Amplicon Sequencing Data Analysis with Qiime 2

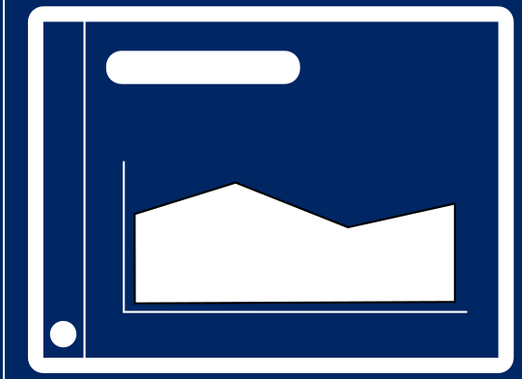
Christian Diener

from the **ISB Microbiome Course 2020**

 cdiener.com cdiener @thaasophobia

Organization of the course

Presentation

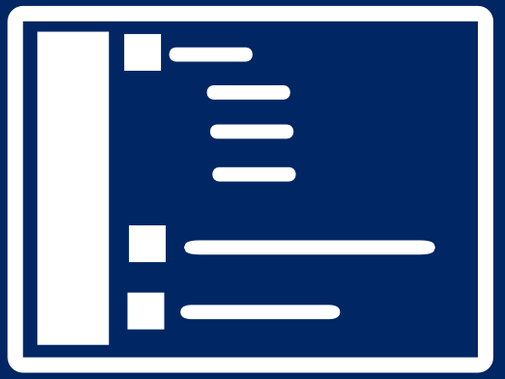


logic
explanations
links

Notebook

technical aspects
materials
visualizations

Chat



support
Q&A

Hold your horses 

Let's get the slides first (use your computer, phone, TV, fridge, anything with a 16:9 screen)

https://gibbons-lab.github.io/isb_course_2020/16S

Qiime

Created ~2010 during the Human Microbiome Project (2007 - 2016) under leadership of Greg Caporaso and Rob Knight.

What is Qiime?

> *QIIME 2 is a powerful, extensible, and decentralized microbiome analysis package with a focus on data and analysis transparency.*

Quantitative insights into Microbial Ecology

So what is it really?

In its essence Qiime is a set of commands to transform microbiome data into intermediate outputs and visualizations.

```
cdiener@moneta [ubc2018]
```

It's commonly used via the command line.

Qiime 2 was introduced 2016 and improves on Qiime 1 based on the experiences during the HMP.

Major changes:

- integrated tracking of data provenance
- semantic type system
- extendable plugin system
- multiple user interfaces (in progress)

Where to find help?

Qiime 2 comes with a lot of help starting from https://qiime2.org such as tutorials, general documentation and a user forum to ask questions.

Artifacts, actions and visualizations

Qiime 2 manages artifacts which is basically intermediate data that is fed to actions to either produce other artifacts or visualizations.

Remember

Artifacts can be intermediate steps, but Visualizations are end points meant for human consumption ☝ .

