

КАЗАНСКИЙ (ПРИВОЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Институт фундаментальной медицины и биологии

А.Э. СВЕРДРУП, Л.Л. ФРОЛОВА, И.И. ЗАДОРИНА

**СБОРКА И АННОТАЦИЯ ПРОКАРИОТИЧЕСКОГО ГЕНОМА
С ИСПОЛЬЗОВАНИЕМ WEB-СЕРВИСА GALAXY**

Учебно-методическое пособие по дисциплине

«Б1.В.04 Спецпрактикум по прикладным методам в биологии»

06.03.01 Биология (бакалавр)

КАЗАНЬ

2024

УДК 004.9

ББК 28.0

С23

*Печатается по рекомендации учебно-методической комиссии
Института фундаментальной медицины и биологии КФУ
(протокол № 1 от 29.08.2024 г.)*

Рецензенты:

д.б.н., зав. каф. Каюмов А.Р.
кафедра генетики ИФМиБ

Свердруп А.Э., Фролова Л.Л., Задорина И.И.

**С23 Сборка и аннотация прокариотического генома с
использованием web-сервиса Galaxy:** учебно-методическое пособие /
А.Э.Свердруп, Л.Л.Фролова, И.И.Задорина // Казанский федеральный
университет, 2024. – 54 с.

В учебно-методическом пособии приведены основные возможности web-сервиса Galaxy по сборке и аннотации необработанных данных высокопроизводительного секвенирования прокариотических геномов. Рекомендовано для изучения дисциплины: Б1.В.04 «Спецпрактикум по прикладным методам в биологии» по направлению подготовки 06.03.01 Биология (бакалавр), а также при подготовке курсовой работы по специальности, научно-исследовательской работы и выпускной квалификационной работы медицинских и биологических направлений.

УДК 004.9

ББК 28.0

© Свердруп А.Э., Фролова Л.Л., Задорина И.И.

© ФГАОУ ВО КФУ, 2024

Содержание

1. Введение	4
2. Протокол сборки генома	7
3. Сборка генома с использованием web-сервиса Galaxy	9
3.1. Регистрация и загрузка данных	10
3.2. Контроль качества секвенирования.....	13
3.2.1. Оценка качества прямых ридов	13
3.2.2. Оценка качества обратных ридов	19
3.3. Тримминг	22
3.4. Контроль качества ридов после тримминга	27
3.5. Сборка генома.....	32
3.6. Контроль качества сборки генома	36
4. Аннотация прокариотического генома в программе Prokka.....	40
5. Определение таксономии исследуемого генома	42
5.1. Извлечение последовательности <i>16S рPHK</i>	42
5.2. Определение таксономии по последовательности <i>16S рPHK</i> в программе blastn	43
6. Аннотация генома в программе RAST.....	46
7. Заключение.....	53
8. Список литературы	53
9. Список электронных источников.....	54

1. Введение

Сборка генома – последний этап экспериментальной работы в технологии NGS – секвенирования нового (второго) поколения (рис.1).

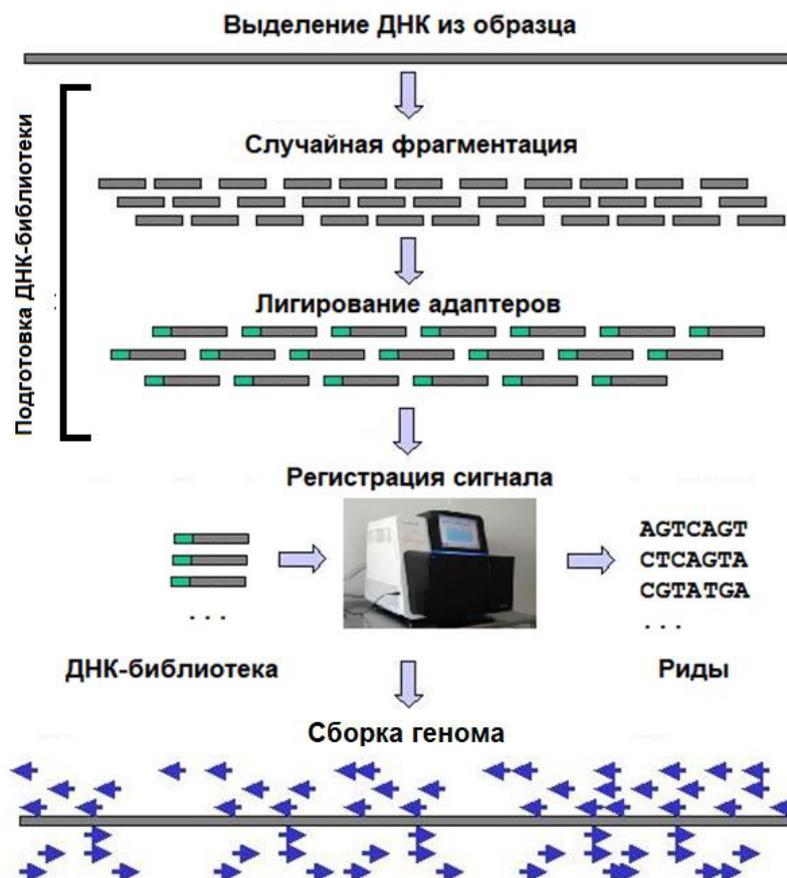


Рис. 1. Основные этапы NGS

Как известно, высокая производительность технологий NGS достигается за счёт массового *параллельного* секвенирования *фрагментов* ДНК (вместо «последовательного» прочтения *целой* молекулы) [1]. Технологии NGS несколько различаются между собой, однако во всех случаях для проведения реакции секвенирования к фрагментам ДНК необходимо присоединение синтетических олигонуклеотидов – *адаптеров*, которые делают возможными манипуляции с молекулами ДНК, например, универсальный адаптер Illumina – AGATCGGAAGAG (https://support.illumina.com/content/dam/illumina-support/help/Illumina_DRAGEN_Bio_IT_Platform_v3_7_1000000141465/Content/SW/Informatics/Dragen/FastQC_Adapter_Kmer_files_fDG.htm). Адаптеры нужны

для прикрепления фрагментов к платформе для секвенирования, также они могут включать в себя так называемые штрих-коды (в случае платформы Illumina – индекс) для идентификации определенных ДНК-фрагментов (например, для различения при загрузке нескольких образцов одновременно). В результате NGS-секвенирования получают массив данных, содержащий огромное количество коротких прочтений фрагментов ДНК вместе с адаптерами – *ридов*.

В связи с этим возникает необходимость восстановления исходной полной последовательности из совокупности ридов, из которых предварительно необходимо удалить последовательности адаптеров. Удаление адаптеров необходимо, так как они могут оставаться в рядах и исказить результаты анализа.

Для сборки и аннотации данные высокопроизводительного секвенирования должны быть представлены в виде двух файлов с прямыми и с обратными рядами (рис.2) в формате FASTQ (*.fastq), которые могут быть упакованы в архивы GZIP (*.fastq.gz).

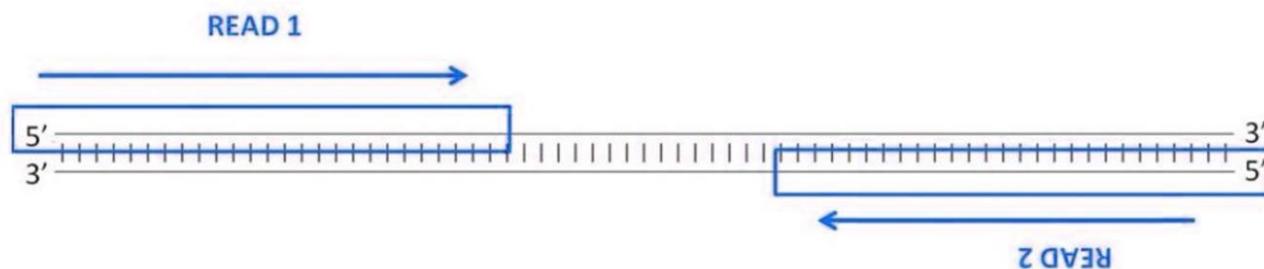


Рис. 2. Прямое (Read 1) и обратное (Read 2) прочтения последовательности

Прямое и обратное прочтения сохраняются в отдельных файлах. Направление прочтений указано в именах файлов – *forward.fastq.gz или *_R1_001.fastq.gz для прямого прочтения и *reverse.fastq.gz или *_R2_001.fastq.gz для обратного прочтения.

В данном пособии в качестве примера используются данные полногеномного секвенирования лактобацилл *Lactobacillus hilgardii* S145 на платформе Illumina MiSeq, которые можно скачать по ссылке (<https://disk.yandex.ru/d/IdsHAMnd8gcagg>):

Lacto_S145_R1_001.fastq.gz (159 МБ)

Lacto_S145_R2_001.fastq.gz (202 МБ)

Сборка генома может осуществляться различными программами и сервисами, один из которых – web-сервис Galaxy [2].

Galaxy – web-сервис с открытым исходным кодом для анализа геномных данных, активно разрабатывается на языке Python с 2005 года командой Galaxy, включающей сотрудников Университета Пенсильвании, Университета Джона Хопкинса, Орегонского университета здоровья и науки и сообщества Galaxy [3]. С каждым годом набор инструментов значительно расширяется и к настоящему времени включает также инструменты для анализа экспрессии генов, протеомики, эпигеномики, транскриптомики и других областей биоинформатики.

Galaxy предоставляет средства для построения многоэтапного вычислительного анализа с дружественным графическим интерфейсом пользователя [4].

Galaxy является платформой для интеграции биологических данных, поддерживает загрузку данных с компьютера пользователя по URL-адресу и напрямую со многих онлайн-ресурсов, таких как UCSC Genome Browser, BioMart и InterMine, поддерживает ряд широко используемых форматов биологических данных и конвертацию между этими форматами.

Galaxy доступен в виде бесплатного общедоступного веб-сервера (www.usegalaxy.org). Пользователи могут создавать учётные записи и сохранять истории, рабочие процессы и наборы данных на сервере.

В качестве программного обеспечения с открытым исходным кодом Galaxy также можно загрузить и установить в виде локальной копии и настроить под конкретные задачи (<https://galaxyproject.org/admin/get-galaxy/>).

2. Протокол сборки генома

Стандартные протоколы сборки широко доступны в Интернете (<https://galaxy-au-training.github.io/tutorials/modules/spades/>, <https://training.galaxyproject.org/training-material/topics/assembly/tutorials/general-introduction/tutorial.html>). Практическая работа по сборке прокариотического генома также доступна в виде цифрового образовательного ресурса на сайте Казанского федерального университета (Биоинформатика, id: 5808).

Протокол сборки генома включает в себя следующие этапы:

- Первичная оценка качества ридов

Первичная оценка включает статистики по ридам: данные о качестве прочтения каждого нуклеотида в риде (низкие значения указывают на высокую вероятность ошибки), процентное содержание каждого из четырех нуклеотидов и неопознанного нуклеотида N, совпадение нуклеотидной последовательности с известными последовательностями адаптеров.

- Предварительная обработка ридов

На этапе предварительной обработки ридов проводят обрезку адаптеров и фильтрацию ридов. В ходе фильтрации удаляются риды низкого качества и риды, слишком короткие после обрезки.

- Повторная оценка качества ридов после обрезки адаптеров и фильтрации ридов

Повторная оценка включает те же параметры, что и первичная оценка, и необходима для контроля эффективности выполненной фильтрации ридов.

- Сборка генома

Сборка генома – процесс объединения коротких ридов в одну или несколько длинных последовательностей (контигов и скаффолдов) для восстановления полногеномной последовательности ДНК.

- Контроль качества сборки полученного генома

После сборки генома проводится контроль качества по статистикам, вычисляемым на основе длин контигов. На качество сборки указывает

непрерывность/целостность сборки. Чем длиннее контиги и меньше их количество, тем выше качество сборки.

- Аннотация генома

Аннотация генома – процесс поиска и описания геномных элементов в последовательности. Аннотация генома включает информацию о генах, кодируемых ими продуктах, и их расположении в аннотируемой последовательности.

- Определение таксономии

После сборки генома определяют его таксономическую принадлежность – идентификация в одном геноме нескольких таксонов свидетельствует о контаминации образца (вместо генома получен метагеном). В таком случае необходимо повторное выделение и секвенирование ДНК.

3. Сборка генома с использованием web-сервиса Galaxy

Внешний вид главной страницы Galaxy (usegalaxy.org) представлен на рис. 3. Как видно из рис. 3, интерфейс Galaxy состоит из трёх частей – списка инструментов в левой части окна (табл. 1), текущей страницы в центре окна (главная страница/параметры выбранного инструмента/предпросмотр данных) и панели «История» в правой части окна, содержащей ссылки на загруженные и вычисленные наборы данных.

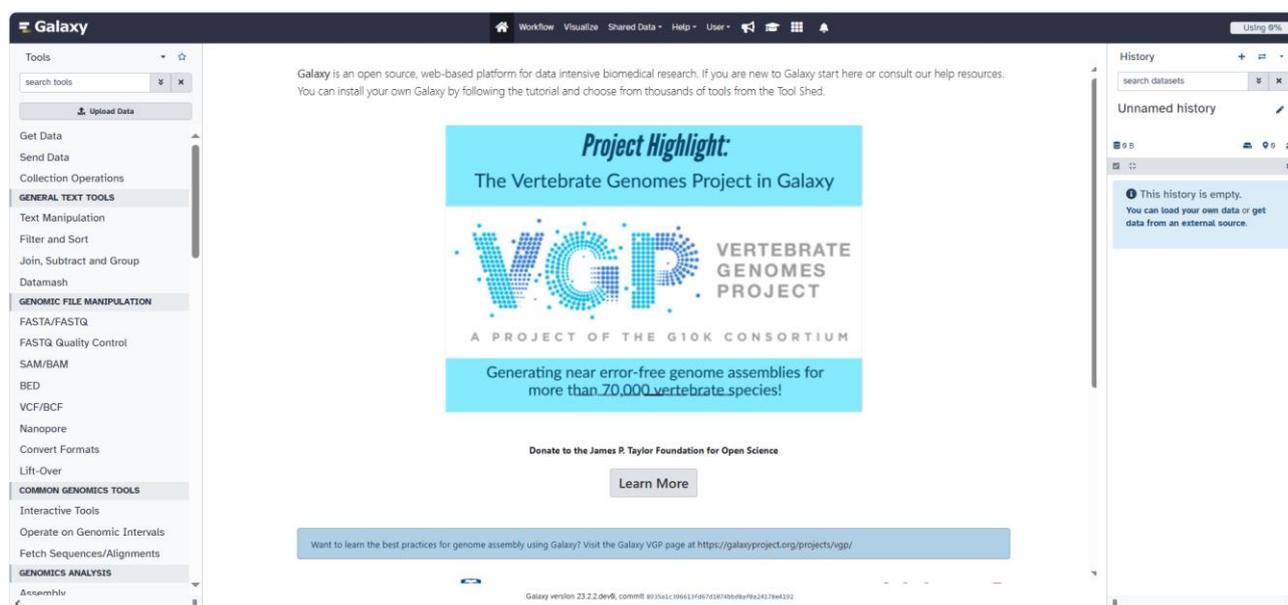


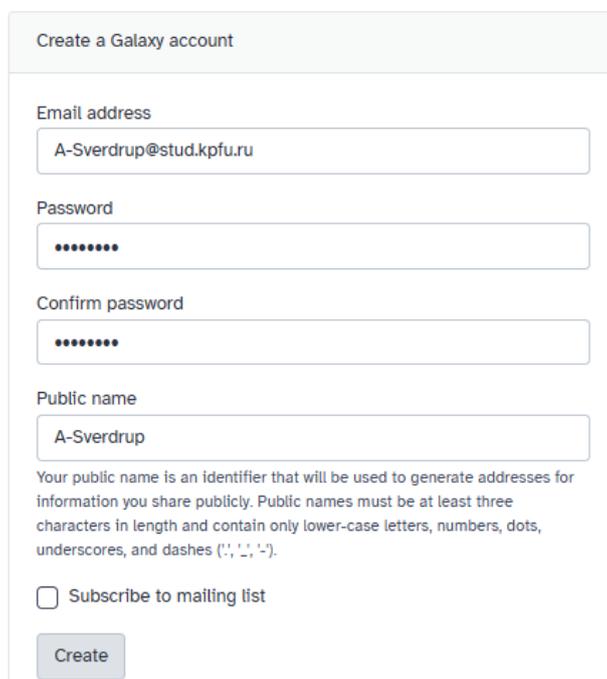
Рис. 3. Главная страница Galaxy

Таблица 1. Основные инструменты для сборки и аннотации генома

Upload data	Загрузка файлов в формате FASTQ
FastQC	Контроль качества секвенирования
Trimmomatic	Обрезка адаптеров и фильтрация ридов
SPAdes	Сборка генома
Quast	Контроль качества сборки генома
Prokka	Аннотация прокариотического генома
RAST	Аннотация прокариотического генома
blastn	Определение таксономии

3.1. Регистрация и загрузка данных

Для работы с web-сервисом Galaxy (usegalaxy.org) необходима регистрация учётной записи (рис. 4). Учётную запись необходимо подтвердить по e-mail.



The image shows a web form titled "Create a Galaxy account". It contains the following fields and elements:

- Email address:** A text input field containing "A-Sverdrup@stud.kpfu.ru".
- Password:** A text input field with masked characters (dots).
- Confirm password:** A text input field with masked characters (dots).
- Public name:** A text input field containing "A-Sverdrup".
- Instructions:** A paragraph of text explaining that the public name is an identifier used for generating addresses for public information, with requirements for length and character set (lower-case letters, numbers, dots, underscores, and dashes).
- Subscription:** A checkbox labeled "Subscribe to mailing list" which is currently unchecked.
- Create:** A button at the bottom of the form.

