

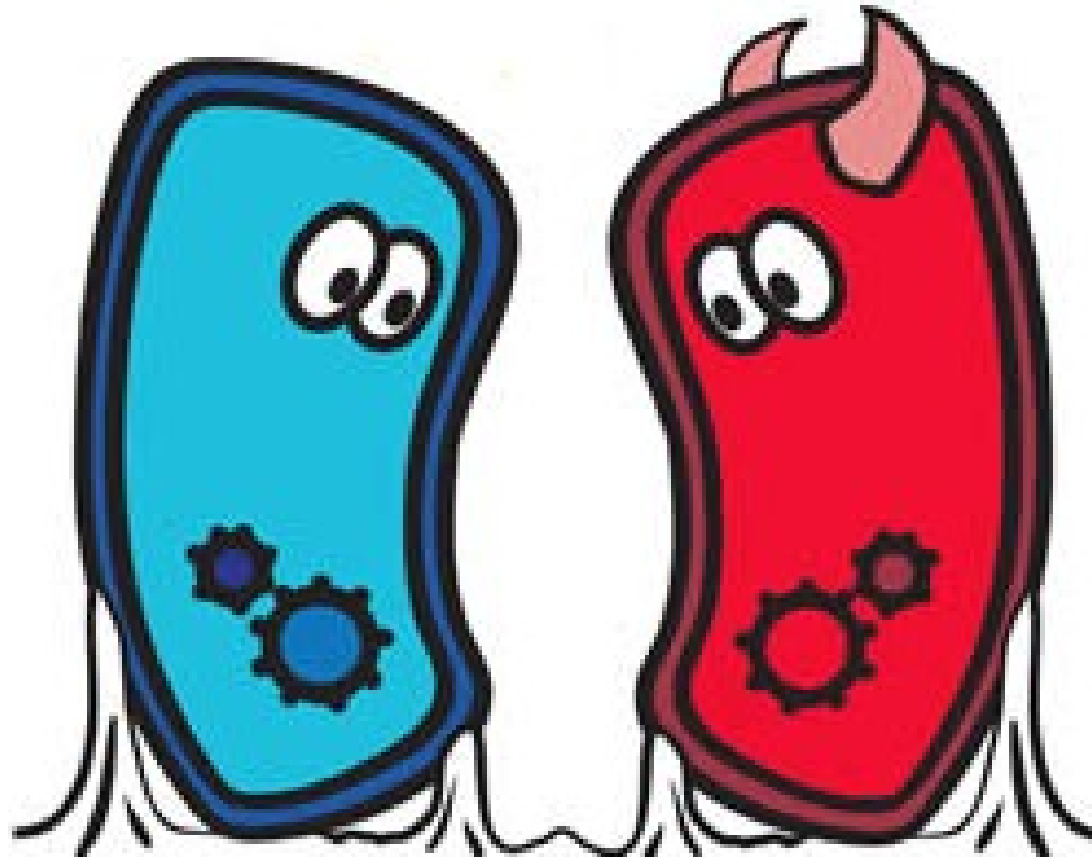
Поиск участков патогенности в геноме штамма *E.coli* 0104:H4

Руководитель проекта: Михаил Райко



В 2011 году в Германии произошла страшная эпидемия острой кишечной инфекции. Пострадало более 3'000 человек.

Почему новый штамм вызывает такие тяжёлые симптомы?



Задачи

1. Собрать геном нового штамма
2. Сравнить с родственным штаммом
3. Найти причину аномальной патогенности

Этапы геномного проекта

1. Убедиться в качестве исходных ридов (FastQC)
2. Собрать контиги (SPAdes)
3. Проверить качество сборки (Quast)
4. Собрать скаффолд с помощью референсного генома (Abacas, Contiguator)
5. Проаннотировать геном (Prokka, RAST)
6. Сравнить с референсным геномом (Mauve)

