

## Отчёт по проекту

Я выбрала человека, структуру ДНК G4\_seq\_Li\_K, гистоновую метку H3K36me3, тип клеток H7, id файлов ENCFF063DDB и ENCFF864HTI, сборка hg38.

### 1. Скачивание и обработка файлов.

```
mkdir final_project  
  
cd final_project  
  
echo ENCFF063DDB > ids.txt  
  
echo ENCFF864HTI >> ids.txt  
  
cat ids.txt | xargs -tI{} wget  
https://www.encodeproject.org/files/{}/@@download/{}.bed.gz
```

Для дальнейшей обработки нам понадобятся только первые 5 столбцов файла (номер хромосомы, координаты начала и конца участка, имя, score), кроме того, файл надо распаковать.

```
cat ids.txt | xargs -tI{} sh -c "zcat {}.bed.gz | cut -f1-5 > H3K36me3_H7.{}.hg38.bed"
```

Далее хотим перевести координаты из hg38 в hg19. Для этого надо сначала скачать дополнительный файл.

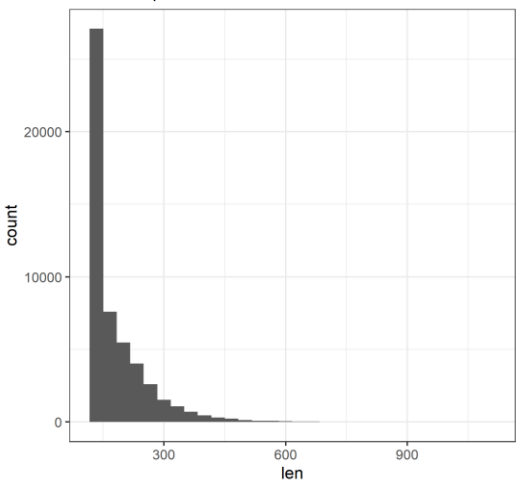
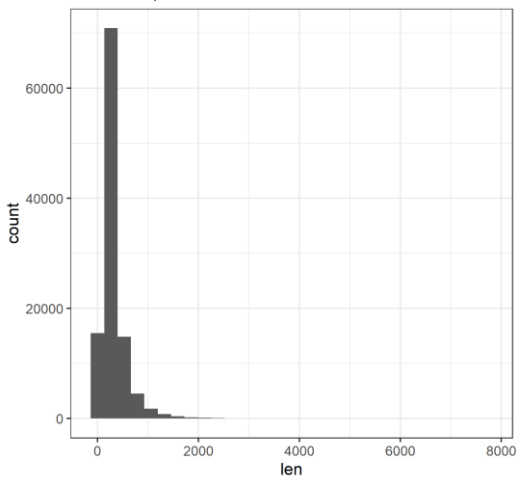
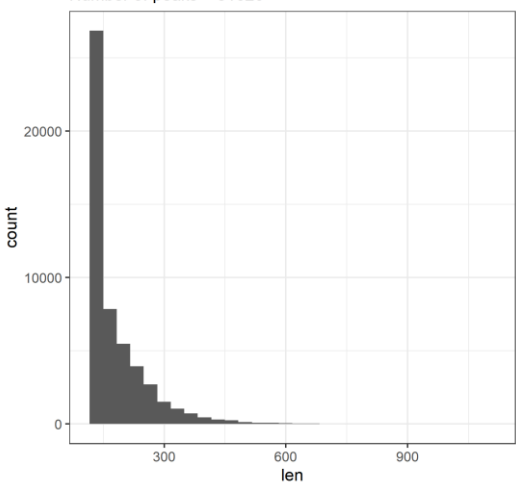
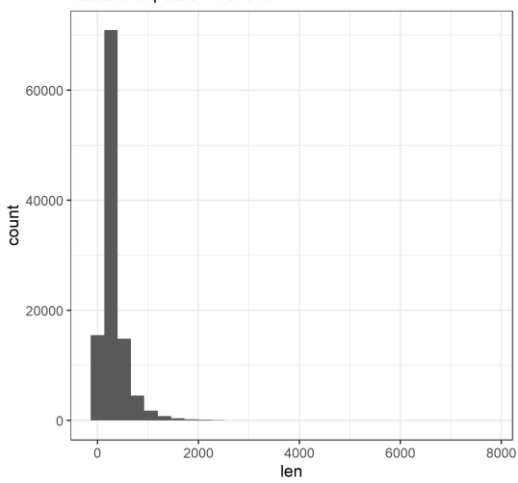
```
wget  
https://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/hg38ToHg19.over.chain.gz
```

