

Open Chromatin and Methylation

Лев Мазаев | мАДБМ18

Этап 1: поиск данных на ENCODE

CTCF ChIP-Seq

Experiment search

Clear Filters

Assay type

DNA binding 19

Selected filters: TF ChIP-seq

Search

TF ChIP-seq 19

Assay title

Selected filters: TF ChIP-seq

Search

Status

Selected filters: released

released 19

Project

GGR 12

ENCODE 7

RFA

Showing 19 of 19 results

Add all items to cart

TF ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 8 hours

Target: CTCF
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 8 hours

Experiment ENCSR432AXE

TF ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 7 hours

Target: CTCF
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 7 hours

Experiment ENCSR173IEA

TF ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 6 hours

Target: CTCF
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 6 hours

Experiment ENCSR287UXS

Взял эксперимент [ENCSR876UZD](#) - *Homo sapiens* A549 treated with 100 nM dexamethasone for 5 hours. Файлы bed narrowPeak: [ENCFF702GWV](#) и [ENCFF921XYL](#).

H3K4me3 ChIP-Seq

Experiment search

Clear Filters

Assay type

DNA binding 15

Selected filters: Histone ChIP-seq

Search

Histone ChIP-seq 15

Assay title

Selected filters: released

released 15

Status

Selected filters: released

released 15

Project

GGR 12

ENCODE 3

RFA

GGR 12

ENCODE2 3

Genome assembly

Selected filters: GRCh38

GRCh38 15

hg19 3

Target category

histone 15

narrow histone mark 15

Showing 15 of 15 results

Add all items to cart

Histone ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 12 hours

Target: H3K4me3
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 12 hours

Experiment ENCSR944WVU

Histone ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 10 hours

Target: H3K4me3
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 10 hours

Experiment ENCSR139GDM

Histone ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 8 hours

Target: H3K4me3
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 8 hours

Experiment ENCSR618MUP

Histone ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 7 hours

Target: H3K4me3
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 7 hours

Experiment ENCSR285FZP

Histone ChIP-seq of A549
Homo sapiens A549 treated with 100 nM dexamethasone for 6 hours

Target: H3K4me3
Lab: Tim Reddy, Duke
Project: GGR
Treatment: dexamethasone (CHEBI:41879) for 6 hours

Experiment ENCSR646OPC

В задании указано, что нужно взять broadPeak для ChIP-Seq, но не во всех экспериментах есть такие данные. Подходящий эксперимент - [ENCSR000DPD](#) - *Homo sapiens* A549. Файлы bed broadPeak: **ENCFF001WUL** и **ENCFF001WUM**. Как выяснилось позднее, сами данные размечены по геному hg19, то есть нужно будет перевести их в hg38, используя [hgLiftOver](#).

[ATAC-Seq](#)

Showing 5 of 5 results

Experiment	Assay type	Description	Lab	Project	Treatment	Status
ENCSR265ZXX	ATAC-seq	ATAC-seq of A549 <i>Homo sapiens</i> A549 treated with 100 nM dexamethasone for 12 hours	Tim Reddy, Duke	GGR	dexamethasone (CHEBI:41879) for 12 hours	released
ENCSR074AHH		ATAC-seq of A549 <i>Homo sapiens</i> A549 treated with 100 nM dexamethasone for 8 hours	Tim Reddy, Duke	GGR	dexamethasone (CHEBI:41879) for 8 hours	released
ENCSR288YMH		ATAC-seq of A549 <i>Homo sapiens</i> A549 treated with 100 nM dexamethasone for 4 hours	Tim Reddy, Duke	GGR	dexamethasone (CHEBI:41879) for 4 hours	released
ENCSR139OYS		ATAC-seq of A549 <i>Homo sapiens</i> A549 treated with 100 nM dexamethasone for 1 hour	Tim Reddy, Duke	GGR	dexamethasone (CHEBI:41879) for 1 hour	released
ENCSR220ASC		ATAC-seq of A549 <i>Homo sapiens</i> A549	Tim Reddy, Duke	GGR		released

Взял эксперимент [ENCSR074AHH](#) - *Homo sapiens* A549 treated with 100 nM dexamethasone for 8 hours. Файлы bed narrowPeak: **ENCFF846JAO**, **ENCFF153HNH** и **ENCFF198GUB**.

[BS-Seq](#)

Showing 1 of 1 results

Experiment	Assay type	Description	Lab	Project	Status
ENCSR481JIW	DNA methylation	WGBS of A549 <i>Homo sapiens</i> A549	Richard Myers, HAIB	ENCODE	released

Тут только один эксперимент: [ENCSR481JIW](#) - whole-genome shotgun bisulfite sequencing (WGBS) - *Homo sapiens* A549. Файлы methylation state at CpG: **ENCFF005TID** и **ENCFF003JVR**.

