# “What causes antibiotic resistance?”

Tags

## Made by Maria Uzun and Alisa Fedorenko

(wait a while for the pictures to load)

## 1. Get the data

Create a directory for Project 1 materials, and inside it create a new directory for raw data

```
mkdir Project1
cd Project1/
mkdir raw_data
cd raw_data
```

Download the reference sequence of the parental (unevolved, not resistant to antibiotics) E. coli strain (fna and gff)

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gff.gz
```

Download raw Illumina sequencing reads from shotgun sequencing of an E. coli strain that is resistant to the antibiotic ampicillin:

Go to https://figshare.com/articles/dataset/amp_res_2_fastq_zip/10006541/3

Press the button “Download all” and download raw sequence reads into “raw_data” directory
Unarchive downloaded data to the “Project1” directory

```
sudo apt-get install unzip #optionally
```

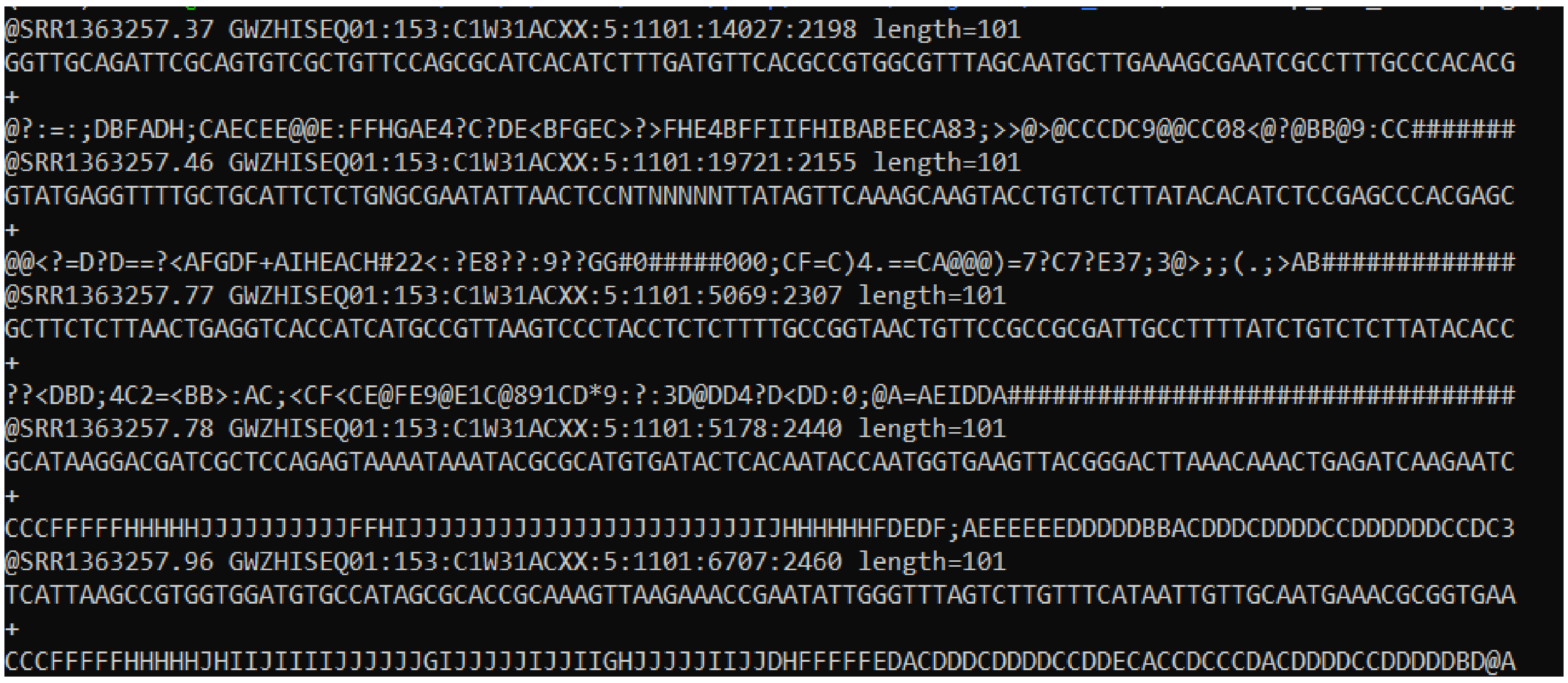
```
gzip -d -c GCF_000005845.2_ASM584v2_genomic.fna.gz > ../GCF_000005845.2_ASM584v2_genomic.fna
gzip -d GCF_000005845.2_ASM584v2_genomic.gff.gz -c > ../GCF_000005845.2_ASM584v2_genomic.gff
unzip 10006541.zip
```

## 2. Inspect raw sequencing data manually

Exit from “raw_data” directory

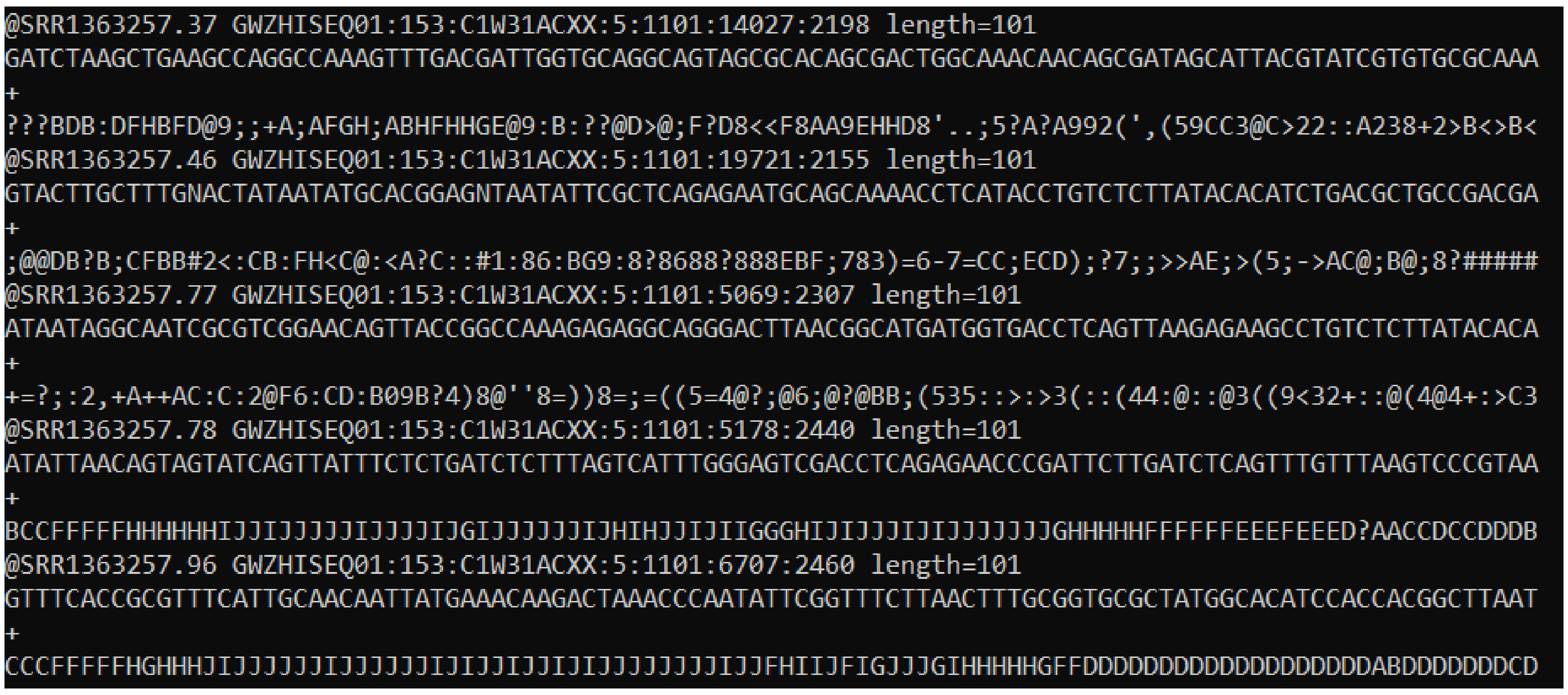
Open the sequence files and verify that the format is correct

```
zcat amp_res_1.fastq.gz|head -20
```



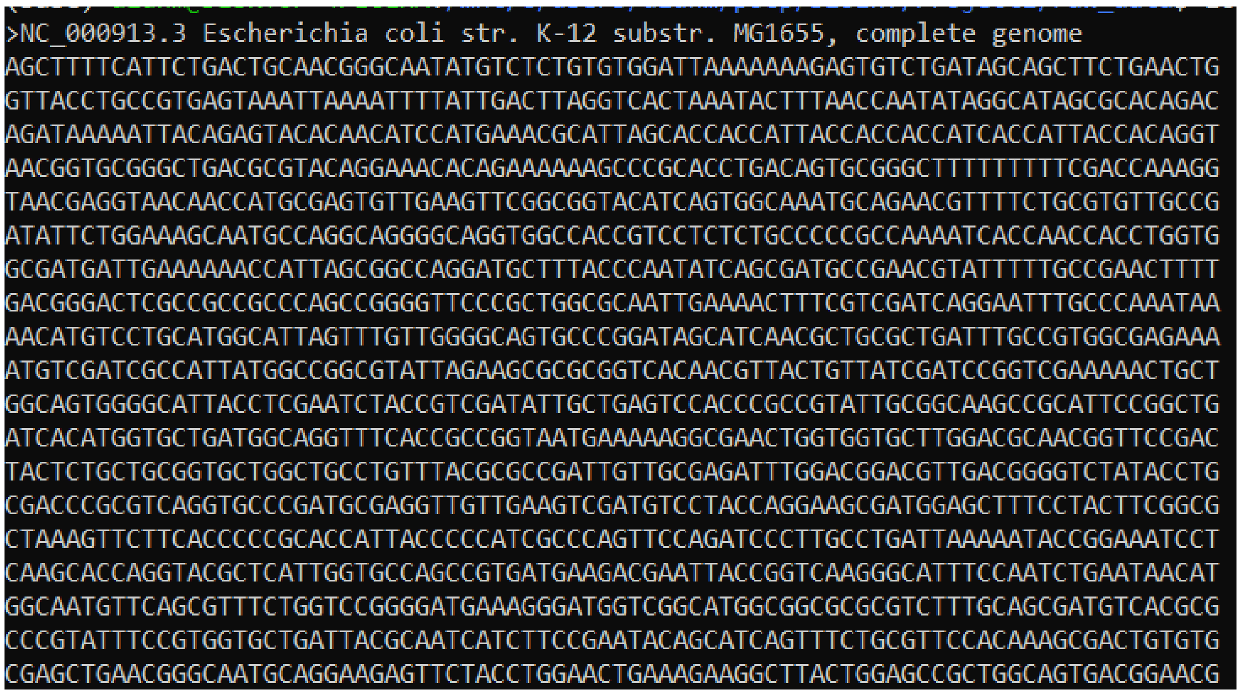
```
@SRR1363257.37 GWZHISEQ01:153:C1W31ACXX:5:1101:14027:2198 length=101
GGTTGCAGATTCGCAGTGTCGCTGTTCCAGCGCATCACATCTTTGATGTTCACGCCGTGGCGTTTAGCAATGCTTGAAAGCGAATCGCCTTTGCCCACACG
+
@?:=:;DBFADH;CAECEE@@E:FFHGAE4?C?DE<BFGEC>?>FHE4BFFIIFHIBABEECA83;>>@>@CCCDC9@@CC08<@?@BB@9:CC#######
@SRR1363257.46 GWZHISEQ01:153:C1W31ACXX:5:1101:19721:2155 length=101
GTATGAGGTTTTGCTGCATTCTCTGNGCGAATATTAACTCCNTNNNNNTTATAGTTCAAAGCAAGTACCTGTCTCTTATACACATCTCCGAGCCCACGAGC
+
@@<?=D?D==?<AFGDF+AIHEACH#22<:?E8??:9??GG#0#####000;CF=C)4.==CA@@@)=7?C7?E37;3@>;;(.;>AB#############
@SRR1363257.77 GWZHISEQ01:153:C1W31ACXX:5:1101:5069:2307 length=101
GCTTCTCTTAACTGAGGTCACCATCATGCCGTTAAGTCCCTACCTCTCTTTTGCCGGTAACTGTTCCGCCGCGATTGCCTTTTATCTGTCTCTTATACACC
+
??<DBD;4C2=<BB>:AC;<CF<CE@FE9@E1C@891CD*9:?:3D@DD4?D<DD:0;@A=AEIDDA##################################
@SRR1363257.78 GWZHISEQ01:153:C1W31ACXX:5:1101:5178:2440 length=101
GCATAAGGACGATCGCTCCAGAGTAAAATAAATACGCGCATGTGATACTCACAATACCAATGGTGAAGTTACGGGACTTAAACAAACTGAGATCAAGAATC
+
CCCFFFFFHHHHHJJJJJJJJJFFHIJJJJJJJJJJJJJJJJJJJJJJJJIJHHHHHHFDEDF;AEEEEEEDDDDDBBACDDDCDDDDCCDDDDDDCCDC3
@SRR1363257.96 GWZHISEQ01:153:C1W31ACXX:5:1101:6707:2460 length=101
TCATTAAGCCGTGGTGGATGTGCCATAGCGCACCGCAAAGTTAAGAAACCGAATATTGGGTTTAGTCTTGTTTCATAATTGTTGCAATGAAACGCGGTGAA
+
CCCFFFFFHHHHHJHIIJIIIIJJJJJJGIJJJJJIJJIIGHJJJJJIIJJDHFFFFFEDACDDDCDDDDCCDDECACCDCCCDACDDDDCCDDDDDBD@A
```

