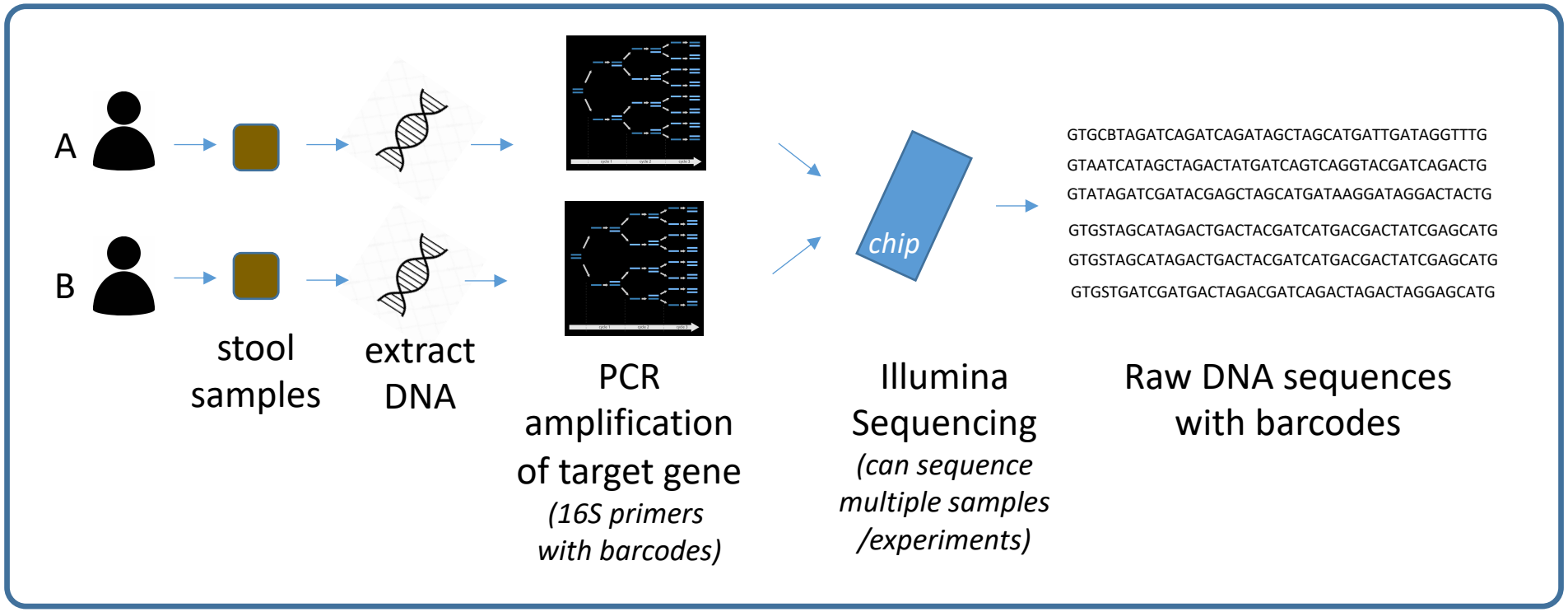
Module 4

Importing and demultiplexing

Module Outcomes

1. Import different types of sequencing files into QIIME2 with or without a manifest file
2. Demultiplex using QIIME2

Generating the data



QIIME2 – Bioinformatics Tool

Correspondence | Published: 24 July 2019

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2

Evan Bolyen, Jai Ram Rideout, [...] J. Gregory Caporaso

Nature Biotechnology **37**, 852–857(2019) | Cite this article

An Author Correction to this article was published on 09 August 2019

This article has been updated

To the Editor – Rapid advances in DNA-sequencing and bioinformatics technologies in the past two decades have substantially improved understanding of the microbial world. This

QIIME2 Moving Pictures Tutorial

https://docs.qiime2.org/2020.8/tutorials/moving-pictures/

Caporaso *et al. Genome Biology* 2011, **12**:R50
http://genomebiology.com/2011/12/5/R50

Genome Biology

RESEARCH Open Access

Moving pictures of the human microbiome

J Gregory Caporaso[1], Christian L Lauber[2], Elizabeth K Costello[3], Donna Berg-Lyons[2], Antonio Gonzalez[4], Jesse Stombaugh[1], Dan Knights[4], Pawel Gajer[5], Jacques Ravel[5], Noah Fierer[2,6], Jeffrey I Gordon[7] and Rob Knight[1,8*]

