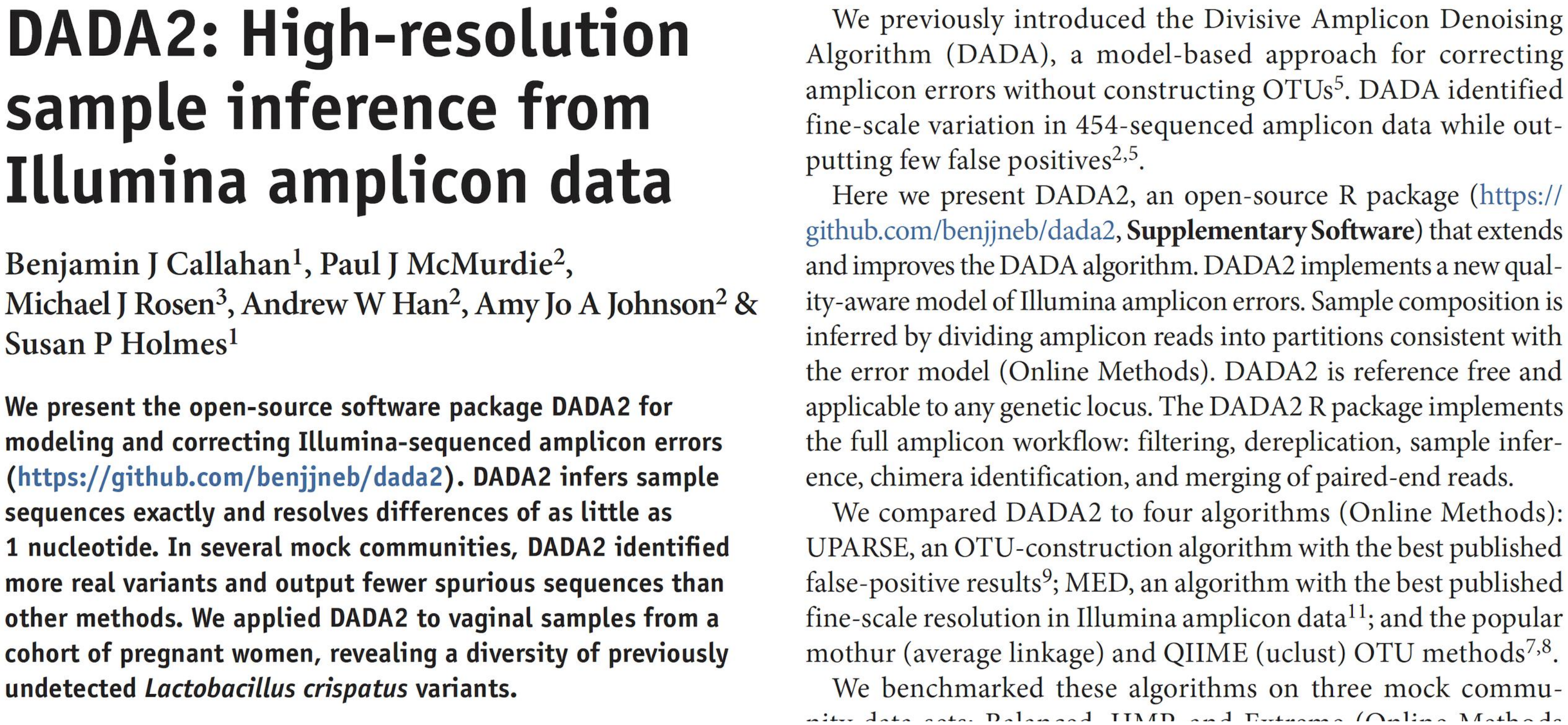
Microbiome workshop

dada2 workflow (fastq to ASV)

Anujit Sarkar
Postdoctoral Scholar
COPH, CON
Genomics Program, USF
June 11, 2020

Analysis of 16S microbiome (fastq to ASV table or bacterial abundance table)

- Fastq files are obtained immediately after 16S rRNA sequencing
- We will analyze the fastq files using dada2 (https://github.com/benjjneb/dada2)



DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as 1 nucleotide. In several mock communities, DADA2 identified more real variants and output fewer spurious sequences than other methods. We applied DADA2 to vaginal samples from a cohort of pregnant women, revealing a diversity of previously undetected *Lactobacillus crispatus* variants.

We previously introduced the Divisive Amplicon Denoising Algorithm (DADA), a model-based approach for correcting amplicon errors without constructing OTUs[5]. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives[2,5].

Here we present DADA2, an open-source R package (https://github.com/benjjneb/dada2, **Supplementary Software**) that extends and improves the DADA algorithm. DADA2 implements a new quality-aware model of Illumina amplicon errors. Sample composition is inferred by dividing amplicon reads into partitions consistent with the error model (Online Methods). DADA2 is reference free and applicable to any genetic locus. The DADA2 R package implements the full amplicon workflow: filtering, dereplication, sample inference, chimera identification, and merging of paired-end reads.

