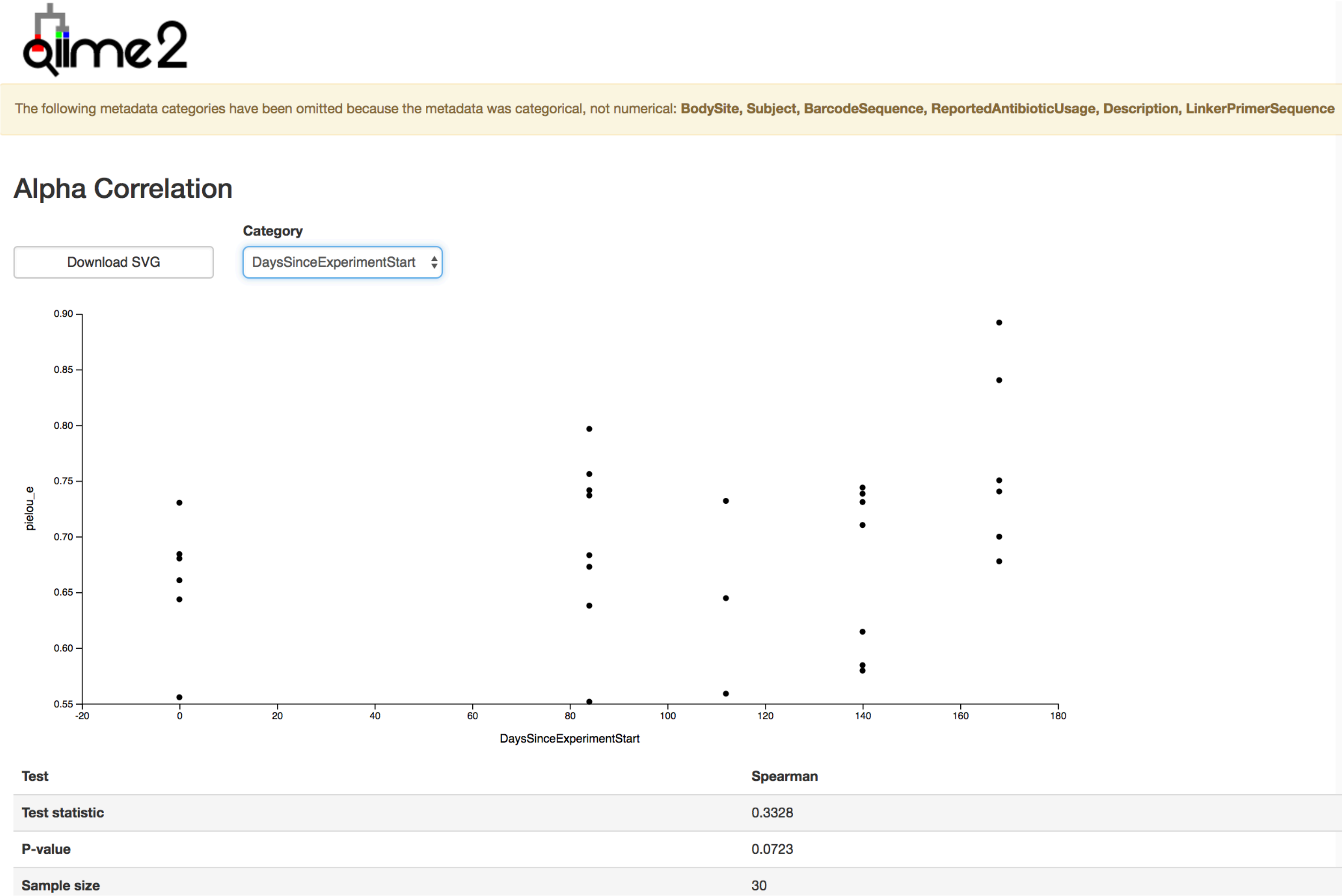
- My answers:
  - Is BodySite value associated with significant differences in phylogenetic diversity?
    - Yes, with p < 1 E-3
  - Which two sites have the most significantly difference in phylogenetic diversity distributions?
    - Left palm is (equally) most significantly different from gut and tongue
      - Consider: any idea why perhaps left palm but not right?
  - Is Subject value associated with significant differences in phylogenetic diversity?
    - No, p value = 0.24

