ShadeLab Bioinformatics Workshop

Week 1

Sequencing platforms, amplicon (16S rRNA, ITS) processing with USEARCH

https://github.com/ShadeLab/workshop

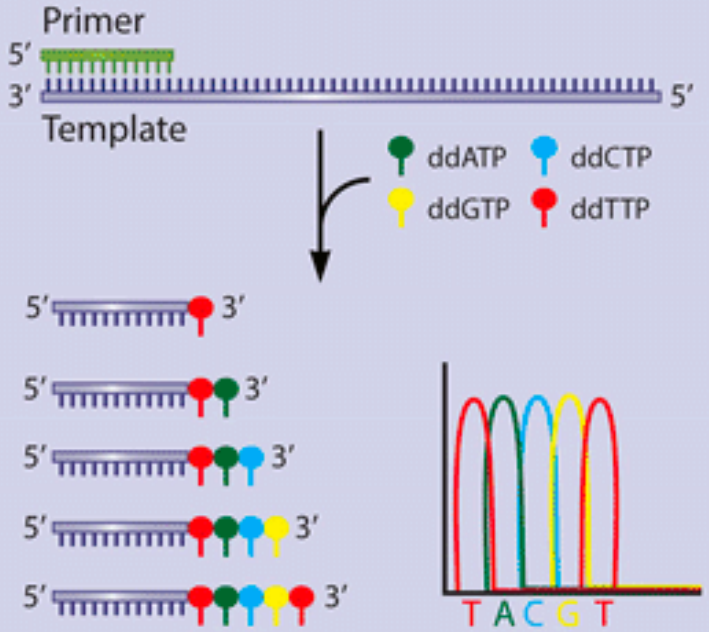
Checklist

- GitHub
  - git commands *clone*, *add*, *commit*, *push*, *pull*
- HPCC
  - directories
  - commands *ssh*, *cd*, *ls*, *mkdir*, *rm*, *cp, mv, scp*

- Install R and RStudio before next workshop

Sequencing platforms

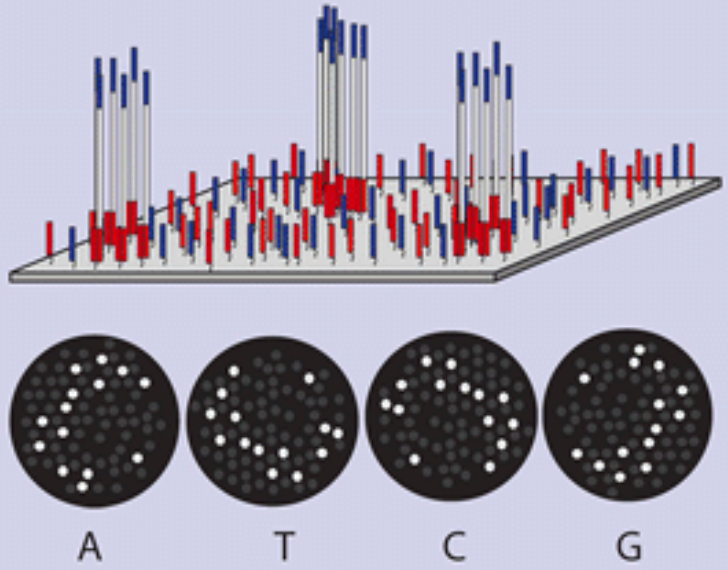
First Generation
Shotgun Sequencing



- Sequencing by synthesis
- High accuracy
- Long read lengths
- Relatively small amount of data generated

e.g., ABI capillary sequencer (ABI)