```
zcat amp_res_2.fastq.gz|head -20
```



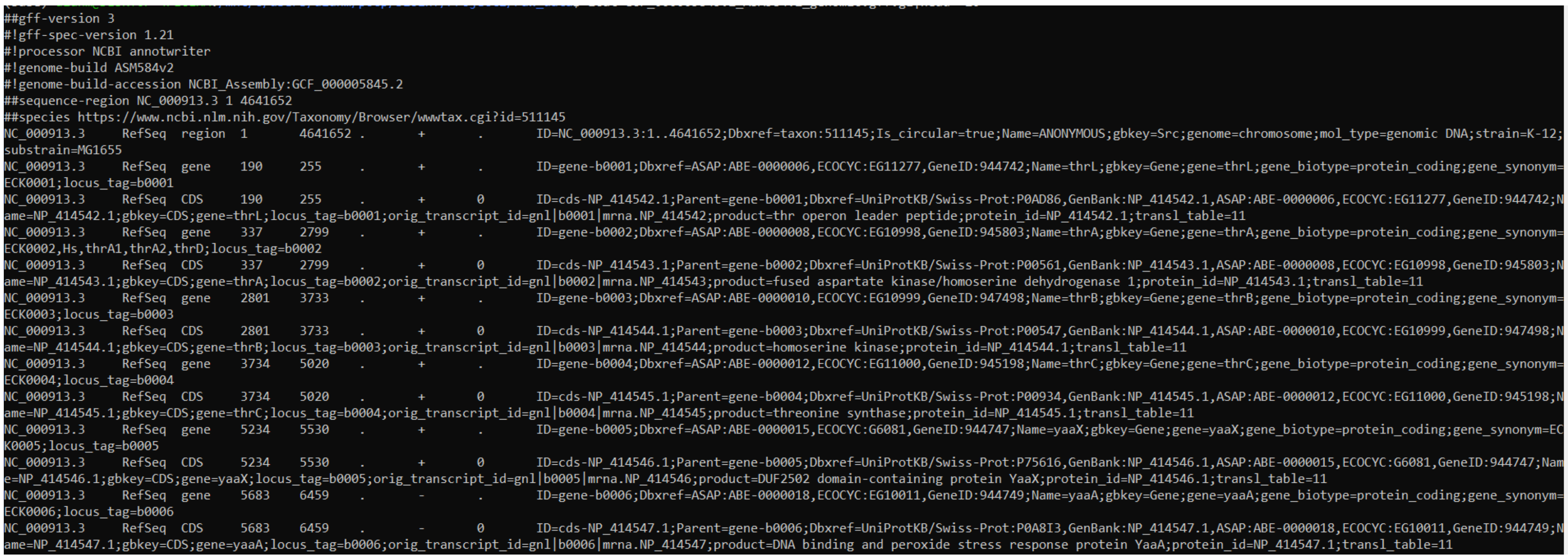
```
@SRR1363257.37 GWZHISEQ01:153:C1W31ACXX:5:1101:14027:2198 length=101
GATCTAAGCTGAAGCCAGGCCAAAGTTTGACGATTGGTGCAGGCAGTAGCGCACAGCGACTGGCAAACAACAGCGATAGCATTACGTATCGTGTGCGCAAA
+
???BDB:DFHBFD@9;;+A;AFGH;ABHFHHGE@9:B:??@D>@;F?D8<<F8AA9EHHD8'..;5?A?A992(',(59CC3@C>22::A238+2>B<>B<
@SRR1363257.46 GWZHISEQ01:153:C1W31ACXX:5:1101:19721:2155 length=101
GTACTTGCTTTGNACTATAATATGCACGGAGNTAATATTCGCTCAGAGAATGCAGCAAAACCTCATACCTGTCTCTTATACACATCTGACGCTGCCGACGA
+
;@@DB?B;CFBB#2<:CB:FH<C@:<A?C::#1:86:BG9:8?8688?888EBF;783)=6-7=CC;ECD);?7;;>>AE;>(5;->AC@;B@;8?#####
@SRR1363257.77 GWZHISEQ01:153:C1W31ACXX:5:1101:5069:2307 length=101
ATAATAGGCAATCGCGTCGGAACAGTTACCGGCCAAAGAGAGGCAGGGACTTAACGGCATGATGGTGACCTCAGTTAAGAGAAGCCTGTCTCTTATACACA
+
+=?;:2,+A++AC:C:2@F6:CD:B09B?4)8@''8=))8=;=((5=4@?;@6;@?@BB;(535::>:>3(::(44:@::@3((9<32+::@(4@4+:>C3
@SRR1363257.78 GWZHISEQ01:153:C1W31ACXX:5:1101:5178:2440 length=101
ATATTAACAGTAGTATCAGTTATTTCTCTGATCTCTTTAGTCATTTGGGAGTCGACCTCAGAGAACCCGATTCTTGATCTCAGTTTGTTTAAGTCCCGTAA
+
BCCFFFFFHHHHHHIJJIJJJJJIJJJJIJGIJJJJJJIJHIHJJIJIIGGGHIJIJJJIJIJJJJJJJGHHHHHFFFFFFEEEFEEED?AACCDCCDDDB
@SRR1363257.96 GWZHISEQ01:153:C1W31ACXX:5:1101:6707:2460 length=101
GTTTCACCGCGTTTCATTGCAACAATTATGAAACAAGACTAAACCCAATATTCGGTTTCTTAACTTTGCGGTGCGCTATGGCACATCCACCACGGCTTAAT
+
CCCFFFFFHGHHHJIJJJJJJIJJJJJJIJIJJIJJIJIJJJJJJJJIJJFHIIJFIGJJJGIHHHHHGFFDDDDDDDDDDDDDDDDDDDABDDDDDDDCD
```

```
zcat GCF_000005845.2_ASM584v2_genomic.fna.gz|head -20
```



```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTG
GTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGT
AACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGG
TAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCG
ATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTG
GCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAA
AACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAA
ATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCT
GGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTG
ATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGAC
TACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTG
CGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCG
CTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCT
CAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACAT
GGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCACGCG
CCCGTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGAATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTG
CGAGCTGAACGGGCAATGCAGGAAGAGTTCTACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACG
```

```
zcat GCF_000005845.2_ASM584v2_genomic.gff.gz|head -20
```



```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build ASM584v2
#!genome-build-accession NCBI_Assembly:GCF_000005845.2
##sequence-region NC_000913.3 1 4641652
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=511145
NC_000913.3     RefSeq  region  1       4641652 .       +       .       ID=NC_000913.3:1..4641652;Dbxref=taxon:511145;Is_circular=true;Name=ANONYMOUS;gbkey=Src;genome=chromosome;mol_type=genomic DNA;strain=K-12;
substrain=MG1655
NC_000913.3     RefSeq  gene    190     255     .       +       .       ID=gene-b0001;Dbxref=ASAP:ABE-0000006,ECOCYC:EG11277,GeneID:944742;Name=thrL;gbkey=Gene;gene=thrL;gene_biotype=protein_coding;gene_synonym=
ECK0001;locus_tag=b0001
NC_000913.3     RefSeq  CDS     190     255     .       +       0       ID=cds-NP_414542.1;Parent=gene-b0001;Dbxref=UniProtKB/Swiss-Prot:P0AD86,GenBank:NP_414542.1,ASAP:ABE-0000006,ECOCYC:EG11277,GeneID:944742;N
ame=NP_414542.1;gbkey=CDS;gene=thrL;locus_tag=b0001;orig_transcript_id=gnl|b0001|mrna.NP_414542;product=thr operon leader peptide;protein_id=NP_414542.1;transl_table=11
NC_000913.3     RefSeq  gene    337     2799    .       +       .       ID=gene-b0002;Dbxref=ASAP:ABE-0000008,ECOCYC:EG10998,GeneID:945803;Name=thrA;gbkey=Gene;gene=thrA;gene_biotype=protein_coding;gene_synonym=
ECK0002,Hs,thrA1,thrA2,thrD;locus_tag=b0002
NC_000913.3     RefSeq  CDS     337     2799    .       +       0       ID=cds-NP_414543.1;Parent=gene-b0002;Dbxref=UniProtKB/Swiss-Prot:P00561,GenBank:NP_414543.1,ASAP:ABE-0000008,ECOCYC:EG10998,GeneID:945803;N
ame=NP_414543.1;gbkey=CDS;gene=thrA;locus_tag=b0002;orig_transcript_id=gnl|b0002|mrna.NP_414543;product=fused aspartate kinase/homoserine dehydrogenase 1;protein_id=NP_414543.1;transl_table=11
NC_000913.3     RefSeq  gene    2801    3733    .       +       .       ID=gene-b0003;Dbxref=ASAP:ABE-0000010,ECOCYC:EG10999,GeneID:947498;Name=thrB;gbkey=Gene;gene=thrB;gene_biotype=protein_coding;gene_synonym=
ECK0003;locus_tag=b0003
NC_000913.3     RefSeq  CDS     2801    3733    .       +       0       ID=cds-NP_414544.1;Parent=gene-b0003;Dbxref=UniProtKB/Swiss-Prot:P00547,GenBank:NP_414544.1,ASAP:ABE-0000010,ECOCYC:EG10999,GeneID:947498;N
ame=NP_414544.1;gbkey=CDS;gene=thrB;locus_tag=b0003;orig_transcript_id=gnl|b0003|mrna.NP_414544;product=homoserine kinase;protein_id=NP_414544.1;transl_table=11
NC_000913.3     RefSeq  gene    3734    5020    .       +       .       ID=gene-b0004;Dbxref=ASAP:ABE-0000012,ECOCYC:EG11000,GeneID:945198;Name=thrC;gbkey=Gene;gene=thrC;gene_biotype=protein_coding;gene_synonym=
ECK0004;locus_tag=b0004
NC_000913.3     RefSeq  CDS     3734    5020    .       +       0       ID=cds-NP_414545.1;Parent=gene-b0004;Dbxref=UniProtKB/Swiss-Prot:P00934,GenBank:NP_414545.1,ASAP:ABE-0000012,ECOCYC:EG11000,GeneID:945198;N
ame=NP_414545.1;gbkey=CDS;gene=thrC;locus_tag=b0004;orig_transcript_id=gnl|b0004|mrna.NP_414545;product=threonine synthase;protein_id=NP_414545.1;transl_table=11
NC_000913.3     RefSeq  gene    5234    5530    .       +       .       ID=gene-b0005;Dbxref=ASAP:ABE-0000015,ECOCYC:G6081,GeneID:944747;Name=yaaX;gbkey=Gene;gene=yaaX;gene_biotype=protein_coding;gene_synonym=EC
K0005;locus_tag=b0005
NC_000913.3     RefSeq  CDS     5234    5530    .       +       0       ID=cds-NP_414546.1;Parent=gene-b0005;Dbxref=UniProtKB/Swiss-Prot:P75616,GenBank:NP_414546.1,ASAP:ABE-0000015,ECOCYC:G6081,GeneID:944747;Nam
e=NP_414546.1;gbkey=CDS;gene=yaaX;locus_tag=b0005;orig_transcript_id=gnl|b0005|mrna.NP_414546;product=DUF2502 domain-containing protein YaaX;protein_id=NP_414546.1;transl_table=11
NC_000913.3     RefSeq  gene    5683    6459    .       -       .       ID=gene-b0006;Dbxref=ASAP:ABE-0000018,ECOCYC:EG10011,GeneID:944749;Name=yaaA;gbkey=Gene;gene=yaaA;gene_biotype=protein_coding;gene_synonym=
ECK0006;locus_tag=b0006
NC_000913.3     RefSeq  CDS     5683    6459    .       -       0       ID=cds-NP_414547.1;Parent=gene-b0006;Dbxref=UniProtKB/Swiss-Prot:P0A8I3,GenBank:NP_414547.1,ASAP:ABE-0000018,ECOCYC:EG10011,GeneID:944749;N
ame=NP_414547.1;gbkey=CDS;gene=yaaA;locus_tag=b0006;orig_transcript_id=gnl|b0006|mrna.NP_414547;product=DNA binding and peroxide stress response protein YaaA;protein_id=NP_414547.1;transl_table=11
```

Open the entire fasta reference file. Do you notice anything different about it? I see circular chromosome sequence

