

QAA_Working_Notes

Helpful

For all steps below, process the two libraries separately.

My data assignments are:

Grace 15_3C_mbnl_S11_L008 17_3E_fox_S13_L008

Henceforth referred to as 15 and 17.

Files found here:

```
/projects/bgmp/shared/2017_sequencing/demultiplexed/  
# Specific paths:  
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R1_001  
.fastq.gz  
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R2_001  
.fastq.gz  
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R1_001.  
fastq.gz  
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R2_001.  
fastq.gz
```

Repo here:

```
/projects/bgmp/ghach/bioinfo/Bi623/QAA
```

Tutorials for command line interface

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon-flipped/lessons/05_qc_running_fastqc_interactively.html#:~:text=To%20run%20the%20FastQC%20program,the%20command%20to%20run%20FastQC.

fastqc seqfile1, seqfile2

Part 1

Creating environment, installing fast QC

```
conda create --name QAA
conda activate QAA
conda install bioconda::fastqc
```

Answered yes to all prompts, verified version:

```
fastqc --version
```

Output: FastQC v0.12.1

Parameters to run fast QC

--extract If set then the zipped output file will be uncompressed in the same directory after it has been created. If --delete is also specified then the zip file will be removed after the contents are unzipped.

Set these both to true?

--o output directory

--svg Save the graphs in the report in SVG format.

Set to true, I like svgs

-a Specifies a non-default file which contains the list of

--adapters adapter sequences which will be explicitly searched against the library. The file must contain sets of named adapters in the form name[tab]sequence. Lines prefixed with a hash will be ignored.

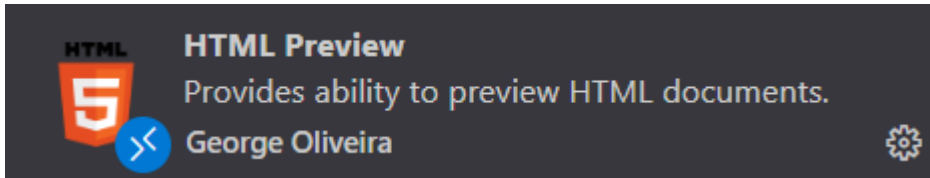
If this isn't relevant now, it will be at some point

Ran this command:

```
fastqc
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R1_001
.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R2_001
.fastq.gz -o output
```

Should specify delete option to delete intermediate zip files.

I don't know how to deal with html files here. I'm going to try an extension:



This works.

Control shift v to turn html file into a preview, same command to toggle off.

For the sake of experimentation, I'm also going to make folder test_output and run the same command with altered parameters:

```
fastqc
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R1_001
.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R2_001
.fastq.gz -o test_output --svg --extract --delete
```

^ This is the better way to do it. unzips everything into a nice folder

Final commands used:

```
mkdir -p 15_output
mkdir -p 17_output
fastqc
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R1_001
.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R2_001
.fastq.gz -o 15_output --svg --extract --delete
fastqc
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R1_001.
fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R2_001.
fastq.gz
```

Realized QAA has nothing in it. Remedied by activating environment:

```
conda install python=3.12
pip install matplotlib
```

```
pip install numpy # this didn't work because numpy comes along with matplotlib
```

Added sbatch scripts run_demux_plotting.sh run_fastqc.sh.

Stupid how fast the latter is compared to the former. Note that fastQC is coded in Java, generally a lot faster than python, and also likely optimized (taking the name on faith)

Part 2

```
conda install bioconda::cutadapt
cutadapt --version
conda install bioconda::trimmomatic
trimmomatic -version
```

Proceeded by selecting y, versions are correct.

Run requires naming the adapters:

For paired-end reads:

```
cutadapt -a ADAPT1 -A ADAPT2 [options] -o out1.fastq -p out2.fastq
in1.fastq in2.fastq
```

Replace "ADAPTER" with the actual sequence of your 3' adapter. IUPAC wildcard characters are supported. All reads from input.fastq will be written to output.fastq with the adapter sequence removed. Adapter matching is error-tolerant. Multiple adapter sequences can be given (use further -a options), but only the best-matching adapter will be removed.

Hunting for adapters

from:

https://support-docs.illumina.com/SHARE/AdapterSequences/Content/SHARE/AdapterSeq/Illumina_DNA/IlluminaUDIndexes.htm

We get these:

Index 1 (i7) Adapters -> affixed to R2

CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG

Index 2 (i5) Adapters

AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCTGCGCAGCGTC

But grepping these left me empty-handed because the reverse complement is what's actually sequenced.