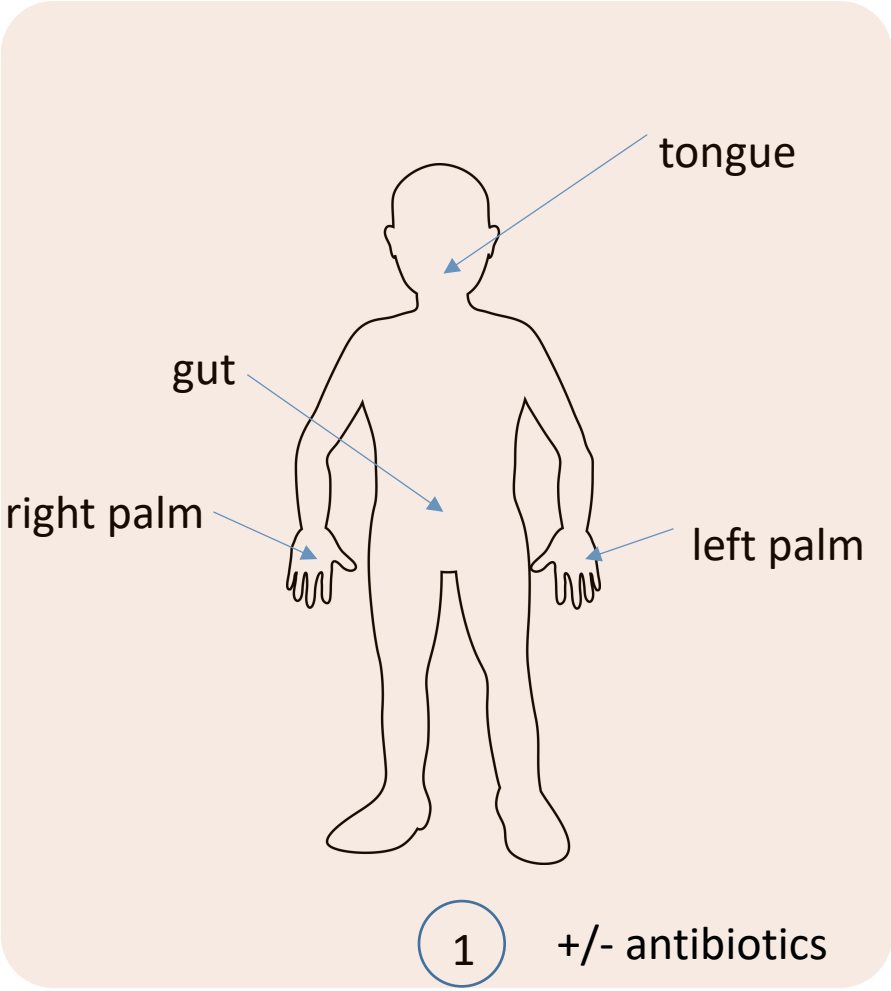
Abstract

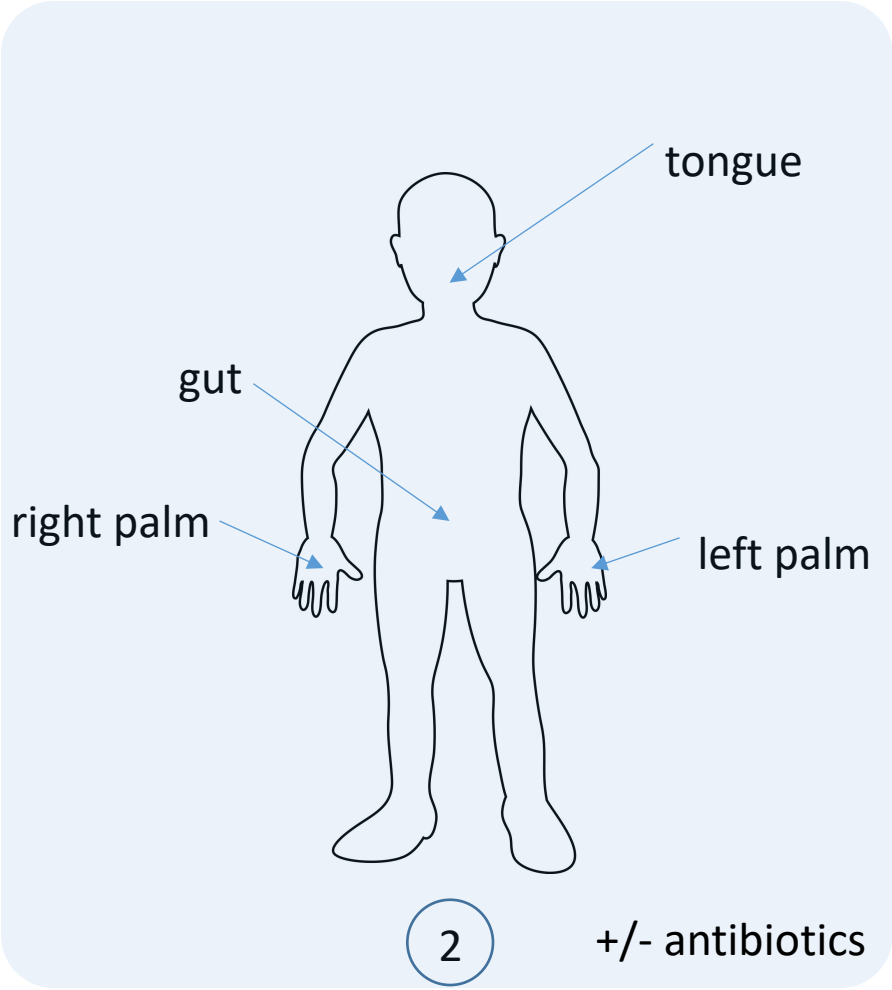
Background: Understanding the normal temporal variation in the human microbiome is critical to developing treatments for putative microbiome-related afflictions such as obesity, Crohn's disease, inflammatory bowel disease and malnutrition. Sequencing and computational technologies, however, have been a limiting factor in performing dense time series analysis of the human microbiome. Here, we present the largest human microbiota time series analysis to date, covering two individuals at four body sites over 396 timepoints.