```
zcat GCF_000005845.2_ASM584v2_genomic.fna.gz
```

**How many reads are in each fastq file?**

```
conda install -c bioconda seqkit
seqkit stats amp_res_1.fastq.gz
```

| File | Format | Type | **Num_seqs** | Sum_len | Min_len | Avg_len | Max_len |
|---|---|---|---|---|---|---|---|
| amp_res_1.fastq.gz | FASTQ | DNA | **455,876** | 46,043,476 | 101 | 101 | 101 |

```
seqkit stats amp_res_2.fastq.gz
```

| File | Format | Type | **Num_seqs** | Sum_len | Min_len | Avg_len | Max_len |
|---|---|---|---|---|---|---|---|
| amp_res_2.fastq.gz | FASTQ | DNA | **455,876** | 46,043,476 | 101 | 101 | 101 |

## 3. Inspect raw sequencing data with FastQC. Filtering the reads.

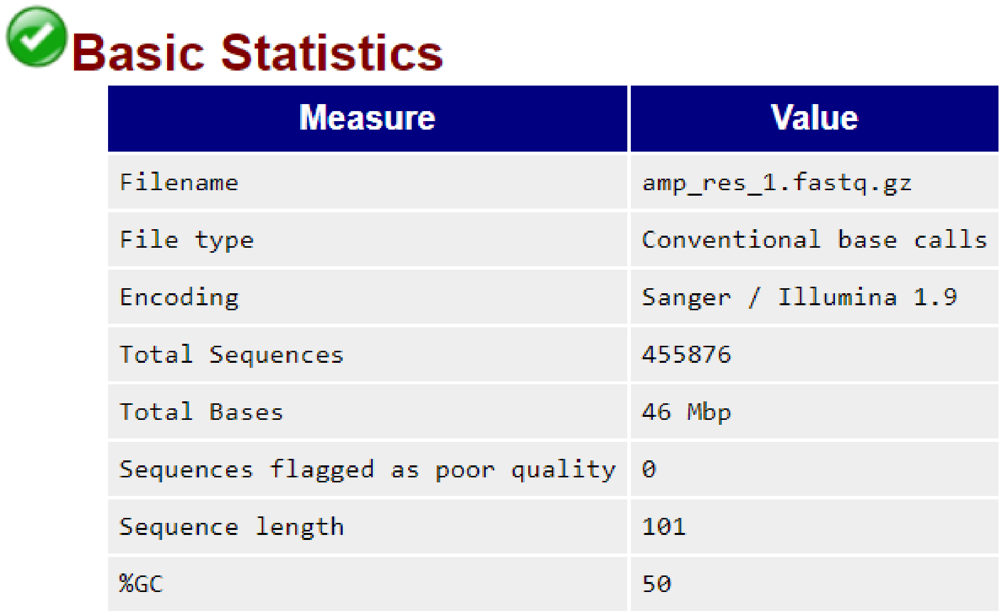
```
cd ..
mkdir fastqc
cd fastqc
fastqc -o ./ ../raw_data/amp_res_1.fastq.gz ../raw_data/amp_res_2.fastq.gz
```

amp_res_1_fastqc.html

amp_res_2_fastqc.html

**Do the basic statistics match what you calculated for the number of reads last time?**

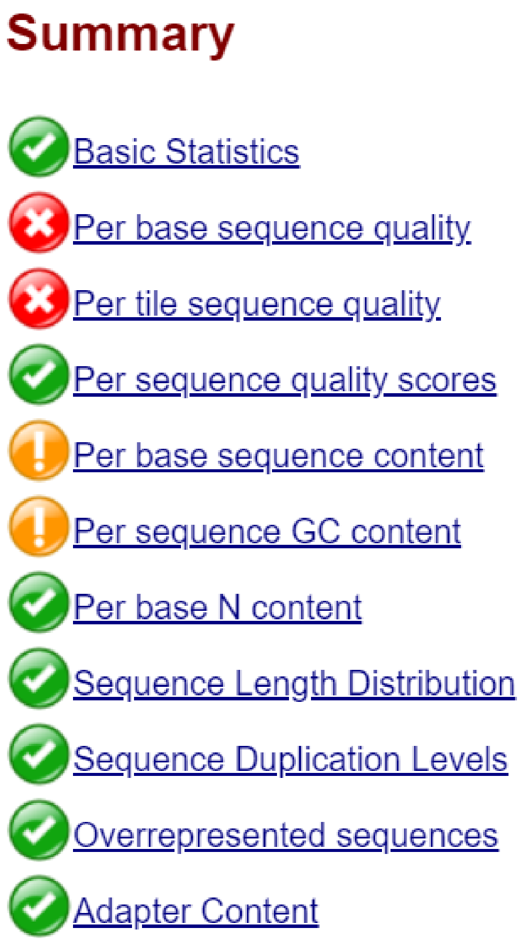
Yes, it matches



Basic Statistics

| Measure | Value |
|---|---|
| Filename | amp_res_1.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 455876 |
| Total Bases | 46 Mbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 101 |
| %GC | 50 |

In generated files, we see red circles at **Per base sequence quality** and at **Per tile sequence quality**

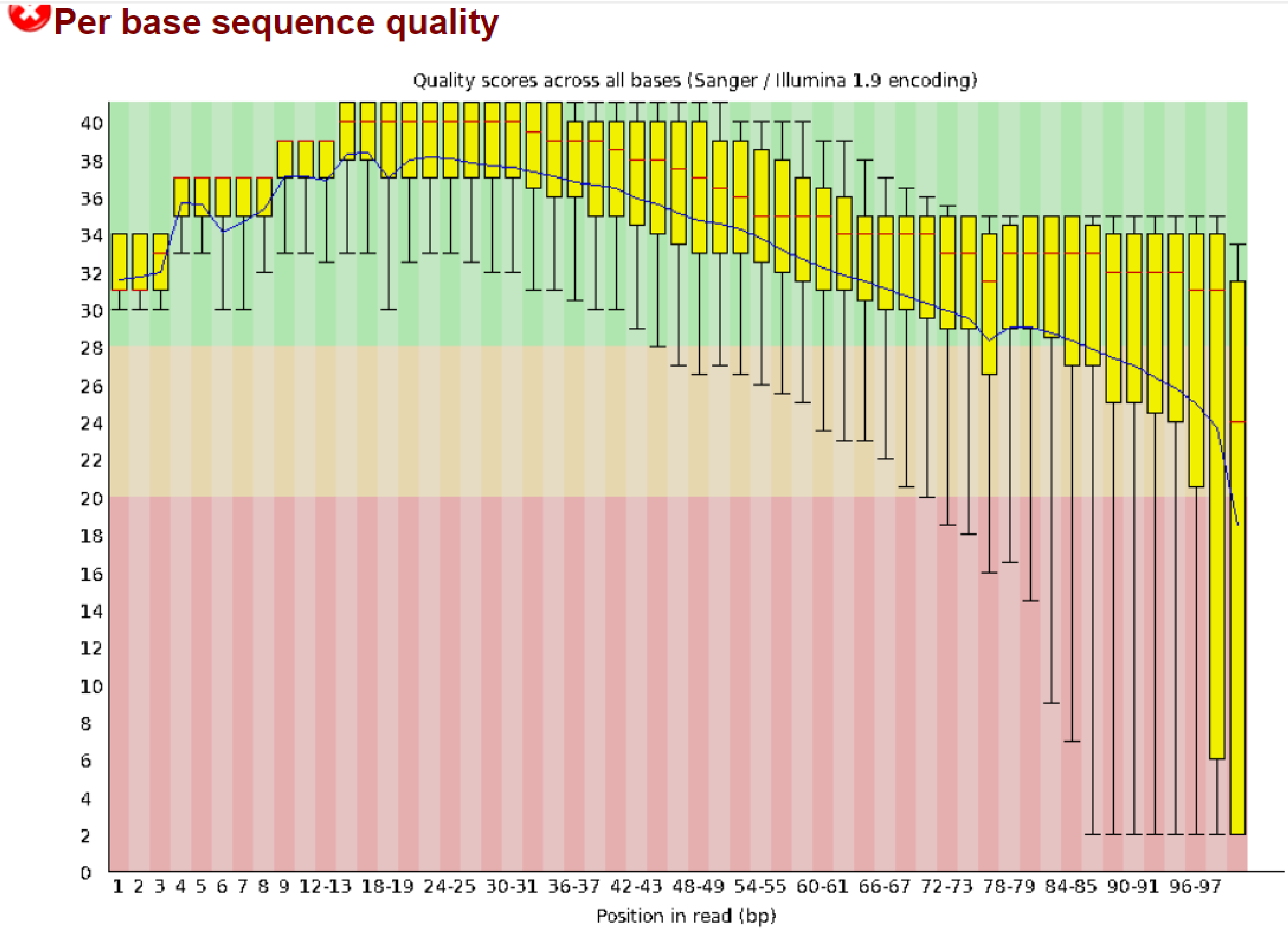


## Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

**What do they mean?**

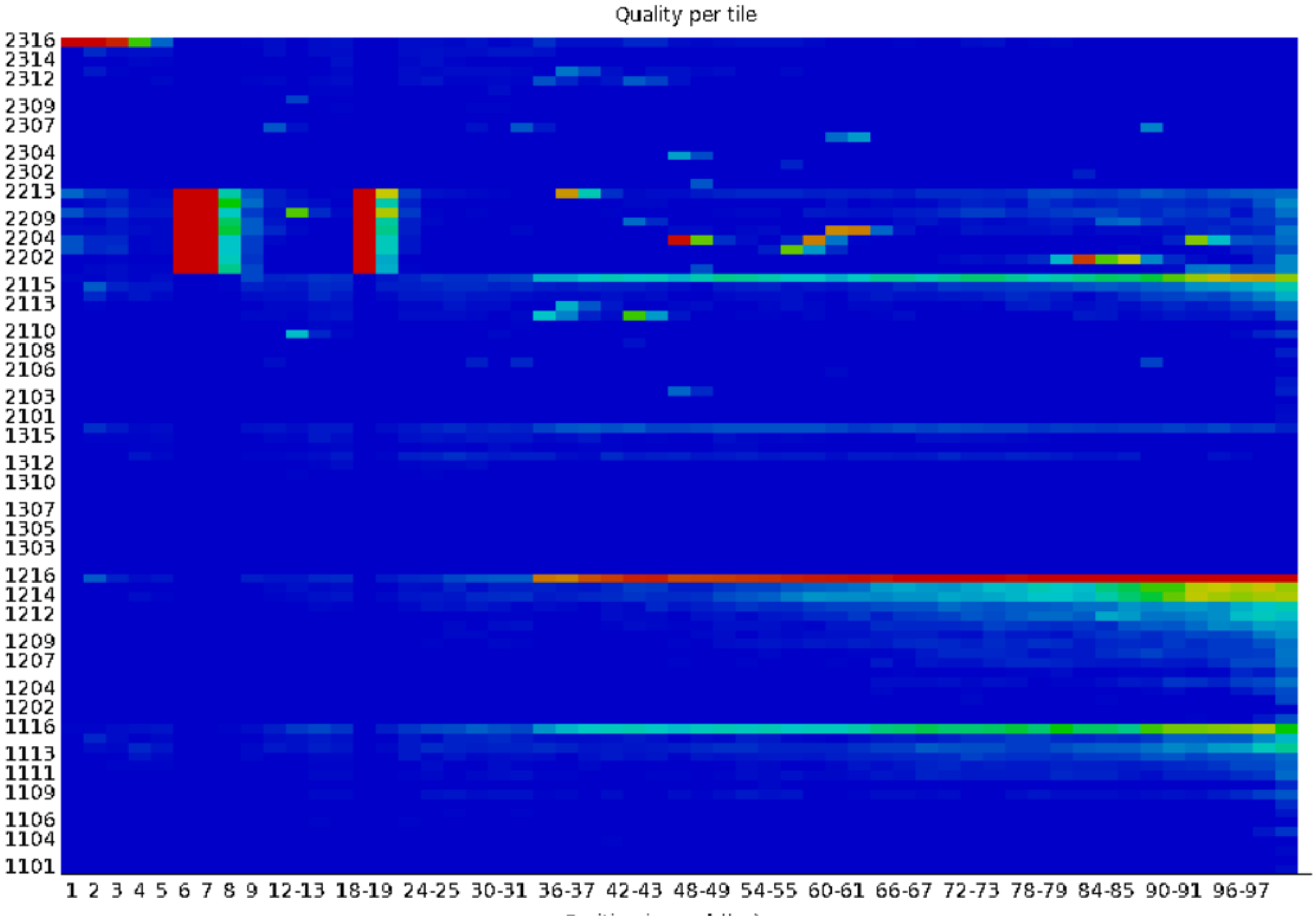
Red bars in the "Per Base Sequence Quality" section usually indicate that in our sequences there are reads where the quality scores drop significantly. These quality drops may suggest that there are regions in sequences with lower accuracy, potentially due to sequencing errors, adapter contamination, or other issues.



## ✗ Per base sequence quality

Red circles in the "Per Tile Sequence Quality" section indicate lower quality or unusual patterns on individual tiles of the flow cell (e.g. air bubbles or reagent problems).

## Per tile sequence quality



**What we should do about anything FastQC identified as unusual?**

We should explore them accurately, and, when possible, eliminate them.

## 4. Filtering the reads

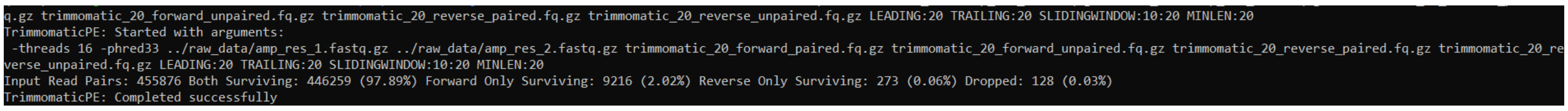
Install Trimmomatic using conda.