R2 Actual Adapter Sequence:

- GTGTAGATCTCGGTGGTCGCCGTATCATT
 - *R1 Actual Adapter Sequence:
- ATCTCGTATGCCGTCTTCTGCTTG

but IDK which of these you use for cutadapt.

First, trying above sequences:

```
cutadapt -a ATCTCGTATGCCGTCTTCTGCTTG -A GTGTAGATCTCGGTGGTCGCCGTATCATT -o
17_R1_adapters_removed.fastq -p 17_R2_adapters_removed.fastq
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R1_001.
fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R2_001.
fastq.gz
```

hmm the actual number of sequences trimmed is paltry and my sequences don't match the ones in the assignment description.

Eventually found the correct adapters here:

<https://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2019/03/illumina-adapter-sequences-2019-1000000002694-10.pdf>

IDT for Illumina TruSeq DNA and RNA UD Indexes

These unique dual (UD) index adapters are arranged in the plate to enforce the recommended pairing strategy.

Adapter Trimming

The following sequences are used for adapter trimming.

Read 1

```
AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
```

Read 2

```
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
```

Index Adapters

Index 1 (i7) Adapters

```
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC [i7] ATCTCGTATGCCGTCTTCTGCTTG
```

Index 2 (i5) Adapters

```
AATGATACGGCGACCACCGAGATCTACAC [i5] ACACTCTTTCCCTACACGACGCTCTTCCGATCT
```

****Adapter trimming happens only at the 3' end! The adapters on the 5' end aren't sequenced, because the sequencing starts at the first base of the DNA insert. IFF the insert is too short, does some of the adapter on the opposite end get sequenced, that's what we're removing.**

Rerunning:

```
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o 17_R1_adapters_removed.fastq -p
17_R2_adapters_removed.fastq
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R1_001.
fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/17_3E_fox_S13_L008_R2_001.
fastq.gz
```

Output:

```
Finished in 59.628 s (5.060 µs/read; 11.86 M reads/minute).
```

```
=== Summary ===
```

```
Total read pairs processed:          11,784,410
```

Read 1 with adapter: 1,024,588 (8.7%)
Read 2 with adapter: 1,104,503 (9.4%)
Pairs written (passing filters): 11,784,410 (100.0%)

Total basepairs processed: 2,380,450,820 bp

Read 1: 1,190,225,410 bp

Read 2: 1,190,225,410 bp

Total written (filtered): 2,335,751,295 bp (98.1%)

Read 1: 1,168,027,279 bp

Read 2: 1,167,724,016 bp

=== First read: Adapter 1 ===

Sequence: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA; Type: regular 3'; Length: 33;
Trimmed: 1024588 times

Minimum overlap: 3

No. of allowed errors:

1-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-33 bp: 3

Bases preceding removed adapters:

A: 16.1%

C: 29.9%

G: 34.4%

T: 13.1%

none/other: 6.4%

Overview of removed sequences

length	count	expect	max.err	error	counts
3	225955	184131.4	0	225955	
4	66666	46032.9	0	66666	
5	35777	11508.2	0	35777	
6	25568	2877.1	0	25568	
7	24280	719.3	0	24280	
8	23497	179.8	0	23497	
9	23284	45.0	0	23009	275
10	23486	11.2	1	22633	853
11	22250	2.8	1	21493	757
12	22031	0.7	1	21267	764
13	21493	0.2	1	20780	713

14	20992	0.0	1	20277 715
15	20729	0.0	1	19970 759
16	20104	0.0	1	19316 788
17	19444	0.0	1	18610 834
18	18735	0.0	1	17954 761 20
19	17917	0.0	1	17041 840 36
20	17612	0.0	2	16716 789 107
21	16784	0.0	2	15947 745 92
22	16414	0.0	2	15556 749 109
23	16116	0.0	2	15207 805 104
24	15528	0.0	2	14709 716 103
25	15233	0.0	2	14396 735 102
26	14526	0.0	2	13654 772 100
27	13947	0.0	2	13057 770 110 10
28	13344	0.0	2	12519 718 95 12
29	12508	0.0	2	11698 695 104 11
30	12041	0.0	3	11191 707 105 38
31	11601	0.0	3	10771 714 84 32
32	10994	0.0	3	10186 650 104 54
33	10295	0.0	3	9573 584 105 33
34	9834	0.0	3	9170 543 92 29
35	9290	0.0	3	8657 527 76 30
36	8851	0.0	3	8263 477 88 23
37	8509	0.0	3	7954 478 61 16
38	8200	0.0	3	7673 443 57 27
39	7677	0.0	3	7192 412 45 28
40	7305	0.0	3	6858 392 45 10
41	6616	0.0	3	6218 344 38 16
42	6034	0.0	3	5664 325 38 7
43	5472	0.0	3	5140 284 35 13
44	4932	0.0	3	4638 266 23 5
45	4696	0.0	3	4393 273 24 6
46	4063	0.0	3	3826 202 29 6
47	3834	0.0	3	3607 191 24 12
48	3522	0.0	3	3326 164 25 7
49	3301	0.0	3	3076 190 28 7
50	3044	0.0	3	2873 154 14 3
51	2647	0.0	3	2492 133 19 3
52	2488	0.0	3	2350 117 14 7
53	2071	0.0	3	1937 111 14 9