Далее используем команду liftOver.

```
cat ids.txt | xargs -tI{} liftOver  
H3K36me3_H7.{}.hg38.bed hg38ToHg19.over.chain.gz  
H3K36me3_H7.{}.hg19.bed H3K36me3_H7.{}.unmapped.bed
```

### 2. Гистограммы.

Используем R, чтобы посмотреть на распределение длин участков и проконтролировать, что после применения liftOver оно не изменилось.

	ENCFF063DDB	ENCFF864HTI
hg19	<p>H3K36me3_H7.ENCFF063DDB.hg19 Number of peaks = 51599</p> 	<p>H3K36me3_H7.ENCFF864HTI.hg19 Number of peaks = 109348</p> 
hg38	<p>H3K36me3_H7.ENCFF063DDB.hg38 Number of peaks = 51620</p> 	<p>H3K36me3_H7.ENCFF864HTI.hg38 Number of peaks = 109376</p> 

После liftOver распределение не изменилось, поэтому обрезать по длине не будем.

### 3. Создание github-репозитория.

Для начала создадим репозиторий на самом сайте GitHub:

[https://github.com/WhiteTeaDragon/hse21\\_H3K36me3\\_G4\\_human](https://github.com/WhiteTeaDragon/hse21_H3K36me3_G4_human). Далее настроим гит на сервере.

```
git config --global user.name WhiteTeaDragon
```

```
git config --global user.email AlexSend57@gmail.com
```

```
git config --global core.autocrlf input
```

```
git config --global color.ui auto
```

Склонируем репозиторий на сервер.

```
mkdir github
```

```
cd github
```

```
git clone
```

```
https://github.com/WhiteTeaDragon/hse21\_H3K36me3\_G4\_human.git
```

Теперь добавим в репозиторий файлы, полученные на 1 шаге.

```
cd hse21_H3K36me3_G4_human
```

```
mkdir data
```

```
echo hg38 > ../../hgs.txt
```

```
echo hg19 >> ../../hgs.txt
```

```
cat ../../ids.txt | xargs -tI{} sh -c "cat  
../../hgs.txt | xargs -tI% cp -v  
../../H3K36me3_H7.{}.%.bed data"
```

Отошлём их на сервер.

```
git add .
```

```
git commit -m "initial commit"
```

```
git push
```

Теперь склонируем репозиторий на личный компьютер. Я использую Windows 10, где можно пользоваться командной строкой.

```
cd
```

```
C:\Users\Alexandra\Documents\Биоинформатика\final_project
```

```
mkdir github
```

```
cd github
```

```
git clone
```

```
https://github.com/WhiteTeaDragon/hse21\_H3K36me3\_G4\_human.git
```