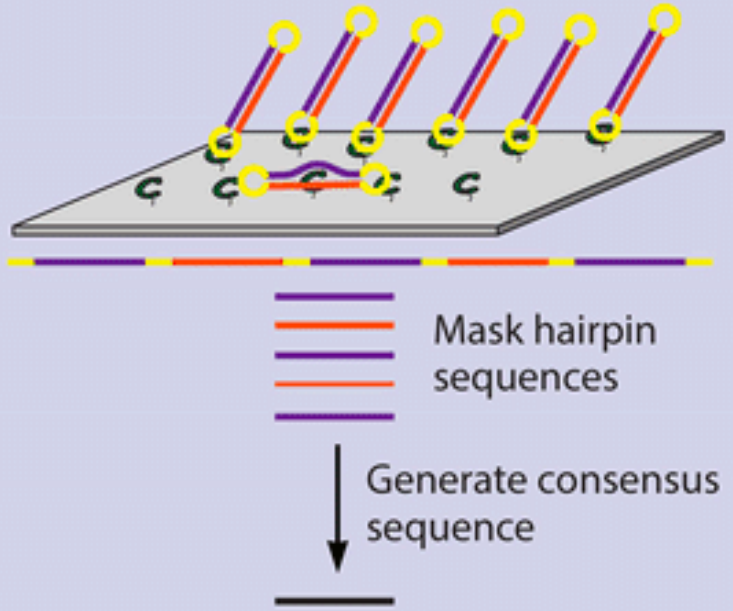
Second Generation
Massively Parallel Sequencing



- Sequencing by synthesis
- Amplified templates are generated during sequencing, reducing the requirements for starting material
- High accuracy
- Short read lengths

e.g., MiSeq (Illumina), Ion Torrent (Thermo Fisher Scientific)

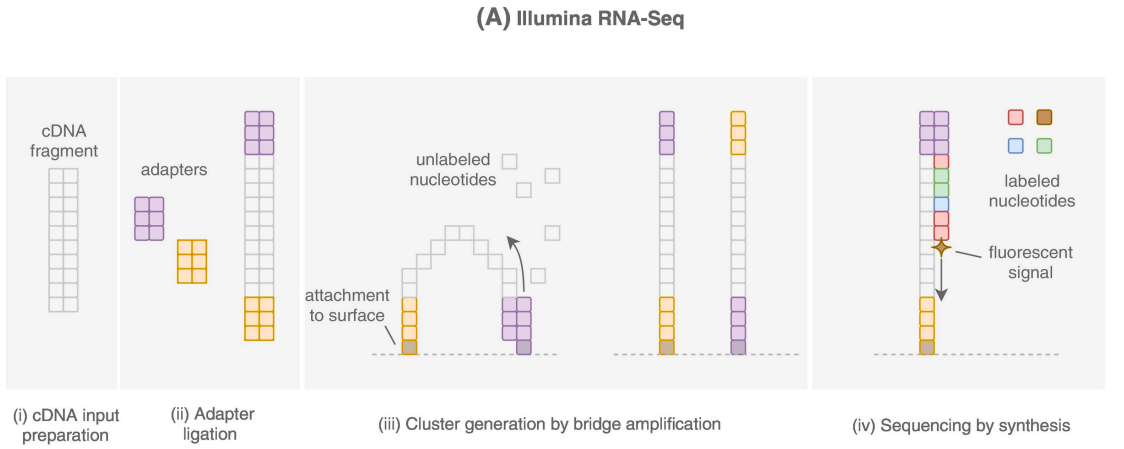
Third Generation
Single-molecule Sequencing

