Results

Contents

As you can see in the table below, we have **343434** reads in the sample taken from *Escherichia coli* and *251128* reads for *Citrobacter freundii*. If you compare each library in proportion to the total number of bases, you will see that *Escherichia coli* library has more nucleotides and more reads.

Table 1: Raw Data Statistics

File Name	Number of Reads	Number of Bases
1-F5-96_S1_L001_R1_001	171717	38124416
1-F5-96_S1_L001_R2_001	171717	38253111
4-F20-96_S2_L001_R1_001	125564	27633180
4-F20-96_S2_L001_R2_001	125564	27727284

After trimming, we get 4 different outputs for each library. For each one of them, we have forward and reverse reads. For those reads, we have paired and unpaired reads.

For *Escherichia coli* library, `trimmomatic` drops **443** reads for forward strand and drops **8394** reads for reverse strand. If we look at *Citrobacter freundii*, **321** reads are dropped by `trimmomatic` for forward strand and **5926** reads are dropped for reverse strand. You can see the detail numbers in the table below.

Table 2: Statistics After Trimming

File Name	F5 Reads	F20 Reads	F20/F5
forward_paired	163188	119552	73.26%
forward_unpaired	8086	5691	70.37%
forward dropped	**443**	**321**	**72.26%**
reverse_paired	163188	119552	73.26%
reverse_unpaired	135	86	63.31%
reverse dropped	**8394**	**5926**	**70.59%**
Total Reads After Trimming	334597	244881	73.19%
Total Reads	343434	251128	73.12%

Also, you can see the comparasion of the raw reads and the trimmed reads in the histogram below.

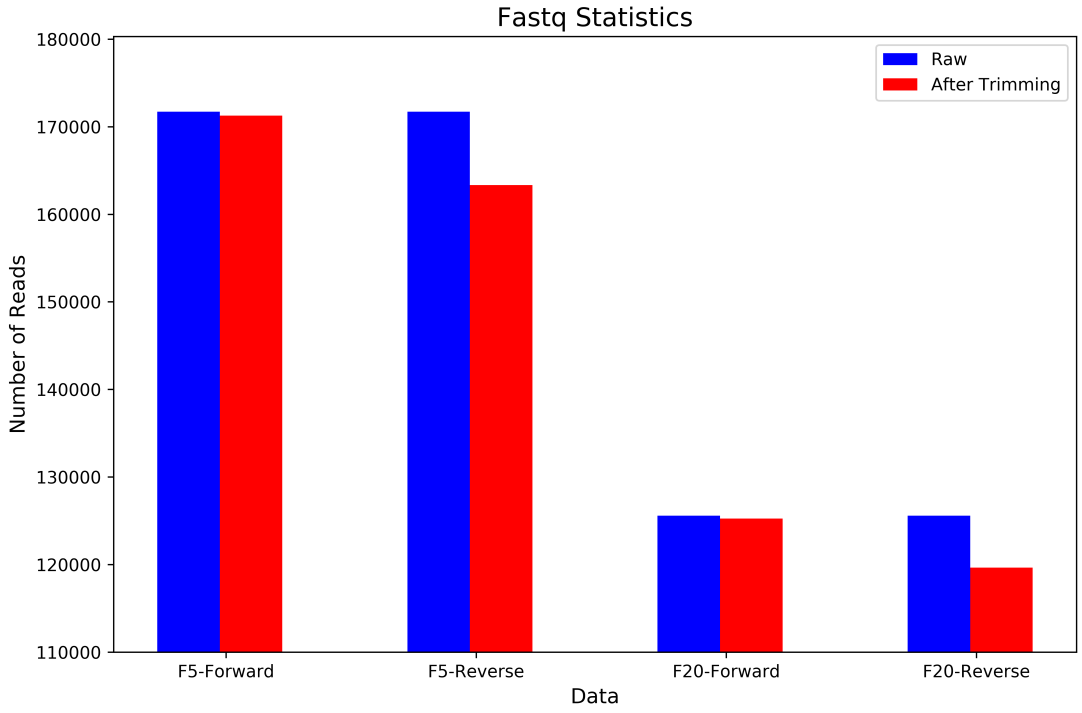


Figure 1: Number Of Reads Before and After Trimming

In the figure below, you can see the read lengths before and after trimming in both libraries. Mostly, the reads are conserved with the length of 250.

