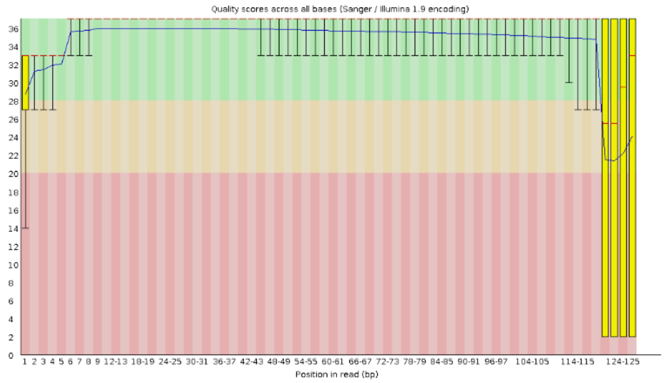
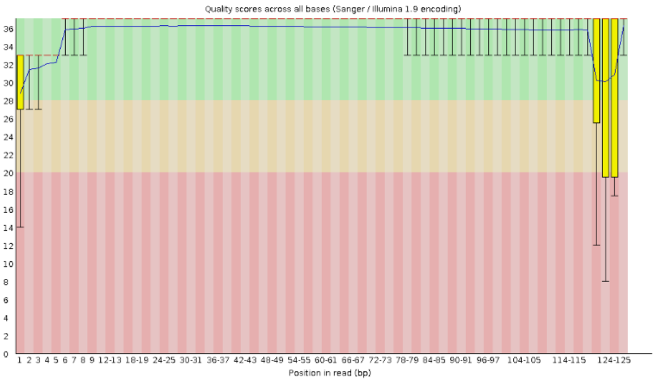
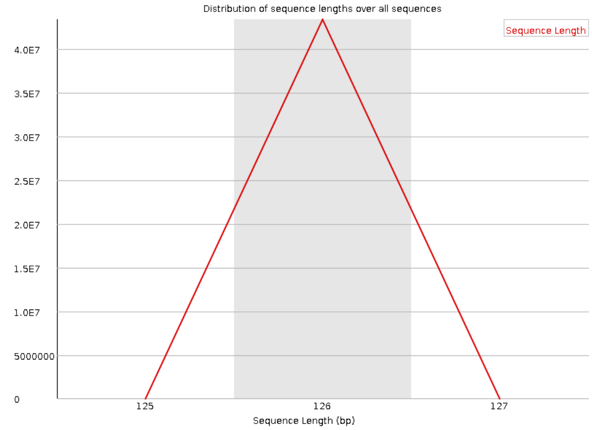
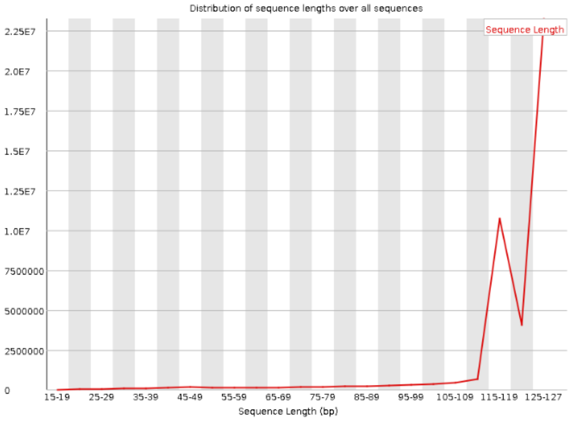
Assignment 4

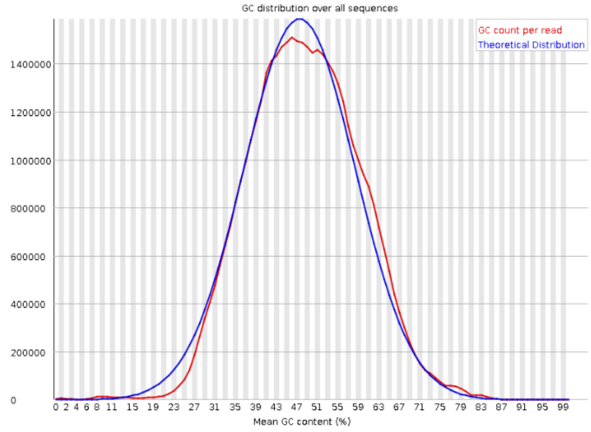
1. FastQC Before Trimming (Left) and after trimming with Phred score 30 (Right)
    a) SRR6188779 (il6-tnf)