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



The following metadata categories have been omitted because the metadata was categorical, not numerical: **BodySite, Subject, BarcodeSequence, ReportedAntibioticUsage, Description, LinkerPrimerSequence**

Alpha Correlation

Category

DaysSinceExperimentStart

Test	Spearman
Test statistic	0.3328
P-value	0.0723
Sample size	30

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, 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 ☺

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, 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

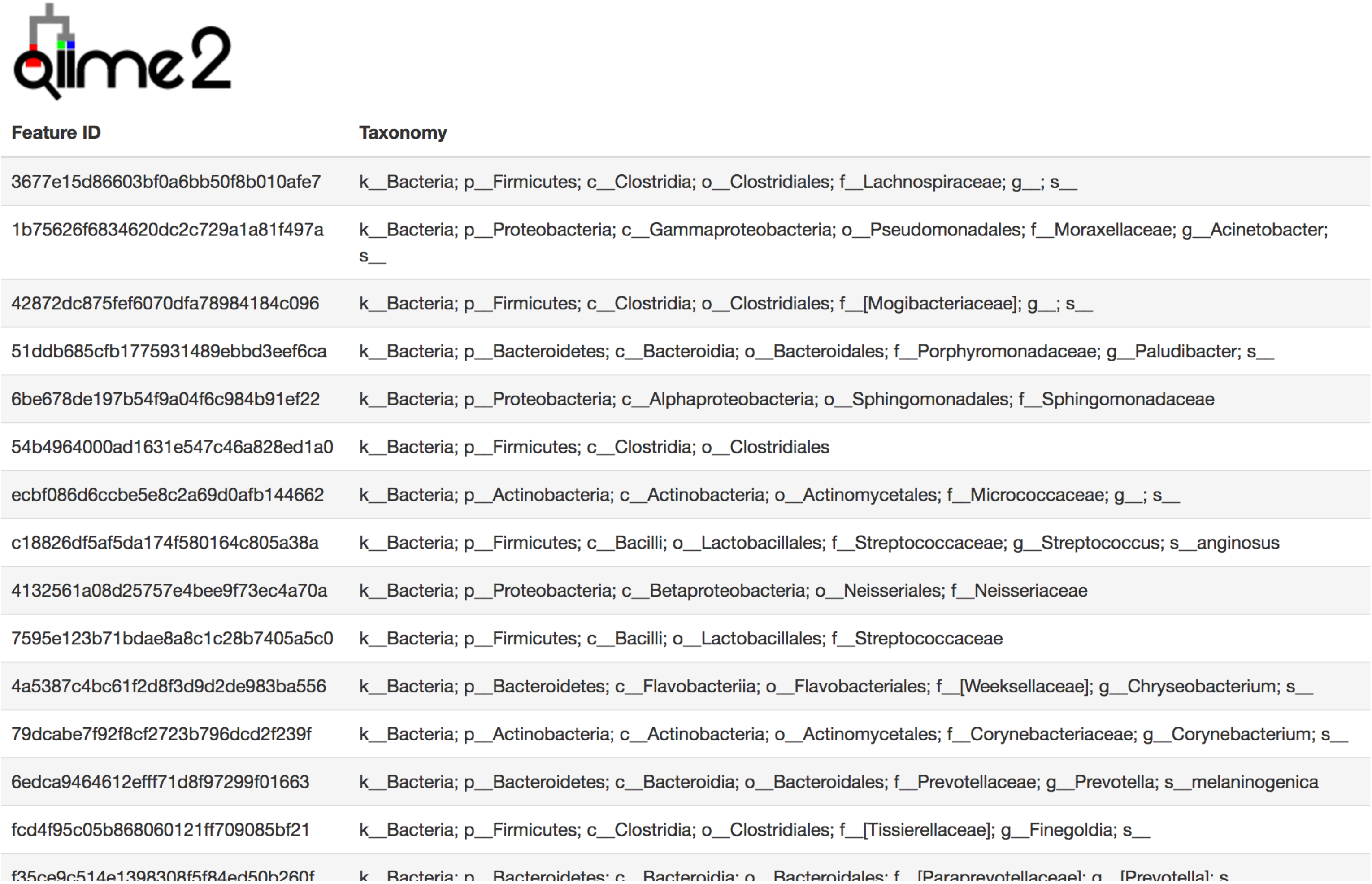
Taxonomy: Expectation Vs Reality

