

Семинар 13.05.21

0. Выбираем организм (человек, hg19), гистоновую метку, тип клеток и два соответствующих файла с гистоновыми метками:

гистоновая метка	Тип клеток	гистон. метка (.bed файл 1)	гистон. метка (.bed файл 2)
H3K27me3	MCF-7	ENCFF669NUD	ENCFF959VSM

1. Заходим на сервер: 92.242.58.92, 32222

2. wget гистоновые метки

3. Обрезаем ненужные столбцы: `zcat ENCFF669NUD.bed.gz | cut -f1-5 > nud.hg38.bed`

4. Надо конвертировать в 19ую версию сборки генома, для этого используем liftOver:

```
evbulatova@laboratory02:~/project$ liftOver vsm.hg38.bed hg38ToHg19.over.chain vsm.hg19.bed vsm.unmapped.bed
Reading liftover chains
Mapping coordinates
evbulatova@laboratory02:~/project$ liftOver nud.hg38.bed hg38ToHg19.over.chain nud.hg19.bed nud.unmapped.bed
Reading liftover chains
Mapping coordinates
evbulatova@laboratory02:~/project$ ls
ENCFF669NUD.bed.gz      nud.hg19.bed          vsm.hg19.bed
ENCFF959VSM.bed.gz     nud.hg38.bed          vsm.hg38.bed
hg38ToHg19.over.chain  nud.unmapped.bed      vsm.unmapped.bed
```

.chain -- как отличаются сборки, надо скачать официальную версию:

<https://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/hg38ToHg19.over.chain.gz>

Unmapped файлы -- то, что не смог конвертировать liftOver с причинами:

```

evbulatova@laboratory02:~/project$ less nud.unmapped.bed
#Partially deleted in new
chr1    148459321      148459563      Peak_11596     107
#Split in new
chr1    149190898      149191165      Peak_24306     84
#Split in new
chr1    205947238      205957922      Peak_219       337
#Deleted in new
chr1    205958493      205960438      Peak_10588     110
#Deleted in new
chr1    205960806      205962951      Peak_4918      139

```

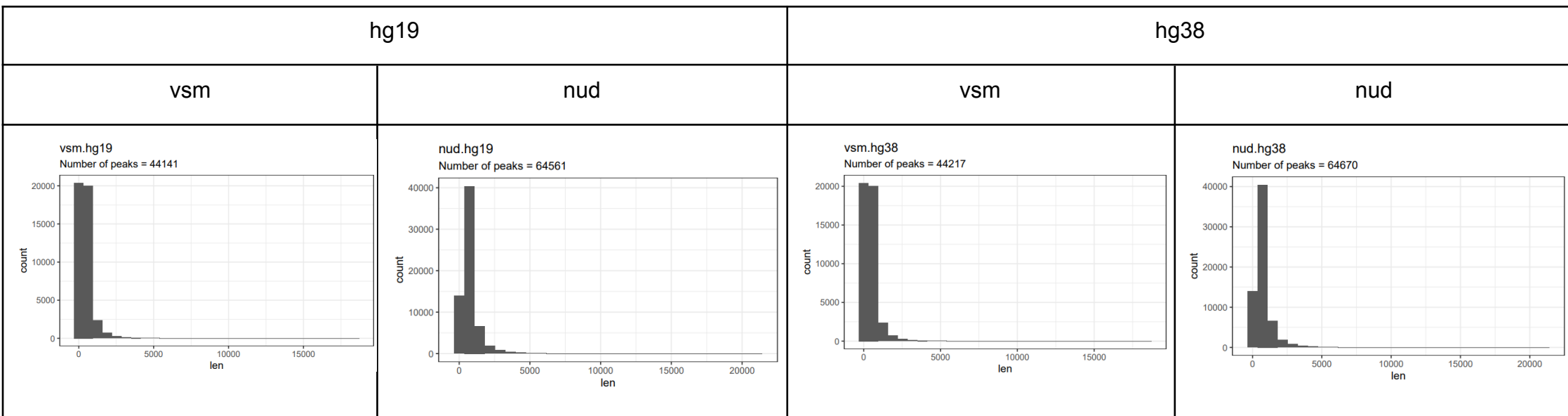
5. Далее файлы перекидываем к себе на компьютер и переходим к написанию кода на R. Мы хотим построить гистограммы длины участков. Сначала просто строим их:

```

1 library(ggplot2)
2 library(dplyr)
3
4 setwd("C:/source/bio/project")
5
6 NAME <- 'vsm.hg38'
7 OUT_DIR <- 'results/'
8
9 bed_df <- read.delim(paste0('data/', NAME, '.bed'), as.is = TRUE, header = FALSE)
10 colnames(bed_df) <- c('chrom', 'start', 'end', 'name', 'score')
11 bed_df$len <- bed_df$end - bed_df$start
12 head(bed_df)
13
14 ggplot(bed_df) +
15   aes(x = len) +
16   geom_histogram() +
17   ggtitle(NAME, subtitle = sprintf('Number of peaks = %s', nrow(bed_df))) +
18   theme_bw()
19 ggsave(paste0('len_hist.', NAME, '.pdf'), path = OUT_DIR)
20