In the following figure, you can see quality values for each position in the reads with BoxWhisker type plots. Before trimming process, some reads inside the range of reasonable quality (orange) in both libraries. Trimming process eleminate these reads and our reads inside the range of good quality (green) only.

In the coming figure, you can see the GC content distribution over all sequences in both libraries. We expect that to see a roughly normal distribution but, some reads have lower mean GC content and some other reads have higher mean GC content.

We aligned our reads with `NCBI RefSeq Database` plasmids. You can see in the table below the total number of records of `NCBI RefSeq Database` are **15076**.

Table 3: Number of NCBI RefSeq Records

File Name	Number of Recors
plasmid.1.1.genomic.fna	4597
plasmid.2.1.genomic.fna	3234
plasmid.3.1.genomic.fna	2524

File Name	Number of Recors
plasmid.4.1.genomic.fna	3023
plasmid.5.1.genomic.fna	1698
Total	15076

The summary of the results is listed in the below table. We put a threshold for the mapping reads number which is `1000`. We use this number because we want to clear results. The threshold can be changed from the person who wants to follow this protocol. Also, we decided that at least half of the reference genome should be covered by our reads. Hence, the threshold for the breadth of coverage is 50%‘.

Table 4: Coverage Statistics

Library	Plasmids with >1K Mapping Reads	Plasmids where Breadth of Coverage >50%
Escherichia coli (F5)	173	45
Citrobacter freundii (F20)	135	36

These results can be interpreted wrongly, because we do not know the answers to the following questions:

1. Are the mapped reads separated in the plasmid?
2. Are the reads only mapped to particular regions of the reference plasmid?

The breadth of coverage results give us the answers of questions listed above. The following table shows the plasmids in the database with higher breadth of coverage.

Table 5: Top 5 Best Candidates

Library	Accession	Breadth of Coverage (%)
Escherichia coli (F5)	NC_025175.1	90.2%
Escherichia coli (F5)	NC_024956.1	88.3%
Escherichia coli (F5)	NC_025139.1	79.7%
Citrobacter freundii (F20)	NC_019049.1	77.7%
Escherichia coli (F5)	NC_004998.1	60.2%

We choose the plasmid with higher breadth of coverage (`NC_025175.1`) for further study. We also included in this study the plasmid `NC_025138.1`, that was suggested by one of the authors of the original paper, who also isolated the samples we study.

We find the number of mapping reads using `samtools view -F 4` command. As you can see in the table below, more reads map to the plasmid `NC_025175.1` than to `NC_025138.1`, in both libraries. The difference between the number of mapping reads is **243443** for *Escherichia coli* (F5) and **165817** for *Citrobacter freundii* (F20).