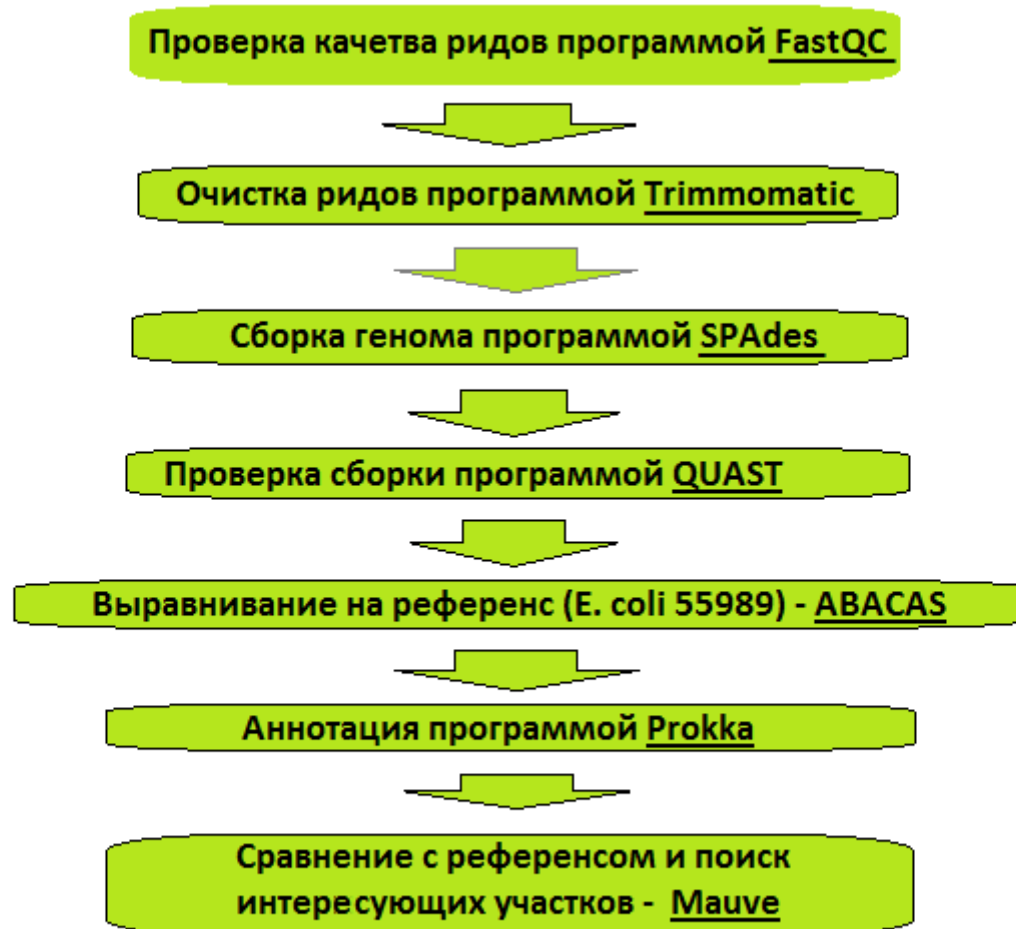
Группа №1

Доменикан Александр
Шишкина Элина
Бутенко Николай
Зайнагтдинова Галия

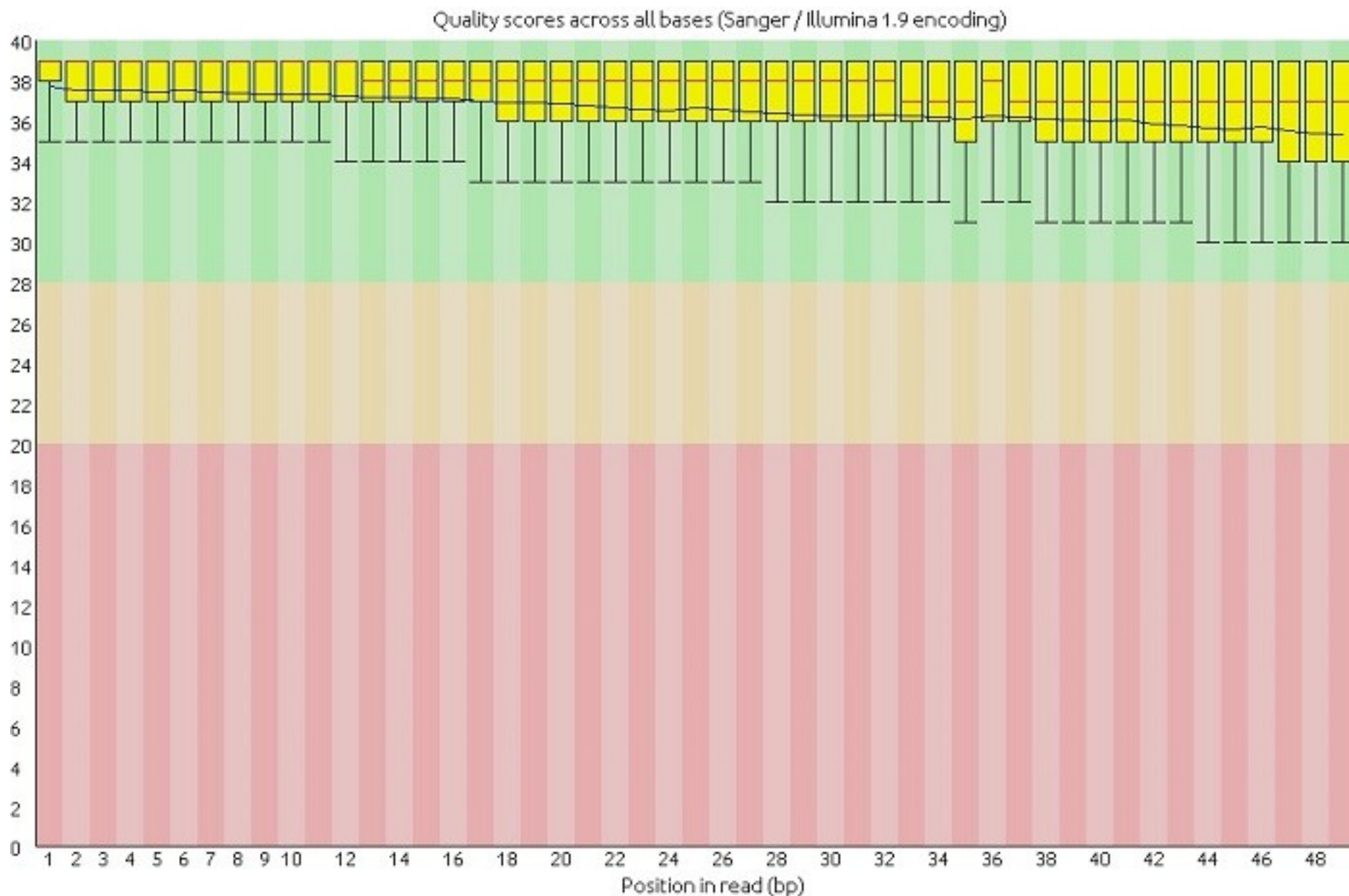