Рис. 4. Форма регистрации

После регистрации и входа в систему для загрузки данных нажмите кнопку *Upload Data*, расположенную под строкой поиска панели инструментов (рис. 5).

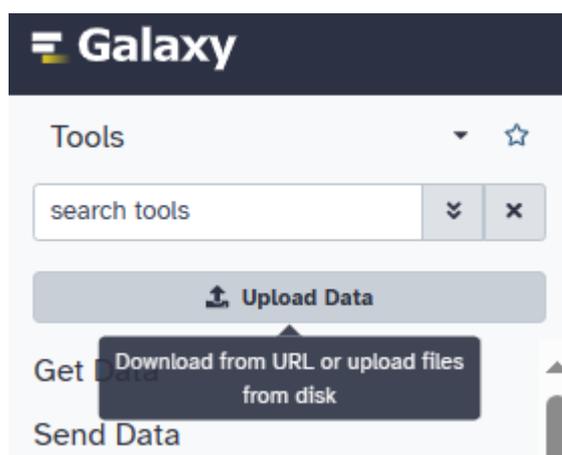


Рис. 5. Выбор команды *Upload data*

В окне *Upload from Disk or Web* нажмите кнопку *Choose local file* и выберите два файла для загрузки в формате FASTQ:

Lacto_S145_R1_001.fastq.gz (159 МБ);

Lacto_S145_R2_001.fastq.gz (203 МБ).

Проверьте правильность выбора и нажмите кнопку *Start* для загрузки файлов на сервер (рис. 6).

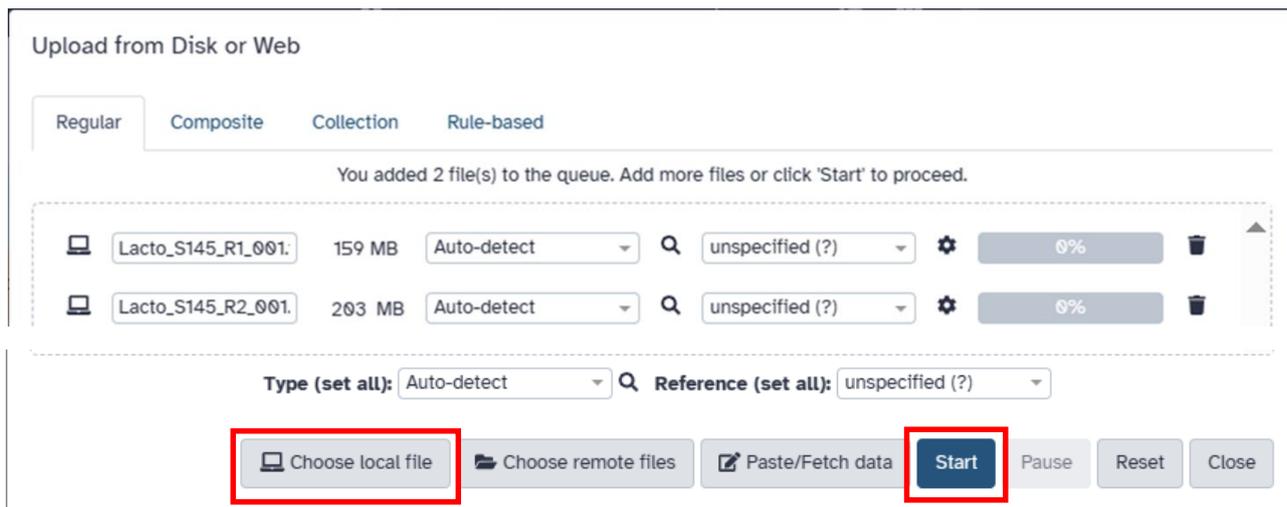


Рис. 6. Загрузка файлов на сервер

По окончании загрузки нажмите кнопку *Close*, загруженные файлы появятся на панели истории (рис. 7).

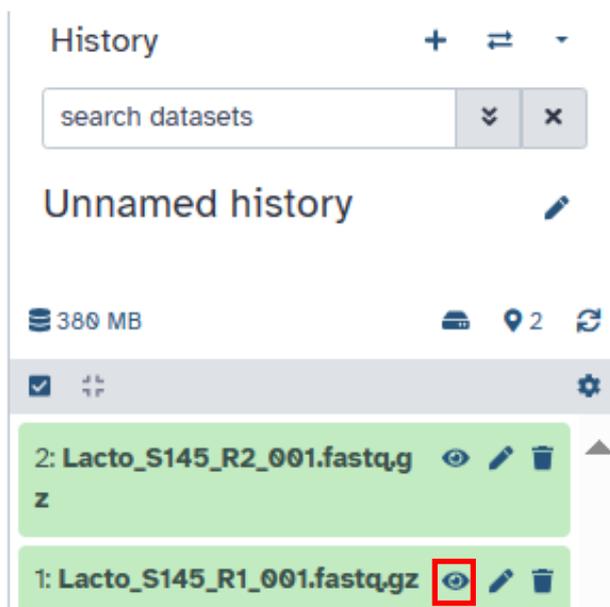


Рис. 7. Загруженные файлы на панели истории

Само значение качества Phred Quality Score рассчитывается как отрицательный десятичный логарифм вероятности ошибки при прочтении, умноженный на 10 (рис.10). Например, значение Phred Quality Score, равное 30, соответствует вероятности ошибки $10^{-\frac{30}{10}} = 0,001 = 0,1\%$ и точности прочтения 99,9%.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Рис.10. Соответствие Phred Quality Score, вероятности ошибки и точности прочтения

3.2. Контроль качества секвенирования

Для контроля качества секвенирования на панели инструментов в разделе **GENOMIC FILE MANIPULATION** нажмите **FASTQ Quality Control** и в выпадающем списке выберите **FastQC** (рис. 11). Так как **FastQC** принимает только один входной файл, описанный алгоритм необходимо повторить отдельно для прямых ридов и для обратных ридов.

3.2.1. Оценка качества прямых ридов

Для оценки качества прямых ридов в поле **Raw read data from your current history** выберите загруженный файл с прямыми ридами (Lacto_S145_R1_001.fastq.gz). Оставьте остальные настройки по умолчанию и нажмите ► **Run Tool**.

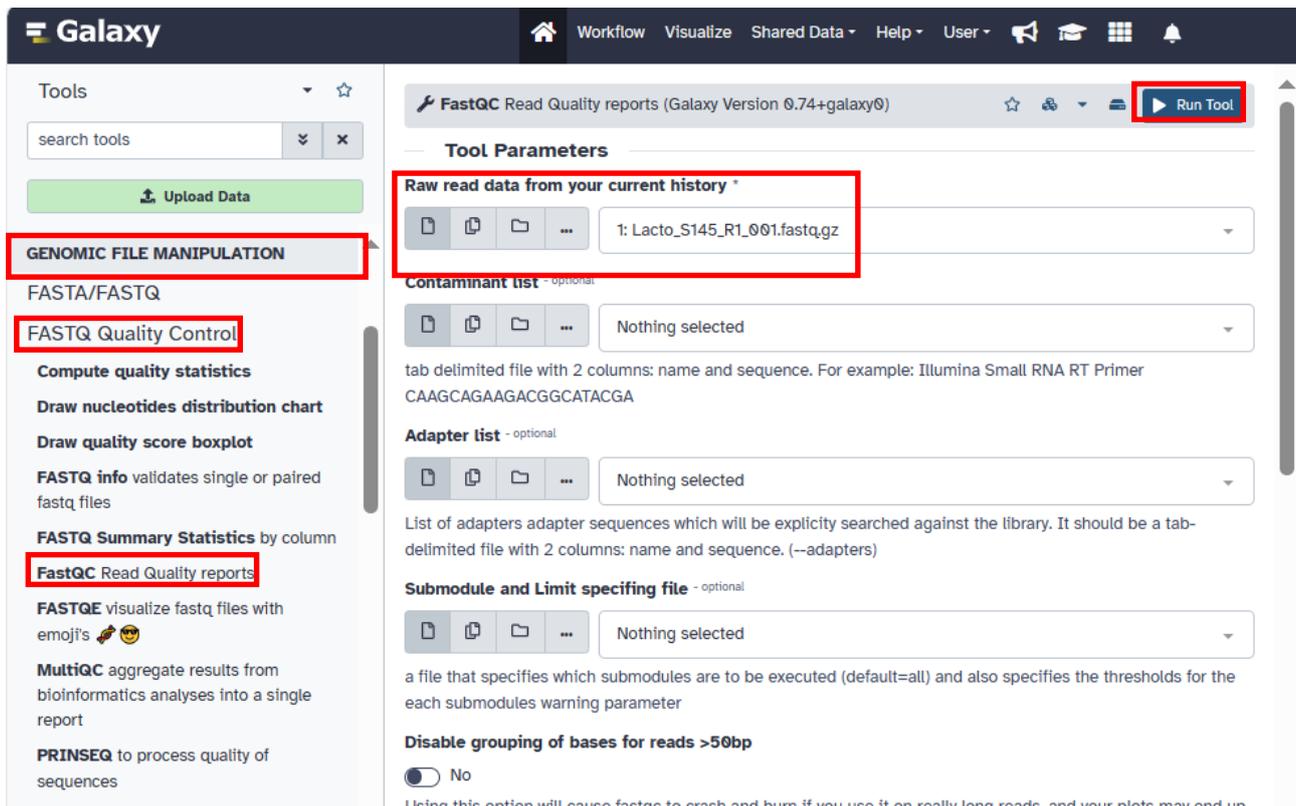


Рис. 11. FastQC в панели инструментов и параметры выполнения

Процесс выполнения будет отображён в панели истории (рис. 12).

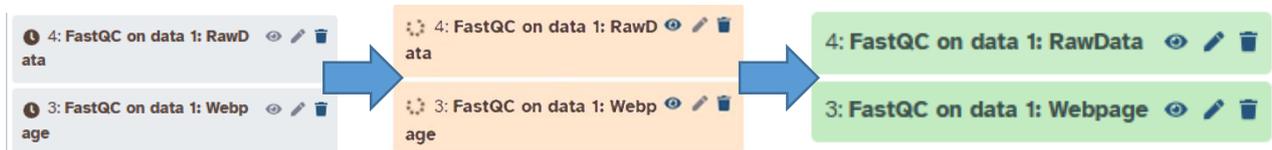


Рис. 12. Процесс выполнения: в очереди (слева), вычисление (в центре), результаты (справа)

Для просмотра результатов нажмите кнопку  для файла *FastQC on data №...: Webpage* (рис. 13).

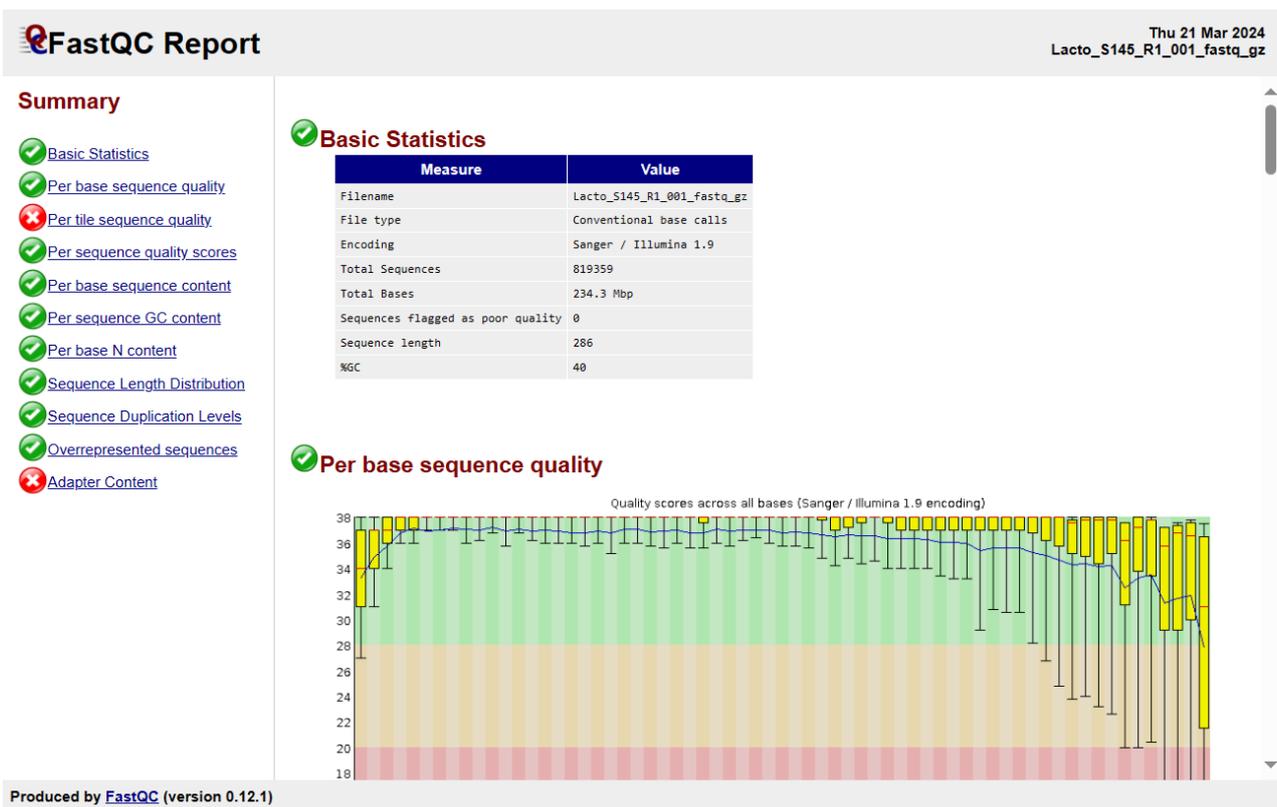


Рис. 13. Отчёт по качеству секвенирования *FastQC* для
Lacto_S145_R1_001.fastq.gz

Как видно из рис. 13, программа *FastQC* проводит оценку качества секвенирования по целому ряду параметров. Особое внимание стоит уделить *Per base sequence quality*, *Per base N content*, *Per base sequence content*, *Overrepresented sequences* и *Adapter content*.

Per base sequence quality (рис. 14): на графике по оси X откладывается позиция нуклеотида в ряде, по оси Y – статистическая оценка качества:

- среднее значение – синяя ломаная;
- медиана – красные засечки;
- интерквартиль (50% выборки) – желтые прямоугольники;
- предельные значения (80% выборки) – черные линии с засечками на концах.

На графике качество ридов разделено на три зоны:

- >28 (зелёная зона) – высокое качество,
- 20-28 (жёлтая зона) – приемлемое качество, необходима фильтрация/обрезка ридов,
- <20 (красная зона) – низкое качество, необходимо провести эксперимент заново (выделение ДНК, подготовку ДНК-библиотеки, секвенирование).

«Золотым стандартом» качества для платформы Illumina принято считать значения качества выше 30.

На качество прочтений может влиять множество факторов, от качества самих библиотек (свежести реактивов, контаминации библиотек, прочтение похожих друг на друга последовательностей), до оптики прибора (не распознает цвет свечения нуклеотида), а также наличие артефактов – ошибок, вызванных неправильными молекулярными взаимодействиями, например, образованием димеров из адаптеров.

Сборка генома из ридов низкого качества приводит к получению слишком коротких контигов, непригодных для дальнейшего анализа/аннотации, что может привести к получению недостоверных либо неправильных результатов. Например, типичный белок содержит около 350 аминокислот, что соответствует нуклеотидной последовательности гена длиной около 1 тыс.п.н. – покрываемой 4 ридами длиной 300 п.н. или 7 ридами длиной 150 п.н. Если эти риды не окажутся объединены в один контиг, то данный ген не будет обнаружен при аннотации, из чего можно сделать неправильный вывод.

Как видно из рис. 14, в нашем примере бóльшая часть данных попадает в зелёную зону. Снижение качества начинается с позиции 220 и является приемлемым до позиции 260, в самом конце рида (позиции 260-286) качество критическое, что говорит об необходимости **тримминга** (обрезка и фильтрация ридов) (раздел 3.3).