54	1881	0.0	3	1783 84 11 3
55	1615	0.0	3	1525 75 14 1
56	1351	0.0	3	1279 65 6 1
57	1387	0.0	3	1306 73 8
58	1214	0.0	3	1143 65 2 4
59	1099	0.0	3	1032 55 11 1
60	1039	0.0	3	978 49 10 2
61	963	0.0	3	911 37 14 1
62	912	0.0	3	864 40 7 1
63	773	0.0	3	721 43 7 2
64	683	0.0	3	641 36 4 2
65	547	0.0	3	513 27 6 1
66	527	0.0	3	494 27 5 1
67	456	0.0	3	433 18 5
68	445	0.0	3	425 20
69	408	0.0	3	383 23 1 1
70	372	0.0	3	353 18 0 1
71	330	0.0	3	311 14 1 4
72	297	0.0	3	285 10 2
73	245	0.0	3	233 10 2
74	206	0.0	3	190 11 5
75	157	0.0	3	141 14 2
76	103	0.0	3	91 11 1
77	85	0.0	3	81 3 1
78	75	0.0	3	66 9
79	75	0.0	3	66 7 1 1
80	28	0.0	3	25 1 2
81	34	0.0	3	32 2
82	15	0.0	3	14 1
83	19	0.0	3	18 1
84	14	0.0	3	14
85	21	0.0	3	19 1 1
86	8	0.0	3	8
87	17	0.0	3	17
88	12	0.0	3	12
89	25	0.0	3	24 1
90	11	0.0	3	10 0 0 1
91	17	0.0	3	14 2 1
92	5	0.0	3	4 0 0 1
93	12	0.0	3	11 1

94	9	0.0	3	6 3
95	13	0.0	3	13
96	13	0.0	3	12 1
97	4	0.0	3	4
98	12	0.0	3	11 0 1
99	5	0.0	3	5
100	10	0.0	3	9 1
101	65502	0.0	3	6 58739 6317 440

=== Second read: Adapter 2 ===

Sequence: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT; Type: regular 3'; Length: 33;
Trimmed: 1104503 times

Minimum overlap: 3

No. of allowed errors:

1-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-33 bp: 3

Bases preceding removed adapters:

A: 16.3%

C: 32.8%

G: 37.0%

T: 7.9%

none/other: 5.9%

Overview of removed sequences

length	count	expect	max.err	error	counts
3	287280	184131.4	0		287280
4	79052	46032.9	0	79052	
5	37530	11508.2	0	37530	
6	26927	2877.1	0	26927	
7	24555	719.3	0	24555	
8	23609	179.8	0	23609	
9	23651	45.0	0	23225	426
10	23709	11.2	1	22832	877
11	22464	2.8	1	21739	725
12	22184	0.7	1	21611	573
13	21583	0.2	1	21051	532
14	21079	0.0	1	20572	507