Собственно Pipeline



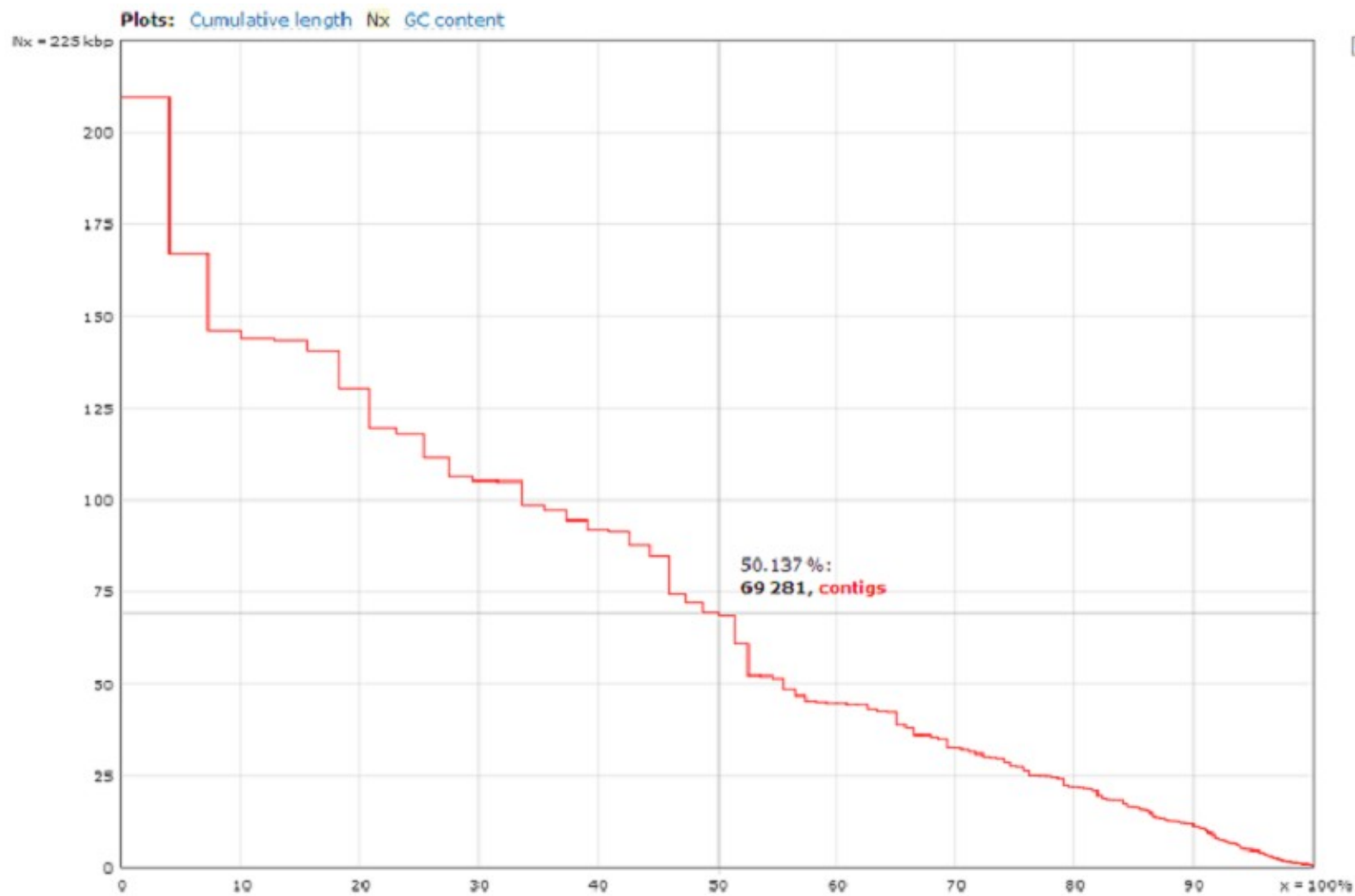
Pipeline



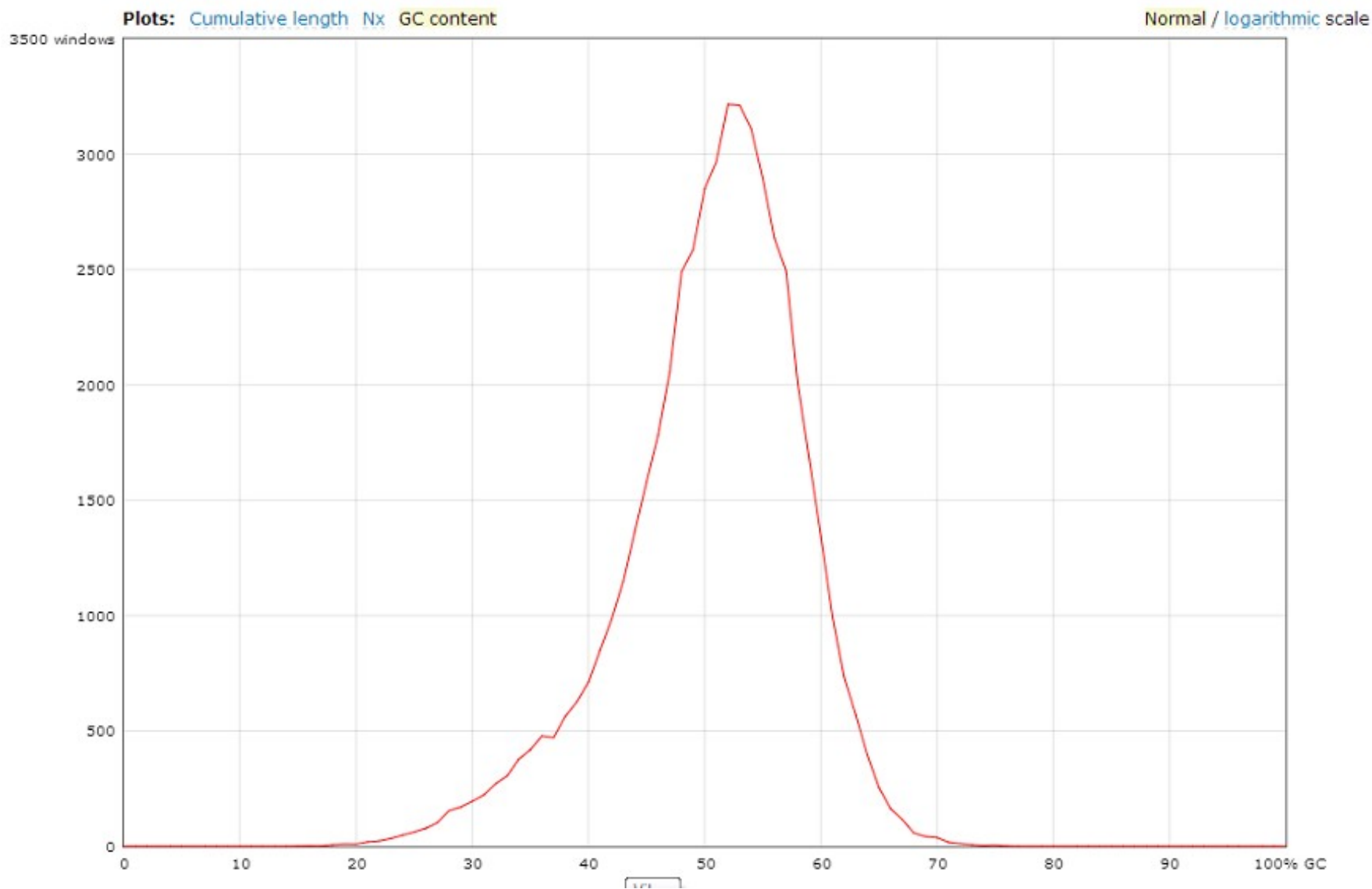
Проверка качества ридов



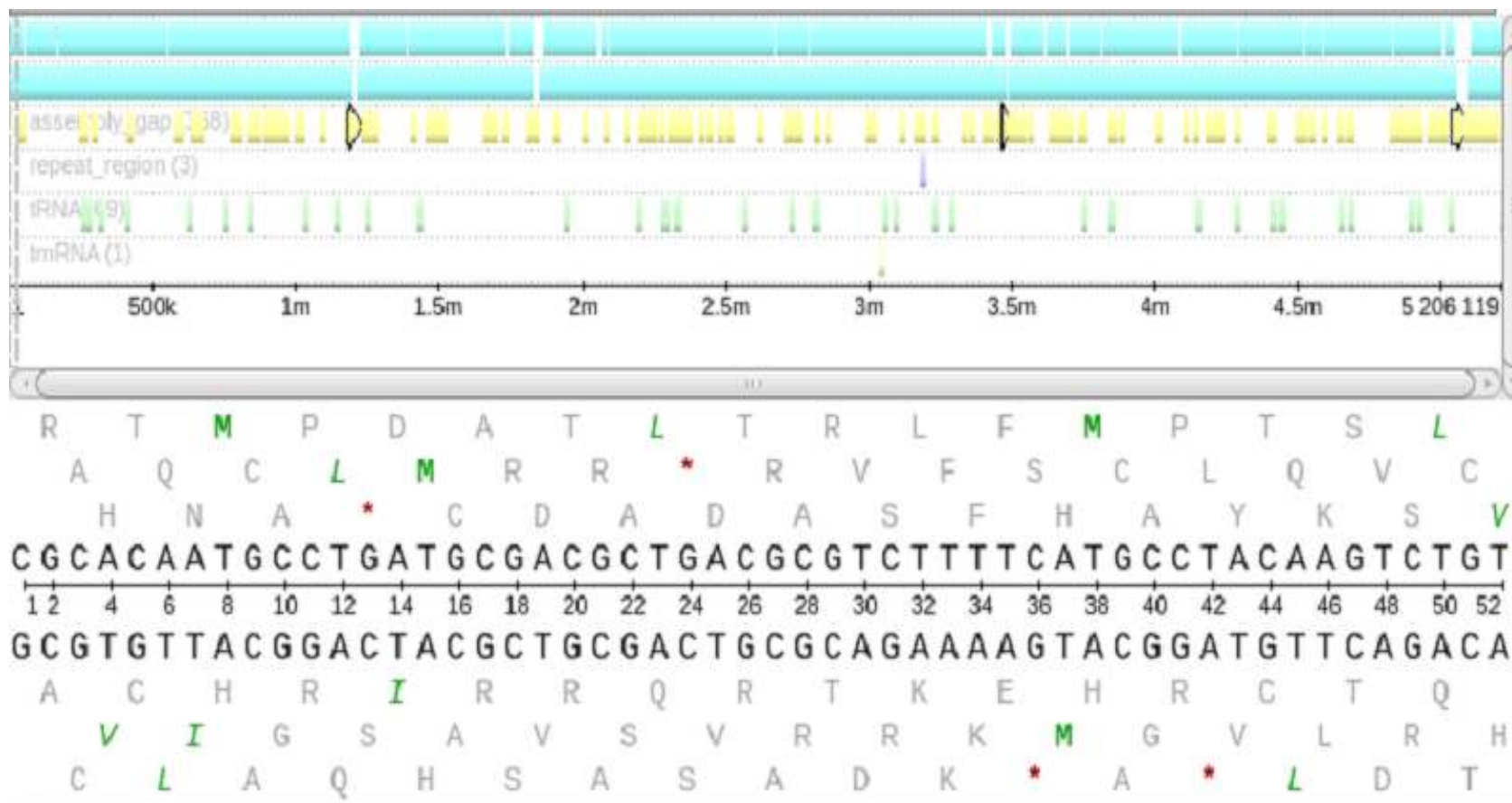