```

Результат:



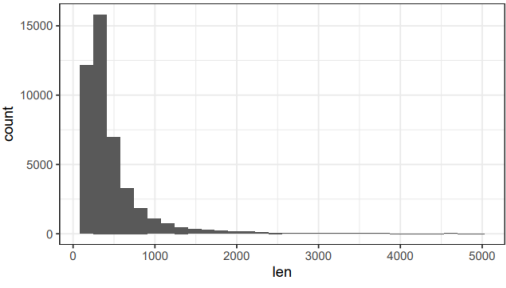
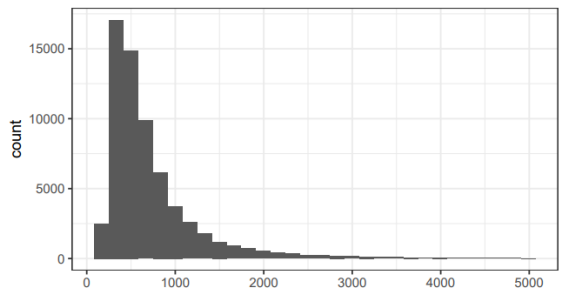
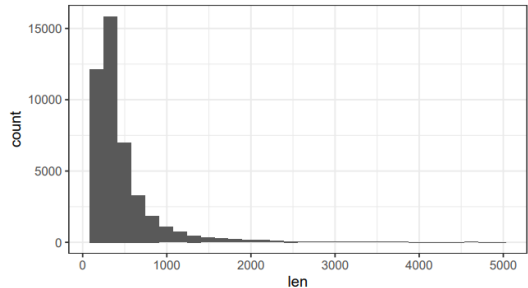
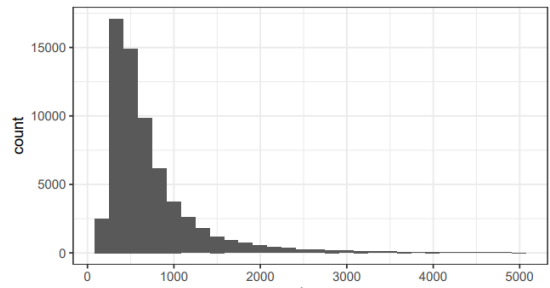
Видим, что есть выбросы. Чтобы с этим бороться, просто откинем слишком большие длины:

```

14 bed_df <- bed_df %>%
15   arrange(-len) %>%
16   filter(len < 5000)
17
18 ggplot(bed_df) +
19   aes(x = len) +
20   geom_histogram() +
21   ggtitle(NAME, subtitle = sprintf('Number of peaks = %s', nrow(bed_df))) +
22   theme_bw()
23 ggsave(paste0('len_hist.', NAME, '.filtered.pdf'), path = OUT_DIR)
24
25 bed_df %>%
26   select(-len) %>%
27   write.table(file='data/nud.hg19.filtered.bed',
28             col.names = FALSE, row.names = FALSE, sep = '\\t', quote = FALSE)
29