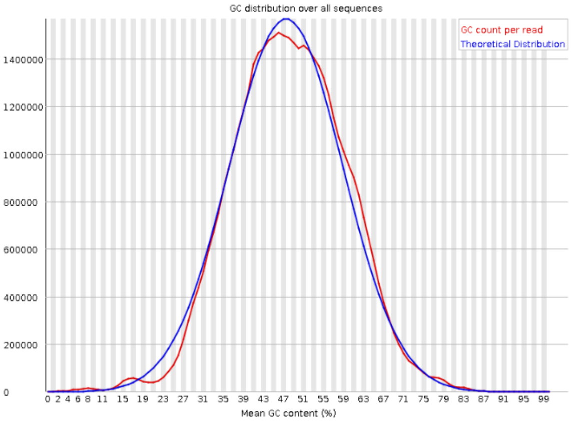






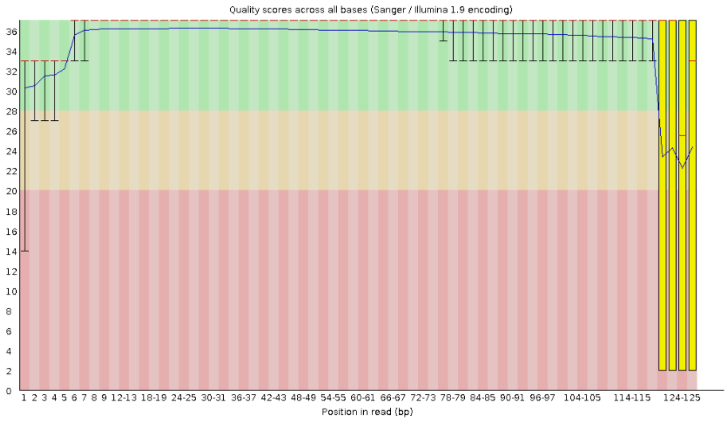


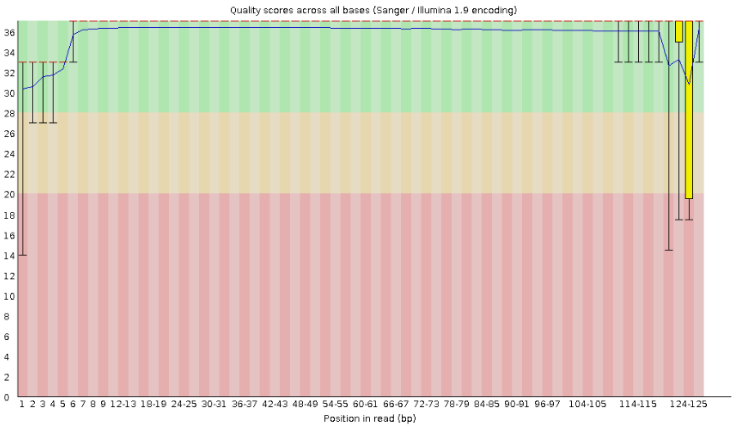


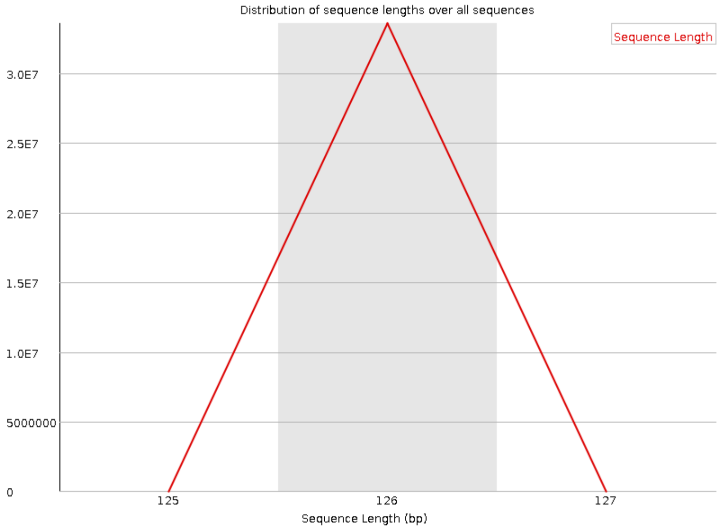


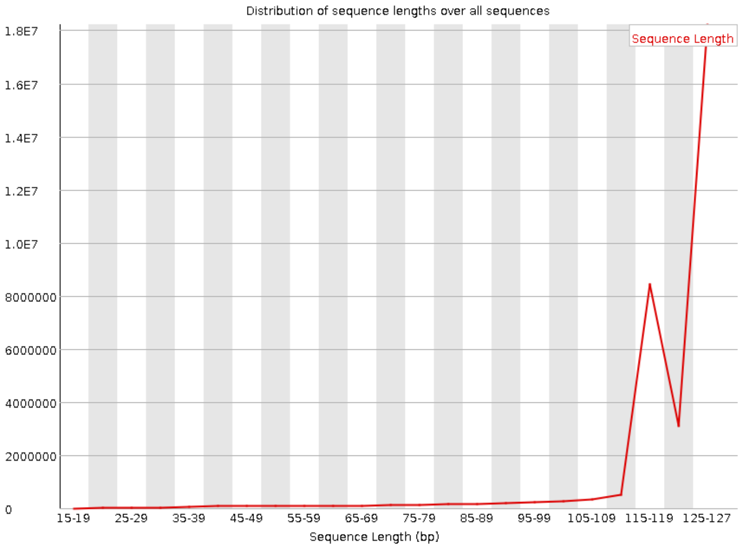
For the sequence quality distributions, we see that the before trimming the right side has a wider spread in the less desirable red region, but after trimming the spread is reduced. There is still some outliers in the red region, but the mean of the sequence fall within an acceptable quality score. The peak of the length distribution is around 126 in both distributions. However, after trimming there is a secondary peak at around 117, which means that a good portion of the reads have been reduced in length following the trimming process. I do notice that for the GC distribution has a dip on the left that falls below the theoretical distribution, but I am still not sure how to interpret this. Most of the examples in biostars show that this is not something to be concerned with. The peak of the GC distribution also falls below the theoretical distribution.