Скачивание производилось с помощью следующей команды:

```
xargs -L 1 curl -O -L < files.txt
```

Этап 2: bed -> bigWig

Далее, распакуем все файлы с помощью `gunzip`, добавим в названия типа эксперимента и переведём их в формат bedGraph. То есть, должно остаться только 4 колонки: chr, start, end, score. Также попутно удалим все неосновные хромосомы, используя регулярное выражение `^chr[0-9YX][0-9]?` для первой колонки:

```

for file in `ls -1 ./bed`
do
    awk '$1 ~ /^[chr][0-9YX][0-9]?$/ {print $1"\t"$2"\t"$3"\t"$5}'
    ./bed/$file \
        | sort -k1,1 -k2,2n > ./bedGraph/${file::-3}bedGraph
done

```

bedGraph для H3K4me3 реплик переразметим на геном hg38 используя [hgLiftOver](#) и заново отсортируем (ещё пришлось объединить по 5-10 пересекающихся регионов в каждом файле при помощи bedtools merge). Теперь переведём все bedGraph в формат bigWig:

```

fetchChromSizes hg38 > hg38.chrom.sizes # и очистим от лишних хромосом

for file in `ls -1 ./bedGraph`
do
    ./bedGraphToBigWig ./bedGraph/${file} hg38.chrom.sizes
    ./bigWig/${file::-8}bigWig
done

```

Результат:

```
(chr) leo@MS7922:~/BioData/NGS_HW7$ ls -htsR
.:
total 3,0M
4,0K bigWig  4,0K bedGraph  4,0K bed  4,0K toBigWig.sh   12K hg38.chrom.sizes  4,0K toBedGraph.sh  3,0M bedGraphToBigWig

./bigWig:
total 748M
720K chip_h3k4me3_ENCFF001WUM.bigWig  636K chip_ctcf_ENCFF702GWV.bigWig  1,1M atac_ENCFF846JAO.bigWig
652K chip_h3k4me3_ENCFF001WUL.bigWig  371M bs_ENCFF005TID.bigWig  1,2M atac_ENCFF198GUB.bigWig
608K chip_ctcf_ENCFF921XYL.bigWig  371M bs_ENCFF003JVR.bigWig  1,1M atac_ENCFF153HNH.bigWig

./bedGraph:
total 2,9G
1,7M chip_h3k4me3_ENCFF001WUL.bedGraph  1,1M chip_ctcf_ENCFF702GWV.bedGraph  1,9M atac_ENCFF846JAO.bedGraph
1,9M chip_h3k4me3_ENCFF001WUM.bedGraph  1,5G bs_ENCFF005TID.bedGraph  2,1M atac_ENCFF198GUB.bedGraph
1,2M chip_ctcf_ENCFF921XYL.bedGraph  1,5G bs_ENCFF003JVR.bedGraph  2,0M atac_ENCFF153HNH.bedGraph

./bed:
total 7,0G
3,5G bs_ENCFF003JVR.bed  3,3M chip_h3k4me3_ENCFF001WUL_hg19.bed  7,7M atac_ENCFF846JAO.bed
3,5G bs_ENCFF005TID.bed  8,3M atac_ENCFF198GUB.bed  2,9M chip_ctcf_ENCFF921XYL.bed
3,7M chip_h3k4me3_ENCFF001WUM_hg19.bed  7,9M atac_ENCFF153HNH.bed  2,8M chip_ctcf_ENCFF702GWV.bed

```

Этап 3: Графики обогащения

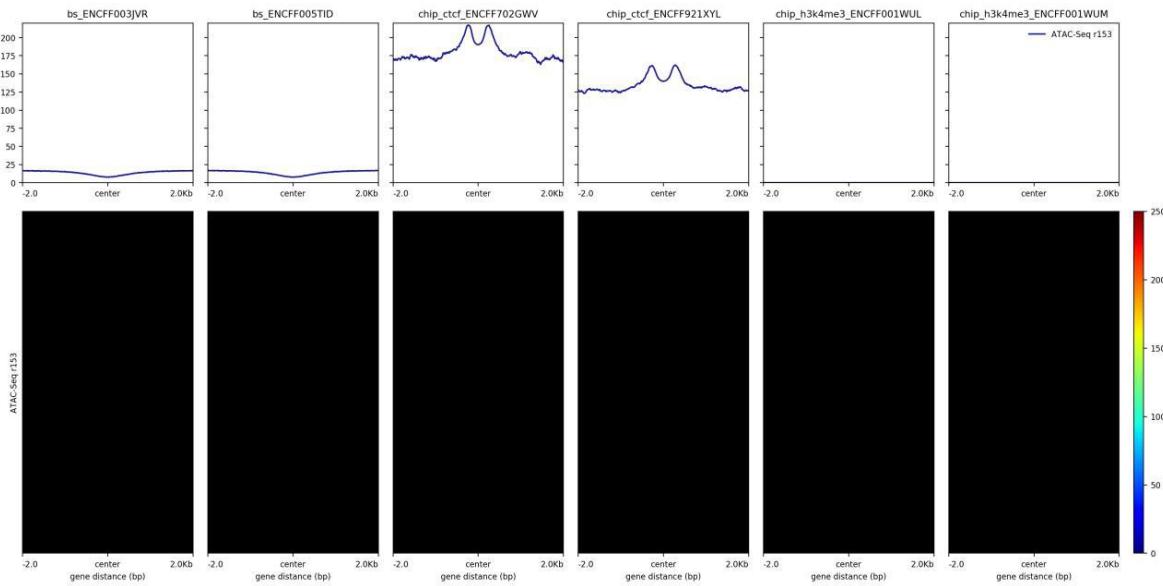
Уберём bigWig для ATAC-Seq из соответствующей папки и построим графики обогащения для всех данных вокруг пиков первой реплики ATAC-Seq:

```

computeMatrix reference-point -S ./bigWig/* -R ./bed/atac_ENCFF153HNH.bed \
--referencePoint center -a 2000 -b 2000 -out me153.tab.gz -p 8
plotHeatmap -m me153.tab.gz -out me153.png --heatmapHeight 15 --heatmapWidth 7
--colorMap jet --sortRegion ascend --regionsLabel 'ATAC-Seq r153'