```

Посмотрим на отфильтрованные результаты:

hg19		hg38	
vsm	nud	vsm	nud
<div><div>vsm.hg19</div><div>Number of peaks = 44053</div></div>	<div><div>nud.hg19</div><div>Number of peaks = 64314</div></div>	<div><div>vsm.hg38</div><div>Number of peaks = 44128</div></div>	<div><div>nud.hg38</div><div>Number of peaks = 64422</div></div>

Семинар 20.05.21

```
evbulatova@laboratory02:~/project$ cd git/hse21_H3K27me3_ZDNA_human/
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ mkdir data
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ cp -v ~/project/vsm.hg19.bed
cp: missing destination file operand after '/home/evbulatova/project/vsm.hg19.bed'
Try 'cp --help' for more information.
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ cp -v ~/project/vsm.hg19.bed data/
'/home/evbulatova/project/vsm.hg19.bed' -> 'data/vsm.hg19.bed'
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ cp -v ~/project/nud.hg19.bed data/
'/home/evbulatova/project/nud.hg19.bed' -> 'data/nud.hg19.bed'
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ cp -v ~/project/nud.hg38.bed data/
'/home/evbulatova/project/nud.hg38.bed' -> 'data/nud.hg38.bed'
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ cp -v ~/project/vsm.hg38.bed data/
'/home/evbulatova/project/vsm.hg38.bed' -> 'data/vsm.hg38.bed'
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ ls
data  README.md
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ cd data
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ ls
nud.hg19.bed  nud.hg38.bed  vsm.hg19.bed  vsm.hg38.bed
```

```
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ git add data/
```

```
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ git commit -m "initial .bed files"
[main 063d679] initial .bed files
 4 files changed, 217589 insertions(+)
 create mode 100644 data/nud.hg19.bed
 create mode 100644 data/nud.hg38.bed
 create mode 100644 data/vsm.hg19.bed
 create mode 100644 data/vsm.hg38.bed
```

```
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human$ git push
Username for 'https://github.com': qwerty-Bk
Password for 'https://qwerty-Bk@github.com':
Counting objects: 7, done.
Delta compression using up to 8 threads.
Compressing objects: 100% (7/7), done.
Writing objects: 100% (7/7), 2.65 MiB | 259.00 KiB/s, done.
Total 7 (delta 0), reused 0 (delta 0)
To https://github.com/qwerty-Bk/hse21_H3K27me3_ZDNA_human.git
 e432274..063d679  main -> main
```

```
kateb@LAPTOP-2F8BJAUE MINGW64 ~/Documents/bio
$ git clone https://github.com/qwerty-Bk/hse21_H3K27me3_ZDNA_human.git
Cloning into 'hse21_H3K27me3_ZDNA_human'...
remote: Enumerating objects: 10, done.
remote: Counting objects: 100% (10/10), done.
remote: Compressing objects: 100% (8/8), done.
remote: Total 10 (delta 0), reused 7 (delta 0), pack-reused 0
Receiving objects: 100% (10/10), 2.65 MiB | 2.69 MiB/s, done.






kateb@LAPTOP-2F8BJAUE MINGW64 ~/Documents/bio
$ ls
hse21_H3K27me3_ZDNA_human/

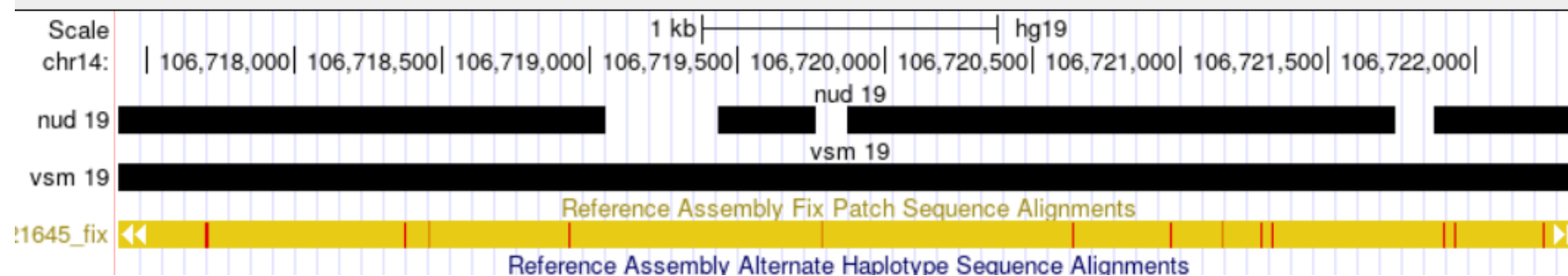
kateb@LAPTOP-2F8BJAUE MINGW64 ~/Documents/bio
$ cd hse21_H3K27me3_ZDNA_human/

kateb@LAPTOP-2F8BJAUE MINGW64 ~/Documents/bio/hse21_H3K27me3_ZDNA_human (main)
$ ls
README.md  data/
```

```
$ ls
filtered_peaks.r  not_filtered_peaks.r

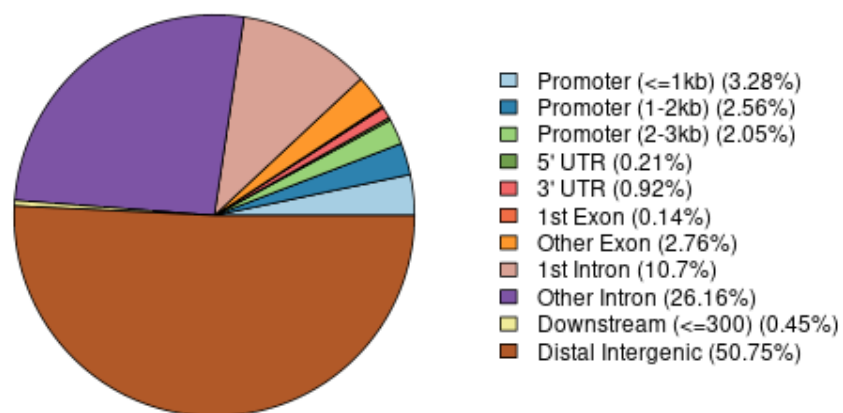
kateb@LAPTOP-2F8BJAUE MINGW64 ~/Documents/bio/hse21_H3K27me3_ZDNA_human/src (main)
```