15	20784	0.0	1	20143 641
16	20170	0.0	1	19626 544
17	19532	0.0	1	18908 624
18	18815	0.0	1	18062 746 7
19	17998	0.0	1	17386 599 13
20	17658	0.0	2	16908 669 81
21	16843	0.0	2	16105 657 81
22	16499	0.0	2	15813 622 64
23	16152	0.0	2	15491 578 83
24	15590	0.0	2	14847 663 80
25	15271	0.0	2	14511 663 97
26	14568	0.0	2	13851 622 95
27	14001	0.0	2	13356 540 103 2
28	13388	0.0	2	12746 553 85 4
29	12565	0.0	2	11905 564 88 8
30	12075	0.0	3	11447 536 69 23
31	11623	0.0	3	10739 743 108 33
32	11044	0.0	3	10417 493 108 26
33	10317	0.0	3	9727 473 78 39
34	9869	0.0	3	9297 451 85 36
35	9309	0.0	3	8797 424 65 23
36	8884	0.0	3	8408 384 61 31
37	8533	0.0	3	8029 408 62 34
38	8221	0.0	3	7735 375 66 45
39	7687	0.0	3	7255 339 65 28
40	7338	0.0	3	6925 324 60 29
41	6627	0.0	3	6298 259 38 32
42	6044	0.0	3	5720 261 42 21
43	5480	0.0	3	5196 235 26 23
44	4966	0.0	3	4700 210 36 20
45	4708	0.0	3	4444 220 28 16
46	4080	0.0	3	3892 155 24 9
47	3857	0.0	3	3650 166 20 21
48	3536	0.0	3	3332 175 21 8
49	3326	0.0	3	3154 128 26 18
50	3066	0.0	3	2890 143 25 8
51	2675	0.0	3	2514 128 20 13
52	2511	0.0	3	2380 104 21 6
53	2083	0.0	3	1971 89 13 10
54	1913	0.0	3	1803 85 11 14

55	1635	0.0	3	1506 102 18 9
56	1379	0.0	3	1292 71 11 5
57	1409	0.0	3	1317 68 17 7
58	1241	0.0	3	1168 59 10 4
59	1126	0.0	3	1043 62 13 8
60	1054	0.0	3	992 54 4 4
61	982	0.0	3	912 50 9 11
62	933	0.0	3	863 51 10 9
63	790	0.0	3	739 34 7 10
64	704	0.0	3	656 32 12 4
65	566	0.0	3	516 38 8 4
66	543	0.0	3	503 33 6 1
67	472	0.0	3	434 28 8 2
68	463	0.0	3	432 25 4 2
69	427	0.0	3	396 19 10 2
70	386	0.0	3	356 17 9 4
71	349	0.0	3	314 23 6 6
72	312	0.0	3	283 22 3 4
73	257	0.0	3	233 16 5 3
74	215	0.0	3	190 20 3 2
75	169	0.0	3	156 9 3 1
76	111	0.0	3	100 8 1 2
77	95	0.0	3	83 9 2 1
78	82	0.0	3	70 9 2 1
79	78	0.0	3	73 1 3 1
80	30	0.0	3	26 2 2
81	38	0.0	3	33 1 3 1
82	16	0.0	3	15 1
83	19	0.0	3	18 1
84	16	0.0	3	15 0 0 1
85	23	0.0	3	21 1 1
86	7	0.0	3	6 1
87	18	0.0	3	15 2 1
88	12	0.0	3	8 4
89	24	0.0	3	22 2
90	10	0.0	3	9 1
91	17	0.0	3	14 2 1
92	4	0.0	3	4
93	12	0.0	3	11 1
94	12	0.0	3	6 3 0 3

95	13	0.0	3	10 3
96	15	0.0	3	10 2 1 2
97	6	0.0	3	2 2 1 1
98	13	0.0	3	8 2 3
99	5	0.0	3	4 0 1
100	10	0.0	3	1 5 2 2
101	65176	0.0	3	6 58156 6462 552

Running:

```
cutadapt -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o 15_R1_adapters_removed.fastq -p
15_R2_adapters_removed.fastq
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R1_001
.fastq.gz
/projects/bgmp/shared/2017_sequencing/demultiplexed/15_3C_mbnl_S11_L008_R2_001
.fastq.gz
```

Output:

```
Processing paired-end reads on 1 core ...
Done          00:00:38      7,806,403 reads @   4.9 µs/read;  12.29 M
reads/minute
Finished in 38.104 s (4.881 µs/read; 12.29 M reads/minute).

=== Summary ===

Total read pairs processed:          7,806,403
  Read 1 with adapter:              417,810 (5.4%)
  Read 2 with adapter:              477,359 (6.1%)
Pairs written (passing filters):    7,806,403 (100.0%)

Total basepairs processed: 1,576,893,406 bp
  Read 1:   788,446,703 bp
  Read 2:   788,446,703 bp
Total written (filtered):  1,566,917,010 bp (99.4%)
  Read 1:   783,587,227 bp
  Read 2:   783,329,783 bp

=== First read: Adapter 1 ===
```

Sequence: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA; Type: regular 3'; Length: 33;
Trimmed: 417810 times

Minimum overlap: 3

No. of allowed errors:

1-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-33 bp: 3

Bases preceding removed adapters:

A: 17.0%

C: 29.2%

G: 37.9%

T: 15.8%

none/other: 0.2%

Overview of removed sequences

length	count	expect	max.err	error	counts
3	152119	121975.0	0	152119	
4	40634	30493.8	0	40634	
5	18152	7623.4	0	18152	
6	11433	1905.9	0	11433	
7	10612	476.5	0	10612	
8	9967	119.1	0	9967	
9	9897	29.8	0	9695	202
10	9771	7.4	1	9375	396
11	9272	1.9	1	8950	322
12	8958	0.5	1	8669	289
13	8435	0.1	1	8183	252
14	7935	0.0	1	7658	277
15	7804	0.0	1	7549	255
16	7350	0.0	1	7094	256
17	7259	0.0	1	6983	276
18	6676	0.0	1	6426	245 5
19	6321	0.0	1	6049	264 8
20	5859	0.0	2	5566	261 32
21	5824	0.0	2	5576	220 28
22	5455	0.0	2	5191	227 37
23	5009	0.0	2	4748	227 34
24	4769	0.0	2	4530	211 28
25	4490	0.0	2	4245	216 29

26	4138	0.0	2	3888 210 40
27	3984	0.0	2	3763 197 23 1
28	3806	0.0	2	3583 197 24 2
29	3369	0.0	2	3158 184 26 1
30	3224	0.0	3	3015 173 23 13
31	2883	0.0	3	2713 136 24 10
32	2616	0.0	3	2464 123 19 10
33	2510	0.0	3	2359 130 14 7
34	2256	0.0	3	2110 124 15 7
35	2010	0.0	3	1886 107 13 4
36	1944	0.0	3	1835 93 13 3
37	1836	0.0	3	1717 95 16 8
38	1672	0.0	3	1577 82 10 3
39	1575	0.0	3	1486 77 9 3
40	1355	0.0	3	1283 60 9 3
41	1188	0.0	3	1123 58 5 2
42	1074	0.0	3	1024 45 4 1
43	986	0.0	3	927 50 6 3
44	898	0.0	3	850 38 9 1
45	829	0.0	3	781 45 3
46	821	0.0	3	770 43 5 3
47	696	0.0	3	663 26 6 1
48	692	0.0	3	641 40 8 3
49	663	0.0	3	627 28 7 1
50	571	0.0	3	537 28 3 3
51	497	0.0	3	459 31 7
52	456	0.0	3	431 18 5 2
53	437	0.0	3	416 15 5 1
54	351	0.0	3	322 26 1 2
55	362	0.0	3	338 18 4 2
56	303	0.0	3	289 10 2 2
57	313	0.0	3	297 12 3 1
58	279	0.0	3	263 12 3 1
59	261	0.0	3	243 16 2
60	268	0.0	3	258 5 2 3
61	239	0.0	3	228 7 4
62	196	0.0	3	180 14 1 1
63	216	0.0	3	197 16 3
64	197	0.0	3	186 11
65	152	0.0	3	142 9 0 1

66	155	0.0	3	146 5 2 2
67	122	0.0	3	121 1
68	120	0.0	3	112 5 2 1
69	92	0.0	3	85 6 1
70	84	0.0	3	81 0 2 1
71	68	0.0	3	66 2
72	83	0.0	3	80 3
73	70	0.0	3	65 3 1 1
74	50	0.0	3	46 4
75	31	0.0	3	29 0 1 1
76	29	0.0	3	23 4 2
77	15	0.0	3	14 1
78	13	0.0	3	12 1
79	15	0.0	3	15
80	6	0.0	3	6
81	6	0.0	3	6
82	4	0.0	3	3 1
83	7	0.0	3	6 1
84	1	0.0	3	1
85	6	0.0	3	6
86	5	0.0	3	5
87	2	0.0	3	2
88	1	0.0	3	1
89	2	0.0	3	2
90	2	0.0	3	2
91	3	0.0	3	3
92	3	0.0	3	3
93	1	0.0	3	1
95	2	0.0	3	2
96	2	0.0	3	1 1
101	686	0.0	3	2 600 81 3

=== Second read: Adapter 2 ===

Sequence: AGATCGGAAGAGCGTCGTAGGGAAAGAGTGT; Type: regular 3'; Length: 33;
Trimmed: 477359 times

Minimum overlap: 3

No. of allowed errors:

1-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-33 bp: 3

Bases preceding removed adapters:

A: 19.7%

C: 29.8%

G: 41.0%

T: 9.4%

none/other: 0.1%

Overview of removed sequences

length	count	expect	max.err	error counts		
3	199425	121975.0	0	199425		
4	47599	30493.8	0	47599		
5	19744	7623.4	0	19744		
6	12401	1905.9	0	12401		
7	10799	476.5	0	10799		
8	10053	119.1	0	10053		
9	10101	29.8	0	9767	334	
10	9962	7.4	1	9460	502	
11	9490	1.9	1	9136	354	
12	9057	0.5	1	8791	266	
13	8483	0.1	1	8292	191	
14	8005	0.0	1	7815	190	
15	7863	0.0	1	7612	251	
16	7412	0.0	1	7224	188	
17	7300	0.0	1	7070	230	
18	6732	0.0	1	6426	304	2
19	6370	0.0	1	6165	201	4
20	5896	0.0	2	5661	200	35
21	5856	0.0	2	5582	242	32
22	5504	0.0	2	5276	197	31
23	5056	0.0	2	4850	179	27
24	4802	0.0	2	4575	197	30
25	4514	0.0	2	4303	183	28
26	4180	0.0	2	3950	201	29
27	4006	0.0	2	3820	154	32
28	3839	0.0	2	3635	176	28
29	3401	0.0	2	3220	150	28 3
30	3242	0.0	3	3080	134	19 9
31	2915	0.0	3	2662	202	33 18

32	2641	0.0	3	2494 116 18 13
33	2537	0.0	3	2380 118 30 9
34	2287	0.0	3	2147 108 23 9
35	2041	0.0	3	1923 89 18 11
36	1959	0.0	3	1832 100 21 6
37	1865	0.0	3	1738 94 27 6
38	1702	0.0	3	1590 89 14 9
39	1613	0.0	3	1512 73 16 12
40	1389	0.0	3	1311 56 13 9
41	1217	0.0	3	1138 56 12 11
42	1093	0.0	3	1040 38 12 3
43	1004	0.0	3	948 37 12 7
44	914	0.0	3	854 44 6 10
45	862	0.0	3	789 48 14 11
46	845	0.0	3	796 34 5 10
47	718	0.0	3	672 39 4 3
48	714	0.0	3	646 50 14 4
49	690	0.0	3	637 35 13 5
50	592	0.0	3	548 32 7 5
51	515	0.0	3	467 34 11 3
52	474	0.0	3	438 26 4 6
53	470	0.0	3	433 22 9 6
54	373	0.0	3	340 21 5 7
55	377	0.0	3	342 22 6 7
56	319	0.0	3	293 19 6 1
57	330	0.0	3	307 14 6 3
58	305	0.0	3	278 18 4 5
59	281	0.0	3	255 14 10 2
60	288	0.0	3	261 16 9 2
61	258	0.0	3	238 15 3 2
62	215	0.0	3	199 9 6 1
63	239	0.0	3	209 23 3 4
64	214	0.0	3	194 10 8 2
65	173	0.0	3	150 16 4 3
66	166	0.0	3	150 8 4 4
67	142	0.0	3	125 10 6 1
68	130	0.0	3	117 10 1 2
69	107	0.0	3	93 6 6 2
70	93	0.0	3	84 5 3 1
71	80	0.0	3	72 5 1 2

72	100	0.0	3	86 7 5 2
73	84	0.0	3	70 8 2 4
74	61	0.0	3	54 3 2 2
75	39	0.0	3	33 2 3 1
76	33	0.0	3	30 1 1 1
77	18	0.0	3	15 2 0 1
78	16	0.0	3	16
79	22	0.0	3	16 3 0 3
80	7	0.0	3	6 1
81	8	0.0	3	7 0 1
82	4	0.0	3	4
83	7	0.0	3	7
84	2	0.0	3	1 0 0 1
85	10	0.0	3	6 1 2 1
86	5	0.0	3	4 1
87	3	0.0	3	2 0 0 1
88	1	0.0	3	1
89	2	0.0	3	2
90	3	0.0	3	2 1
91	4	0.0	3	3 0 1
92	3	0.0	3	2 1
93	1	0.0	3	1
95	2	0.0	3	2
96	2	0.0	3	1 1
98	1	0.0	3	0 0 0 1
99	2	0.0	3	0 0 2
101	680	0.0	3	0 597 73 10

Trimming

Usage:

```
PE [-version] [-threads <threads>] [-phred33|-phred64] [-trimlog
<trimLogFile>] [-summary <statsSummaryFile>] [-quiet] [-validatePairs] [-
basein <inputBase> | <inputFile1> <inputFile2>] [-baseout <outputBase> |
<outputFile1P> <outputFile1U> <outputFile2P> <outputFile2U>] <trimmer1>...
```

Resource:

<http://www.usadellab.org/cms/?page=trimmomatic>

" For paired-end data, two input files are specified, and 4 output files, 2 for the 'paired' output

where both reads survived the processing, and 2 for corresponding 'unpaired' output where a read survived, but the partner read did not."

