HW1

1) Найдите в интернете и скачайте на кластер fastqc. Проверьте все входные данные при помощи fastqc.

Скачиваем fastqc с сайта, распаковываем, меняем режим и запускаем цикл для всех файлов.

```
> wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.9.zip
> unzip fastqc_v0.11.9.zip
> chmod +x FastQC/fastqc
> FastQC/fastqc -f fastq -o ./rna-seq/rna-fastqc /mnt/local/vse2020/shared/rna.seq.data/*
> for i in `ls -1 /mnt/local/vse2020/home/anastasia_rastvorova/rna-seq/rna-fastqc`; do
unzip /mnt/local/vse2020/home/anastasia_rastvorova/rna-seq/rna-fastqc/$i; done
```

2) Найдите в интернете и скачайте бинарники для последних версий hisat2.

```
> wget ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/downloads/hisat2-2.1.0-Linux_x86_64.zip
> unzip hisat2-2.1.0-Linux_x86_64.zip
```

3) Зайти на ensembl.org → downloads → Download data via FTP → скачать последовательность 19 хромосомы мыши и её аннотацию в формате gtf (для всего генома).

```
> wget ftp://ftp.ensembl.org/pub/release-101/fasta/mus_musculus/dna/Mus_musculus.GRCm38.
dna.chromosome.19.fa.gz
> gzip -d Mus_musculus.GRCm38.dna.chromosome.19.fa.gz
> mv Mus_musculus.GRCm38.dna.chromosome.19.fa chr19.fa #переименовала для удобства

> wget ftp://ftp.ensembl.org/pub/release-101/gtf/mus_musculus/Mus_musculus.GRCm38.101.
chr.gtf.gz
> gzip -d Mus_musculus.GRCm38.101.gtf.gz
```

Можно было не распаковывать, но я зачем-то распаковала.

4) Отфильтруйте из аннотации только 19ую хромосому при помощи команды grep -P '^19\t'.

```
> grep -P '^19\t' Mus_musculus.GRCm38.101.gtf > chr19.gtf
```

5) Постройте индекс по последовательности 19-ой хромосомы при помощи команды hisat2-build (без координат сайтов)

```
> hisat2-2.1.0/./hisat2-build hisat2-build chr19.fa index_chr19
```

6) Прокартируйте все fq файлы (начните с одного) на 19-ую хромосому при помощи hisat2 не допуская обрезания ридов и сообщив hisat2 координаты сайтов сплайсинга.

Получим сайты сплайсинга.

```
> hisat2-2.1.0/./hisat2_extract_splice_sites.py Mus_musculus.GRCm38.101.gtf >
splicing_sites.txt
```

Запуск hisat2 я решила это делать вручную, поэтому каждую команду подавала построчно.

```
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/B14.5.fq.gz |
samtools view -bS - > B14.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/B15.5.fq.gz |
samtools view -bS - > B15.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/B17.5.fq.gz |
samtools view -bS - > B17.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/B20.fq.gz |
samtools view -bS - > B20.bam
```

```
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/B34.fq.gz |
samtools view -bS - > B34.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/C14.5.fq.gz |
samtools view -bS - > C14.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/C15.5fq.gz |
samtools view -bS - > C15.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/C17.5.fq.gz |
samtools view -bS - > C17.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/C15.5.fq.gz |
samtools view -bS - > C15.5.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/C20.fq.gz |
samtools view -bS - > C20.bam
> hisat2-2.1.0/./hisat2 --no-softclip -x index_chr19 --known-splicesite-infile
splicing_sites_mus.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/C34.fq.gz |
samtools view -bS - > C34.bam
```

7) Выберите случайно один образец.

Я выбрала B17.5.

8) Сколько ридов картируется в регион 19:12485000-12490000 в этом образце?