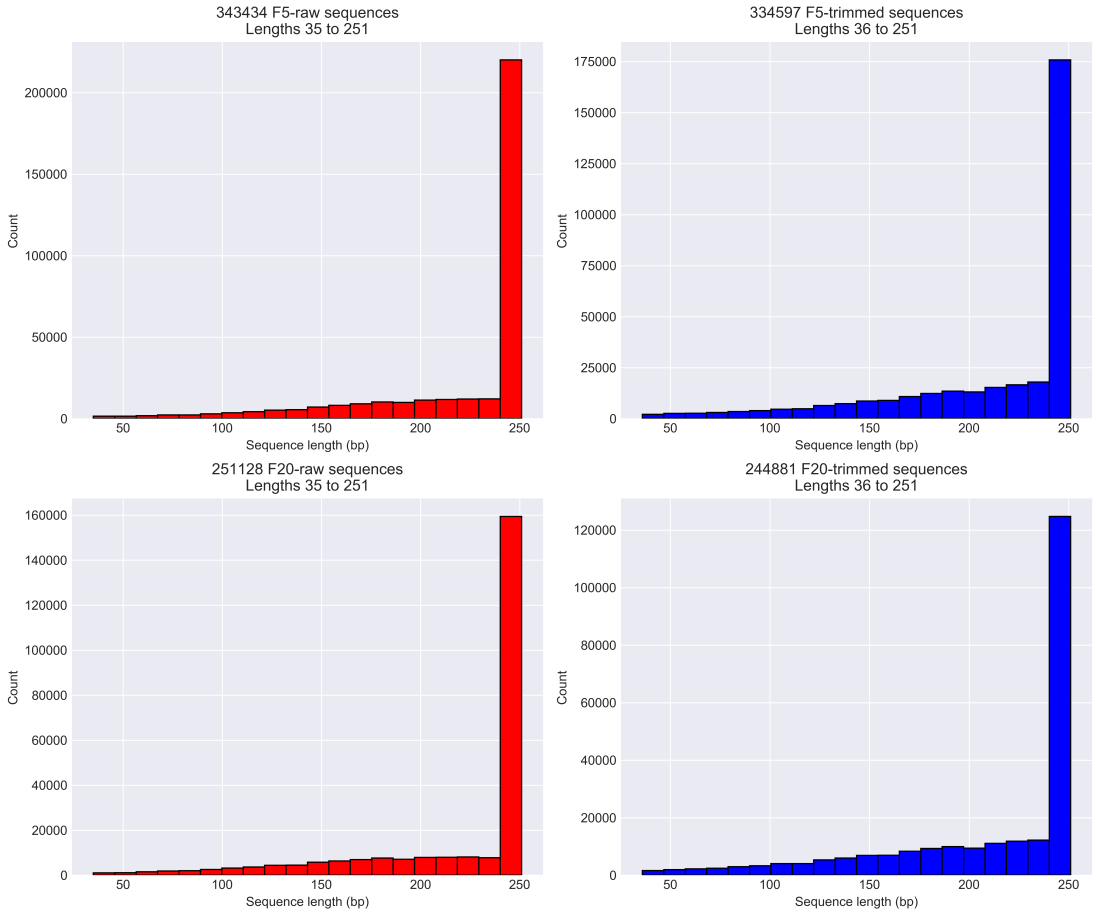
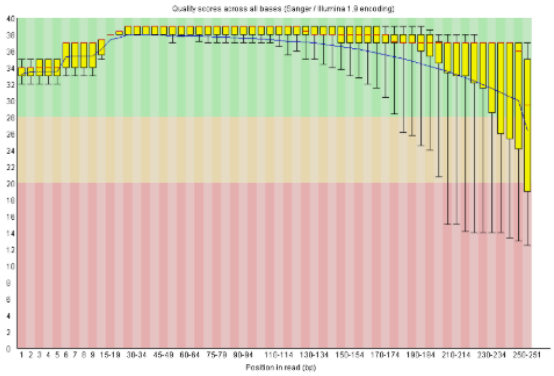
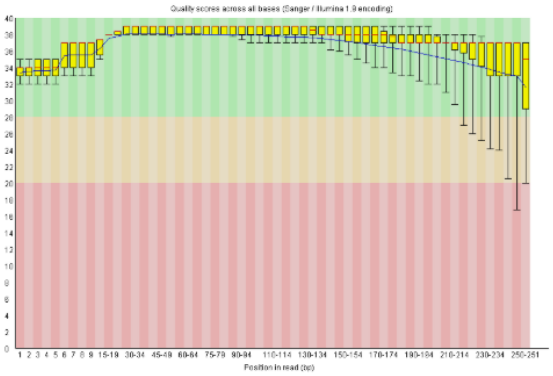