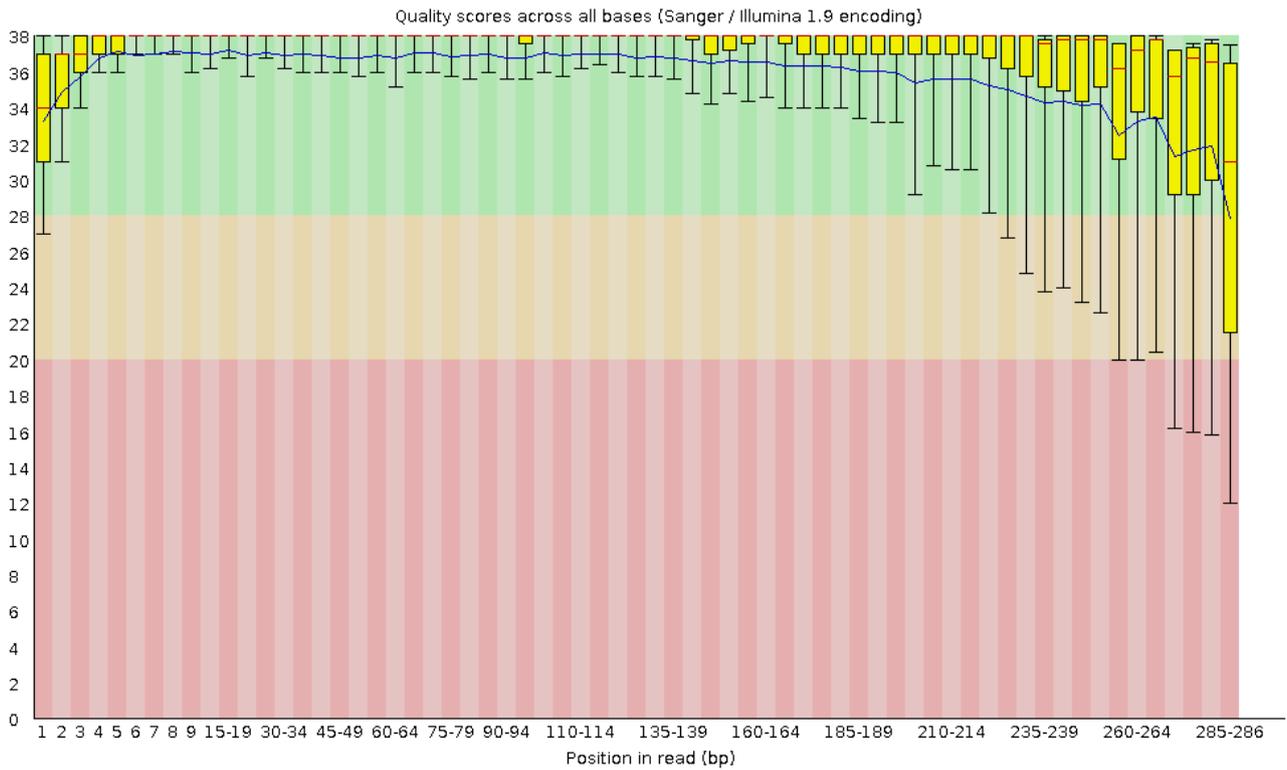


Рис. 14. *Per base sequence quality* для Lacto_S145_R1_001.fastq.gz

Per base N content (рис. 15) указывает на количество неизвестного нуклеотида N в каждой позиции. Наличие неопознанных нуклеотидов в последовательности нежелательно и указывает на ошибки секвенирования.

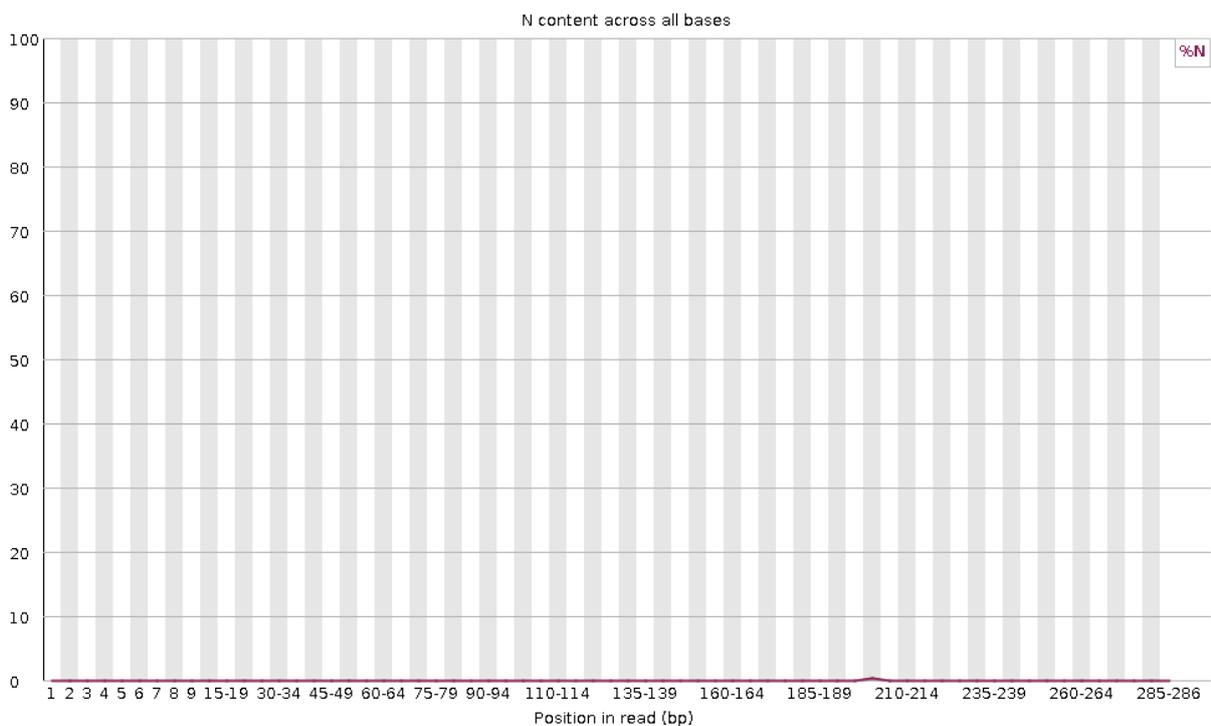


Рис. 15. *Per base N content* для Lacto_S145_R1_001.fastq.gz

Как видно из рис. 15, в нашем примере неопознанные нуклеотиды полностью отсутствуют – красная линия графика совпадает с осью X.

Per base sequence content (рис. 16) указывает на долю каждого нуклеотида (А, Т, С, G) в каждой позиции ряда в процентах.

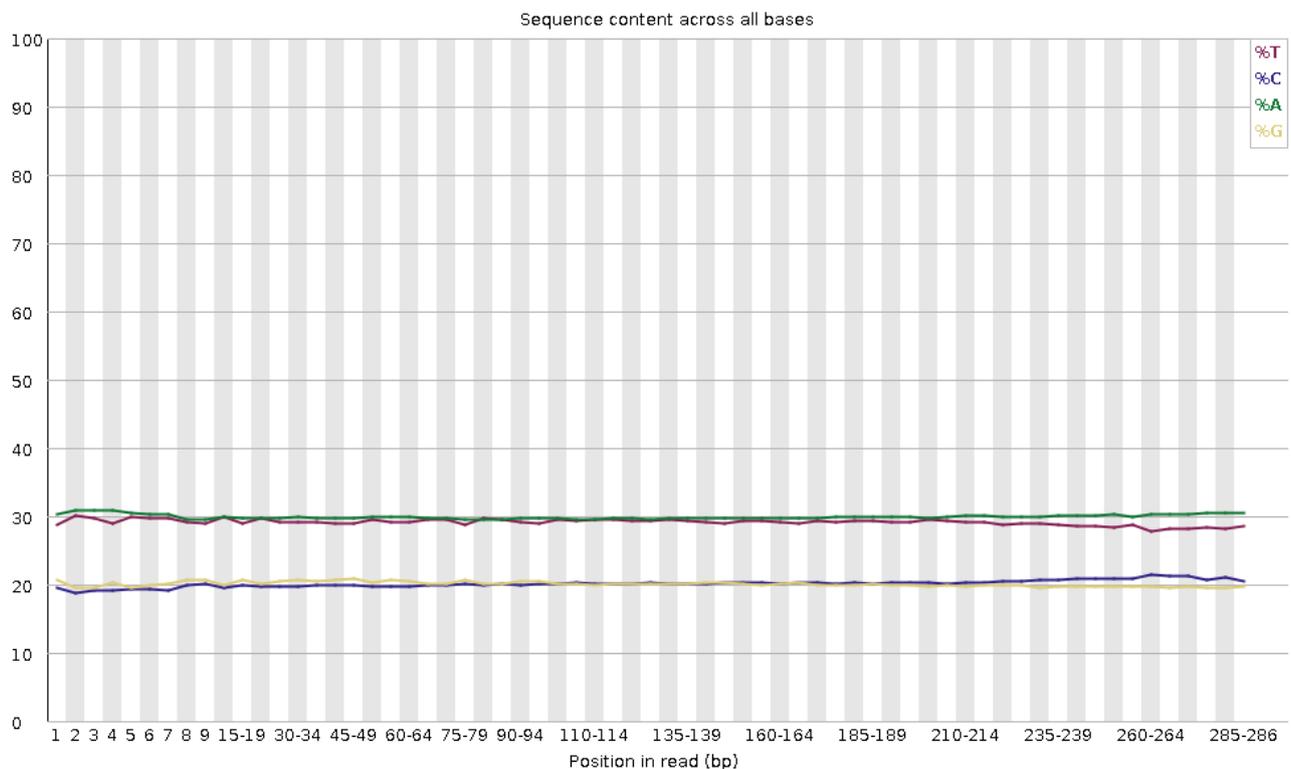


Рис. 16. *Per base sequence content* для Lacto_S145_R1_001.fastq.gz

Overrepresented sequences (рис. 17) представляет в табличном виде чрезвычайно часто встречающиеся в рядах последовательности, содержание которых, как правило, связано с наличием артефактов. Отсутствие информации в данном разделе указывает на хороший результат.



Рис. 17. *Overrepresented sequences* для Lacto_S145_R1_001.fastq.gz

Adapter content (рис. 18) представляет в графическом виде содержание адаптеров в последовательности.

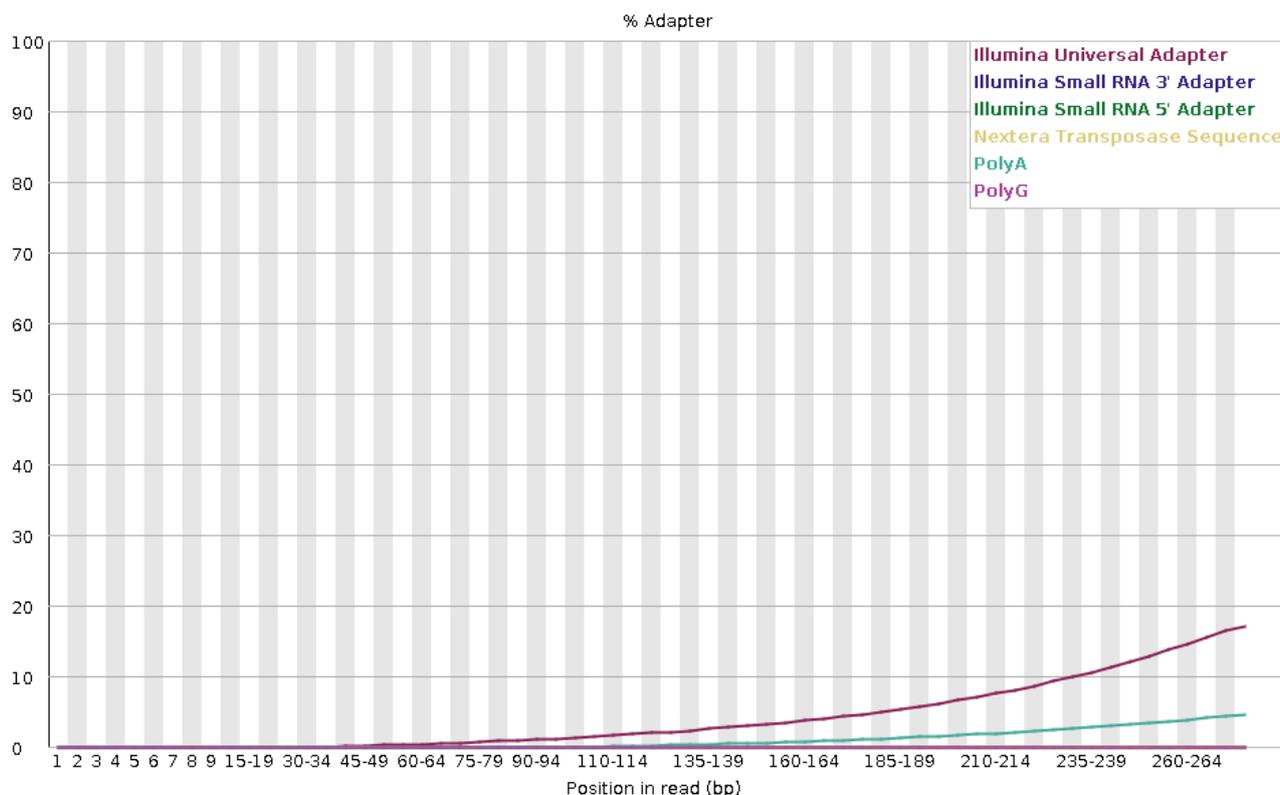


Рис. 18. *Adapter content* для Lacto_S145_R1_001.fastq.gz

Как видно из рис. 18, в рядах Lacto_S145_R1_001.fastq.gz обнаружены адаптеры (Illumina Universal Adapter – бордовая линия на графике), что свидетельствует о необходимости **тримминга** (обрезки адаптеров и фильтрации рядов).

3.2.2. Оценка качества обратных ридов

Для оценки качества обратных ридов в поле *Raw read data from your current history* выберите загруженный файл с обратными рядами (Lacto_S145_R2_001.fastq.gz). Оставьте остальные настройки по умолчанию и нажмите ► *Run Tool*.

Ниже на рис.19-23 представлены результаты анализа. Как видно из рис. 19, качество обратных ридов (Lacto_S145_R2_001.fastq.gz) значительно ниже, чем для прямых ридов (Lacto_S145_R1_001.fastq.gz) что говорит

об необходимости **тримминга** (обрезка и фильтрация ридов) (раздел 3.3). Доли нуклеотидов в последовательности, количество неопознанных нуклеотидов (0), отсутствие артефактов, содержание адаптеров (Illumina Universal Adapter), совпадает для прямых и обратных ридов (рис. 20-23).