Добавим в репозиторий код, строящий гистограммы, и сами гистограммы (картинки проще добавить в клон репозитория через визуальный интерфейс, так как xargs сложно найти на Windows).

```
mkdir hse21_H3K36me3_G4_human/src

copy ..\1seminar.R
hse21_H3K36me3_G4_human/src\1seminar.R

cd hse21_H3K36me3_G4_human

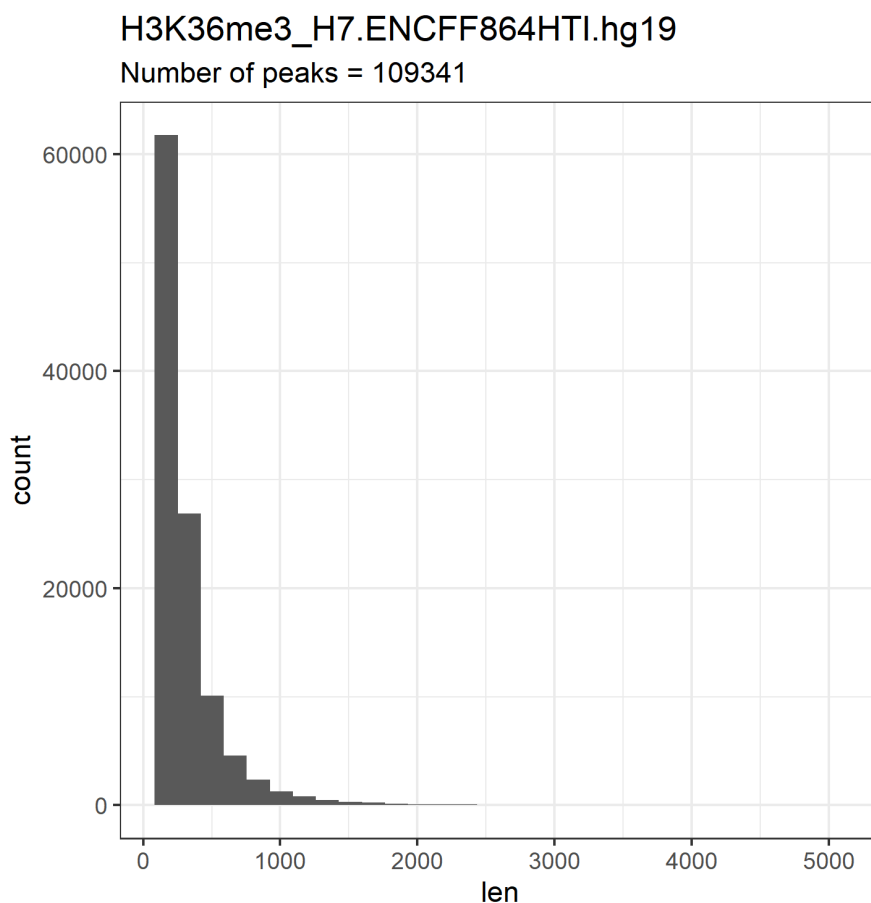
git add .

git commit -m "code and images"

git push
```

#### 4. Фильтрация пиков.

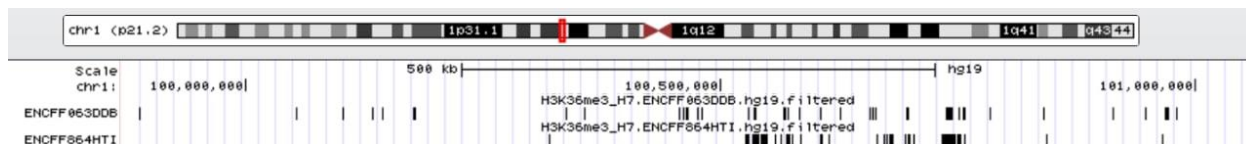
Заметим, что на гистограммах для ENCFF864HTI есть пики длины 8000. Отфильтруем пики так, чтобы длина была не больше 5000. Получится такой график:



Код и полученные файлы добавим на гитхаб (на Windows это удобнее всего делать не через командную строку, а через визуальный интерфейс, поэтому команды я здесь не пишу).

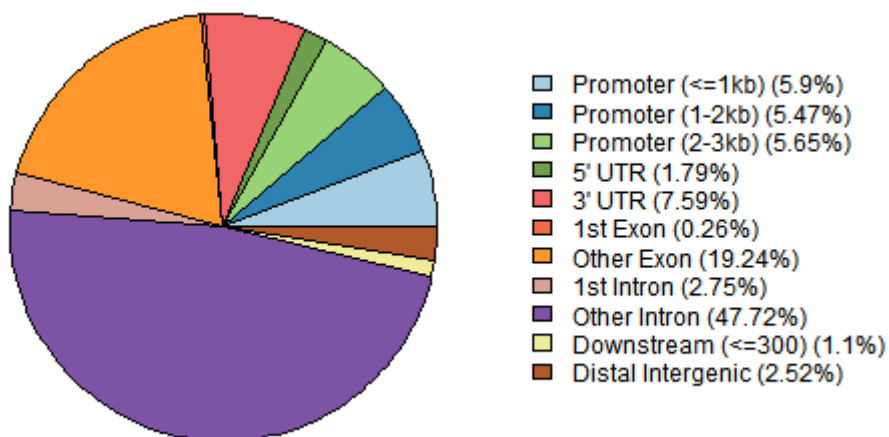
## 5. Визуализация в геномном браузере.