Проверка качества сборки (N50)



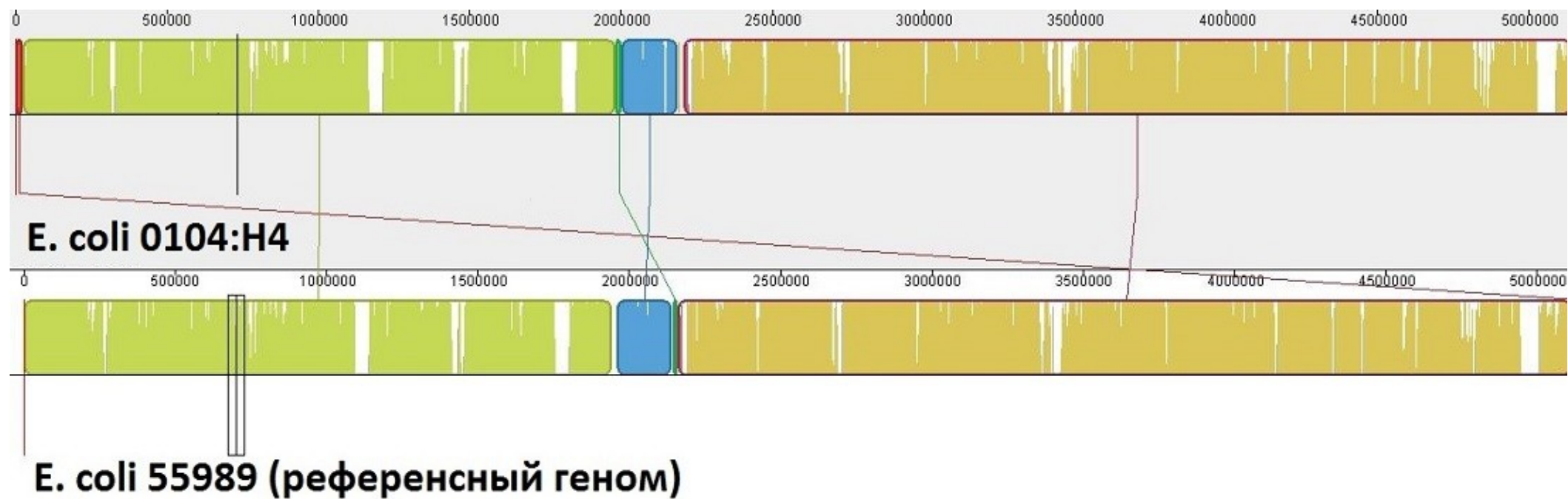
Проверка качества сборки (GC контент)



Визуализация аннотации (Ugene)



