Скачивание референсной последовательности

1. Находим организм в описании эксперемента

   Organism: Escherichia coli O8:H36

SRX3519594: Whole genome Illumina MiSeq sequence of Escherichia coli serovar O8:H36

1 ILLUMINA (Illumina MiSeq) run: 822,970 spots, 299.9M bases, 185.4Mb downloads

External Id: EXT00290142

Design: MiSeq deep shotgun sequencing of cultured isolate.

Submitted by: FDA Center for Food Safety and Applied Nutrition (CFSAN)

Study: GenomeTrakr Project: US Food and Drug Administration

PRJNA230969 • SRP058582 • All experiments • All runs

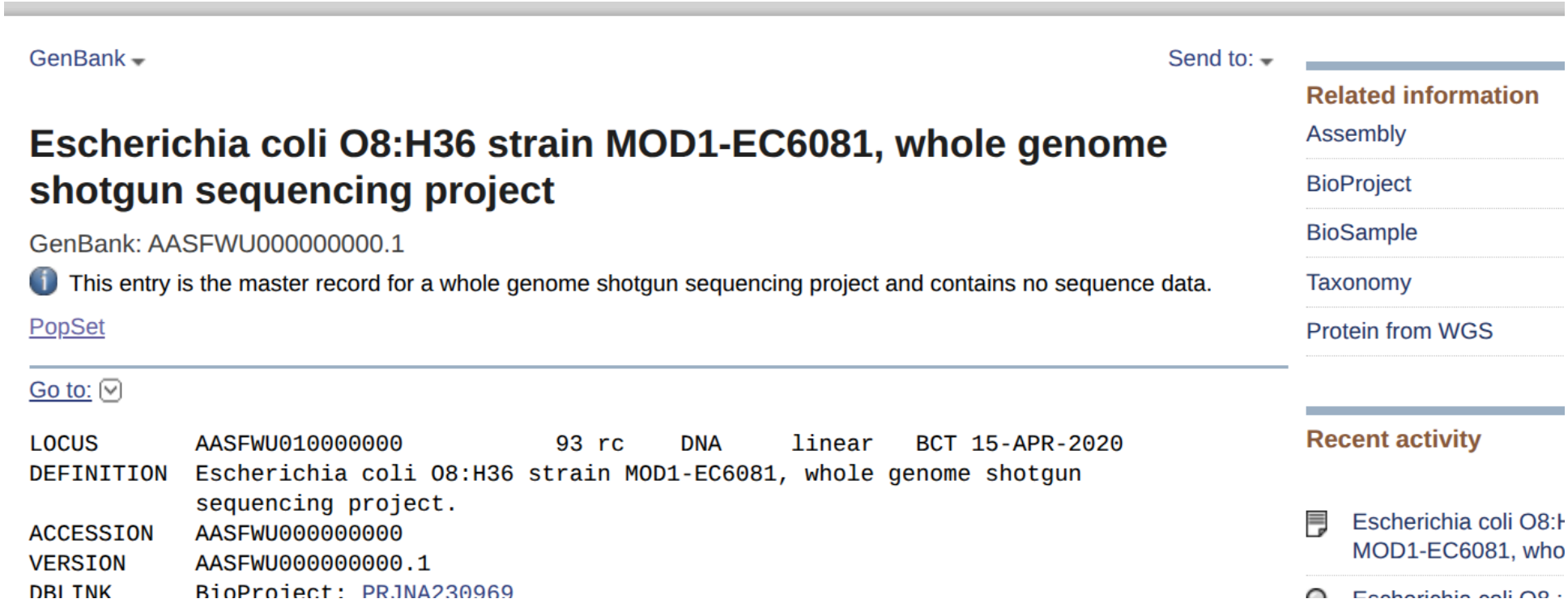
show Abstract

Sample:

SAMN08273716 • SRS2798506 • All experiments • All runs

Organism: Escherichia coli O8:H36

2. В базе нуклеотидов на NCBI находим организм



GenBank ▾ Send to: ▾

Escherichia coli O8:H36 strain MOD1-EC6081, whole genome shotgun sequencing project

GenBank: AASFWU000000000.1

This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.