Можно визуализировать в геномном браузере два итоговых файла с пиками (H3K36me3\_H7.ENCFF864HTI.hg19.filtered.bed и H3K36me3\_H7.ENCFF063DDB.hg19.bed).

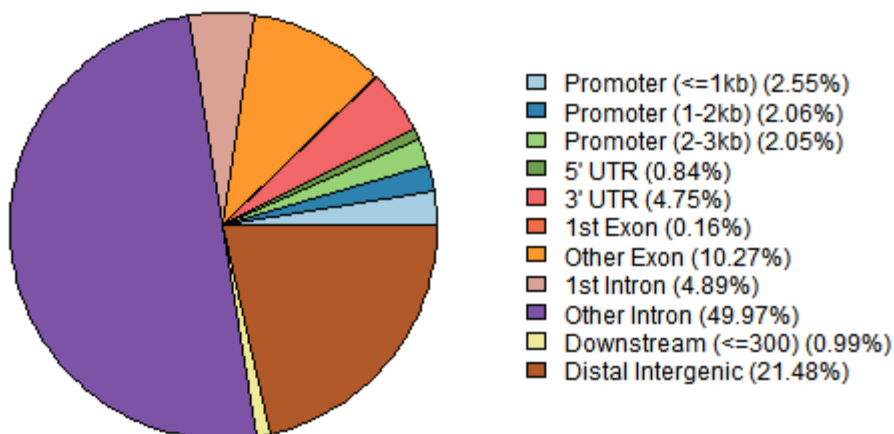


## 6. Графики с распределением пиков по частям генома.

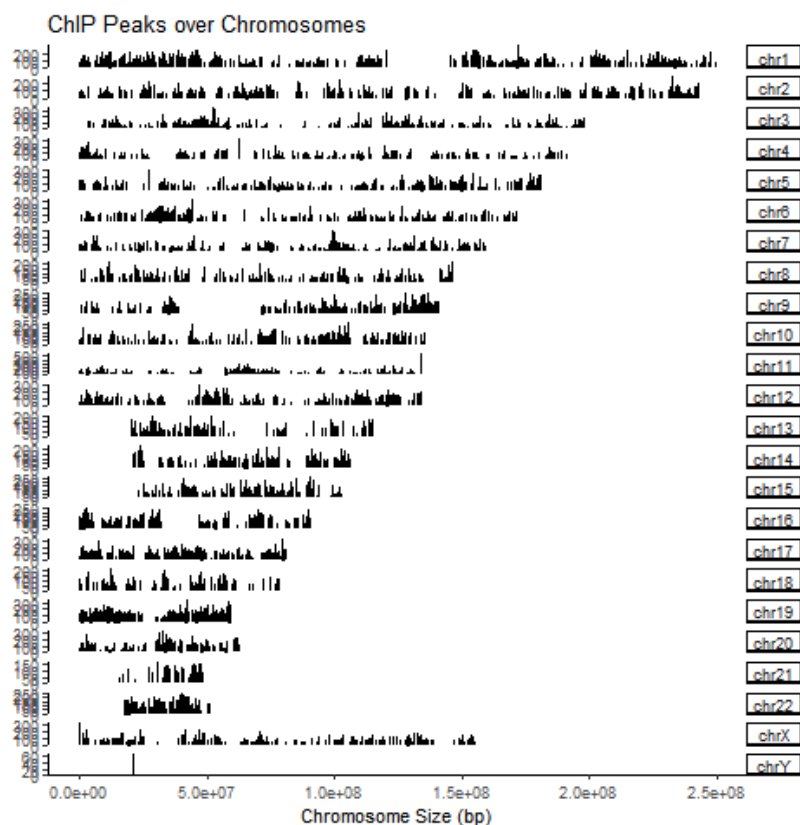
С помощью скрипта на R можно построить круговые диаграммы. Для ENCFF864HTI:



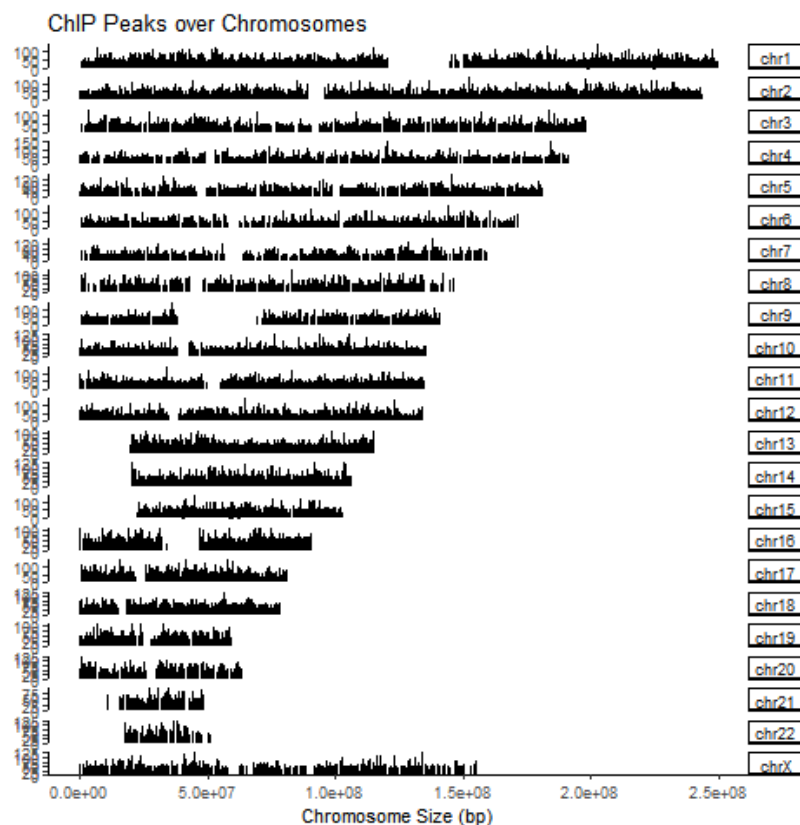
Для ENCFF063DDB:



Также можно посмотреть на распределение пиков по хромосомам.  
Для ENCFF864HTI:



Для ENCFF063DDB:



Эти картинки и генерирующий их код тоже отправим в репозиторий.

## 7. Скачивание файла со вторичной структурой.

Для начала скачаем bed-файлы для G4\_seq\_Li\_K и сольём их в один файл.

```
cd final_project
```

```
mkdir g4
```

```
cd g4
```

```
wget
```

```
"https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSM3003539&format=file&file=GSM3003539%5FHomo%5Fall%5Fw15%5Fth%2D1%5Fminus%2Ehits%2Emax%2EK%2Ew50%2E25%2Ebed%2Egz"
```

Для удобства я переименовываю скачанный файл.

```
mv
```

```
index.html\?acc\=GSM3003539\&format\=file\&file\=GSM3003539_Homo_all_w15_th-1_minus.hits.max.K.w50.25.bed.gz  
GSM3003539_Homo_all_w15_th-1_minus.hits.max.K.w50.25.bed.gz
```

```
wget
```

```
"https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSM3003539&format=file&file=GSM3003539%5FHomo%5Fall%5Fw15%5Fth%2D1%5Fplus%2Ehits%2Emax%2EK%2Ew50%2E25%2Ebed%2Egz"
```

```
mv
```

```
index.html\?acc\=GSM3003539\&format\=file\&file\=GSM3003539_Homo_all_w15_th-1_plus.hits.max.K.w50.25.bed.gz  
GSM3003539_Homo_all_w15_th-1_plus.hits.max.K.w50.25.bed.gz
```

```
zcat *hits.max.K.w50.25.bed.gz | sort -k1,1 -k2,2n |  
bedtools merge > g4.merged.bed
```