We compared DADA2 to four algorithms (Online Methods): UPARSE, an OTU-construction algorithm with the best published false-positive results[9]; MED, an algorithm with the best published fine-scale resolution in Illumina amplicon data[11]; and the popular mothur (average linkage) and QIIME (uclust) OTU methods[7,8].

We benchmarked these algorithms on three mock community data sets: Balanced, HMP and Extreme (Online Methods

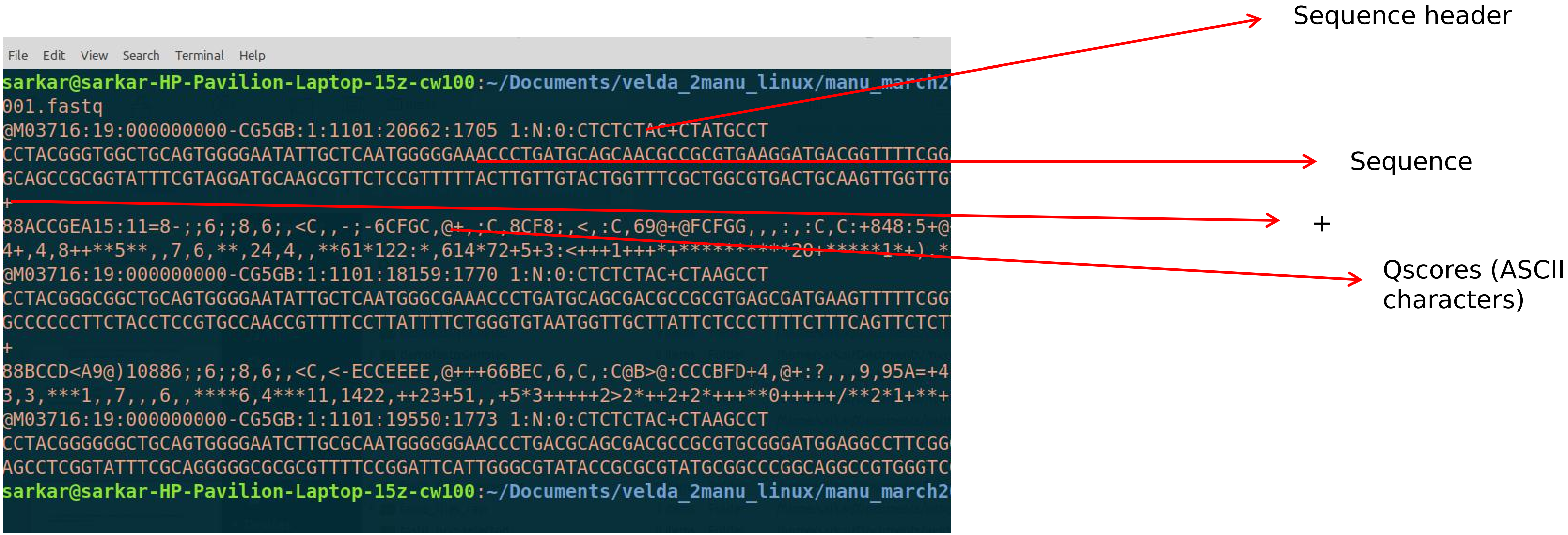
Install dada2

-
  ```
  if (!requireNamespace("BiocManager", quietly=TRUE)) install.packages("BiocManager")
  BiocManager::install("dada2")
  ```

-
  ```
  > packageVersion("dada2")
  ```

-
  ```
  > library(dada2)
  ```