```

Вот что получилось:

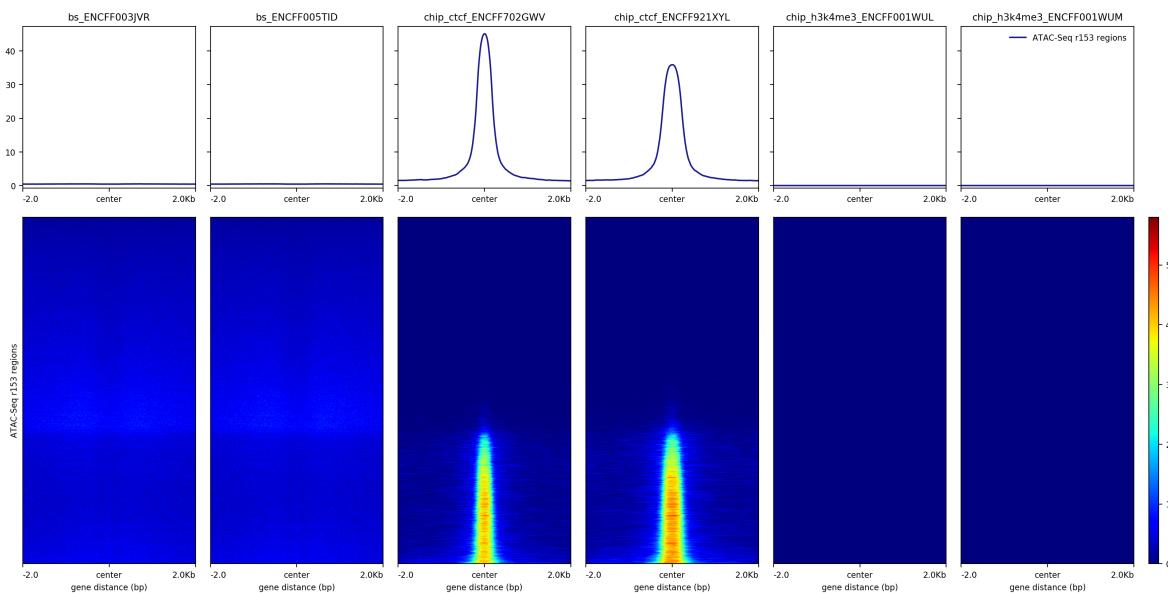


Как видим, получилось ничего. В результате гугления, выяснилось, что когда геном очень неплотно покрыт сигналом, как в нашем случае (кроме данных WGBS), нужно использовать дополнительную опцию `--missingDataAsZero`. Во всех файлах, кроме WGBS очень мало записей:

```
(base) leo@MS7922:~/BioData/NGS_HW7/bed$ wc -l *
 73440 atac_ENCFF153HNH.bed
 77204 atac_ENCFF198GUB.bed
 71579 atac_ENCFF846JAO.bed
 58607924 bs_ENCFF003JVR.bed
 58607924 bs_ENCFF005TID.bed
 41213 chip_ctcf_ENCFF702GWV.bed
 42573 chip_ctcf_ENCFF921XYL.bed
 67014 chip_h3k4me3_ENCFF001WUL_hg19.bed
 75873 chip_h3k4me3_ENCFF001WUM_hg19.bed
117664744 total
```

Повторный запуск с новой опцией:

```
computeMatrix reference-point -S ./bigWig/* -R ./bed/atac_ENCFF153HNH.bed --
referencePoint center -a 2000 -b 2000 -out me153.tab.gz -p 8 --
missingDataAsZero
plotHeatmap -m me153.tab.gz -out me153.png --heatmapHeight 15 --heatmapWidth 7
--regionsLabel 'ATAC-Seq r153 regions' --colorMap jet --sortRegion ascend
```



Вот теперь, какой-никакой результат есть. Благодаря опции `--sortRegions ascend` регионы сортированы **по возрастанию среднего значения**. Первые две колонки - Bisulfite Sequencing для CpG метилирования, наблюдается слабый сигнал для всех пиков ATAC-Seq (целиком по каждому пику), и он везде примерно одинаковый. Видимо у этих данных слабое покрытие. Вторые две колонки - ChIP-Seq транскрипционного фактора CTCF, примерно для трети ATAC-Seq регионов сигнал очень существенный в центре, но при этом узкий. При этом в тех регионах, где сигнал начинает обрываться, видна и некоторая раздельная полоса в данных BS. Видимо это из-за сортировки регионов: после раздела BS начинает влиять на среднее по региону, чем ChIP-Seq для CTCF. Для данных H3K4me3 - не видно ничего, так как score всех регионов в этих broadPeak-файлах - нулевой.

```
(base) leo@MS7922:~/BioData/NGS_HW7$ cut -f 5 ./bed/chip_h3k4me3_ENCFF001WUL_hg19.bed | sort | uniq -c
 67014 0
(base) leo@MS7922:~/BioData/NGS_HW7$ cut -f 5 ./bed/chip_h3k4me3_ENCFF001WUM_hg19.bed | sort | uniq -c
 75873 0
```