PopSet

Go to:

```
LOCUS       AASFWU010000000           93 rc    DNA     linear   BCT 15-APR-2020
DEFINITION  Escherichia coli O8:H36 strain MOD1-EC6081, whole genome shotgun
            sequencing project.
ACCESSION   AASFWU000000000
VERSION     AASFWU000000000.1
DBLINK      BioProject: PRJNA230969
```

Related information

Assembly

BioProject

BioSample

Taxonomy

Protein from WGS

Recent activity

Escherichia coli O8:H
MOD1-EC6081, who

3. Переходим во вкладку `Assembly`

4. Нажимаем `Download` и скачиваем `fasta` и `gff` файлы

NCBI Datasets

Genome assembly PDT000275040.2

Download datasets API FTP

Submitted GenBank assembly	GCA_012306995.1
Taxon	Escherichia coli O8:H36
Strain	MOD1-EC6081
WGS project	AASFWU01
Submitter	FDA/CFSAN
Date	Apr 15, 2020

Genome notes

Индексация референсной последовательности

```
bowtie2-build GCA_012306995.1_PDT000275040.2_genomic.fna bowtie_index/yeast
```

Выравнивание чтений на геном

```
bowtie2 -x bowtie_index/yeast -1 SRR6427360/trimmed_fasta/SRR6427360_1.fastq -2 SRR6427360/trimmed_fasta/SRR6427360_2.fastq -S aligned/output.sam --fast -p 8 --time
```

Результат выполнения команды

```
Time loading reference: 00:00:00
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:00
Multiseed full-index search: 00:00:46
744032 reads; of these:
  744032 (100.00%) were paired; of these:
    519849 (69.87%) aligned concordantly 0 times
    218155 (29.32%) aligned concordantly exactly 1 time
    6028 (0.81%) aligned concordantly >1 times
    ----
    519849 pairs aligned concordantly 0 times; of these:
      482852 (92.88%) aligned discordantly 1 time
    ----
    36997 pairs aligned 0 times concordantly or discordantly; of these:
      73994 mates make up the pairs; of these:
        32262 (43.60%) aligned 0 times
        20133 (27.21%) aligned exactly 1 time
        21599 (29.19%) aligned >1 times
97.83% overall alignment rate
Time searching: 00:00:47
Overall time: 00:00:47
```

Итоговый процент выравнивания

`97.83% overall alignment rate` - означает, что из всех поданных на вход ридов (пар и одиночных) 97.83% успешно сопоставились с референсным геномом дрожжей.

Конвертирование и индексирование `bam` файла

Конвертирование

```
samtools view -Sb aligned/output.sam > aligned/output.bam
```

Сортировка и индексирование `bam` файла

```
samtools sort aligned/output.bam -o aligned/output_sorted.bam
```

QC-отчет

```
qualimap bamqc -bam aligned/output_sorted.bam -gff genomic.gff -c -nw 400 -hm 3
-outdir aligment_qc
```