Purpose of this task



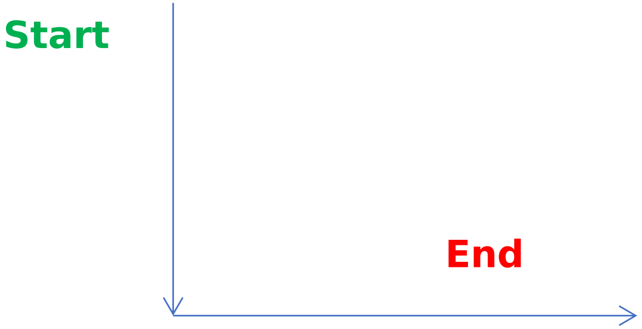
```
sarkar@sarkar-HP-Pavilion-Laptop-15z-cw100:~/Documents/velda_2manu_linux/manu_march2
001.fastq
@M03716:19:000000000-CG5GB:1:1101:20662:1705 1:N:0:CTCTCTAC+CTATGCCT
CCTACGGGTGGCTGCAGTGGGGAATATTGCTCAATGGGGGAAACCCTGATGCAGCAACGCCGCGTGAAGGATGACGGTTTTCGG
GCAGCCGCGGTATTTCGTAGGATGCAAGCGTTCTCCGTTTTTACTTGTTGTACTGGTTTCGCTGGCGTGACTGCAAGTTGGTTG
+
88ACCGEA15:11=8-;;6;;8,6;,<C,,-;-6CFGC,@+,;C,8CF8;,<,:C,69@+@FCFGG,,,:,:C,C:+848:5+@
4+,4,8++**5**,,7,6,**,24,4,,**61*122:*,614*72+5+3:<+++1+++*+**********20;*****1*+),*
@M03716:19:000000000-CG5GB:1:1101:18159:1770 1:N:0:CTCTCTAC+CTAAGCCT
CCTACGGGCGGCTGCAGTGGGGAATATTGCTCAATGGGCGAAACCCTGATGCAGCGACGCCGCGTGAGCGATGAAGTTTTTCGG
GCCCCCCTTCTACCTCCGTGCCAACCGTTTTCCTTATTTTCTGGGTGTAATGGTTGCTTATTCTCCCTTTTCTTTCAGTTCTCT
+
88BCCD<A9@)10886;;6;;8,6;,<C,<-ECCEEEE,@+++66BEC,6,C,:C@B>@:CCCBFD+4,@+:?,,,9,95A=+4
3,3,***1,,7,,,6,,****6,4***11,1422,++23+51,,+5*3+++++2>2*++2+2*+++**0+++++/**2*1+**+
@M03716:19:000000000-CG5GB:1:1101:19550:1773 1:N:0:CTCTCTAC+CTAAGCCT
CCTACGGGGGGCTGCAGTGGGGAATCTTGCGCAATGGGGGAACCCTGACGCAGCGACGCCGCGTGCGGGATGGAGGCCTTCGG
AGCCTCGGTATTTCGCAGGGGGCGCGCGTTTTCCGGATTCATTGGGCGTATACCGCGCGTATGCGGCCCGGCAGGCCGTGGGTC
sarkar@sarkar-HP-Pavilion-Laptop-15z-cw100:~/Documents/velda_2manu_linux/manu_march2
```

Sequence header

Sequence

+

Qscores (ASCII characters)



Sample	Streptococcus	Veilonella	Prevotella
Sample1	25	4	45
Sample2	14	0	25
Sample3	42	32	0

What do you need before starting the analysis

- R and Rstudio with dada2 installed

- Demultiplexed paired-end fastq files (preferably from Illumina, for this workshop) stored in a folder/directory

- An empty folder where all your results will be exported

- A 16S rRNA database (Greengenes, Silva or 16S RDP) downloaded and stored in a folder

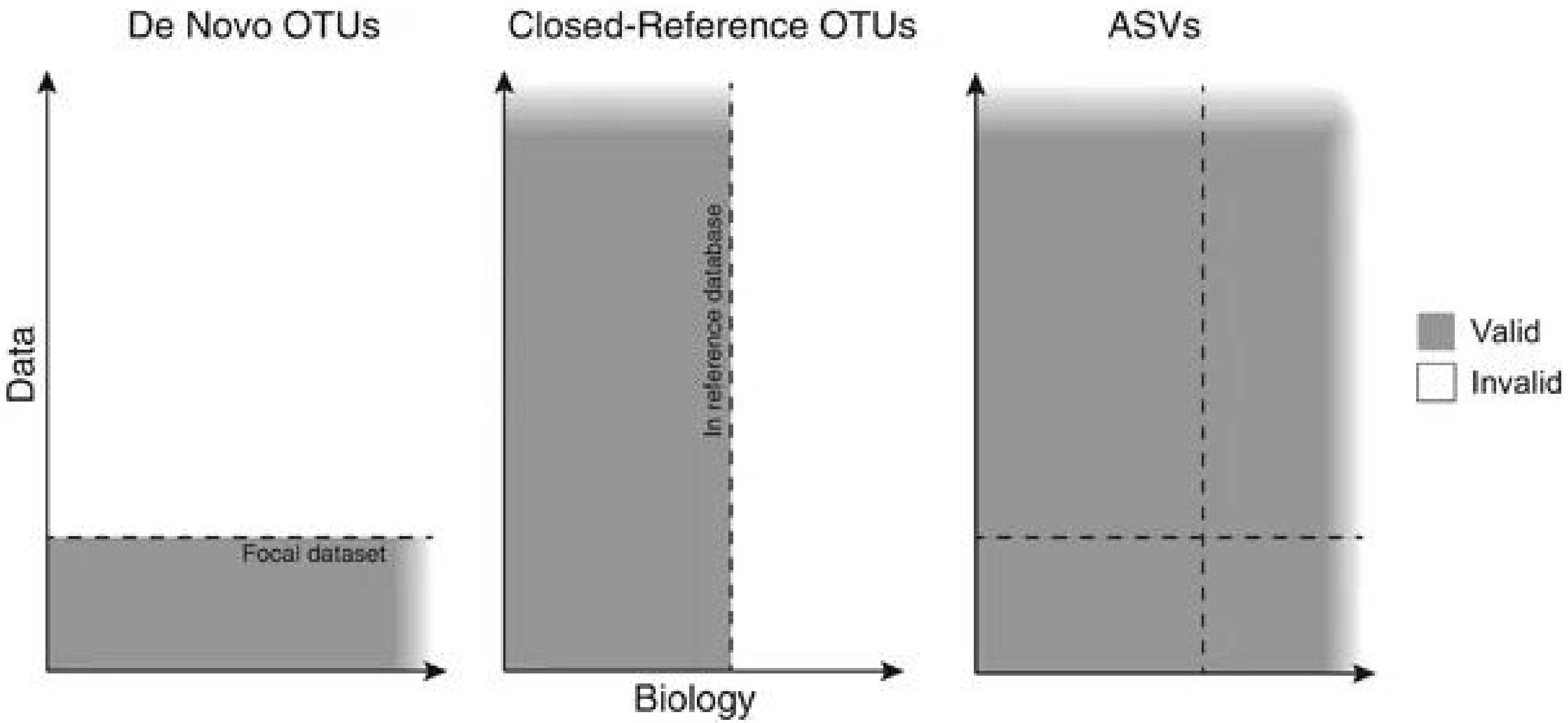
- The path of all the files and folders mentioned above