Figure 2: Histogram of Reads Length

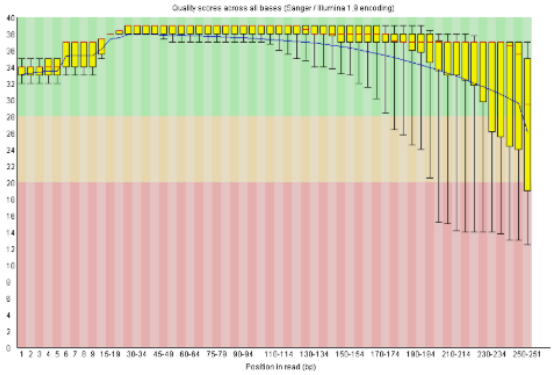
F5 Raw Reads



F5 Trimmed Reads



F20 Raw Reads



F20 Trimmed Reads

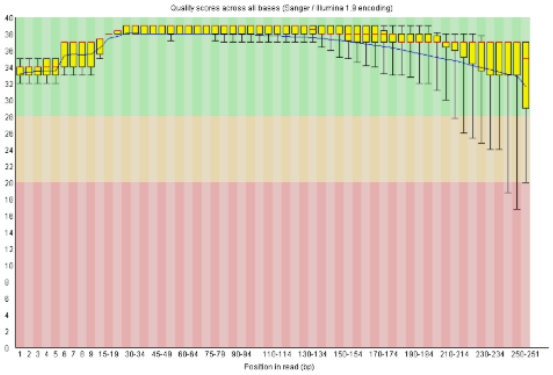
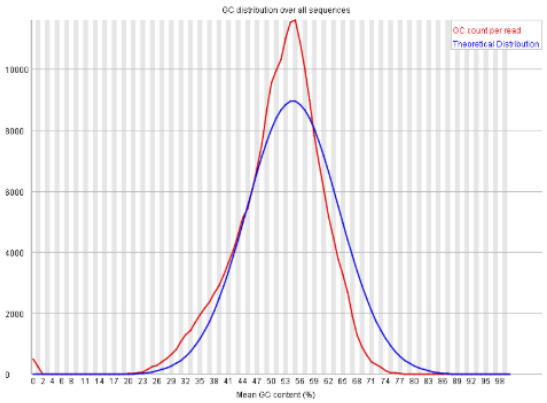
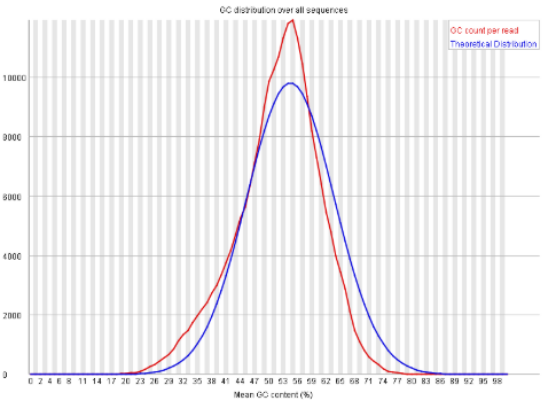


