RiboSeq

Лев Мазаев | мАДБМ18

Считал все на локальной машине.

Этап 1: скачивание и распаковка файлов образцов

Первая задача - скачать SRA-файлы. Я решил выбрать образцы с нокдауном метилтрансферазы METTL3. На GEO-странице эксперимента внизу можно перейти в SRA Run Selector:

Platforms (1)	GPL9185 Illumina Genome Analyzer (Mus musculus)
Samples (14)	GSM2717358 Ribo_seq_Scram_rep1
Less...	GSM2717359 Ribo_seq_Scram_rep2
	GSM2717360 Ribo_seq_shMETTL3_rep1
	GSM2717361 Ribo_seq_shMETTL3_rep2
	GSM2717362 Ribo_seq_shABCF1_rep1
	GSM2717363 Ribo_seq_shABCF1_rep2
	GSM2717364 Ribo_seq_Torin
	GSM2717365 RNA_seq_Scram_rep1
	GSM2717366 RNA_seq_Scram_rep2
	GSM2717367 RNA_seq_shMETTL3_rep1
	GSM2717368 RNA_seq_shMETTL3_rep2
	GSM2717369 RNA_seq_shABCF1_rep1
	GSM2717370 RNA_seq_shABCF1_rep2
	GSM2717371 RNA_seq_Torin

Relations

BioProject	PRJNA395723
SRA	SRP113540

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

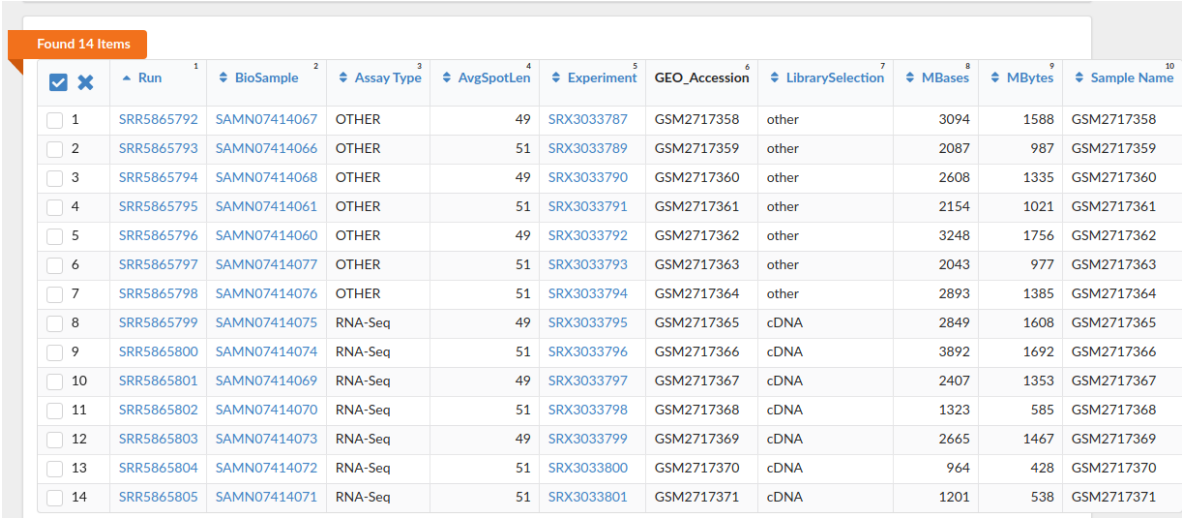
Supplementary file	Size	Download	File type/resource
GSE101865_RAW.tar	2.1 Mb	(http)(custom)	TAR (of TXT)

SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file

Там мы видим все образцы (Samples) и их соответствующие SRR-записи:



Found 14 Items

	Run	BioSample	Assay Type	AvgSpotLen	Experiment	GEO_Accession	LibrarySelection	MBases	MBytes	Sample Name
1	SRR5865792	SAMN07414067	OTHER	49	SRX3033787	GSM2717358	other	3094	1588	GSM2717358
2	SRR5865793	SAMN07414066	OTHER	51	SRX3033789	GSM2717359	other	2087	987	GSM2717359
3	SRR5865794	SAMN07414068	OTHER	49	SRX3033790	GSM2717360	other	2608	1335	GSM2717360
4	SRR5865795	SAMN07414061	OTHER	51	SRX3033791	GSM2717361	other	2154	1021	GSM2717361
5	SRR5865796	SAMN07414060	OTHER	49	SRX3033792	GSM2717362	other	3248	1756	GSM2717362
6	SRR5865797	SAMN07414077	OTHER	51	SRX3033793	GSM2717363	other	2043	977	GSM2717363
7	SRR5865798	SAMN07414076	OTHER	51	SRX3033794	GSM2717364	other	2893	1385	GSM2717364
8	SRR5865799	SAMN07414075	RNA-Seq	49	SRX3033795	GSM2717365	cDNA	2849	1608	GSM2717365
9	SRR5865800	SAMN07414074	RNA-Seq	51	SRX3033796	GSM2717366	cDNA	3892	1692	GSM2717366
10	SRR5865801	SAMN07414069	RNA-Seq	49	SRX3033797	GSM2717367	cDNA	2407	1353	GSM2717367
11	SRR5865802	SAMN07414070	RNA-Seq	51	SRX3033798	GSM2717368	cDNA	1323	585	GSM2717368
12	SRR5865803	SAMN07414073	RNA-Seq	49	SRX3033799	GSM2717369	cDNA	2665	1467	GSM2717369
13	SRR5865804	SAMN07414072	RNA-Seq	51	SRX3033800	GSM2717370	cDNA	964	428	GSM2717370
14	SRR5865805	SAMN07414071	RNA-Seq	51	SRX3033801	GSM2717371	cDNA	1201	538	GSM2717371

Скачиваем SRR586579(4|5) (GSM271736(0|1) - Ribo_seq_shMETTL3_rep(1|2)) и SRR586580(1|2) (GSM271736(7|8) - RNA_seq_shMETTL3_rep(1|2)):

```
mkdir ~/BioData/RiboSeq
cd ~/BioData/Riboseq
prefetch -v SRR5865794
prefetch -v SRR5865795
prefetch -v SRR5865801
prefetch -v SRR5865802
```

Далее, в процессе скачивания посмотрим, что делают команды `esearch` и `efetch`:

```
(riboseq) leo@MS7922:~/ME/HSE/NGS/5$ esearch -db sra -query PRJNA395723 | efetch -format runinfo | head -n 2
Run,ReleaseDate,LoadDate,spots,bases,spots_with_mates,avgLength,size_MB,AssemblyName,download_path,Experiment,LibraryName,LibraryStrategy,LibrarySelection,LibrarySource,LibraryLayout,InsertSize,InsertDev,Platform,Model,SRAStudy,BioProject,Study_Pubmed_id,ProjectID,Sample,BioSample,SampleType,TaxID,ScientificName,SampleName,g1k_pop_code,source,g1k_analysis_group,Subject_ID,Sex,Disease,Tumor,Affection_Status,Analyte_Type,Histological_Type,Body_Site,CenterName,Submission,dbgap_study_accession,Consent,RunHash,ReadHash
SRR5865792,2017-07-27 15:39:06,2017-07-25 14:31:52,63161402,3094908698,0,49,1588,,https://sra-download.ncbi.nlm.nih.gov/traces/sra50/SRR/005728/SRR5865792,SRX3033787,,OTHER,other,TRANSCRIPTOMIC,SINGLE,0,0,ILLUMINA,Illumina Genome Analyzer,SRP113540,PRJNA395723,2,395723,SRS2382513,SAMN07414067,simple,10090,Mus musculus,GSM2717358,,,,,,,no,,,,GEO,SRA591747,,public,4188AEEB23A891933E4AAAE909E4E530,C71561E9AD98C85508957F5D62AED15F
(riboseq) leo@MS7922:~/ME/HSE/NGS/5$ esearch -db gds -query "GSE101865[ACCN]" | efetch -format docsum -mode json | head -n 10
{
    "header": {
        "type": "esummary",
        "version": "0.3"
    },
    "result": {
        "uids": [
            "200101865",
            "100009185",
            "302717371",
```

Из них при помощи ruby-скрипта можно получить табличку для выяснения номеров SRA, но не стал этого делать, так как всё есть в SRA Run Selector.

Распакуем скачанные SRA-архивы (предварительно собрав их в одной папке) с помощью `parallel-fastq-dump` (работает гораздо быстрее, чем обычный `fastq-dump`):

```
parallel-fastq-dump -s *.sra -t 8 --split-files
```

Сравнение размеров файлов:

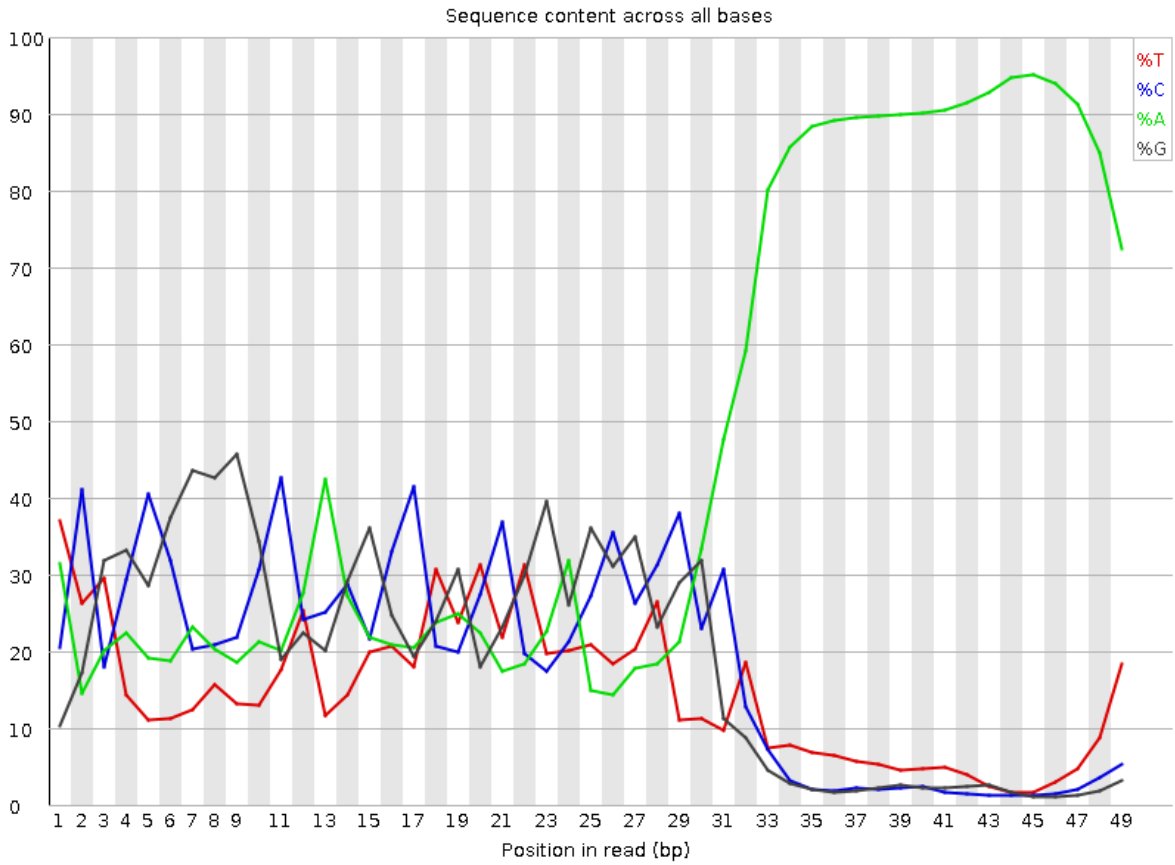
```
(riboseq) leo@MS7922:~/BioData/RiboSeq$ du -ah
12G     ./SRR5865794_1.fastq
586M    ./SRR5865802.sra
12G     ./SRR5865801_1.fastq
9,7G    ./SRR5865795_1.fastq
1,4G    ./SRR5865801.sra
1022M   ./SRR5865795.sra
1,4G    ./SRR5865794.sra
6,0G    ./SRR5865802_1.fastq
43G     .
```

Этап 2: отрезка адаптеров и QC

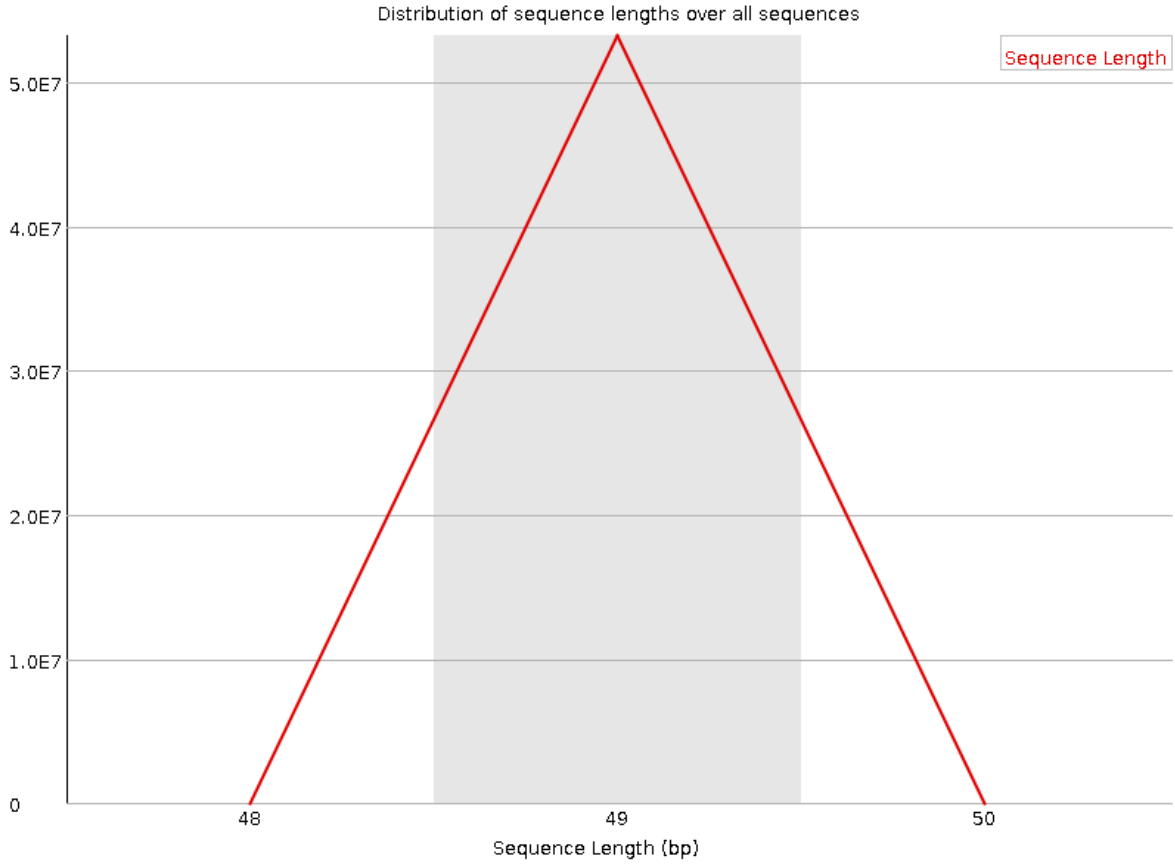
Первичный анализ последовательностей:

```
fastqc -t 8 ~/BioData/RiboSeq/*.fastq
```

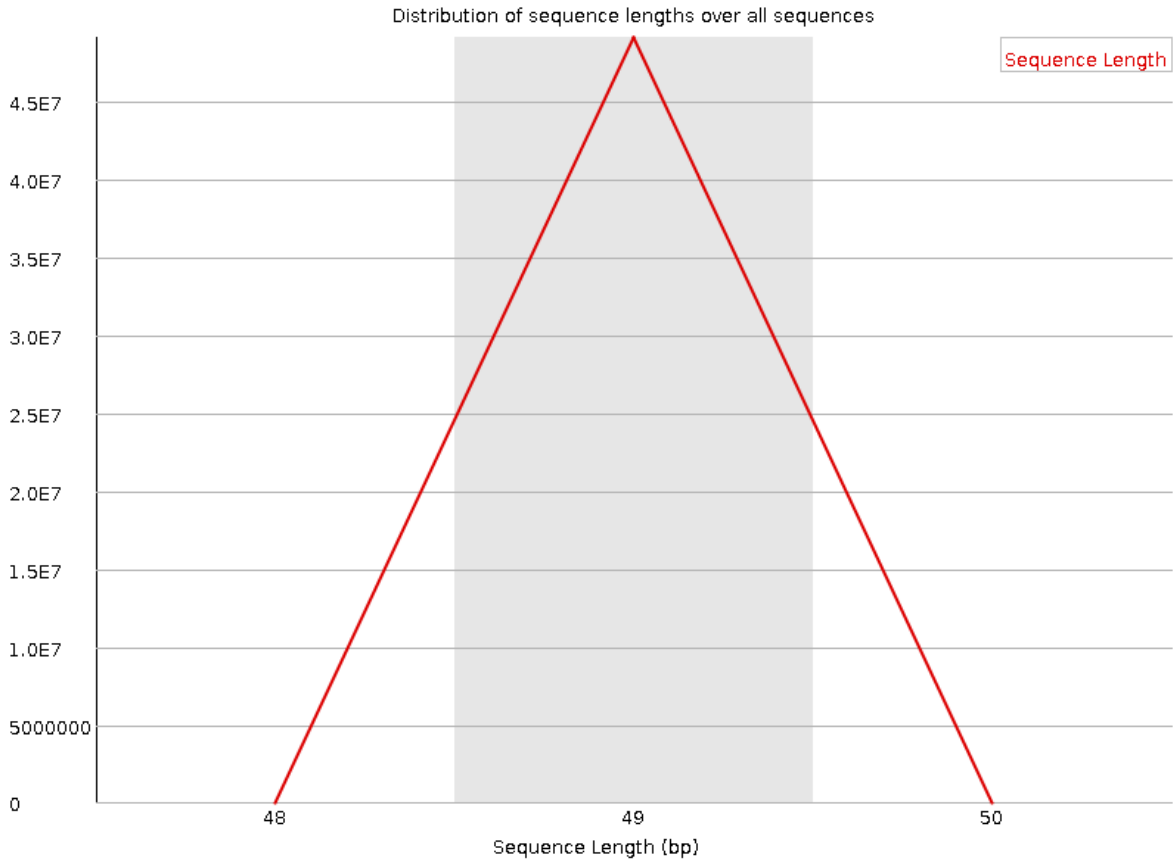
Длина рида в первых репликах - 49, во вторых - 51. Заметен полиА-адаптер в первых репликах (рис. для RiboSeq rep1):



Распределение длин прочтений (RiboSeq rep1):



RNASeq rep1:



Для вторых реплик тоже самое с пиком на 51. Теперь произведём отрезку адаптеров.

Data processing	For replicate 1, the 3' adaptor AAAAAAAAAA and low quality bases were trimmed by Cutadapt. The trimmed reads were mapped to Mouse transcriptom (GRCm38.83) by Bowtie, using parameters: -a --best -m1 --strata. Two mismatches were permitted. For replicate 2, the 3' adaptor CTGTAGGCACCATCAAT were trimmed by Cutadapt. The first 6 nucleotide and the last 4 nucleotides were also clipped after adaptor removal. The trimmed reads were mapped to Mouse transcriptom (GRCm38.83) by Bowtie, using parameters: -a --best -m1 --strata. Two mismatches were permitted. The PCR duplicates were removed from the mapped reads using the methods in Lecanda et al., 2016 (PMID: 27450428) Genome_build: GRCm38.83 Supplementary_files_format_and_content: RPKM files were generated. Each line represents one mRNAs. RPKM values were separated by Tabs.