Results: We find that despite stable differences between body sites and individuals, there is pronounced variability in an individual's microbiota across months, weeks and even days. Additionally, only a small fraction of the total taxa found within a single body site appear to be present across all time points, suggesting that no core temporal microbiome exists at high abundance (although some microbes may be present but drop below the detection threshold). Many more taxa appear to be persistent but non-permanent community members.

Conclusions: DNA sequencing and computational advances described here provide the ability to go beyond infrequent snapshots of our human-associated microbial ecology to high-resolution assessments of temporal variations over protracted periods, within and between body habitats and individuals. This capacity will allow us to define normal variation and pathologic states, and assess responses to therapeutic interventions.

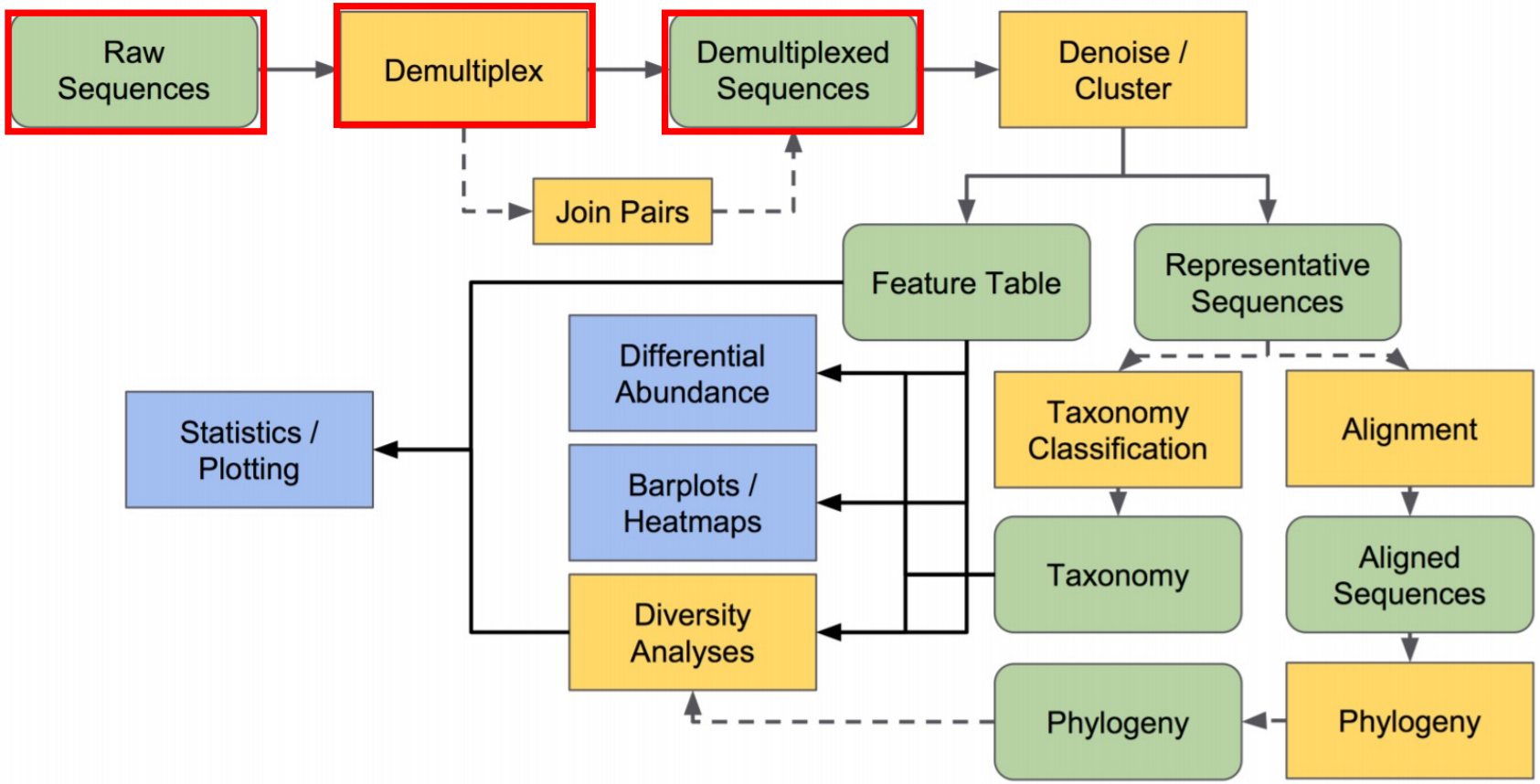
Moving pictures of the human microbiome





Time Points (days)