```
conda install -c bioconda trimmomatic
cd ..
mkdir trimmomatic
cd trimmomatic
```

Trimmomatic was run in paired-end mode, with the following parameters:
- Cut bases off the start of a read if quality is below 20
- Cut bases off the end of a read if quality is below 20
- Trim reads using a sliding window approach, with window size 10 and average quality within the window 20.
- Drop the read if it is below length 20.

```
trimmomatic PE -threads 16 -phred33 ../raw_data/amp_res_1.fastq.gz ../raw_data/amp_res_2.fastq.gz trimmomatic_20_forward_paired.fq.gz trimmomatic_20_forward_unpaired.fq.gz trimmomatic_20_reverse_paired.fq.gz trimmomatic_20_reverse_unpaired.fq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20
```

Trimmomatic reports back stats, on how many paired reads were kept



```
q.gz trimmomatic_20_forward_unpaired.fq.gz trimmomatic_20_reverse_paired.fq.gz trimmomatic_20_reverse_unpaired.fq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20
TrimmomaticPE: Started with arguments:
 -threads 16 -phred33 ../raw_data/amp_res_1.fastq.gz ../raw_data/amp_res_2.fastq.gz trimmomatic_20_forward_paired.fq.gz trimmomatic_20_forward_unpaired.fq.gz trimmomatic_20_reverse_paired.fq.gz trimmomatic_20_re
verse_unpaired.fq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20
Input Read Pairs: 455876 Both Surviving: 446259 (97.89%) Forward Only Surviving: 9216 (2.02%) Reverse Only Surviving: 273 (0.06%) Dropped: 128 (0.03%)
TrimmomaticPE: Completed successfully
```

Manual checking

```
seqkit stats trimmomatic_20_forward_paired.fq.gz
```

| File | Format | Type | **Num_seqs** | Sum_len | Min_len | Avg_len | Max_len |
|---|---|---|---|---|---|---|---|
| trimmomatic_20_forward_paired.fq.gz | FASTQ | DNA | **446,259** | 42,003,868 | 20 | 94.1 | 101 |

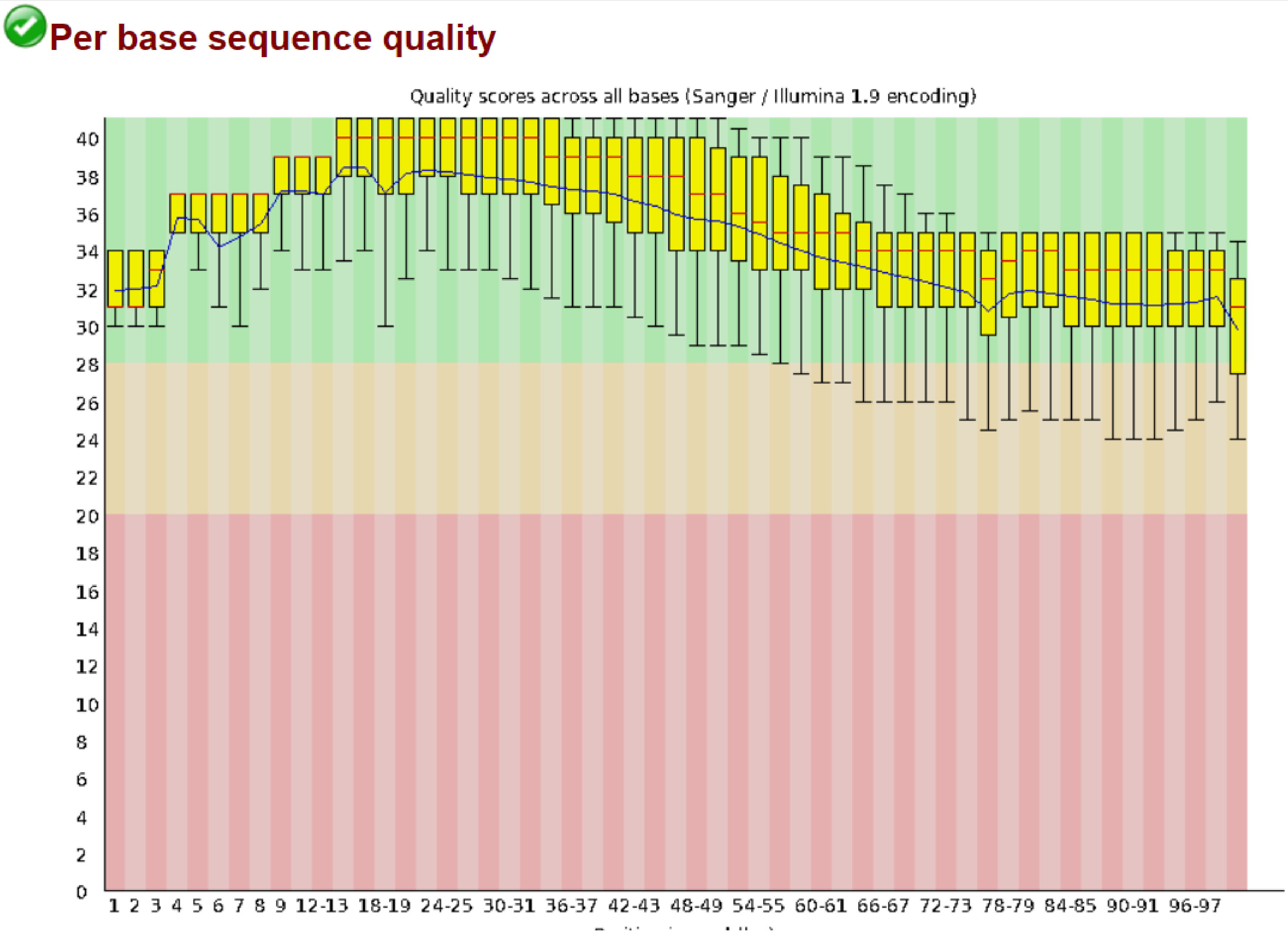
```
seqkit stats trimmomatic_20_reverse_paired.fq.gz
```

| File | Format | Type | **Num_seqs** | Sum_len | Min_len | Avg_len | Max_len |
|---|---|---|---|---|---|---|---|
| trimmomatic_20_reverse_paired.fq.gz | FASTQ | DNA | **446,259** | 41,649,154 | 20 | 93.3 | 101 |

To see how trimming affected the overall quality of the data, we repeated the fastqc analysis from section 3, but this time on the _1P.fq and _2P.fq files.

```
fastqc trimmomatic_20_forward_paired.fq.gz trimmomatic_20_reverse_paired.fq.gz -o ../fastqc/
```

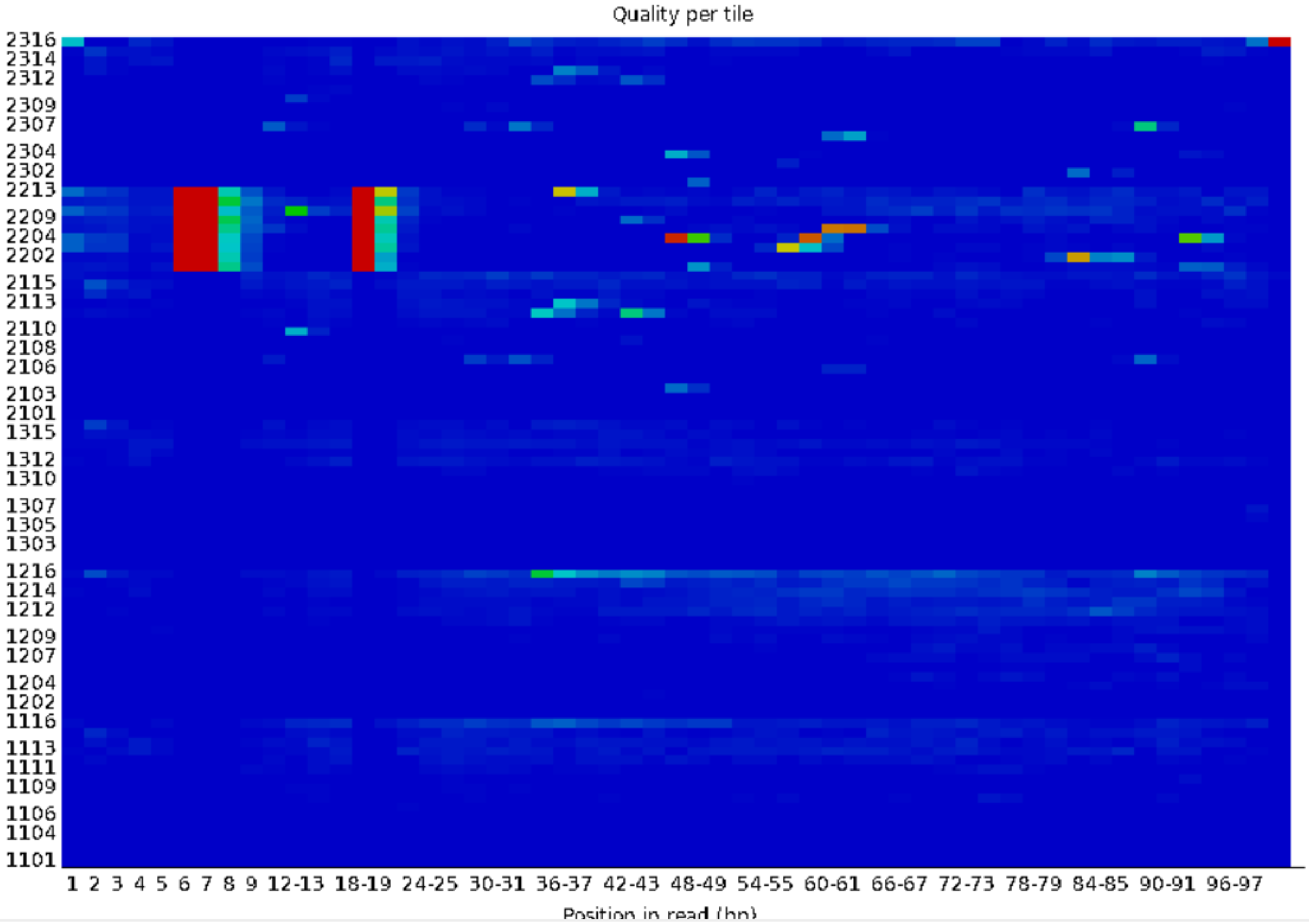
As a result, we see that according to **Per base sequence quality** everything was cut/trimmed correctly.



Per base sequence quality

**Per tile sequence quality** was also improved, except for one part, which, apparently, was caused by flowcell problems.

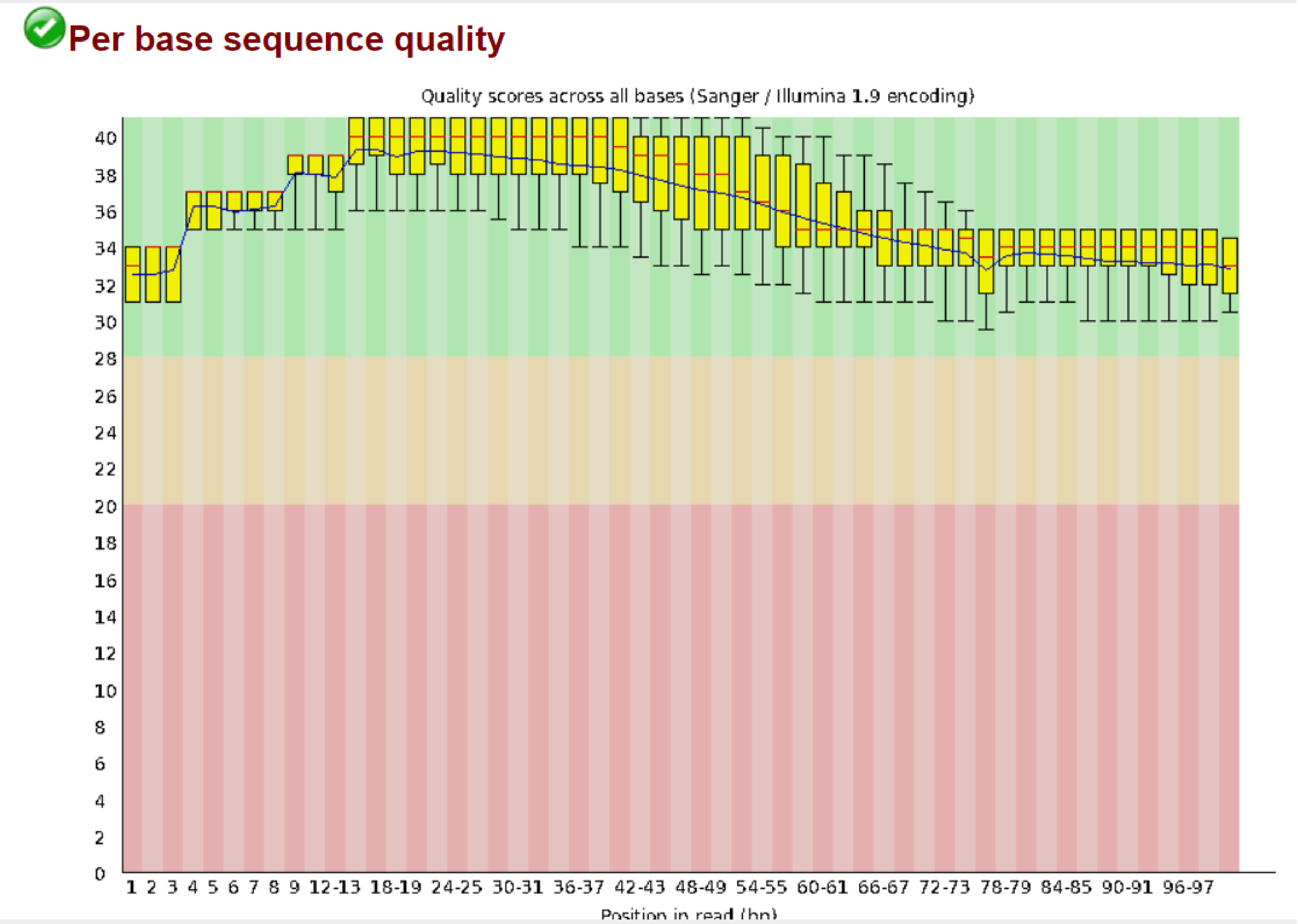
## Per tile sequence quality



Next, we increased the quality score at all steps to 30