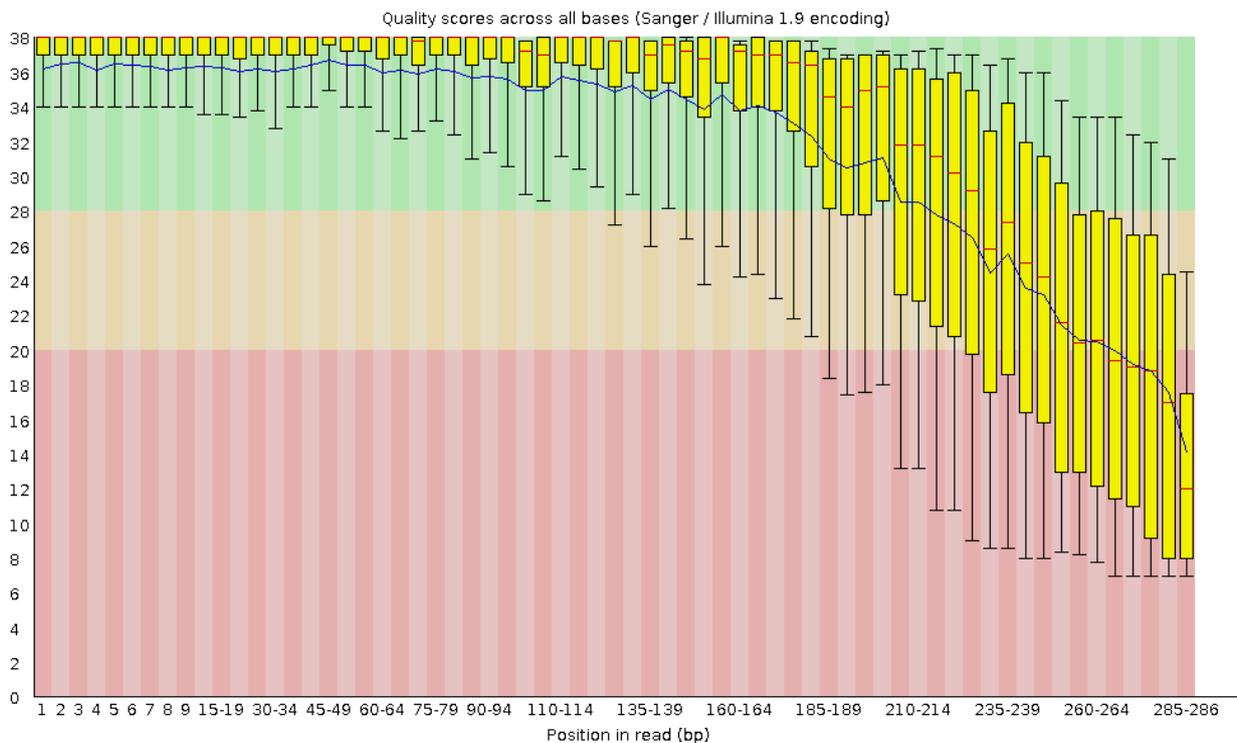


Рис. 19. *Per base sequence quality* для Lacto_S145_R2_001.fastq.gz

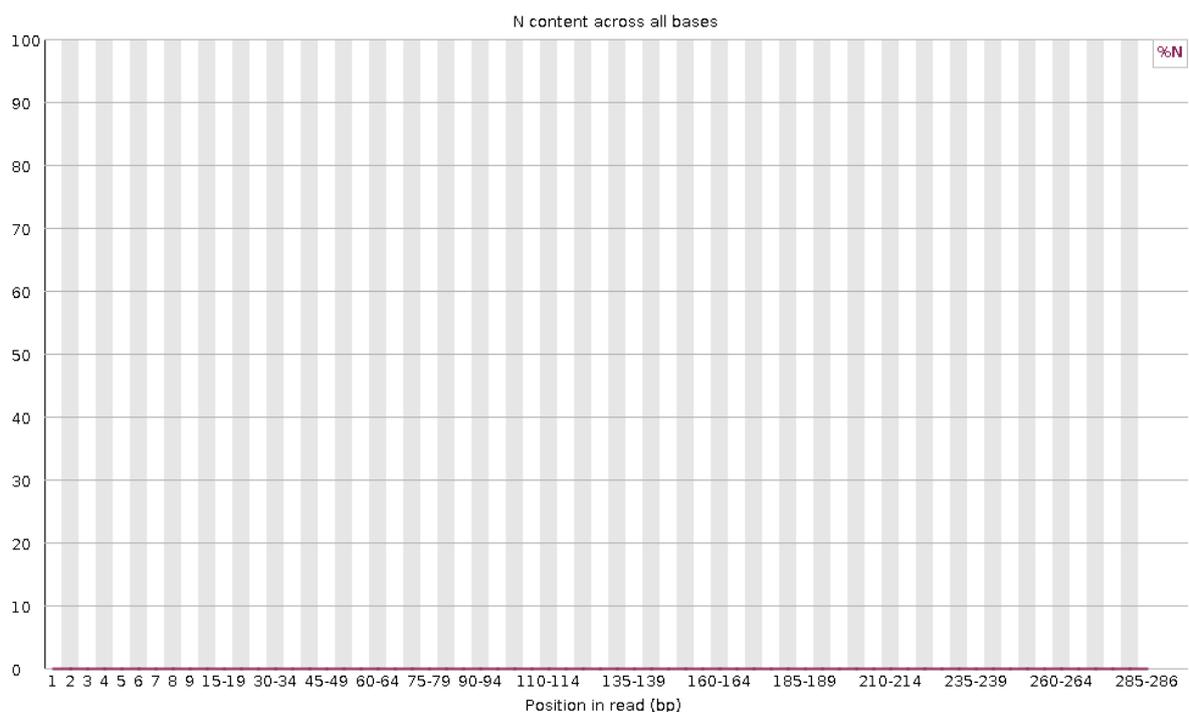


Рис. 20. *Per base N content* для Lacto_S145_R2_001.fastq.gz

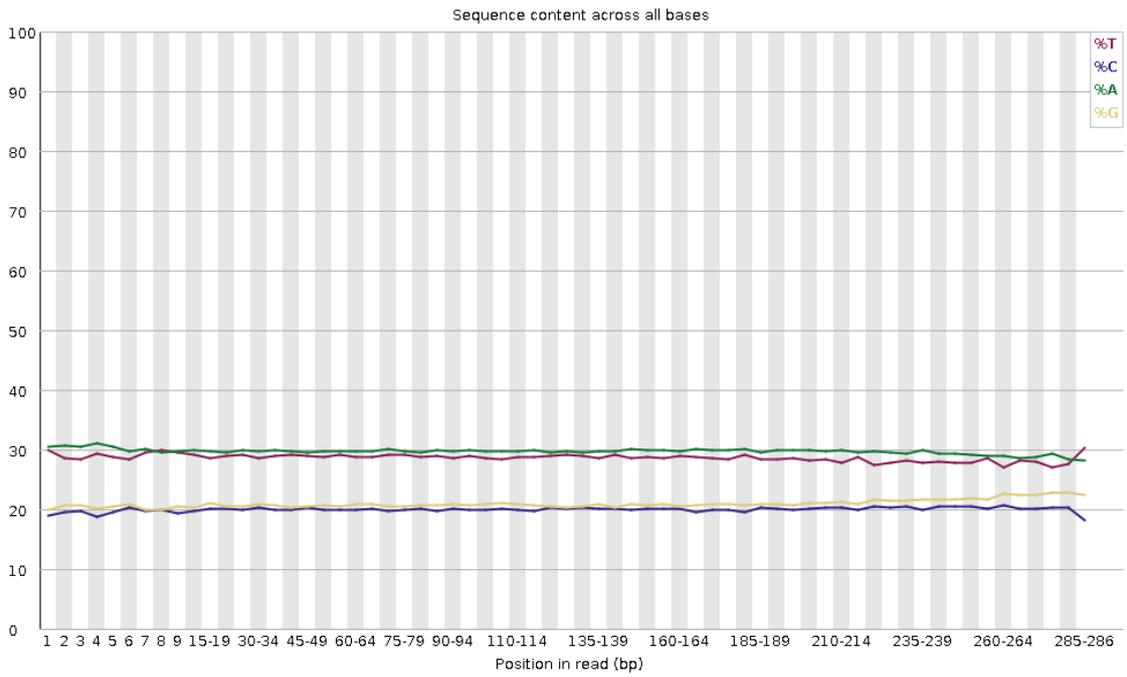


Рис. 21. *Per base sequence content* для Lacto_S145_R2_001.fastq.gz



Рис. 22. *Overrepresented sequences* для Lacto_S145_R2_001.fastq.gz

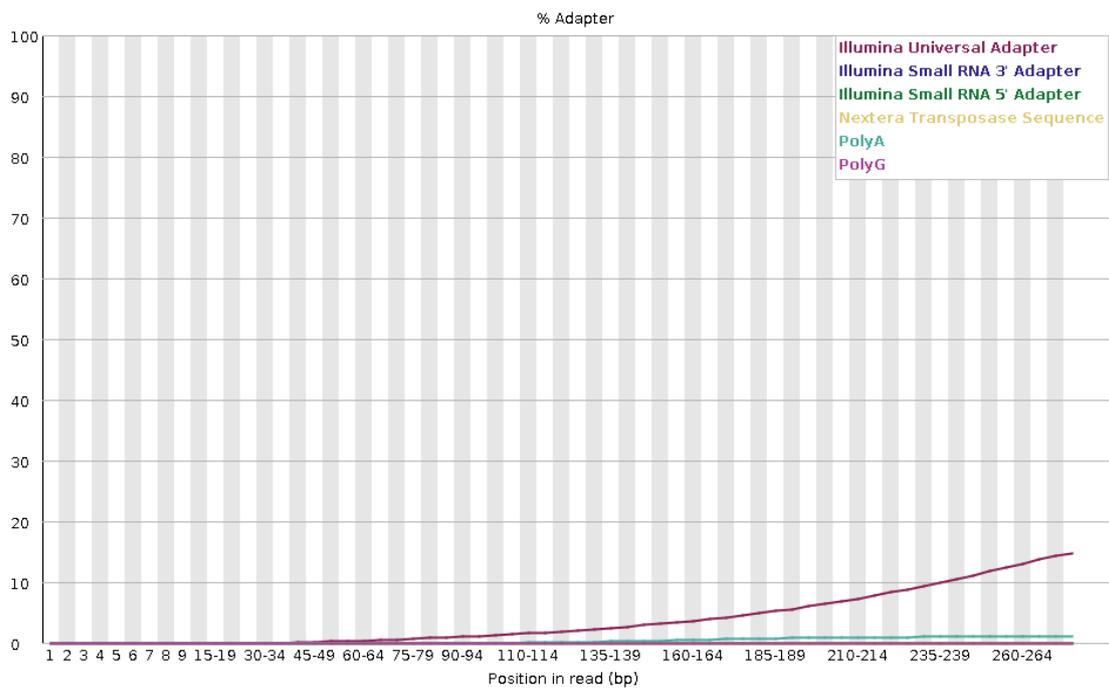


Рис. 23. *Adapter content* для Lacto_S145_R2_001.fastq.gz

3.3. Тримминг

Для проведения тримминга на панели инструментов в разделе **GENOMIC FILE MANIPULATION** нажмите **FASTQ Quality Control** и в выпадающем списке выберите **Trimmomatic** (рис. 24).

В пункте **Single-end or Paired-end reads?** выберите вариант **Paired-end (Two separate input files)**. В поле **Input FASTQ file (R1/first of pair)** выберите загруженный файл с прямыми ридами (`Lacto_S145_R1_001.fastq.gz`), в поле **Input FASTQ file (R2/second of pair)** – файл с обратными ридами (`Lacto_S145_R2_001.fastq.gz`).

Включите переключатель **Perform initial ILLUMINACLIP step?** (Yes/No), затем в пункте **Select standard adapter sequences or provide custom?** выберите вариант **Custom**. В появившемся поле **Custom adapter sequences in fasta format** введите последовательность Illumina Universal Adapter – `AGATCGGAAGAG`. В поле **How accurate the match between any adapter etc. sequence must be against a read *** введите 8.

В разделе **Trimmomatic Operation** настройте операцию **1: Trimmomatic Operation** (табл.2). Добавьте еще 3 операции кнопкой + **Insert Trimmomatic operation** и настройте их в соответствии с параметрами (табл.2).

Оставьте остальные настройки по умолчанию и нажмите ► **Run Tool**. Прогресс выполнения и результаты появятся в панели истории (рис. 25).

Для просмотра обрезанных ридов нажмите кнопку  (рис. 25). По умолчанию выводится только первый 1 МБ данных (рис. 26).

Таблица 2. Параметры Trimmomatic Operation
<http://www.usadellab.org/cms/?page=trimmomatic>

1: Trimmomatic Operation	
Select Trimmomatic operation to perform:	<i>Cut bases off the start of a read, if below a threshold quality (LEADING),</i>
Minimum quality required to keep a base *:	3
2: Trimmomatic Operation	
Select Trimmomatic operation to perform:	<i>Cut bases off the end of a read, if below a threshold quality (TRAILING),</i>
Minimum quality required to keep a base *:	3
3: Trimmomatic Operation	
Select Trimmomatic operation to perform:	<i>Sliding window trimming (SLIDINGWINDOW)</i>
Number of bases to average across *:	4
Average quality required *:	15
4: Trimmomatic Operation	
Select Trimmomatic operation to perform:	<i>Drop reads below a specified length (MINLEN)</i>
Minimum length of reads to be kept *:	36

Galaxy

Workflow Visualize

Tools

search tools

Upload Data

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

Compute quality statistics

Draw nucleotides distribution chart

Draw quality score boxplot

FASTQ info validates single or paired fastq files

FASTQ Summary Statistics by column

FastQC Read Quality reports

FASTQE visualize fastq files with emoji's 🍌🍌

MultiQC aggregate results from bioinformatics analyses into a single report

PRINSEQ to process quality of sequences

Trimmomatic flexible read trimming tool for Illumina NGS data

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

Lift-Over

COMMON GENOMICS TOOLS

Interactive Tools

Operate on Genomic Intervals

Fetch Sequences/Alignments

GENOMICS ANALYSIS

Assembly

Annotation

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.38.1)

Tool Parameters

Single-end or paired-end reads?

Paired-end (two separate input files)

Input FASTQ file (R1/first of pair) *

1: Lacto_S145_R1_001.fastq.gz

accepted formats

Input FASTQ file (R2/second of pair) *

1: Lacto_S145_R1_001.fastq.gz

accepted formats

Perform initial ILLUMINA CLIP step?

Yes

Cut adapter and other illumina-specific sequences from the read

Select standard adapter sequences or provide custom?

Custom

Custom adapter sequences in fasta format - optional

AGATCGGAAGAG

Write sequences in the fasta format.

Maximum mismatch count which will still allow a full match to be performed *

2

How accurate the match between the two "adapter ligated" reads must be for PE palindrome read alignment *

30

How accurate the match between any adapter etc. sequence must be against a read *

8

Minimum length of adapter that needs to be detected (PE specific/palindrome mode) *

8

Always keep both reads (PE specific/palindrome mode)?

Yes

See help below

Рис. 24. *Trimmomatic* в панели инструментов и настройки выполнения

Galaxy

Workflow Visualize

Tools

search tools

Upload Data

Draw quality score boxplot

FASTQ info validates single or paired fastq files

FASTQ Summary Statistics by column

FastQC Read Quality reports

FASTQE visualize fastq files with emoji's 🍌🍌

MultiQC aggregate results from bioinformatics analyses into a single report

PRINSEQ to process quality of sequences

Trimmomatic flexible read trimming tool for Illumina NGS data

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

Lift-Over

COMMON GENOMICS TOOLS

Interactive Tools

Operate on Genomic Intervals

Fetch Sequences/Alignments

GENOMICS ANALYSIS

Assembly

Annotation

Mapping

Variant Calling

ChIP-seq

RNA-seq

GENOMICS TOOLKITS

METAGENOMICS

DEPRECATED TOOLS

Multiple Alignments

Phenotype Association

Evolution

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.38.1)

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Cut bases off the start of a read, if below a threshold quality (LEADING)

Minimum quality required to keep a base *

3

Bases at the start of the read with quality below the threshold will be removed

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Cut bases off the end of a read, if below a threshold quality (TRAILING)

Minimum quality required to keep a base *

3

Bases at the end of the read with quality below the threshold will be removed

3: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across *

4

Average quality required *

15

4: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept *

36

+ Insert Trimmomatic Operation

Рис. 24. Продолжение настройки *Trimmomatic*

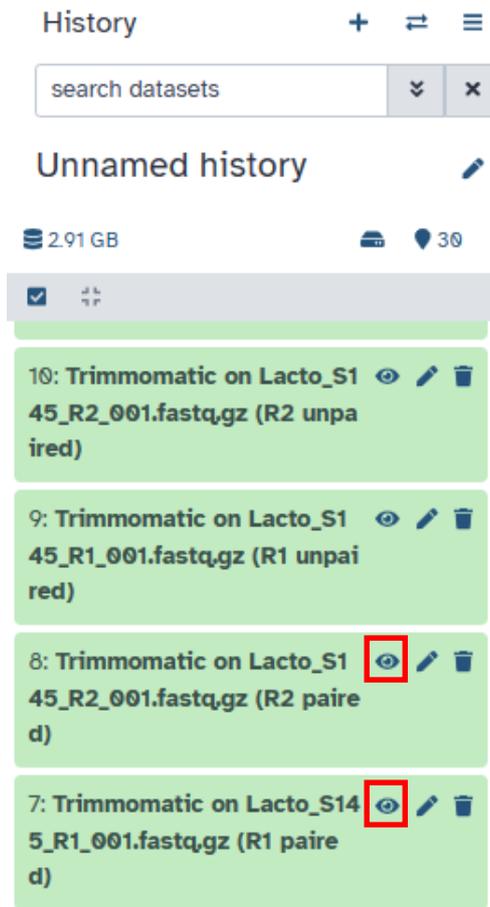


Рис. 25. Результаты тримминга: *R1 paired* – обработанные прямые риды, *R2 paired* – обработанные обратные риды, *R1 unpaired* и *R2 unpaired* – удаленные прямые и обратные риды соответственно

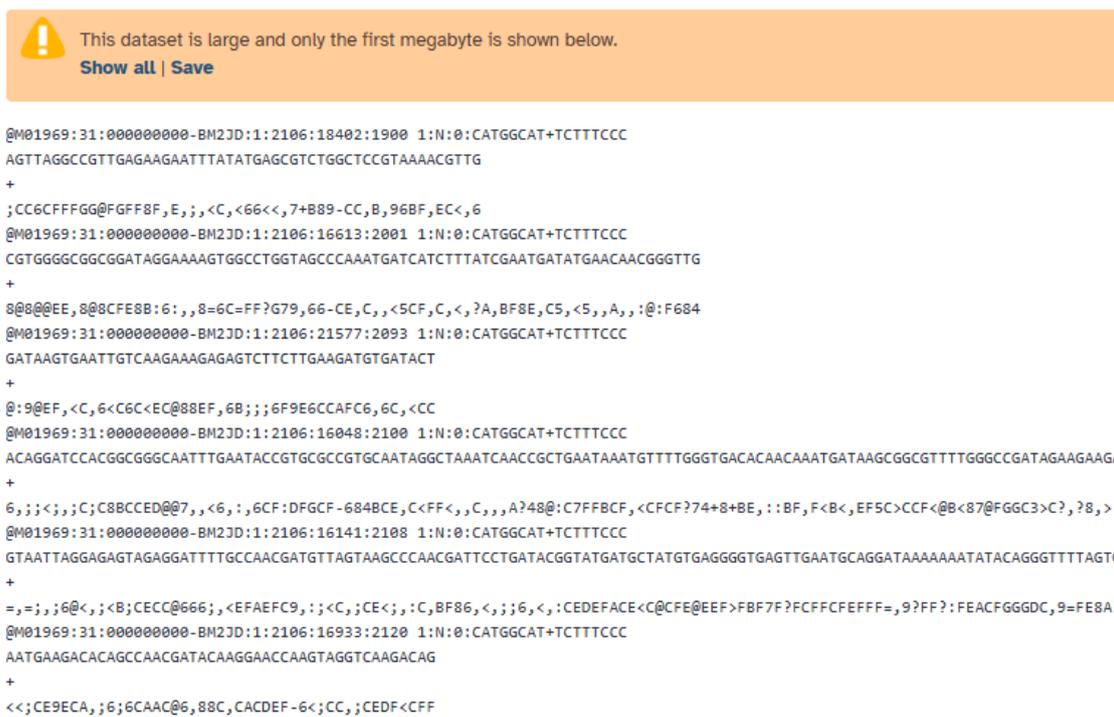


Рис. 26. Просмотр результатов тримминга

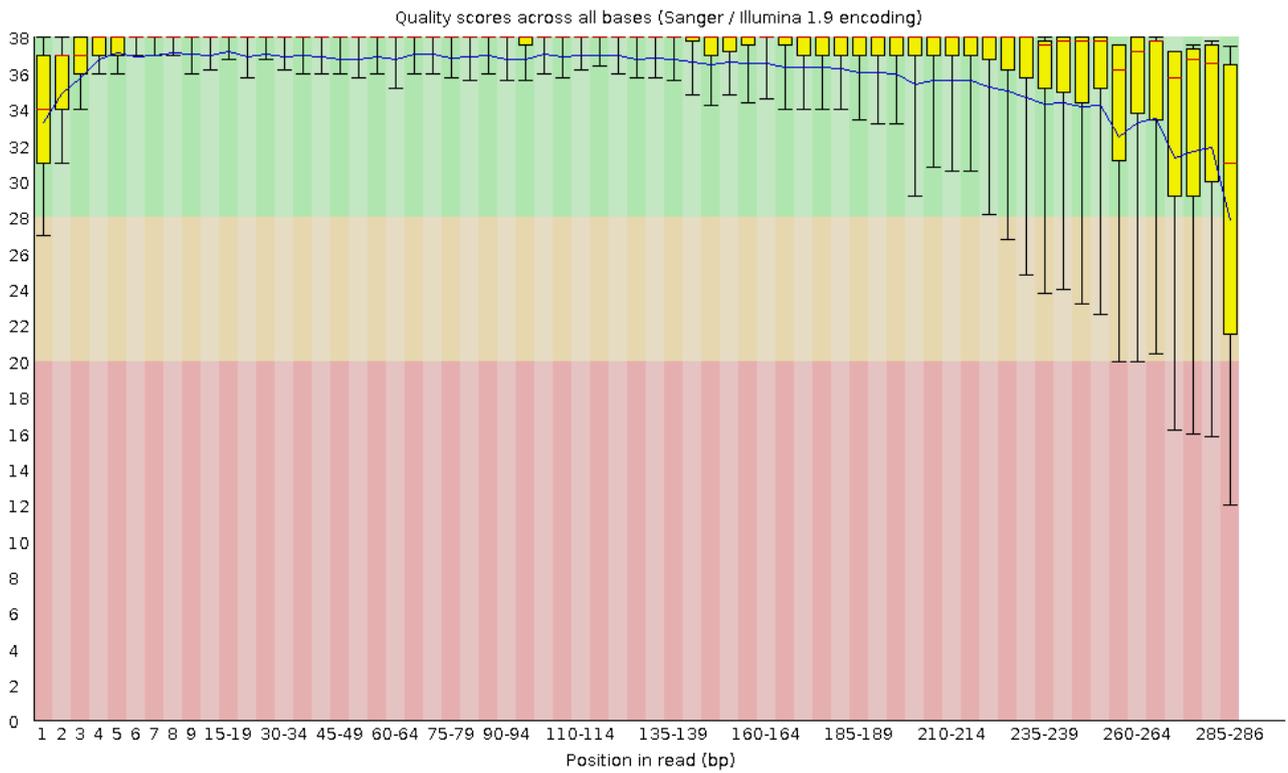
3.4. Контроль качества ридов после тримминга

Для контроля качества ридов после тримминга на панели инструментов в разделе **GENOMIC FILE MANIPULATION** нажмите **FASTQ Quality Control** и в выпадающем списке выберите **FastQC**.