Figure 3: FastQC Per Base Sequence Quality BoxWhisker Plots Comparasion

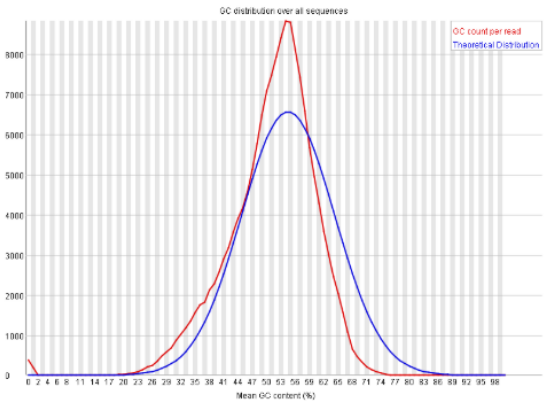
F5 Raw Reads



F5 Trimmed Reads



F20 Raw Reads



F20 Trimmed Reads

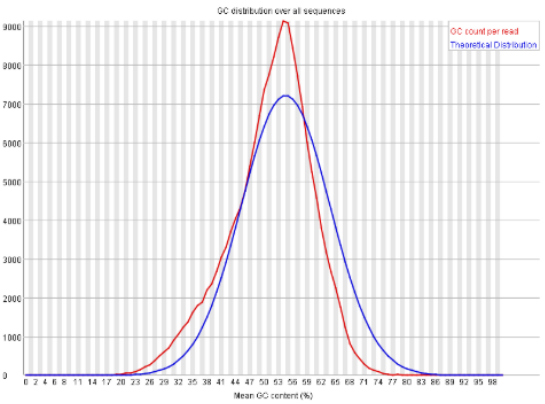


Figure 4: FastQC Per Sequence GC Content Plots Comparasion

Table 6: Number of Mapping Reads with Best Candidates

Library	Accession	Number of Mapping Reads	Total Reads After Trimming/Number of Mapping Reads
Escherichia coli (F5)	NC_025175.1	277167	82.84%
Escherichia coli (F5)	NC_025138.1	33724	10.08%
Citrobacter freundii (F20)	NC_025175.1	187823	76.70%
Citrobacter freundii (F20)	NC_025138.1	22006	8.99%

After choosing the best candidate, we want see the depth of coverage for each position.