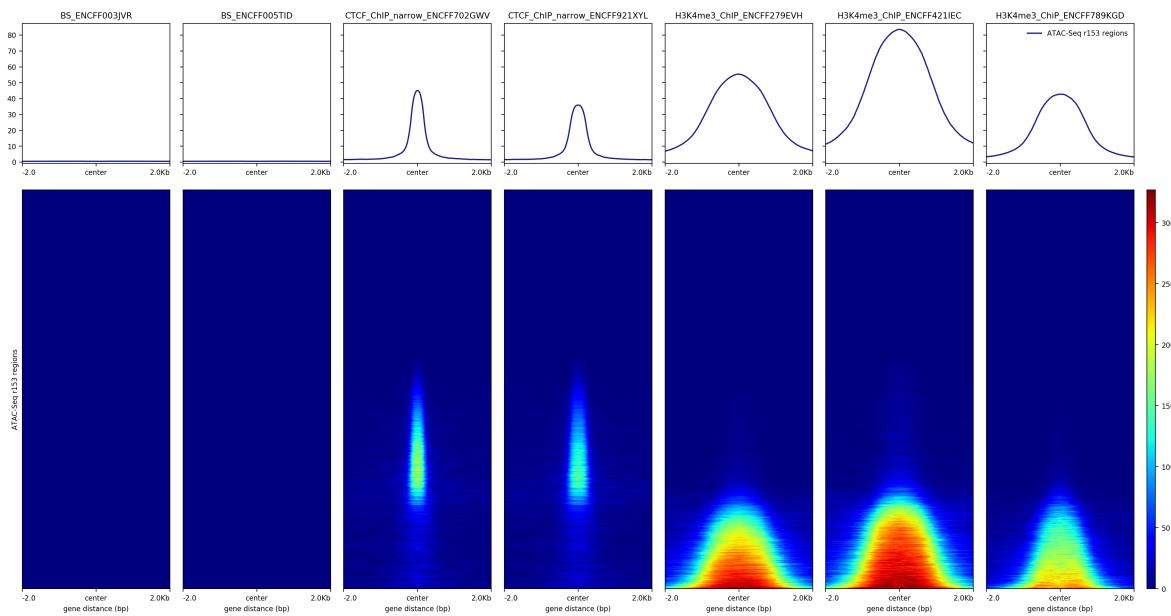
Поэтому эти данные просто не подходят для такого анализа, попробуем их заменить на что-нибудь другое, правда придется отказаться от broadPeak (таких данных очень мало и ENCODE обозначает их как плохие по качеству, еще они все размечены по hg19) в пользу narrowPeak.

Скачаем [ENCSR524UOX](#) - *Homo sapiens* A549 treated with 100 nM dexamethasone for 5 hours, H3K4me3 ChIP-seq. Файлы bed narrowPeak: **ENCFF279EVH**, **ENCFF421IEC** и **ENCFF789KGD**. Итого, получилось следующее:

```
(chr) leo@MS7922:~/BioData/NGS_HW7/bigWig$ ls -hts
total 744M
464K H3K4me3_ChIP_ENCFF789KGD.bigWig 608K CTCF_ChIP_narrow_ENCFF921XYL.bigWig 371M BS_ENCFF003JVR.bigWig
444K H3K4me3_ChIP_ENCFF421IEC.bigWig 636K CTCF_ChIP_narrow_ENCFF702GWV.bigWig
368K H3K4me3_ChIP_ENCFF279EVH.bigWig 371M BS_ENCFF005TID.bigWig
```

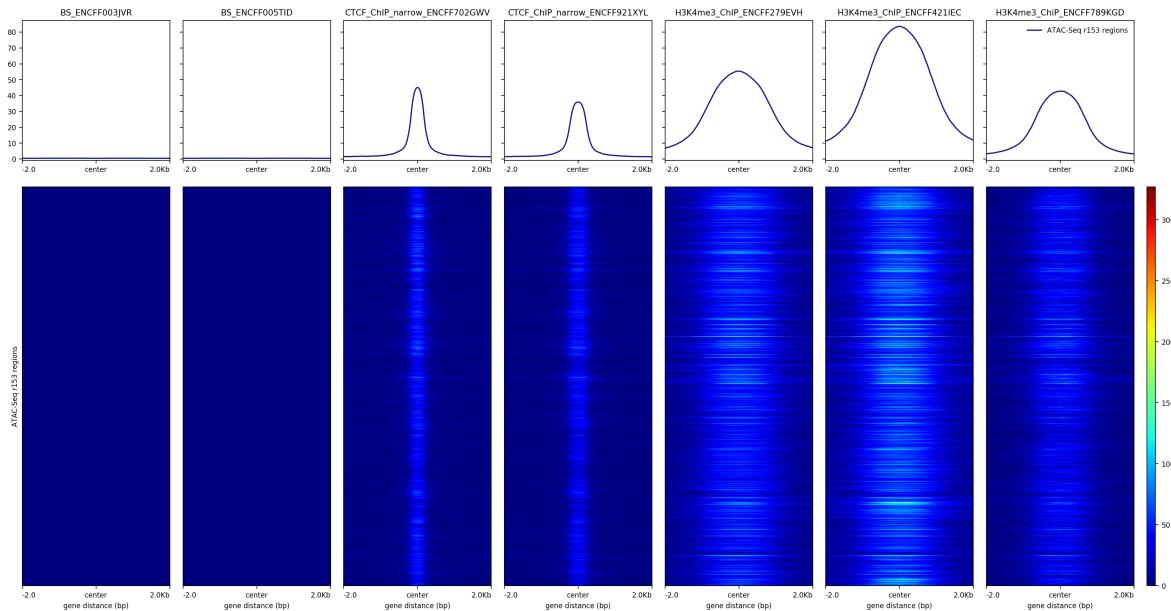
Будем запускать все эти файлы против каждой реплики ATAC-Seq:

```
computeMatrix reference-point -S ./bigWig/* -R ./atac/atac_ENCFF153HNH.bed --
referencePoint center -a 2000 -b 2000 -out me153.tab.gz -p 8 --
missingDataAsZero
plotHeatmap -m me153.tab.gz -out me153.png --heatmapHeight 20 --heatmapWidth 7
--regionsLabel 'ATAC-Seq r153 regions' --colorMap jet --sortRegion ascend
```