	qwerty-Bk r code	
	data	initial .bed files
	results	r code
	src	r code
	README.md	Initial commit

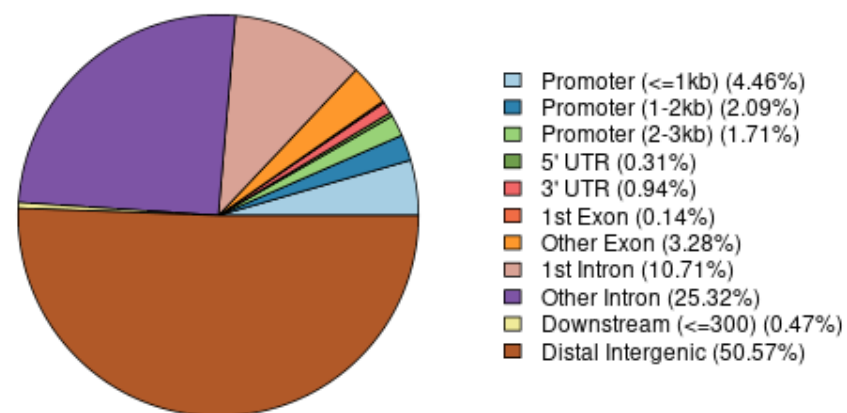


Чипсикер на расположение пиков гистоновой метки относительно аннотированных генов:

vsm



nud



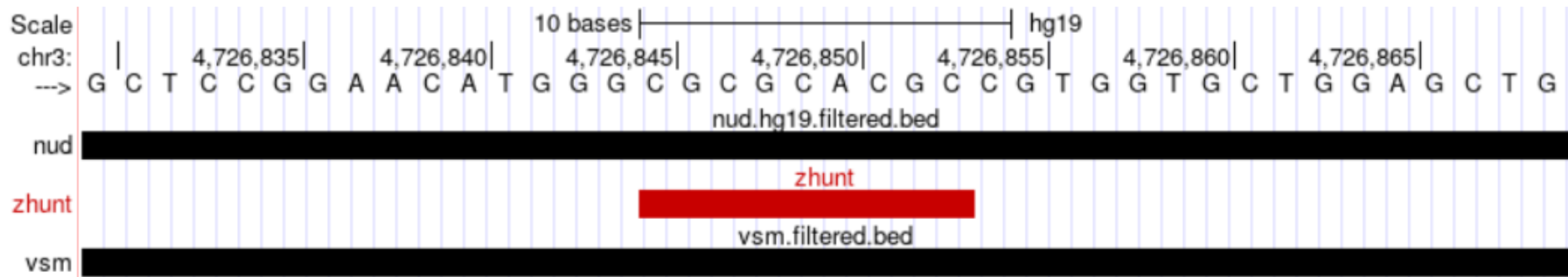
Семинар 27.05.21

Мы выбрали одну из вторичных структур, я взяла Z-Hunt (сначала я рассмотрела Deer-Z, но он давал меньше пересечений с метками).

Вообще, даже с Z-Hunt пересечений оказалось крайне мало.

За час максимальное найденное мной пересечение имеет длину 9 нуклеотидов:

chr3:4,724,895-4,728,894



Сливаем два файла:

```
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ cat *19.filtered.bed | sort -k1,1 -k2,2n |  
bedtools merge > merge.hg19.bed
```

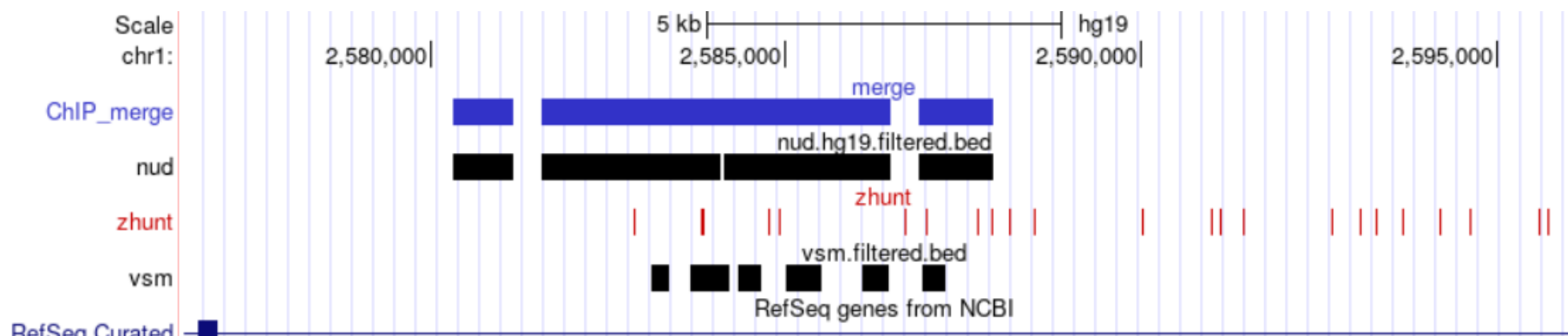
Строк в слиянии не столько, сколько строк в сумме, так как многие регионы пересекаются:

```

evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ ls *19.filtered.bed -l
-rw-rw-r-- 1 evbulatova evbulatova 2446950 мая 27 12:30 nud.hg19.filtered.bed
-rw-rw-r-- 1 evbulatova evbulatova 1658594 мая 27 12:30 vsm.hg19.filtered.bed
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ ls merge.hg19.bed -l
-rw-rw-r-- 1 evbulatova evbulatova 1905174 мая 27 12:31 merge.hg19.bed

```

Снова визуализируем:



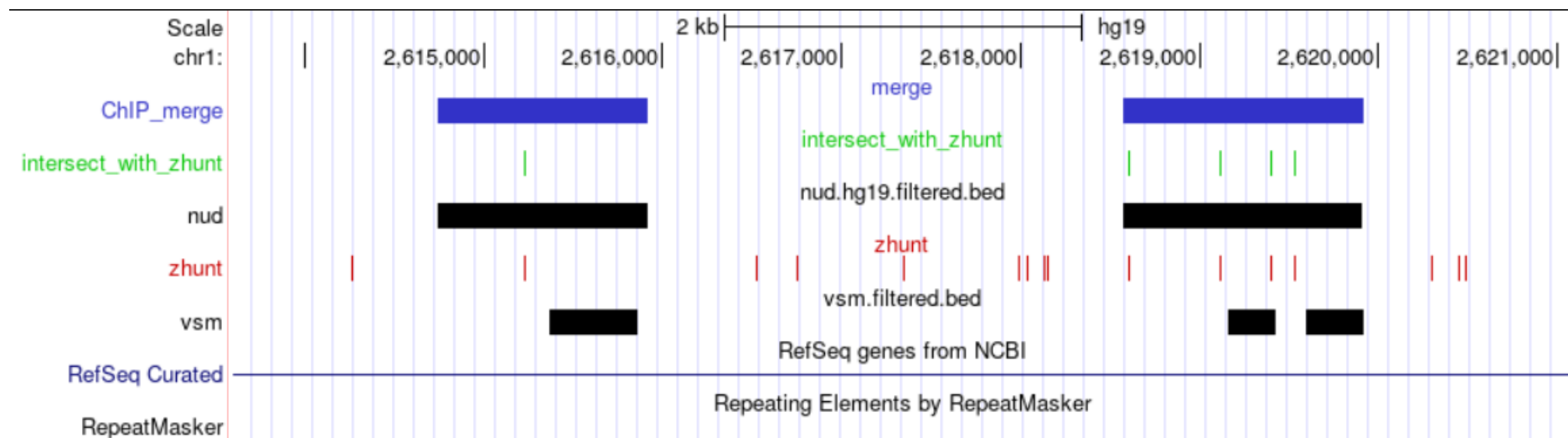
Пересекаем с zhunt:

```

evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ bedtools intersect -a zhunt.bed -b merge.hg19.
bed > intersect_with_zhunt.bed
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ ls merge.hg19.bed -l
-rw-rw-r-- 1 evbulatova evbulatova 1905174 мая 27 12:31 merge.hg19.bed
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ ls zhunt.bed -l
-rw-rw-r-- 1 evbulatova evbulatova 9838318 мая 27 12:30 zhunt.bed
evbulatova@laboratory02:~/project/git/hse21_H3K27me3_ZDNA_human/data$ ls intersect_with_zhunt.bed -l
-rw-rw-r-- 1 evbulatova evbulatova 233665 мая 27 12:46 intersect_with_zhunt.bed

```

Как видно, файл не очень большой. Если бы мы искали участки, где zhunt пересекается и с тем, и с другим, то нашли бы совсем мало.