Сначала отсортируем и проиндексируем файл.

```
> samtools sort B17.5.bam -o B17.5.sorted.bam
> samtools index B17.5.sorted.bam
> samtools view -c B17.5.sorted.bam 19:12485000-12490000
> samtools view B17.5.sorted.bam 19:12485000-12490000 | wc -l
```

Прокартировалось 165 ридов.

9) Сколько из них картируется только в одно место генома.

```
> samtools view -q 60 -c B17.5.sorted.bam 19:12485000-12490000
```

163

10) Сколько ридов картровалось без замен? Сколько с 1, 2 и т. д. заменами?

Посмотрим замены по CIGAR командой, где X – мисметчи:

```
> samtools view B17.5.sorted.bam 19:12485000-12490000 | awk '$6 ~ /X/ || $1 ~ /^@/' | wc -l
```

0

Ридов с заменами нет.

11) Сколько ридов картировалось на экзон-экзонные границы?

N – это пропущенный регион.

```
> samtools view B17.5.sorted.bam 19:12485000-12490000 | awk '$6 ~ /N/ || $1 ~ /^@/' | wc -l
```

58

HW2

1) Прокартируйте все образцы при помощи hisat2.

Прокартировали в предыдущем ДЗ.

2) Соберите транскрипты при помощи stringtie для каждого образца используя аннотацию из ensembl (-G).

```
> wget http://ccb.jhu.edu/software/stringtie/dl/stringtie-2.1.4.Linux_x86_64.tar.gz
> tar -xzf stringtie-2.1.4.Linux_x86_64.tar.gz
```

Отсортируем все .bam файлы.

```
> for i in `ls -1 /mnt/local/vse2020/home/anastasia_rastvorova/rna-seq/sort`; do samtools sort /mnt/local/vse2020/home/anastasia_rastvorova/rna-seq/sort/$i -o sorted.$i; done
```

И для каждого запустим Stringtie.

```
> for i in `ls -1 rna-seq/sort`; do stringtie-2.1.4.Linux_x86_64/stringtie rna-seq/sort/$i -o stringtie.$i.gtf -G rna-seq/chr19.gtf; done
```

Получим новую аннотацию.

```
> ls *gtf > *string.gtf.list
> stringtie-2.1.4.Linux_x86_64/stringtie --merge *string.gtf.list -G rna-seq/chr19.gtf -o new_annot_chr19.gtf
```

3) Перекартируйте риды используя новую аннотацию.

Получим новые сайты сплайсинга.

```
> hisat2-2.1.0/./hisat2_extract_splice_sites.py rna-seq/stringtie_sort/new_annot_chr19.gtf > rna-seq/new_splicing_sites.txt
```

Запустим картирование в цикле.

```
> for i in `ls -1 /mnt/local/vse2020/shared/rna.seq.data/`; do hisat2-2.1.0/./hisat2 --no-softclip -x rna-seq/index_chr19 --known-splicesite-infile rna-seq/new_splicing_sites.txt -p 2 -U /mnt/local/vse2020/shared/rna.seq.data/$i | samtools view -bS - > rna-seq/new_bam/$i.bam; done
```

Отсортируем.

```
> for i in `ls -1 rna-seq/new_bam `; do samtools sort rna-seq/new_bam/$i -o rna-seq/new_bam/new.sort.$i; done
```

Запустим Stringtie на новых файлах.

```
> for i in `ls -1 rna-seq/new_bam`; do stringtie-2.1.4.Linux_x86_64/stringtie rna-seq/new_bam/$i -o rna-seq/new_gtf/$i.gtf -G rna-seq/chr19.gtf -e; done
```

4) Оцените экспрессию генов в каждом образце при помощи stringtie, получите таблицу read counts при помощи prepDE.py

```
> ls rna-seq/new_gtf/*gtf > rna-seq/new_gtf/*new.gtf.list
> stringtie-2.1.4.Linux_x86_64/prepDE.py -i rna-seq/new_gtf/*new.gtf.list -l 101 -t rna-seq/trans_counts.csv -g  rna-seq/gene_counts.csv
```