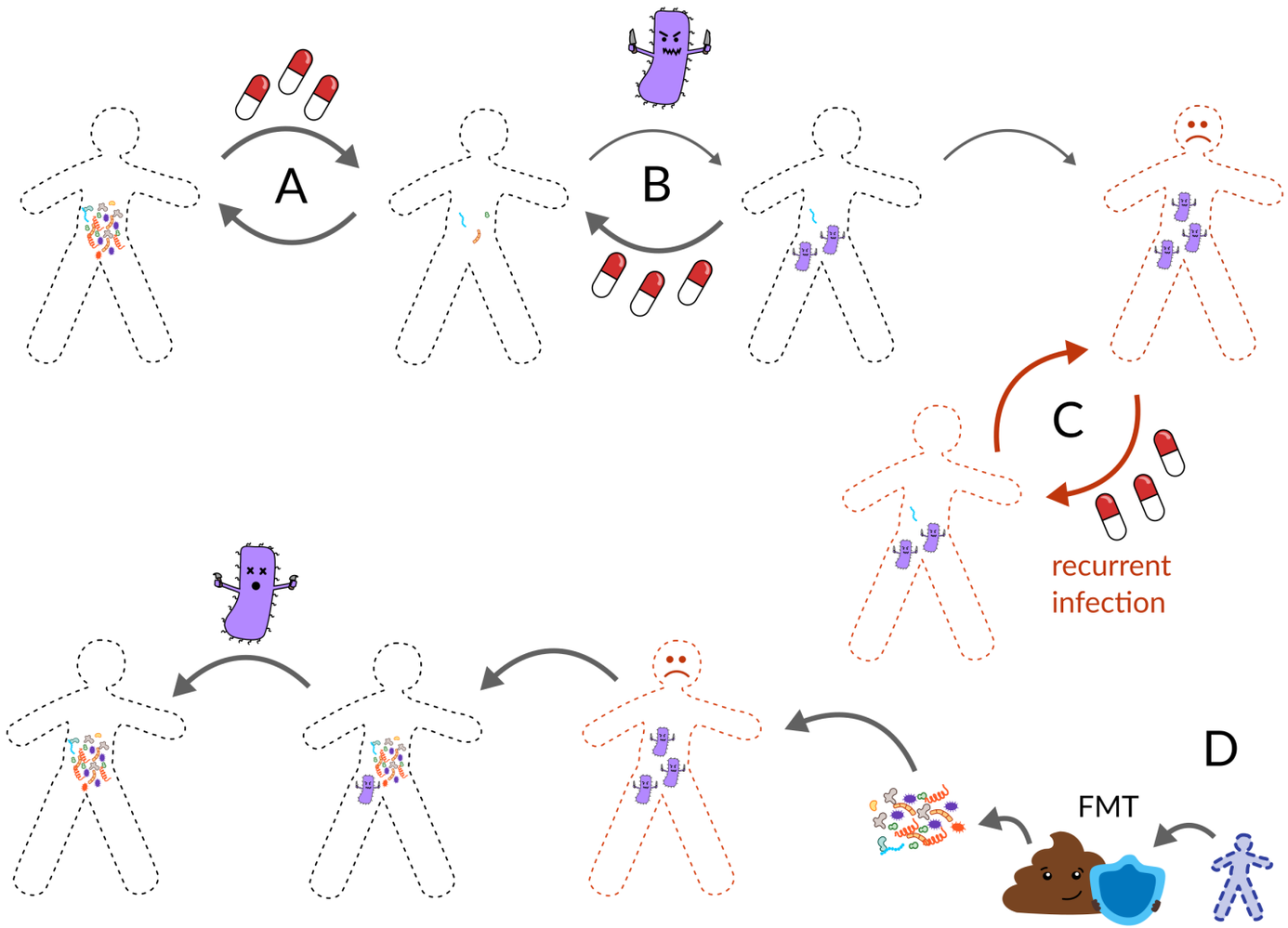
Analyzing the microbial composition during recurrent *C. diff* infection

16S amplicon sequencing data of the V4 region from fecal samples

4 healthy donors and 4 individuals with recurrent infection.

https://doi.org/10.1186/s40168-015-0070-0

The *C. diff* infection cycle



courtesy of Stephanie Swegle

Setup

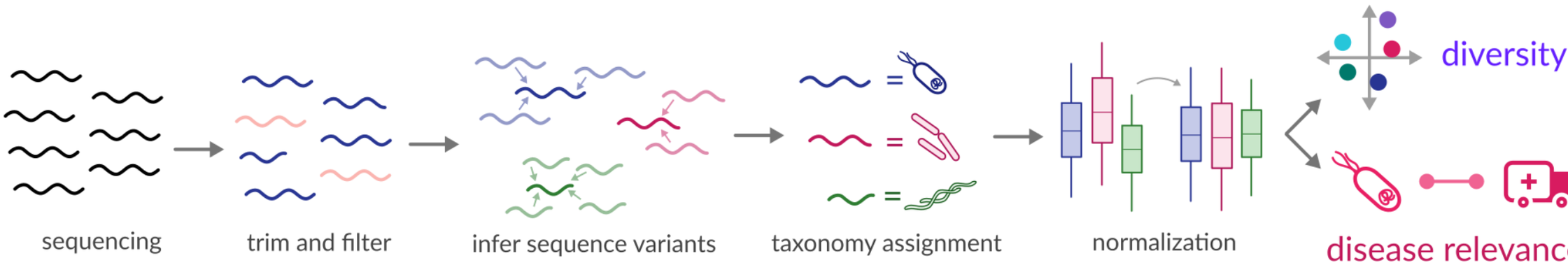
 Let’s switch to the notebook and get started

Wait... what?