Адаптер первой реплики - AAAAAAAAAA:

```
cutadapt -a AAAAAAAAAA -q 20 --minimum-length 20 -j 8 -o
riboseq1.trimmed.fastq.gz --trimmed-only SRR5865794_1.fastq
cutadapt -a AAAAAAAAAA -q 20 --minimum-length 20 -j 8 -o
rnaseq1.trimmed.fastq.gz SRR5865801_1.fastq
```

Адаптер второй реплики - CTGTAGGCACCATCAAT (minimum length 30, так как потом будем ещё отрезать баркоды):

```
cutadapt -a CTGTAGGCACCATCAAT -q 20 --minimum-length 30 -j 8 -o
riboseq2.trimmed.fastq.gz --trimmed-only SRR5865795_1.fastq
cutadapt -a CTGTAGGCACCATCAAT -q 20 --minimum-length 30 -j 8 -o
rnaseq2.trimmed.fastq.gz SRR5865802_1.fastq
```

Далее, нужно дедуплицировать вторые реплики:

```
seqkit rmdup -s -j 8 riboseq2.trimmed.fastq.gz | gzip -c >
riboseq2.dedup.fastq.gz
```

[INFO] 18088740 duplicated records removed. То есть 46% записей было удалено из исходных 38801308 в `riboseq2.trimmed.fastq.gz`.

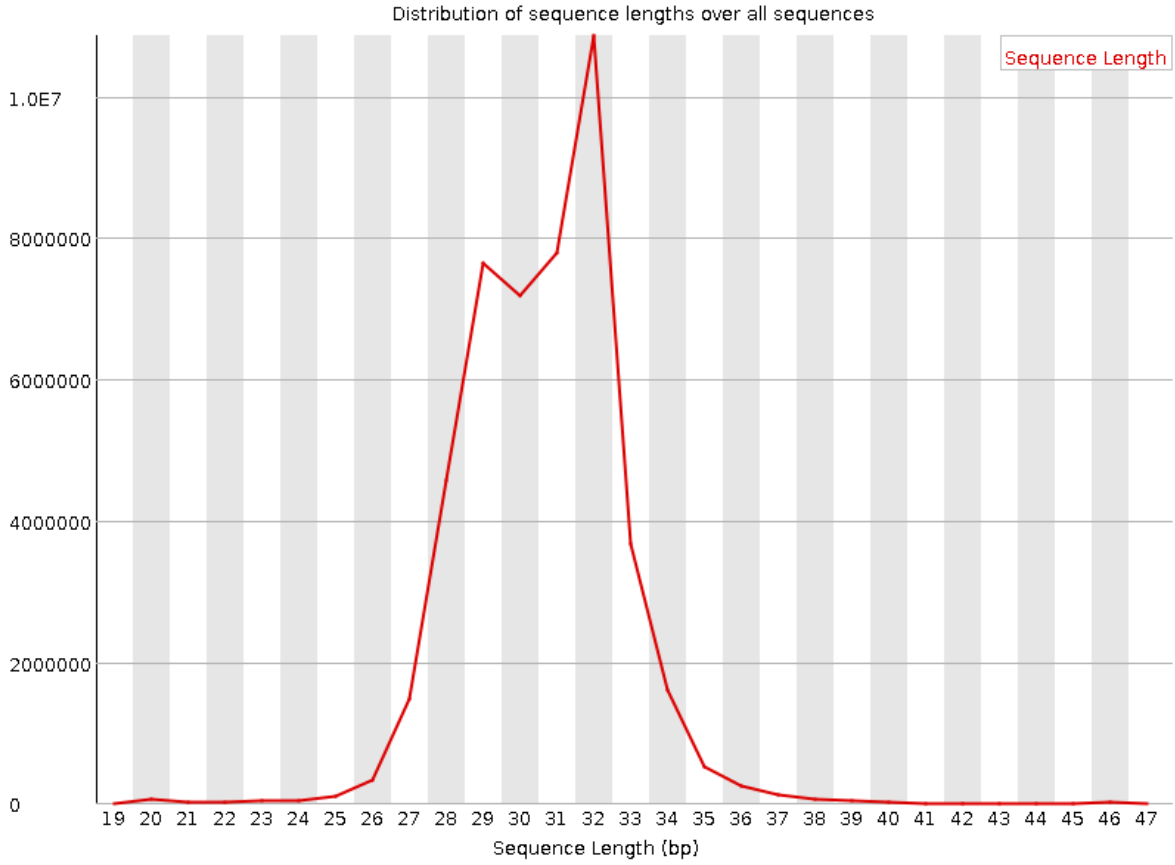
```
seqkit rmdup -s -j 8 rnaseq2.trimmed.fastq.gz | gzip -c >
rnaseq2.dedup.fastq.gz
```

[INFO] 9371952 duplicated records removed. То есть 36% записей было удалено из исходных 25769506 в `rnaseq2.trimmed.fastq.gz`

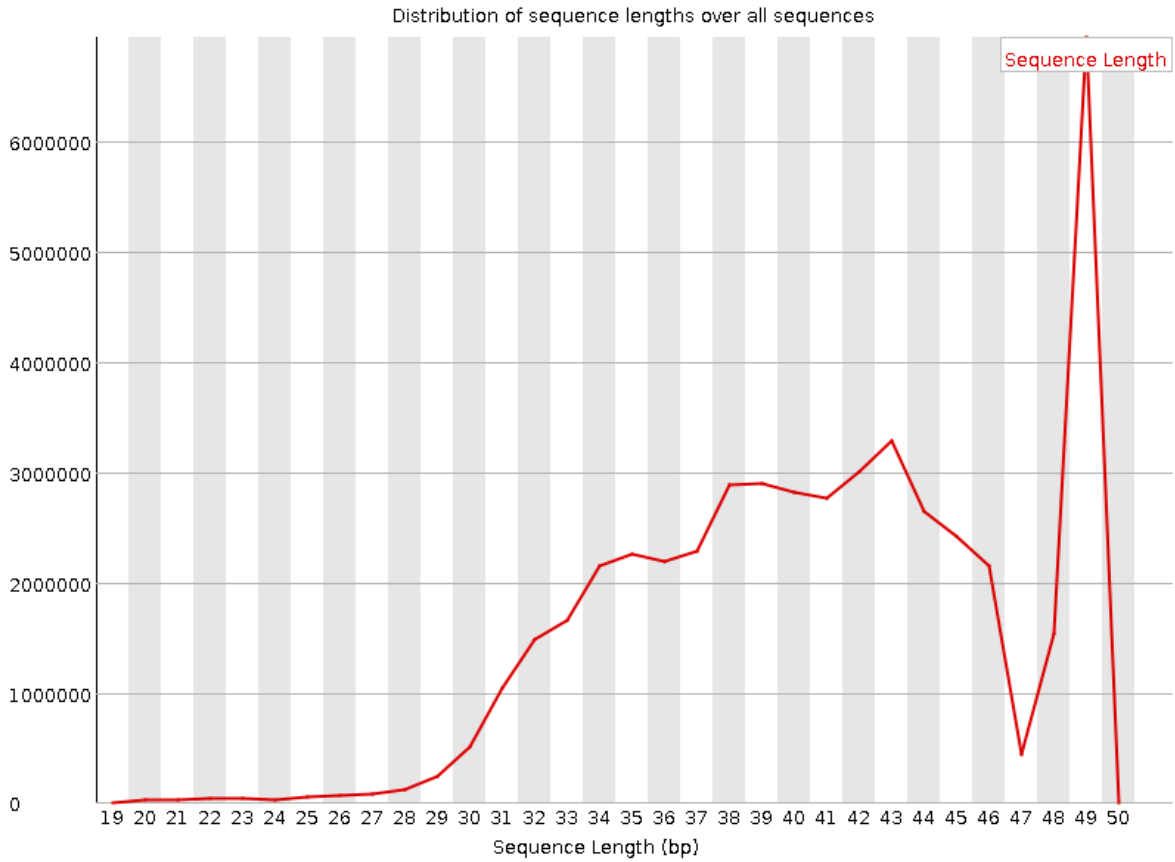
Теперь удалим баркоды:

```
cutadapt -u 6 -j 8 riboseq2.dedup.fastq.gz | cutadapt -u -4 -j 8 - -o
riboseq2.final.fastq.gz
cutadapt -u 6 -j 8 rnaseq2.dedup.fastq.gz | cutadapt -u -4 -j 8 - -o
rnaseq2.final.fastq.gz
```

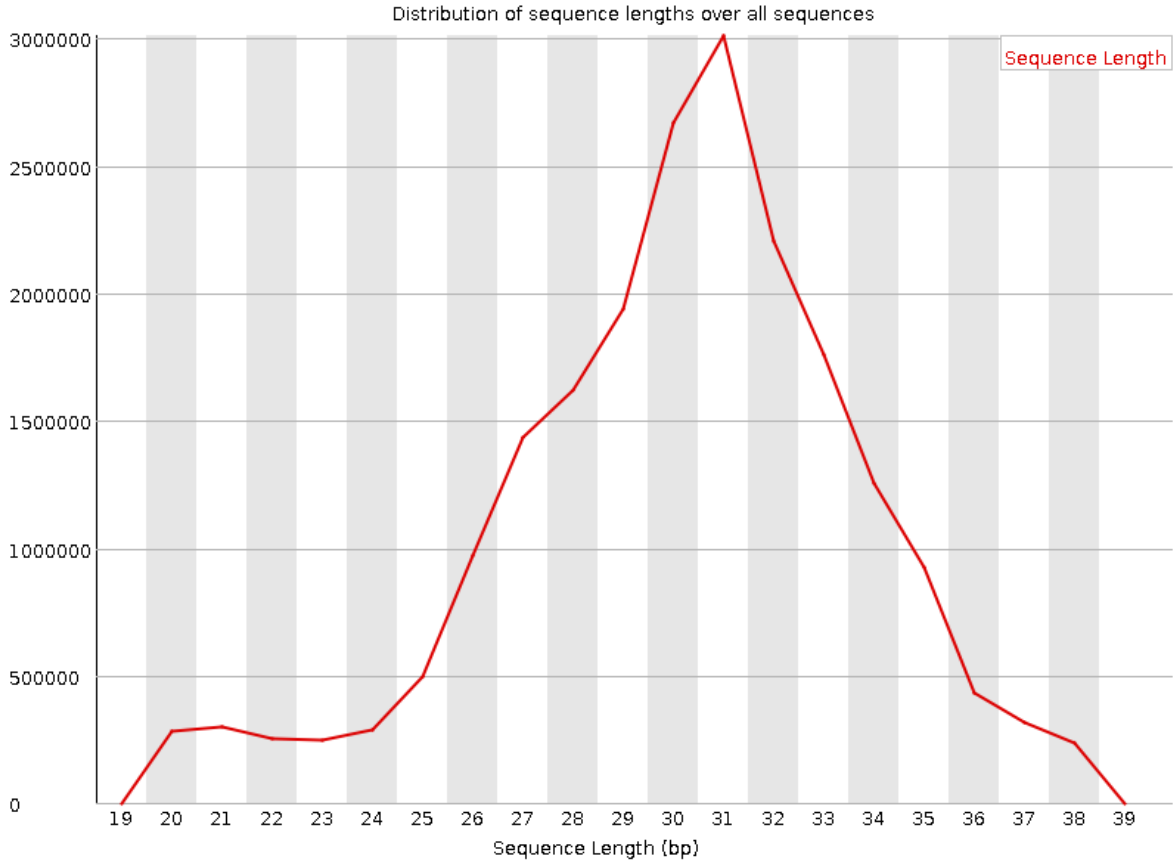
FastQC после всех операций. RiboSeq rep1:



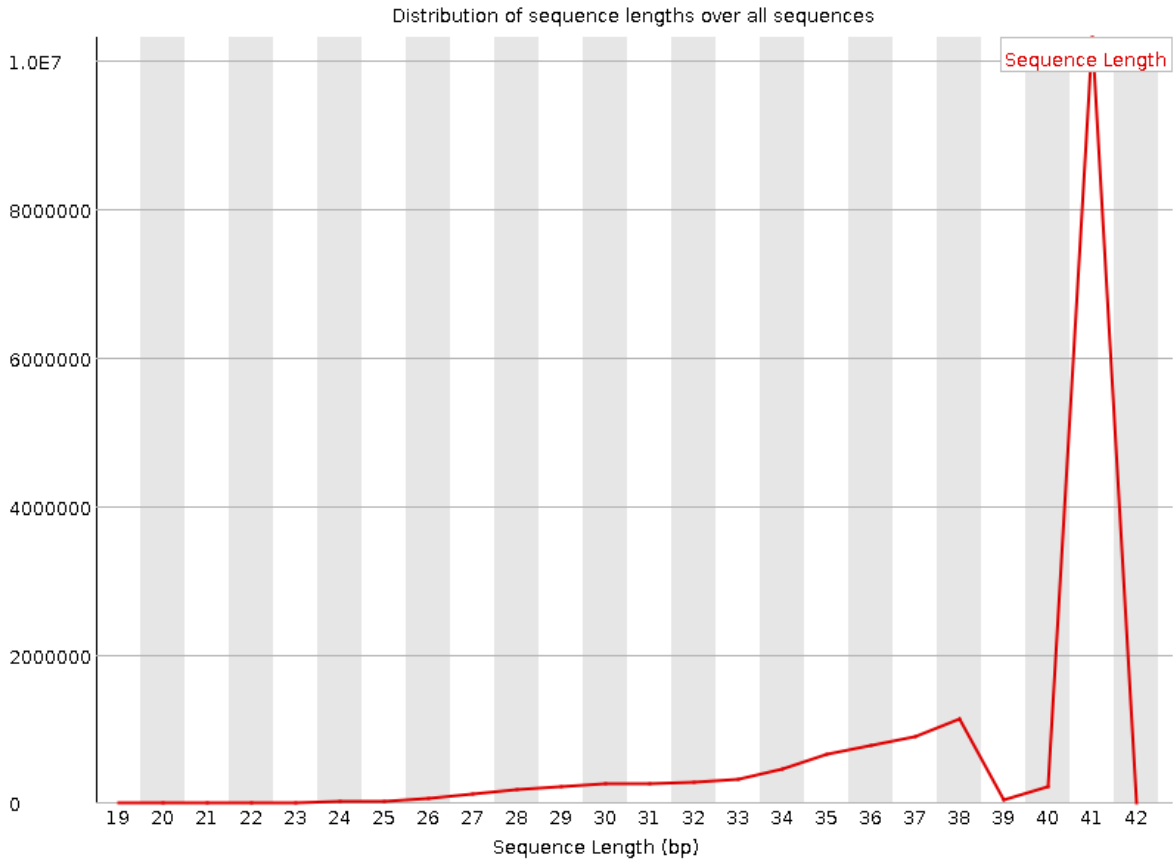
RNASeq rep1:



RiboSeq rep2:



RNASeq rep2:



Видим, что характерная длина рибосомных футпринтов - 25-32 пн. Из первой реплики RNASeq можно сделать вывод о том, что авторы старались сделать фрагменты РНК близкими по длине к рибосомным.

Этап 3: проверим долю рибосомной РНК

Построим индекс `bowtie` (скачал rRNA_euk.fasta с сервера):

```
bowtie-build --threads 8 rRNA_euk.fasta rRNA
```

Получим оценки доли рРНК для каждого из образцов, полученных выше:

```
(bio) leo@MS7922:~/BioData/RiboSeq$ bowtie --sam -p 8 ./bwt/rRNA riboseq1.trimmed.fastq.gz --chunkmbs 10000 > /dev/null
# reads processed: 46789965
# reads with at least one reported alignment: 23056356 (49.28%)
# reads that failed to align: 23733609 (50.72%)
Reported 23056356 alignments
(bio) leo@MS7922:~/BioData/RiboSeq$ bowtie --sam -p 8 ./bwt/rRNA rnaseq1.trimmed.fastq.gz --chunkmbs 10000 > /dev/null
# reads processed: 48273908
# reads with at least one reported alignment: 29740959 (61.61%)
# reads that failed to align: 18532949 (38.39%)
Reported 29740959 alignments
(bio) leo@MS7922:~/BioData/RiboSeq$ bowtie --sam -p 8 ./bwt/rRNA riboseq2.final.fastq.gz --chunkmbs 10000 > /dev/null
# reads processed: 20712568
# reads with at least one reported alignment: 12134426 (58.58%)
# reads that failed to align: 8578142 (41.42%)
Reported 12134426 alignments
(bio) leo@MS7922:~/BioData/RiboSeq$ bowtie --sam -p 8 ./bwt/rRNA rnaseq2.final.fastq.gz --chunkmbs 10000 > /dev/null
# reads processed: 16397554
# reads with at least one reported alignment: 10441188 (63.68%)
# reads that failed to align: 5956366 (36.32%)
Reported 10441188 alignments
```

