Bioinformatics Lessons Schedule

- RNA-seq
- single cell RNA-seq
- RRBS

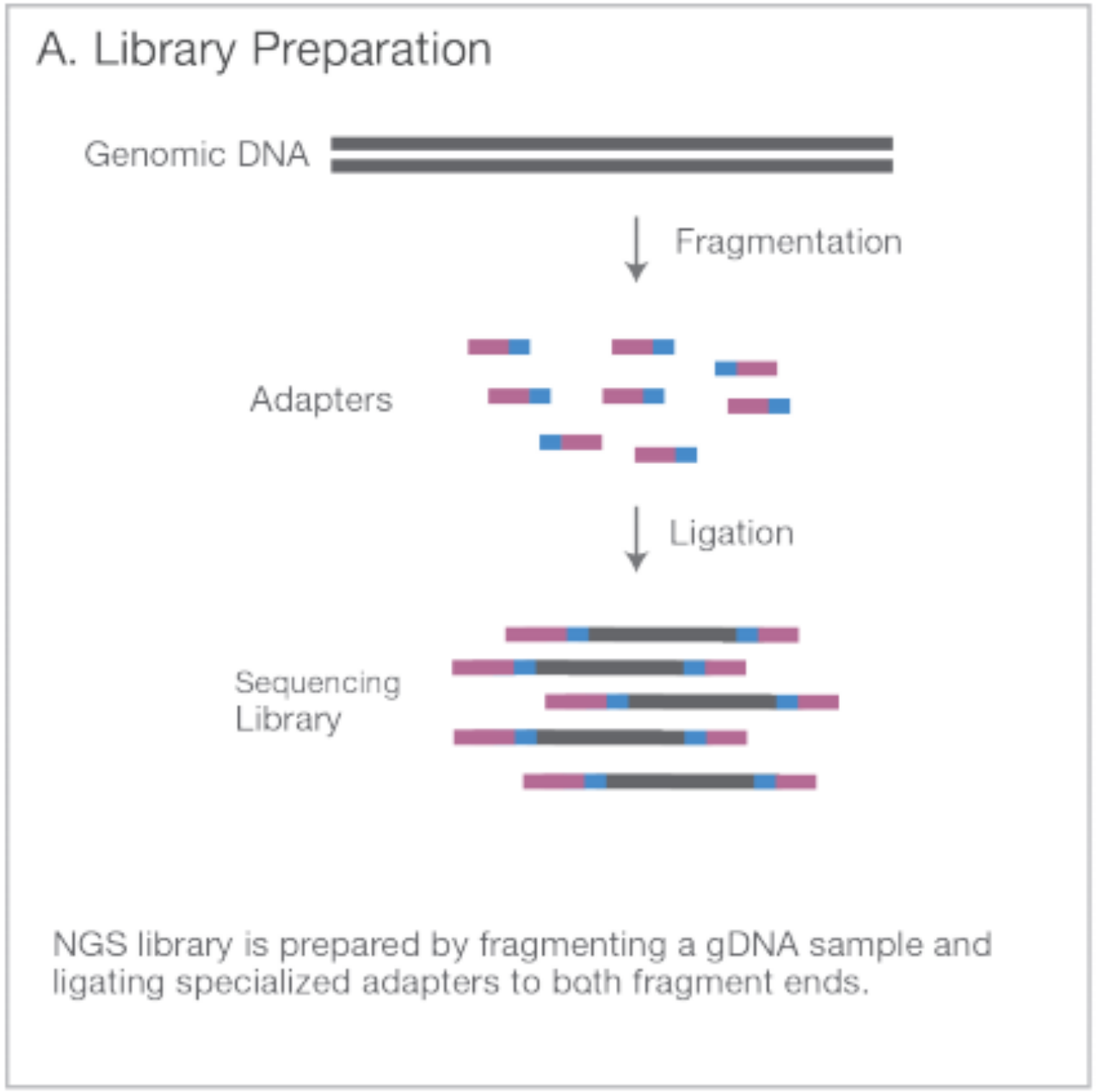
Date	Subject
11-26	No lesson, week of Thanksgiving
12-03	Server Basics, repeat of Week 5, continued
12-10	Basic Git
12-17	How to run software on the server
12-24	Christmas break
12-31	Christmas break
01-07	Process RNA-seq
01-14	Process RNA-seq, continued
01-21	Analyze RNA-seq
01-28	Analyze RNA-seq, continued

RNA-seq

Quick Review: How does Illumina sequencing work?