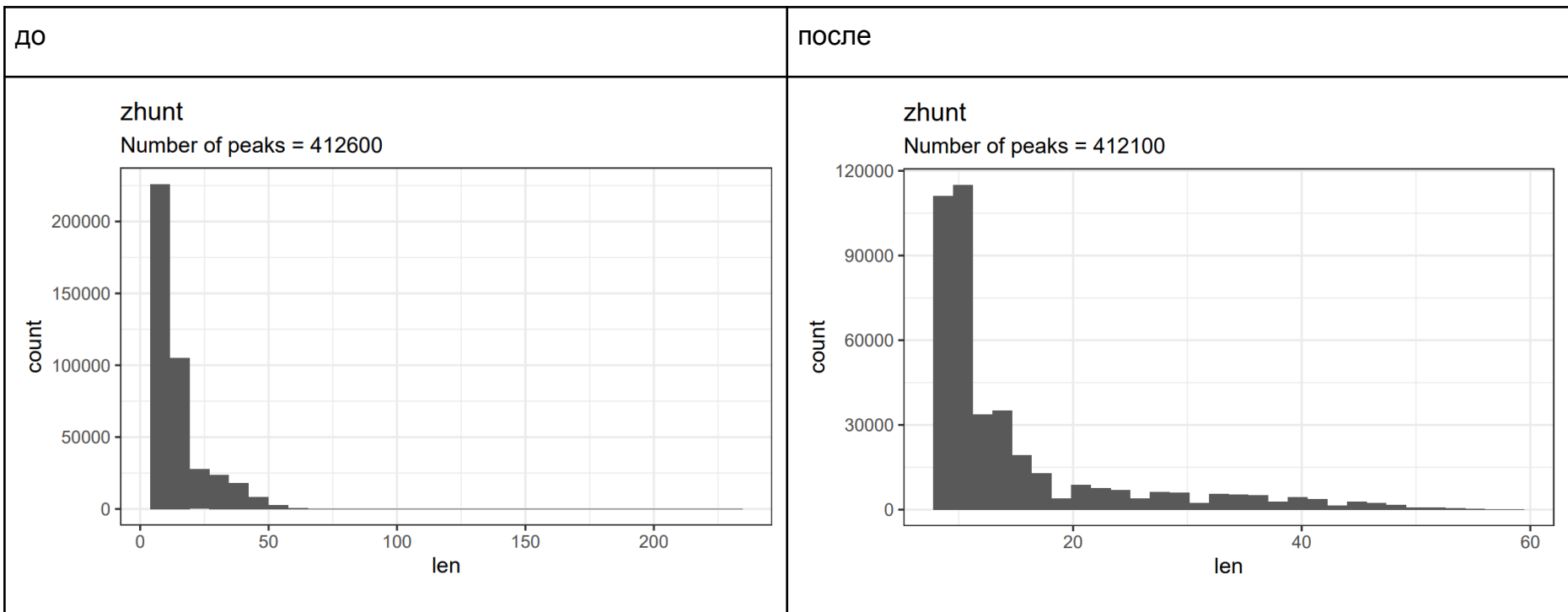


Сессия сохранена в <https://genome.ucsc.edu/s/kateb/project21>.

