

Скачивание референсной последовательности

- Находим организм в описании эксперимента

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](#)

- В базе нуклеотидов на [NCBI](#) находим организм

GenBank ▾ Send to: ▾

Related information

Assembly
BioProject
BioSample
Taxonomy
Protein from WGS

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

GenBank: AASFU000000000.1

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

[PopSet](#)

Go to:

LOCUS	AASFU010000000	93	rc	DNA	linear	BCT	15-APR-2020
DEFINITION	Escherichia coli O8:H36 strain MOD1-EC6081, whole genome shotgun sequencing project.						
ACCESSION	AASFU000000000						
VERSION	AASFU000000000.1						
DRI TNK	BioProject: PRJNA230969						

Recent activity

Escherichia coli O8:H36 strain MOD1-EC6081, who

Escherichia coli O8:H36 strain MOD1-EC6081, who

- 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 AASFNU01

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 alignment_qc
```

Анализ отчета

1. Эффективность выравнивания

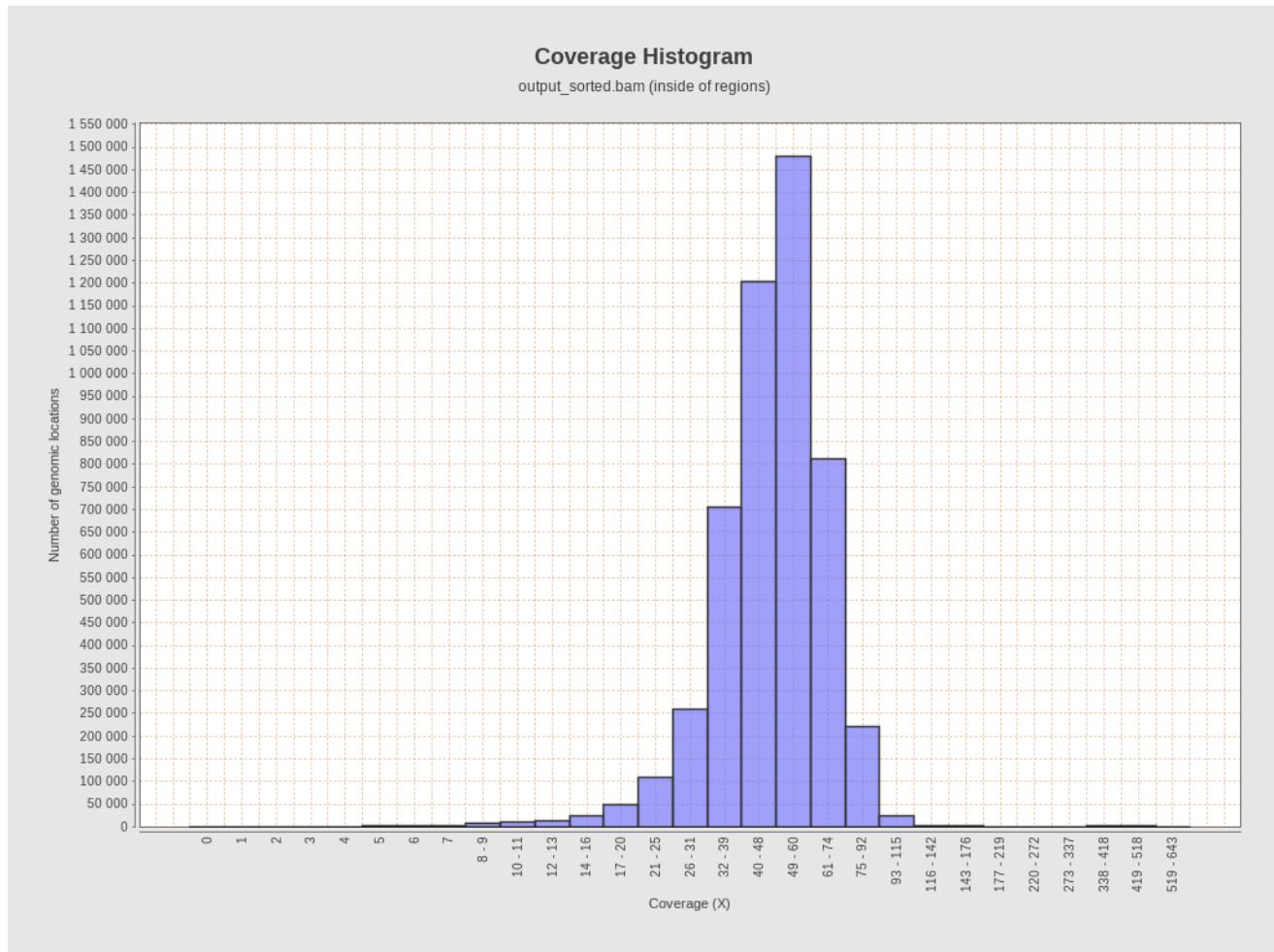
Reference size	4 936 175
Number of reads	1 488 064
Mapped reads	1 455 802 / 97,83%
Unmapped reads	32 262 / 2,17%

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

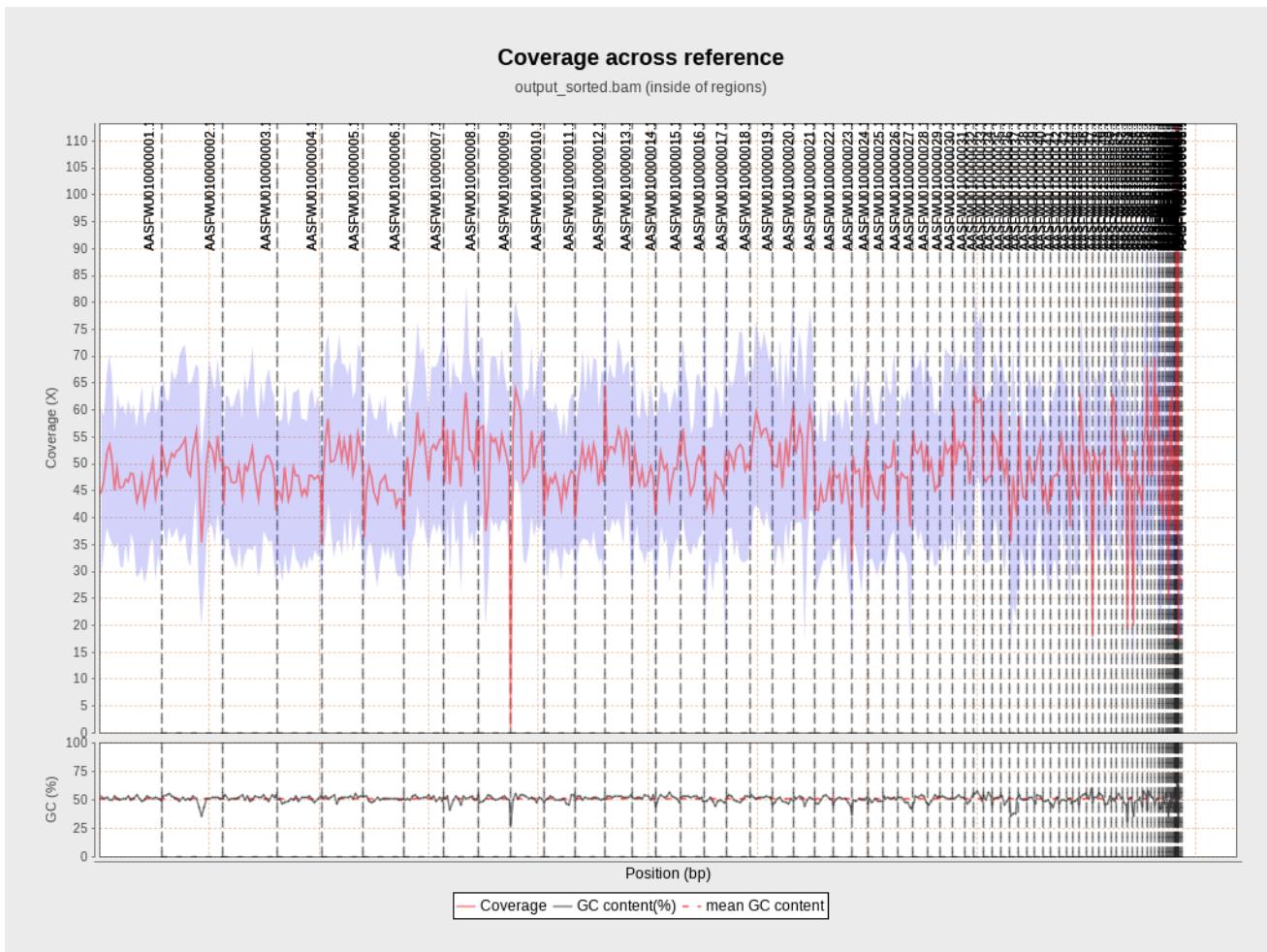