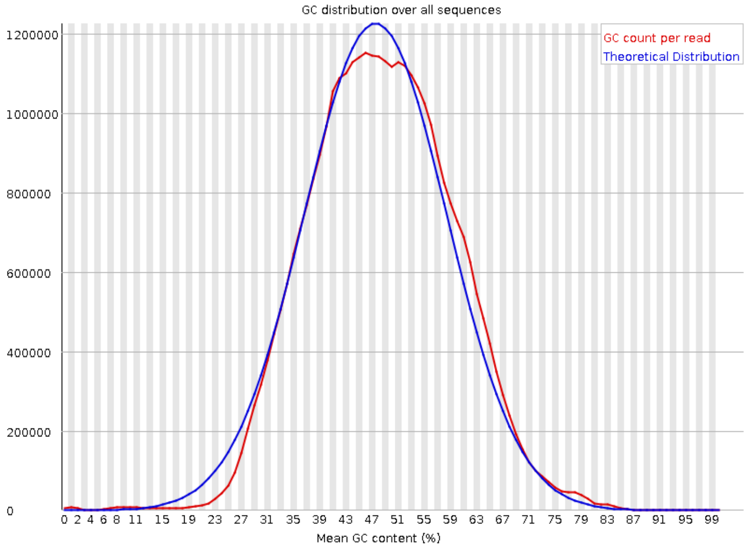
b) SRR6188780 (il6-tnf)