```
trimmomatic PE -threads 16 -phred33 ../raw_data/amp_res_1.fastq.gz ../raw_data/amp_res_2.fastq.gz trimmomatic_30_forward_paired.fq.gz trimmomatic_30_forward_unpaired.fq.gz trimmomatic_30_reverse_paired.fq.gz trimmomatic_30_reverse_unpaired.fq.gz LEADING:30 TRAILING:30 SLIDINGWINDOW:10:30 MINLEN:30
fastqc trimmomatic_30_forward_paired.fq.gz trimmomatic_30_reverse_paired.fq.gz -o ../fastqc/
```

As a result we see, that all reads are in green zone according to their quality scores, which is great.



Per base sequence quality

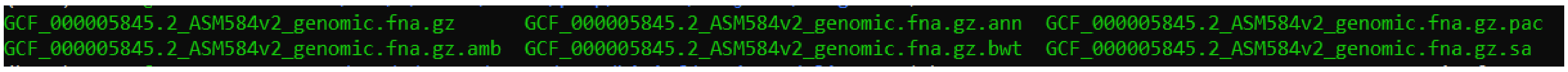
## 5 . Aligning sequences to reference

### 5.1 Index the reference file

Run bwa index on the reference sequence with the default options

```
cd ..
mkdir alignment
cd alignment
#conda install -c bioconda bwa
cp ../raw_data/GCF_000005845.2_ASM584v2_genomic.fna.gz .
gzip -d -c GCF_000005845.2_ASM584v2_genomic.fna.gz > GCF_000005845.2_ASM584v2_genomic.fna
bwa index GCF_000005845.2_ASM584v2_genomic.fna
ls
```

Creates index files

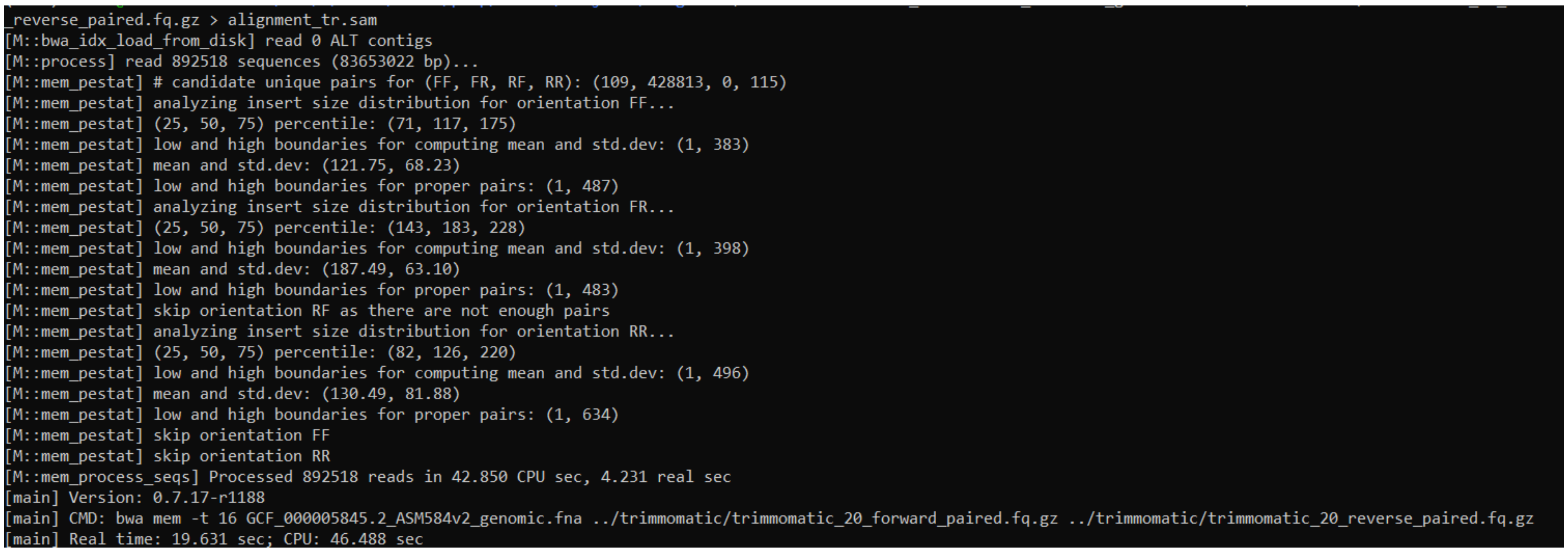


```
GCF_000005845.2_ASM584v2_genomic.fna.gz      GCF_000005845.2_ASM584v2_genomic.fna.gz.ann  GCF_000005845.2_ASM584v2_genomic.fna.gz.pac
GCF_000005845.2_ASM584v2_genomic.fna.gz.amb  GCF_000005845.2_ASM584v2_genomic.fna.gz.bwt  GCF_000005845.2_ASM584v2_genomic.fna.gz.sa
```

### 5.2 Aligning reads

```
bwa mem -t 16 GCF_000005845.2_ASM584v2_genomic.fna ../trimmomatic/trimmomatic_20_forward_paired.fq.gz ../trimmo
matic/trimmomatic_20_reverse_paired.fq.gz > alignment.sam
```

If nothing happens, press Enter 😁



```
_reverse_paired.fq.gz > alignment_tr.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 892518 sequences (83653022 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (109, 428813, 0, 115)
[M::mem_pestat] analyzing insert size distribution for orientation FF...
[M::mem_pestat] (25, 50, 75) percentile: (71, 117, 175)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 383)
[M::mem_pestat] mean and std.dev: (121.75, 68.23)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 487)
[M::mem_pestat] analyzing insert size distribution for orientation FR...
[M::mem_pestat] (25, 50, 75) percentile: (143, 183, 228)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 398)
[M::mem_pestat] mean and std.dev: (187.49, 63.10)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 483)
[M::mem_pestat] skip orientation RF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation RR...
[M::mem_pestat] (25, 50, 75) percentile: (82, 126, 220)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 496)
[M::mem_pestat] mean and std.dev: (130.49, 81.88)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 634)
[M::mem_pestat] skip orientation FF
[M::mem_pestat] skip orientation RR
[M::mem_process_seqs] Processed 892518 reads in 42.850 CPU sec, 4.231 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem -t 16 GCF_000005845.2_ASM584v2_genomic.fna ../trimmomatic/trimmomatic_20_forward_paired.fq.gz ../trimmomatic/trimmomatic_20_reverse_paired.fq.gz
[main] Real time: 19.631 sec; CPU: 46.488 sec
```

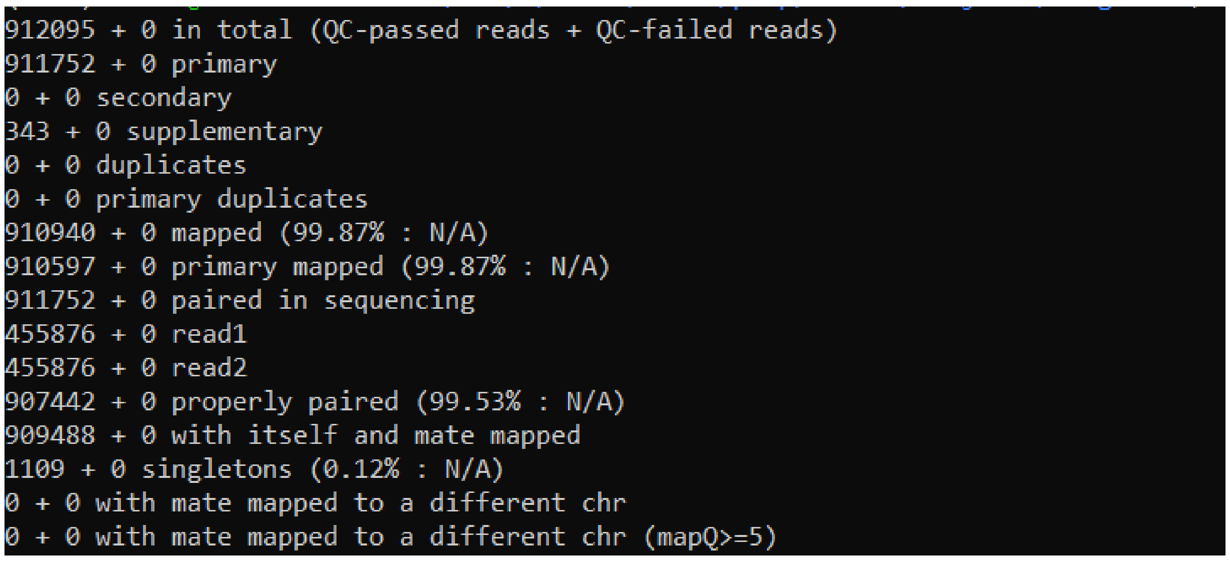
### 5.3. Compress SAM file

To compress and sort the sam file with the commands below.
A compressed sam file is called a bam file

```
samtools view -S -b alignment.sam > alignment.bam
```

To get some basic statistics:

```
samtools flagstat alignment.bam
```



```
912095 + 0 in total (QC-passed reads + QC-failed reads)
911752 + 0 primary
0 + 0 secondary
343 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
910940 + 0 mapped (99.87% : N/A)
910597 + 0 primary mapped (99.87% : N/A)
911752 + 0 paired in sequencing
455876 + 0 read1
455876 + 0 read2
907442 + 0 properly paired (99.53% : N/A)
909488 + 0 with itself and mate mapped
1109 + 0 singletons (0.12% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

What percentage of reads are mapped?