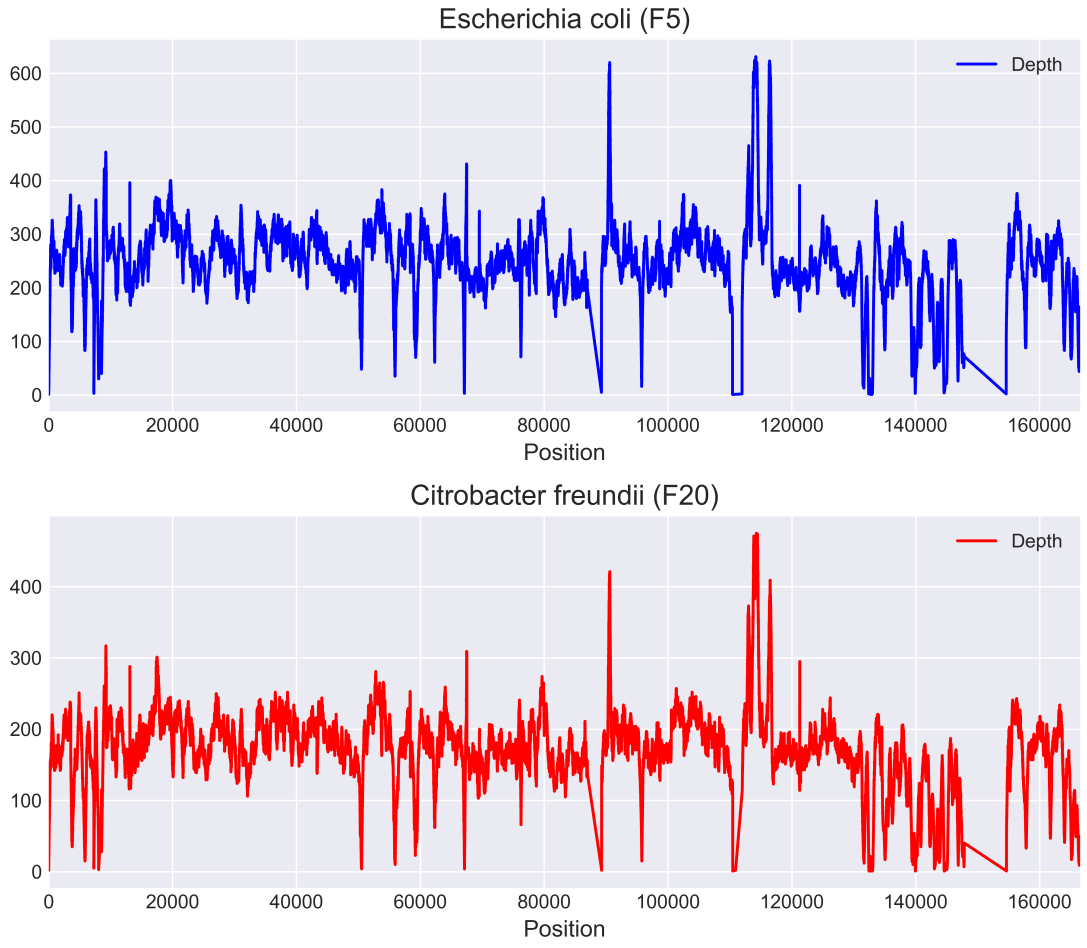


Figure 5: Depth of Coverage for Each Position

In the below table, we introduced the exact positions that are not covered from our reads.

Table 7: Positions With 0 Depth For NC_025175.1

Library	Position Start	Position End	End-Start
Escherichia coli (F5)	8061	8417	356
Escherichia coli (F5)	87006	89274	2268
Escherichia coli (F5)	110447	111986	1539
Escherichia coli (F5)	132435	132485	50
Escherichia coli (F5)	132738	132968	230
Escherichia coli (F5)	147876	154647	6771
Citrobacter freundii (F20)	8062	8417	355
Citrobacter freundii (F20)	87006	89274	2268
Citrobacter freundii (F20)	110447	110885	438
Citrobacter freundii (F20)	110903	111987	1084
Citrobacter freundii (F20)	132738	132896	158
Citrobacter freundii (F20)	132922	132964	42
Citrobacter freundii (F20)	139942	139959	17
Citrobacter freundii (F20)	144620	144652	32
Citrobacter freundii (F20)	147876	154650	6774

Also, some genes that are not covered from our reads. In the following table, you can see which genes that are not covered from our reads.

Table 8: Genes That Are Not Covered From Our Reads

Library	Accession	Strand	Locus Tag	Product
Both	NC_025175.1	(-)	D616_p149014	hypothetical protein
Both	NC_025175.1	(+)	D616_p149136	IS629 transposase
Both	NC_025175.1	(+)	D616_p149137	Mobile element protein
Both	NC_025175.1	(+)	D616_p149138	hypothetical protein
Both	NC_025175.1	(+)	D616_p149139	Retron-type RNA-directed DNA polymerase
Both	NC_025175.1	(-)	D616_p149173	AroG
Both	NC_025175.1	(-)	D616_p149174	hypothetical protein
Both	NC_025175.1	(-)	D616_p149207	hypothetical protein
Citrobacter freundii (F20)	NC_025175.1	(-)	D616_p149217	Relaxase /helicase
Citrobacter freundii (F20)	NC_025175.1	(+)	D616_p149224	Mercuric resistance operon coregulator, MerD
Both	NC_025175.1	(+)	D616_p149228	TniB NTP-binding protein
Both	NC_025175.1	(+)	D616_p149229	hypothetical protein
Both	NC_025175.1	(-)	D616_p149230	macrolide 2-phosphotransferase, mph(B)
Both	NC_025175.1	(-)	D616_p149231	hydrolase, alpha/beta fold family

Library	Accession	Strand	Locus Tag	Product
Both	NC_025175.1	(-)	D616_p149232	Transcriptional regulator, TetR family
Both	NC_025175.1	(-)	D616_p149233	hypothetical protein
Both	NC_025175.1	(-)	D616_p149234	Beta-lactamase
Both	NC_025175.1	(+)	D616_p149235	hypothetical protein
Both	NC_025175.1	(-)	D616_p149236	GGDEF family protein
Both	NC_025175.1	(+)	D616_p149237	TniB NTP-binding protein