1. Data for analysis

>Download roommate's results (Illumina single-end sequencing run)
Filename:
http://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/001/SRR1705851/

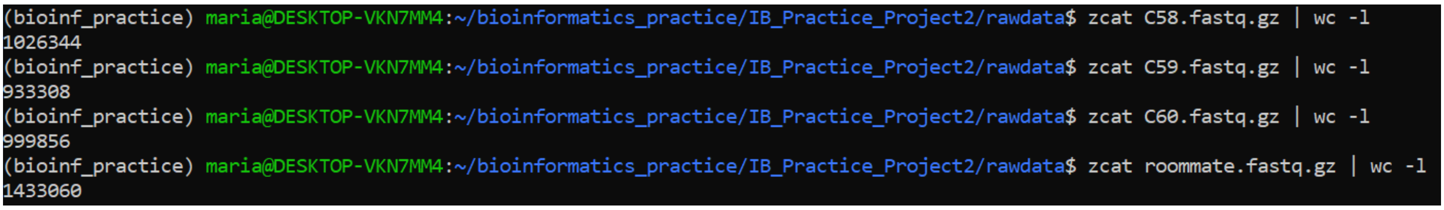
>Make fasta file with reference sequence
Filename: reference.fasta
https://www.ncbi.nlm.nih.gov/nuccore/KF848938.1?report=fasta

>Download fastq data for the three controls (from sequencing of isogenic reference samples) from SRA FTP:
using
wget <link>
SRR1705858: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/008/SRR1705858/SRR1705858.fastq.gz saved as C58.fastq.gz
SRR1705859: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/009/SRR1705859/SRR1705859.fastq.gz saved as C59.fastq.gz
SRR1705860: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR170/000/SRR1705860/SRR1705860.fastq.gz saved as C60.fastq.gz

Check the length of referense FASTA
> wc reference.fasta
1794 smb = about 550 AA

**Count the lines for initial fastq.gz files**

**zcat {filename} | wc -l**



```
(bioinf_practice) maria@DESKTOP-VKN7MM4:~/bioinformatics_practice/IB_Practice_Project2/rawdata$ zcat C58.fastq.gz | wc -l
1026344
(bioinf_practice) maria@DESKTOP-VKN7MM4:~/bioinformatics_practice/IB_Practice_Project2/rawdata$ zcat C59.fastq.gz | wc -l
933308
(bioinf_practice) maria@DESKTOP-VKN7MM4:~/bioinformatics_practice/IB_Practice_Project2/rawdata$ zcat C60.fastq.gz | wc -l
999856
(bioinf_practice) maria@DESKTOP-VKN7MM4:~/bioinformatics_practice/IB_Practice_Project2/rawdata$ zcat roommate.fastq.gz | wc -l
1433060
```

1 ноября 2023 г.    19:43

## Check the reads quality for roommate and 3 control fastq files

```
fastqc -o . {file1} {file2} {file3} {file4}
fastqc -o . C58.fastq.gz C59.fastq.gz C60.fastq.gz roommate.fastq.gz
```

3 Control have 35-38 Mbp, satisfied 'per base' and 'per sequence' quality and look very similar

failed Checkpoints:

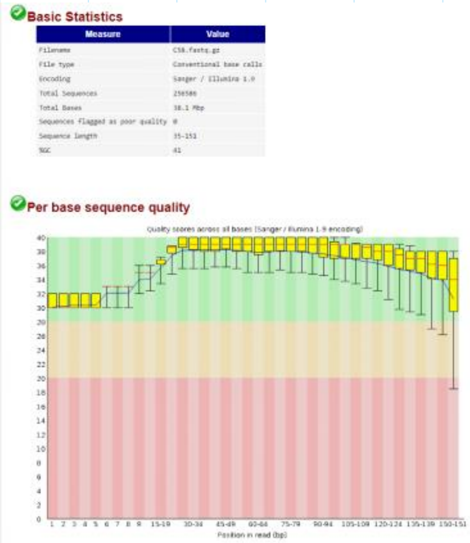
[FAIL]Per base sequence content
[FAIL]Per sequence GC content
[FAIL]Sequence Duplication Levels
[FAIL]Overrepresented sequences

C58 Warning checkpoint:

[WARNING]Sequence Length Distribution

## Control data

Control C58 File


**Basic Statistics**

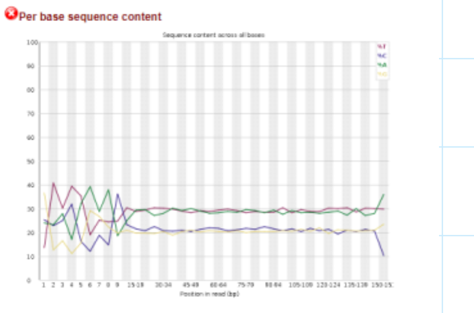
| Measure | Value |
|---|---|
| Filename | C58.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 256586 |
| Total Bases | 38.1 Mbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 35-151 |
| %GC | 41 |