Анализ отчета

1. Эффективность выравнивания

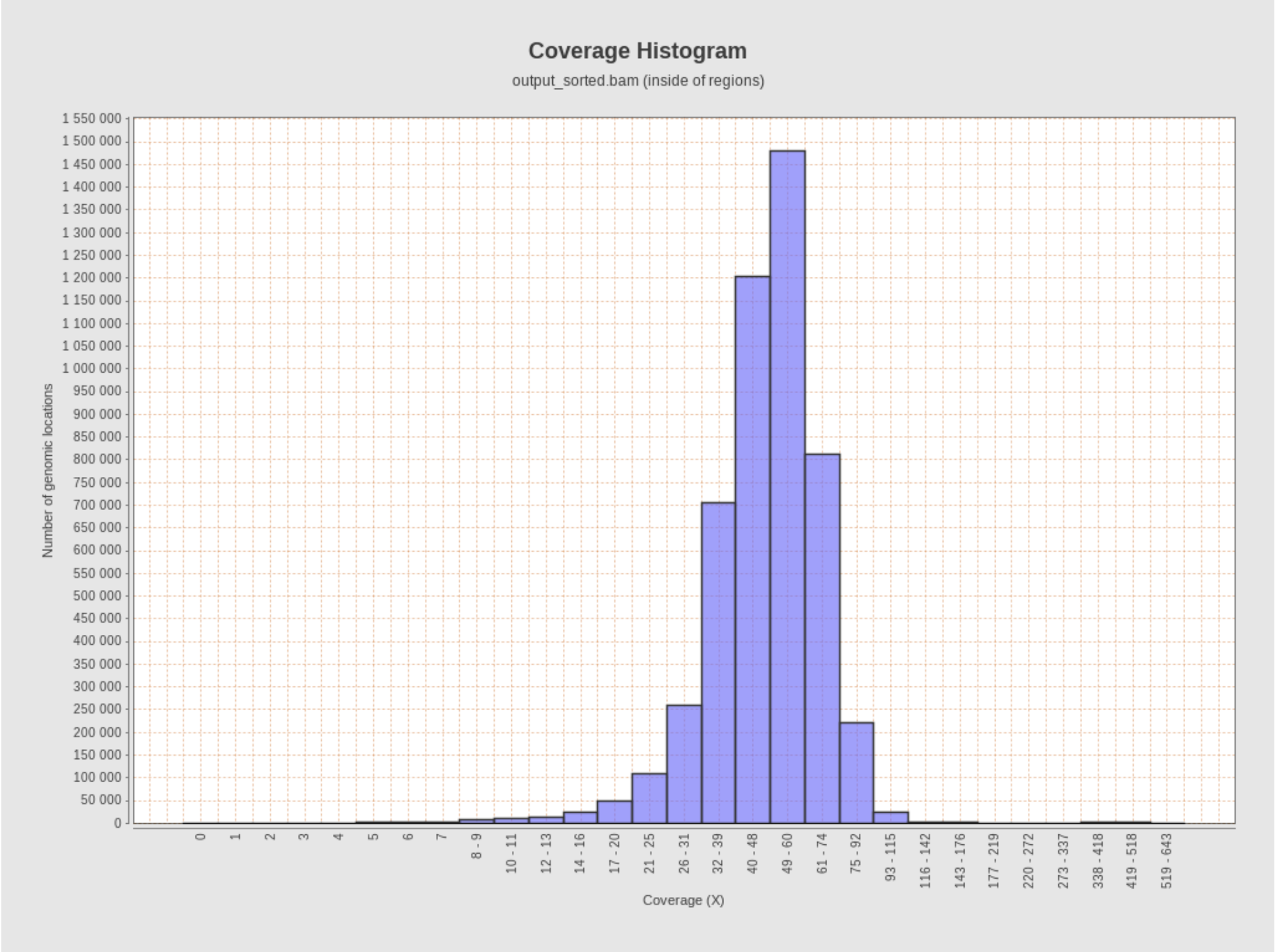
Reference size	4 936 175
Number of reads	1 488 064
Mapped reads	1 455 802 / 97,83%
Unmapped reads	32 262 / 2,17%

Высокий процент успешного выравнивания ридов `97.83%` говорит о хорошем качестве образца и правильном выборе референсного генома

2. Равномерность покрытия

Coverage (inside of regions)