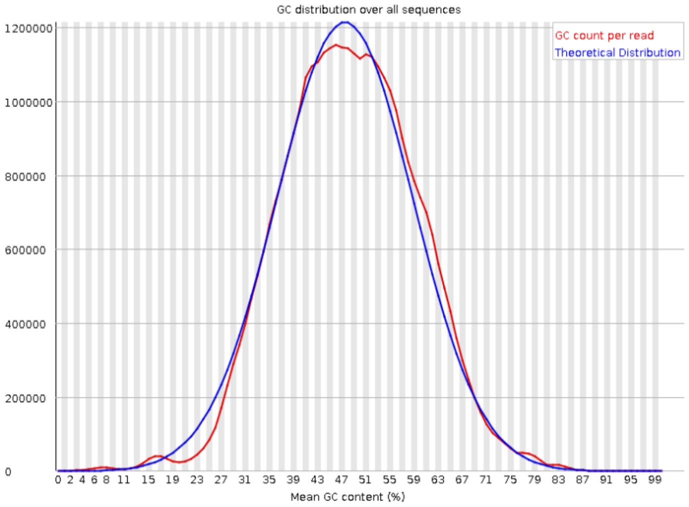






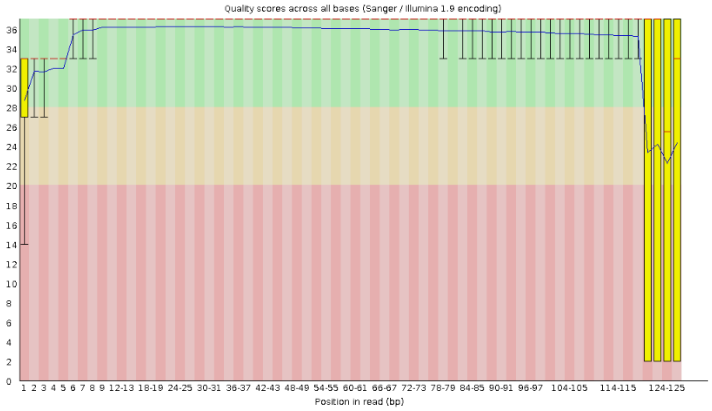


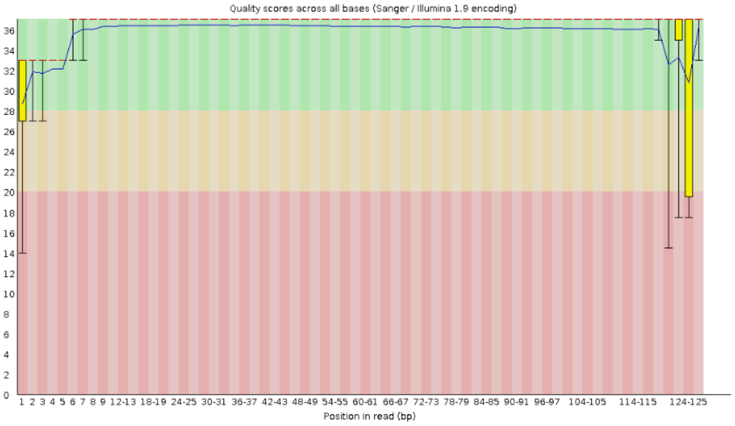


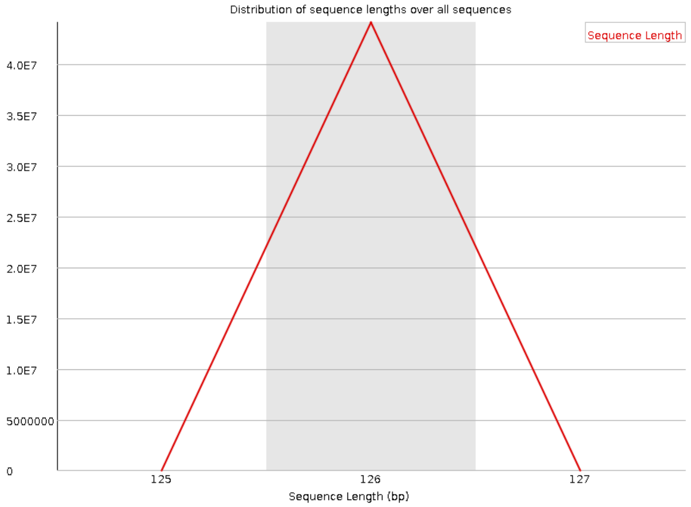


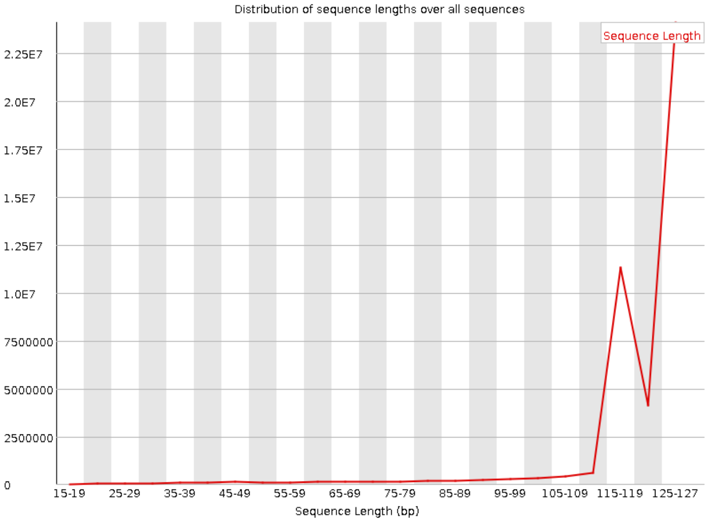
The similarity between this set of FastQC results and those from (a) were so small that I thought I was looking at the same data. But when I looked at the y-axis on each of the plots that's where I could see the biggest difference. The trimming once again had the same effect in reducing the spread of the 123-127 positions to within a more desirable quality score. The sequence length distribution peaks again at 126, but after trimming a few of the sequences are shortened to about 117 read lengths. And again we see the same pattern with the GC distribution.

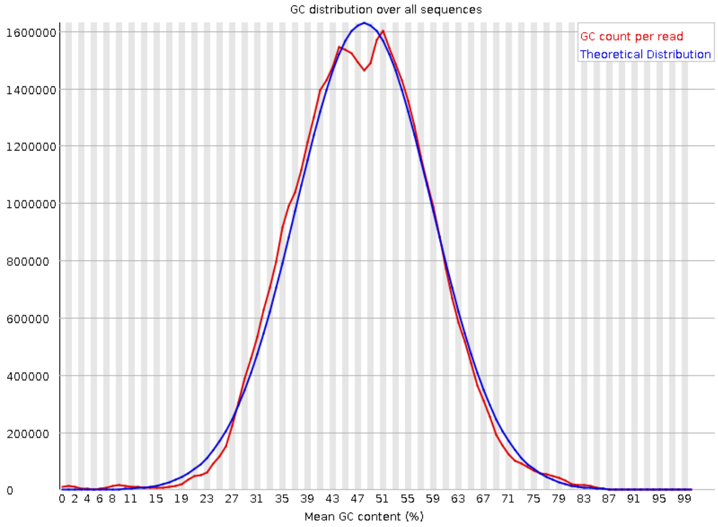
c) SRR6188783 (r848)