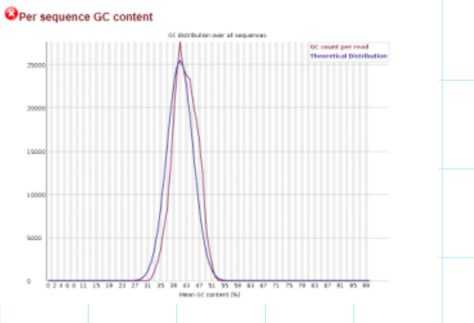
**Per base sequence quality**

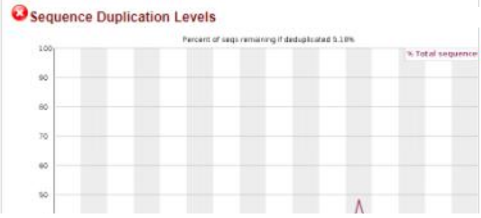
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

C58 failed Checkpoints:

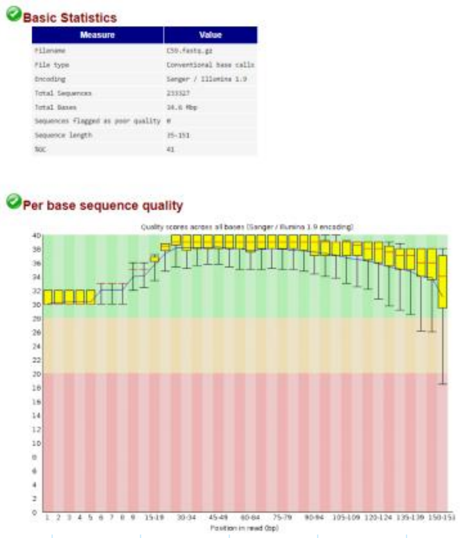
[FAIL]Per base sequence content
[FAIL]Per sequence GC content
[FAIL]Sequence Duplication Levels
[FAIL]Overrepresented sequences







Control C59 File


**Basic Statistics**

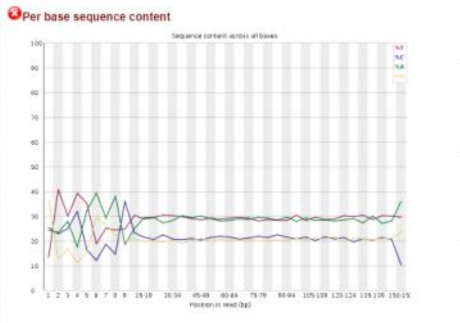
| Measure | Value |
|---|---|
| Filename | C59.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 233327 |
| Total Bases | 34.6 Mbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 35-151 |
| %GC | 41 |

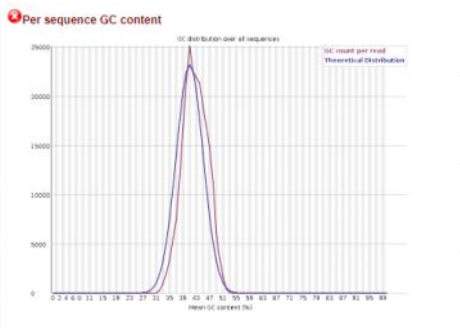
**Per base sequence quality**

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

C58 failed Checkpoints:

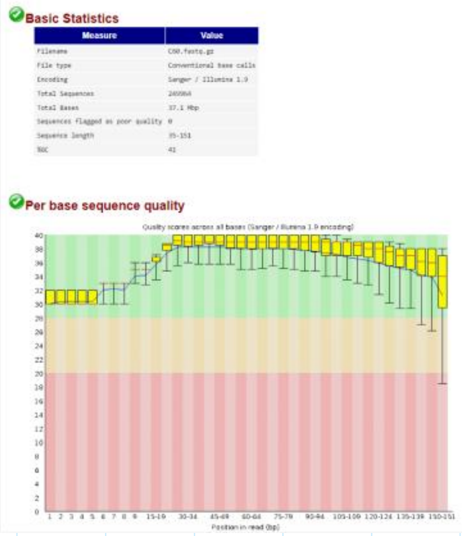
[FAIL]Per base sequence content
[FAIL]Per sequence GC content
[FAIL]Sequence Duplication Levels
[FAIL]Overrepresented sequences