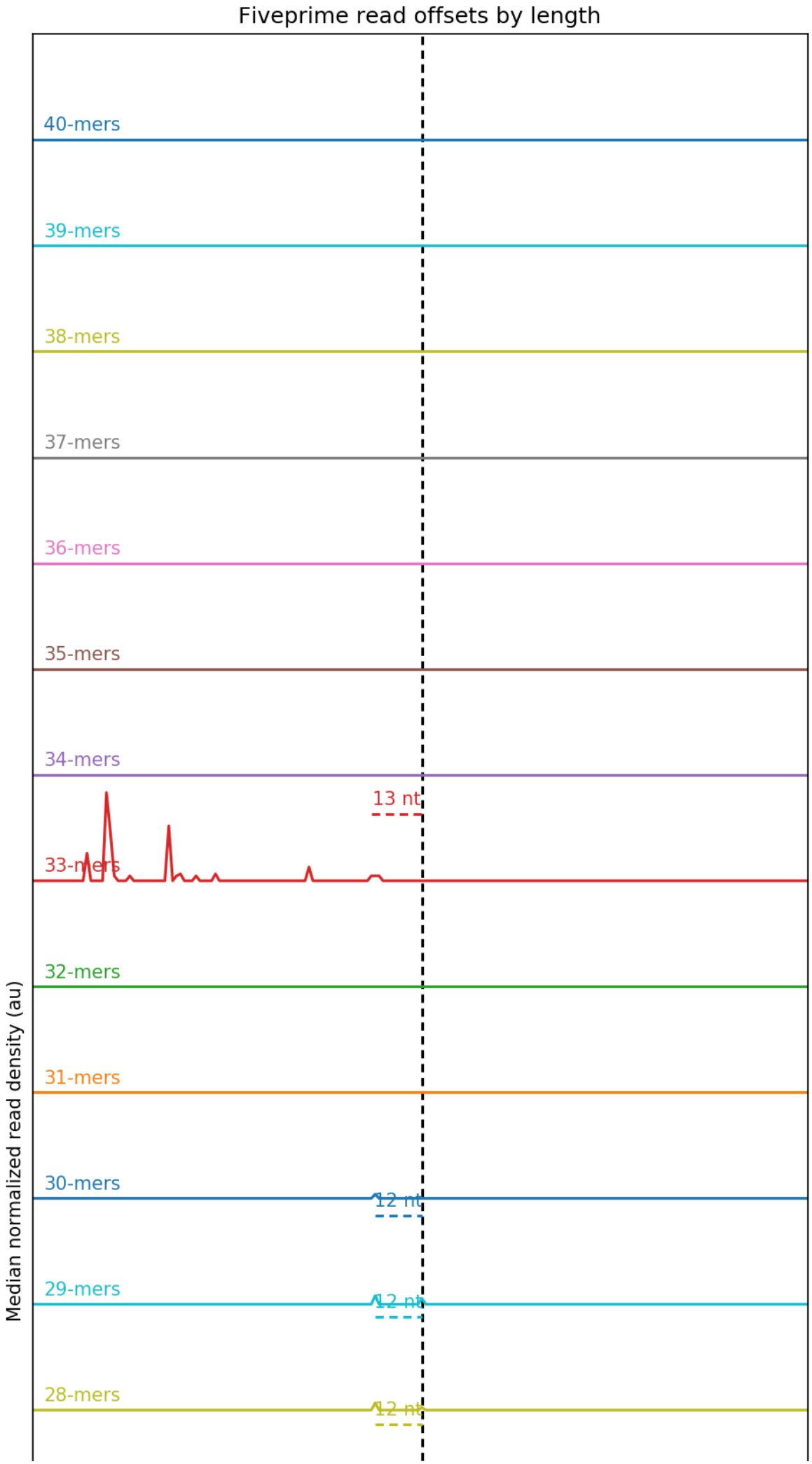
Этап 4: картирование на реальный геном

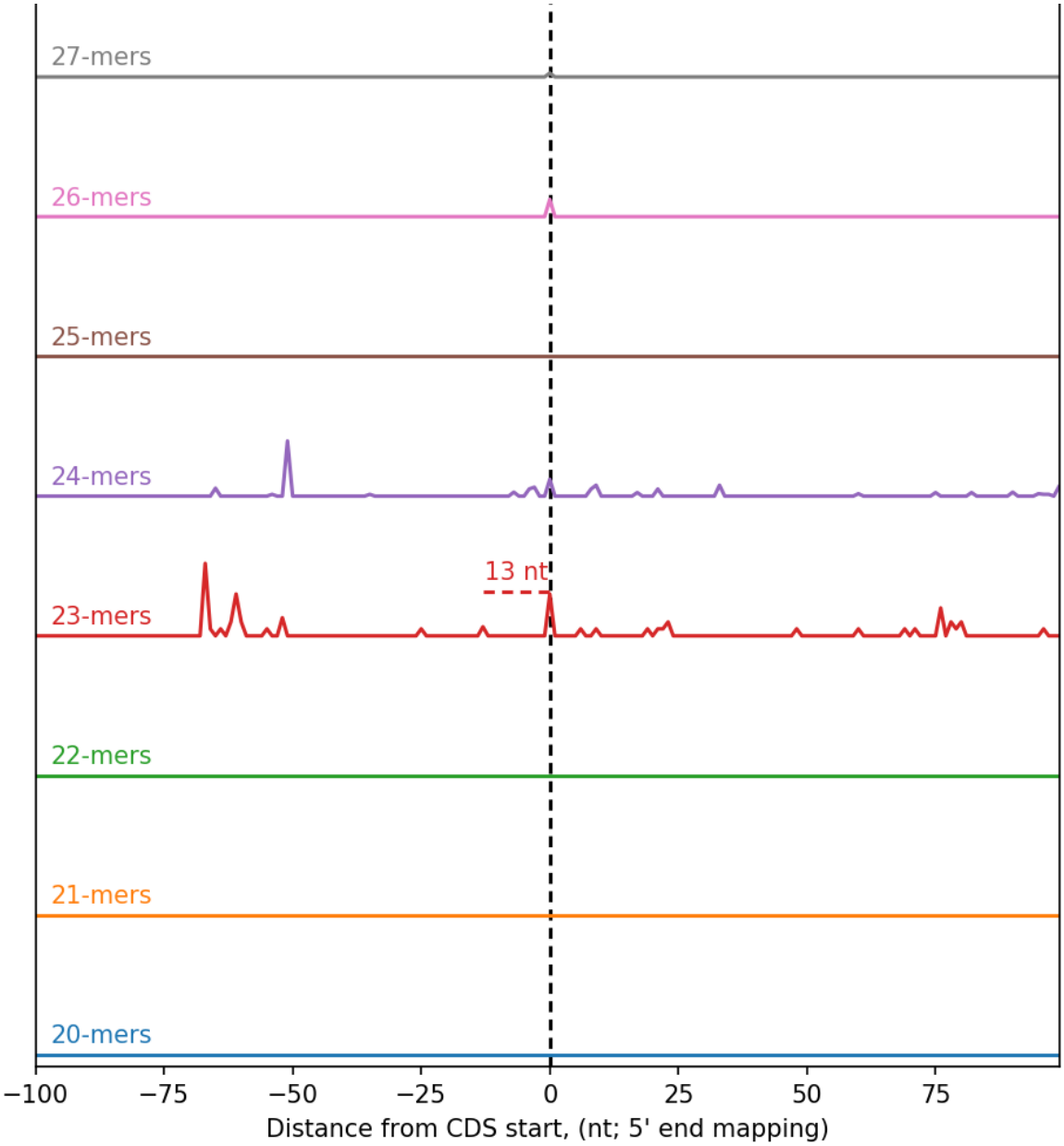
Использовал готовые файлы.

Этап 5: фазирование прочтений и метагенные профили

Фазирование по RiboSeq нокдауну METTL3 без aggregate:

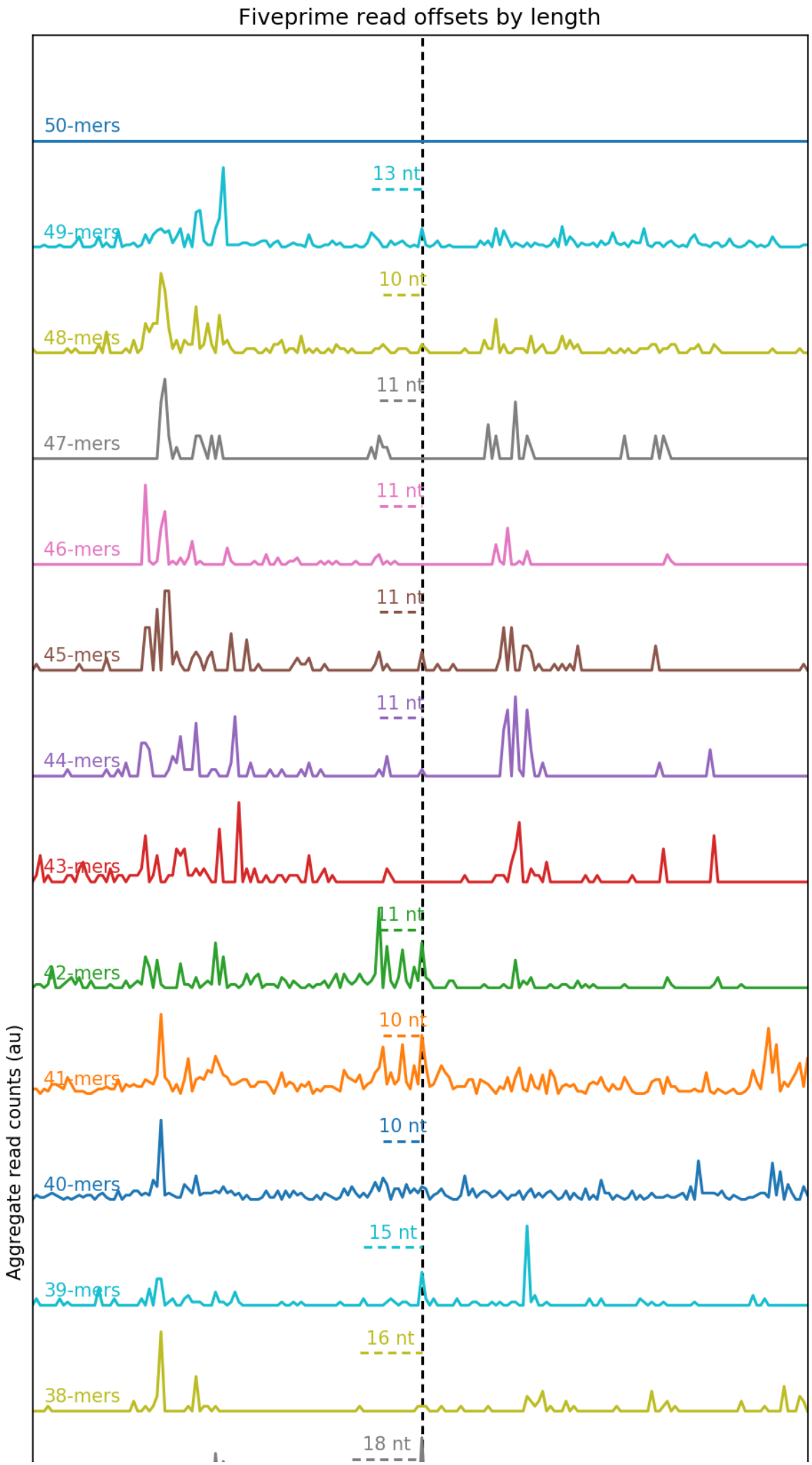
```
psite ~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt psite_test --
countfile_format BAM --count_files
~/BioData/RiboSeq/olbams/METTL3_ribo_Coots2017_m_r1.bam
~/BioData/RiboSeq/olbams/METTL3_ribo_Coots2017_m_r2.bam --min_length 20 --
max_length 40 --constrain 10 18 --min_count 10 --default 14
```