## 8. Визуализация в геномном браузере.

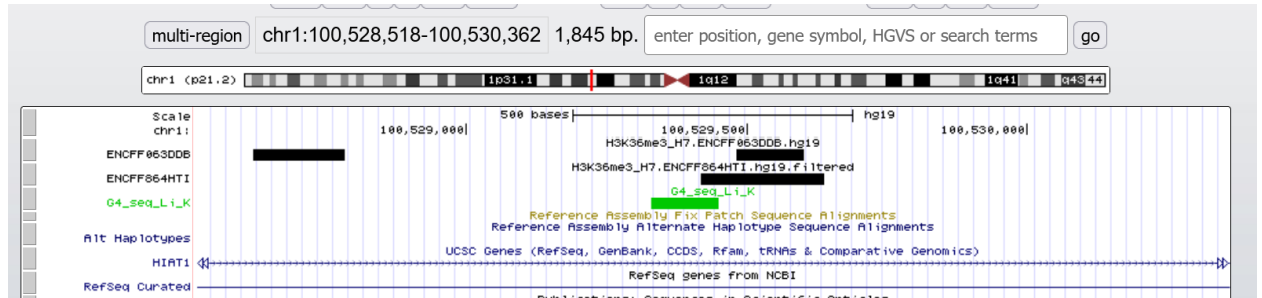
Далее этот файл можно загрузить на гитхаб, после чего визуализировать его в геномном браузере вместе с экспериментами.

```
cd ../github/hse21_H3K36me3_G4_human
git pull
cp ../../g4/g4.merged.bed data
git add .
git commit -m "merged g4 added"
git push
```

В геномный браузер первые 2 файла я загрузила с компьютера, а разметку g4 добавила так:

```
track visibility=dense name="G4_seq_Li_K"
color=0,200,0 description="G4_seq_Li_K"
https://raw.githubusercontent.com/WhiteTeaDragon/hse21\_H3K36me3\_G4\_human/master/data/g4.merged.bed
```

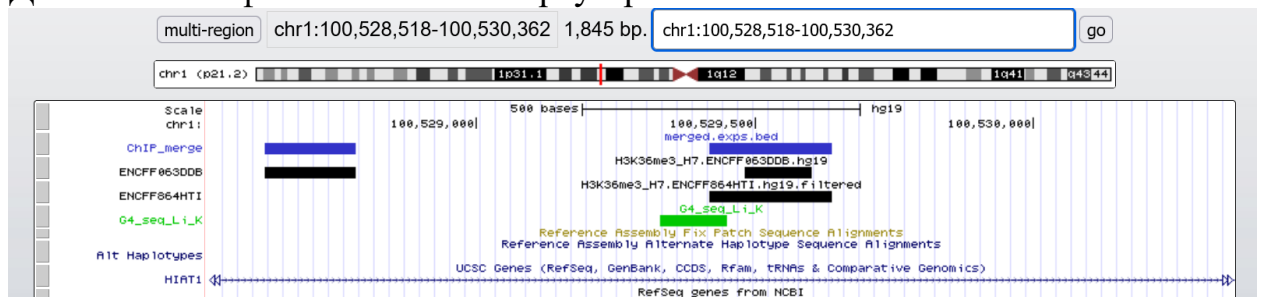
Получилось найти пересечение пиков из экспериментов и g4 в интроне на позиции chr1:100,528,518-100,530,362.