QIIME2 workflow

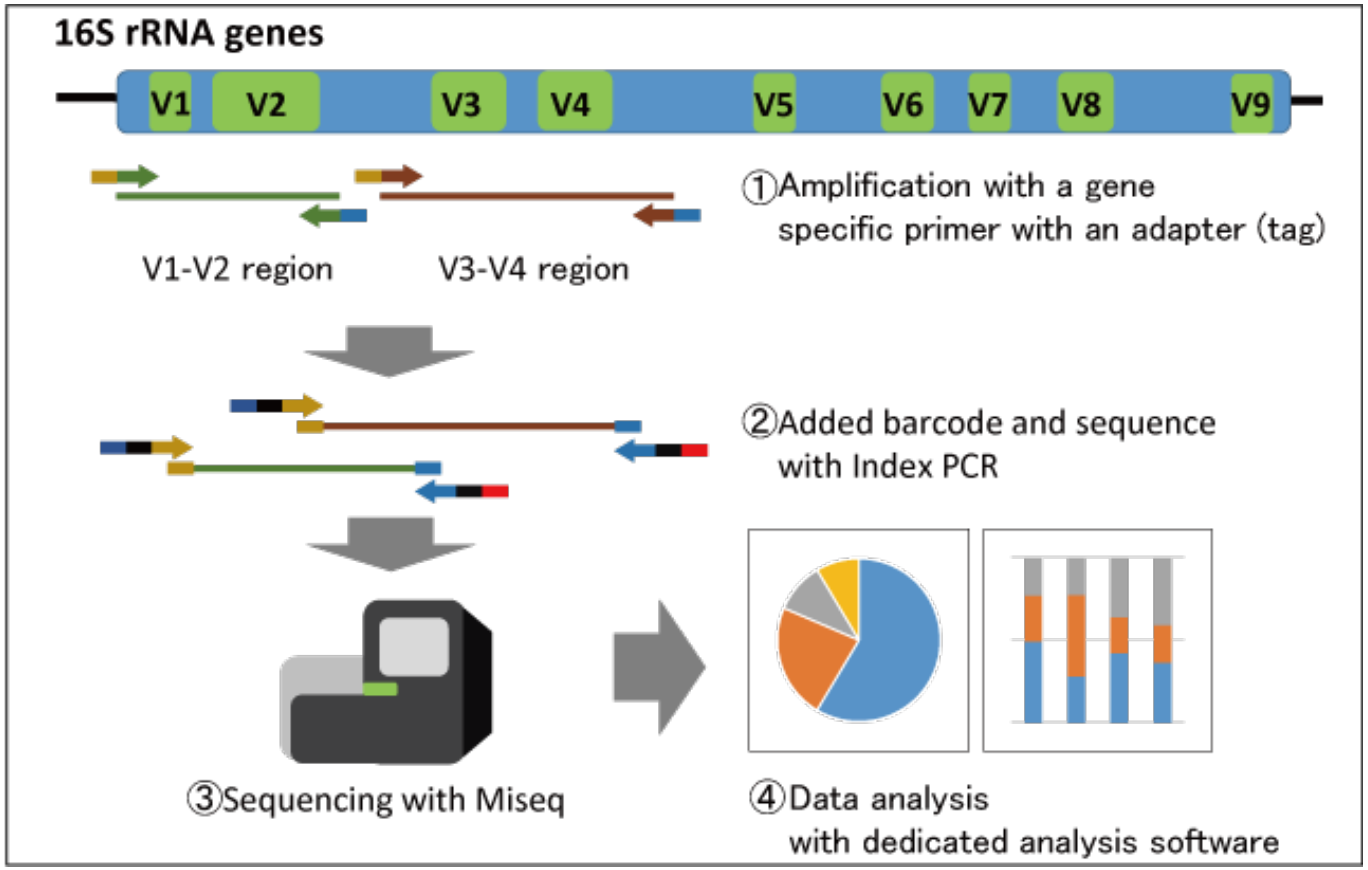


Yellow: processing steps
Green: inputs/outputs
Blue: R analysis

https://docs.qiime2.org/2018.8/tutorials/overview/#demultiplexing

Input: Raw sequences

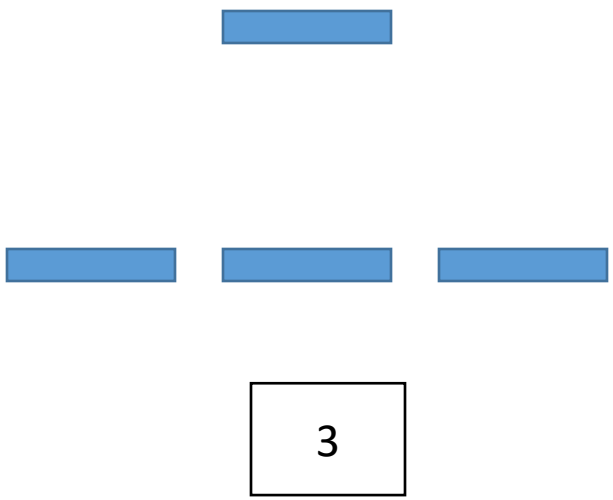
- Short reads: you set the parameter for the length (typically 150-300 bps long)
- Barcoded based on sample
- Covers some of the variable regions (9 in total, total of 1500bp long for 16S)



