Microbiome workshop

dada2 workflow (fastq to ASV)

Anujit Sarkar
Postdoctoral Scholar
COPH, CON
Genomics Program, USF
Dec 03, 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.

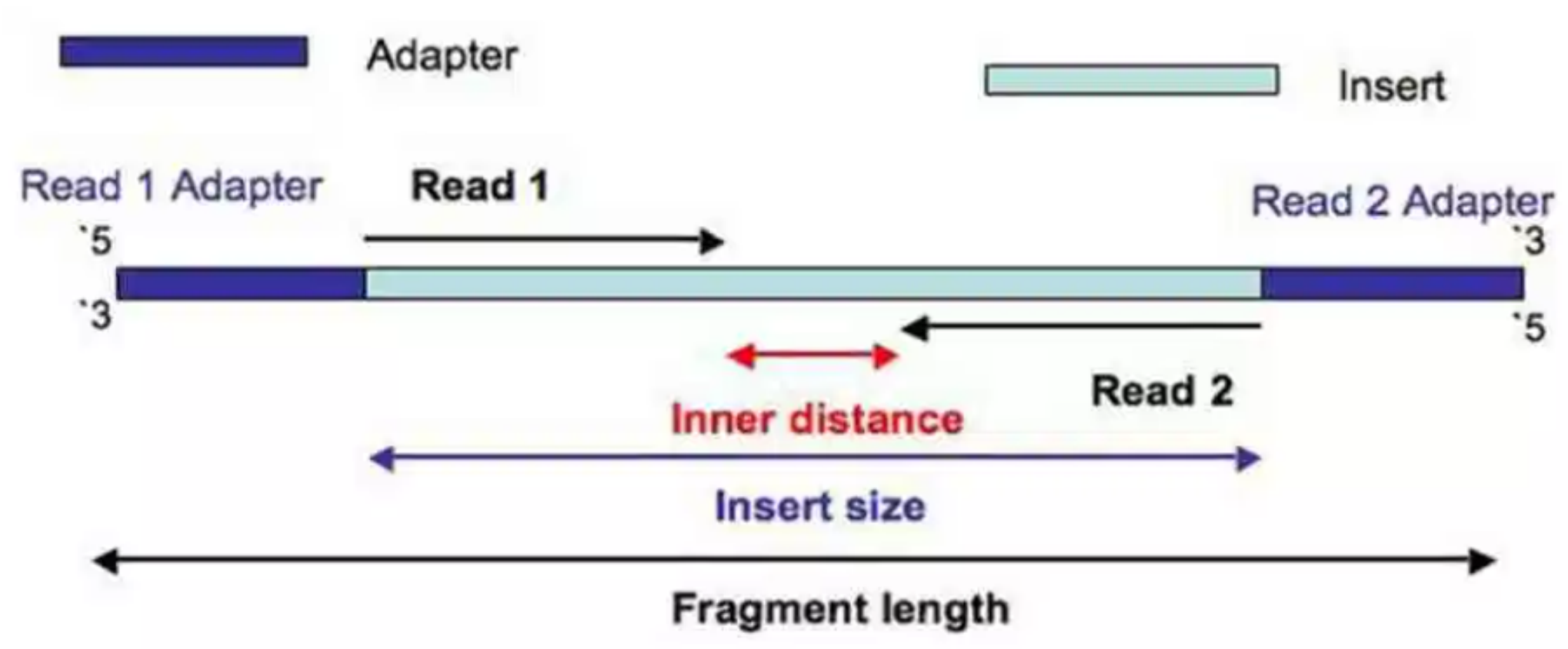
We previously introduced the Divisive Amplicon Denoising Algorithm (DADA), a model-based approach for correcting amplicon errors without constructing OTUs[5]. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives[2,5].

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].