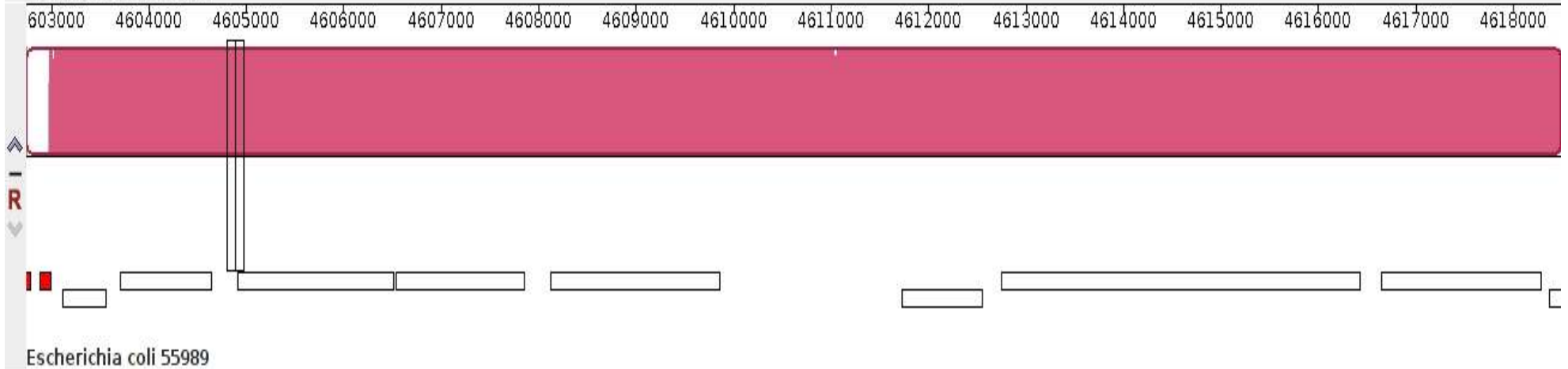
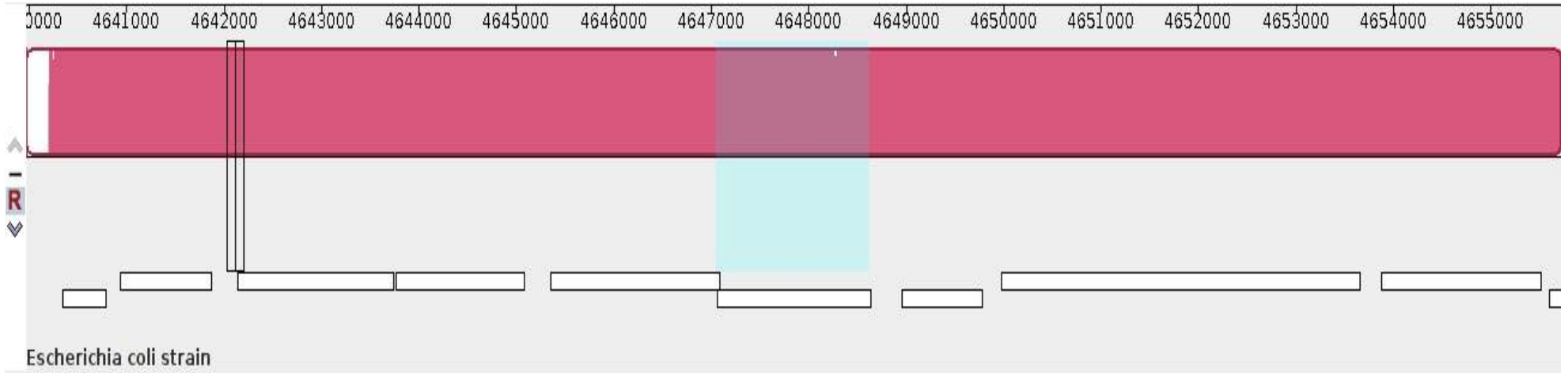
Сравнение с геномом E. coli 55989



Как мы искали нужный участок?
Ручками.



Note: 'enterotoxin'



Результат. Выравнивание BLASTом на гомологичные белки

NCBI

Conserved Domains

SH3 SH2 SH1

Conserved domains on [lc|92020]

View Standard Results

Local query sequence

Graphical summary show options

Query seq. Specific hits Superfamilies Multi-domains

Toxin_15

Toxin_15 superfamily

pfam07906

[Specific hit] pfam07906, ShET2 enterotoxin, N-terminal region ;The members of this family are sequences that are similar to the N-terminal half of the ShET2 enterotoxin produced by Shigella flexneri and Escherichia coli. This protein was found to confer toxigenicity in the Ussing chamber, and the N-terminal region was found to be important for the protein's enterotoxic effect. It is thought to be a hydrophobic protein that forms inclusion bodies within the bacterial cell, and may be secreted by the Mxi system. Most members of this family are annotated as putative enterotoxins, but one member is a regulator of acetyl CoA synthetase, and another two members are annotated as ankyrin-like regulatory proteins and contain Ank repeats (pfam00023).

List of domain hits

Name	Accession	Description
[+] Toxin_15	pfam07906	ShET2 enterotoxin, N-terminal region. The members of this family are are sex
[+] PRK12409	PRK12409	D-amino acid dehydrogenase small subunit; Provisional

Blast search parameters

Data Source: Live blast search RID = XN13AUNB01R

User Options: Database: CDSEARCH/cdd v3.11 Low complexity filter: yes Composition Based Adj

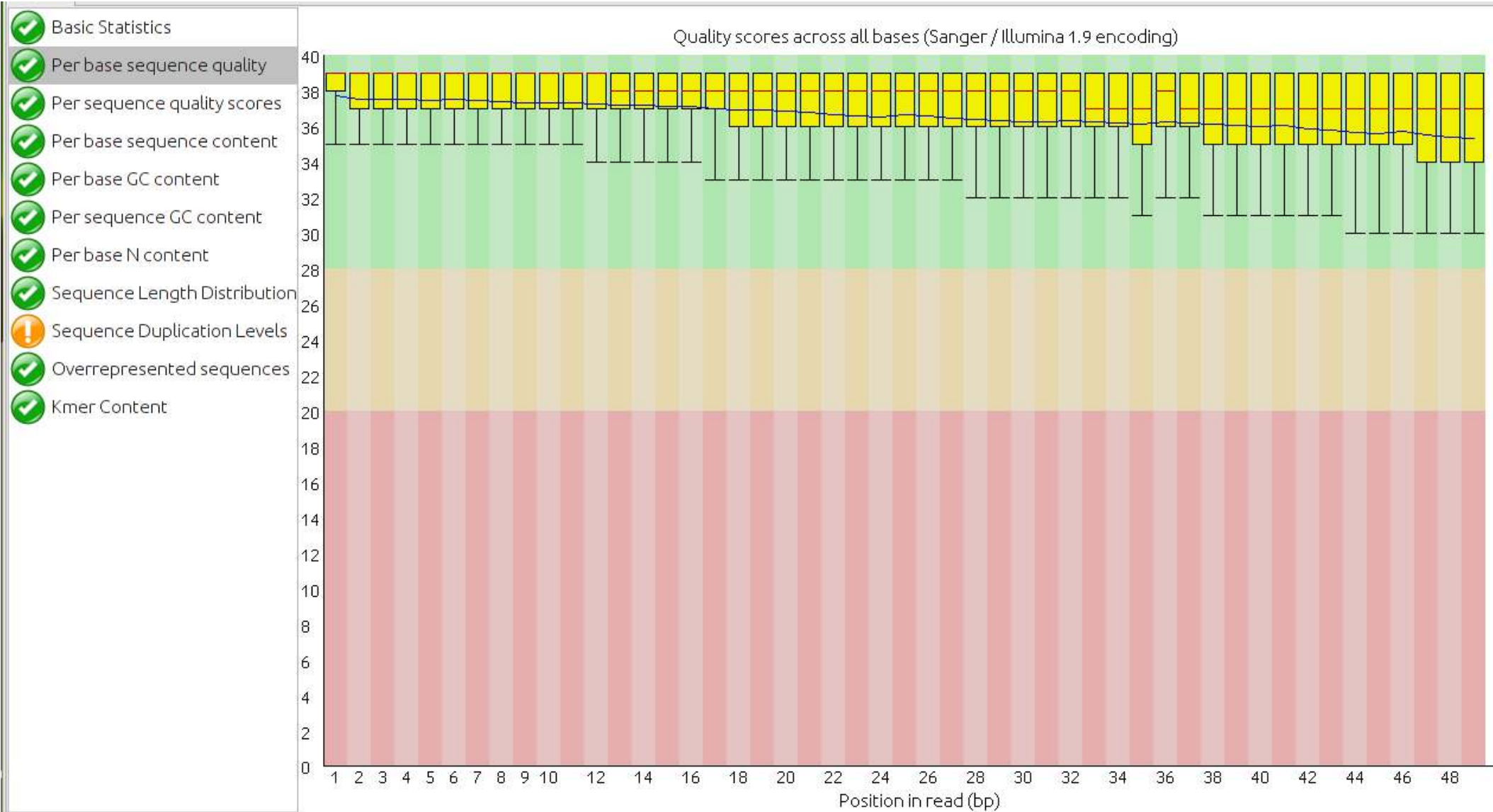
References:

- Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", *Nucleic Acids Res.*39(D)225-9.
- Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", *Nucleic Acids Res.*37(D)205-10.
- Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", *Nucleic Acids Res.*32(W)327-331.

Группа №2

- Голокоз Александр
- Гусак Юлия
- Котенко Анастасия
- Малиновский Илья
- Ненарокова Анна
- Орлова Марина

Качество исходных ридов



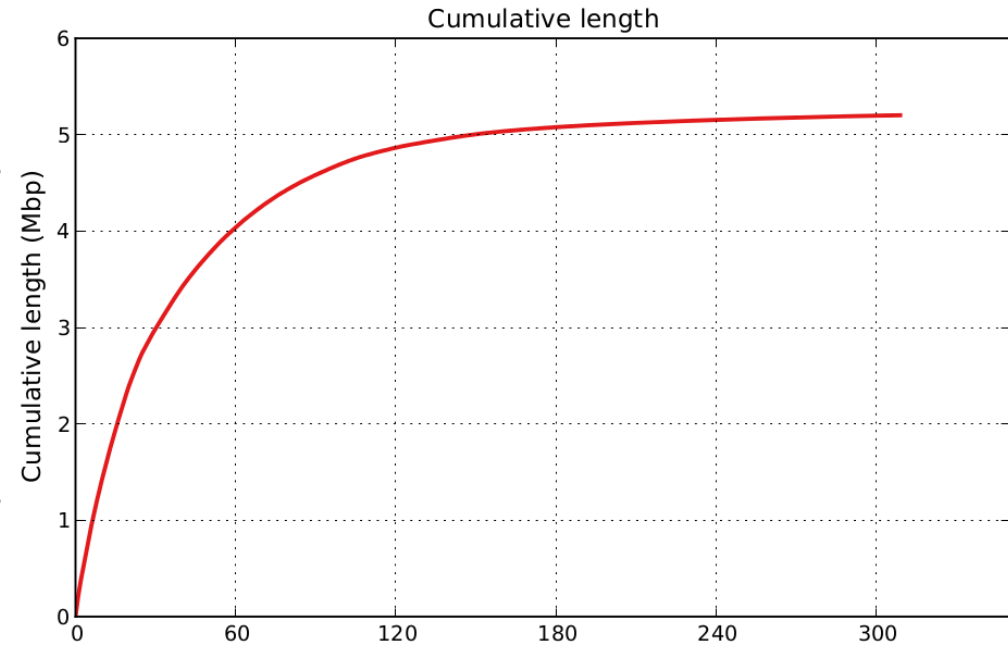
Отчёт о качестве сборки

Statistics without reference ≡ contigs

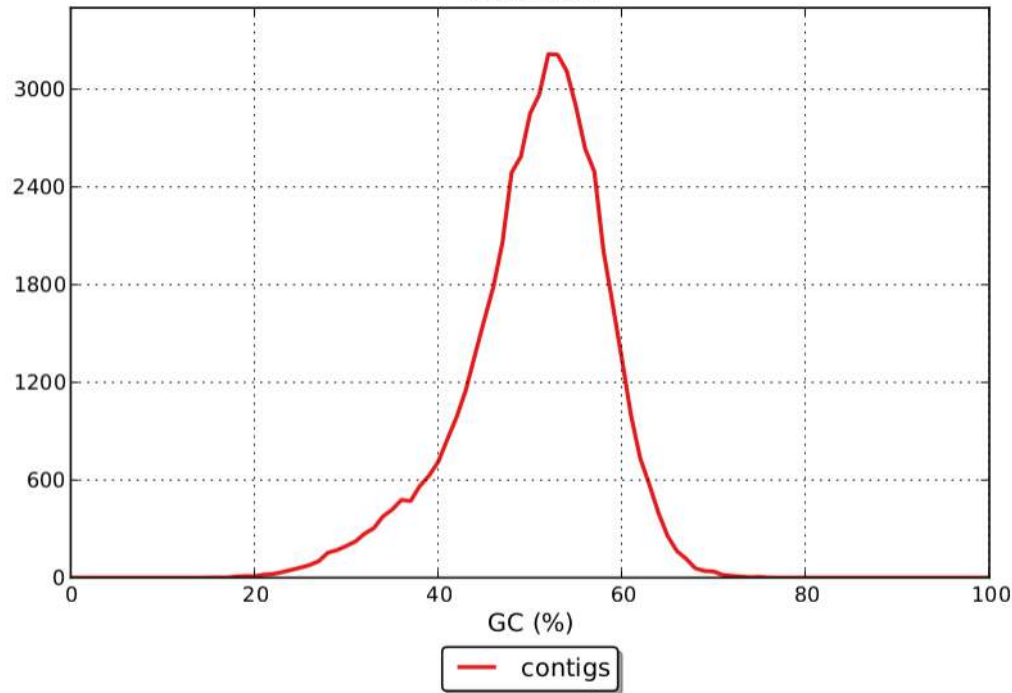
# contigs	309
Largest contig	209 682
Total length	5 201 829
N50	69 281