https://www.repertoire.co.jp/en/research/technology/16srrnainfo/

Clarification of Terminology

- Read: individual short sequences
- Library: all the reads per sample
- Sequencing depth: the size of the library, ie. library size



Sequence File Formats

Record name; starts with a "@"

fastq file:

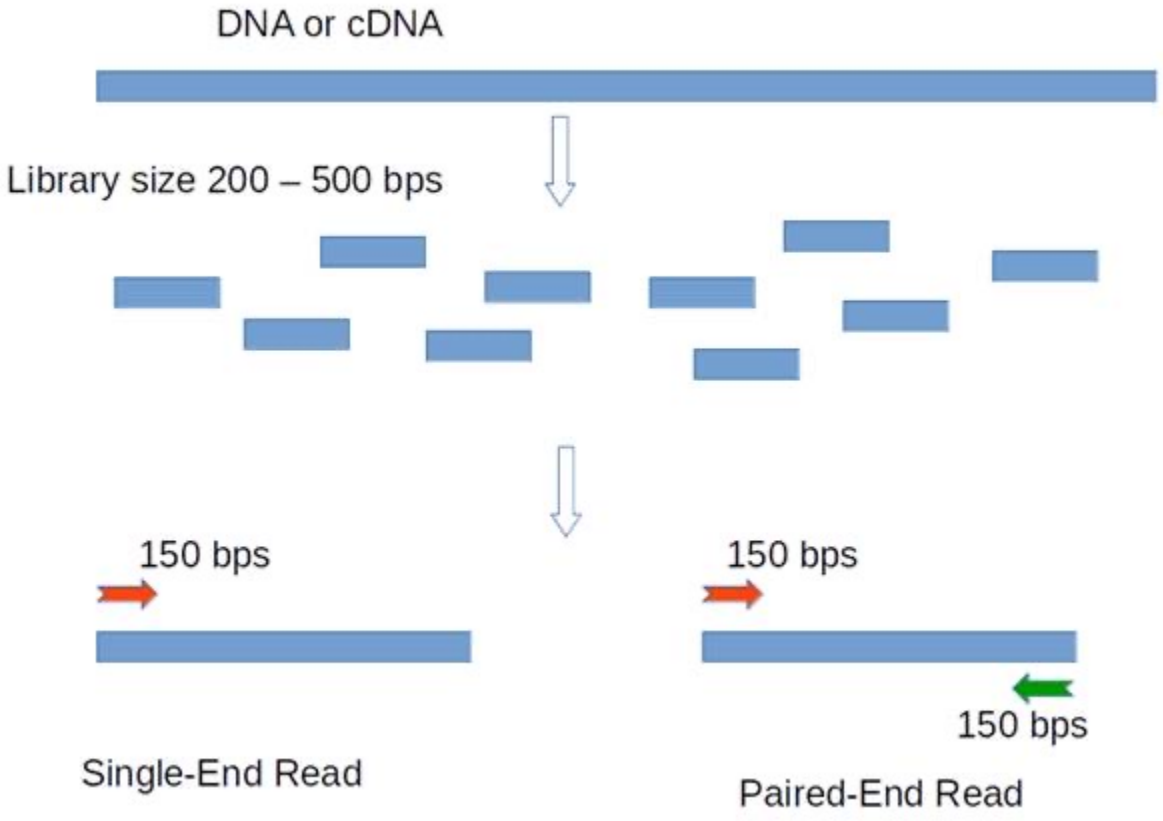
```
@sequence_1
TTTCCGGGGCACATAATCTTCAGCCGGGCGC
+
9C;=;=<9@4868>9:67AA<9>65<=>591
@sequence_1
TCAGCCGGGCCTTCAGCCGGGGCACATAATA
+
('%3(&&&%.....
```