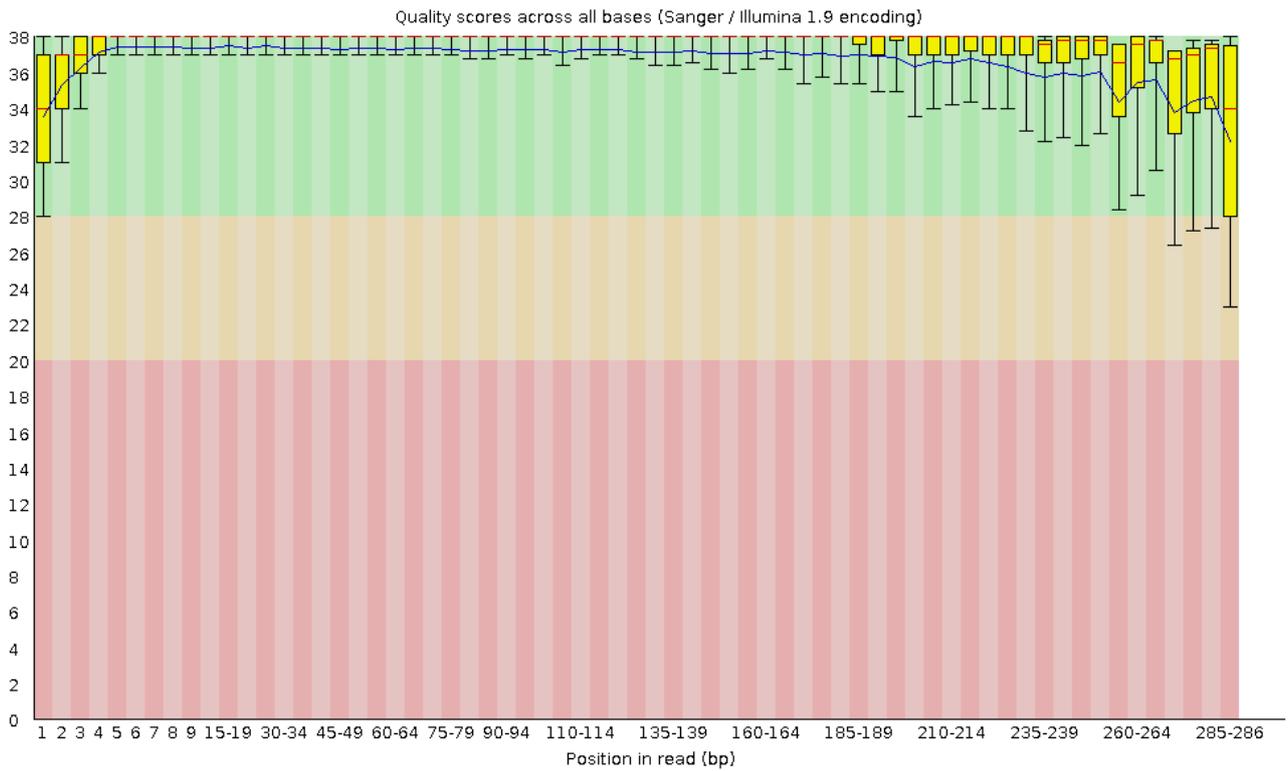
В поле **Raw read data from your current history** выберите с прямыми ридами **Trimmomatic on Lacto_S145_R1_001.fastq.gz** и нажмите ► **Run Tool**, затем выполните программу ещё раз с обратными ридами: **Trimmomatic on Lacto_S145_R2_001.fastq.gz: R2 paired**

Ниже приведено сравнение оценки качества ридов до и после тримминга для прямых (рис. 27-28) и обратных (рис. 29-30) ридов.

После удаления адаптеров качество ридов повысилось (рис. 27, 29), содержание адаптеров, как и ожидалось, упало до нуля (рис. 28, 30).



a)

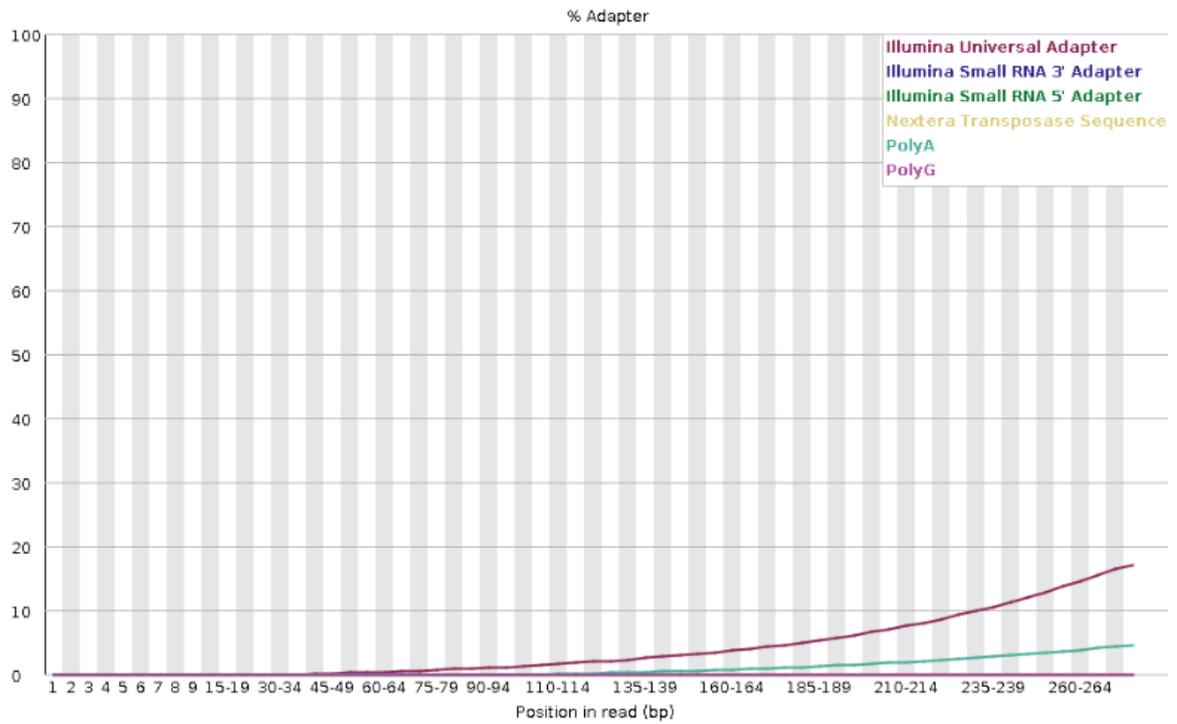


б)

Рис. 27. *Per base sequence quality* для Lacto_S145_R1_001.fastq.gz

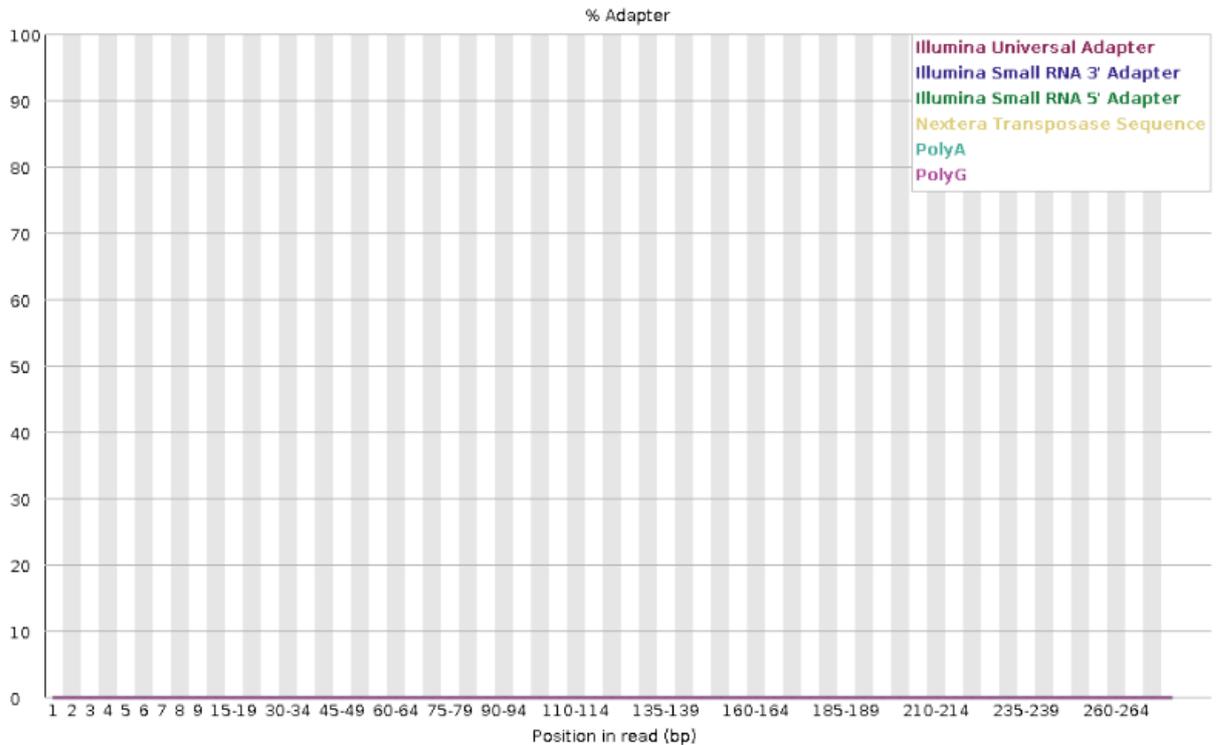
а) до тримминга, б) после тримминга

✘ Adapter Content



а)

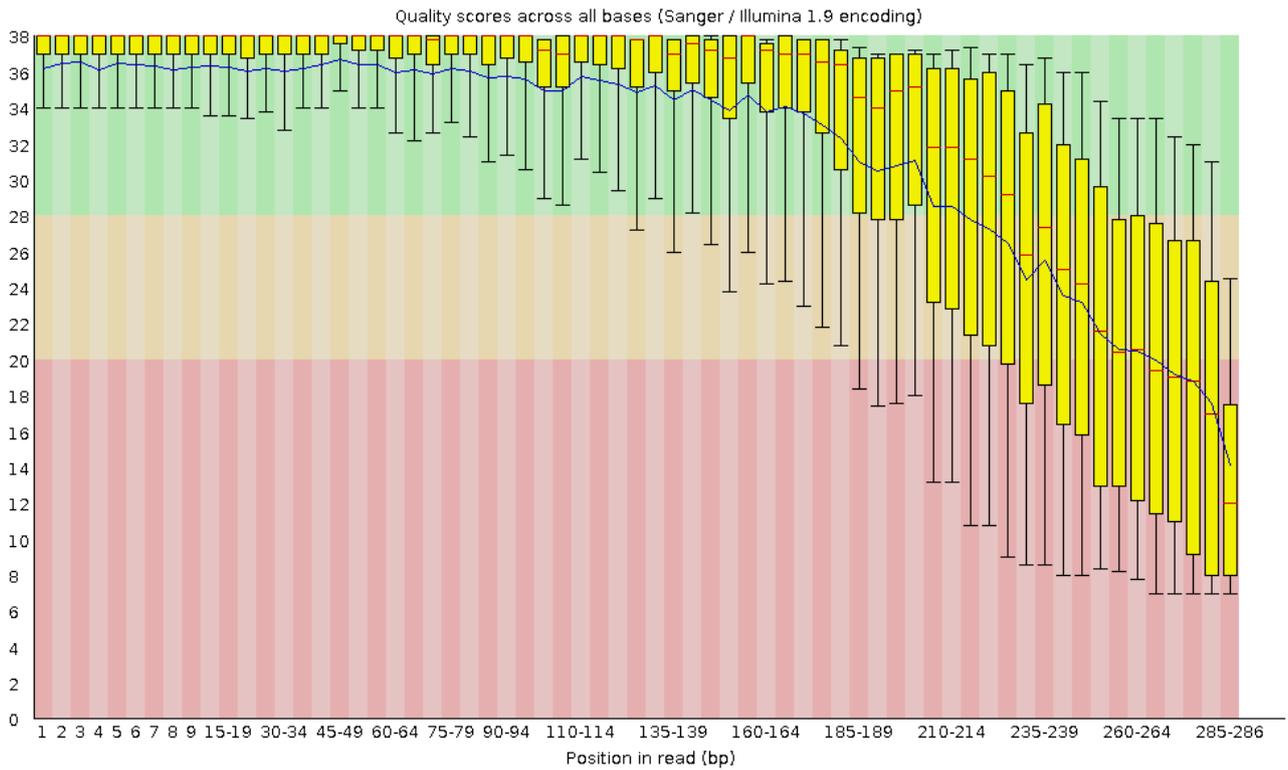
✔ Adapter Content



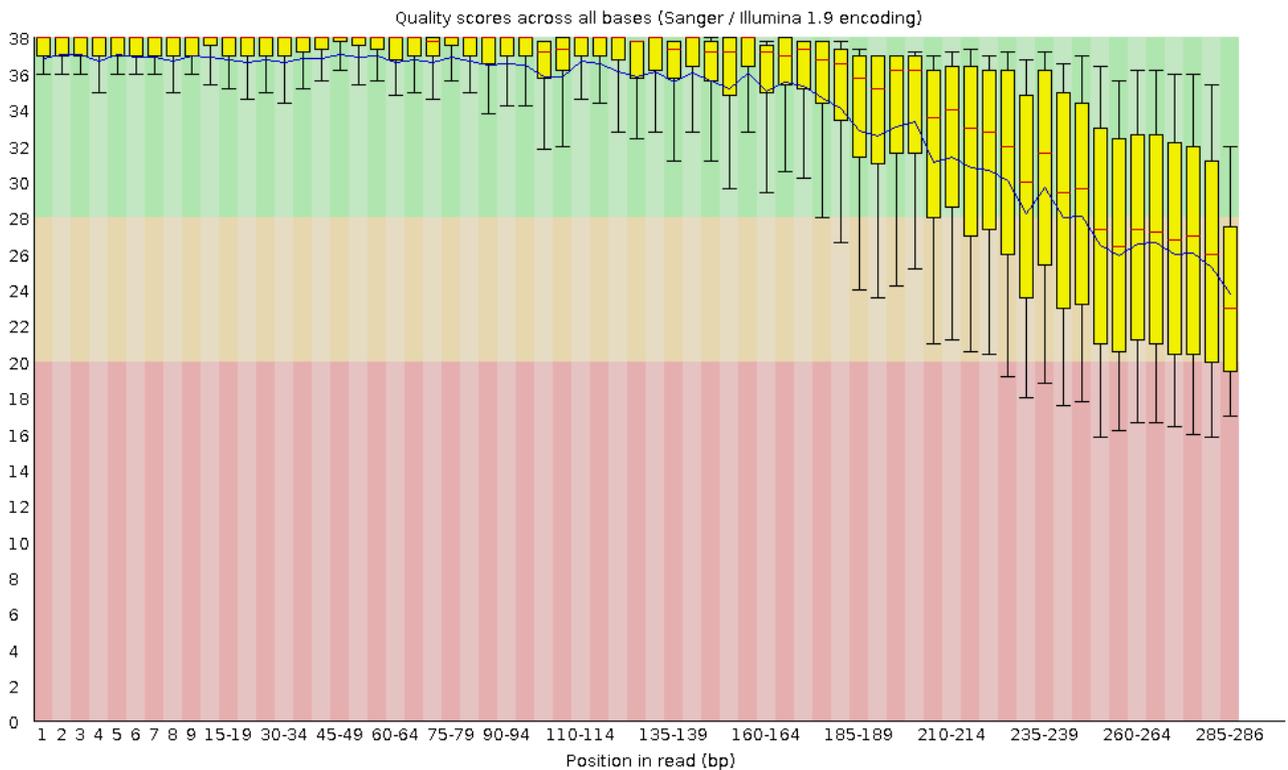
б)