99.87%

### 5.4 Sort and index BAM file

a) Sort bam file by sequence coordinate on reference

```
samtools sort alignment.bam -o alignment_sorted.bam
```

b) Index bam file for faster search:

```
samtools index alignment_sorted.bam
```

## 6. Variant calling

The solution is to make an intermediate file type called an mpileup, because it goes through each position and "piles up" the reads, tabulating the number of bases that match or don't match the reference. Mpileup requires a sorted, indexed bam file. For this, run the basic command below.

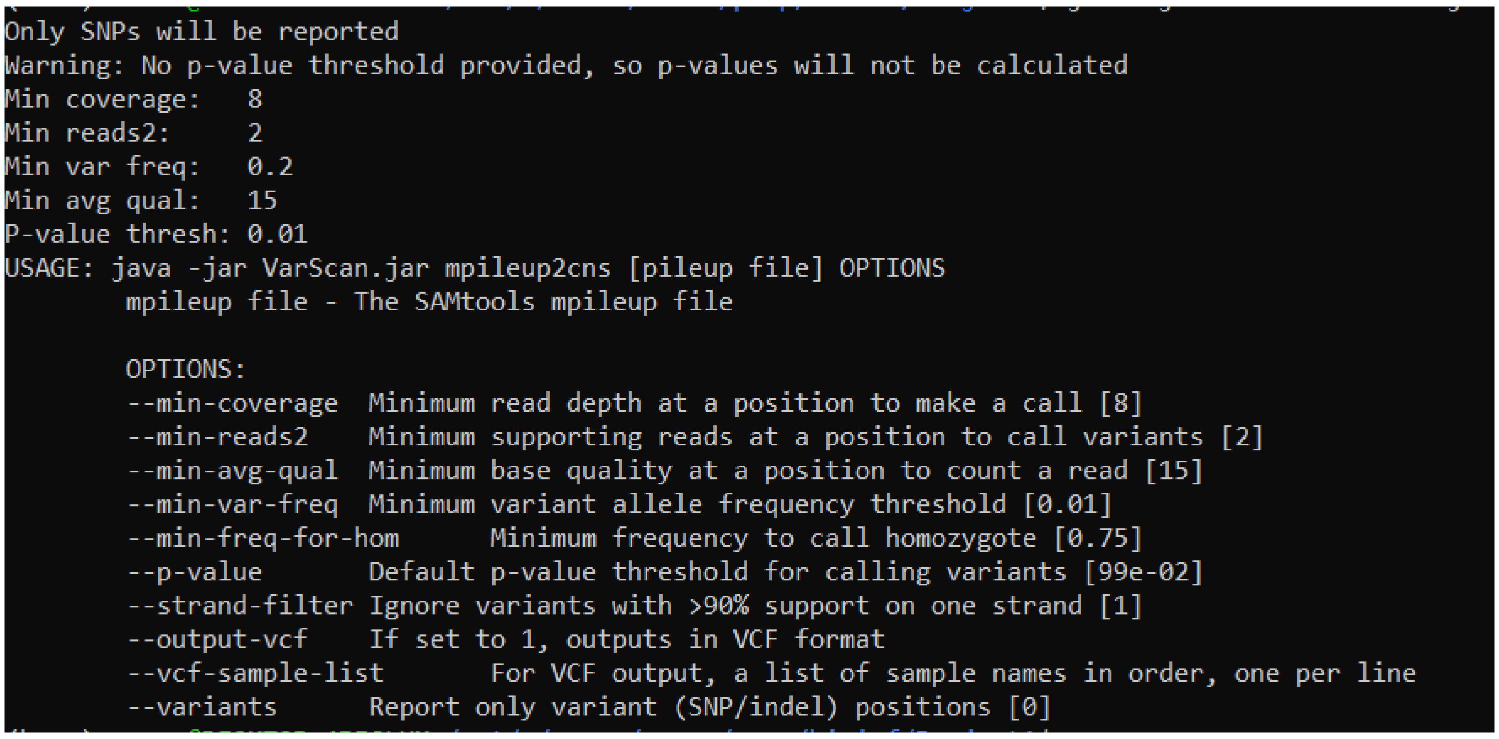
```
cd ..
mkdir variant_calling
cd variant_calling
samtools mpileup -f ../alignment/GCF_000005845.2_ASM584v2_genomic.fna ../alignment/alignment_sorted.bam > my.mp
ileup
```

To call actual variants, we will be using a program called VarScan (variant scanner)

Download version 2.4.0 to the variant_calling directory from here https://github.com/dkoboldt/varscan/blob/master/VarScan.v2.4.0.jar

Eo call help for a command mpileup2snp

```
java -jar VarScan.v2.4.0.jar mpileup2snp -h
```



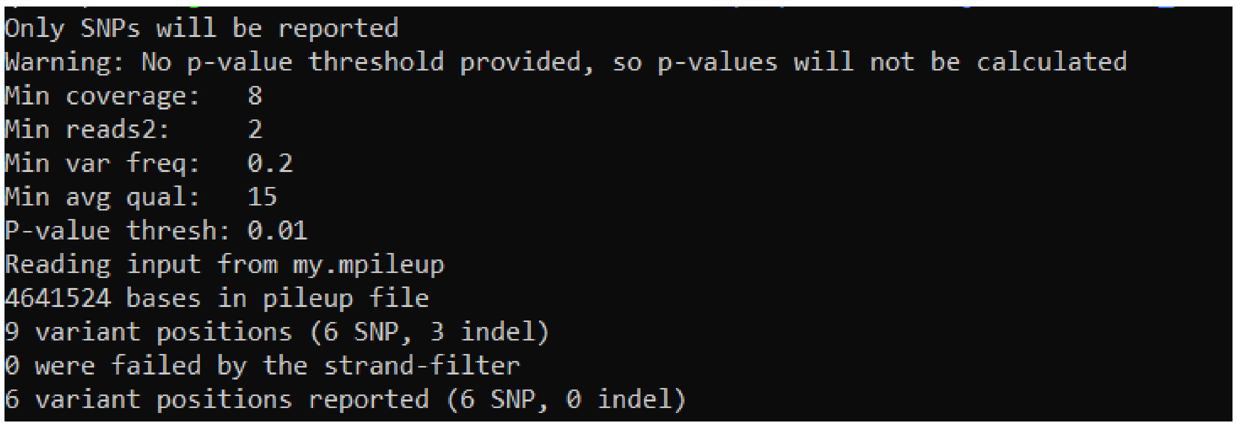
```
Only SNPs will be reported
Warning: No p-value threshold provided, so p-values will not be calculated
Min coverage:   8
Min reads2:     2
Min var freq:   0.2
Min avg qual:   15
P-value thresh: 0.01
USAGE: java -jar VarScan.jar mpileup2cns [pileup file] OPTIONS
        mpileup file - The SAMtools mpileup file

        OPTIONS:
        --min-coverage  Minimum read depth at a position to make a call [8]
        --min-reads2    Minimum supporting reads at a position to call variants [2]
        --min-avg-qual  Minimum base quality at a position to count a read [15]
        --min-var-freq  Minimum variant allele frequency threshold [0.01]
        --min-freq-for-hom      Minimum frequency to call homozygote [0.75]
        --p-value       Default p-value threshold for calling variants [99e-02]
        --strand-filter Ignore variants with >90% support on one strand [1]
        --output-vcf    If set to 1, outputs in VCF format
        --vcf-sample-list       For VCF output, a list of sample names in order, one per line
        --variants      Report only variant (SNP/indel) positions [0]
```

For SNPs checking run the following command

```
java -jar VarScan.v2.4.0.jar mpileup2snp my.mpileup --min-var-freq 0.2 --variants --output-vcf 1 > VarScan_resu
lts.vcf
```

6 variant positions reported (6 SNP, 0 indel)



```
Only SNPs will be reported
Warning: No p-value threshold provided, so p-values will not be calculated
Min coverage:   8
Min reads2:     2
Min var freq:   0.2
Min avg qual:   15
P-value thresh: 0.01
Reading input from my.mpileup
4641524 bases in pileup file
9 variant positions (6 SNP, 3 indel)
0 were failed by the strand-filter
6 variant positions reported (6 SNP, 0 indel)
```

## 7. Variant effect prediction

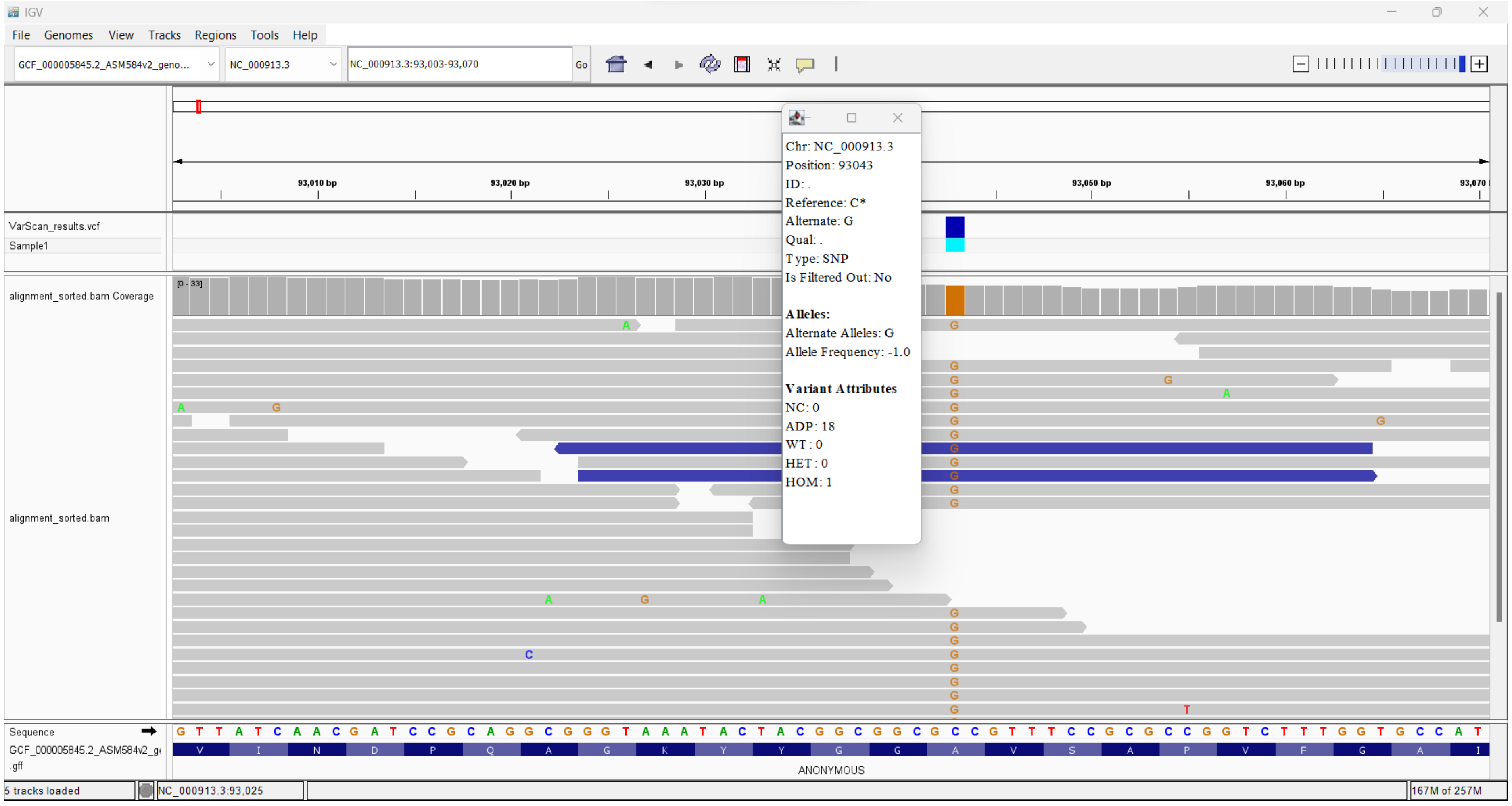
Data visualization in IGV browser:

Select “Genomes”, ”Load Genome from File” and select our reference genome. Then select “File”, “Open from file” and select your BAM file. Add also vcf file and annotation in gff format.