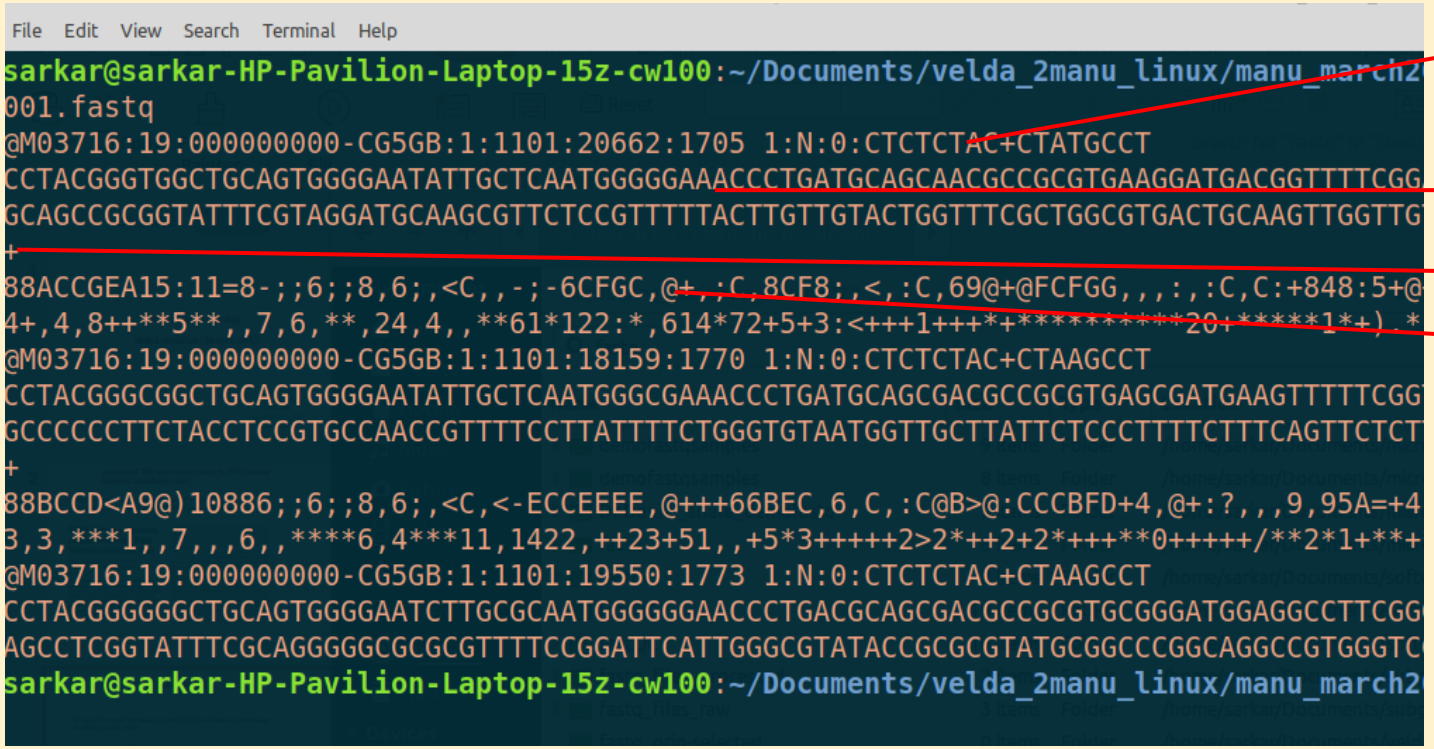
We benchmarked these algorithms on three mock commu-