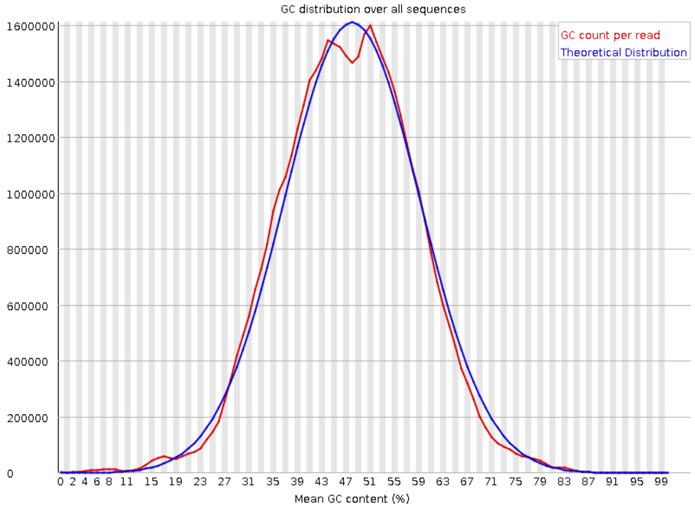






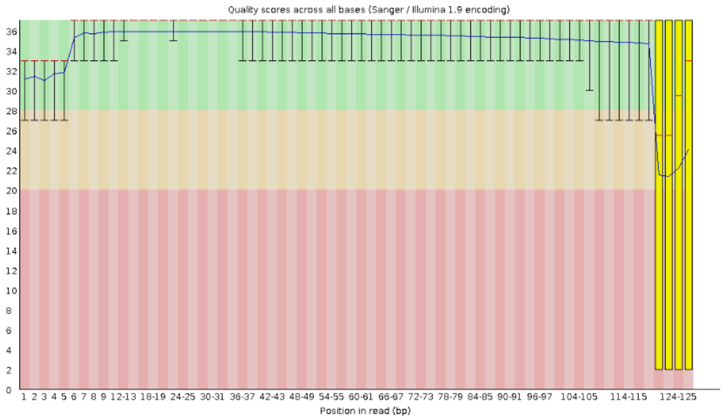


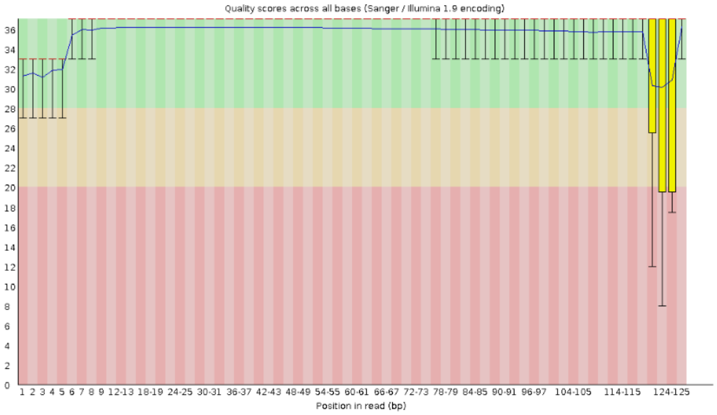


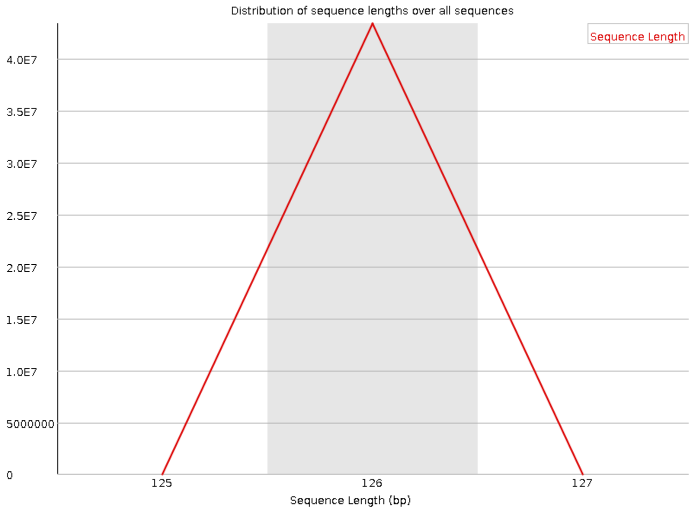


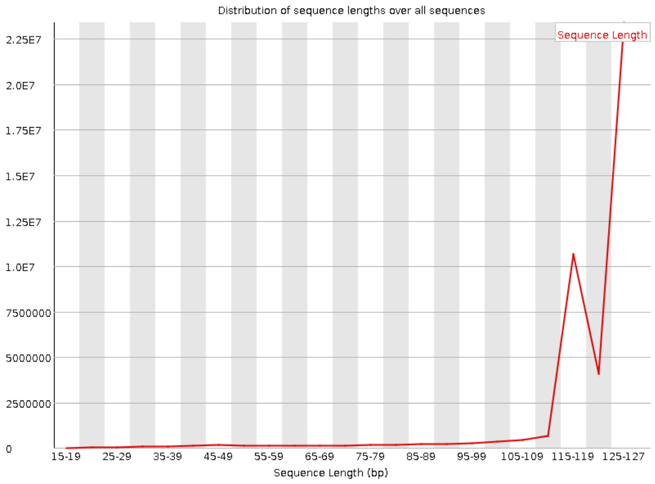
We are seeing the same patterns once again for the r848 treated sequences. Trimming once again reduces the spread of quality for sites 123-127. Sequence length had a similar pattern after trimming, indicating that a good portion of the read lengths were reduced to between 115-119 bp's.