Major steps for analysis (all in RStudio)

- Setup your environment for the analysis
- Apply quality filters to discard bad sequences
- Learn error rates from your data
- Infer Amplicon sequence variants from your forward and reverse sequences
- Merge your paired-end filtered sequences
- Make a table of the sequence variants (ASVs) in your data
- Remove chimeric sequences
- Track your workflow to monitor loss of sequences
- Assign taxonomy to each ASV based on reference database
- Save ASV taxonomy, ASV sequences and ASV distribution in your samples
- Rarefy ASV table to equal depths (optional)
- Remove ASVs whose total count is zero (optional)

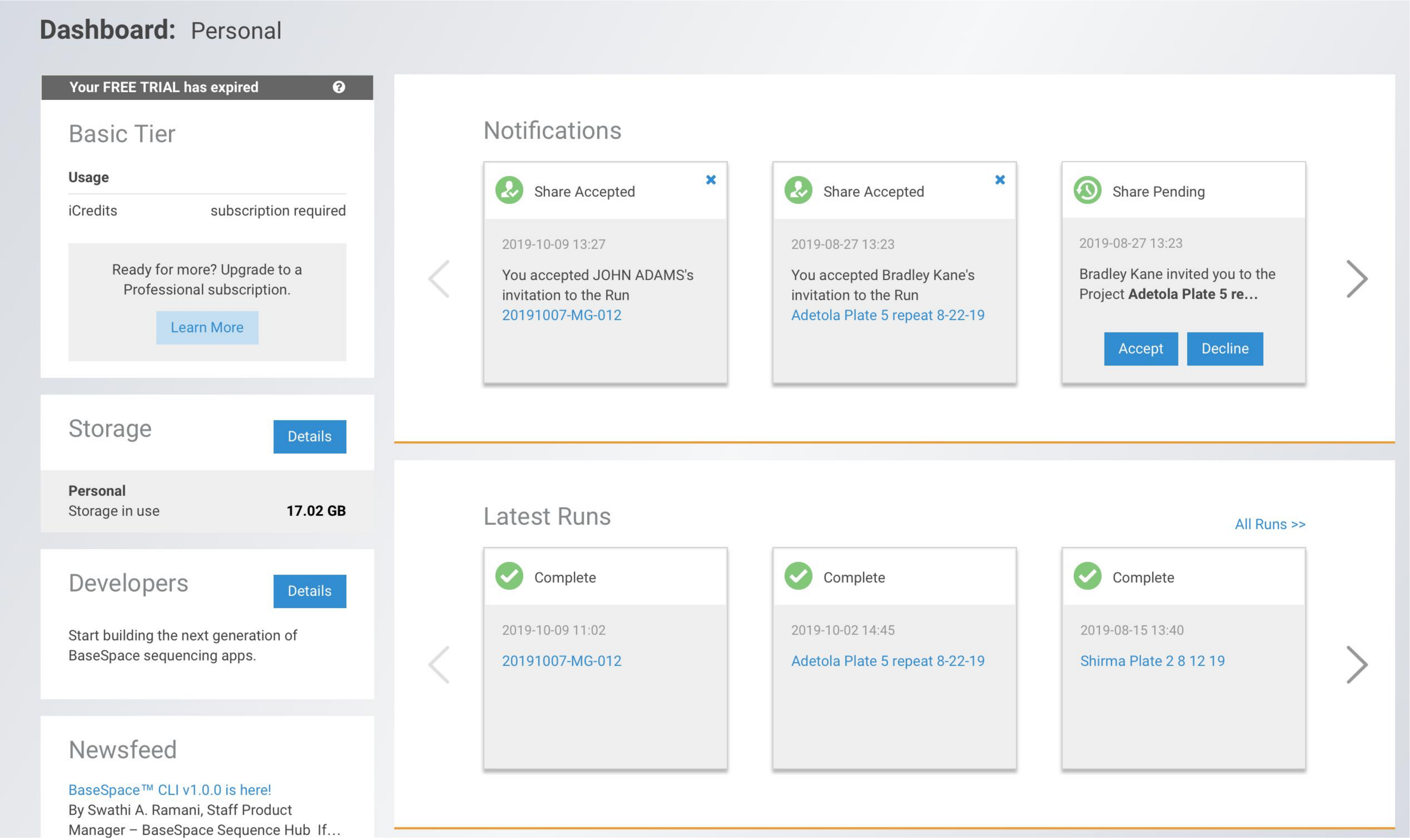
OTU vs ASV

- ASVs are truly of biological origin
- ASVs can identify up to single nucleotide differences



https://www.nature.com/articles/ismej2017119#Sec2

Get data from Illumina BaseSpace



Dashboard: Personal

Your FREE TRIAL has expired

Basic Tier

Usage

iCredits	subscription required

Ready for more? Upgrade to a Professional subscription.

Learn More

Storage

Details

Personal

Storage in use	**17.02 GB**

Developers

Details

Start building the next generation of BaseSpace sequencing apps.

Newsfeed

BaseSpace™ CLI v1.0.0 is here!