Рис. 28. *Adapter content* для Lacto_S145_R1_001.fastq.gz

а) до тримминга, б) после тримминга



a)

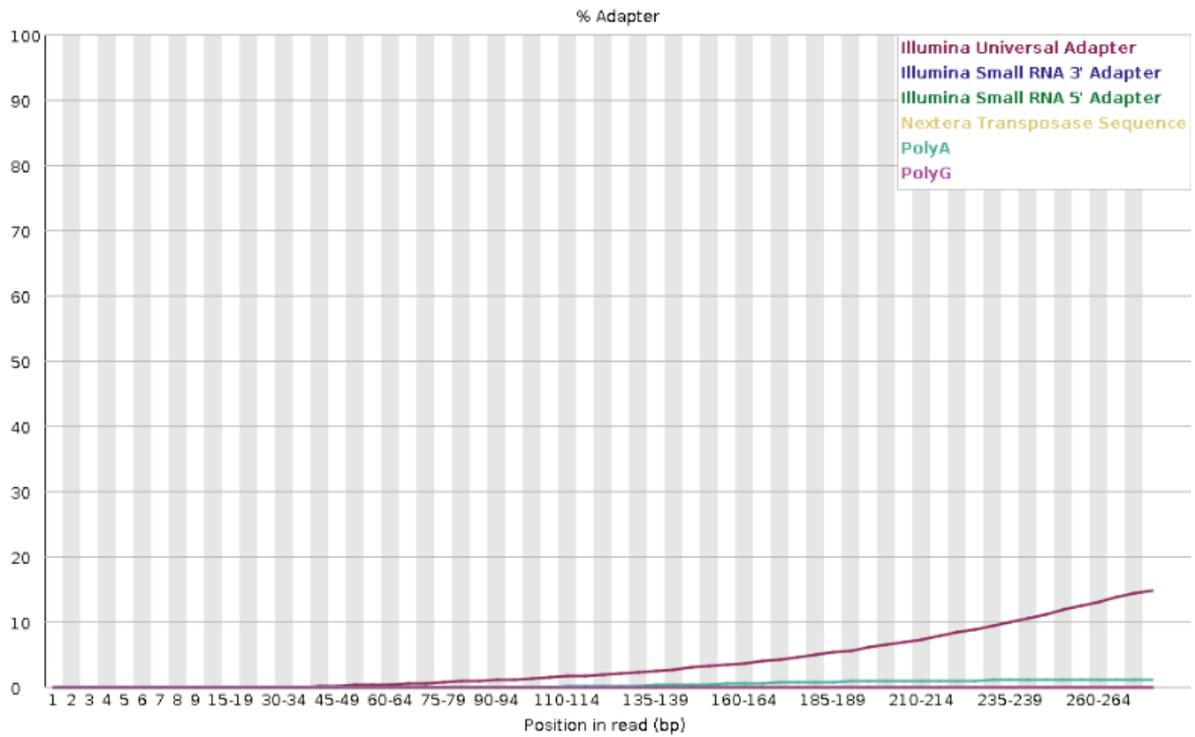


б)

Рис. 29. *Per base sequence quality* для Lacto_S145_R2_001.fastq.gz

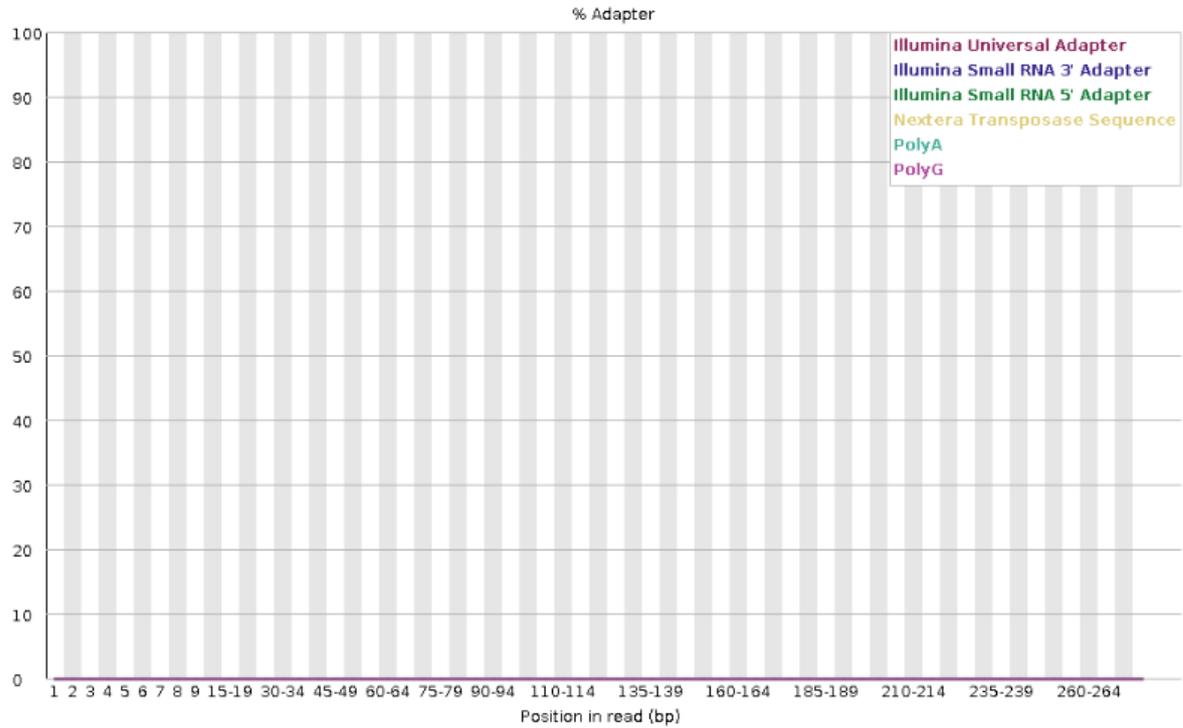
а) до тримминга, б) после тримминга

✖ Adapter Content



а)

✔ Adapter Content



б)

Рис. 30. *Adapter content* для Lacto_S145_R2_001.fastq.gz:

а) до тримминга, б) после обрезки

3.5. Сборка генома

Для сборки генома на панели инструментов в разделе **GENOMICS ANALYSIS** нажмите **Assembly** и в выпадающем списке выберите **SPAdes** (рис. 31).

В поле **FASTA/FASTQ file(s): forward reads** выберите результат работы Trimmomatic с прямыми рядами **Trimmomatic on Lacto_S145_R1_001.fastq.gz: R1 paired**.

В поле **FASTA/FASTQ file(s): reverse reads** выберите результат работы Trimmomatic с обратными рядами **Trimmomatic on Lacto_S145_R2_001.fastq.gz: R2 paired**.

The screenshot displays the configuration page for the SPAdes genome assembler. On the left, a 'Tools' sidebar lists various bioinformatics tools, with 'SPAdes genome assembler for genomes of regular and single-cell projects' highlighted in a red box. The main configuration area on the right includes several sections: 'Tool Parameters' with 'Operation mode' set to 'Assembly and error correction'; 'Single-end or paired-end short-reads' set to 'Paired-end: individual datasets'; 'FASTA/FASTQ file(s): forward reads' set to '7: Trimmomatic on Lacto_S145_R1_001.fastq.gz (R1 paired)'; 'FASTA/FASTQ file(s): reverse reads' set to '8: Trimmomatic on Lacto_S145_R2_001.fastq.gz (R2 paired)'; 'Type of paired-reads' set to 'Default (-pe)'; and 'Select orientation of reads' set to 'FR (-> <-)'. Other options like 'Use an additional set of short-reads' (Disabled) and 'Additional read files' (empty) are also visible.

Рис. 31. Настройка SPAdes

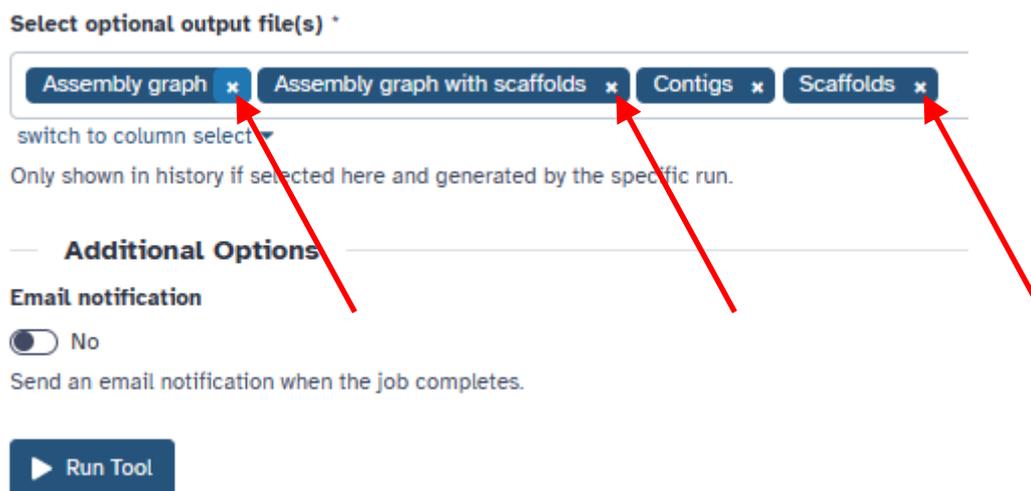


Рис. 32. Настройка SPAdes (продолжение)

В пункте *Select optional output file(s)* удалите кнопкой × все автоматически выбранные варианты, кроме *Contigs* (рис. 32).

Оставьте остальные настройки по умолчанию и нажмите ► *Run Tool*.

Сборка генома – самый долгий и вычислительно сложный этап, выполнение программы занимает значительное время (от часа и более). Сборка осуществляется из обрезанных отфильтрованных ридов после тримминга. Консенсусная последовательность коротких ридов образует контиг. Последовательность, объединяющая несколько контигов с неизвестной последовательностью известного размера (NNNNN) между ними образуют скаффолд (рис.33).

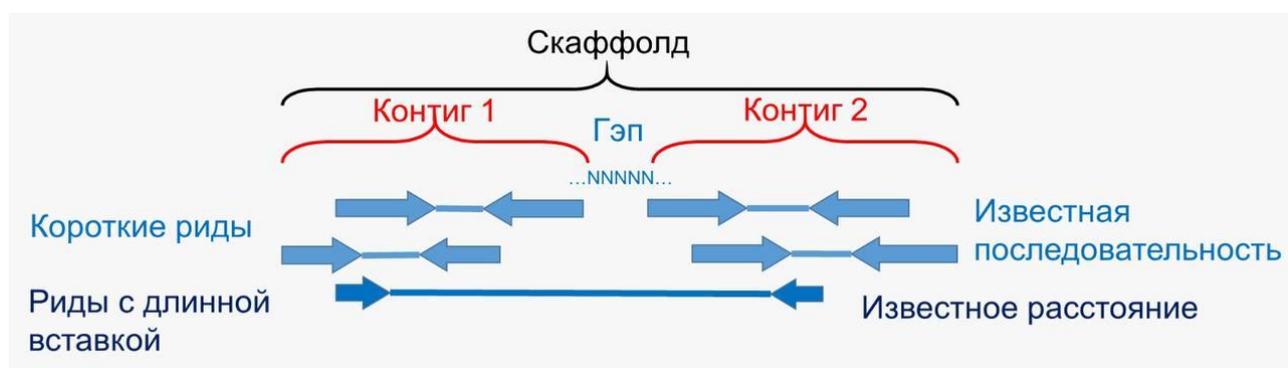


Рис. 33. Принцип сборки генома: риды – контиги – скаффолды

Прогресс выполнения и результат сборки генома (один файл под названием *SPAdes on data ... and data...: Contigs*) появится в панели истории (рис. 34).

Для просмотра собранных контигов (*Contigs*) нажмите кнопку . По умолчанию выводится только первый мегабайт данных (рис. 35).

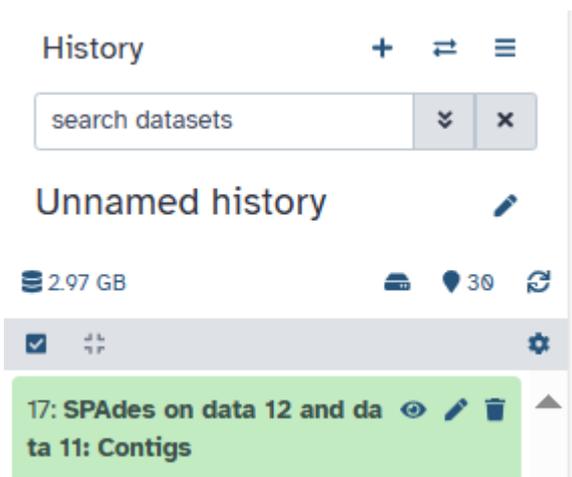


Рис. 34. Результат сборки генома на панели истории



Рис. 35. Просмотр результата сборки генома

Как видно из рис.35, контиги представлены в формате FASTA. В заголовке указан номер контига (NODE...), его длина (length...) и средняя степень покрытия (cov...), которая указывает, сколько в среднем ридов перекрывает каждую позицию в контиге (рис.36). Высоким значением покрытия для сборки генома *de novo* является значение в пределах 50-100.

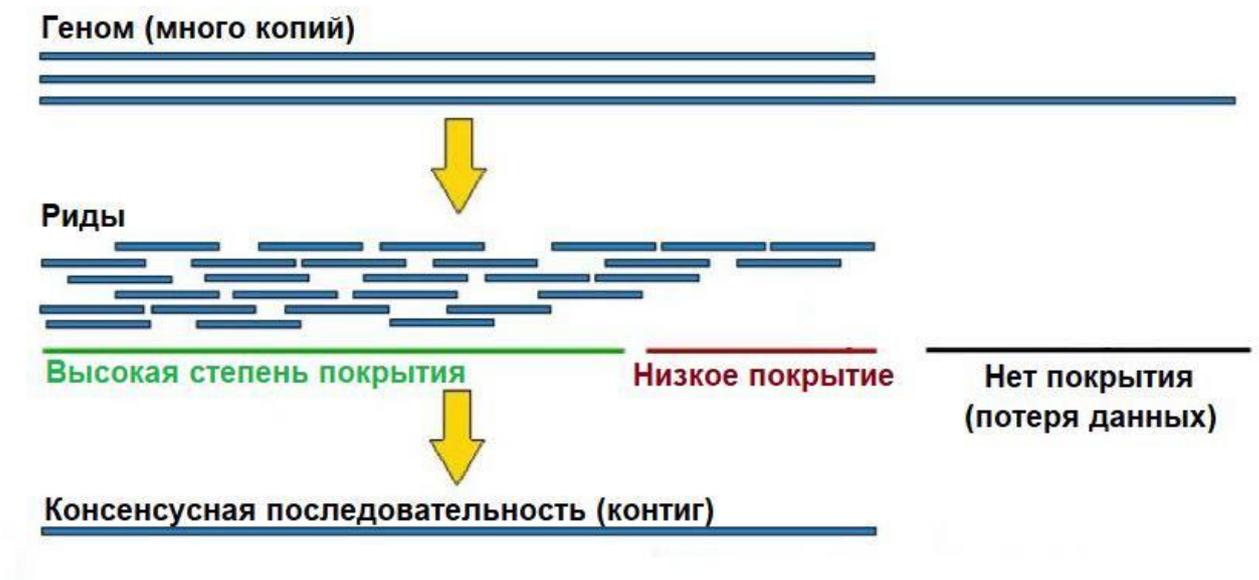


Рис. 36. Принцип сборки генома: степень покрытия контига ридами

3.6. Контроль качества сборки генома

Для контроля качества сборки генома на панели инструментов в разделе *GENOMICS ANALYSIS* нажмите *Assembly* и в выпадающем списке выберите *Quast* (рис. 37).

В поле *Contigs/scaffolds file* * выберите результат сборки генома *SPAdes on data ... and data...: Contigs*.

Оставьте остальные настройки по умолчанию и нажмите ► *Run Tool*. Прогресс выполнения и результаты появятся в панели истории (рис. 38). Для просмотра отчёта по качеству сборки генома нажмите кнопку 👁 (рис. 39).