	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 taxa tabulate \
  --i-data taxonomy.qza \
  --o-visualization taxonomy.qzv
```

Taxonomic Assignment Tabulation View



qiime2

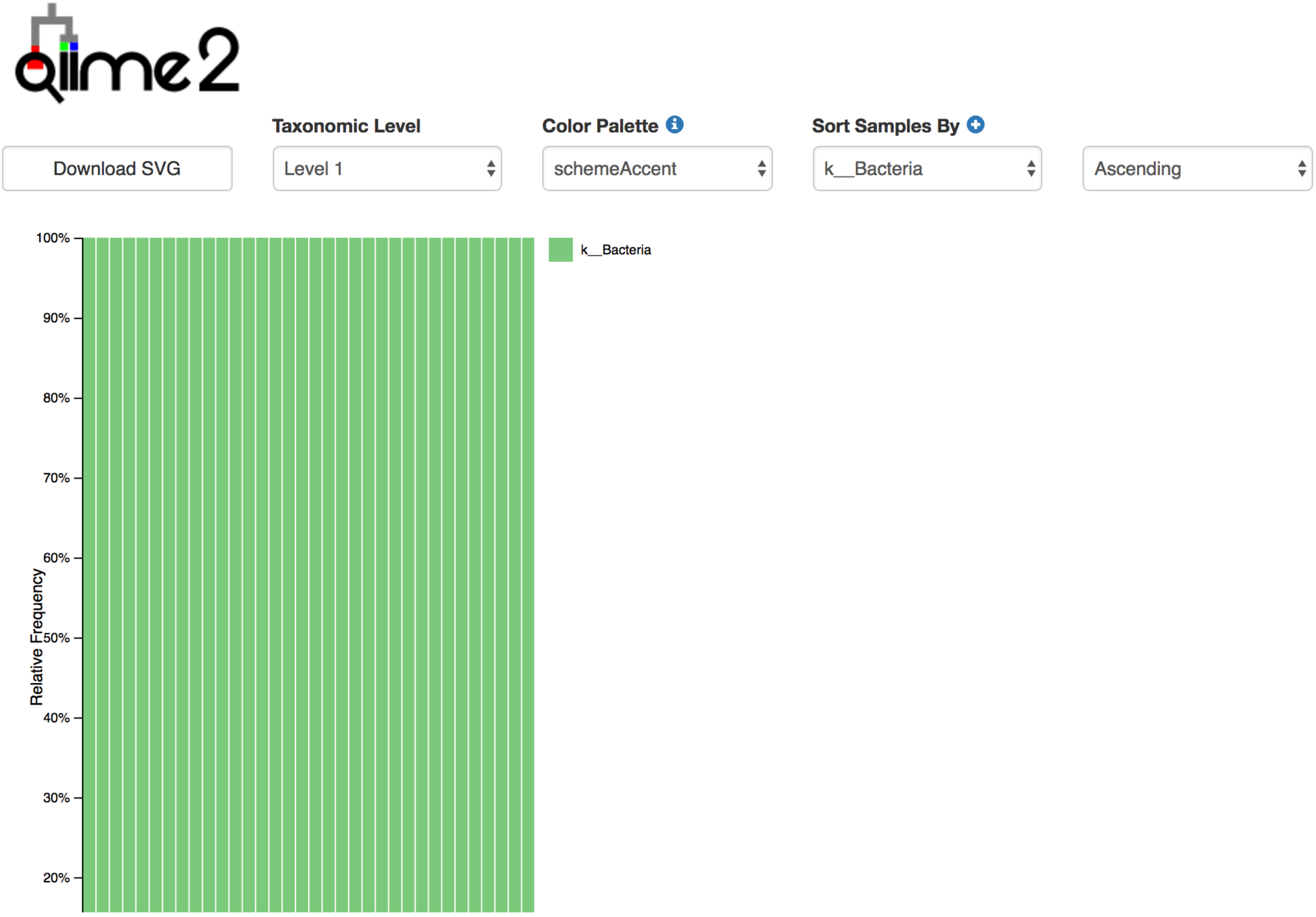
Feature ID	Taxonomy
3677e15d86603bf0a6bb50f8b010afe7	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
1b75626f6834620dc2c729a1a81f497a	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Acinetobacter; s__
42872dc875fef6070dfa78984184c096	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



Download SVG

Taxonomic Level: Level 1

Color Palette: schemeAccent

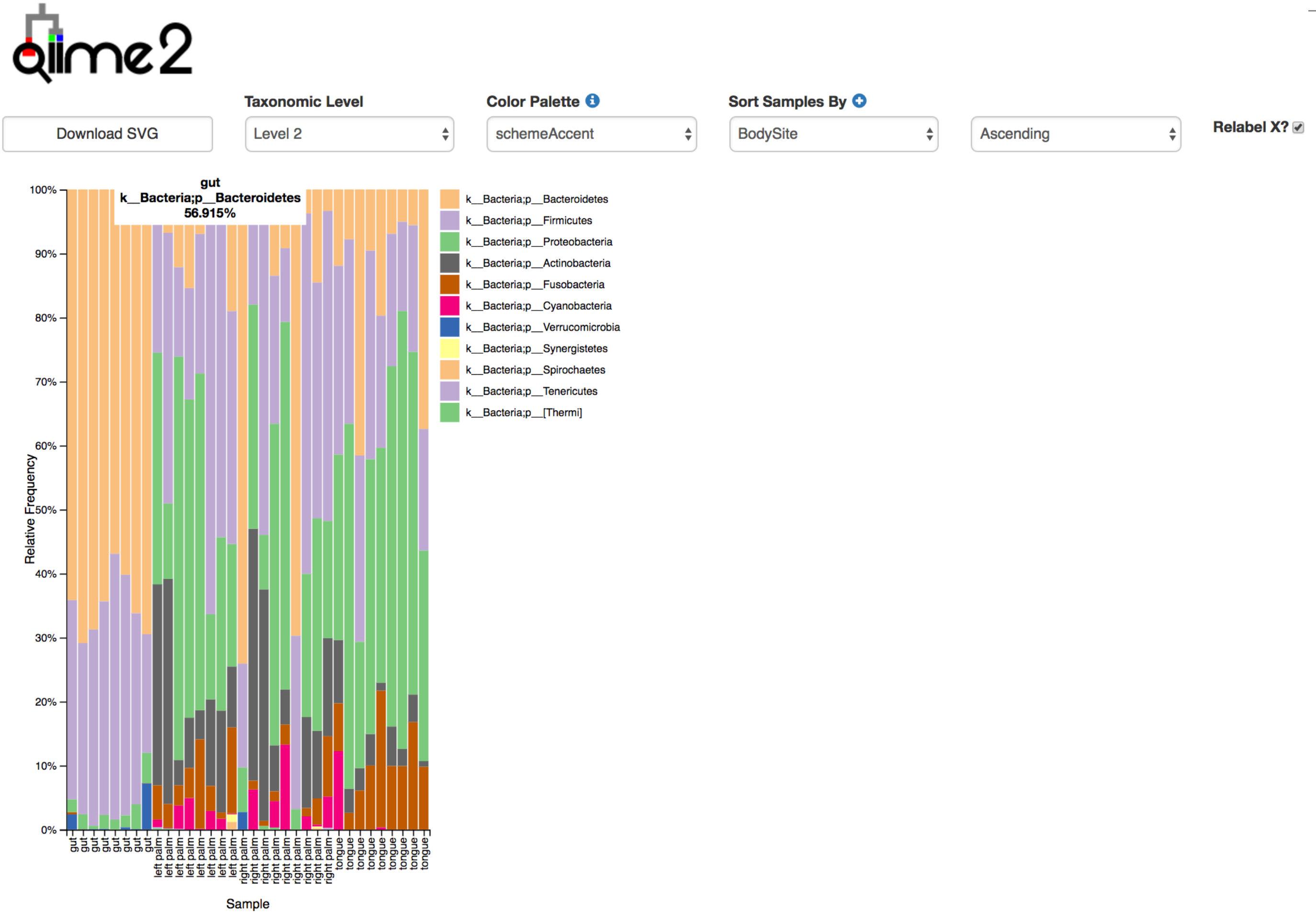
Sort Samples By: k__Bacteria

Ascending

Exercise: Taxonomic Assignment