Средняя глубина покрытия 50.05, что достаточно хорошо

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



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%

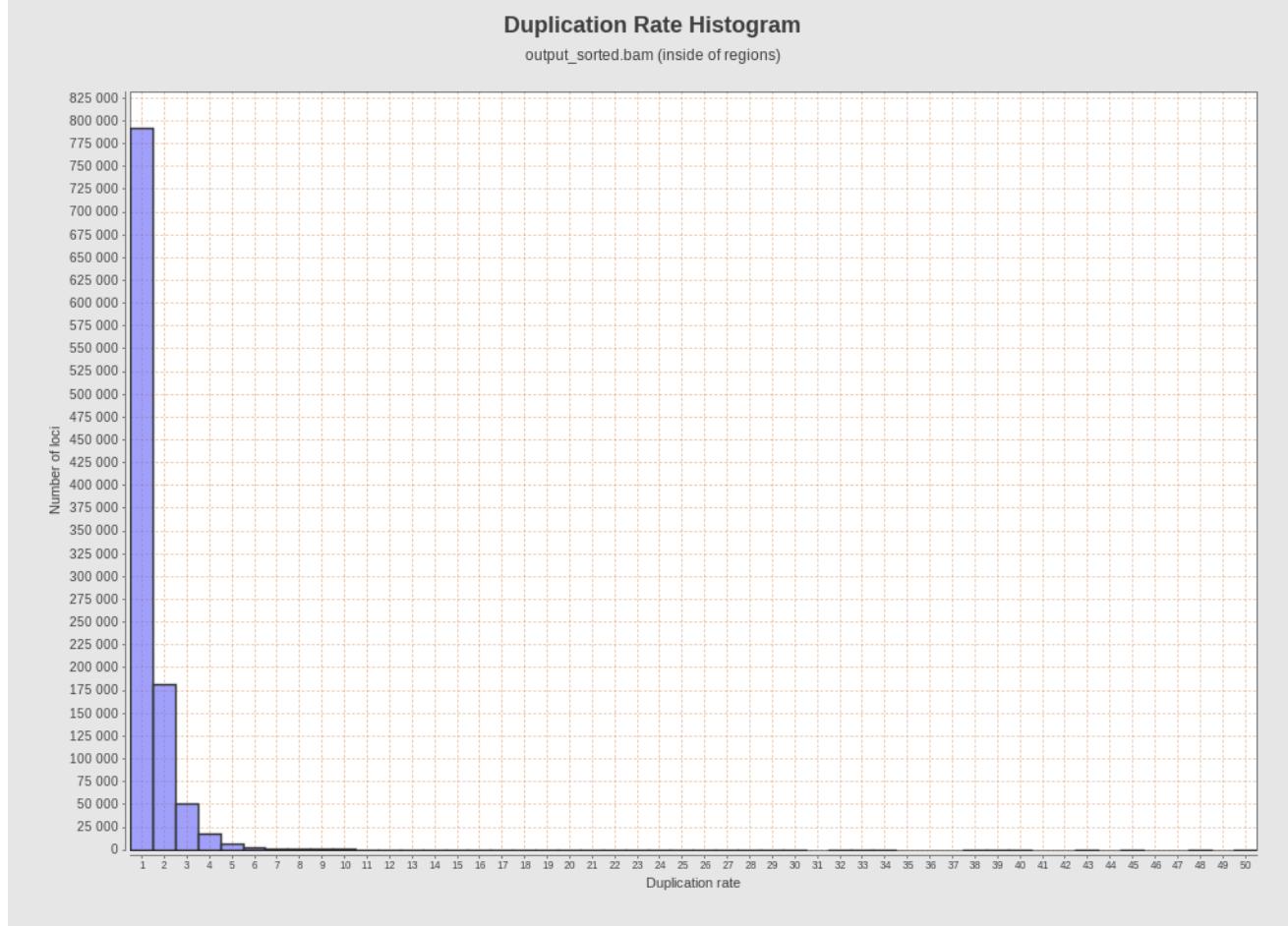
Общая частота ошибок: 0.84%

Мисматчи : 1,914,859

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

Делеции 8,534 (0.56% ридов содержат делеции)