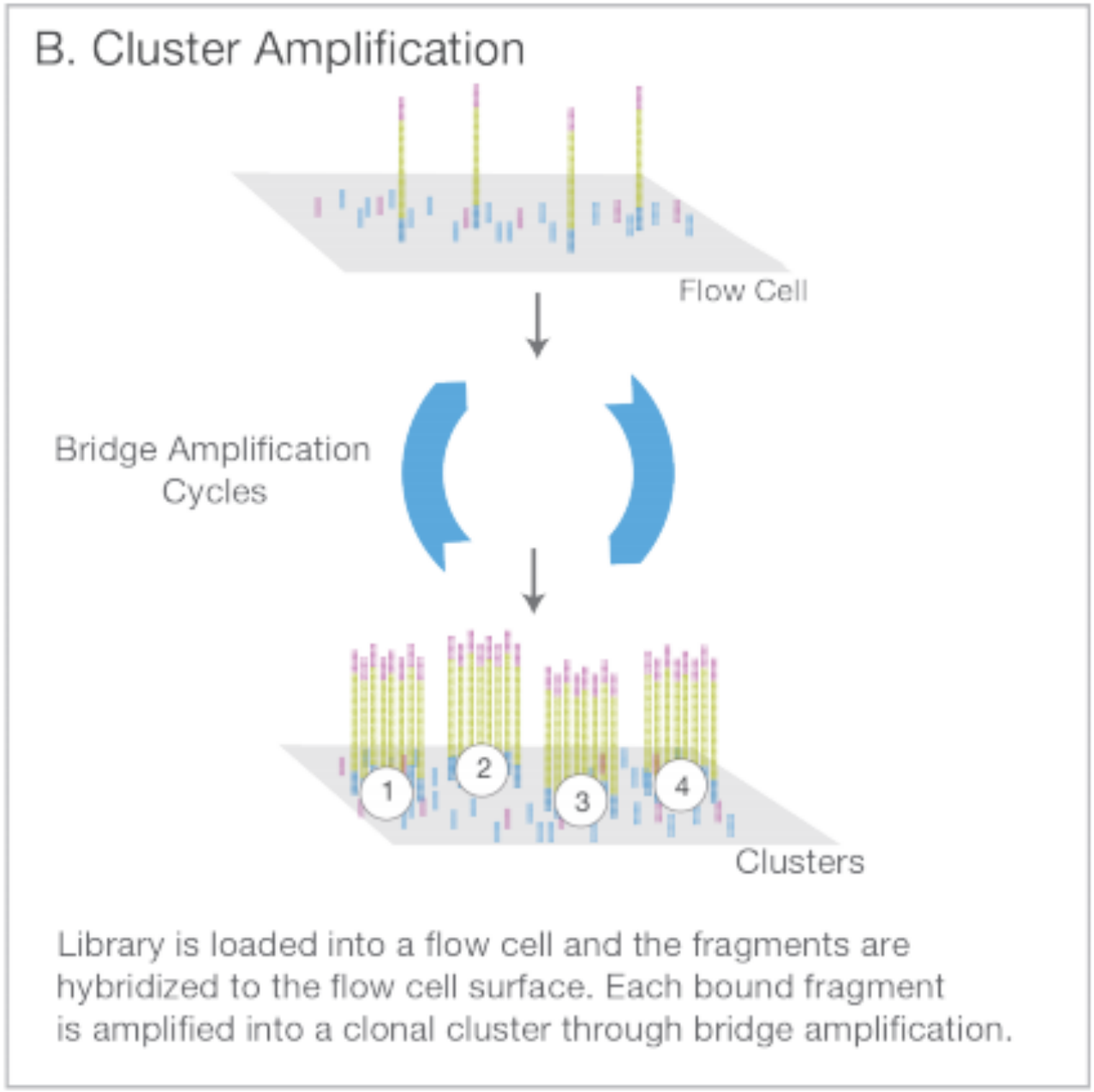
Illumina Sequencing

- General overview
- For RNA-seq, extract RNA and remove ribosomal RNA as well

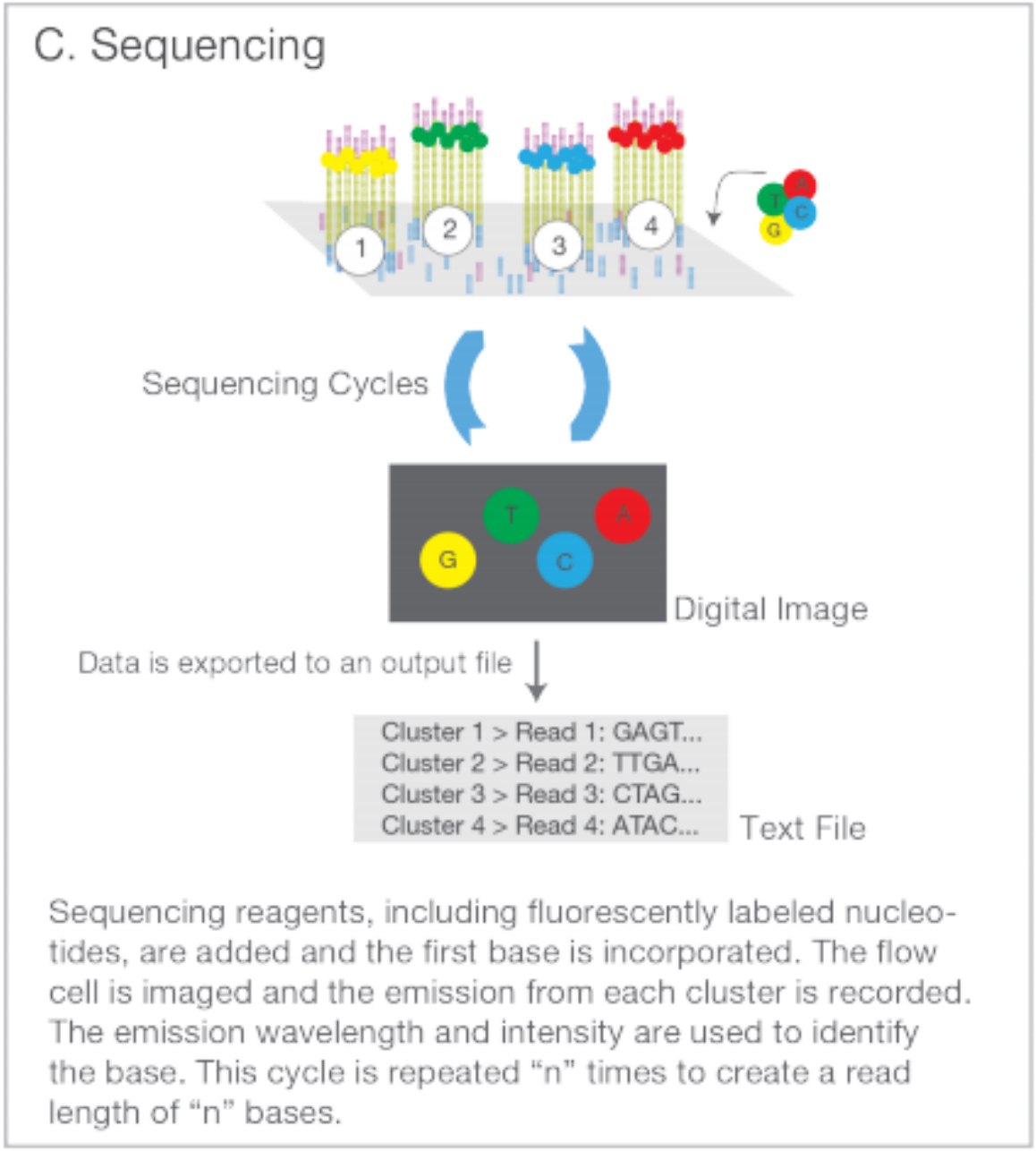


Illumina Sequencing

Make tiles of identical DNA to read

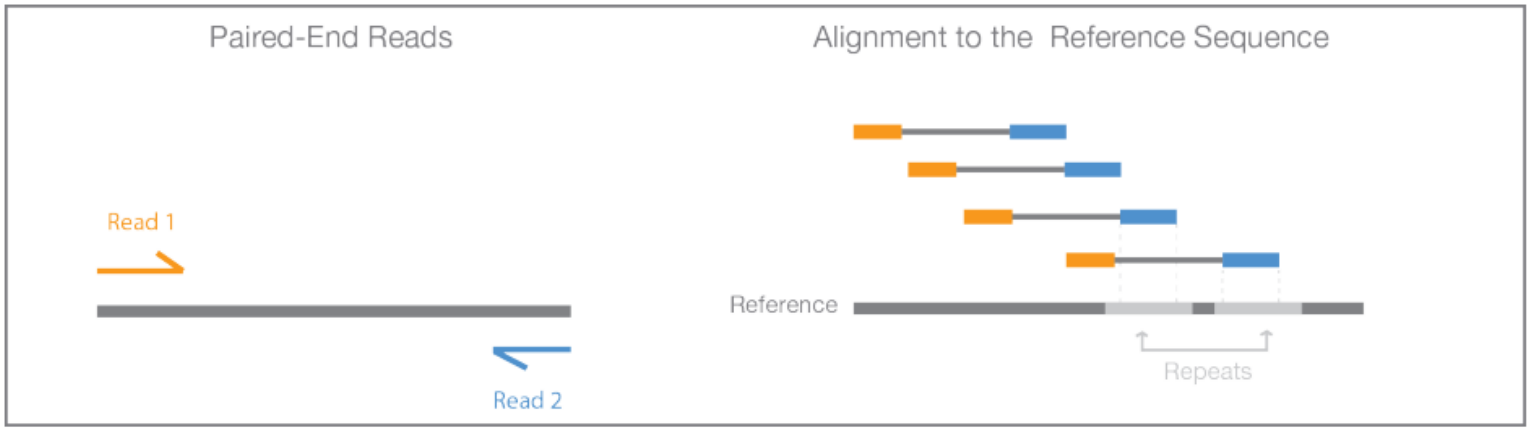


Illumina Sequencing



Paired-End Sequencing

- Sequence both ends of the fragment
- Because sequencing is always 5’ to 3’, the read pairs will be in the opposite orientation
- 90% of the time, the programs you use will be aware of the difference in orientation and take care of it for you
- Because the distance between the pairs is known (depends on the sequence length you asked for) mapping is more accurate, especially in highly repetitive regions of the genome
- For RNA-seq, paired end reads are necessary if you want to look at alternative splicing
- More expensive than single end sequencing



What does raw sequencing data look like?

FastQ Files

FastQ Files

- Fastq files (usually) end in either `fastq.gz` or `fq.gz` (or they can be missing the .gz extension)

FastQ Files

- Fastq files (usually) end in either `fastq.gz` or `fq.gz` (or they can be missing the .gz extension)
- File names will have some combination of the following information (depends on the sequencer):
  - Sample ID
  - Lane
  - Read number
  - Unique ID from the company or sequencer