DNA sequence

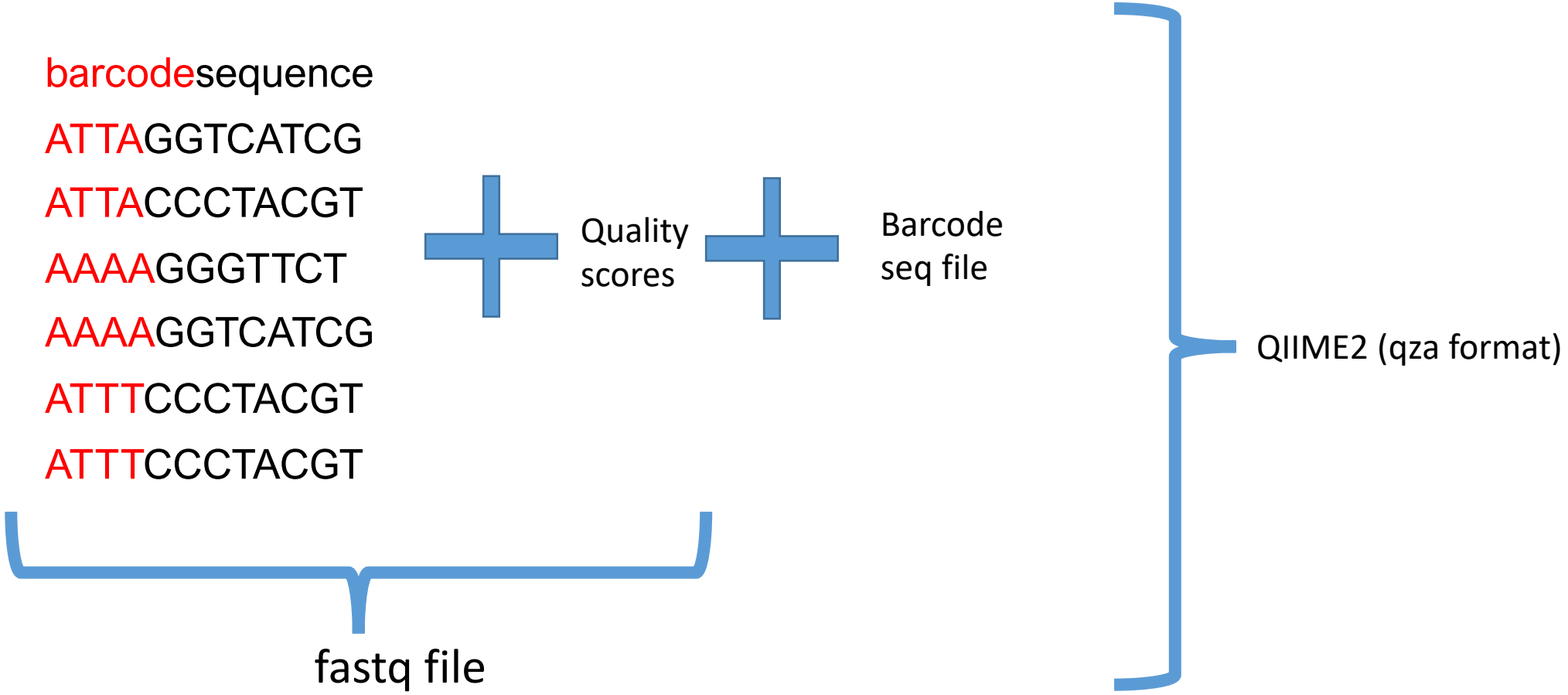
Phred score

Types of sequences

- **Single-end**: sequence from one end of the fragment only
- **Paired-end:** sequence from both end



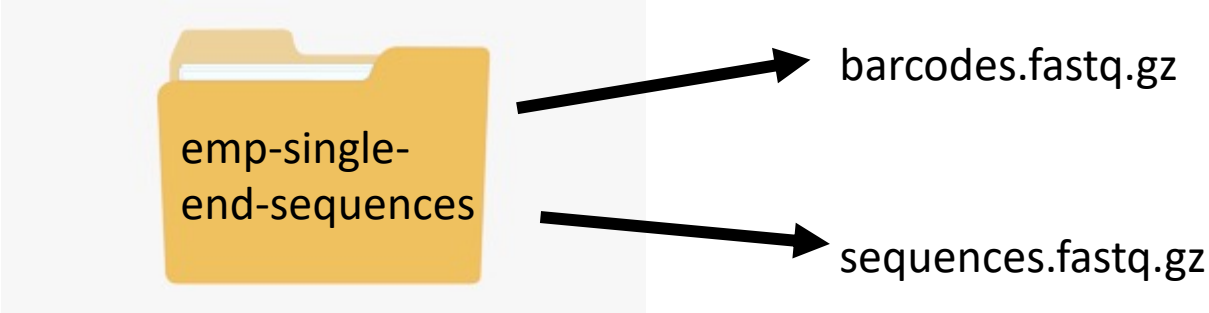
Importing your data to Qiime2 (visual)



Syntax of Qiime2 Commands

- Qiime2 commands start with `qiime`
- `\` refer to command continued on next line
- `--` indicates a verbose command
- `i` refers to input file
- `o` for output file
- `m` for metadata/information needed to process data

Dissecting the code




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

calling on the QIIME2 tool called ”import”
specifying what type of data we have
helping locate all the relevant files
naming our file output file (type qza)