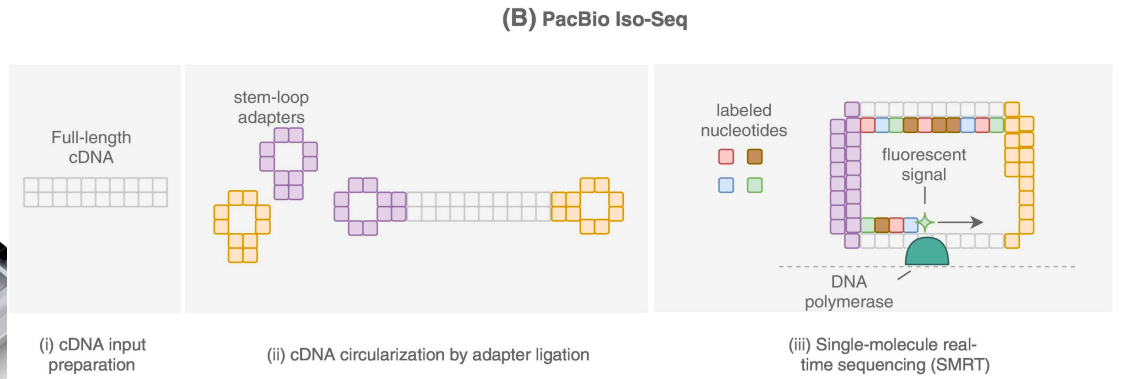


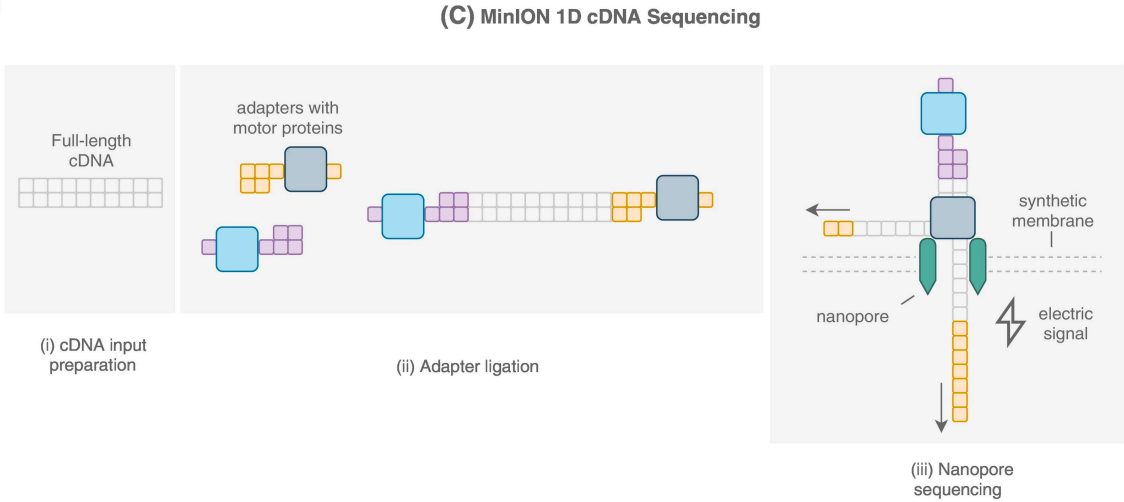
- Single-molecule templates
- Low accuracy
- Long read lengths

e.g., Single-Molecule Real-Time (SMRT) — Sequencing (Pacific Biosciences), MinION (Oxford Nanopore Technologies)