FastQ Files

- Fastq files (usually) end in either `fastq.gz` or `fq.gz` (or they can be missing the .gz extension)
- File names will have some combination of the following information (depends on the sequencer):
  - Sample ID
  - Lane
  - Read number
  - Unique ID from the company or sequencer

```
[kkeith@cbix rnaseq_data]$ ll
total 46700
-rw-r--r--. 1 kkeith research 4276449 Dec 18 10:55 dac1_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4441834 Dec 18 10:55 dac1_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 4118184 Dec 18 10:55 dac2_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4296786 Dec 18 10:55 dac2_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 4336091 Dec 18 10:56 dac3_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4519748 Dec 18 10:56 dac3_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 3652875 Dec 18 10:57 siC1_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 3830628 Dec 18 10:57 siC1_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 3941656 Dec 18 10:58 siC2_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4103667 Dec 18 10:58 siC2_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 3078529 Dec 18 10:59 siC3_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 3198840 Dec 18 10:59 siC3_chr21_R2.fastq.gz
```

FastQ Files

- Fastq files (usually) end in either `fastq.gz` or `fq.gz` (or they can be missing the .gz extension)
- File names will have some combination of the following information (depends on the sequencer):
  - Sample ID
  - Lane
  - Read number
  - Unique ID from the company or sequencer

```
[kkeith@cbix rnaseq_data]$ ll
total 46700
-rw-r--r--. 1 kkeith research 4276449 Dec 18 10:55 dac1_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4441834 Dec 18 10:55 dac1_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 4118184 Dec 18 10:55 dac2_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4296786 Dec 18 10:55 dac2_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 4336091 Dec 18 10:56 dac3_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4519748 Dec 18 10:56 dac3_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 3652875 Dec 18 10:57 siC1_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 3830628 Dec 18 10:57 siC1_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 3941656 Dec 18 10:58 siC2_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 4103667 Dec 18 10:58 siC2_chr21_R2.fastq.gz
-rw-r--r--. 1 kkeith research 3078529 Dec 18 10:59 siC3_chr21_R1.fastq.gz
-rw-r--r--. 1 kkeith research 3198840 Dec 18 10:59 siC3_chr21_R2.fastq.gz
```

```
[kkeith@cbix Rawdata]$ ll [fm]*/*.fq.gz
-rw-r--r--. 1 jjelinek research 4994684738 Dec 10 20:02 f45y4/f45y4_CKDL190143587-1a-6_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 5329097223 Dec 10 20:03 f45y4/f45y4_CKDL190143587-1a-6_H723FBBXX_L1_2.fq.gz
-rw-r--r--. 1 jjelinek research 4378414462 Dec 10 20:01 f53y6/f53y6_CKDL190143587-1a-12_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 4688210826 Dec 10 20:02 f53y6/f53y6_CKDL190143587-1a-12_H723FBBXX_L1_2.fq.gz
-rw-r--r--. 1 jjelinek research 4353424157 Dec 10 20:01 f61y8/f61y8_CKDL190143587-1a-19_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 4589695705 Dec 10 20:01 f61y8/f61y8_CKDL190143587-1a-19_H723FBBXX_L1_2.fq.gz
-rw-r--r--. 1 jjelinek research 4595389700 Dec 10 20:00 m38y1/m38y1_CKDL190143587-1a-2_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 4905936742 Dec 10 20:01 m38y1/m38y1_CKDL190143587-1a-2_H723FBBXX_L1_2.fq.gz
-rw-r--r--. 1 jjelinek research 3745371888 Dec 10 20:00 m45y3/m45y3_CKDL190143587-1a-5_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 3963991106 Dec 10 20:00 m45y3/m45y3_CKDL190143587-1a-5_H723FBBXX_L1_2.fq.gz
-rw-r--r--. 1 jjelinek research 4237208070 Dec 10 20:01 m53y5/m53y5_CKDL190143587-1a-7_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 4497568575 Dec 10 20:02 m53y5/m53y5_CKDL190143587-1a-7_H723FBBXX_L1_2.fq.gz
-rw-r--r--. 1 jjelinek research 4061248079 Dec 10 20:00 m61y7/m61y7_CKDL190143587-1a-16_H723FBBXX_L1_1.fq.gz
-rw-r--r--. 1 jjelinek research 4308294343 Dec 10 20:01 m61y7/m61y7_CKDL190143587-1a-16_H723FBBXX_L1_2.fq.gz
```

What does a raw read look like?