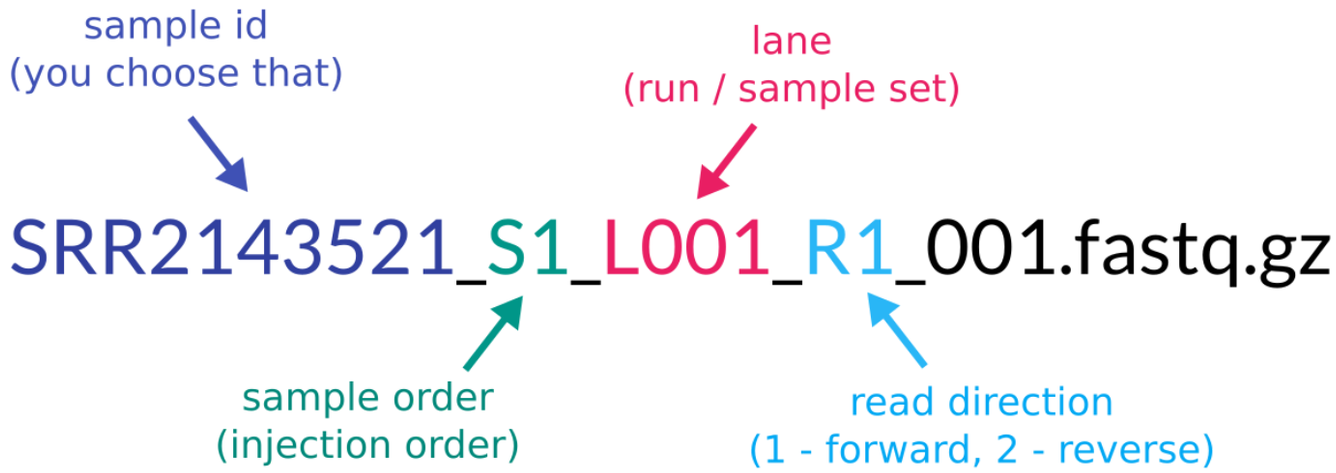
All output we generate can be found in the `treasure_chest` folder at

https://github.com/gibbons-lab/isb_course_2020/treasure_chest

What will we do today?



Illumina FastQ files (Basespace)




```
@SRR2143527.13917 13917 length=251
TACGTAGGTGGCGAGCGTTATCCGGAATTATTGGGCGTAAA...
+
BBBBAF?A@D2BEEEGGGFGGGHGGGCFGFHHCFHCEFGGH...
```

We have our raw sequencing data but Qiime 2 only operates on artifacts. How do we convert our data to an artifact?

Our first Qiime 2 commands

We can import the data with the `import` action. For that we have to give Qiime 2 a manifest (list of raw files) and tell it what type of data we are importing and what type of artifact we want.

 let's jump back to the open Colaboratory notebook...

View a Qiime 2 visualization

There are two ways to look at a Qiime 2 visualization:

- visit https://view.qiime2.org and load the file
- use `qiime tools view [file.qzv]` if you have Qiime 2 installed

🤔 What do you observe across the read? Where would you truncate the reads?

Qiime 2 commands can become pretty long. Here some pointers to remember the structure of a command:

```
qiime plugin action --i-argument1 ... --o-argument2 ...
```

Argument types usually begin with a letter denoting their meaning:

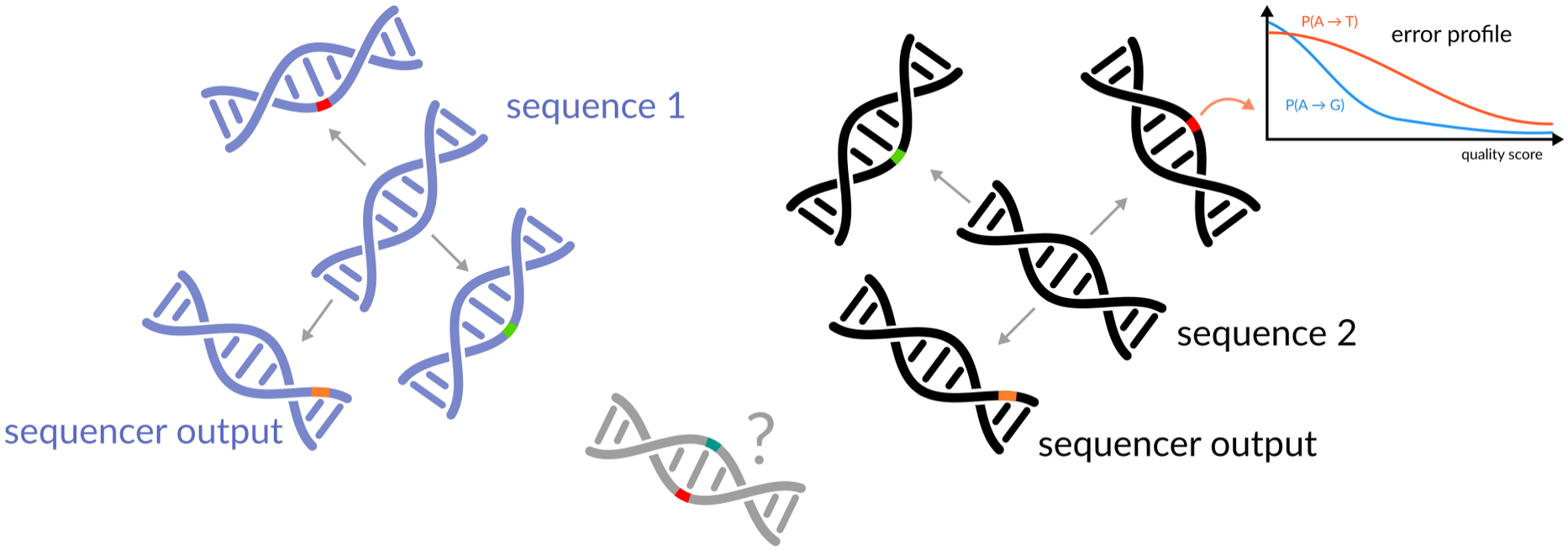
- `--i-...` = input files
- `--o-...` = output files
- `--p-...` = parameters
- `--m-...` = metadata

Time to bring in the big guns 💣 ⚡

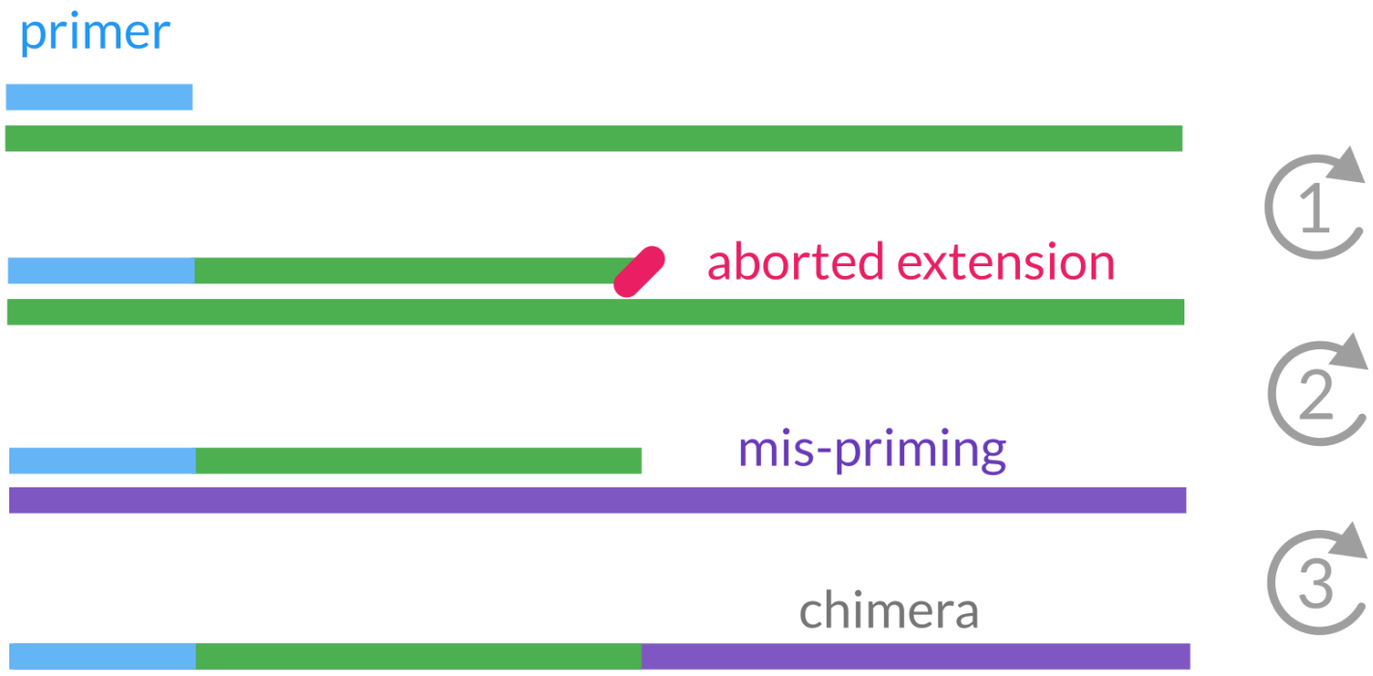
We will now run the DADA2 plugin which will do 3 things:

1. filter and trim the reads
2. find the most likely original sequences in the sample (ASVs)
3. remove chimeras
4. count the abundances

Identifying alternative sequence variants (ASVs)



PCR chimeras



We now have a table containing the counts for each ASV in each sample. We also have a list of ASV

🤔 Do you have an idea what we could do with those two data sets? What quantities might we be interested in?

Relationship between ASVs

One of the basic things we might want to see is how the sequences across all samples are related t
one another. We are interested in their phylogeny.

You can visualize your tree using iTOL (https://itol.embl.de/).

Diversity

In microbial community analysis we are usually interested in two different diversity quantities, **alp**
diversity and **beta diversity**.

Alpha diversity

How diverse is a single sample?

- how many taxa do we observe (richness)? → #observed taxa
- are taxa equally abundant or are there rare/dominant taxa? → Shannon, Evenness

Beta diversity

How different are two or more samples/donors/sites from each other?

- how many taxa are shared between samples? → Jaccard index
- do shared taxa have the same abundance? → Bray-Curtis distance
- do samples share genetically similar taxa? → UniFrac, Faith PD

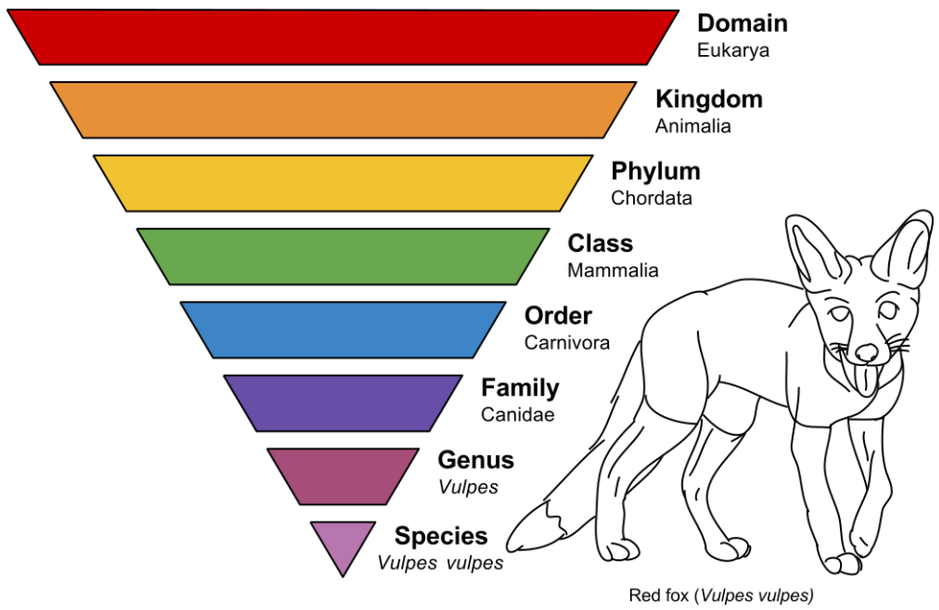
Principal Coordinate Analysis

But what organisms are there in our sample?

We are still just working with sequences and have no idea what **organisms** those correspond to.

🤔 What would you do to go from a sequence to an organism/bacteria?

Taxonomic ranks



Even though just looking for our sequence in a database of known genes seems like the best idea that does not work great in practice. Why?

More elaborate methods use subsequences (k-mers) and their counts to predict the lineage/taxonomy with machine learning methods. For 16S amplicon fragments this provides bette generalization.

Your turn

What is the relationship between particular taxa and the disease state?

And we are done 👏

Thanks!