Paired-end sequencing of 16S amplicons



https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/

Purpose of this task

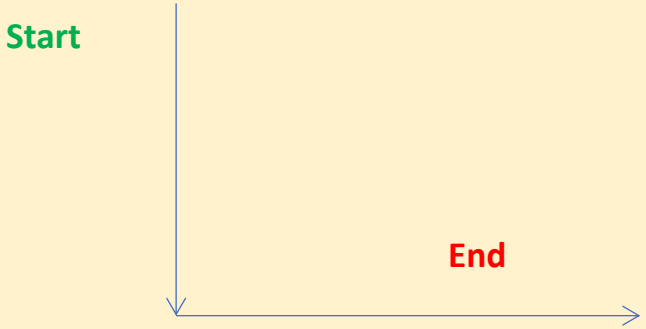


Sequence header

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

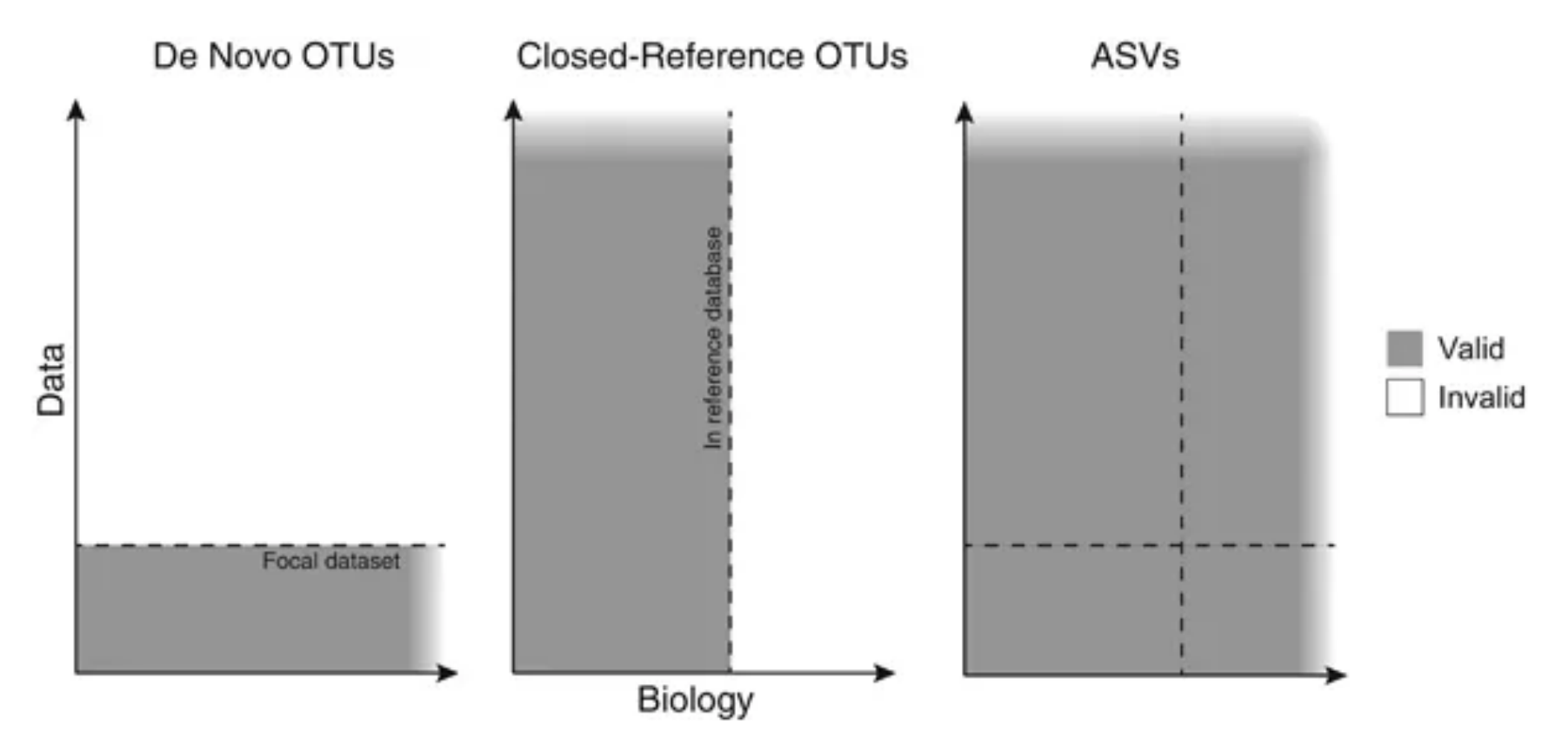
- R and Rstudio with dada2 installed
- Demultiplexed paired-end fastq files (preferably from Illumina, for this workshop) stored in a folder/directory
- An empty folder where all your results will be exported
- A 16S rRNA database (Greengenes, Silva or 16S RDP) downloaded and stored in a folder
- The 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)