Самостоятельная работа

Анализ вторичной структуры

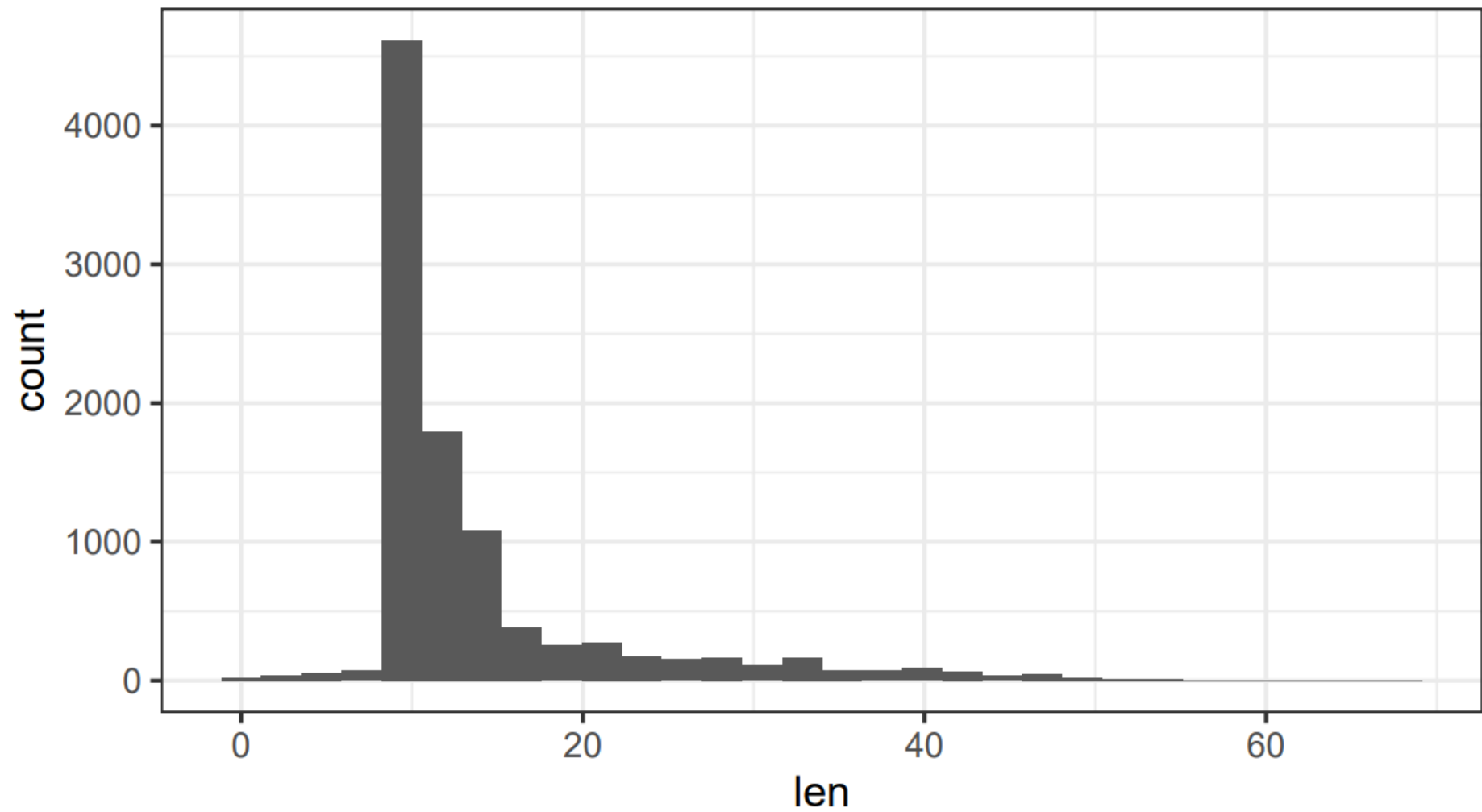
Распределение пиков zhunt до обрезания и после:



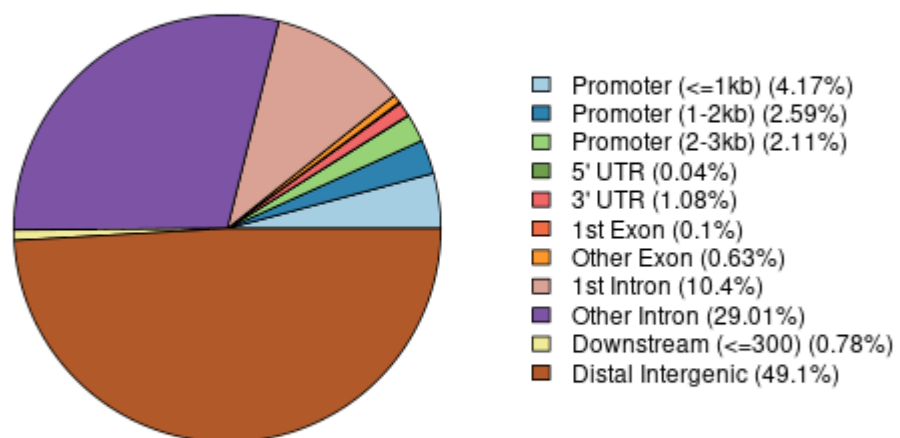
Распределение длин пересечений:

intersect_with_zhunt

Number of peaks = 9801



Чипсикер:



Анализ пересечений (продолжение)

Для анализа пересечений запускаем код на ассоциацию пересечений с ближайшими генами. Общее количество уникальных генов -- 298; пиков -- 502.

Проводим GO анализ, сохраняем данные таблицей.

Наименьшие значения FDR, т.е. наиболее значимые категории:

1. multicellular organism development
2. system development
3. anatomical structure development
4. multicellular organismal process
5. developmental process

Закидываем все полученные данные на гитхаб.