1. Для анализа был выбран белок UxuR, содержащий 2 файла с экспериментальными ридами и 1 контроль.

Был проведен контроль качества прочтений с помощью fastqc, multiqc (html отчеты прикреплены отдельно)

```
> ./FastQC/fastqc UxuR_1.fastq
> ./FastQC/fastqc UxuR_2.fastq
> ./FastQC/fastqc UxuR_control.fastq
>  mutiqc .
```

2. Экстрагируйте риды и обрежьте их с помощью Trimmomatic

Запуск trimmomatic осуществлялся в Single End Mode, т.к. риды не парные, типа SE50

```
> trimmomatic SE -phred33 UxuR_1.fastq UxuR_1_trim.fastq LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
> Input Reads: 58606277 Surviving: 57850993 (98.71%) Dropped: 755284 (1.29%)

> trimmomatic SE -phred33 UxuR_2.fastq UxuR_2_trim.fastq LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
> Input Reads: 23901482 Surviving: 23764685 (99.43%) Dropped: 136797 (0.57%)

> trimmomatic SE -phred33 UxuR_control.fastq UxuR_control_trim.fastq
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
> Input Reads: 22987727 Surviving: 22872857 (99.50%) Dropped: 114870 (0.50%)
```

3. Наложите на геном E. coli K-12 MG1655 genome с помощью Bowtie

Скачивание генома:

```
> wget
"https://www.ncbi.nlm.nih.gov/sviewer/viewer.cgi?tool=portal&save=file&log$=s
eqview&db=nuccore&report=fasta&id=545778205&extrafeat=null&conwithfeat=on&hid
e-cdd=on" -O EColi_genome.fasta
```

Построение индексации для картирования:

```
> bowtie2-build EColi_genome.fasta EColi_index
```

Картирование ридов из UxuR_1, UxuR_2, Uxu_R_control на геном:

```
> bowtie2 -x EColi_index -U UxuR_1_trim.fastq -S UxuR_1.sam
> … 65.44% overall alignment rate

> bowtie2 -x EColi_index -U UxuR_2_trim.fastq -S UxuR_2.sam
> … 84.81% overall alignment rate

> bowtie2 -x EColi_index -U UxuR_control_trim.fastq -S UxuR_control.sam
> … 66.96% overall alignment rate
```

4. Найдите пики с помощью MACS2 и сделайте картинку/табличку с генами, к которым они относятся

Переведем .sam файлы в .bam формат с помощью samtools. BAM файлы подаются на вход MACS2 и после этого можно удалить *.sam для экономии места.

```
> samtools view -S -b -o UxuR_1.bam UxuR_1.sam
> samtools view -S -b -o UxuR_2.bam UxuR_2.sam
> samtools view -S -b -o UxuR_control.bam UxuR_control.sam
```

Поиск пиков UxuR_1 vs UxuR_control и UxuR_2 vs UxuR_control :

```
> macs2 callpeak -t UxuR_1.bam -c UxuR_control.bam -f BAM -g 4.7e+6 --extsize
147 --nomodel --keep-dup=auto --outdir macs_UxuR_1_peaks

> macs2 callpeak -t UxuR_2.bam -c UxuR_control.bam -f BAM -g 4.7e+6 --extsize
147 --nomodel --keep-dup=auto --outdir macs_UxuR_2_peaks
```

Пришлось добавить подсказки macs2, иначе он не выводил пики

Для создания таблицы с генами необходимо было посмотреть на пересечения координат полученных пиков и координат генов в аннотации. Для этого была скачана аннотация генома E.coli с NCBI и в питоне отсортирована и преобразована в файл .bed формата.

```
> bedtools closest -a /macs_UxuR_1_peaks/NA_peaks.narrowPeak -b
ann_sorted.bed > UxuR_1_annotated.bed

> bedtools closest -a /macs_UxuR_2_peaks/NA_peaks.narrowPeak -b
ann_sorted.bed > UxuR_2_annotated.bed
```

Таблицы в .csv прикреплены отдельно

5. На основании найденных пиков предположите, какие функции могут выполнять белки UxuR

Сопоставив координаты пиков и генов в аннотации видно, что пики чаще располагаются у концов генов или пересекают их. При этом, функционально, это совершенно разные гены. Вероятно, белок UxuR является регуляторным белком в клеточных матричных процессах.