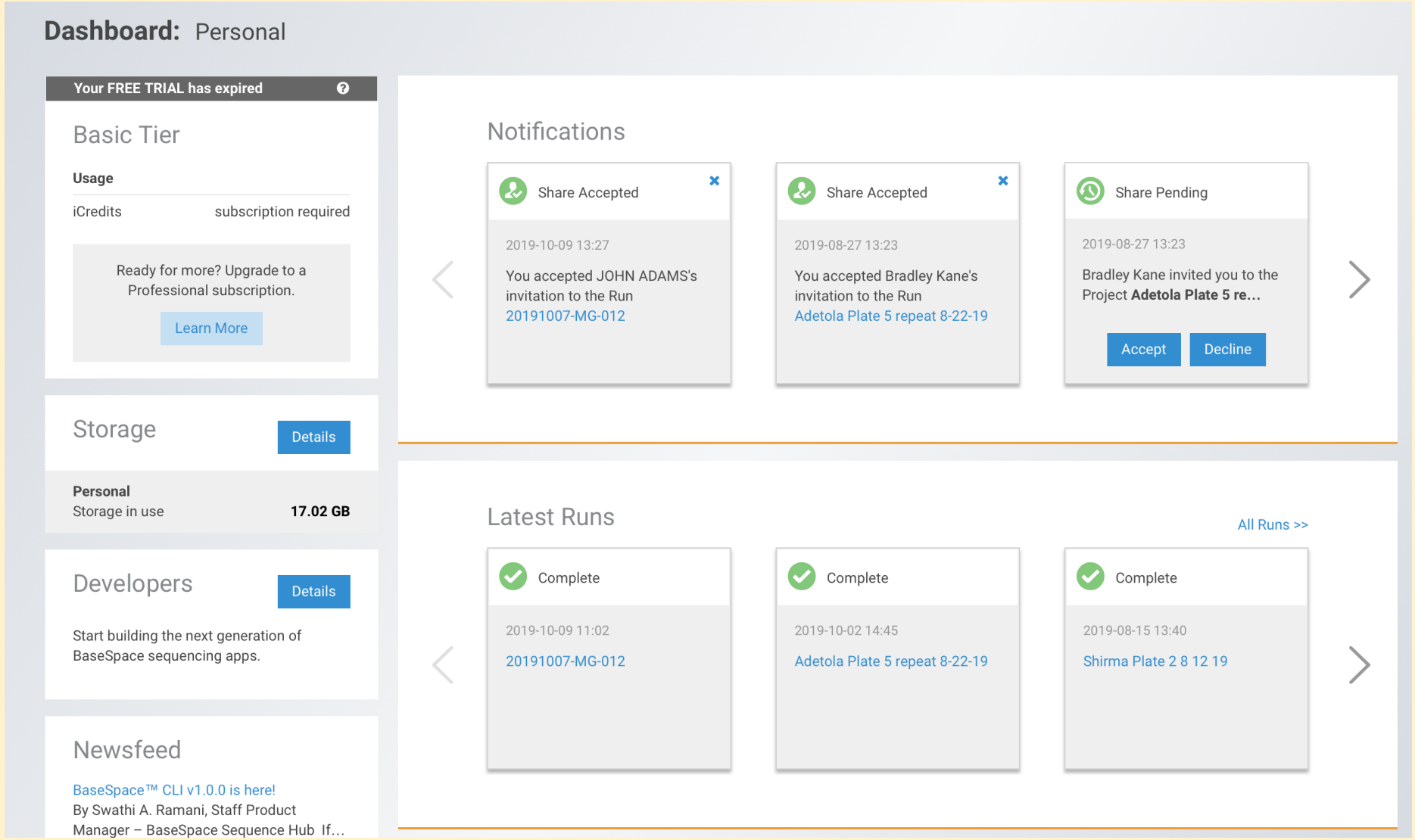
OTU vs ASV

- 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



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

SUMMARY | SAMPLES | CHARTS | METRICS | INDEXING QC | SAMPLE SHEET | FILES

Share | Download | More

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

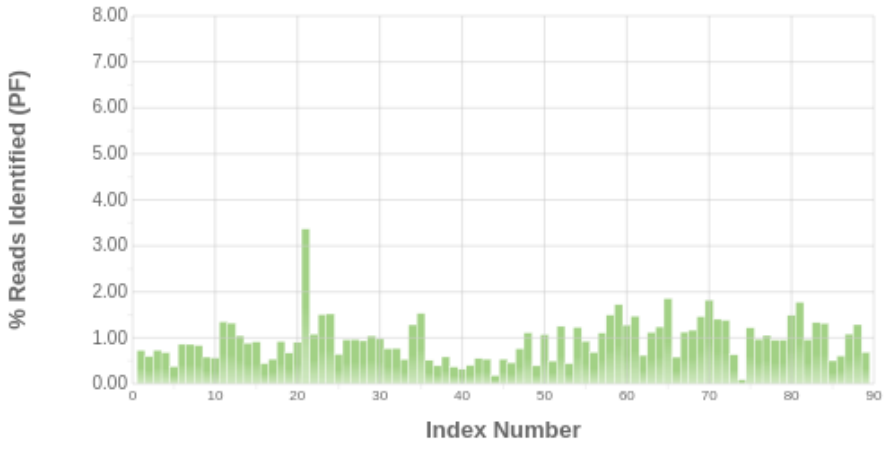
Samples (89)

view all