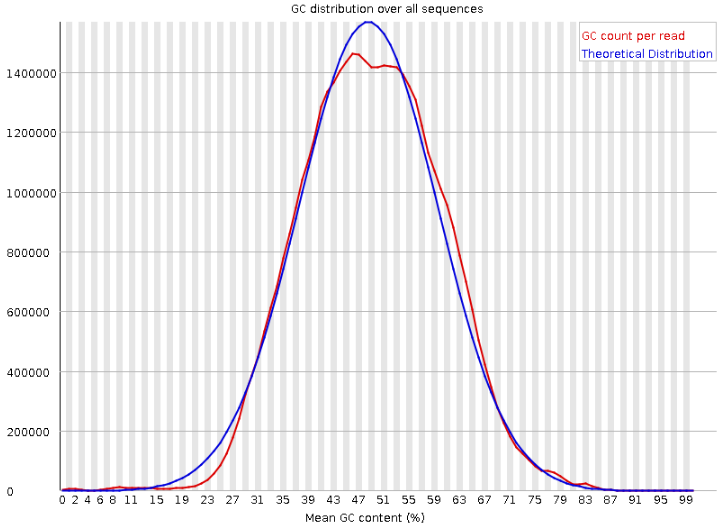
d) SRR6188774 (untreated)

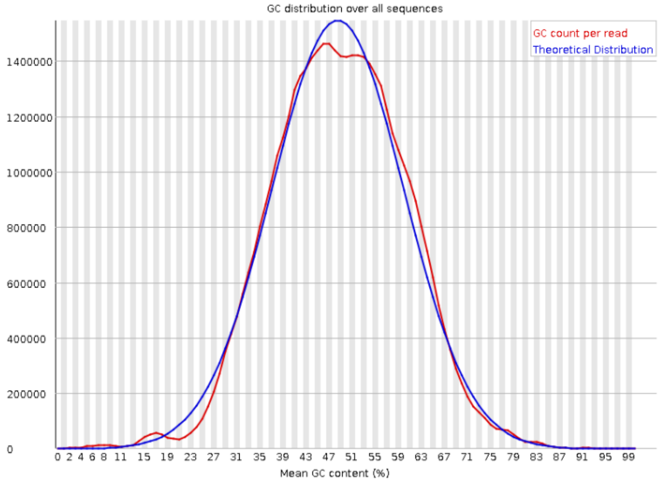












As with the r848 and il6-tnf treated sequences, the untreated sequences have the same patterns. Reading on some threads withing Biostars, this is a common pattern for good quality RNA-Seq reads, which is that the GC distribution falls below the theoretical distribution at the peak. Since the results had such similar patterns across each of the treated/untreated reads, I did not replicate the summaries of the FastQC outputs. However, the history files will be included.

2. Table of history numbers (galaxy file id numbers) and operations

Group	Paired reads	Trim ID's	HISAT2 ID	htseq-count ID
il6-tnf	SRR6188779 - SRR6188780	110,111	122	125
	SRR6188779 - SRR6188781	135,134	146	149
r848	SRR6188783 - SRR6188784	118,119	124	129
	SRR6188783 - SRR6188785	138,139	147	151
Control (Untreated)	SRR6188774 - SRR6188775	114,115	123	127
	SRR6188774 - SRR6188776	142,143	148	153

3. The HISAT2 hit a lot of errors, so I outline the steps in this section.
   I. Paired trimming of (a) SRR6188779 (il6-tnf) and (b) SRR6188780 (il6-tnf): `Error_Step_2I.txt`. Input file: `Inputs_Step_2I.txt`
   II. Since the paired trimming step I did not work, I proceeded to run HISAT2 on the unpaired trimmed reads for of (a) SRR6188779 (il6-tnf) and (b) SRR6188780 (il6-tnf) anyways as was done with the lecture video. Which produced the following error: `Error_Step_2II.txt`. Input file: `Inputs_Step_2II.txt.` I did not expect this to work, because they weren't trimmed with the paired setting, but I had to check anyways. After reading through biostars, it seemed to be an issue with Trim Galore and a cudapt dependency with a recent update. In order to move past this issue, I tried using a recommended similar program: Trimmomatic.
4. Trimmomatic paired trimming input files

I. Treatment il6-tnf:
   1. Paired trimming of SRR6188779 (il6-tnf) and SRR6188780 (il6-tnf) input file: `Trim_inputs_3I1.txt`.
   2. Paired trimming of SRR6188779 (il6-tnf) and SRR6188781 (il6-tnf) input file: `Trim_inputs_3I2.txt.`

II. Treatment r848:
   1. Paired trimming of SRR6188783 (r848) and SRR6188784 (r848) input file: `Trim_Input_3II1.txt`.
   2. Paired trimming of SRR6188783 (r848) and SRR6188785 (r848) input file: `Trim_Input_3II2.txt`.