- LEADING: quality of 3
- TRAILING: quality of 3
- SLIDING WINDOW: window size of 5 and required quality of 15
- MINLENGTH: 35 bases

Be sure to output compressed files and clear out any intermediate files.

- Does file extensions with gz just do this automatically?

Commands:

```
trimmomatic PE -phred33 15_R1_adapters_removed.fastq
15_R2_adapters_removed.fastq 15_R1_adapters_removed_paired.fq.gz
15_R1_adapters_removed_unpaired.fq.gz 15_R2_adapters_removed_paired.fq.gz
15_R2_adapters_removed_unpaired.fq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15
MINLEN:35
```

OUTPUT:

TrimmomaticPE: Started with arguments:

```
-phred33 15_R1_adapters_removed.fastq 15_R2_adapters_removed.fastq
15_R1_adapters_removed_paired.fq.gz 15_R1_adapters_removed_unpaired.fq.gz
15_R2_adapters_removed_paired.fq.gz 15_R2_adapters_removed_unpaired.fq.gz
LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
Input Read Pairs: 7806403 Both Surviving: 7418387 (95.03%) Forward Only
Surviving: 377369 (4.83%) Reverse Only Surviving: 5675 (0.07%) Dropped: 4972
(0.06%)
```

TrimmomaticPE: Completed successfully

```
trimmomatic PE -phred33 17_R1_adapters_removed.fastq
17_R2_adapters_removed.fastq 17_R1_adapters_removed_paired.fq.gz
17_R1_adapters_removed_unpaired.fq.gz 17_R2_adapters_removed_paired.fq.gz
17_R2_adapters_removed_unpaired.fq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15
MINLEN:35
```

OUTPUT:

TrimmomaticPE: Started with arguments:

```
-phred33 17_R1_adapters_removed.fastq 17_R2_adapters_removed.fastq
17_R1_adapters_removed_paired.fq.gz 17_R1_adapters_removed_unpaired.fq.gz
```

```
17_R2_adapters_removed_paired.fq.gz 17_R2_adapters_removed_unpaired.fq.gz
LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:35
Input Read Pairs: 11784410 Both Surviving: 11240766 (95.39%) Forward Only
Surviving: 461940 (3.92%) Reverse Only Surviving: 8602 (0.07%) Dropped: 73102
(0.62%)
TrimmomaticPE: Completed successfully
```

Plot trimmed length distributions for R1 and R2, on the same plot

plot_length_dist.py

I guess put the paired and unpaired distributions on the same plot? Should I designate them as paired and unpaired?

Did it, committed it. Considering doing another bit where I plot cumulative distributions rather than the paired and unpaired reads separately.

Part 3

Matplotlib and numpy already installed above.

star

In QAA environment, run:

```
conda install star -c bioconda
conda activate QAA # prev install boots me from environment
conda install bioconda:htseq
```

Hmmm in retrospect, because of the being booted from the environment thing, I may have actually installed matplotlib and numpy in the base environment. Running:

```
conda install matplotlib
```

From there, all numpy packages were installed.

Making some folders, getting some data, unzipping:

```
mkdir gtf_primary_assembly
mkdir lib_17_ALIGNMENT
mkdir lib_15_ALIGNMENT
```

```
mkdir mouse_GENOME_INDEX
cd gtf_primary_assembly
wget https://ftp.ensembl.org/pub/release-
112/fasta/mus_musculus/dna/Mus_musculus.GRCm39.dna_sm.primary_assembly.fa.gz
wget https://ftp.ensembl.org/pub/release-
112/gtf/mus_musculus/Mus_musculus.GRCm39.112.gtf.gz
gunzip Mus_musculus.GRCm39.112.gtf.gz
Mus_musculus.GRCm39.dna_sm.primary_assembly.fa.gz
```

Corrected ps8 and pushed. Copying over code files:

```
cp /projects/bgmp/ghach/bioinfo/Bi621/PS/PS8/ps8-graceHach/src/alignReads.sh .
cp /projects/bgmp/ghach/bioinfo/Bi621/PS/PS8/ps8-graceHach/src/genGen.sh .
cp /projects/bgmp/ghach/bioinfo/Bi621/PS/PS8/ps8-graceHach/src/parse_sam.py .
```

Edited genGen.sh for this project.

```
sbatch genGen.sh
# Checked log files for successful completion
```

Edited alignReads.sh for two runs.