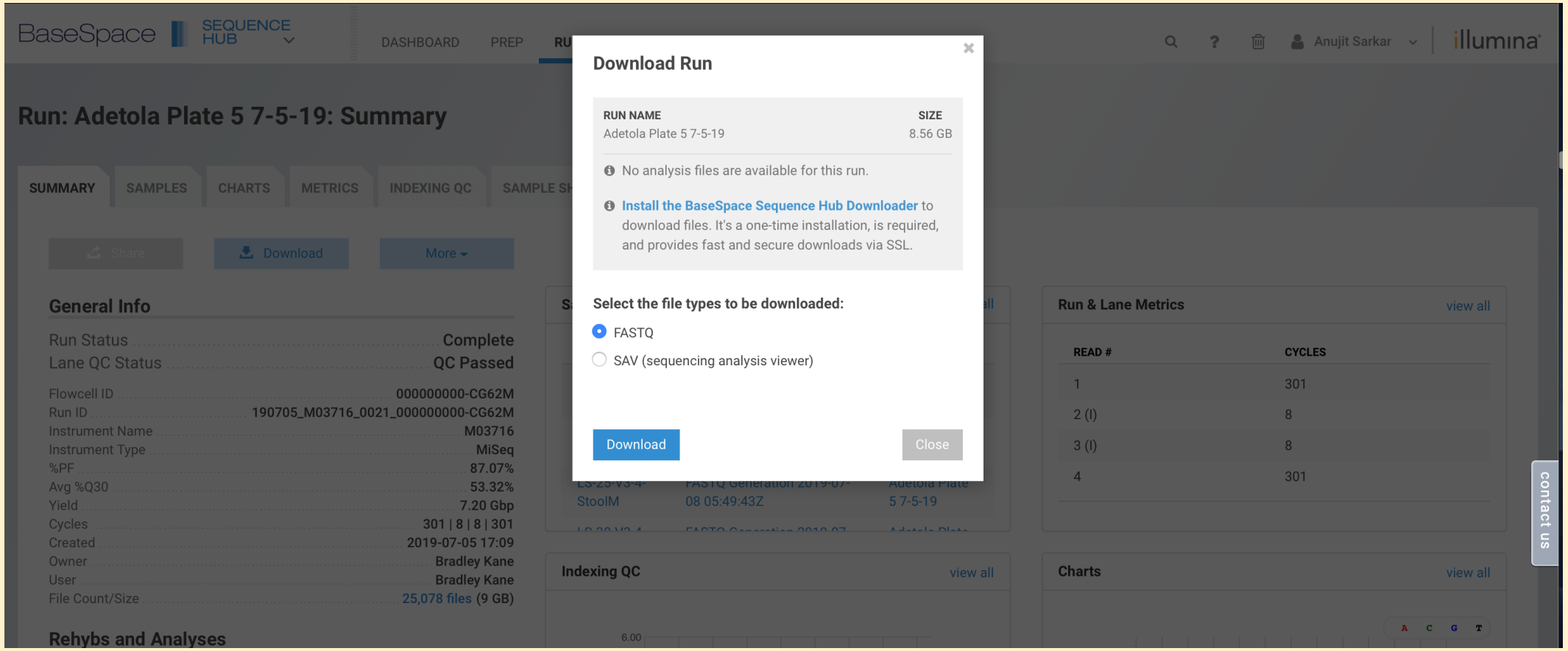
SAMPLE ID	APP	PROJECT
04mid	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
15bas	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
pr04mid	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
14bas	FASTQ Generation 2018-05-29 18:34:28Z	Velda plate 1
21end	FASTQ Generation 2018-05-29 18:34:28Z	Velda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