Results of visualization:

**SNP 1**

- Position: 93043
- Reference:  C
- Alternate: G
- Gene: ftsI
- Effect: Missense variant, changing amino acid from Alanine (Ala) to Glycine (Gly)


IGV
File Genomes View Tracks Regions Tools Help
GCF_000005845.2_ASM584v2_geno...
NC_000913.3
NC_000913.3:93,003-93,070
Go
93,010 bp
93,020 bp
93,030 bp
93,050 bp
93,060 bp
93,070
VarScan_results.vcf
Sample1
alignment_sorted.bam Coverage
[0 - 33]
alignment_sorted.bam
Sequence
GCF_000005845.2_ASM584v2_ge .gff
G T T A T C A A C G A T C C G C A G G C G G G T A A A T A C T A C G G C G G C G C C G T T T C C G C G C C G G T C T T T G G T G C C A T
V I N D P Q A G K Y Y G G A V S A P V F G A I
ANONYMOUS
5 tracks loaded
NC_000913.3:93,025
167M of 257M

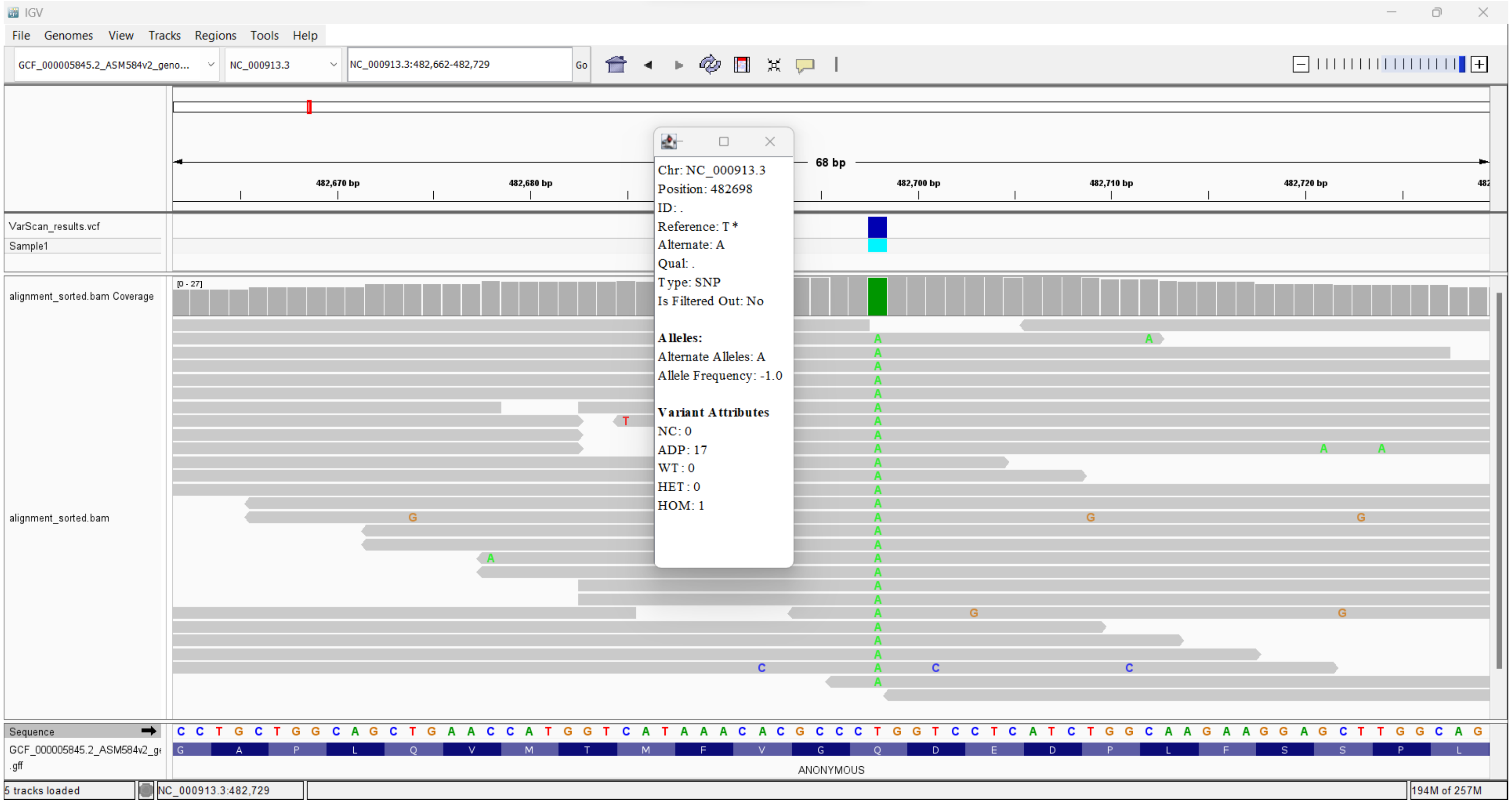
Chr: NC_000913.3
Position: 93043
ID: .
Reference: C*
Alternate: G
Qual: .
Type: SNP
Is Filtered Out: No

**Alleles:**
Alternate Alleles: G
Allele Frequency: -1.0

**Variant Attributes**
NC: 0
ADP: 18
WT: 0
HET: 0
HOM: 1

## SNP 2

- Position: 482698
- Reference:  A
- Alternate: T
- Gene: acrB
- Effect: Missense variant, changing amino acid from Glutamine (Gln/Q) to Leucine (Leu/L)


IGV
File Genomes View Tracks Regions Tools Help
GCF_000005845.2_ASM584v2_geno...
NC_000913.3
NC_000913.3:482,662-482,729
Go
68 bp
482,670 bp
482,680 bp
482,700 bp
482,710 bp
482,720 bp
VarScan_results.vcf
Sample1
alignment_sorted.bam Coverage
[0 - 27]
alignment_sorted.bam
Sequence
GCF_000005845.2_ASM584v2_ge .gff
C C T G C T G G C A G C T G A A C C A T G G T C A T A A A C A C G C C C T G G T C C T C A T C T G G C A A G A A G G A G C T T G G C A G
G A P L Q V M T M F V G Q D E D P L F S S P L
ANONYMOUS
5 tracks loaded
NC_000913.3:482,729
194M of 257M

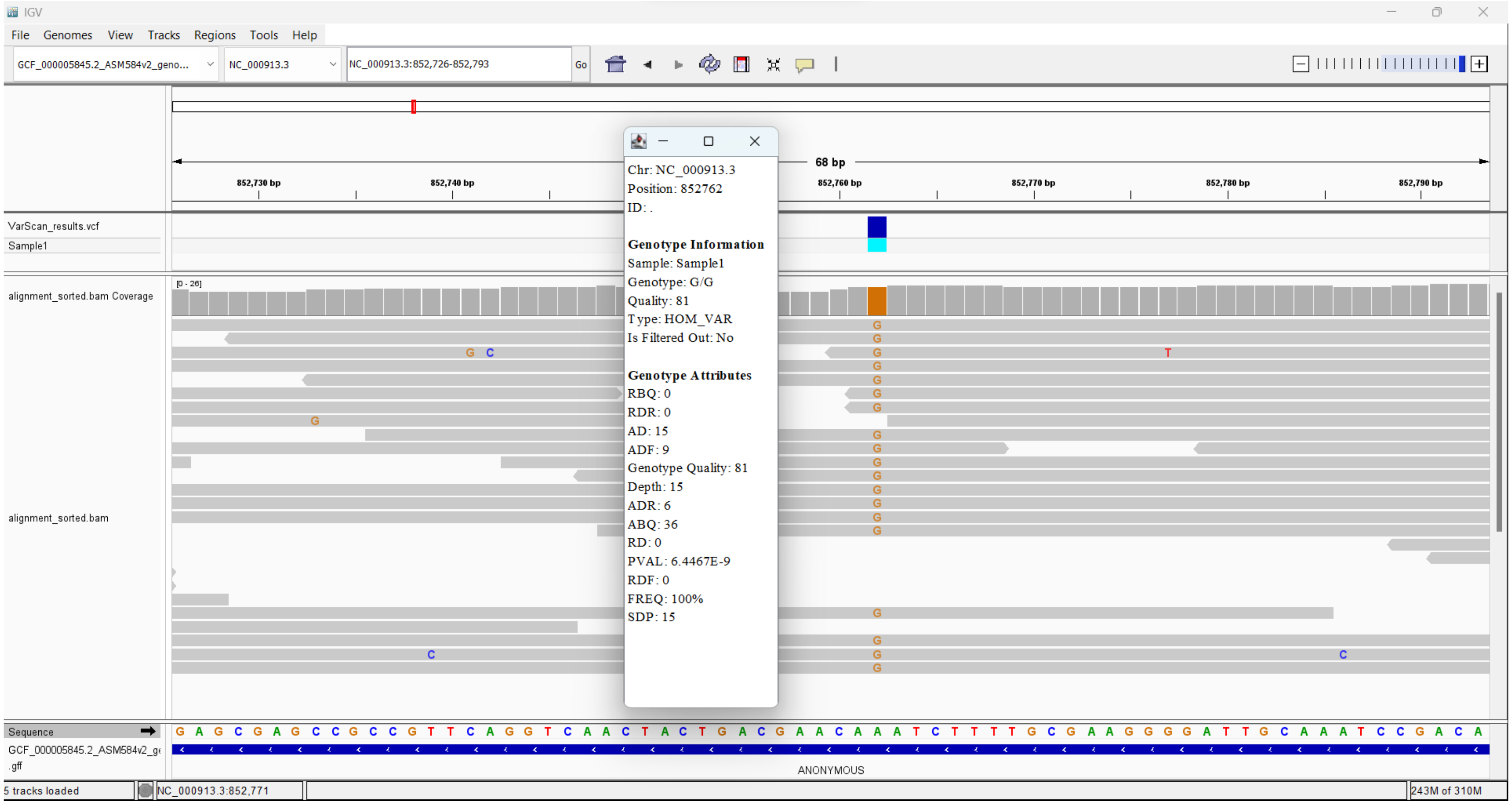
Chr: NC_000913.3
Position: 482698
ID: .
Reference: T*
Alternate: A
Qual: .
Type: SNP
Is Filtered Out: No

**Alleles:**
Alternate Alleles: A
Allele Frequency: -1.0

**Variant Attributes**
NC: 0
ADP: 17
WT: 0
HET: 0
HOM: 1

## SNP 3

- Position: 852762
- Reference:  A
- Alternate: G
- Gene: rybA
- Effect: Upstream gene variant with no amino acid change.



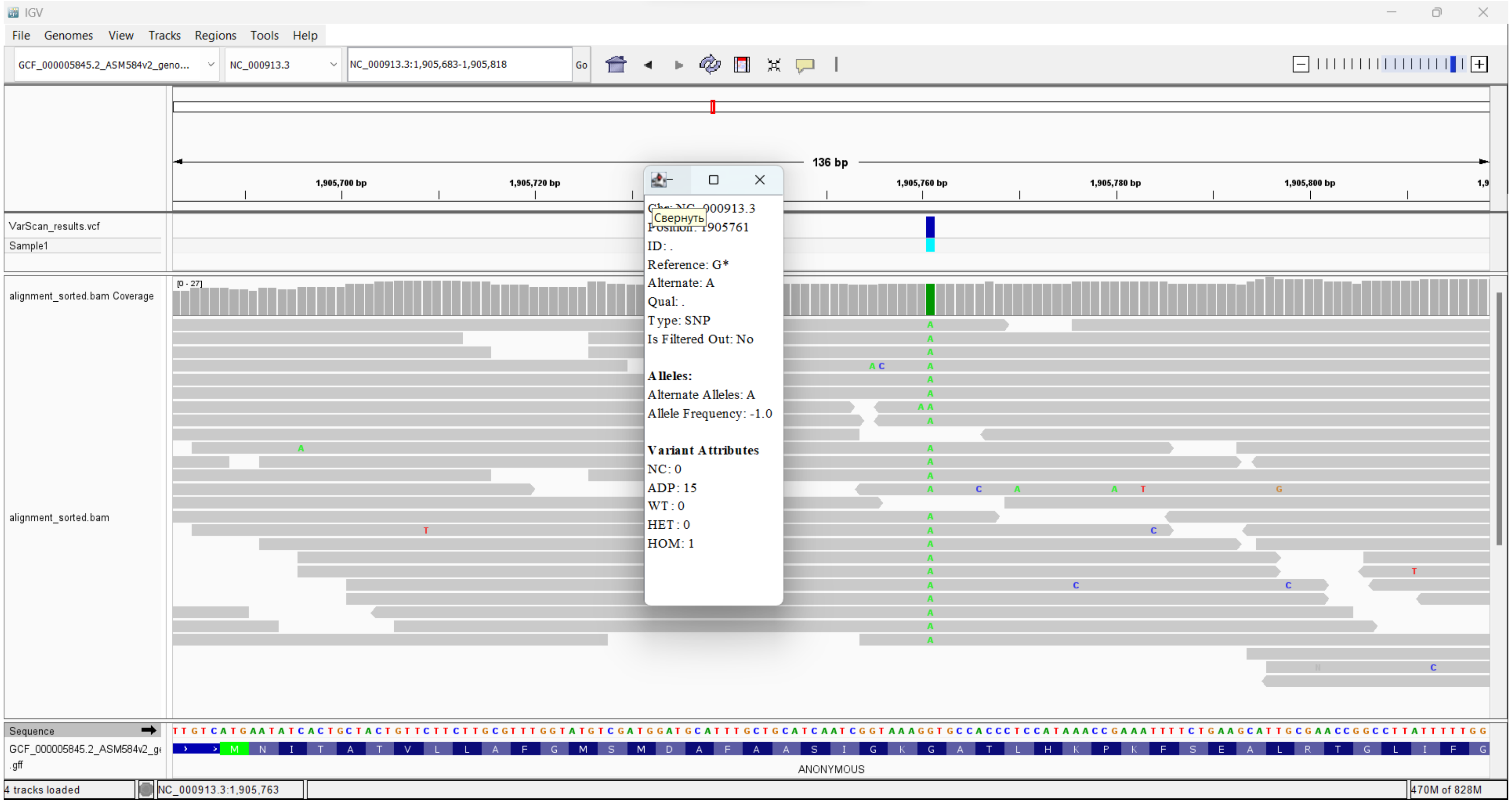
Chr: NC_000913.3
Position: 852762
ID: .