Mean	50,0514
Standard Deviation	19,6403

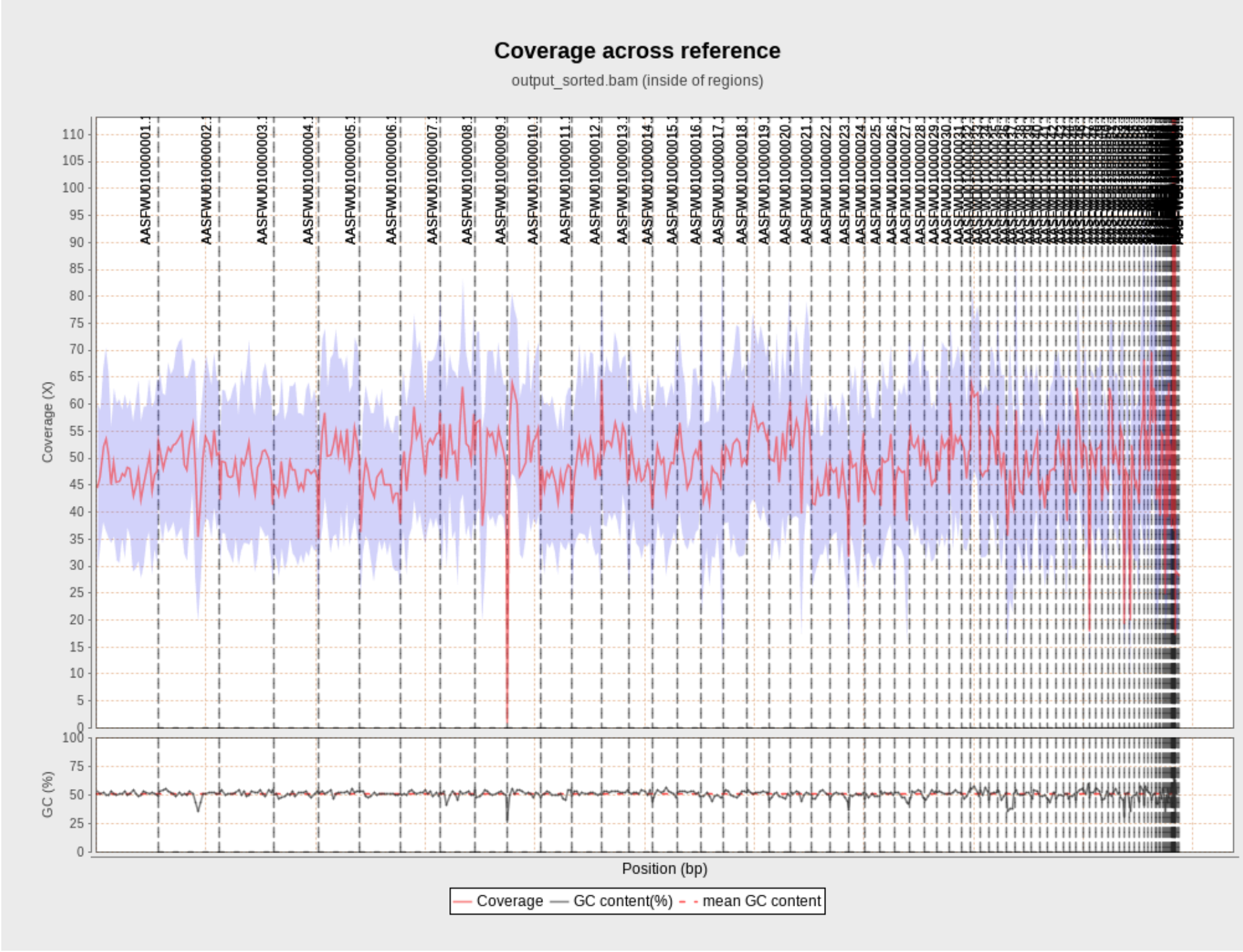

Coverage Histogram
output_sorted.bam (inside of regions)
Number of genomic locations
1 550 000
1 500 000
1 450 000
1 400 000
1 350 000
1 300 000
1 250 000
1 200 000
1 150 000
1 100 000
1 050 000
1 000 000
950 000
900 000
850 000
800 000
750 000
700 000
650 000
600 000
550 000
500 000
450 000
400 000
350 000
300 000
250 000
200 000
150 000
100 000
50 000
0
0
1
2
3
4
5
6
7
8 - 9
10 - 11
12 - 13
14 - 16
17 - 20
21 - 25
26 - 31
32 - 39
40 - 48
49 - 60
61 - 74
75 - 92
93 - 115
116 - 142
143 - 176
177 - 219
220 - 272
273 - 337
338 - 418
419 - 518
519 - 643
Coverage (X)

Средняя глубина покрытия 50.05, что достаточно хорошо

3. Аномалии покрытия

Есть несколько хромосом с экстремально высоким покрытием

AASFWU010000080.1	1021	457455	448,046	155,3103
AASFWU010000083.1	817	110419	135,1518	55,4268
AASFWU010000077.1	1337	585189	437,6881	106,2931
AASFWU010000079.1	1045	354821	339,5416	84,3407


Coverage across reference
output_sorted.bam (inside of regions)
Coverage (X)
GC (%)
Position (bp)
Coverage
GC content(%)
mean GC content

Но их длина короче остальных, возможно это рнк

4. Несовпадения и инделы

Mismatches and indels (inside of regions)

General error rate	0,84%
Mismatches	1 914 859
Insertions	28 519
Mapped reads with at least one insertion	1,46%
Deletions	8 534
Mapped reads with at least one deletion	0,56%
Homopolymer indels	39,46%