Далее объединим два эксперимента в 1 bed-файл.

```
cat H3K36me3_H7.ENCF063DDB.hg19.bed
H3K36me3_H7.ENCF864HTI.hg19.filtered.bed | sort -k1,1
-k2,2n | bedtools merge > merged.exps.bed
```

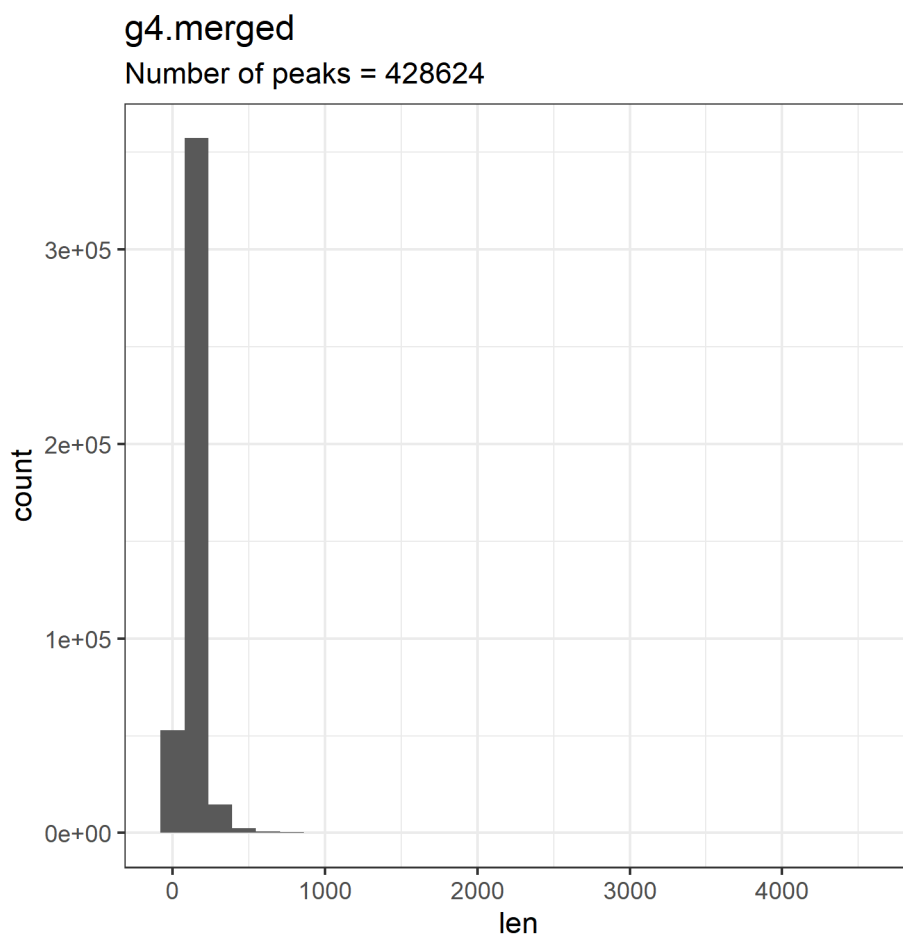
Добавим этот файл в геномный браузер.





## 9. Гистограмма для G4.

Построим гистограмму для файла с участками G4.



Здесь длина пиков примерно такая же, как у пиков из экспериментов – основная масса и там, и там, имеет длину меньше 1000. Добавим эту картинку на гитхаб.

## 10. Пересечение вторичной структуры ДНК и пиков для гистоновой метки.

```
bedtools intersect -a g4.merged.bed -b merged.exps.bed  
> exp.g4.intersection.bed
```

Можно посмотреть на длины файлов:

```
wc g4.merged.bed -l
```

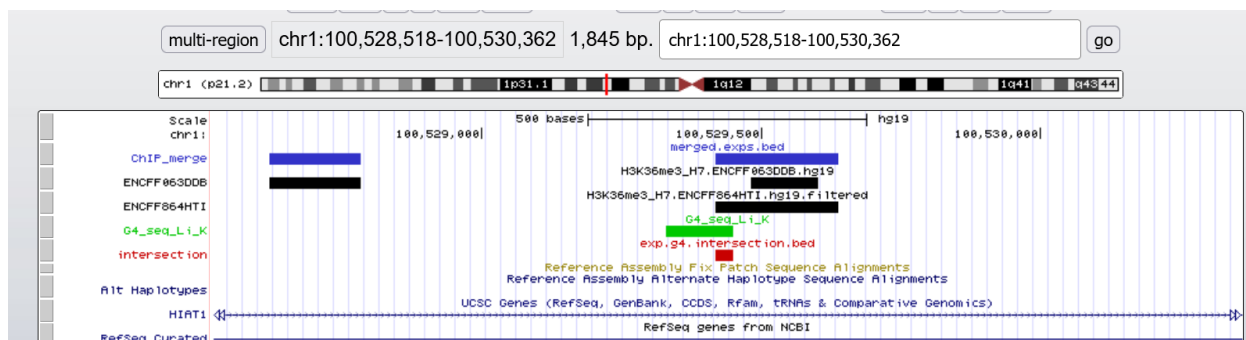
```
wc merged.exps.bed -l
```

```
wc exp.g4.intersection.bed -l
```

```
alsenderovich@laboratory02:~/
428624 g4.merged.bed
alsenderovich@laboratory02:~/
147914 merged.exps.bed
alsenderovich@laboratory02:~/
8527 exp.g4.intersection.bed
```