5. Дубликаты

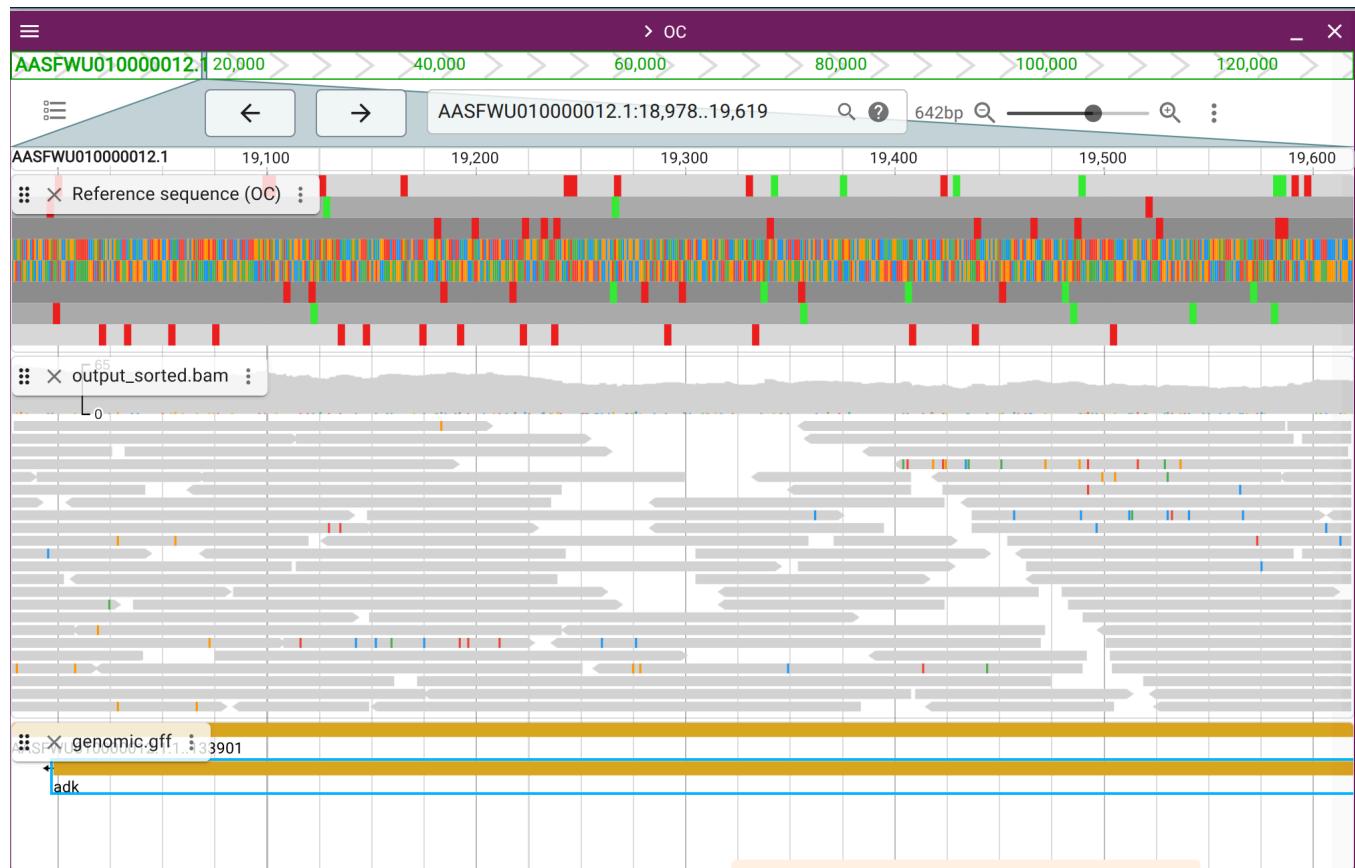


На графике дубликатов видно, что подавляющее большинство позиций в геноме уникальны (огромный пик на значении "1"), а позиции с дубликатами встречаются крайне редко и их очень мало.

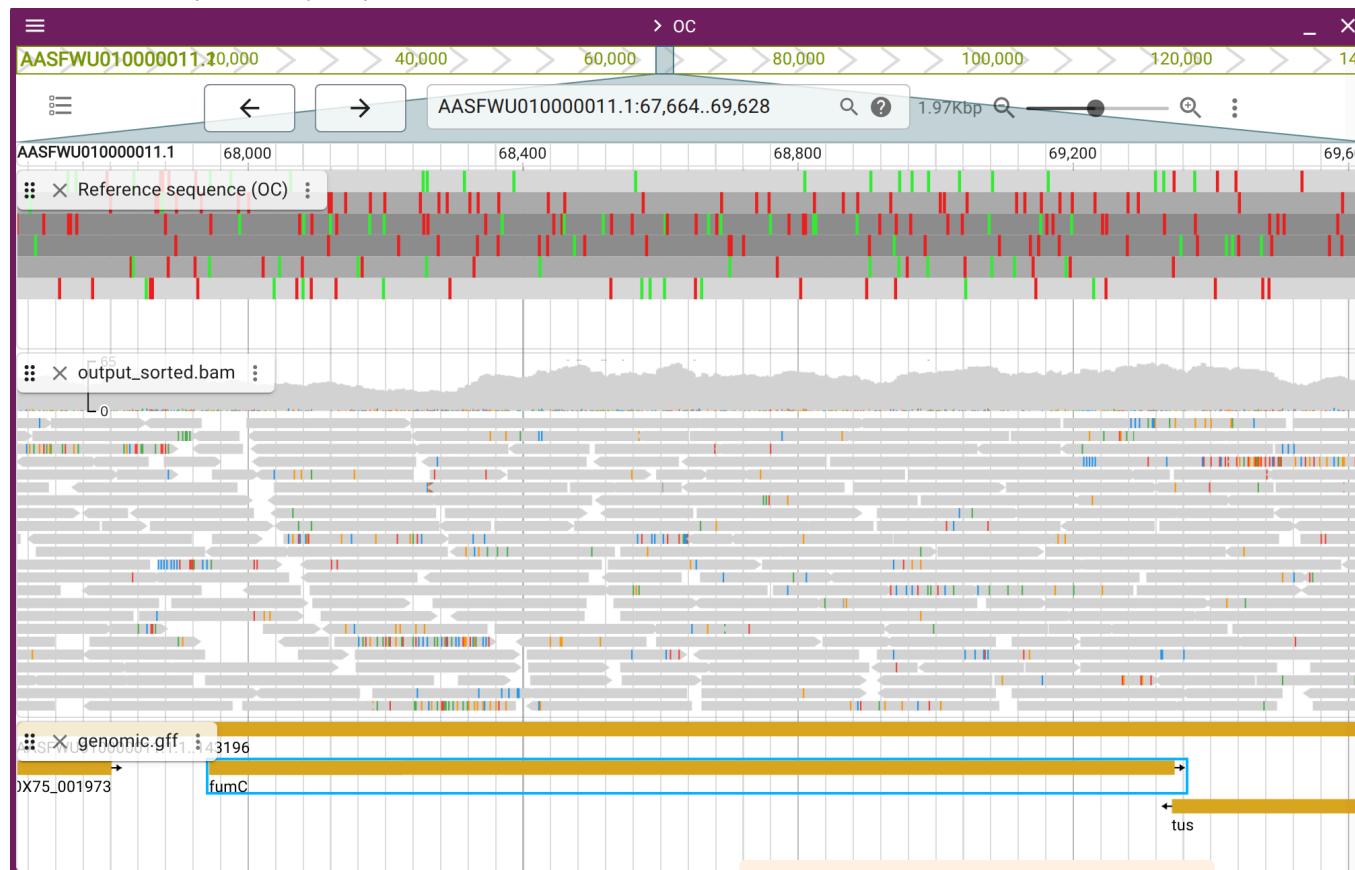
Визуализация данных