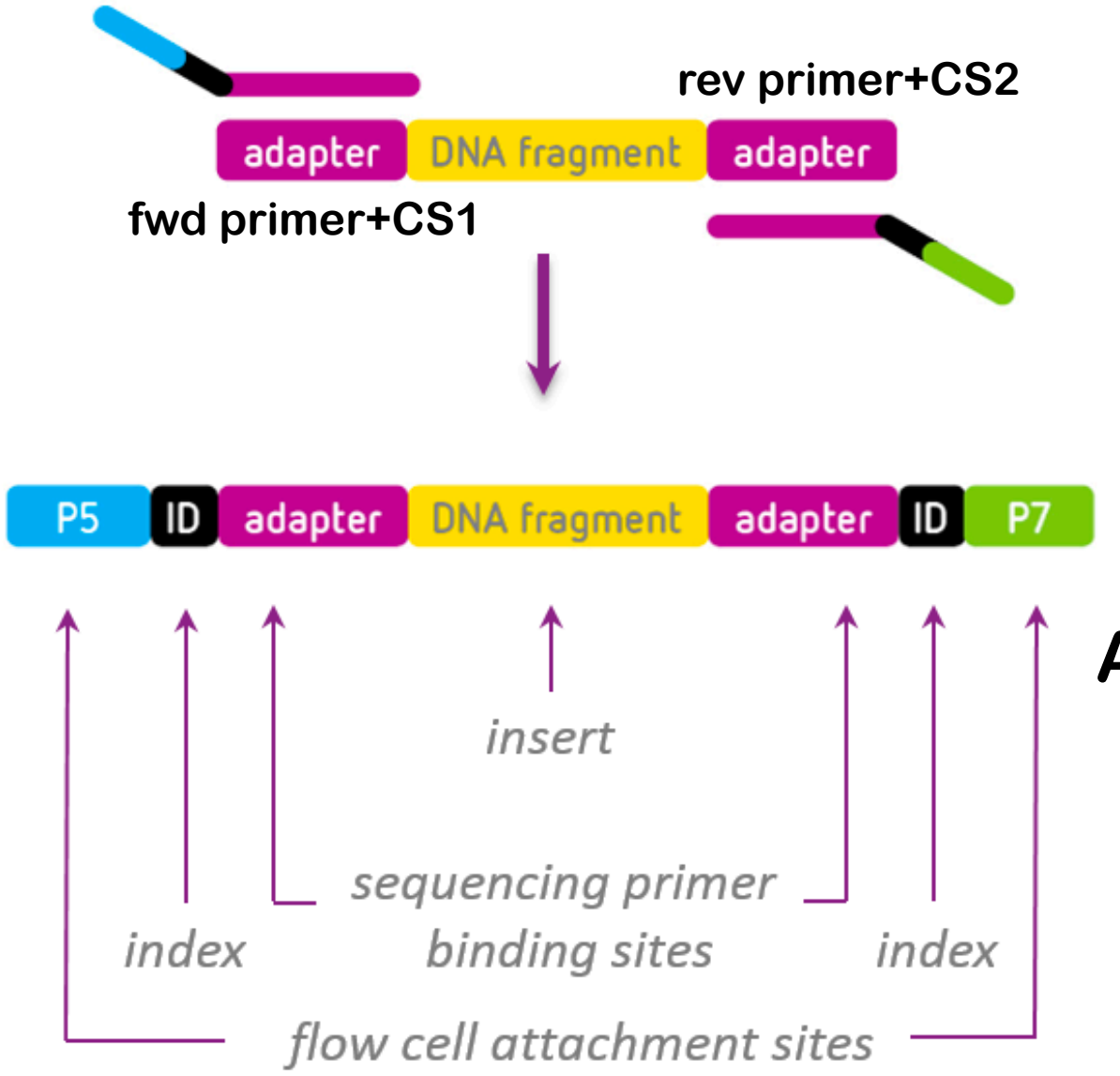
Technologies behind the 2nd and 3rd gen sequencing platforms







Illumina amplicon structure – library



All extra information needs to be removed!

(in some cases even PhiX signal – Illumina internal control)

RTSF (Illumina) sequencing results

… but first how to submit your sample:
https://my.labguru.com/knowledge/experiments/7540
https://rtsf.natsci.msu.edu/genomics/submitting-samples/

You will receive an email from RTSF containing the link to retrieve the files and QC report.

Sequencing is complete for samples submitted to the RTSF Genomics Core, project ID STO8258 (Aggregate_diversity). This project included a mixture of sample types. There were 16 samples of microbial metagenomic DNA submitted for 16 amplicon library preparation. The V4 hypervariable region was amplified using dual indexed, Illumina compatible primers 515f/806r following the Schloss lab protocol (Kozich, JJ, et al. 2013). PCR products were batch normalized using an Invitrogen SequalPrep DNA Normalization plate and the products recovered from the plate were pooled. The pool was QC'd and quantified using a combination of Qubit dsDNA HS, Agilent 4200 TapeStation HS DNA1000 and Kapa Illumina Library Quantification qPCR assays.