III. Controls:
   1. Paired trimming of SRR6188774 (untreated) and SRR6188775 (untreated) input file: `Trim_inputs_3III1`.
   2. Paired trimming of SRR6188774 (untreated) and SRR6188776 (untreated) input file: `Trim_inputs_3III2.txt`.

I felt that this was a representative example of each pair of treatments.

5. After switching to Trimmomatic tool, the HISAT2 software performed on the paired trimmed reads from part 3 without error. The input files are below:

I. Treatment il6-tnf:
   1. HISAT2 for paired reads SRR6188779 (il6-tnf) and SRR6188780 (il6-tnf) input file: `HISAT2_inputs_4I1.txt`.
   2. HISAT2 for paired reads SRR6188779 (il6-tnf) and SRR6188781 (il6-tnf) input file: `HISAT2_inputs_4I2.txt`.

II. Treatment r848:
   1. HISAT2 for paired reads SRR6188783 (r848) and SRR6188784 (r848) input file: `HISAT2_inputs_4II1.txt`.
   2. HISAT2 for paired reads SRR6188783 (r848) and SRR6188785 (r848) input file: `HISAT2_inputs_4II2.txt`.

III. Control:

1. HISAT2 for paired reads SRR6188774 (untreated) and SRR6188775 (untreated) input file: `HISAT2_inputs_4III1.txt`.
2. HISAT2 for paired reads SRR6188774 (untreated) and SRR6188776 (untreated) input file: `HISAT2_inputs_4III2.txt`

6. The input files for the htseq-count programs

I. Treatment il6-tnf:

1. Htseq-count for SRR6188779 (il6-tnf) and SRR6188780 (il6-tnf) input file: `htseq_inputs_5I1.txt`.
2. Htseq-count for SRR6188779 (il6-tnf) and SRR6188781 (il6-tnf) input file: `htseq_inputs_5I1.txt`.

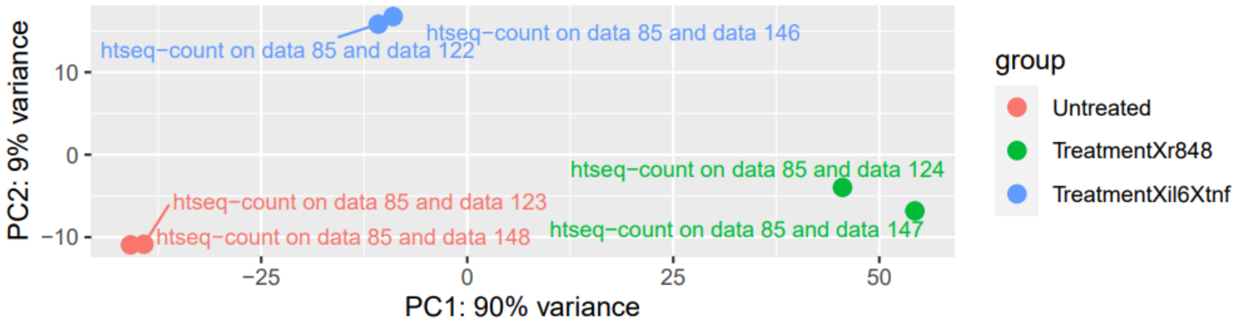
II. Treatment r848:

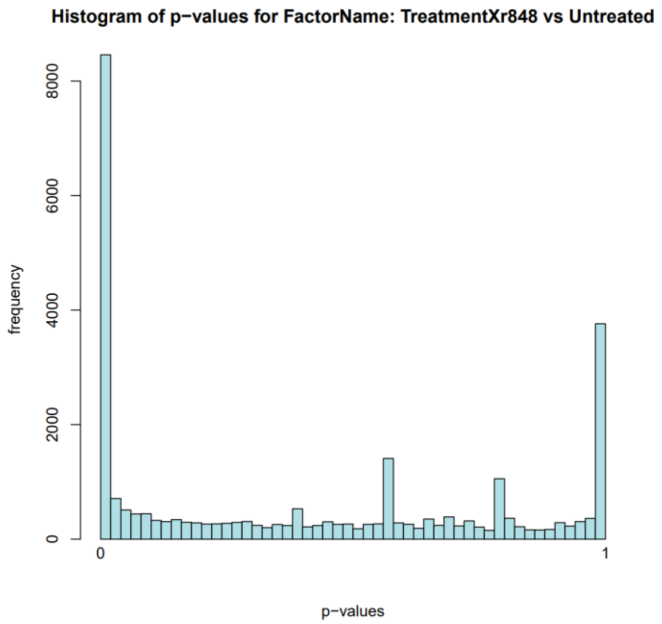
1. Htseq-count for paired reads SRR6188783 (r848) and SRR6188784 (r848) input file: `htseq_inputs_5II1.txt`.
2. Htseq-count for paired reads SRR6188783 (r848) and SRR6188785 (r848) input file: `htseq_inputs_5II2.txt`.