What does a raw read look like?

read ID

```
@SN930:673:HT5JVBCXY:2:2103:18909:8888/1
CTTTATTTCTGCCTTCATTTTGTTATGTACCCAGTAGTCATTCAGGAGCAGGTTGTTCAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAAT
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGIIIIGGGIIIIIIIIIIIIIIIIIIIIIGGGIGIGGIGIGIGIIG
```

What does a raw read look like?

read ID

pair/mate number

```
@SN930:673:HT5JVBCXY:2:2103:18909:8888/1
CTTTATTTCTGCCTTCATTTTGTTATGTACCCAGTAGTCATTCAGGAGCAGGTTGTTCAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAAT
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGIIIIGGGIIIIIIIIIIIIIIIIIIIIIGGGIGIGGIGIGIGIIG
```

What does a raw read look like?

read ID

pair/mate number

```
@SN930:673:HT5JVBCXY:2:2103:18909:8888/1
CTTTATTTCTGCCTTCATTTTGTTATGTACCCAGTAGTCATTCAGGAGCAGGTTGTTCAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAAT
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGIIIIGGGIIIIIIIIIIIIIIIIIIIIIGGGIGIGGIGIGIGIIG
```

sequence

What does a raw read look like?

read ID

pair/mate number

```
@SN930:673:HT5JVBCXY:2:2103:18909:8888/1
CTTTATTTCTGCCTTCATTTTGTTATGTACCCAGTAGTCATTCAGGAGCAGGTTGTTCAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAAT
+
GGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGIIIIGGGIIIIIIIIIIIIIIIIIIIIIGGGIGIGGIGIGIGIIG
```

nothing

sequence

What does a raw read look like?

read ID

pair/mate number

```
@SN930:673:HT5JVBCXY:2:2103:18909:8888/1
CTTTATTTCTGCCTTCATTTTGTTATGTACCCAGTAGTCATTCAGGAGCAGGTTGTTCAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAAT
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGIIIIGGGIIIIIIIIIIIIIIIIIIIIIGGGIGIGGIGIGIGIIG
```

nothing

base quality score

sequence

What does a raw read look like?

read ID

pair/mate number

```
@SN930:673:HT5JVBCXY:2:2103:18909:8888/1
CTTTATTTCTGCCTTCATTTTGTTATGTACCCAGTAGTCATTCAGGAGCAGGTTGTTCAGTTTCCATGTAGTTGAGCAGTTTTGAGTGAGTTTCTTAATCCTGAGTTCTAGTTTGATTGCACTGTGGTCTGAGAGACAGTTTGTTATAAT
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGIIIIGGGIIIIIIIIIIIIIIIIIIIIIGGGIGIGGIGIGIGIIG
```