Sequence Duplication Levels
Percent of seqs remaining if deduplicated 5.42%
% Total sequences

Control C60 File


**Basic Statistics**

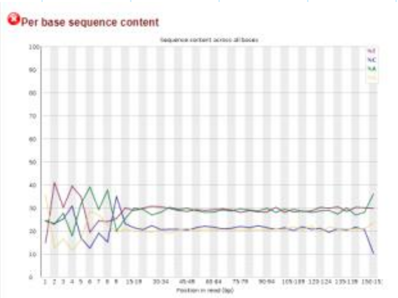
| Measure | Value |
|---|---|
| Filename | C60.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 245964 |
| Total Bases | 37.1 Mbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 35-151 |
| %GC | 41 |

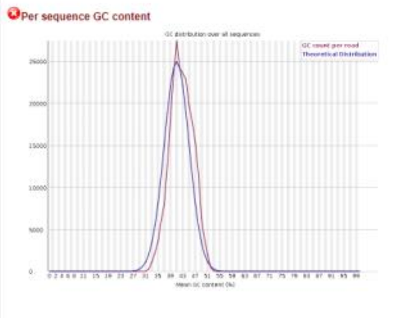
**Per base sequence quality**

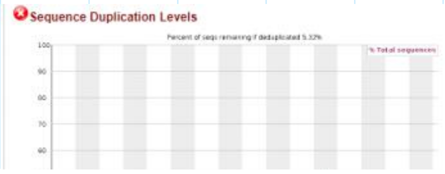
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

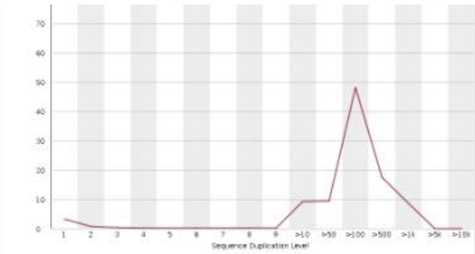
C58 failed Checkpoints:

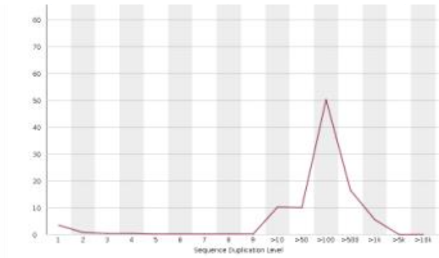
[FAIL]Per base sequence content
[FAIL]Per sequence GC content
[FAIL]Sequence Duplication Levels
[FAIL]Overrepresented sequences