In addition, 36 primary PCR products prepared in your lab were also provided. Products from three (3) different targets were provided. These 1° PCR products were generated using primers with Fluidigm CS1/CS2 universal oligomers at their 5' ends. The Genomics Core performed 2° PCR using dual indexed, Illumina compatible primers which targeted the CS1/CS2 ends of the 1° PCR products. PCR products were batch normalized using an Invitrogen SequalPrep DNA Normalization plate and the products recovered from the plate were pooled. The pool was QC'd and quantified using a combination of Qubit dsDNA HS, Agilent 4200 TapeStation HS DNA1000 and Kapa Illumina Library Quantification qPCR assays. The QC revealed that it was likely the "amp_Block" samples were not amplified during the 2° PCR reactions. A second pool of the normalized PCR products was made, including only products from the AOA and AOB samples. This pool was QC'd as above.

The 16S and AOA/AOB library pools were combined in proportion to the number of libraries in each (16/28). This combined pool was loaded onto an Illumina MiSeq v3 flow cell and sequencing was performed in a 2x300bp paired end format using a MiSeq v3 600 cycle reagent cartridge. Custom sequencing and index primers complementary to the 16S 515f/806 as well as primers complementary to the Fluidigm CS1/CS2 oligos were added to appropriate wells of the reagent cartridge. Base calling was done by Illumina Real Time Analysis (RTA) v1.18.54 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v2.19.1. A summary of the run output is attached below. Basic QC information about your sequence data is provided by the accompanying FastQC reports. Please see the FastQC Tutorial and FAQ (https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/) for information regarding interpretation of these reports.

The number of reads per sample is on average 2X more for the 16S-V4 libraries than for the AOA/AOB amplicons. It is not surprising that different amplicons perform better or worse when sequencing on an Illumina instrument.

Overall the read quality is poor, especially so for Read 2. Poor quality for amplicon libraries sequenced on a MiSeq v3 PE300 run was a concern raised when first discussing this project.

Data may be downloaded using the Shade lab account on the Genomics FTP server as before. Data for this project is in subdirectory 20190920_Amplicon_PE300. You must use secure FTP (FTPS) when connecting to the Genomics server. See the Genomics FAQ for general instructions (https://rtsf.natsci.msu.edu/genomics/data-retrieval/). Sequence data typically remain available on the FTP server for 60 days. It is the responsibility of the researcher to download and store their data long term, including a safe backup copy. The RTSF Genomics Core only guarantees retention of sequence data for one year from the date of availability.