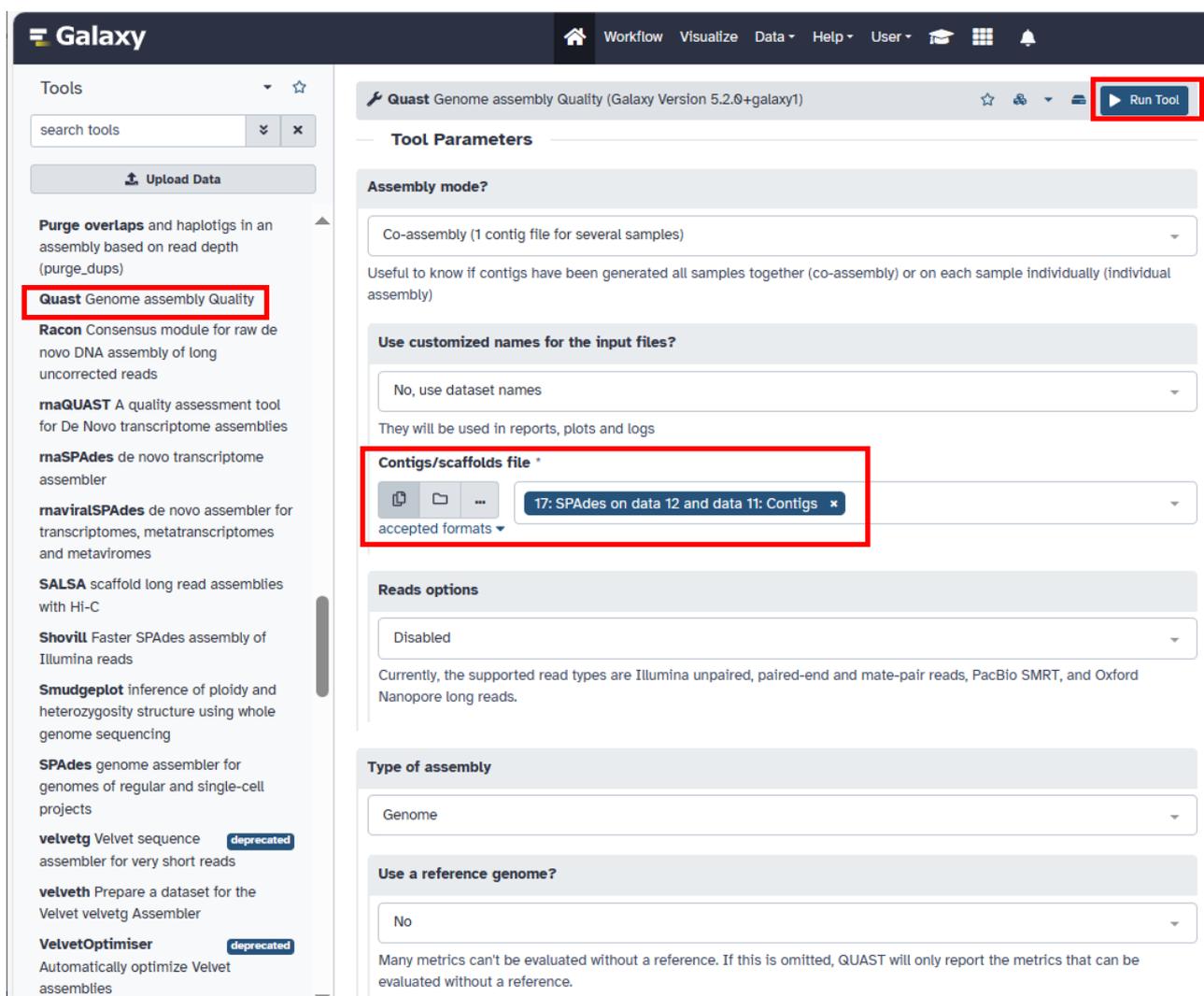


Рис. 37. *Quast* в панели инструментов и настройки выполнения

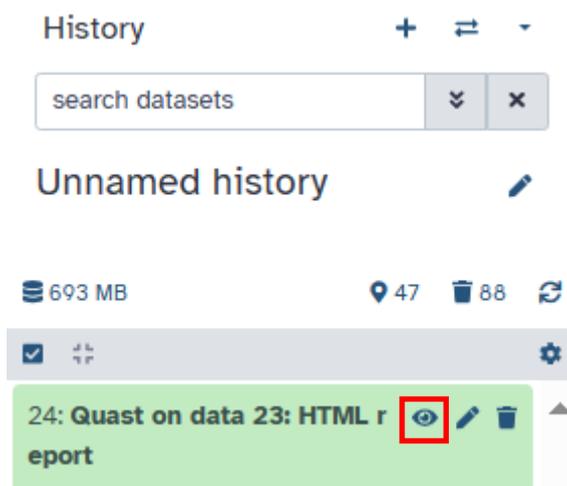


Рис. 38. Результат выполнения программы *Quast* в панели истории

Как видно из рис. 39, отчёт содержит табличные и графические данные.

В таблице приведены статистические данные по количеству значимых контигов с длиной больше 500 п.н. (*# contigs*, 784), общему количеству всех контигов (*# contigs* ≥ 0 bp, 856), количеству контигов длиной более 1000 п.н. (*# contigs* ≥ 1000 bp, 570), длине самого протяжённого контига (*Largest contig*, 240910 п.н.), общей длине значимых (>500 п.н.) контигов (*Total length*, 5029383 п.н.), общей длине всех контигов (*Total length* ≥ 0 bp, 5060818 п.н.), общей длине контигов >1000 п.н. (*Total length* ≥ 1000 bp, 4874706 п.н.), GC-составу (*GC%*, 41.82), количеству и содержанию на 100000 п.н. неопознанных нуклеотидов (*#N's*, 0 и *#N's per 100 kb*, 0) и параметрам, описывающими вклад контигов в общую сборку:

N50 = 52948 п.н. – контиги более длинные, чем это значение, покрывают не менее 50% сборки.

L50 = 26 – количество таких контигов, образующих 50% сборки.

N90 = 1977 п.н., *L90* = 330 – аналогичные параметры для совокупности контигов, образующих 90% сборки.

В отчёт включён график (рис.39), отображающий все N_x -значения для $x = 0-100\%$.

auN = 68125 – площадь под графиком N_x -значений

Подробнее ознакомиться с руководством по QUAST можно по ссылке:
(<https://quast.sourceforge.net/docs/manual.html#sec3.2>)

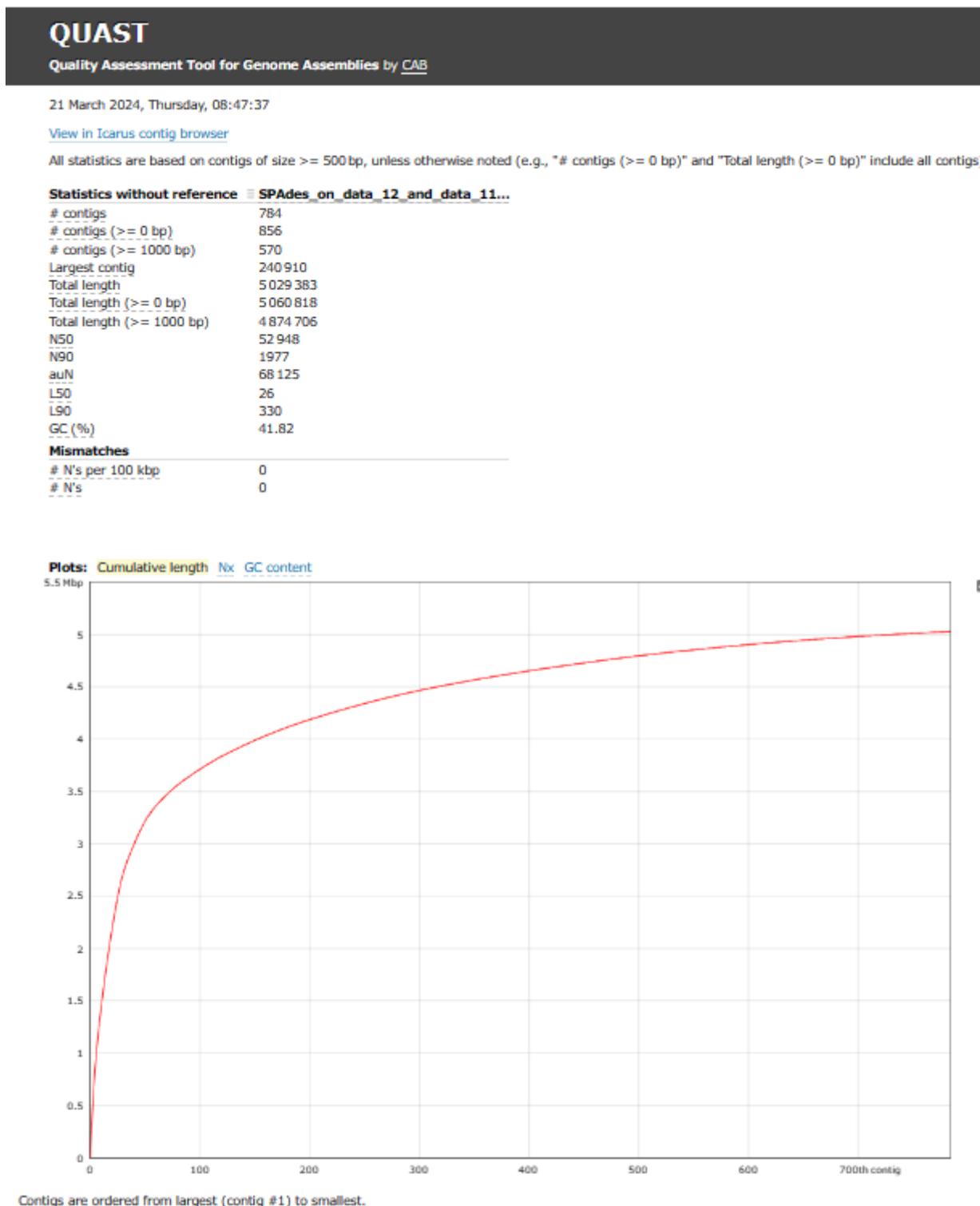


Рис. 39. Отчёт **Quast** о качестве сборки генома – таблица по статистике и кумулятивный график длины

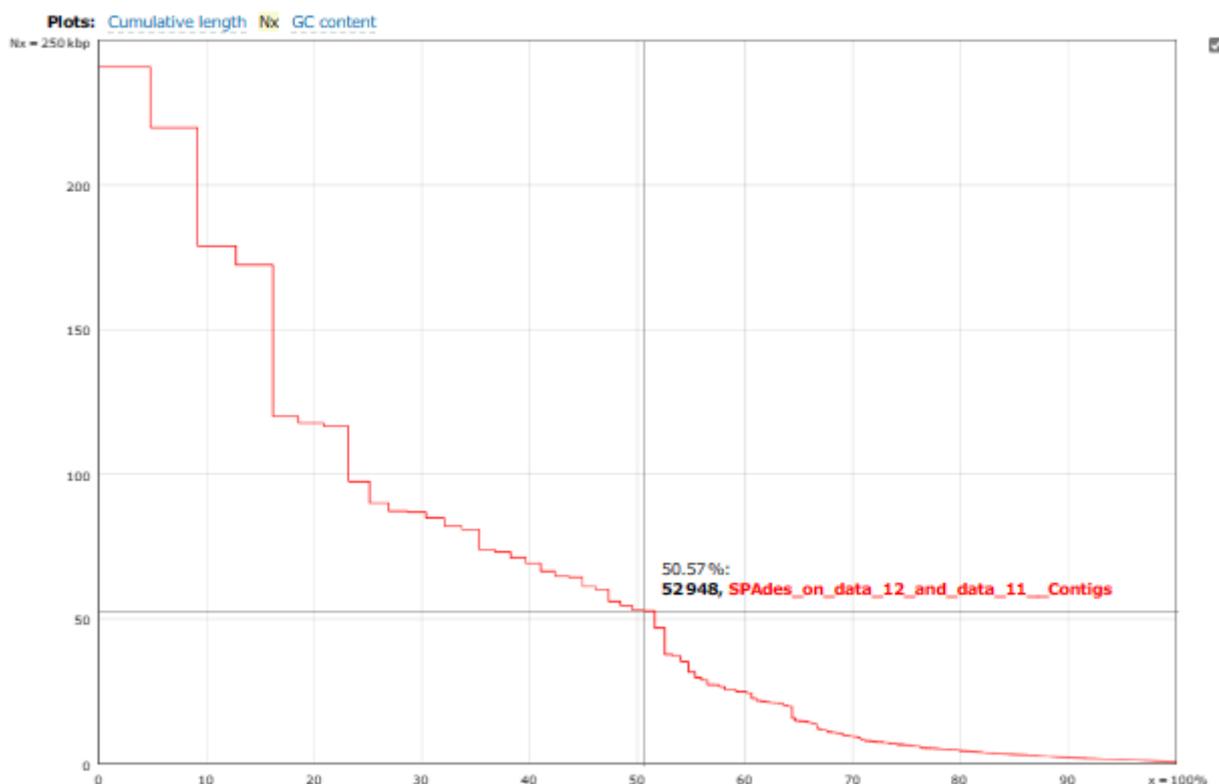


Рис. 39. (продолжение) График Nx -значений, $N50$ выделено

По данным параметрам можно сделать вывод о качестве сборки генома:

$N50$ является мерой целостности генома – качество сборки тем выше, чем ниже значение $N50$ (чем больше $N50$, тем более мелкими кусочками собрался геном), как и *общее количество контигов*. Чем меньше *количество* и *содержание неопознанных нуклеотидов*, тем выше качество.

В случае, если известен референсный геном, можно проверить соответствие примерной *общей длины* и *GC-состава*. Допустимо колебание GC-состава в пределах 1%, более сильные отличия могут свидетельствовать о контаминации секвенированного образца.

4. Аннотация прокариотического генома в программе Prokka

Для аннотации генома на панели инструментов в разделе *GENOMICS ANALYSIS* нажмите *Annotation* и в выпадающем списке выберите *Prokka* (рис. 40).

В поле *Contigs to annotate* * выберите результат сборки генома *SPAdes on data ... and data...: Contigs*.

В поле *Minimum contig size (--mincontiglen)* введите значение 500 для фильтрации контигов (удаление контигов длиной менее 500 п.н.).

Оставьте остальные настройки по умолчанию и нажмите ► *Run Tool*. Прогресс выполнения и результаты аннотации в разных форматах появятся в панели истории (рис. 40).

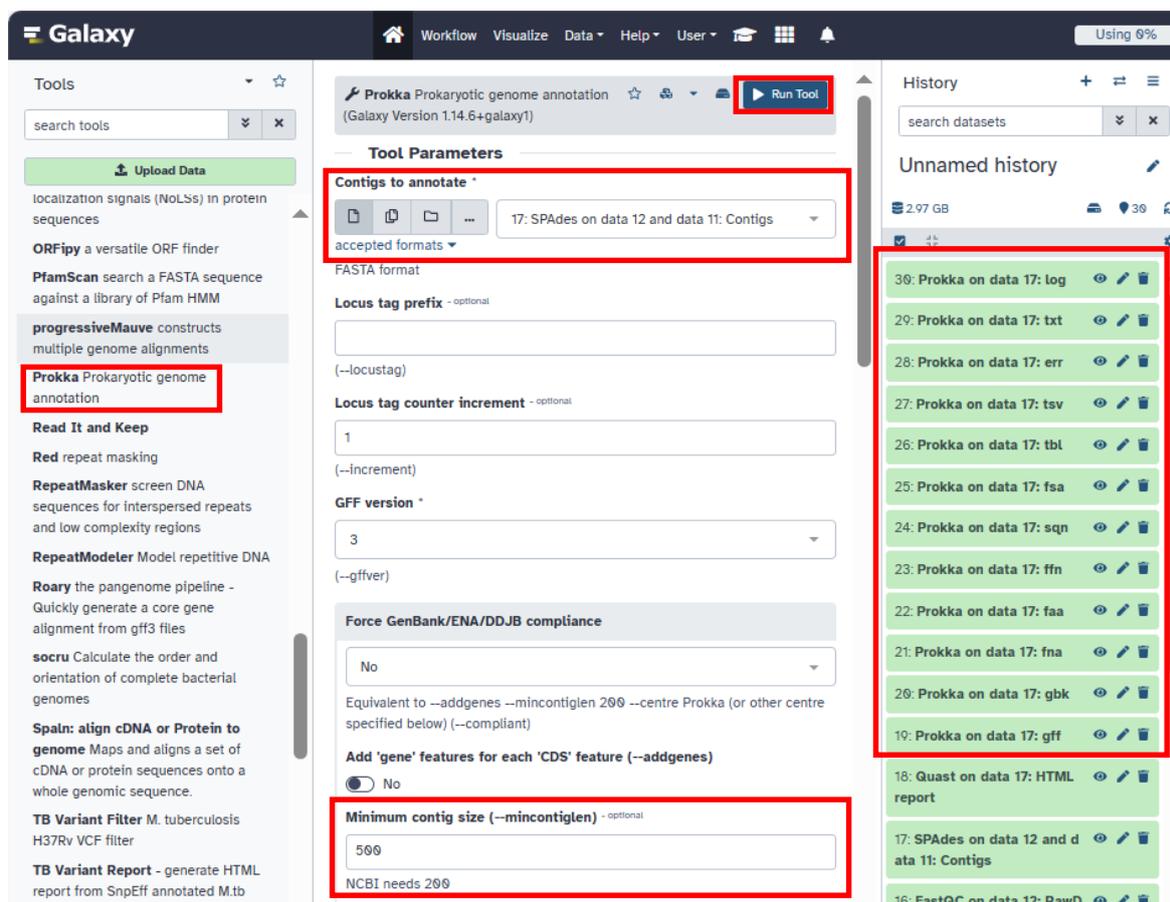
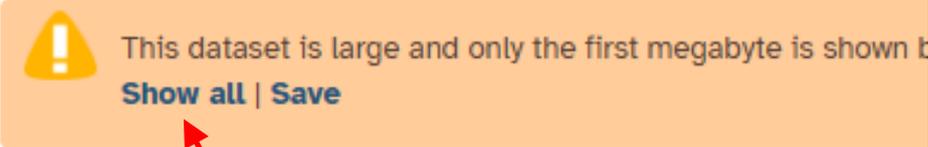


Рис. 40. *Prokka* в панели инструментов, настройки выполнения, результаты аннотации в панели истории