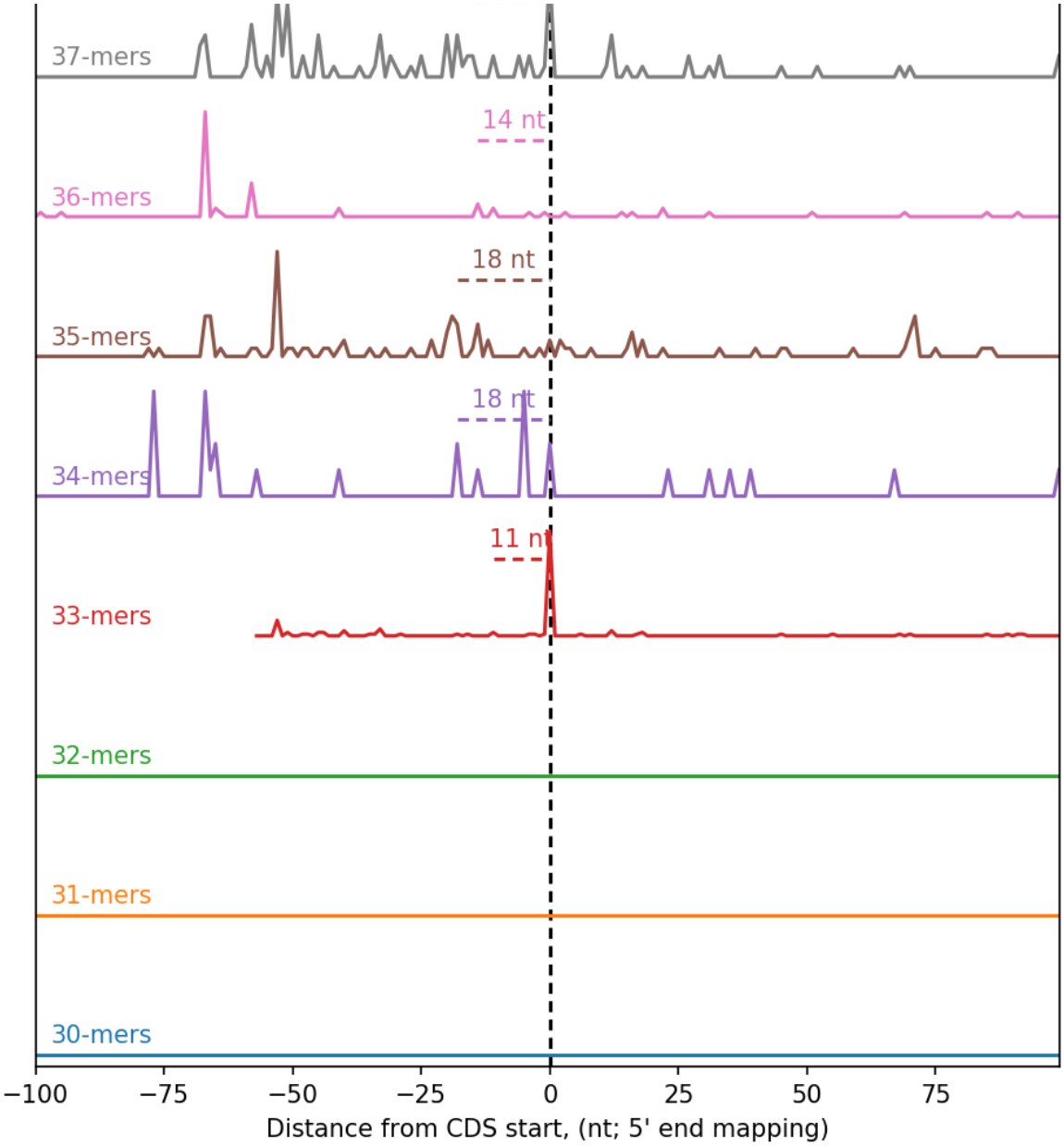

Fiveprime read offsets by length
40-mers
39-mers
38-mers
37-mers
36-mers
35-mers
34-mers
13 nt
33-mers
32-mers
31-mers
30-mers
12 nt
29-mers
12 nt
28-mers
12 nt
Median normalized read density (au)



Фазирование по RNASeq (от 30 до 50):

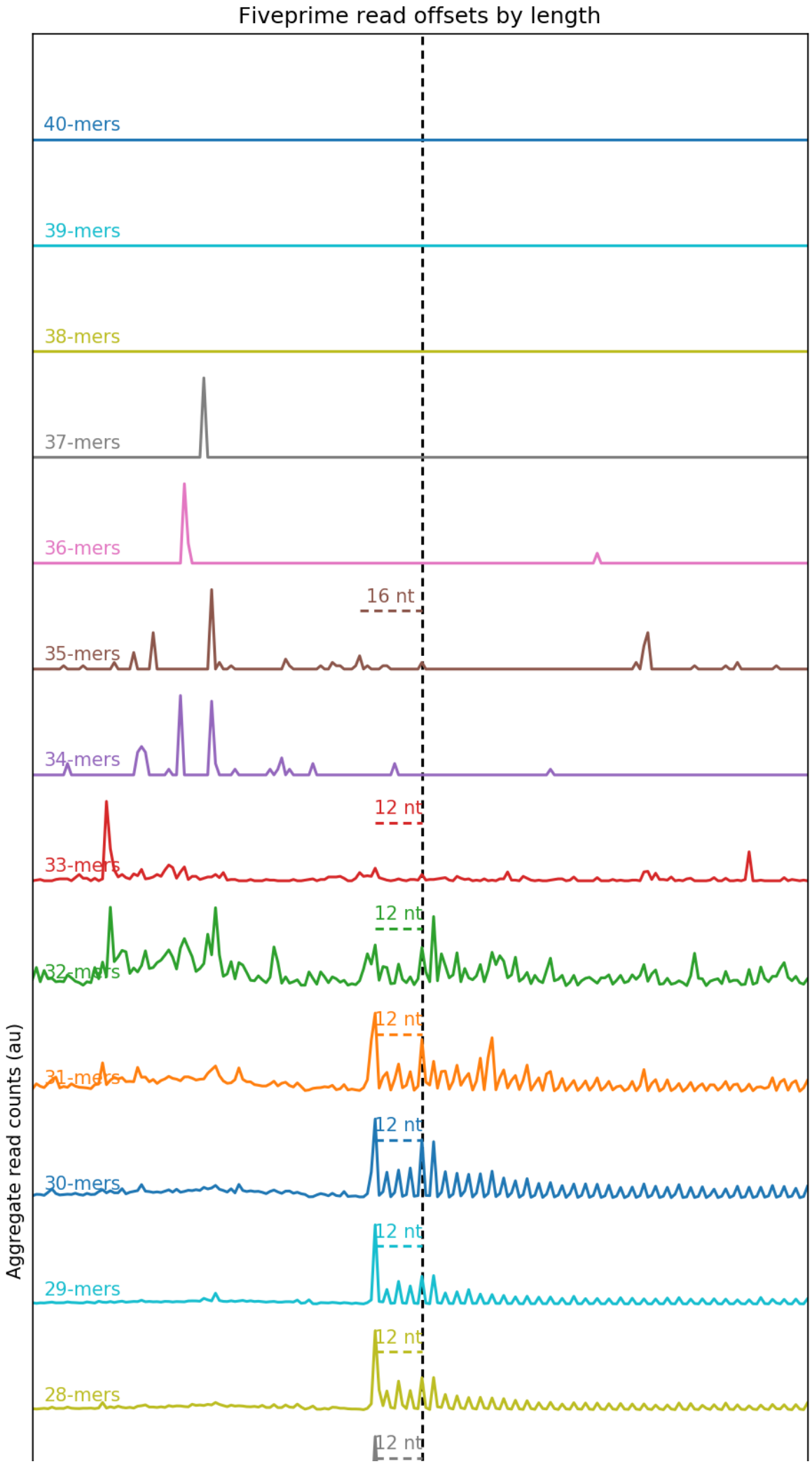
```
psite ~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt psite_test --
countfile_format BAM --count_files
~/BioData/RiboSeq/olbams/METTL3_rna_Coots2017_m_r1.bam
~/BioData/RiboSeq/olbams/METTL3_rna_Coots2017_m_r2.bam --min_length 30 --
max_length 50 --aggregate --constrain 10 18 --min_count 10 --default 14
```

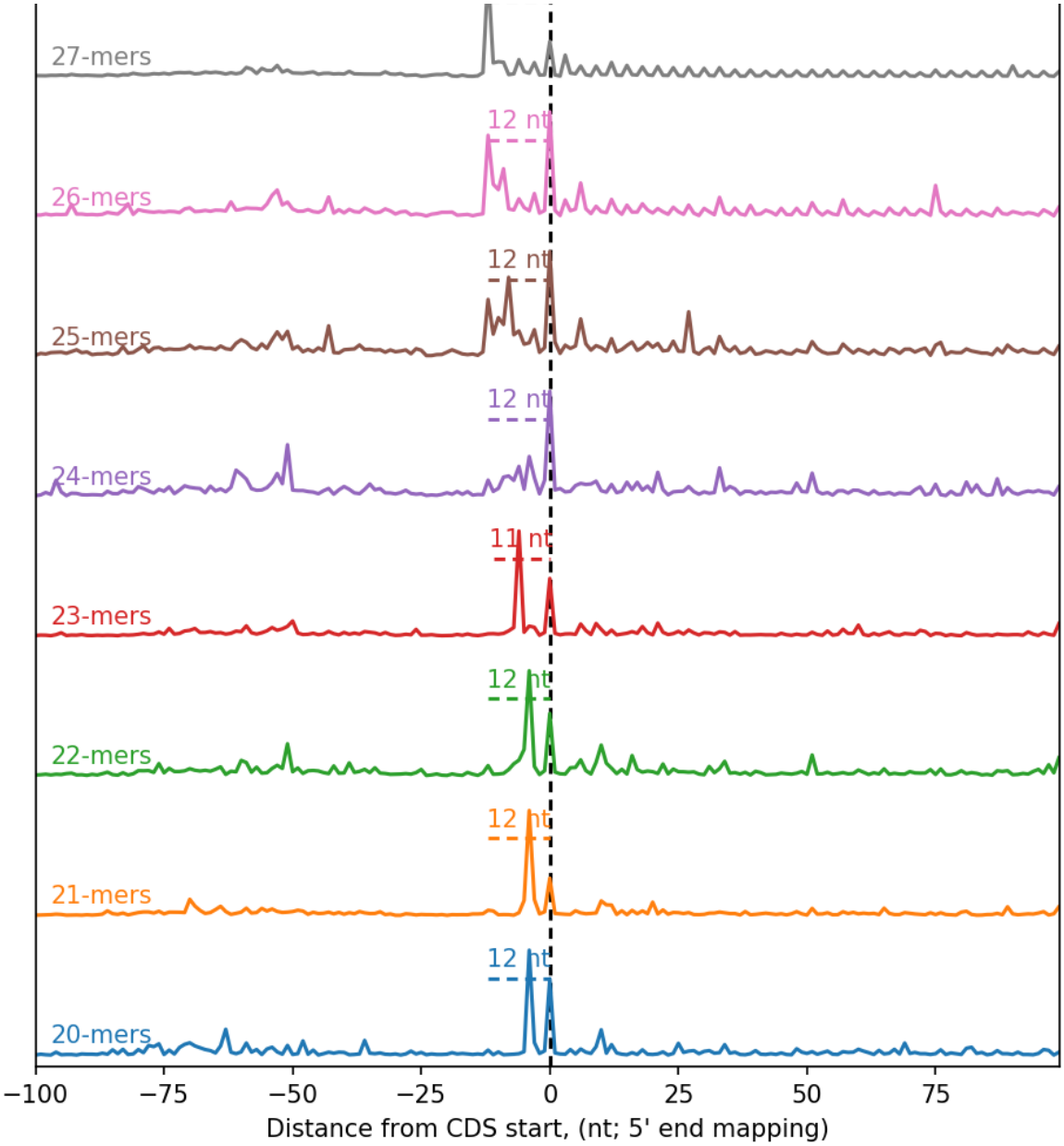

Fiveprime read offsets by length
Aggregate read counts (au)
50-mers
13 nt
49-mers
10 nt
48-mers
11 nt
47-mers
11 nt
46-mers
11 nt
45-mers
11 nt
44-mers
43-mers
11 nt
42-mers
10 nt
41-mers
10 nt
40-mers
15 nt
39-mers
16 nt
38-mers
18 nt