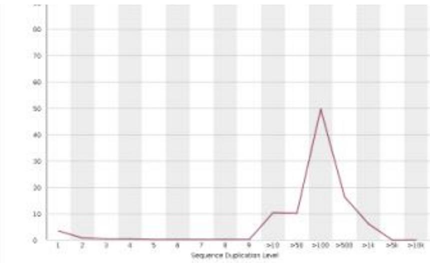


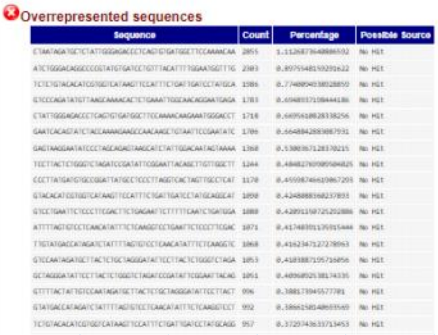






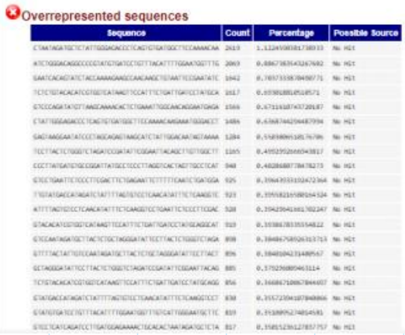






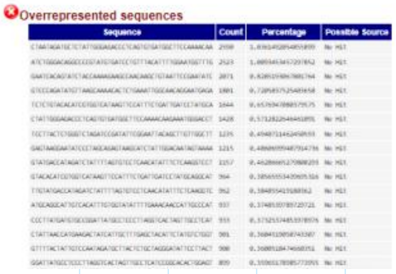
Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| [illegible] | 2855 | [illegible] | No Hit |
| [illegible] | 2303 | [illegible] | No Hit |
| [illegible] | 1986 | [illegible] | No Hit |
| [illegible] | 1783 | [illegible] | No Hit |
| [illegible] | 1718 | [illegible] | No Hit |
| [illegible] | 1706 | [illegible] | No Hit |
| [illegible] | 1368 | [illegible] | No Hit |
| [illegible] | 1244 | [illegible] | No Hit |
| [illegible] | 1170 | [illegible] | No Hit |
| [illegible] | 1090 | [illegible] | No Hit |
| [illegible] | 1088 | [illegible] | No Hit |
| [illegible] | 1071 | [illegible] | No Hit |
| [illegible] | 1068 | [illegible] | No Hit |
| [illegible] | 1053 | [illegible] | No Hit |
| [illegible] | 1051 | [illegible] | No Hit |
| [illegible] | 996 | [illegible] | No Hit |
| [illegible] | 992 | [illegible] | No Hit |
| [illegible] | 957 | [illegible] | No Hit |



Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| [illegible] | 2613 | [illegible] | No Hit |
| [illegible] | 2069 | [illegible] | No Hit |
| [illegible] | 1642 | [illegible] | No Hit |
| [illegible] | 1617 | [illegible] | No Hit |
| [illegible] | 1566 | [illegible] | No Hit |
| [illegible] | 1486 | [illegible] | No Hit |
| [illegible] | 1284 | [illegible] | No Hit |
| [illegible] | 1165 | [illegible] | No Hit |
| [illegible] | 948 | [illegible] | No Hit |
| [illegible] | 925 | [illegible] | No Hit |
| [illegible] | 923 | [illegible] | No Hit |
| [illegible] | 920 | [illegible] | No Hit |
| [illegible] | 919 | [illegible] | No Hit |
| [illegible] | 898 | [illegible] | No Hit |
| [illegible] | 896 | [illegible] | No Hit |
| [illegible] | 885 | [illegible] | No Hit |
| [illegible] | 856 | [illegible] | No Hit |
| [illegible] | 838 | [illegible] | No Hit |
| [illegible] | 833 | [illegible] | No Hit |
| [illegible] | 817 | [illegible] | No Hit |



Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| [illegible] | 2598 | [illegible] | No Hit |
| [illegible] | 2523 | [illegible] | No Hit |
| [illegible] | 2071 | [illegible] | No Hit |
| [illegible] | 1881 | [illegible] | No Hit |
| [illegible] | 1644 | [illegible] | No Hit |
| [illegible] | 1428 | [illegible] | No Hit |
| [illegible] | 1235 | [illegible] | No Hit |
| [illegible] | 1215 | [illegible] | No Hit |
| [illegible] | 1157 | [illegible] | No Hit |
| [illegible] | 964 | [illegible] | No Hit |
| [illegible] | 952 | [illegible] | No Hit |
| [illegible] | 937 | [illegible] | No Hit |
| [illegible] | 933 | [illegible] | No Hit |
| [illegible] | 901 | [illegible] | No Hit |
| [illegible] | 900 | [illegible] | No Hit |
| [illegible] | 899 | [illegible] | No Hit |

Beginning of the list

C58 Warning checkpoint:
[WARNING]Sequence Length Distribution

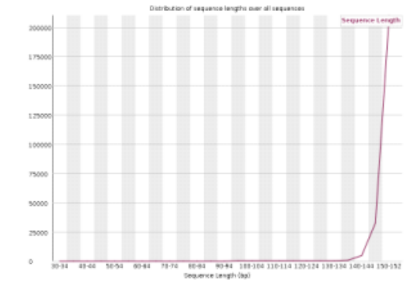
Beginning of the list

C58 Warning checkpoint:
[WARNING]Sequence Length Distribution

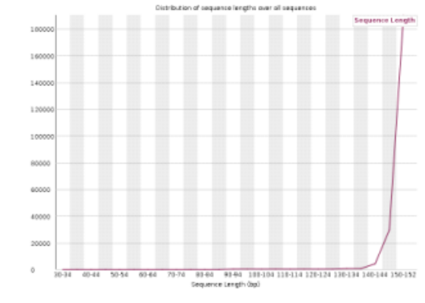
Beginning of the list

C58 Warning checkpoint:
[WARNING]Sequence Length Distribution

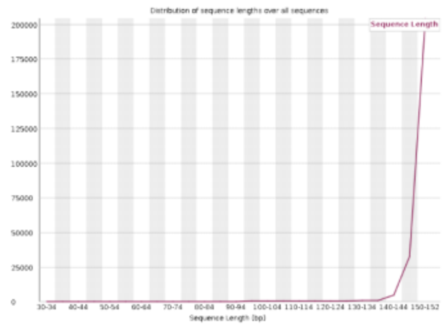
Sequence Length Distribution



Sequence Length Distribution



Sequence Length Distribution

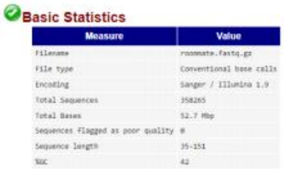


## Data from roommate

Data from roommate presents
358265 sequenses and 52.7 Mbp and different from control files quality characteristics. It performs less overrepresented sequences and more normal per sequence GC content