By Swathi A. Ramani, Staff Product Manager – BaseSpace Sequence Hub If…

Notifications

Share Accepted

2019-10-09 13:27

You accepted JOHN ADAMS's invitation to the Run 20191007-MG-012

Share Accepted

2019-08-27 13:23

You accepted Bradley Kane's invitation to the Run Adetola Plate 5 repeat 8-22-19

Share Pending

2019-08-27 13:23

Bradley Kane invited you to the Project **Adetola Plate 5 re...**

Accept Decline

Latest Runs

All Runs >>

Complete

2019-10-09 11:02

20191007-MG-012

Complete

2019-10-02 14:45

Adetola Plate 5 repeat 8-22-19

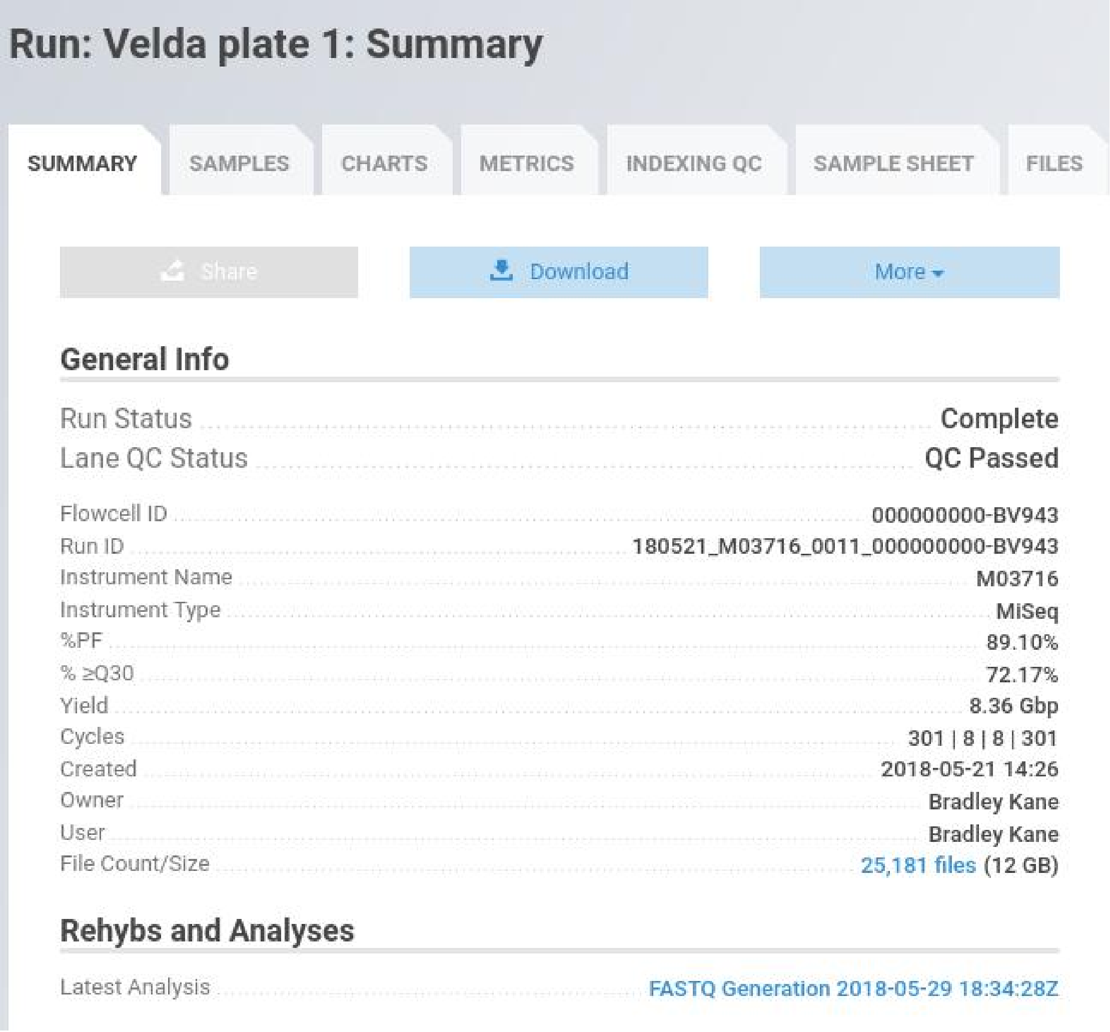
Complete

2019-08-15 13:40

Shirma Plate 2 8 12 19