Regards,

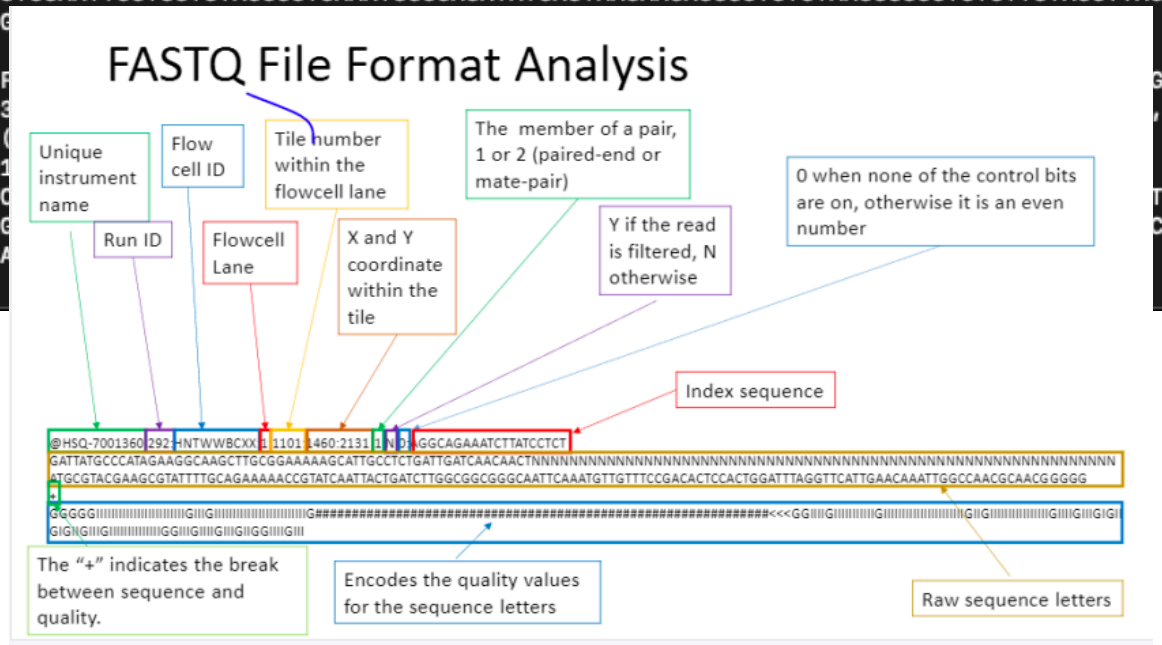
Kevin M. Carr

Download the FASTQ files to /mnt/research/ShadeLab/Sequences/raw_sequences/ along with the email content and the QC file that is attached in email

```
[(base) -bash-4.2$ cd ShadeLab/
[(base) -bash-4.2$ cd Sequence/
[(base) -bash-4.2$ cd raw_sequence/
[(base) -bash-4.2$ ls -lth
total 120K
drwxr-s---  2 bintarti ShadeLab 8.0K Oct 15 14:08 its.tutoria
drwxr-s---  4 stopnise ShadeLab 8.0K Oct  8 15:33 Soil_aggreg
drwxr-s---  4 bintarti ShadeLab 8.0K Sep 13 12:20 Bean_variab
drwxr-sr-x  4 bowsher1 ShadeLab 8.0K Jun 19 22:37 Bowsher
drwxrws---. 5 stopnise ShadeLab 8.0K Jun 19 11:11 Bean_develo
drwxrwsrwx. 4 kearnspa ShadeLab 8.0K Jun 13 14:58 Bean_bioge
drwxr-sr-x  3 stopnise ShadeLab 8.0K Jun 13 14:41 Phyllo
drwxrwsrwx  6 stopnise ShadeLab 8.0K Jun 13 13:21 GLBRC
drwxrwsrwx  8 gradykea ShadeLab 8.0K May  3 10:35 AppleReplan
drwxr-sr-x  2 dooleys1 ShadeLab 8.0K Sep  5  2018 GLBRC_analy
drwxr-sr-x  6 kearnspa ShadeLab 8.0K Apr 23  2018 PRI_Kearns
drwxr-sr-x  4 stopnise ShadeLab 8.0K Aug 17  2017 Soil
drwxr-sr-x  2 shadeash ShadeLab 8.0K Jul 11  2017 JGI_iTags_(
drwxrwsrwx  3 sorens75 ShadeLab 8.0K Jun  8  2017 Centralia
drwx------  2 shadeash ShadeLab 8.0K Aug  4  2016 160701_Jack
(base) -bash-4.2$
```