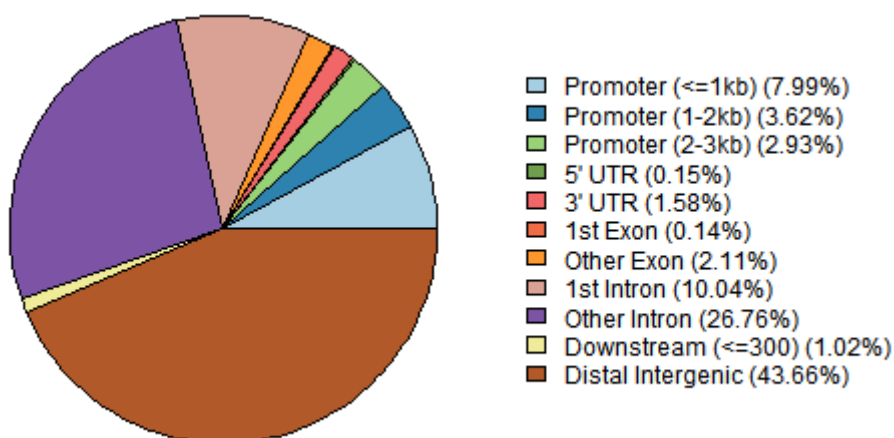
Всего лишь 2 процента G4 пересекаются с пиками из экспериментов. Добавим файлы на гитхаб. Визуализируем пересечение в геномном браузере с помощью строчки

```
track visibility=dense name="intersection"
color=200,0,0 description="exp.g4.intersection.bed"
https://raw.githubusercontent.com/WhiteTeaDragon/hse21-H3K36me3-G4-human/master/data/exp.g4.intersection.bed
```



## 11. Построение круговой диаграммы для вторичной структуры ДНК и пересечения.

Для вторичной структуры ДНК:



Для пересечения:

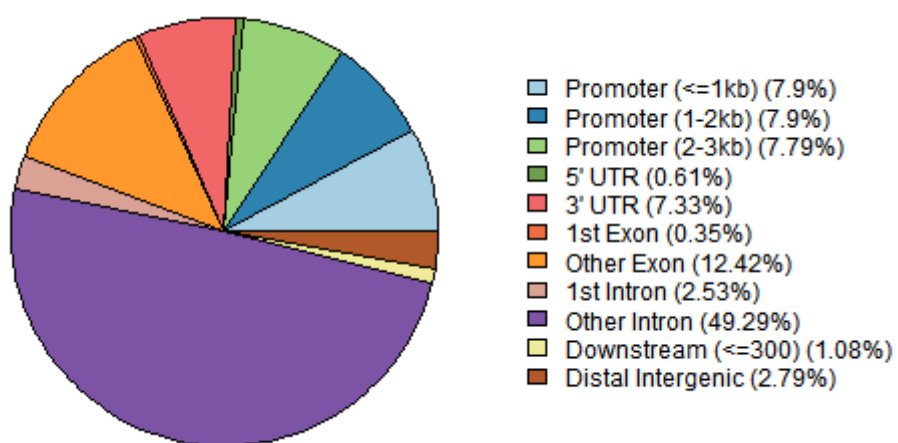


График для пересечения больше всего похож на график для ENCFF864HTI.