Indexing QC

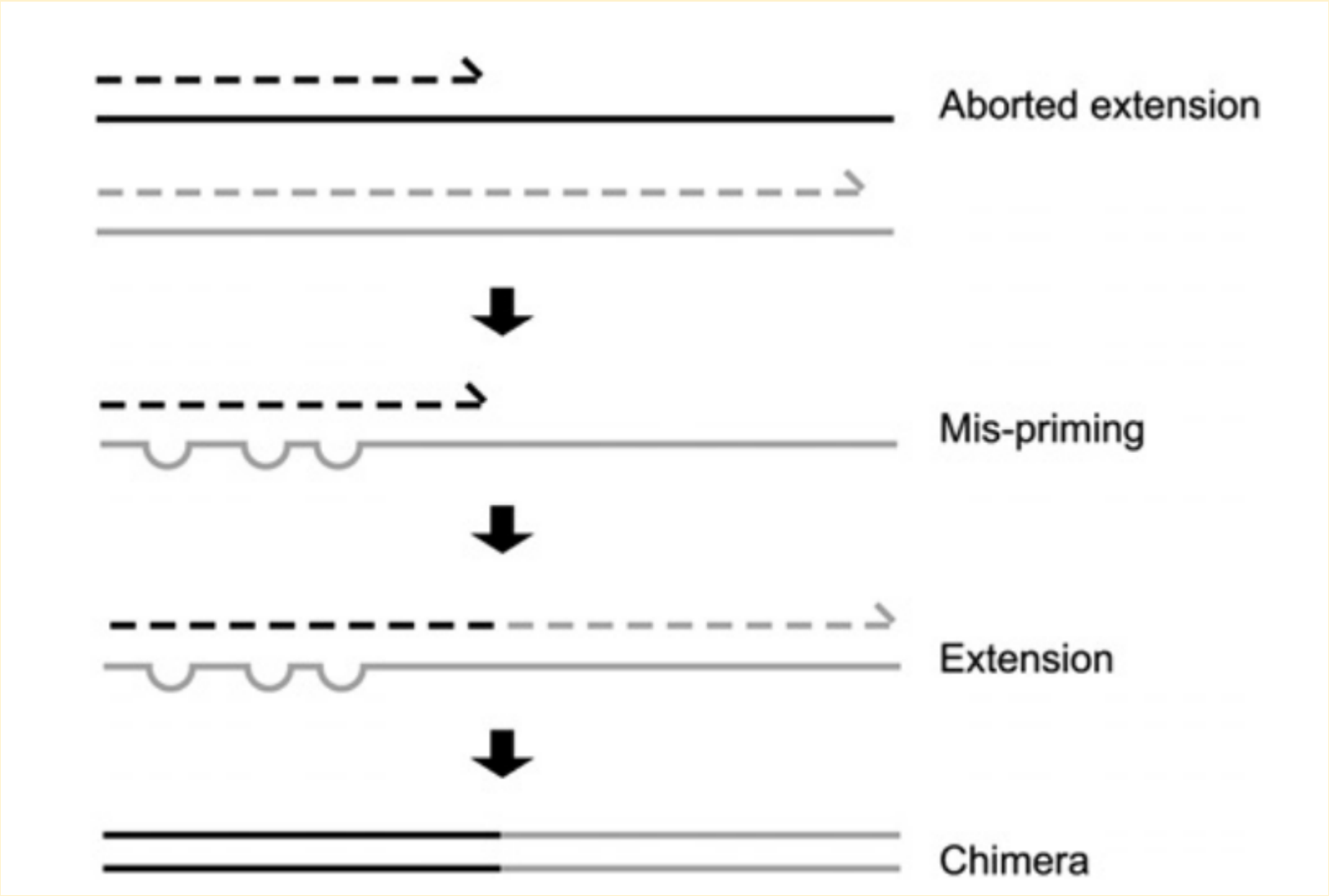
view all



Download Fastq files



Chimera formation during 16S PCR



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


Thank you!