- “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


Download SVG

Taxonomic Level: Level 2

Color Palette: schemeAccent

Sort Samples By: BodySite, Ascending

Relabel X? ☑

qiime2
gut
k__Bacteria;p__Bacteroidetes
56.915%
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Relative Frequency
gut
left palm
right palm
tongue
Sample
k__Bacteria;p__Bacteroidetes
k__Bacteria;p__Firmicutes
k__Bacteria;p__Proteobacteria
k__Bacteria;p__Actinobacteria
k__Bacteria;p__Fusobacteria
k__Bacteria;p__Cyanobacteria
k__Bacteria;p__Verrucomicrobia
k__Bacteria;p__Synergistetes
k__Bacteria;p__Spirochaetes
k__Bacteria;p__Tenericutes
k__Bacteria;p__[Thermi]

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

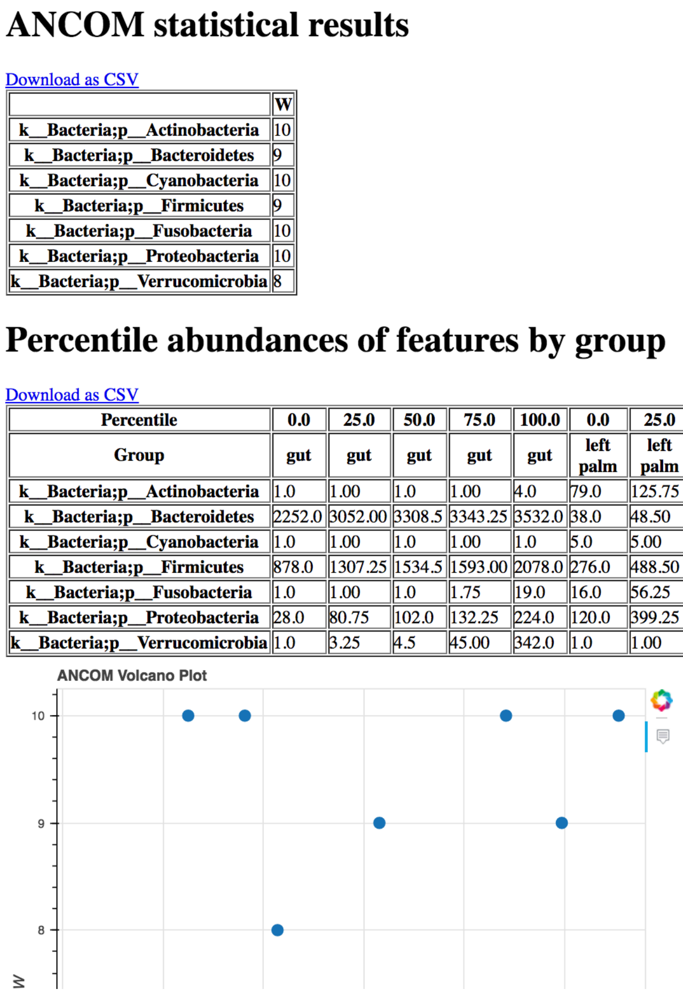
Differential Abundance Analysis

- 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?
  - Remember: microbiome datasets are “compositional” (fixed sum)
  - **Watch out**: “traditional” statistical methods perform badly for this sort of data!
    - E.g., 95% false positives when you expect an FDR of 5%
  - What to use instead?
    - Balance trees (borrowed from geology) are currently the best known option (as of 2017)
      - But they aren’t implemented in QIIME 2 yet
      - So until they are, use previous best known option (as of 2015)
    - ANCOM (ANalysis of Composition Of Microbiomes)

Practicum: Differential Abundance Analysis