https://login.illumina.com/platform-services-manager/?rURL=https://basespace.illumina.com&clientId=basespace&clientVars=aHR0cHM6Ly9iYXNlc3BhY2UuaWxsdW1pbmEuY29tL2Rhc2hib2FyZA&redirectMethod=GET#/

Summary of sequencing run in BaseSpace



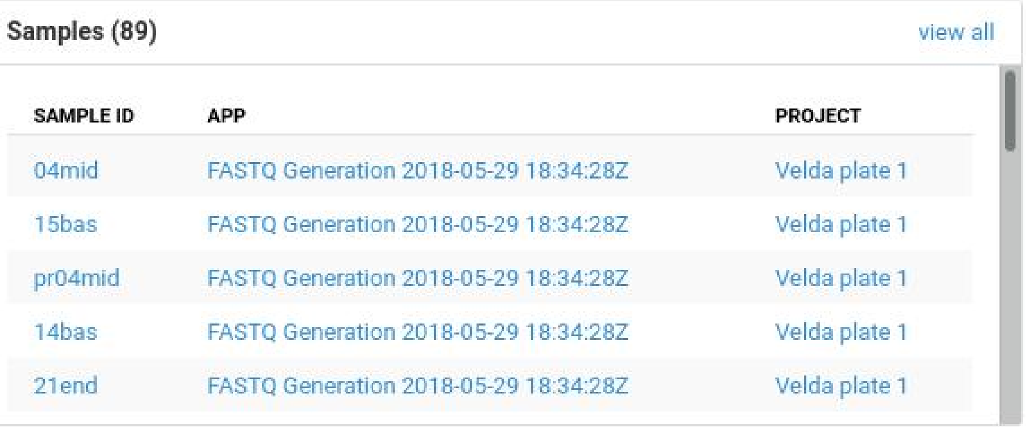
Run: Velda plate 1: Summary

General Info

Run Status	Complete
Lane QC Status	QC Passed
Flowcell ID	000000000-BV943
Run ID	180521_M03716_0011_000000000-BV943
Instrument Name	M03716
Instrument Type	MiSeq
%PF	89.10%
% ≥Q30	72.17%
Yield	8.36 Gbp
Cycles	301 | 8 | 8 | 301
Created	2018-05-21 14:26
Owner	Bradley Kane
User	Bradley Kane
File Count/Size	25,181 files (12 GB)

Rehybs and Analyses

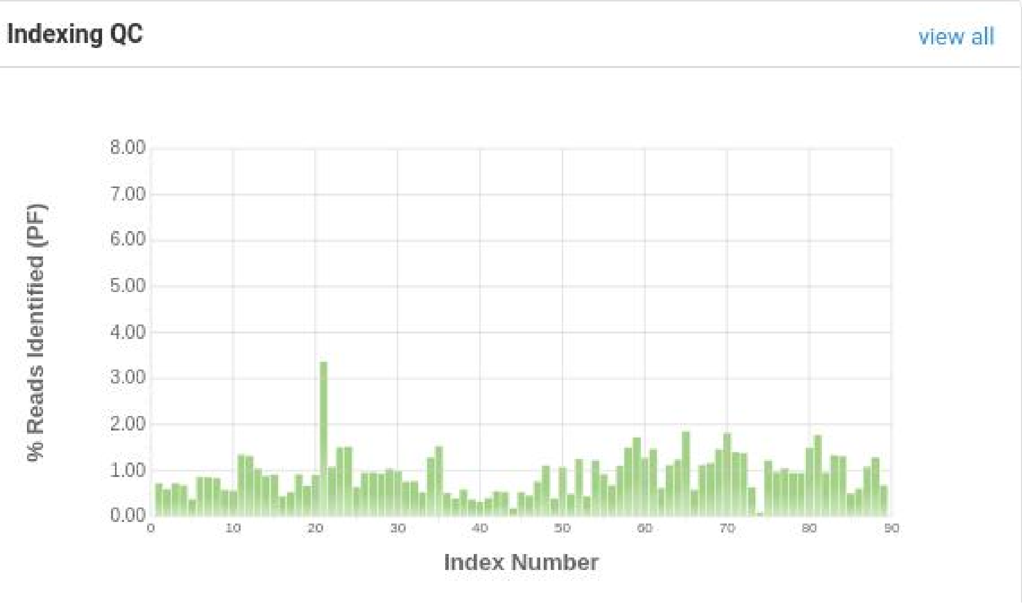
Latest Analysis	FASTQ Generation 2018-05-29 18:34:28Z