Фазирование по всем RiboSeq-образцам:

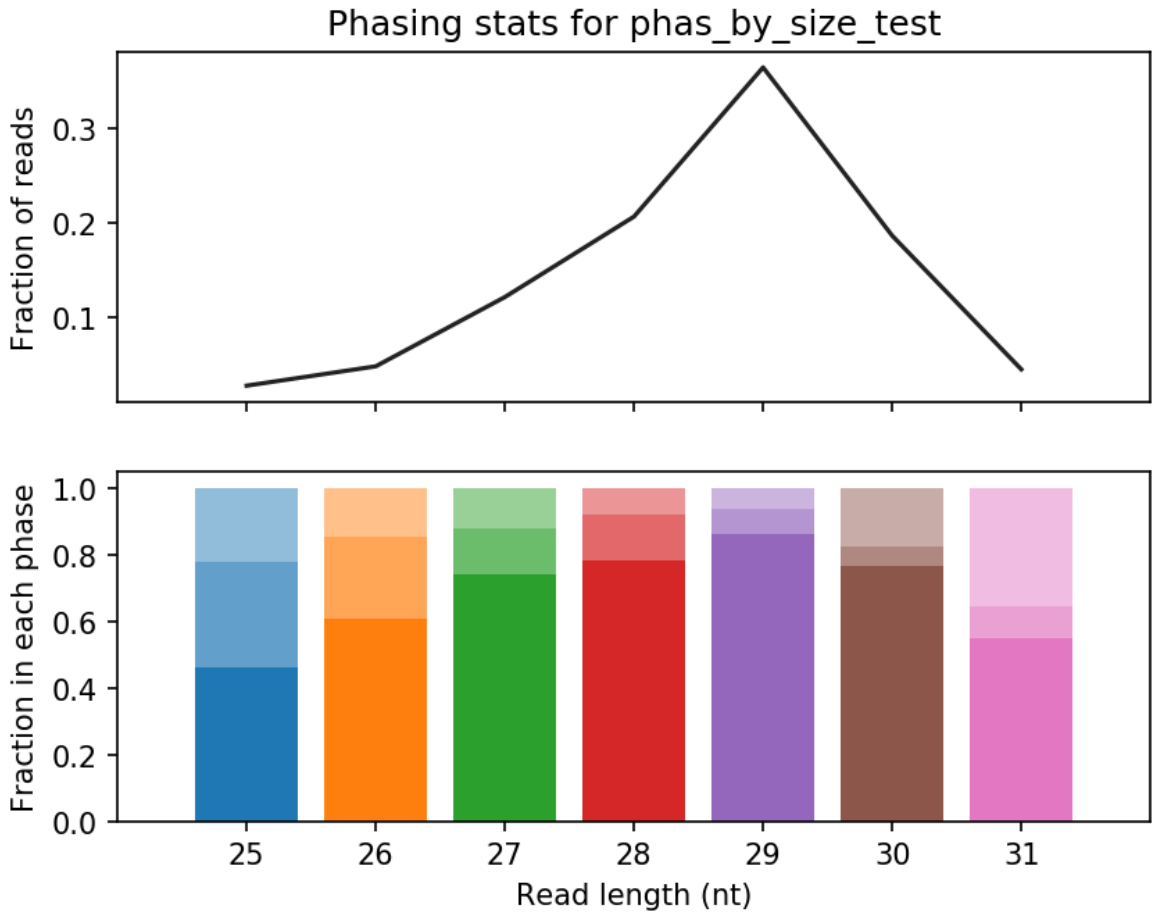
```
psite ~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt psite_test --
countfile_format BAM --countfiles $(find ~/BioData/RiboSeq/bams/ -name
'*ribo*.bam' | tr '\n' ' ') --min_length 20 --max_length 40 --aggregate --
constrain 10 18 --min_count 10 --default 14
```


Fiveprime read offsets by length
40-mers
39-mers
38-mers
37-mers
36-mers
16 nt
35-mers
34-mers
12 nt
33-mers
12 nt
32-mers
12 nt
31-mers
12 nt
30-mers
12 nt
29-mers
12 nt
28-mers
12 nt
Aggregate read counts (au)



Проверка фазирования:

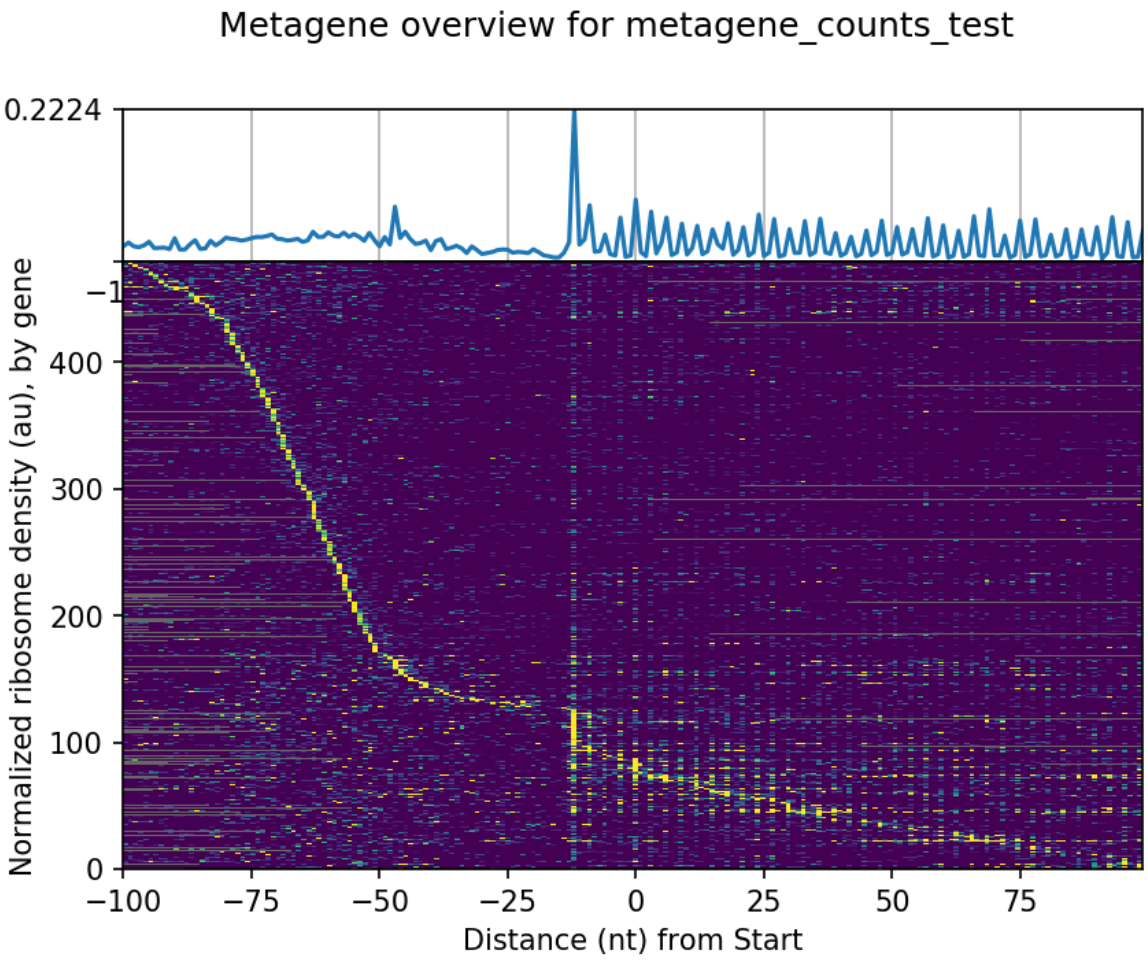
```
phase_by_size --countfile_format BAM --count_files $(find
~/BioData/RiboSeq/bams/ -name '*ribo*.bam' | tr '\n' ' ') --fiveprime_variable
--offset ./psite/psite_test_p_offsets.txt --min_length 25 --max_length 31
~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt phas_by_size_test
```



Получилось достаточно правильное, большая часть ридов ложится на frame0.