Mismatches

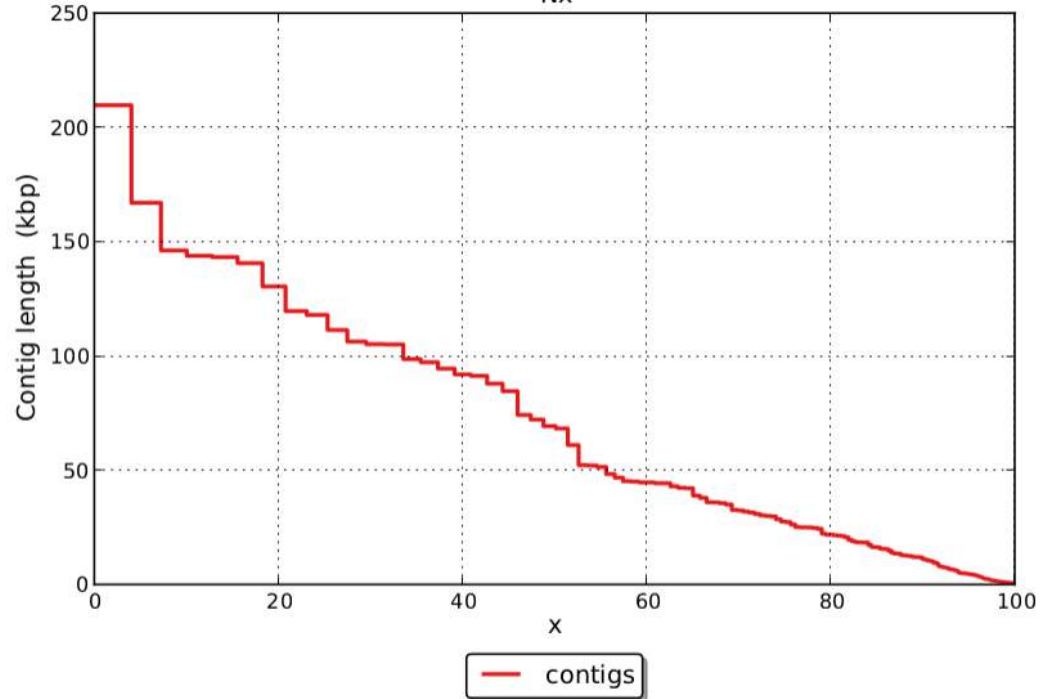
# N's per 100 kbp	0
-------------------	---