6. Используя ChIPMunk, найдите мотив для выбранного белка

Получение .fasta последовательностей участков пиков для UxuR_1, UxuR_2:

```
> bedtools getfasta -fi EColi_genome.fasta -bed
/macs_UxuR_1_peaks/NA_peaks.narrowPeak -fo peaks_UxuR_1.fasta

> bedtools getfasta -fi EColi_genome.fasta -bed
/macs_UxuR_2_peaks/NA_peaks.narrowPeak -fo peaks_UxuR_2.fasta
```

Получение файлов с мотивом для UxuR_1, UxuR_2:

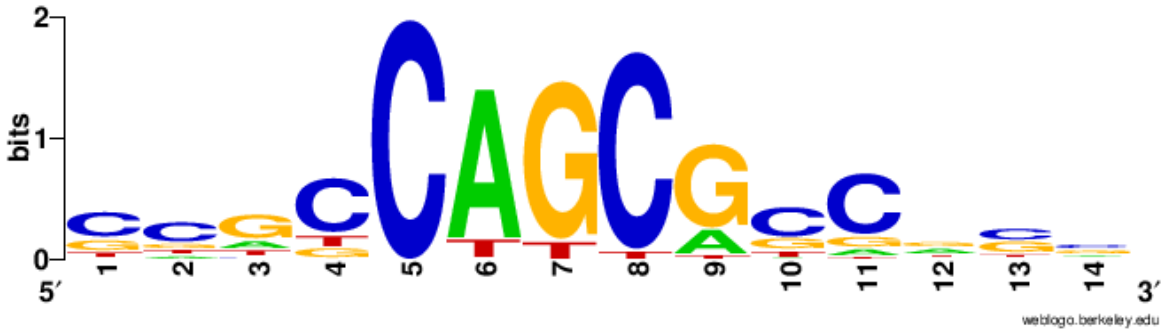
```
> java -cp chipmunk.jar ru.autosome.ChIPMunk s:peaks_UxuR_1.fasta >
motif_seq_UxuR_1.txt

> java -cp chipmunk.jar ru.autosome.ChIPMunk s:peaks_UxuR_2.fasta >
motif_seq_UxuR_2.txt
```

Далее с помощью скрипта на python были экстрагированы последовательности сайтов в формате .csv. Затем они были преобразованы в .fasta формат вручную:

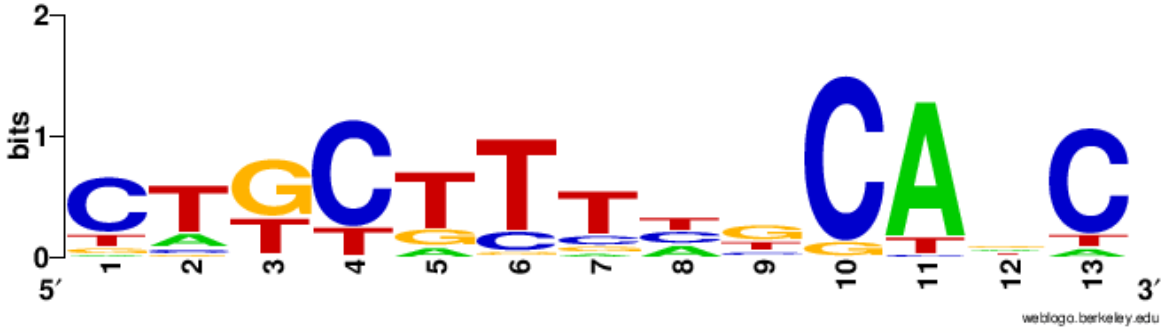
```
> ruby -ne 'puts ">" + $_.split(",").first(2).join("\n")'
motif_seq_UxuR_1.csv > motif_seq_UxuR_1.fasta

> ruby -ne 'puts ">" + $_.split(",").first(2).join("\n")'
motif_seq_UxuR_2.csv > motif_seq_UxuR_2.fasta
```

Данные файлы были поданы в веб. версию weblogo (https://weblogo.berkeley.edu/logo.cgi) для создания логотипа мотива связывающегося белком UxuR в первой и второй реплике эксперимента:

Для motif_seq_UxuR_1.fasta :



Для motif_seq_UxuR_2.fasta :



Интересно, что для UxuR_1 и UxuR_2 мотив получился разный. Возможно, это разные изоформы одного белка.