III. Control:

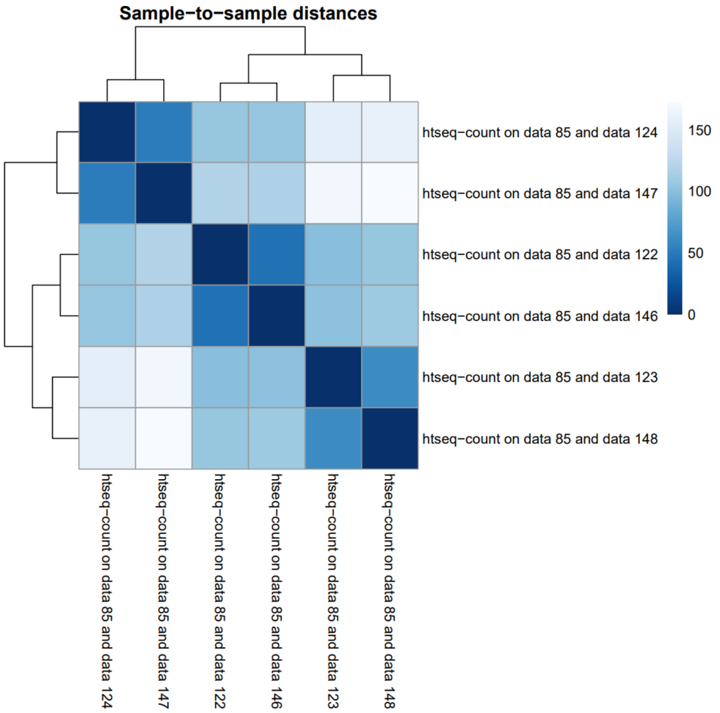
1. Htseq-count for paired reads SRR6188774 (untreated) and SRR6188775 (untreated) input file: `htseq_inputs_5III1.txt`.
2. Htseq-count for paired reads SRR6188774 (untreated) and SRR6188775 (untreated) input file: `htseq_inputs_5III2.txt`.

7. Next, we look at the PCA and histogram graphs resulting from the DESeq2 software





It’s hard to tell if there’s any outliers since I am not accustomed to seeing a PCA plot with such few observations. I’m assuming in practice you would replicate this analysis with much more samples and the outliers would become more obvious. In terms of variability within each group, the first two principal components demonstrate that the il6-tnf group and the untreated control group are much have less variability than the r848 treatment group. This is also seen in the hierarchical clustering heatmap:



The two r848 treatment groups are data 124 and 147 and the heatmap shows these being less correlated with each other than the other two groups.

The top 10-ish lines from the normalized counts table is:

```
     htseq-count on data 85 and data 123   htseq-count on data 85 and data 148   htseq-count on data 85 and data 124
     htseq-count on data 85 and data 147   htseq-count on data 85 and data 122   htseq-count on data 85 and data 146

ENSG00000000005.6   0       0       0       0       0       0

ENSG00000000419.12  220.535221703275       225.353056336052       568.707217726312       570.474005711788
     332.579339205698       284.37937277823

ENSG00000000457.14  97.1106800113855       124.923976881942       42.0687530920833       24.3792310133243
     72.8800403814955       61.7821001562927

ENSG00000000460.17  70.1702978146785       61.2372435695795       15.5810196637346       24.3792310133243
     41.0591776797158       30.8910500781464

ENSG00000000938.13  1485.4801439161        1524.80736488253       3709.8407819352        2574.44679500705
     2233.61926577654       1988.83848885478

ENSG00000000971.16  0.626520516202487      0.612372435695794      0       0       0       0

ENSG00000001036.14  506.228577091609       543.17435046217        306.946087375571       271.422105281677
     505.027885460504       506.068085103751

ENSG00000001084.13  474.902551281485       842.012099081717       855.397979539028       255.169284606128
     262.778737150181       244.402719735923

ENSG00000001167.14  264.391657837449       265.157264656279       468.988691878411       510.338569212255
     484.498296620646       498.799602732422

ENSG00000001460.18  4.38564361341741       3.06186217847897       3.11620393274691       4.87584620266486       0
     0.90856029641607
```

The order of the samples in the normalized counts file based on the history file and column titles are as follows:

1. SRR6188774 - SRR6188775
2. SRR6188774 - SRR6188776
3. SRR6188783 - SRR6188784
4. SRR6188783 - SRR6188785
5. SRR6188779 - SRR6188780
6. SRR6188779 - SRR6188781

1,2 are the control group. 3,4 are r848 treated group. 5,6 is il6-tnf treated group.