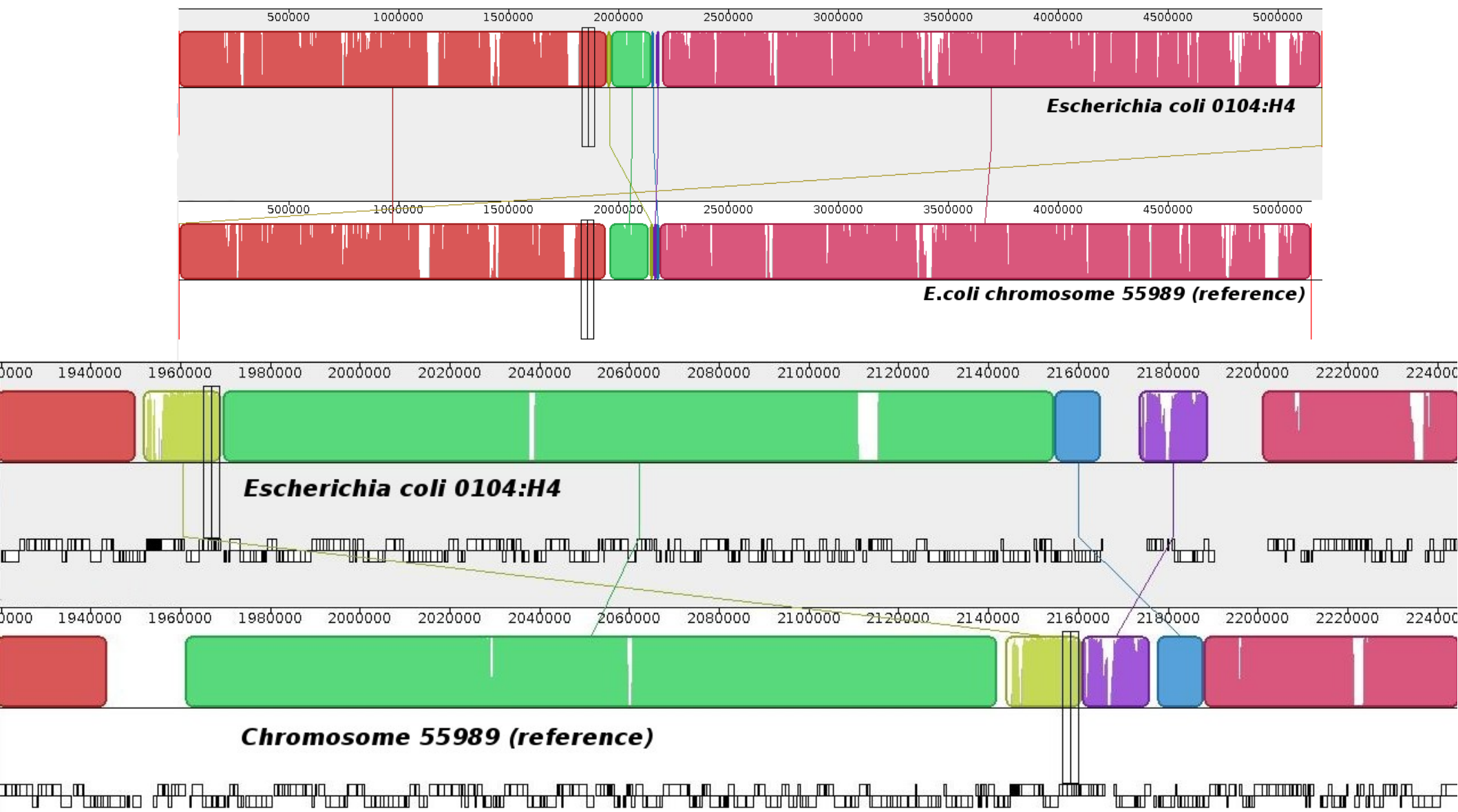
GC content



Nx



Сравнение изучаемого штамма с родственным



Скрипт для автоматической обработки ридов

```
import argparse
from os import system, remove, rename

def processRings(filename, qualityThreshold):
    system("cat " + filename)

    reportDir = "Rings[" + str(filename[-4]) + "].txt"
    leftLowQual = 0
    rightLowQual = 0
    rightBorderSet = False

    # Initializing reads with quality below threshold
    with open(reportDir + "fastq_data.txt", "w") as report:
        data = sys.stdin.readlines()

        for i in range(len(data)):
            if "Sequence length" in data[i]:
                rightLowQual = int(data[i].split()[2]) + 1

            if "Read length" in data[i]:
                j = i + 1
                while ("seq" in data[j]):
                    seq = data[j].split()[0]
                    readNumbers = 0
                    if "in st-Header":
                        readNumbers = seq.split()[1]
                    else:
                        readNumbers = int(seq.split()[1])
                    if float(data[j].split()[2]) < qualityThreshold:
                        if leftLowQual == readNumbers[0] - 1:
                            leftLowQual = readNumbers[0]
                        else:
                            if not rightBorderSet:
                                rightLowQual = readNumbers[1]
                                rightBorderSet = True
                            j = j + 1

            if "seq" in data[i] and "div" in data[i - 1]:
                j = i - 1
                while ("seq" in data[j]):
                    seq = data[j].split()[0]
                    if type(seq) == "int":
                        seq = seq.split()[0]
                    readNumbers.append(seq)

    # Print leftLowQual, rightLowQual
    printLeftLowQual = filename[-4] + "LeftLowQual"
    system("cat " + reportDir + "LeftLowQual.txt")
    # Print rightLowQual
    printRightLowQual = filename[-4] + "RightLowQual"
    system("cat " + reportDir + "RightLowQual.txt")
    # Print all reads with q >= qualityThreshold
    with open(reportDir + "fastq_data.txt", "w") as report:
        with open(reportDir + "fastq_data.txt", "r") as report:
            data = sys.stdin.readlines()
            readStrings = []
            remove = False
            for i in range(len(data)):
                readStrings.append(data[i])
```

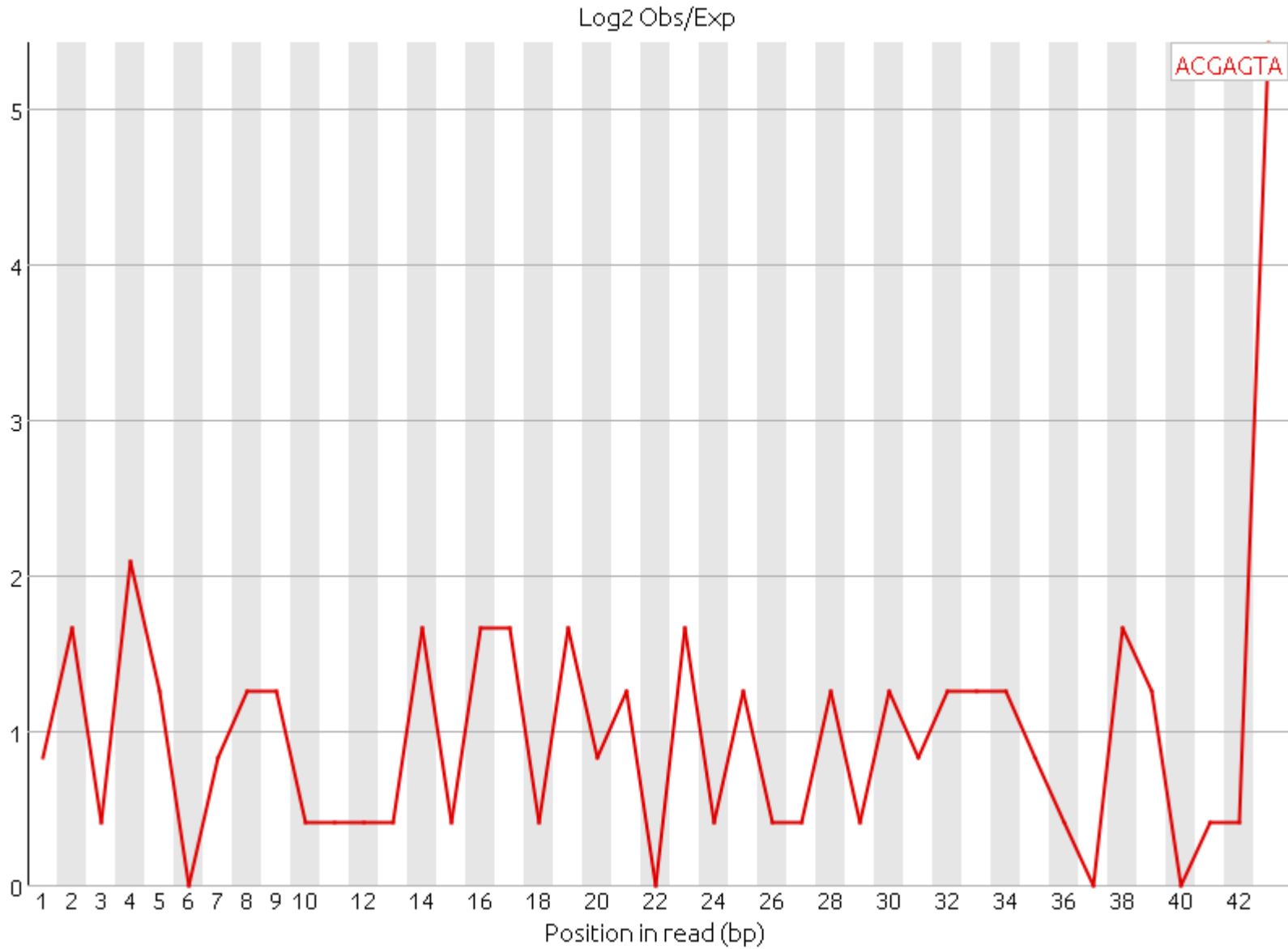
Скрипт для сравнения gbk файлов

1. Сравнивает по названию генов и в output выдает названия новых генов и их белков
2. Сравнивает последовательности белков и выдает названия новых белков

Группа 3

Иванов Кирилл
Фурменков Александр
Митрофанова Ольга

Отчет в FastQC



Приобретенный опыт

- Навыки работы с ранее неизвестным биоинформатическим софтом
- Знакомство с новыми форматами представления данных и их структурой
- Практический скилл сборки генома с проверками качества, ранее неизвестный
- Улучшение мастерства в черной магии, необходимой для получения результатов там, где другие методы не подходят

Результаты

- Найден ген шига-токсина, который отсутствует у референсного штамма
- Найдены гены фимбрий I типа
- Найдены гены других токсинов, которых нет у референсного штамма
- В базе NCBI найдены гомологичные последовательности для несобранных контигов, они принадлежат *E.coli* и родственным видам

Спасибо за внимание!