nothing

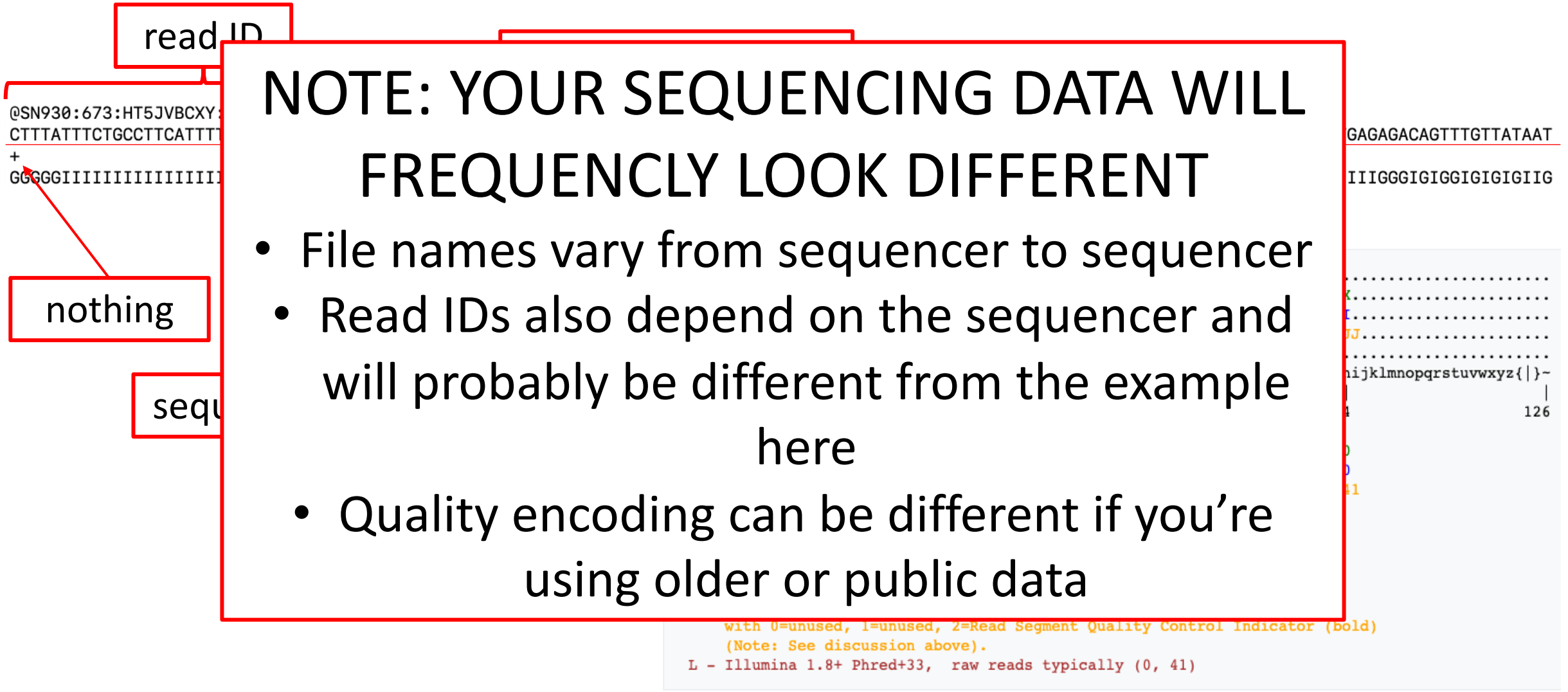
sequence

base quality score

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0.........................26...31.......40
                         -5....0........9..............................40
                               0........9..............................40
                                  3.....9...............................41
 0.2.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

What does a raw read look like?



Setting Up a Project

Open a plain text file to take notes in / document your work

Set Up Your Project Folder

1. Log onto the server

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

5. Go to the `rnaseq_data` directory

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

5. Go to the `rnaseq_data` directory
6. `git init`

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

5. Go to the `rnaseq_data` directory
6. `git init`
7. Make a documentation file

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

5. Go to the `rnaseq_data` directory
6. `git init`
7. Make a documentation file
8. Add a .gitignore

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

5. Go to the `rnaseq_data` directory
6. `git init`
7. Make a documentation file
8. Add a .gitignore
9. Commit

Set Up Your Project Folder

1. Log onto the server
2. Go to your data folder
3. Make a directory, `rnaseq_practice`
4. Type this command into your terminal to copy the practice rnaseq data

```
cp -r /mnt/data/coriell_bioinformatics_server_lessons/coriell_server_lessons/rnaseq/rnaseq_data/ ~/data/rnaseq_practice/
```

5. Go to the `rnaseq_data` directory
6. `git init`
7. Make a documentation file
8. Add a .gitignore
9. Commit
10. Set up and push to remote

Add Your SSH Key to Your GitHub Account

1. Copy your ssh key to your clipboard using ⌘+c or by highlighting the text in PuTTy when you view it using `less`

   ```
   less ~/.ssh/id_rsa.pub
   ```

2. Follow along with me as I show you how to add the key to your GitHub account.