**Genotype Information**
Sample: Sample1
Genotype: G/G
Quality: 81
Type: HOM_VAR
Is Filtered Out: No

**Genotype Attributes**
RBQ: 0
RDR: 0
AD: 15
ADF: 9
Genotype Quality: 81
Depth: 15
ADR: 6
ABQ: 36
RD: 0
PVAL: 6.4467E-9
RDF: 0
FREQ: 100%
SDP: 15

**SNP 4**

- Position: 1905761
- Reference:  G
- Alternate: A
- Gene: mntP
- Effect: Missense variant, changing amino acid from Glycine (Gly) to Aspartic Acid (Asp)



Chr: NC_000913.3
Position: 1905761
ID: .
Reference: G*
Alternate: A
Qual: .
Type: SNP
Is Filtered Out: No

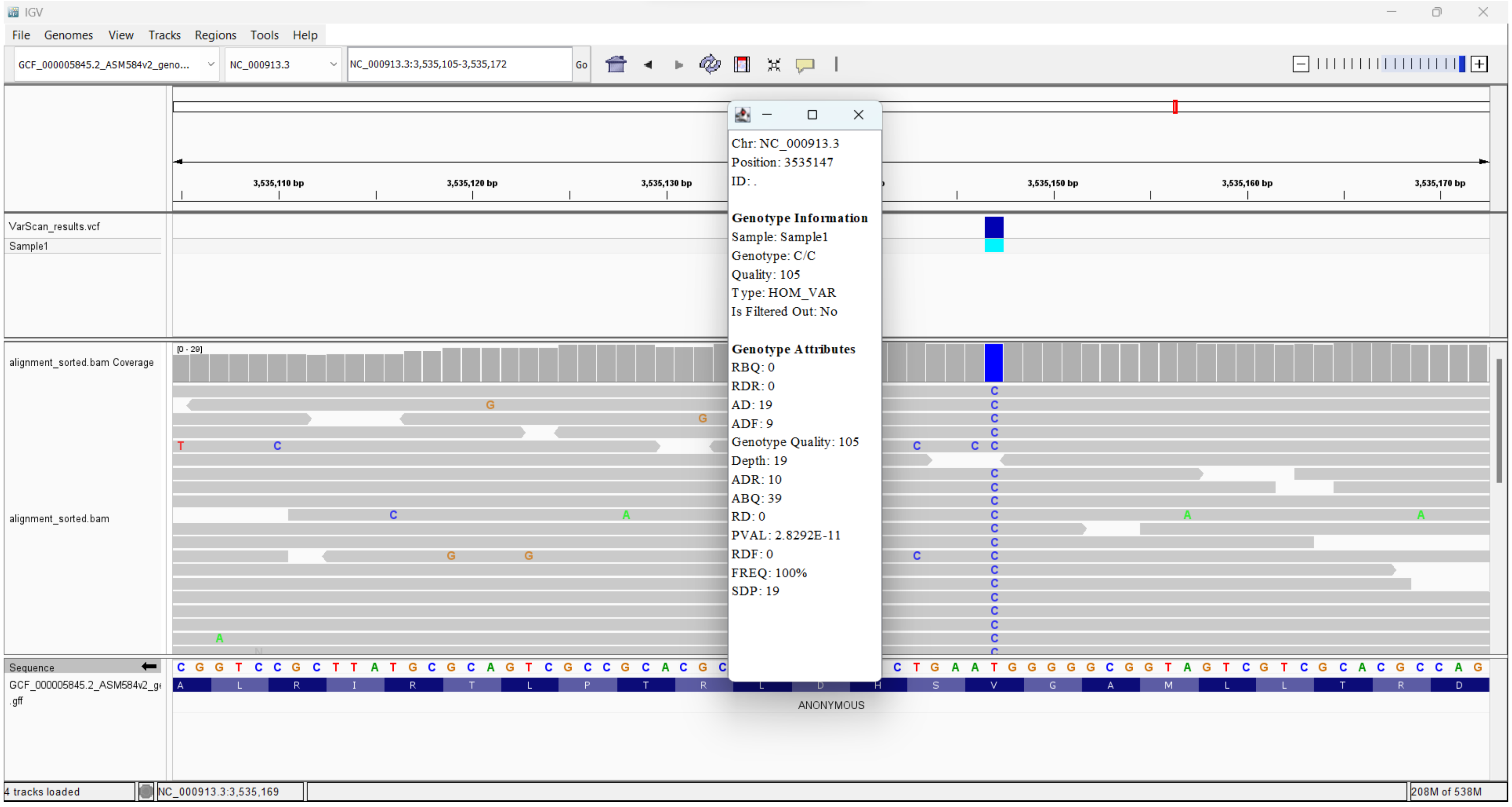
**Alleles:**
Alternate Alleles: A
Allele Frequency: -1.0

**Variant Attributes**
NC: 0
ADP: 15
WT: 0
HET: 0
HOM: 1

**SNP 5**

- Position: 3535147
- Reference:  T
- Alternate: G
- Gene: envZ

- Effect: Missense variant, changing amino acid from Valine (Val) to Glycine (Gly)



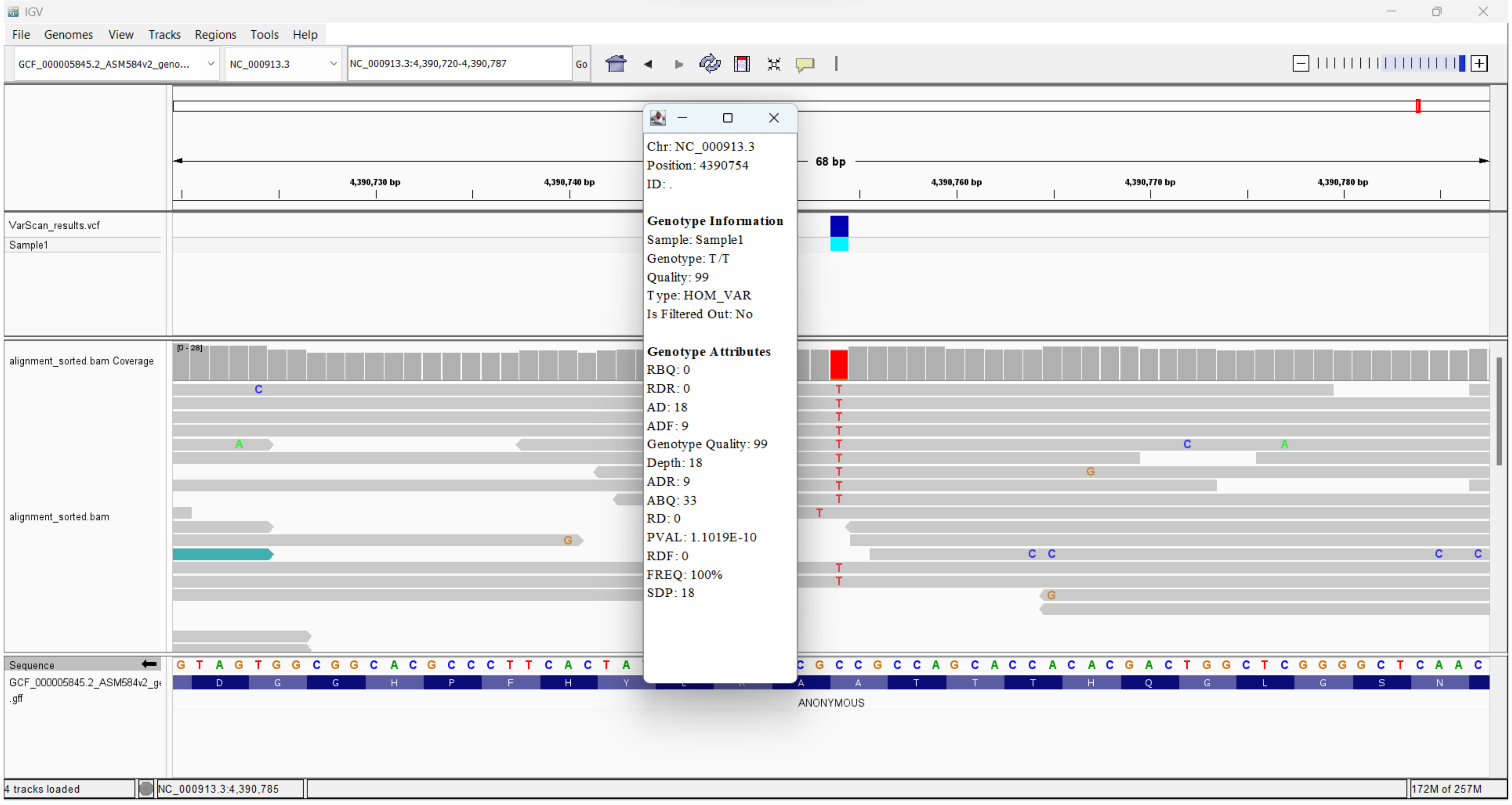
Chr: NC_000913.3
Position: 3535147
ID: .

**Genotype Information**
Sample: Sample1
Genotype: C/C
Quality: 105
Type: HOM_VAR
Is Filtered Out: No

**Genotype Attributes**
RBQ: 0
RDR: 0
AD: 19
ADF: 9
Genotype Quality: 105
Depth: 19
ADR: 10
ABQ: 39
RD: 0
PVAL: 2.8292E-11
RDF: 0
FREQ: 100%
SDP: 19

**SNP 6**

- Position: 4390754
- Reference:  C
- Alternate: A
- Gene: rsgA
- Effect: Synonymous variant, with no change in the amino acid sequence Ala → Ala



Chr: NC_000913.3
Position: 4390754
ID: .

**Genotype Information**
Sample: Sample1
Genotype: T/T
Quality: 99
Type: HOM_VAR
Is Filtered Out: No

**Genotype Attributes**
RBQ: 0
RDR: 0
AD: 18
ADF: 9
Genotype Quality: 99
Depth: 18
ADR: 9
ABQ: 33
RD: 0
PVAL: 1.1019E-10
RDF: 0
FREQ: 100%
SDP: 18

## 8. Automatic SNP annotation

```
conda install -c bioconda snpeff
```

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2
```

```
_genomic.gbff.gz
```

Create empty text file snpEff.config, and add there just one string: k12.genome : ecoli_K12
Create folder for the database

Put there your .gbk file (unzip and rename to genes.gbk)
Create database
Annotate

```
echo "k12.genome : ecoli_K12" > snpEff.config
mkdir -p data/k12
gunzip GCF_000005845.2_ASM584v2_genomic.gbff.gz
cp GCF_000005845.2_ASM584v2_genomic.gbff data/k12/genes.gbk
snpEff build -genbank -v k12
snpEff ann k12 VarScan_results.vcf > VarScan_results_annotated.vcf
```

As a result, there will be a vcf file with additional field "ANN" (for "annotation"),
describing all the effects for each SNP.

VarScan_results_annotated.vcf

Results from automatic annotation correlate with acquired with VarScan