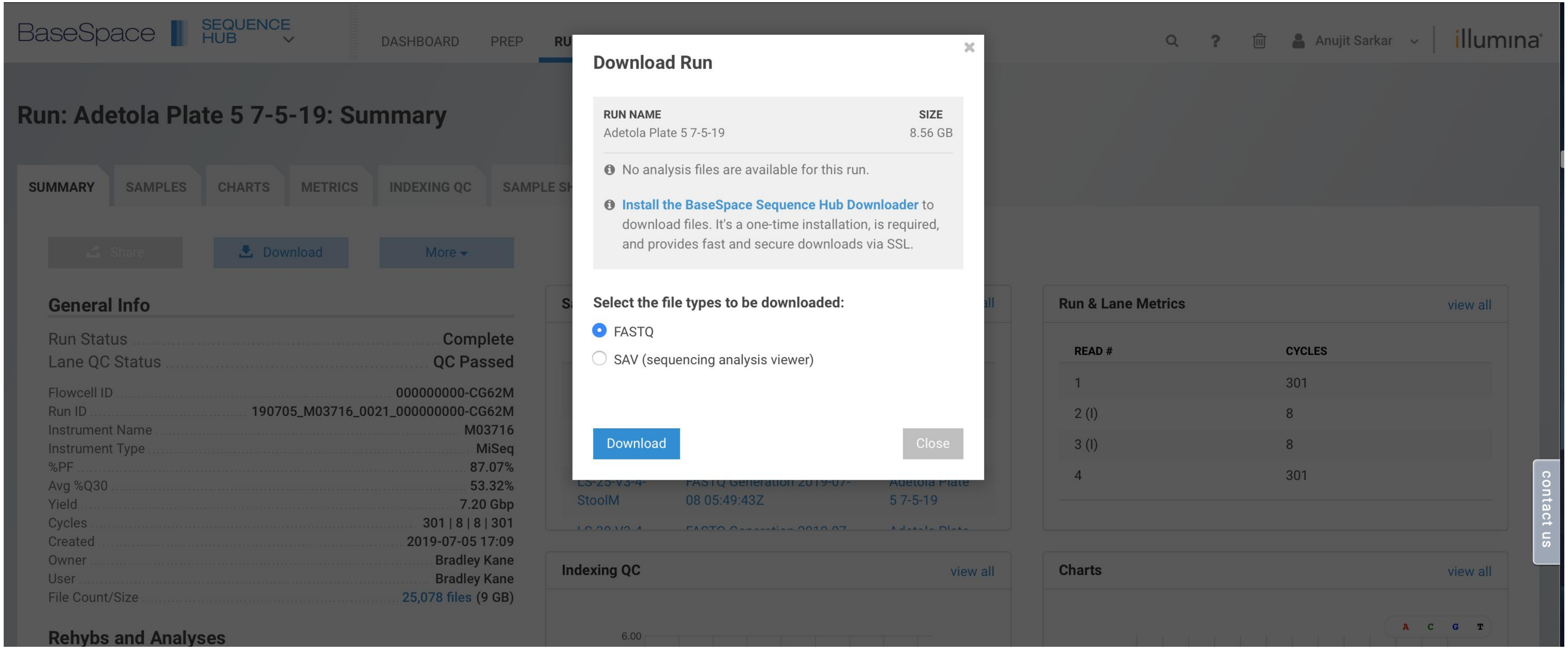
Samples (89)

SAMPLE ID	APP	PROJECT
04mid	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
15bas	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
pr04mid	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
14bas	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
21end	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1

Indexing QC



Download Fastq files



Run: Adetola Plate 5 7-5-19: Summary

Download Run

RUN NAME	SIZE
Adetola Plate 5 7-5-19	8.56 GB

No analysis files are available for this run.

Install the BaseSpace Sequence Hub Downloader to download files. It's a one-time installation, is required, and provides fast and secure downloads via SSL.