One thing I'm not sure about is what to do with the orphaned reads. For the time being, I'm leaving them out.

Ran super fast.

Created shell script parse_sam.sh.

```
chmod 755 parse_sam.sh
./parse_sam.sh

# RESULTS:
RESULTS for lib_17_ALIGNMENT/Aligned.out.sam
Number of reads mapped: 21532811
Number of reads unmapped: 483272
RESULTS for lib_15_ALIGNMENT/Aligned.out.sam
Number of reads mapped: 14436372
Number of reads unmapped: 207582
```

I think I'll modify this script to help answer #14.

htseq

```
htseq-count [options] <alignment_files> <gtf_file>

# ACTUAL COMMANDS RUN

htseq-count --stranded=yes --counts_output=lib17_strandedyes
lib_17_ALIGNMENT/Aligned.out.sam
gtf_primary_assembly/Mus_musculus.GRCm39.112.gtf
htseq-count --stranded=reverse --counts_output=lib17_strandedrev
lib_17_ALIGNMENT/Aligned.out.sam
gtf_primary_assembly/Mus_musculus.GRCm39.112.gtf
htseq-count --stranded=yes --counts_output=lib15_strandedyes
lib_15_ALIGNMENT/Aligned.out.sam
gtf_primary_assembly/Mus_musculus.GRCm39.112.gtf
htseq-count --stranded=reverse --counts_output=lib15_strandedrev
lib_15_ALIGNMENT/Aligned.out.sam
gtf_primary_assembly/Mus_musculus.GRCm39.112.gtf

# Put into script run_htcount.sh
```

From documentation:

"For `stranded=yes` and paired-end reads, the first read has to be on the same strand and the second read on the opposite strand. For `stranded=reverse`, these rules are reversed"

The Strandedness Question

What does it mean for a library to be stranded?

-> keep track of which strand is the coding strand, and which is the template strand.

It's stranded:

[illegible]

AAAAA SOMETHING BAD HAPPENED TO THE ENVIRONMENT. I cannot run
 ./plot_length_dist.sh in QAA environment. I get an incomprehensible error trying to save.
 Versions of matplotlib and numpy are the same:

```
(QAA) [ghach@n0349 QAA]$ conda list -n QAA numpy
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3/envs/QAA:
#
# Name          Version          Build    Channel
numpy           2.1.1            pypi_0   pypi
(QAA) [ghach@n0349 QAA]$ conda list numpy
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3/envs/QAA:
#
# Name          Version          Build    Channel
numpy           2.1.1            pypi_0   pypi
(QAA) [ghach@n0349 QAA]$ conda list matplotlib
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3/envs/QAA:
#
# Name          Version          Build    Channel
matplotlib      3.9.2            py312h7900ff3_0   conda-forge
matplotlib-base  3.9.2            py312h854627b_0   conda-forge
(QAA) [ghach@n0349 QAA]$ conda list -n QAA matplotlib
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3/envs/QAA:
#
# Name          Version          Build    Channel
matplotlib      3.9.2            py312h7900ff3_0   conda-forge
matplotlib-base  3.9.2            py312h854627b_0   conda-forge
```

Addressing the issues with plot_length_dist

Re-cloned the repo in Bi623/QAA_reclone, made environment QAA_redux.

Skipping demux_plotting and fastQC.

Corrected file paths in plot_length_dist.sh, committed and pushed.

Aaaaaaand it just worked. Probably made some major error when creating QAA, but no issues with QAA-redux, created according to the instructions in Assignmnet_2_Replication.

Compared versions and think this may have been the issue:

numpy	1.26.4	py312heda63a1_0	conda-forge	redux
numpy	2.1.1	pypi_0	pypi	orig


```
(base) [ghach@n0349 lab_notebook]$ conda list numpy
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3:
#
# Name          Version          Build          Channel
numpy           1.26.4           py312heda63a1_0  conda-forge
(base) [ghach@n0349 lab_notebook]$ conda list -n QAA_redux numpy
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3/envs/QAA_redux:
#
# Name          Version          Build          Channel
numpy           1.26.4           py312heda63a1_0  conda-forge
(base) [ghach@n0349 lab_notebook]$ conda list -n QAA numpy
# packages in environment at /gpfs/projects/bgmp/ghach/miniforge3/envs/QAA:
#
# Name          Version          Build          Channel
numpy           2.1.1            pypi_0          pypi
```