Используем JBrowse

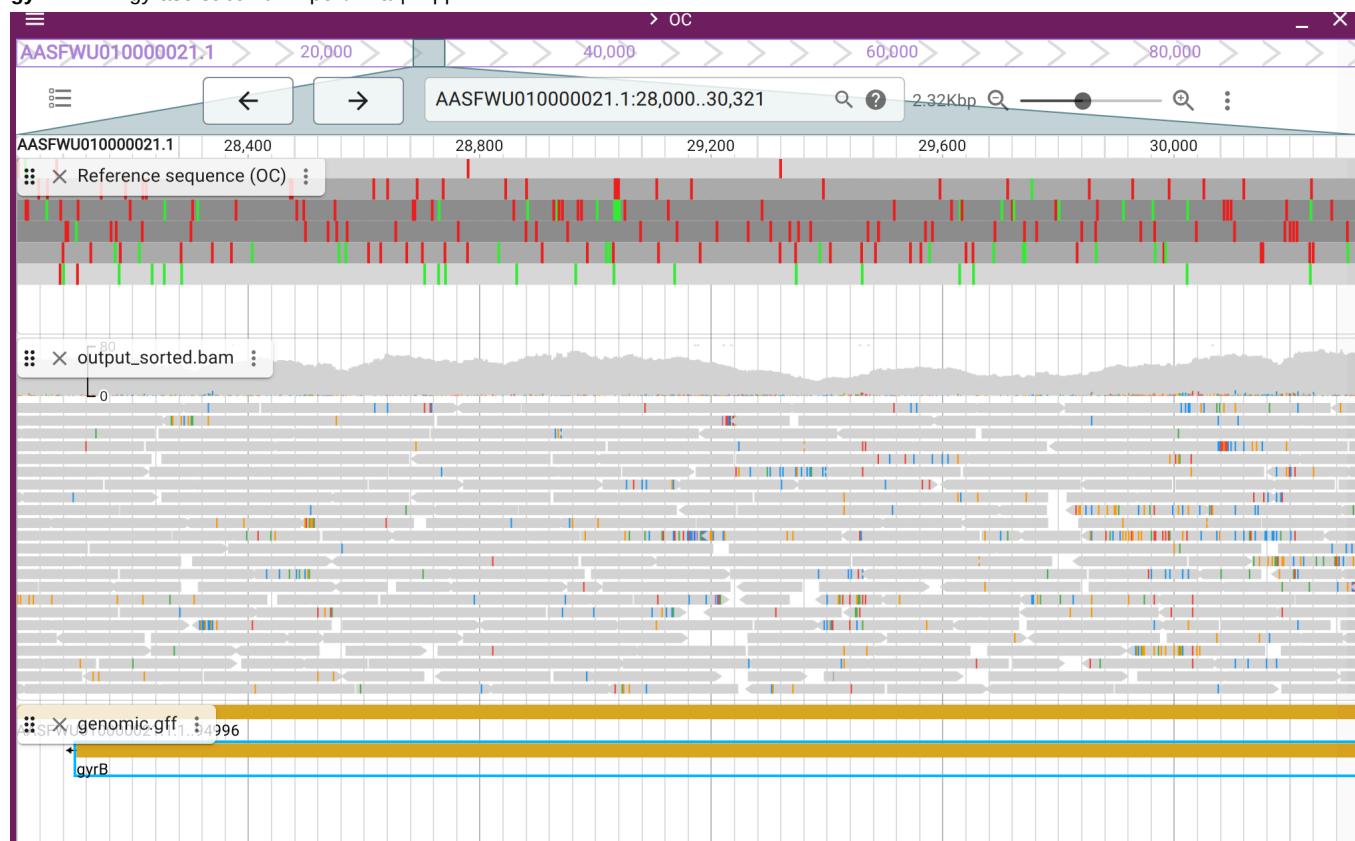
adk - Adenylate kinase - метаболизм энергии



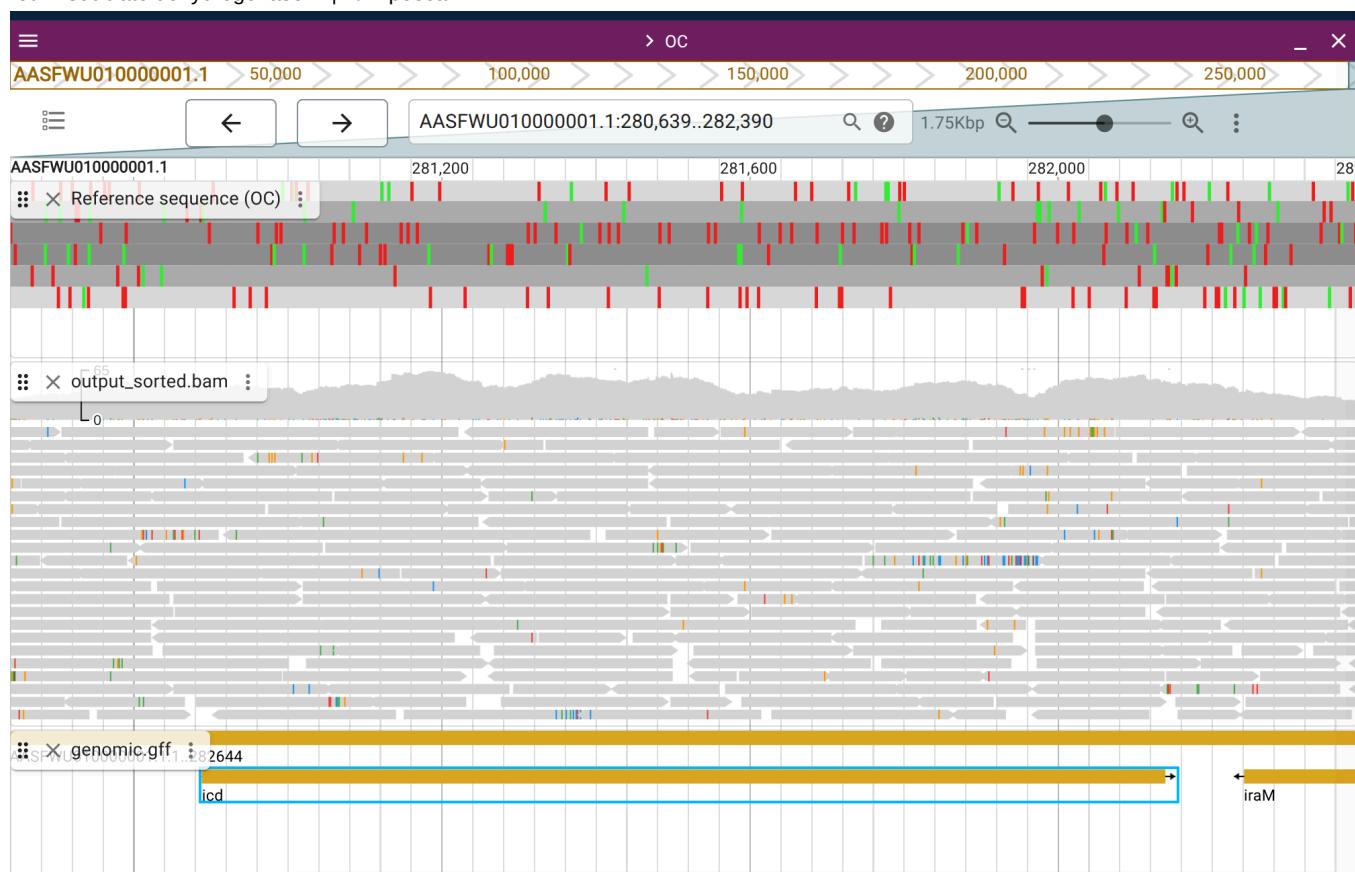
fumC - Fumarate hydratase - цикл Кребса



gyrB - DNA gyrase subunit B - репликация ДНК



icd - Isocitrate dehydrogenase - цикл Кребса



mdh - Malate dehydrogenase - цикл Кребса

