

Отчет.

Для анализа был выбран tetraodon, а именно его 21 хромосома. Был скачан файл в формате .fasta. Далее был загружен файл zhunt3-alan.c
Далее были выполнены следующие команды:

```
!gcc zhunt3-alan.c -lm -o zhunt3
```

```
!chmod a+x zhunt3
```

```
!./zhunt3 12 8 12 chr21.fa
```

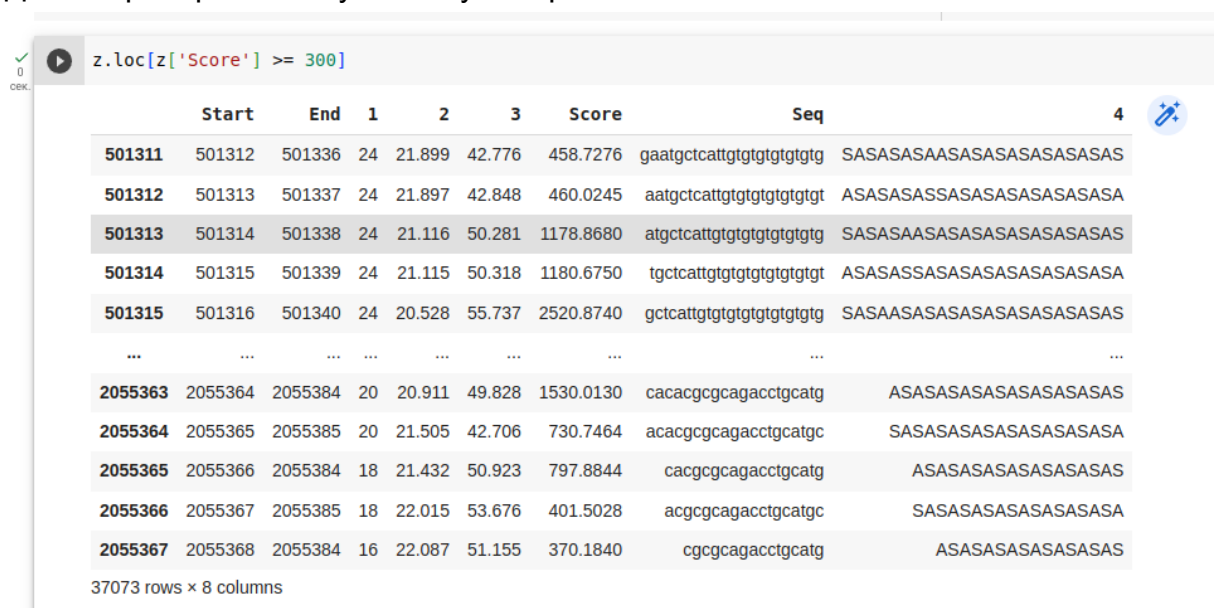
Получен файл chr21.fa.Z-SCORE

1) Для анализа будем использовать pandas:

```
import pandas as pd
```

```
z=pd.read_csv("chr21.fa.Z-SCORE", skiprows=1,  
names=["Start","End","1","2","3","Score","Seq","4"], delim_whitespace=True)
```

Далее проверяем все участки у которых z-score



	Start	End	1	2	3	Score	Seq	4
501311	501312	501336	24	21.899	42.776	458.7276	gaatgctcattgtgtgtgtgtgtg	SASASASAASASASASASASASAS
501312	501313	501337	24	21.897	42.848	460.0245	aatgctcattgtgtgtgtgtgtgt	ASASASASSASASASASASASASAS
501313	501314	501338	24	21.116	50.281	1178.8680	atgctcattgtgtgtgtgtgtgtg	SASASAAASASASASASASASASAS
501314	501315	501339	24	21.115	50.318	1180.6750	tgctcattgtgtgtgtgtgtgtgt	ASASASSASASASASASASASASAS
501315	501316	501340	24	20.528	55.737	2520.8740	gctcattgtgtgtgtgtgtgtgtg	SASAASASASASASASASASASAS
...
2055363	2055364	2055384	20	20.911	49.828	1530.0130	cacacgcgcagacctgcatg	ASASASASASASASASASASAS
2055364	2055365	2055385	20	21.505	42.706	730.7464	acacgcgcagacctgcatgc	SASASASASASASASASASASAS
2055365	2055366	2055384	18	21.432	50.923	797.8844	cacgcgcagacctgcatg	ASASASASASASASASASASAS
2055366	2055367	2055385	18	22.015	53.676	401.5028	acgcgcagacctgcatgc	SASASASASASASASASASASAS
2055367	2055368	2055384	16	22.087	51.155	370.1840	cgcgcagacctgcatg	ASASASASASASASASASASAS

37073 rows x 8 columns

Получается, что участков z-ДНК 37073

2) Теперь приступаем к следующему заданию. Устанавливаем необходимую библиотеку:

```
!pip install biopython
```

```
from Bio import SeqIO
```

Теперь запускаем следующий код, чтобы узнать количество квадруплексов:

```
inp = "chr21.fa"
for r in SeqIO.parse(inp, "fasta"):
    name, seq = r.id, str(r.seq)
    pattern="(?:G{3,}[ATGC]{1,7}){3,}G{3,}"
    ans=[m.start(),m.end(),m.group(0)] for m in re.finditer(pattern,seq)
    len(ans)
```

77

Итого, их всего 77.

- 3) Теперь 3 задание. Определим, куда попадают z-ДНК и квадруплексы. Качаем файл tetNig2.ensGene.gtf.gz распаковываем, загружаем в колаб. Также качаем bedtools.