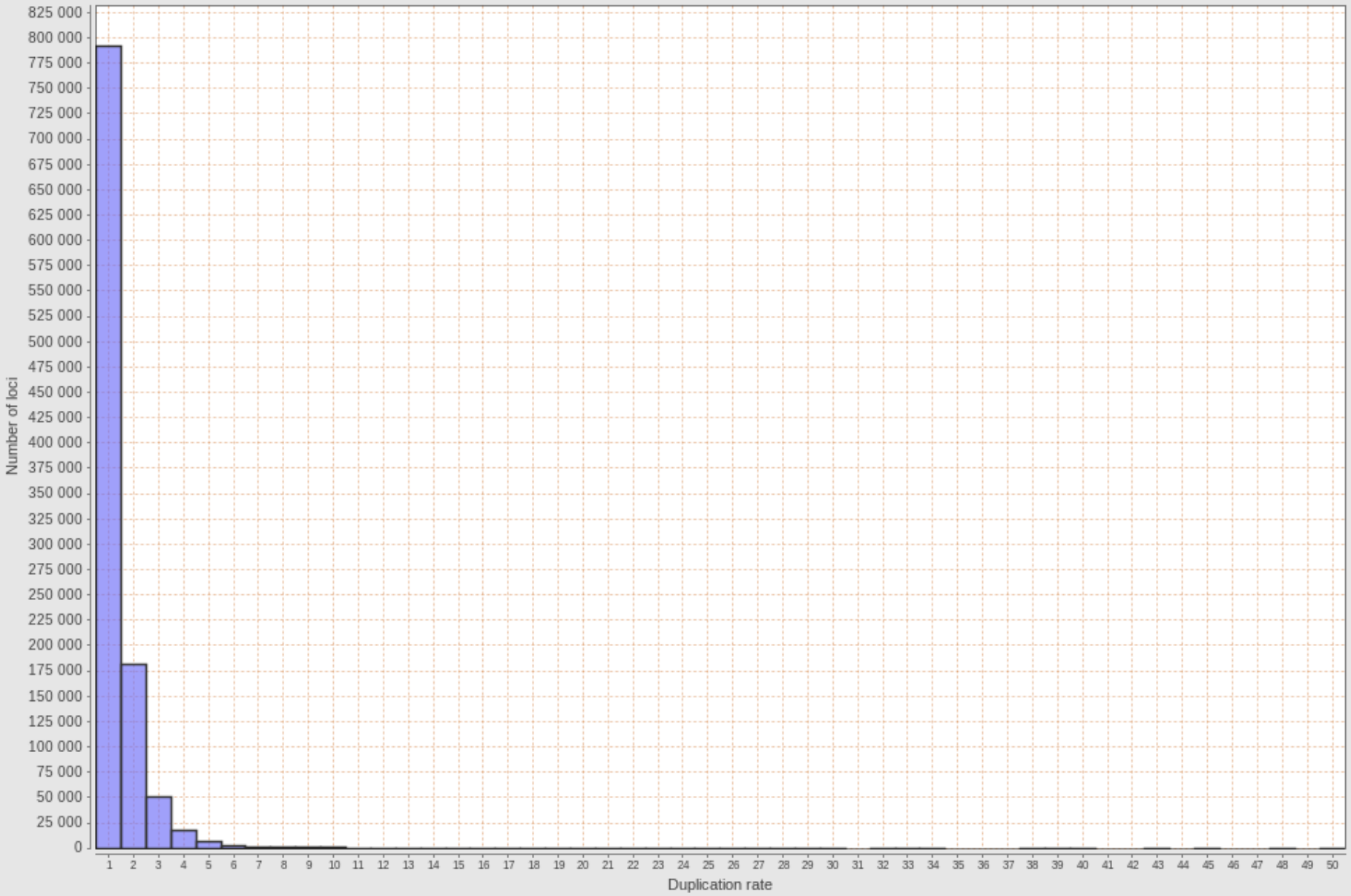
Общая частота ошибок: 0.84%

Мисматчи : 1,914,859

Вставки: 28,519 (1.46% ридов содержат вставки)

Делеции 8,534 (0.56% ридов содержат делеции)

5. Дубликаты


Duplication Rate Histogram
output_sorted.bam (inside of regions)
Number of loci
825 000
800 000
775 000
750 000
725 000
700 000
675 000
650 000
625 000
600 000
575 000
550 000
525 000
500 000
475 000
450 000
425 000
400 000
375 000
350 000
325 000
300 000
275 000
250 000
225 000
200 000
175 000
150 000
125 000
100 000
75 000
50 000
25 000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Duplication rate

На графике дубликатов видно, что подавляющее большинство позиций в геноме уникальны (огромный пик на значении "1"), а позиции с дубликатами встречаются крайне редко и их очень мало.

Визуализация данных

Используем `JBrowse`