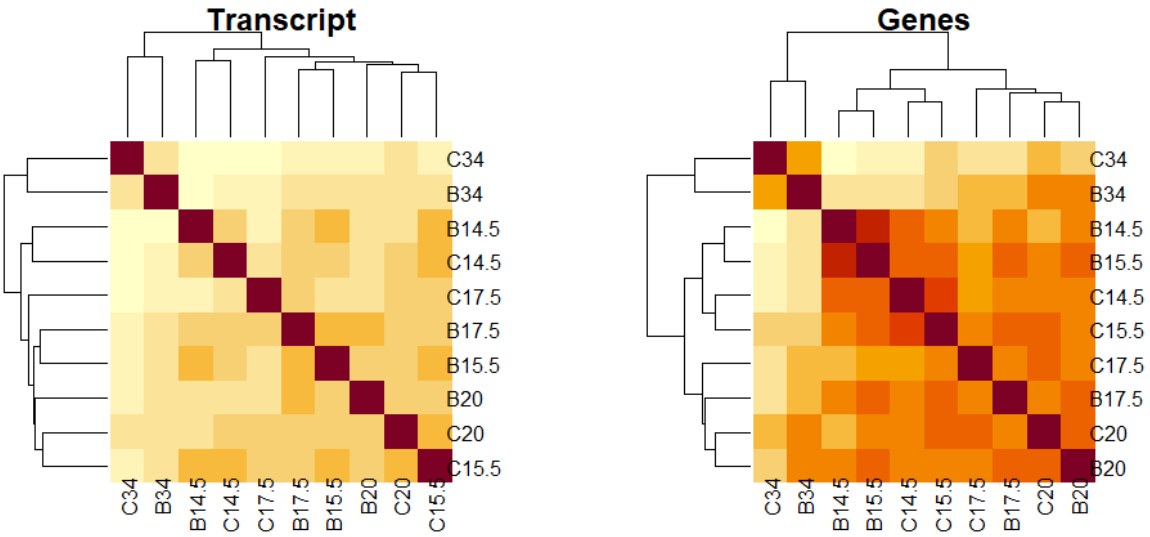
5) Постройте PCA и Heatmaps для образцов.

Далее расчёты делались в R (*hw_Rastvorova_2_3.R*).

Результаты картирования с помощью hisat2. Картирование по аннотации stringtie немного хуже.

	no stringtie	stringtie
B14.5	96.18%	92.13%
B15.5	93.03%	88.16%
B17.5	93.21%	87.45%
B20	95.33%	91.45%
B34	94.96%	91.33%
C14.5	94.73%	91.68%
C15.5	92.59%	87.29%
C17.5	94.98%	90.13%
C20	95.98%	92.51%
C34	94.32%	92.75%

Heatmaps для экспрессии по транскриптам и генам.



Результаты PCA.

```
> #PCA
> summary(prcomp(transcript, center = TRUE, scale = TRUE))
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10
Standard deviation     2.9365 0.74326 0.58665 0.42215 0.34315 0.27105 0.21812 0.17162 0.13911 0.12104
Proportion of Variance 0.8623 0.05524 0.03442 0.01782 0.01177 0.00735 0.00476 0.00295 0.00194 0.00147
Cumulative Proportion  0.8623 0.91754 0.95195 0.96978 0.98155 0.98890 0.99365 0.99660 0.99853 1.00000
> summary(prcomp(genes, center = TRUE, scale = TRUE))
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10
Standard deviation     3.0061 0.72088 0.42750 0.33899 0.24678 0.20990 0.14205 0.10784 0.07888 0.05487
Proportion of Variance 0.9037 0.05197 0.01828 0.01149 0.00609 0.00441 0.00202 0.00116 0.00062 0.00030
Cumulative Proportion  0.9037 0.95563 0.97391 0.98540 0.99149 0.99590 0.99791 0.99908 0.99970 1.00000
```