shadelab — stopnise@dev-intel14:/mnt/research/ShadeLab/Sequence/raw_sequence — ssh stopnise@hpcc.msu....

```
@M02808:590:000000000-CNC3P:1:1101:16191:2302 1:N:0:CGAGAGTT+CTACTATA
TACAGAGGTCCCAAGCGTTGTTCGGATTCACTGGGCGTAAAGGGTGCGTAGGTGGCCAGGTAAGTTTGGTGTGAAAGCTCGGAGCTTAACTCCAAAGAGTCATTGAATA
CTATCTGGCTGGAGGGTCGGAGGGGAGACTGGAATTCTCGTTGTAGCAGTGAAATGCGTAGATATCGAGAGGAACACCAGTGGCGAAGGCGAGTCTCTGGACGACTCCT
GACACTGACGCACGAAAGCTAGGGGAGCAAACAGGATTAGATACCCTTGTAGTCCGGCTGACTGACTCGAGATTTATCTCGT
+
8-8-6,C8EFCEC8@CBFECFFGC+@BFF<FF98,;+@@,<,,:E7@@CFFDFCFGEEGFGGGGGGFFGGGGGGGGGGGGGGCGGGGGGGGGGFGGGFGGGGGGGFGFF
FG<BFEADFF7=,,++@F7+++8::787<FDC88CDFGFG+@DCF9D;>EF9DDBFE>EGGGECDCC+****,>>8FC*,?>719****4:*8DFEGF96+0,))/3;@
))8*6@8@()676>:>E>BFF<FFD(,,(:9700114938):7FC<?0462:6<<((2;023<47:<23(-(().654<BFF
@M02808:590:000000000-CNC3P:1:1101:13000:2304 1:N:0:CGAGAGTT+CTACTATA
TACGGAGGGTGCAAGCGTTATCCGGATTCACTGGGTTTAAAGGGTGCGTAGGCGGATTAGAAAGTCCGTTGTGAAATCTCTGAGCTTAACTCGGAAATTGCGGTGGATA
CTATTAGTCTTGAATACTGTGGAGGTTAGCGGAATATGTCATGTAGCGGTGAAATGCATAGATATTTCATAGAACACCAATTGCGAAGGCAGCTGGCTACACAGATATT
GACGCTGAGGCACGAAAGCGTGGGGATCAAACAGGATTAGAAACCCTGGTAGTCCGGCTGACTGACTCGAGAGTTATCTCGT
+
@-@@9,@EDFEBD9,C@CF;EFFC7,BCF,CEF8,CFFA99CD@F7@FFGEGGGGGGGGFFGGGGGGGGGGGGGGGGGGGGGDGGGGGGGGGGGGGGGGGGGGGFF
GG9EFABFFGF?,,E<EF,A,,,,+@EA=<@7+>@=F@FG9FDFF9D77C=7<BDDD,E7;+@7E,CF,?,,,,3,@>*1@<,<95***4<=:D7,;:,<49;)*0*99
*58)-920()803??DC66B<;ECG0:32</901:14=>2)0*<887(84<9<>>4(6>4:3><0899118:18B5<<<A?F
@M02808:590:000000000-CNC3P:1:1101:22296:2327 1:N:0:CGAGAGTT+CTACTATA
TACGTAGGGGGCAAGCGTTGTCCGGATTTATTGGGCGTAAAGAGCGTGTAGGCGGCCATGTAGGTCCGTTGTGAGAACTCGAGGCTCAACCTCGAGACGCCGATGGAAA
CCATGTGGCTAGAGTCCGGAAGAGGAGAGTGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGTAAGAACACCCGTGTCTAAGGCGGCTCTCTTGTACCTTACT
GACGCTTAGACGCGAAAGCGTGGGGAGCG
+
@-A@C@F>FC78@,C@BFE+BEF7+,6BF
GG=A?D<,AB9,,5C<?C++++,:,,4,3
))2)-4(()/.320;>E>73435E3(22(
@M02808:590:000000000-CNC3P:1
TACGTAGGGGGCAAGCGTTGTCCGGAATC
CTGTCGAGCTCGAGTCCGGAAGAGGCGAG
GACGCTGAGACGCGAAAGCGTGGGGAGCA
:
```



FASTQ file

If PE option, you will receive 2 files per sample, containing *R1*.fastq and *R2* labels

Amplicon processing tools

USEARCH

https://www.drive5.com/usearch/

(alternative: https://github.com/torognes/vsearch)

https://www.mothur.org/

Quantitative Insights Into Microbial Ecology

qiime

http://qiime.org/

DADA2

Amplicon Sequencing. **Exactly.** *Version 1.12*

http://benjjneb.github.io/dada2/index.html

USEARCH 64-bit version