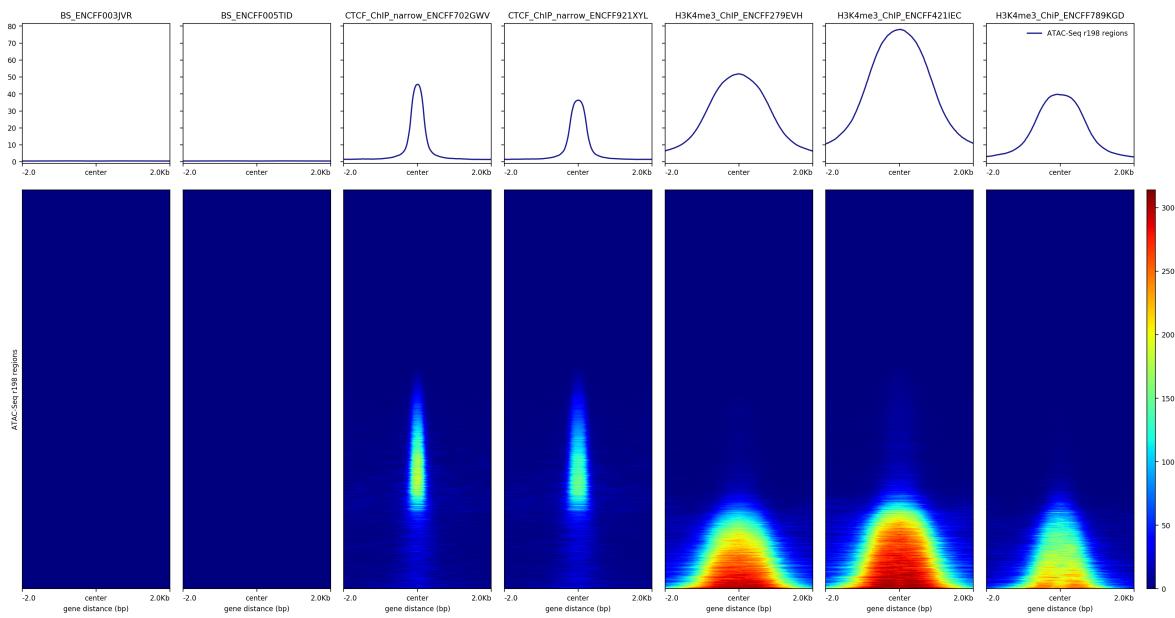
Попробуем сделать картинку без сортировки регионов по значению:

```
plotHeatmap -m me153.tab.gz -out me153_unsorted.png --heatmapHeight 20 --
heatmapWidth 7 --regionsLabel 'ATAC-Seq r153 regions' --colorMap jet --
sortRegion no
```