Failed Checkpoints:
[FAIL]Per base sequence content
[FAIL]Sequence Duplication Levels

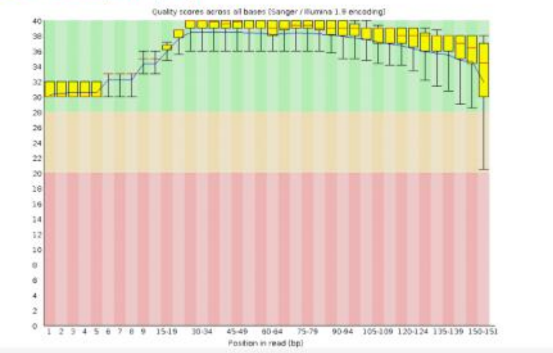
Warning checkpoints:
[WARNING]Per sequence GC content
[WARNING]Sequence Length Distribution
[WARNING]Overrepresented sequences



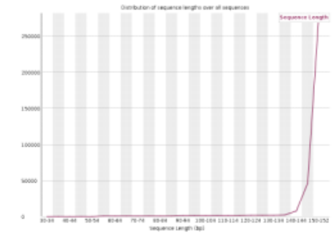
Basic Statistics

| Measure | Value |
|---|---|
| Filename | roommate.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 358265 |
| Total Bases | 52.7 Mbp |
| Sequences flagged as poor quality | 0 |
| Sequence length | 35-151 |
| %GC | 42 |

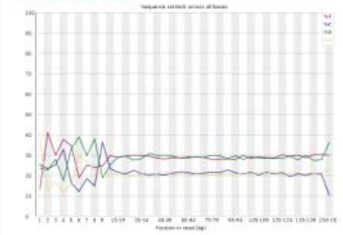
Per base sequence quality



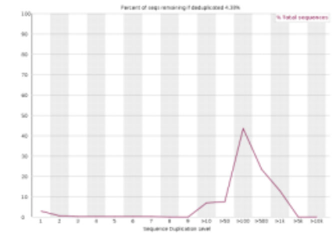
Sequence Length Distribution



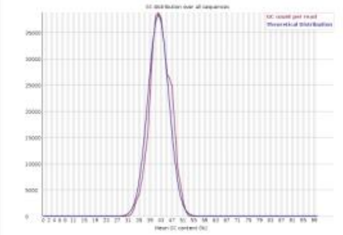
Per base sequence content

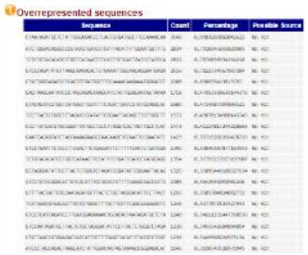


Sequence Duplication Levels



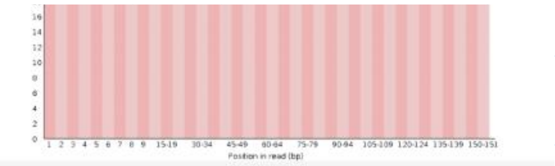
Per sequence GC content

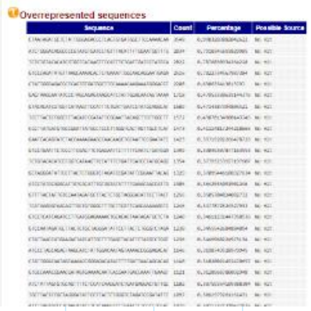




Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |





| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |
| [illegible] | [illegible] | [illegible] | No Hit |

Beginning of the list

# Points for discussion

1 ноября 2023 г.      19:27

1 ноября 2023 г.
19:22
Virus strain:  A/Hong Kong/4801/2014 (H3N2)
//What about vaccination against this strain:
(Need toFind out what strains were in this season’s vaccine. Was that one of the flu strains covered by this vaccine?)
// What about epitopes:
Munoz, Deem 2004
https://drive.google.com/file/d/1xe5-4LxIV4bO4mX6jhvrMAqtpOpkWsXm/view

Structure and receptor binding preferences of recombinant human A(H3N2) virus hemagglutinins

https://www.rcsb.org/structure/4WEA