Another way to import

- Manifest file: spreadsheet that details where all the files are located in your computer

```
sample-id       absolute-filepath
recip.220.WT.OB1.D7     $PWD/demultiplexed_seqs/10483.recip.220.WT.OB1.D7_30_L001_R1_001.fas
recip.290.ASO.OB2.D1    $PWD/demultiplexed_seqs/10483.recip.290.ASO.OB2.D1_27_L001_R1_001.fa
recip.389.WT.HC2.D21    $PWD/demultiplexed_seqs/10483.recip.389.WT.HC2.D21_1_L001_R1_001.fas
recip.391.ASO.PD2.D14   $PWD/demultiplexed_seqs/10483.recip.391.ASO.PD2.D14_5_L001_R1_001.fa
recip.391.ASO.PD2.D21   $PWD/demultiplexed_seqs/10483.recip.391.ASO.PD2.D21_1_L001_R1_001.fa
recip.391.ASO.PD2.D7    $PWD/demultiplexed_seqs/10483.recip.391.ASO.PD2.D7_15_L001_R1_001.fa
recip.400.ASO.HC2.D14   $PWD/demultiplexed_seqs/10483.recip.400.ASO.HC2.D14_32_L001_R1_001.f
recip.401.ASO.HC2.D7    $PWD/demultiplexed_seqs/10483.recip.401.ASO.HC2.D7_22_L001_R1_001.fa
recip.403.ASO.PD2.D21   $PWD/demultiplexed_seqs/10483.recip.403.ASO.PD2.D21_31_L001_R1_001.f
```

Dissecting the code

```
qiime tools import \
  --type "SampleData[SequencesWithQuality]" \
  --input-format SingleEndFastqManifestPhred33V2 \
  --input-path ./manifest.tsv \
  --output-path ./demux_seqs.qza
```

calling on the QIIME2 tool called ”import”

specifying what type of data we have (don’t change these two lines)

helping locate all the relevant files

naming our file output file (type qza)

Demultiplexing (visual)

AAAAGGTCATCG
ATTAGGTCATCG
ATTTCCCTACGT
AAAAGGGTTCT
ATTTCCCTACGT
ATTACCCTACGT

Sample 1
ATTAGGTCATCG
ATTACCCTACGT
Sample 2
AAAAGGGTTCT
AAAAGGTCATCG
Sample 3
ATTTCCCTACGT
ATTTCCCTACGT

Sample 1
GGTCATCG
CCCTACGT
Sample 2
GGGTTCT
GGTCATCG
Sample 3
CCCTACGT
CCCTACGT
`

Metadata file

#SampleID	age	age_units	collection_timestamp	description	geo_loc_na	host_comn	host_gravidity	l
11360.vole.1	2 to 5	years	2016-05-25 12:30	vole feces n	Ukraine	Bank vole	no	
11360.vole.10	2 to 5	years	2016-05-26 13:00	vole feces n	Ukraine	Bank vole	yes	
11360.vole.107	2 to 5	years	2016-05-28 17:08	vole feces n	Ukraine	Bank vole	no	
11360.vole.119	2 to 5	years	2016-05-29 12:10	vole feces n	Ukraine	Bank vole	no	
11360.vole.12	2 to 5	years	2016-05-26 12:50	vole feces n	Ukraine	Bank vole	no	
11360.vole.122	2 to 5	years	2016-05-29 11:59	vole feces n	Ukraine	Bank vole	no	
11360.vole.129	2 to 5	years	2016-05-29 12:31	vole feces n	Ukraine	Bank vole	yes	
11360.vole.130	10	years	2016-05-29 11:28	vole feces n	Ukraine	Bank vole	no	
11360.vole.131	2 to 5	years	2016-05-29 13:30	vole feces n	Ukraine	Bank vole	no	
11360.vole.133	10	years	2016-05-29 13:25	vole feces n	Ukraine	Bank vole	no	
11360.vole.134	2 to 5	years	2016-05-29 13:18	vole feces n	Ukraine	Bank vole	no	
11360.vole.135	2 to 5	years	2016-05-29 14:35	vole feces n	Ukraine	Bank vole	no	
11360.vole.137	2 to 5	years	2016-05-29 13:25	vole feces n	Ukraine	Bank vole	no	
11360.vole.138	1	years	2016-05-29 14:47	vole feces n	Ukraine	Bank vole	no	
11360.vole.14	2 to 5	years	2016-05-26 13:00	vole feces n	Ukraine	Bank vole	no	
11360.vole.140	2 to 5	years	2016-05-30 13:40	vole feces n	Ukraine	Bank vole	no	
11360.vole.146	2 to 5	years	2016-05-30 18:00	vole feces n	Ukraine	Bank vole	yes	
11360.vole.149	1	years	2016-05-31 14:10	vole feces n	Ukraine	Bank vole	no	
11360.vole.15	2 to 5	years	2016-05-26 13:00	vole feces n	Ukraine	Bank vole	no	
11360.vole.150	1	years	2016-05-31 14:05	vole feces n	Ukraine	Bank vole	no	
11360.vole.159	1	years	2016-06-01 14:55	vole feces n	Ukraine	Bank vole	no	
11360.vole.162	2 to 5	years	2016-06-01 15:25	vole feces n	Ukraine	Bank vole	no	