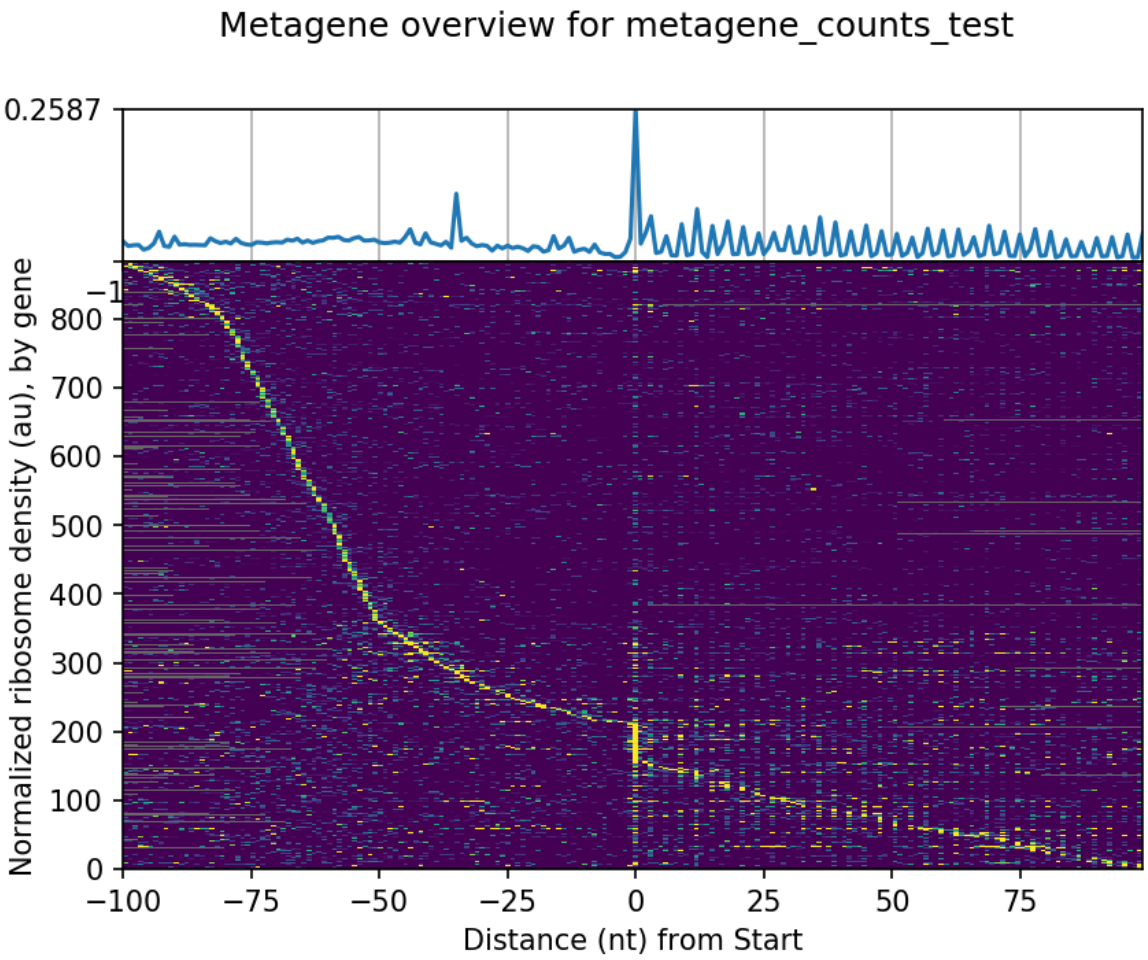
Нефазированный метагенный профиль METTL3 rep2 (длина 25-31):

```
metagene count --countfile_format BAM --count_files
~/BioData/RiboSeq/bams/METTL3_ribo_Coots2017_m_r2.bam --fiveprime --min_length
25 --max_length 31 --min_count 10 --use_mean --landmark Start
~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt metagene_counts_test
```



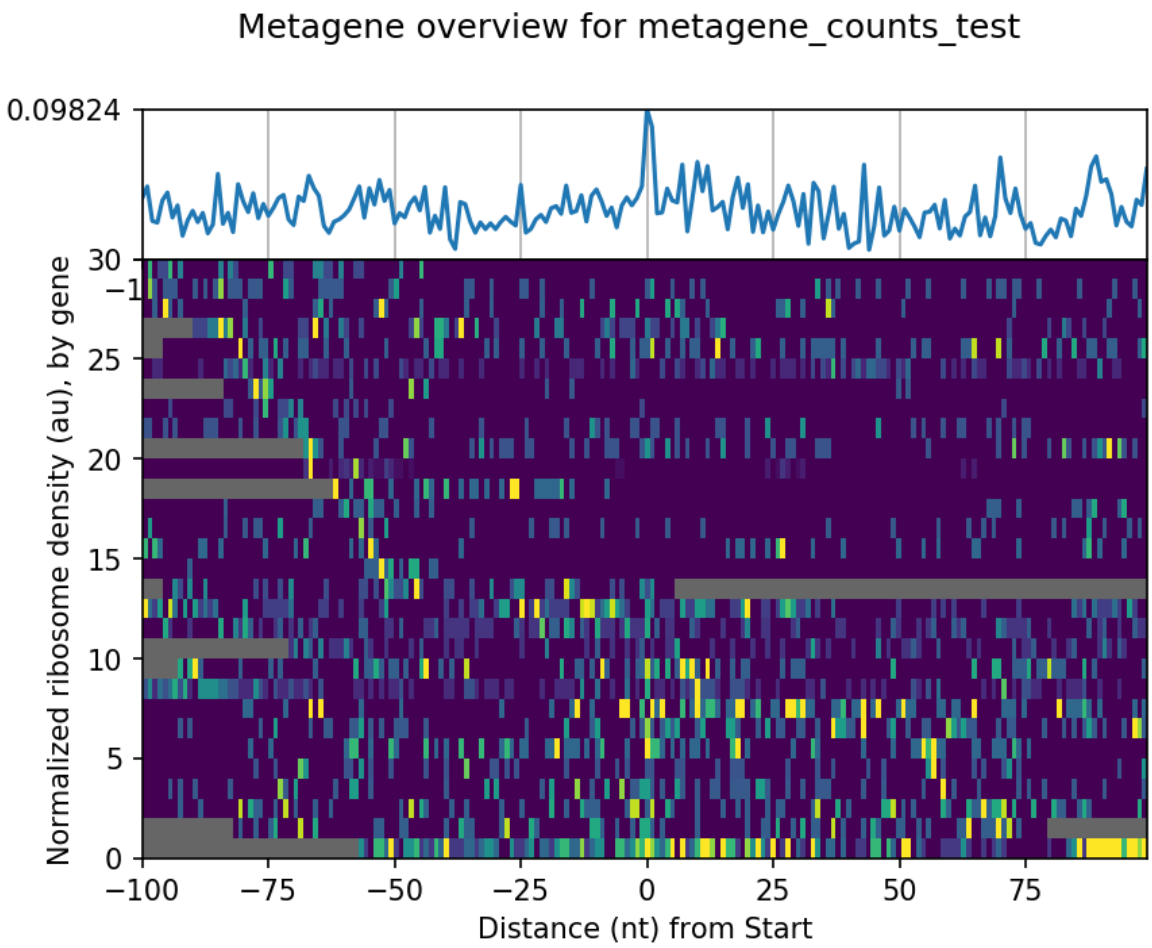
Фазированный метагенный профиль того же образца:

```
metagene count --countfile_format BAM --count_files
~/BioData/RiboSeq/bams/METTL3_ribo_Coots2017_m_r2.bam --fiveprime_variable --
offset ./psite/psite_test_p_offsets.txt --min_length 25 --max_length 31 --
min_count 5 --use_mean --landmark Start
~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt metagene_counts_test
```



Метагенный профиль RNASeq METTL3 rep2:

```
metagene count --countfile_format BAM --count_files
~/BioData/RiboSeq/bams/METTL3_rna_Coots2017_m_r2.bam --fiveprime --min_length
35 --max_length 45 --min_count 10 --use_mean --landmark Start
~/BioData/RiboSeq/plastidmetagene/mouse_start_rois.txt metagene_counts_test
```



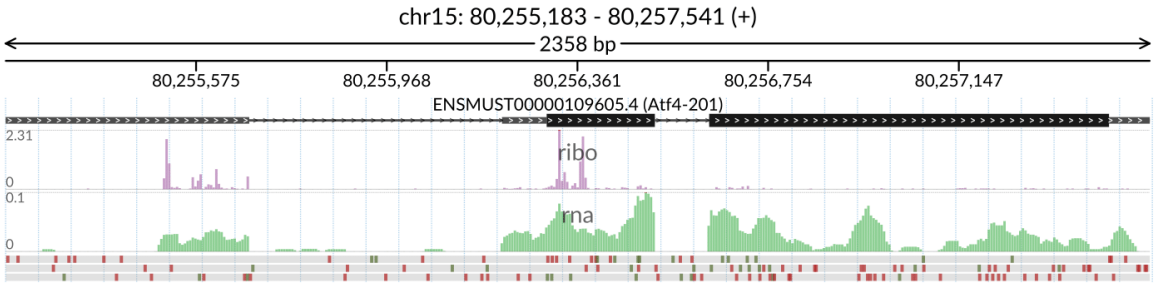
Этап 6: получение bedGraph и визуализация

Сделаем профили RiboSeq и RNASeq для METTL3 rep2:

```
make_wiggle -o METTL3_ribo2 --count_files
~/BioData/RiboSeq/bams/METTL3_ribo_Coots2017_m_r2.bam --normalize --min_length
25 --max_length 31 --fiveprime_variable --offset
../psite/psite_test_p_offsets.txt
make_wiggle -o METTL3_rna2 --count_files
~/BioData/RiboSeq/bams/METTL3_rna_Coots2017_m_r2.bam --normalize --center
bedtools unionbedg -i METTL3_ribo2_fw.wig METTL3_ribo2_rc.wig | csvtk mutate2
-H -t -L 5 -e '$4+$5' | cut -f 4-5 --complement > ribo.bedGraph
bedtools unionbedg -i METTL3_rna2_fw.wig METTL3_rna2_rc.wig | csvtk mutate2 -H
-t -L 5 -e '$4+$5' | cut -f 4-5 --complement > rna.bedGraph
```

Визуализируем ген *ATF4* с помощью `svist4get`:

```
svist4get -gtf ~/BioData/Annotations/gencode.vM23.basic.annotation.gtf -fa
~/BioData/GRCm38.primary_assembly.genome.fa -bg ribo.bedGraph rna.bedGraph -g
ENSMUSG00000042406.8
```



Как можно заметить, данный ген имеет преждевременную рамку считывания в 5'-UTR.