```
qiime taxa collapse \
  --i-table table.qza \
  --i-taxonomy taxonomy.qza \
  --p-level 2 \
  --o-collapsed-table table-level2.qza

qiime composition add-pseudocount \
  --i-table table-l2.qza \
  --o-composition-table comp-table-level2.qza

qiime composition ancom \
  --i-table comp-table-level2.qza \
  --m-metadata-file sample-metadata.tsv \
  --m-metadata-category BodySite \
  --o-visualization ancom-bodysite-level2.qzv
```

Differential Abundance Analysis View



ANCOM statistical results

Download as CSV

	W
k__Bacteria;p__Actinobacteria	10
k__Bacteria;p__Bacteroidetes	9
k__Bacteria;p__Cyanobacteria	10
k__Bacteria;p__Firmicutes	9
k__Bacteria;p__Fusobacteria	10
k__Bacteria;p__Proteobacteria	10
k__Bacteria;p__Verrucomicrobia	8

Percentile abundances of features by group

Download as CSV

Percentile	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
Group	gut	gut	gut	gut	gut	left palm	left palm	left palm	left palm	left palm	right palm	right palm	right palm	right palm	right palm	tongue	tongue	tongue	tongue	tongue
k__Bacteria;p__Actinobacteria	1.0	1.00	1.0	1.00	4.0	79.0	125.75	203.5	388.00	1040.0	1.0	29.0	269.0	716.0	906.0	26.0	51.0	78.0	139.0	199.0
k__Bacteria;p__Bacteroidetes	2252.0	3052.00	3308.5	3343.25	3532.0	38.0	48.50	220.5	247.25	492.0	9.0	48.0	199.0	490.0	985.0	93.0	150.0	257.0	390.0	1970.0
k__Bacteria;p__Cyanobacteria	1.0	1.00	1.0	1.00	1.0	5.0	5.00	27.0	41.50	80.0	1.0	1.0	28.0	69.0	290.0	1.0	1.0	1.0	1.0	212.0
k__Bacteria;p__Firmicutes	878.0	1307.25	1534.5	1593.00	2078.0	276.0	488.50	691.0	828.50	1250.0	60.0	137.0	231.0	2495.0	3111.0	276.0	408.0	508.0	722.0	1000.0
k__Bacteria;p__Fusobacteria	1.0	1.00	1.0	1.75	19.0	16.0	56.25	93.0	191.25	495.0	1.0	7.0	12.0	72.0	558.0	36.0	129.0	157.0	514.0	770.0
k__Bacteria;p__Proteobacteria	28.0	80.75	102.0	132.25	224.0	120.0	399.25	646.5	1010.75	1860.0	28.0	64.0	295.0	1083.0	2248.0	283.0	671.0	749.0	1731.0	3538.0
k__Bacteria;p__Verrucomicrobia	1.0	3.25	4.5	45.00	342.0	1.0	1.00	1.0	1.00	1.0	1.0	1.0	1.0	1.0	19.0	1.0	1.0	1.0	1.0	1.0

ANCOM Volcano Plot

Details: Differential Abundance Analysis

- W-statistic
  - # of other items from which a single item is found to be significantly different
    - With alpha=0.05 by default (can be changed)
- Percentile abundance table:
  - A table of items and their percentile abundances in each group
  - Rows are items
  - Columns are percentile within a group
  - Values are abundance of reads for given percentile for that group