Demultiplexing Code

```
qiime demux emp-single \
--i-seqs emp-single-end-sequences.qza \
--m-barcodes-file sample-metadata.tsv \
--m-barcodes-column barcode-sequence \
--o-per-sample-sequences demux.qza \
--o-error-correction-details demux-details.qza
```

calling on the QIIME2 tool called ”demux emp-single”

inputting your imported file

inputting your metadata file

and indicating which column refers to the barcodes

two outputs: demux.qza and demux-details.qza

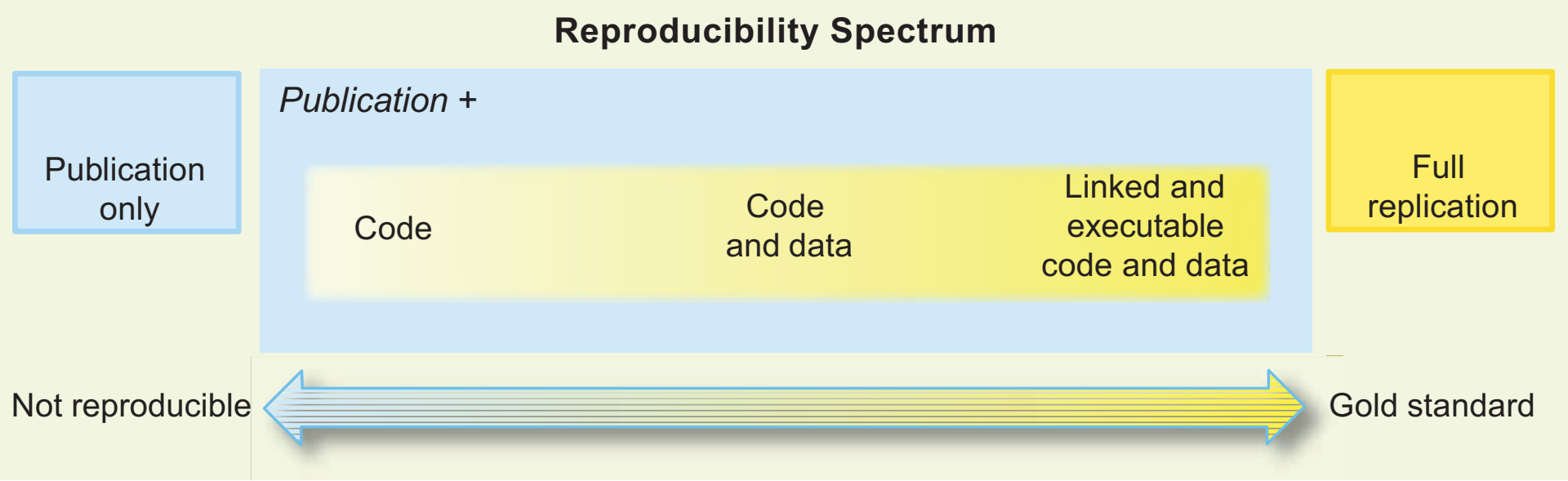
```
qiime demux summarize \
--i-data demux.qza \
--o-visualization demux.qzv
```

converting your qza file to a qzv (visualization file)

Documentation

Scripts and Lab Notebook

Importance of Reproducibility



Peng RD. 2011. Science 334:1226–1227.

Anatomy of shell script

```
#!/bin/bash
```

Indicates which shell to use to interpret the script

```
# Comments start with a hash `#` and are
not executed by the shell
<command>

# Comments often provide context for
commands
<command>
```

Anatomy of shell script

```
#!/bin/bash

# Create a directory for project and navigate
to it
mkdir /data/moving_pictures_tutorial
cd /data/moving_pictures_tutorial

# Import data while working directory is
`/data/moving_pictures_tutorial`
qiime tools import \
 --type EMPSingleEndSequences \
 --input-path
 /mnt/datasets/moving_pictures/emp-single-end-
 sequences \
 --output-path emp-single-end-sequences.qza
```

Edit scripts in plain text with Notepad (Windows 10) or TextEdit (macOS) or on R as a shell script

Script formats

- Using R studio to document scripts as proper scripts
  - Shell script files end with extension *.sh*
  - R script files end with *.R*
- Publication ready scripts
  - Don’t contain parts that are intended for personal use
    - Eg. Transferring files from server to local computer

Digital Lab Notebook

- Internal documentation
- Day-by-day experimentation/activity: What analysis you ran? Who ran the analysis? Reference to the script. Any issues that arise and how you did or did not resolve it
- Organized well enough so that another team can pick up your notebook and understand how you arrived at your findings
- Where to keep your notebook?
  - Consider somewhere where all members can have access to. Share drive.