Select the file types to be downloaded:

● FASTQ

○ SAV (sequencing analysis viewer)

Download   Close

General Info

Run Status	Complete
Lane QC Status	QC Passed
Flowcell ID	000000000-CG62M
Run ID	190705_M03716_0021_000000000-CG62M
Instrument Name	M03716
Instrument Type	MiSeq
%PF	87.07%
Avg %Q30	53.32%
Yield	7.20 Gbp
Cycles	301 | 8 | 8 | 301
Created	2019-07-05 17:09
Owner	Bradley Kane
User	Bradley Kane
File Count/Size	25,078 files (9 GB)

Rehybs and Analyses

LS-25-V3-4-StoolM FASTQ Generation 2019-07-08 05:49:43Z Adetola Plate 5 7-5-19

Run & Lane Metrics

READ #	CYCLES
1	301
2 (I)	8
3 (I)	8
4	301

Indexing QC

Charts

Setup your environment for the analysis

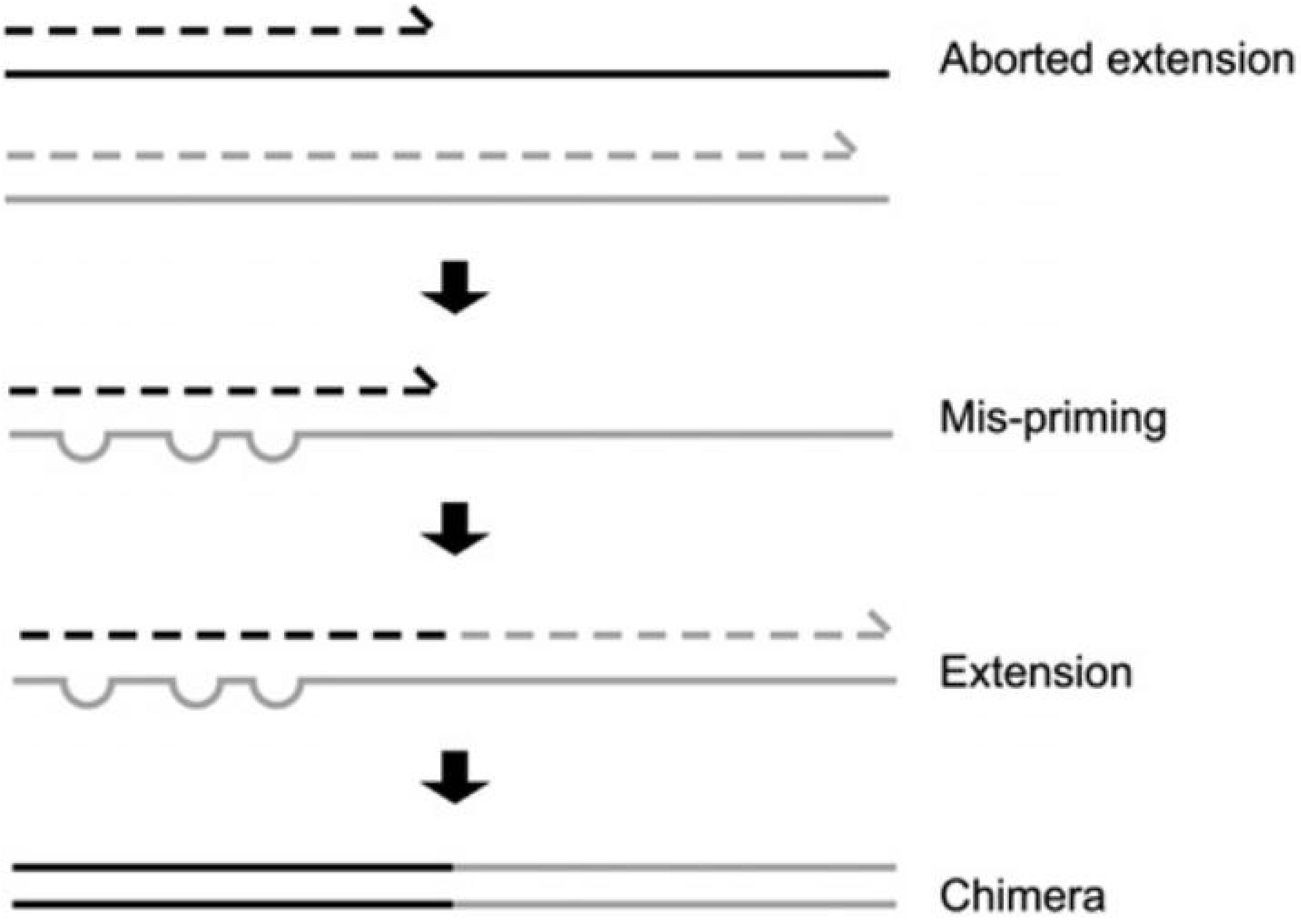
```
# Make a folder to save your results
> setwd("/home/sarkar/Documents/microbiome_workshop/resnew")

# Indicate the location of your fastq files
>demo_microbiome_fasqfiles <-
"/home/sarkar/Documents/microbiome_workshop/demo_fastq/demofastqsamples"

# load dada2
> library(dada2)
> packageVersion("dada2")


# Check if the fastq files are indicated correctly
> list.files(demo_microbiome_fasqfiles)
```

Chimera formation during 16S PCR



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3044863/

Thank you!