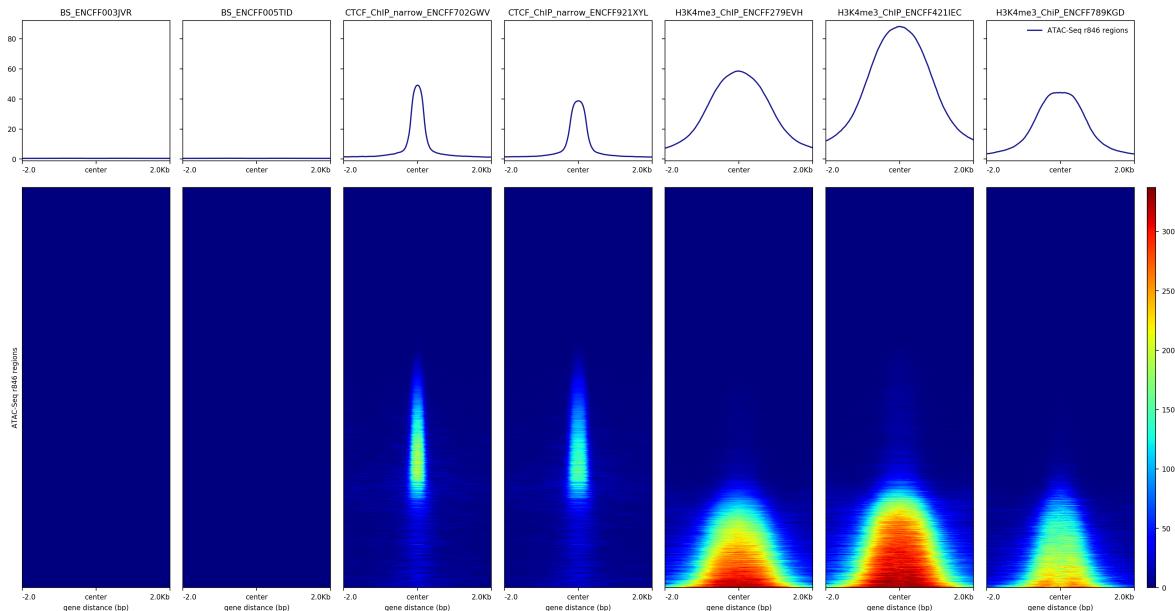
Теперь у H3K4me3 настолько сильный сигнал, что вероятно из-за него не видно BS. Также этот сигнал довольно широкий, простирается почти целиком на все 4000-нуклеотидное окно с центром в пике ATAC-Seq, в то время как сигнал CTCF довольно узкий и сосредоточен только в области пика. Ещё CTCF и H3K4me3 не скоррелированы по регионам - из-за их сортировки по среднему. Тут может быть и биологическая причина: когда метилирование слишком сильное, CTCF не так часто присоединяется к ДНК, а при умеренном метилировании - максимально часто. Другая реплика:

```
computeMatrix reference-point -S ./bigWig/* -R ./atac/atac_ENCFF198GUB.bed --
referencePoint center -a 2000 -b 2000 -out me198.tab.gz -p 8 --
missingDataAsZero
plotHeatmap -m me198.tab.gz -out me198.png --heatmapHeight 20 --heatmapWidth 7
--regionsLabel 'ATAC-Seq r198 regions' --colorMap jet --sortRegion ascend
```



Такой же результат. Последняя ATAC-Seq реплика:

```
computeMatrix reference-point -S ./bigWig/* -R ./atac/atac_ENCFF846JAO.bed --
referencePoint center -a 2000 -b 2000 -out me846.tab.gz -p 8 --
missingDataAsZero
plotHeatmap -m me846.tab.gz -out me846.png --heatmapHeight 20 --heatmapWidth 7
--regionsLabel 'ATAC-Seq r846 regions' --colorMap jet --sortRegion ascend
```



От реплики к реплике результат остается один и тот же. Что в целом можно сказать:

- У Bisulfite Sequencing видимо очень слабое покрытие, сигнал был виден только до того, как мы обновили H3K4me3 данные. Сам сигнал равномерно покрывает всё 4000-нуклеотидное окно с центром в пике ATAC-Seq. Т. е. с биологической точки зрения наблюдается CpG-метилирование всего окна вокруг пика ATAC-Seq.
- CTCF данные дают узкий пик по центру окна. Так как регионы у нас сортированы по среднему значению (и видимо по всем входным трекам), то, глядя на картинки, можно сделать вывод, что для наиболее множественного соединения CTCF с ДНК излишнее метилирование H3K4 вредно. Узкий сигнал CTCF расположен там же, где и пик ATAC-Seq, что логично, так как CTCF садится на ДНК именно в середине участка открытого хроматина, т. е. на пике ATAC-Seq.

- Первые данные по метилированию гистоновых концов оказались очень низкого качества, score у всех регионов был нулевой, поэтому пришлось взять другие. Что видим на этих данных: в хороших по среднему значению регионах сигнал очень высокий по всему окну целиком, с уменьшением среднего значения ширина сигнала начинает падать, амплитуда тоже. Биологически это означает, что есть регионы открытого хроматина, где почти все гистоновые концы метилированы, но есть и такие регионы, где модификации гистоновых концов наблюдаются только в центре, и этого достаточно для доступа к ДНК. При таком "умеренном" сценарии также увеличивается активность CTCF .

Посмотрим, что получится если вычислить матрицу по всем репликам:

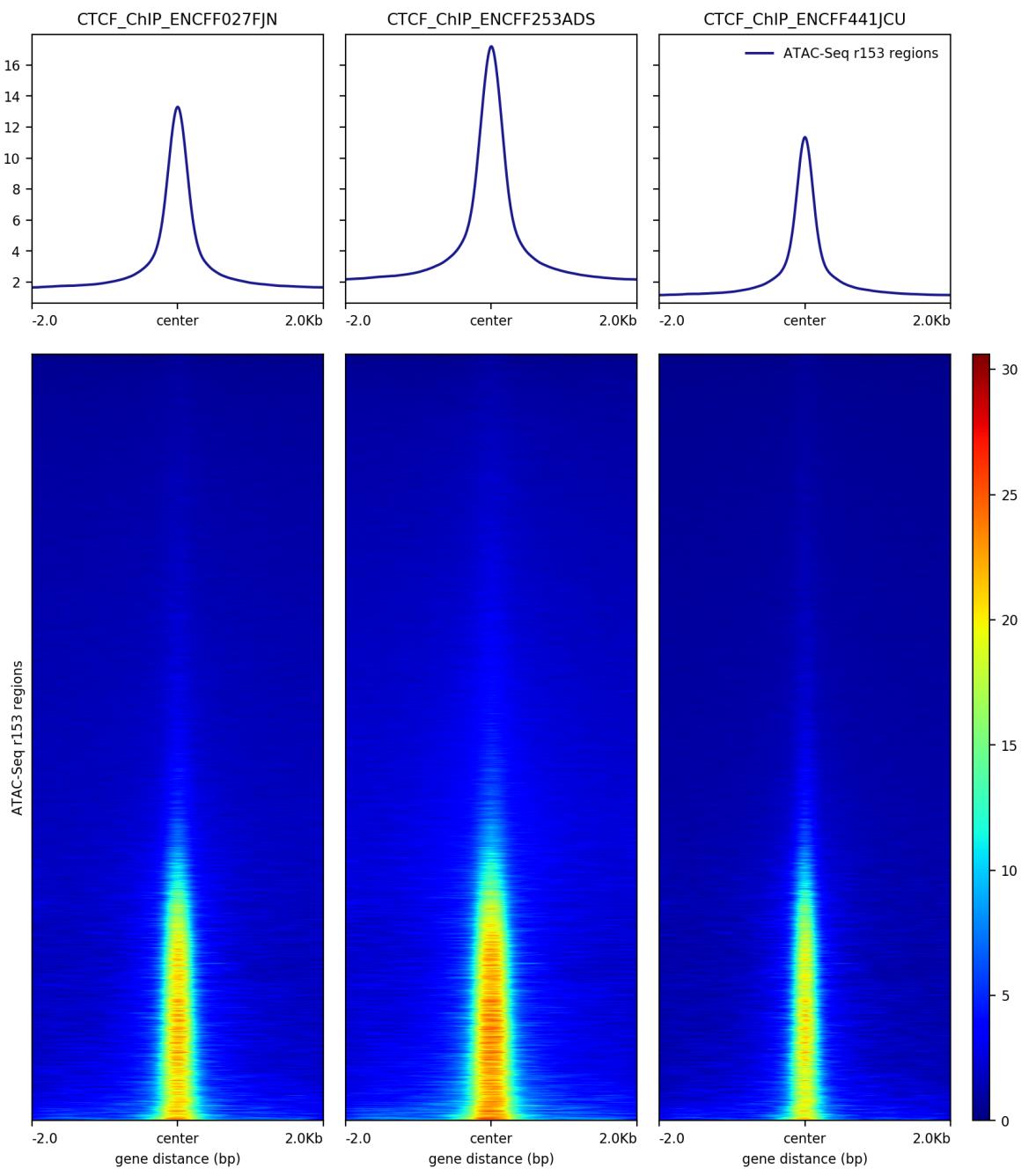
```
computeMatrix scale-regions -S ./bigWig/* -R ./atac/* -b 1000 -a 1000 -out all_scaled.tab.gz -p 8 --missingDataAsZero
plotHeatmap -m all_scaled.tab.gz -out all_scaled.png --heatmapHeight 20 --
heatmapWidth 7 --regionsLabel 'ATAC-Seq scaled' --colorMap jet --sortRegion
ascend
```

К сожалению не удалось:

MemoryError: Unable to allocate array with shape (221515, 2100) and data type float64

Попробуем ещё 3 bigWig из данных CTCF (**ENCFF027FJN**, **ENCFF253ADS**, **ENCFF441JCU** - это именно bigWig, скачанные с [ENCSR876UZD](#), далее их использовать не будем) против первой реплики:

```
computeMatrix reference-point -S ./CTCFdownloadedbigWigs/* -R
./atac/atac_ENCFF153HNH.bed --referencePoint center -a 2000 -b 2000 -out
exp_bw.tab.gz -p 8 --missingDataAsZero
plotHeatmap -m exp_bw.tab.gz -out exp_bw.png --heatmapHeight 20 --heatmapWidth
7 --regionsLabel 'ATAC-Seq r153 regions' --colorMap jet --sortRegion ascend
```



Здесь результат вполне согласуется с тем, что мы видели до того: узкие линии, совпадающие по положению с пиком ATAC-Seq. При этом наблюдается слабый сигнал и вокруг основного: видимо причина в том, что эти bigWig сделаны напрямую с BAM-ов, то есть это просто покрытие ридами, в то время как наши bigWig-и сделаны из данных после процедуры peak calling.

Этап 4: Профили всех сигналов вокруг TSS человека

Воспользуемся сервисом Biomart, чтобы скачать позиции TSS:

New **Count** **Results** URL XML Perl Help

Dataset Human genes (GRCh38.p13) Filters Chromosome/scaffold: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y Attributes Chromosome/scaffold name Transcription start site (TSS) Gene end (bp)	Export all results to File TSV <input type="checkbox"/> Unique results only <input checked="" type="checkbox"/> Go Email notification to View 10 rows as HTML <input type="checkbox"/> Unique results only <table border="1" style="width: 100%; border-collapse: collapse; font-size: 0.8em;"> <thead> <tr> <th>Chromosome/scaffold name</th> <th>Transcription start site (TSS)</th> <th>Gene end (bp)</th> </tr> </thead> <tbody> <tr><td>1</td><td>167005959</td><td>167006077</td></tr> <tr><td>1</td><td>237120807</td><td>237121109</td></tr> <tr><td>1</td><td>26602432</td><td>26602432</td></tr> <tr><td>1</td><td>203729581</td><td>203729705</td></tr> <tr><td>1</td><td>25023503</td><td>25023586</td></tr> <tr><td>1</td><td>99791662</td><td>99791753</td></tr> <tr><td>1</td><td>200054061</td><td>200054165</td></tr> <tr><td>1</td><td>206748092</td><td>206748092</td></tr> <tr><td>1</td><td>25964197</td><td>25964300</td></tr> <tr><td>1</td><td>207708981</td><td>207708981</td></tr> </tbody> </table>	Chromosome/scaffold name	Transcription start site (TSS)	Gene end (bp)	1	167005959	167006077	1	237120807	237121109	1	26602432	26602432	1	203729581	203729705	1	25023503	25023586	1	99791662	99791753	1	200054061	200054165	1	206748092	206748092	1	25964197	25964300	1	207708981	207708981
Chromosome/scaffold name	Transcription start site (TSS)	Gene end (bp)																																
1	167005959	167006077																																
1	237120807	237121109																																
1	26602432	26602432																																
1	203729581	203729705																																
1	25023503	25023586																																
1	99791662	99791753																																
1	200054061	200054165																																
1	206748092	206748092																																
1	25964197	25964300																																
1	207708981	207708981																																

Таким образом, есть bed-файл без стренда, соответственно первое число всегда будет браться в качестве начала региона. Дополнение названий хромосом для соответствия аннотации UCSC, а также `bedtools merge`, чтобы сократить число регионов (те, что пересекаются - скорее всего один и тот же ген с разными TSS):

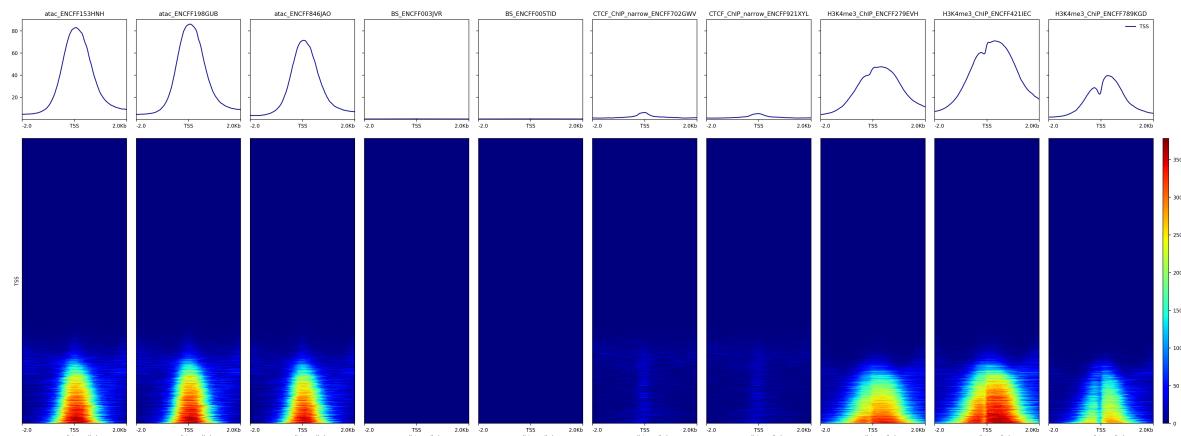
```
awk '{a="chr"$0; print a}' mart_export.txt | sort -k1,1 -k2,2n >
~/BioData/NGS_HW7/tss_raw.bed
bedtools merge tss_raw.bed > tss.bed
```

Также Теперь к уже использованным ранее bigWig, которые мы получили из разных bed-файлов, добавим ещё 3 новых с ATAC-Seq данных:

```
(chr) leo@MS7922:~/BioData/NGS_HW7$ ls -hts ./bigWig/
total 747M
464K H3K4me3_ChIP_ENCFF789KGD.bigWig      636K CTCF_ChIP_narrow_ENCFF702GWV.bigWig  1,2M atac_ENCFF198GUB.bigWig
444K H3K4me3_ChIP_ENCFF421IEC.bigWig       371M BS_ENCFF005STID.bigWig           1,1M atac_ENCFF153HNH.bigWig
368K H3K4me3_ChIP_ENCFF279EVH.bigWig        371M BS_ENCFF003JVR.bigWig
608K CTCF_ChIP_narrow_ENCFF921XYL.bigWig   1,1M atac_ENCFF846JAO.bigWig
```

И сделаем матрицу против TSS:

```
computeMatrix reference-point -S ./bigWig/* -R tss.bed --referencePoint TSS -a
2000 -b 2000 -out tss.tab.gz -p 8 --missingDataAsZero
plotHeatmap -m tss.tab.gz -out tss.png --heatmapHeight 20 --heatmapWidth 7 --
regionsLabel 'TSS' --colorMap jet --sortRegion ascend
```



В данном случае центр каждого окна - точка начала транскрипции. Все данные - ATAC-Seq, CTCF ChIP-Seq и H3K4me3 ChIP-Seq центрируются на этой же точке. Это говорит о том, что это открытый хроматин, что здесь CTCF, и что присутствует метилирование концов гистонов. Данным BS видимо опять не хватает интенсивности.