Для просмотра результатов аннотации нажмите кнопку 👁 напротив соответствующих файлов. Как видно из рис. 40, *Prokka* выводит 12 файлов с

результатами в разных форматах. Для дальнейшей работы необходим только формат *ffn* – *Prokka on data...: ffn* (рис. 41) – нуклеотидные последовательности **всех** аннотированных последовательностей в формате FASTA.



```
>JBGHFDFB_00001 putative protein
ATGGTTAAAACACATGCAATAGAAATTTTTGATGGTTATGCGACTGAATATGATCAGTGG
TTTCTGGCCAATAAGCAGATATTTCTGAGTGAACATAAACTTTTAAAATCTGTTTTAGAC
TTAGGCTCAACAAAACATGCTTATCAATCGGTTGTGGCAGCGGGTTATTTGAAGATACC
CTCCACCGCCAATATGGCCTGCCACTTTTTGATGGTGTGAACCATCTACCGATATGGCA
CAGATTGCAAAAAGCGGGGATTAAACGTTCCGCTTGGAAAAGCAGAAAGTATTAATCTG
CCCACAAGACAGTATGATACCATTACTTTAATGGCAGCTCCAGTTATATTAAGGACTTA
TTGGCAGCTTATCAAAATTGTATCAAAGCACTCAAGTCCGGCGGCCATTTTATTTAATT
GACGTGCCAAAAGAAAGTGCTTATGGATTAATGTATATGCTTGCCAAAATTCAGGGTAAT
TATCAATCAGAAGAATTGGCAAATGTTCTTCCGAAATGGCCATACCCAATTGAACTGGTT
GATTCTGCATACTGGCATAACCACCTAGAAAAAAGCACCTTTAGAAAACGAGCTTCAG
CTACGACATTTAAGGTTCAAACAAACCTTGGCAGCAAACCCACTCTACACAAACGATAGT
GTTGAAGAACCTGAAGATGGCTTTTCAAAGGGTGGTTATGTCGCAATAATTGCCGAAAAA
CCTTAA
>JBGHFDFB_00002 Peptide methionine sulfoxide reductase MsrA
ATGGAAGATACAGCAATTTTCCCGGTGGGTGTTTCTGGTGTATGGTTAAACCGTTTGAT
ACCATGCCCGGTATTAATAAAGTTATTTCCGGATACACGGGTGGTCATGTAGCAAAATCCG
AGTACGAACAGGTTAGCAGCCATACAACGGGACATACCGAGGCCGTTAAAATCAATTTT
GATCCAGATATTATTAGTTATAAGCAACTGGTTCAAATTTATTGGCAGCAAACAGATCCA
ACAGATGCAATGGGACAGTTTCAGGATCGAGGAGATAACTATCGTCCAGTGATCTTTGTG
AAAGATGAAGCGCAACGAAAAATTGCGGAGGCTTCCAAAAAAGCTTTAGAGGAAACCGGT
ATGTTTGACCAGCCGATTGTCACTCAAATGAAGAGGCCAAACCTTTTATCCAGCAGAA
GAGGAACATCAACAGTTTTATAAAAAAGAATCCATTTTCGCTTCCAAATGGAGGAAGCTGGT
GGACGTGAAAAATTTGTGAAAGAACTGGCAGTCAAAAACAAAAATAA
>JBGHFDFB_00003 Peptide methionine sulfoxide reductase MsrB
ATGGTTGAGAAAAAAGATTTAAAACATCGATTAACACCAGAACAATATGCGGTTACCCAA
GAGGCCGCTACGGAGGCACCCTTTAGTGGTCAATATGATCATTTTTATCAAGAGGGTATT
TATGTTGACGTCGTAAGTGGTGAACCTTTGTTTTATCTAAGGACAAATATGACGCTGGG
TGTGGATGGCCATCATTTACCAAACCAATTGAGCAAACAAGTATTCATAAGAATTTGGAT
ACTTCTACGGGATGATTAGAGAGGAAGTTCAAAGTAATGATGCTCATTCTCATCTAGGG
CATGTGTTTGGCGATGGGCCAAAAGAAGCAGGCCGCCAACGATATTGTATCAATCCGCG
GCCTTGAAGTTCATTTCCACAGACGAGCTTGAAGAGGCAGGGTACGGTCAGTATAAACCA
ATGTTTAACTGA
```

Рис. 41. Фрагмент файла в формате *ffn*

5. Определение таксономии исследуемого генома

5.1. Извлечение последовательности *16S pPHK*

Для извлечения последовательности из собранного и аннотированного генома откройте предпросмотр файла *Prokka on data...: ffn*, нажмите *Show All* (рис. 41) и выполните поиск на странице (по умолчанию **Ctrl+F**) по запросу «16S ribosomal RNA».

Выделите и скопируйте нуклеотидную последовательность гена в формате FASTA (рис. 42).

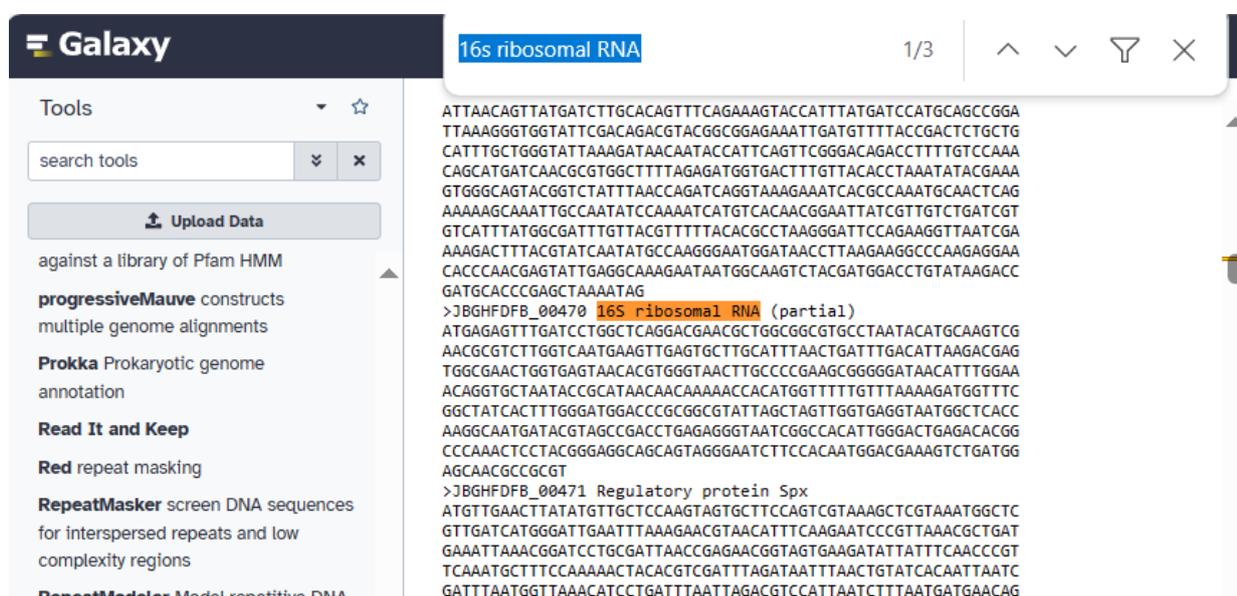


Рис. 42. Поиск последовательностей *16S ribosomal RNA*

Вставьте и сохраните последовательность в текстовый файл на компьютере (рис. 43).



Рис. 43. Создание и сохранение текстового файла с последовательностью *16S pPHK*

Если в геноме обнаружено несколько последовательностей *16S pPHK* (рис. 42), то крайне вероятно контаминация. В таком случае необходимо провести идентификацию **каждой** из последовательностей *16S pPHK* в программе blastn, и, если они принадлежат разным видам, провести разделение метагенома.

5.2. Определение таксономии по последовательности *16S pPHK* в программе blastn

Программа blastn включена в набор инструментов Galaxy, однако используемые базы данных (nt 17-Apr-2014) сильно устарели, поэтому предпочтительнее использование инструмента blastn (рис. 44) на сайте NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) [5].

The image shows the NCBI BLAST website interface for a blastn search. The page title is "BLAST® » blastn suite". The main heading is "Standard Nucleotide BLAST". The interface is divided into several sections:

- Enter Query Sequence:** This section contains a text input field with the sequence: `>JBGHDFB_04708 16S ribosomal RNA (partial)
TGATACGTAGCCGACCTGAGAGGGTAATCGGCCACATTGGGACTGAGA
CACGGCCAAAC
TCCTACGGGAGGCAGCAGTAGGGAATCTTCCACAATGGACGAAAGTCT`. Below the input field are options for "Or, upload file" and "Job Title" (JBGHDFB_04708 16S ribosomal RNA (partial)).
- Choose Search Set:** This section includes a "Database" dropdown menu set to "Standard databases (nr etc.)", a "Nucleotide collection (nr/nt)" dropdown menu, and various optional filters like "Exclude" and "Limit to".
- Program Selection:** This section has a radio button selected for "Highly similar sequences (megablast)".
- BLAST Button:** A prominent blue button labeled "BLAST" is located at the bottom left of the configuration area.

Рис. 44. Настройка выполнения blastn

Вставьте одну последовательность *16S pPHK* в поле *Enter Query Sequence*, выберите *Nucleotide collection nr/nt* в поле *Database*, выберите *Highly similar sequences (megablast)* в поле *Program Selection* и нажмите на кнопку *BLAST*. Ниже приведены полученные результаты (рис. 45)

Повторите для остальных последовательностей (если они есть).

BLAST® » blastn suite » results for RID-ZXT5GGKK016

Home Recent Results Saved Strategies Help

[< Edit Search](#) Save Search Search Summary ▾

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title JBGHFDFB_00470 16S ribosomal RNA (partial)
RID ZXT5GGKK016 Search expires on 03-24 22:18 pm Download All ▾
Program BLASTN Citation ▾
Database nt See details ▾
Query ID lcl|Query_3386961
Description JBGHFDFB_00470 16S ribosomal RNA (partial)
Molecule type dna
Query Length 433
Other reports Distance tree of results MSA viewer ⓘ

Filter Results

Organism only top 20 will appear exclude
 Type common name, binomial, taxid or group name
 + Add organism

Percent Identity to **E value** to **Query Coverage** to

Filter **Reset**

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download ▾ Select columns ▾ Show 100 ▾ ⓘ

select all 100 sequences selected GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> Lentilactobacillus hilgardii strain LMG 07934 chromosome, complete genome	Lentilactobacillus hilgardii	800	4003	100%	0.0	100.00%	2771862	CP050262.1
<input checked="" type="checkbox"/> Lentilactobacillus hilgardii strain FLUB chromosome, complete genome	Lentilactobacillus hilgardii	800	4003	100%	0.0	100.00%	3071102	CP047121.1
<input checked="" type="checkbox"/> Lentilactobacillus hilgardii strain LH500 chromosome, complete genome	Lentilactobacillus hilgardii	800	3970	100%	0.0	100.00%	2654177	CP044119.1
<input checked="" type="checkbox"/> Lactobacillus hilgardii strain M3-4 16S ribosomal RNA gene, partial sequence	Lentilactobacillus hilgardii	797	797	99%	0.0	100.00%	1540	KF030792.1

Рис. 45. Результаты поиска *blastn* гомологичных последовательностей к *JBGHFDFB_00470 16S ribosomal RNA*

Как видно из рис. 45, последовательность по гену *16S pPHK* при $E\text{-value}=0.0$ (вероятность ошибки равна нулю) со 100% совпадением нуклеотидных последовательностей (Per Ident) соответствует известной нуклеотидной последовательности *16S pPHK Lentilactibacillus hilgardii* в БД GenBank. Таким образом, секвенированная последовательность относится к *Lentilactibacillus hilgardii*. Для извлечения Таксоному ID нажмите ссылку в поле Accession, чтобы перейти к аннотации известной последовательности в БД GenBank (рис. 45). Таксоному ID приведено в поле `/db_xref="taxon:1588"` (рис. 46).

Lentilactobacillus hilgardii strain LMG 07934 chromosome, complete genome

GenBank: CP050262.1

[FASTA](#) [Graphics](#)[Go to:](#) 

LOCUS CP050262 2771862 bp DNA circular BCT 29-MAR-2020
DEFINITION Lentilactobacillus hilgardii strain LMG 07934 chromosome, complete genome.
ACCESSION CP050262
VERSION CP050262.1
DBLINK BioProject: [PRJNA609644](#)
BioSample: [SAMN14262734](#)
KEYWORDS .
SOURCE Lentilactobacillus hilgardii
ORGANISM [Lentilactobacillus hilgardii](#)
Bacteria; Bacillota; Bacilli; Lactobacillales; Lactobacillaceae; Lentilactobacillus.
REFERENCE 1 (bases 1 to 2771862)
AUTHORS Zhuravleva,D.E., Iskhakova,Z.I., Ozhegov,G.D., Angelov,A., Engerer,C., Khusnutdinova,D.R., Gogoleva,N.E., Forchhammer,K. and Kayumov,A.R.
TITLE The whole genome of Lactobacillus brevis subsp. gravesensis ATCC 27305
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 2771862)
AUTHORS Zhuravleva,D.E., Iskhakova,Z.I., Ozhegov,G.D., Angelov,A., Engerer,C., Khusnutdinova,D.R., Gogoleva,N.E., Forchhammer,K. and Kayumov,A.R.
TITLE Direct Submission
JOURNAL Submitted (19-MAR-2020) Institute of Fundamental Medicine and Biology, Department of Genetics, Kazan Federal University, Kremlyovskaya St, 18, Kazan 420008, Russia
COMMENT ##Genome-Assembly-Data-START##
Assembly Date :: 13-JAN-2020
Assembly Method :: Unicycler v. 0.4.8-beta
Genome Representation :: Full
Expected Final Version :: Yes
Genome Coverage :: 201.0x
Sequencing Technology :: Illumina MiSeq; Oxford Nanopore MiniION
##Genome-Assembly-Data-END##
FEATURES Location/Qualifiers
source 1..2771862
/organism="Lentilactobacillus hilgardii"
/mol_type="genomic DNA"
/strain="LMG 07934"
/isolation_source="Fermented beverages, wine"
/culture_collection="LMG:07934"
/db_xref="taxon:1588"
/geo_loc_name="Germany"
/collection_date="1967"
gene 58..1410
/gene="dnaA"
/locus_tag="G8J22_00001"
CDS 58..1410
/gene="dnaA"

Рис. 46. Аннотация известной нуклеотидной последовательности *16S pPHK**Lentilactobacillus hilgardii* в GenBank

6. Аннотация генома в программе RAST

RAST (<https://rast.nmpdr.org/>) – Rapid Annotation using Subsystem Technology – программа для аннотации генома.

Для работы в программе RAST необходимо скачать файл **SPAdes on data... and data...: Contigs** из **Galaxy** (рис. 47). Нажмите на название файла в панели истории, чтобы открыть меню. Нажмите кнопку , чтобы загрузить файл на компьютер.

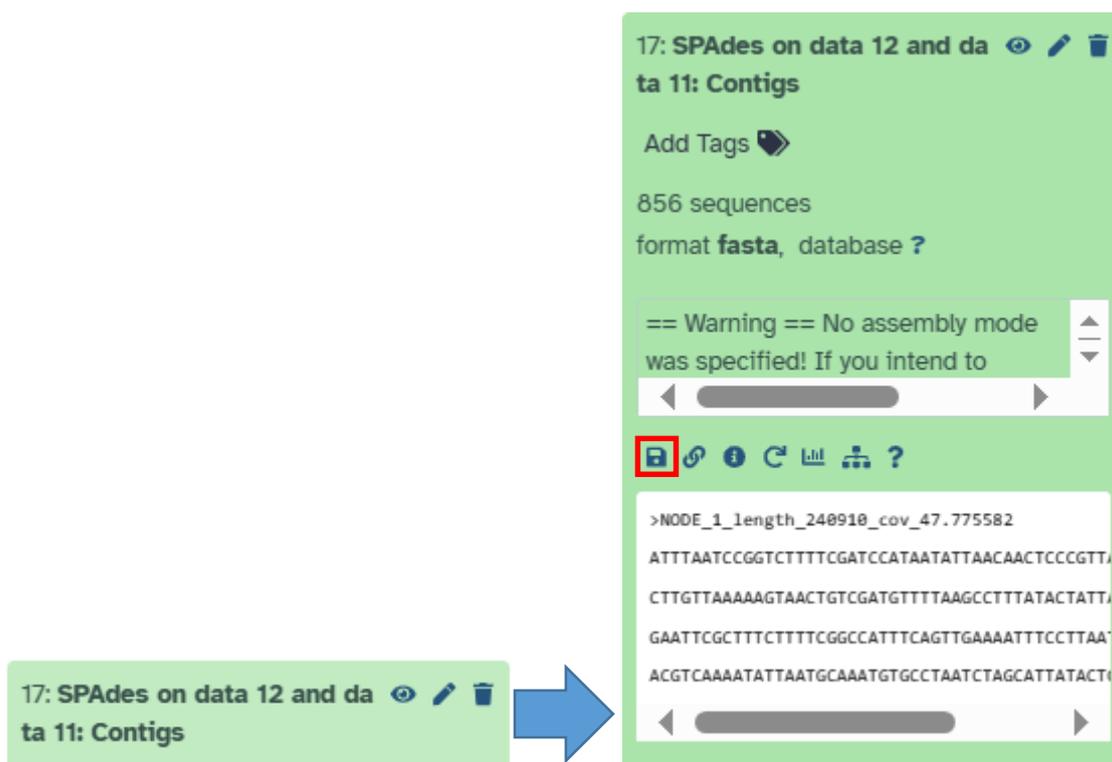


Рис. 47. Скачивание файла **SPAdes on data... and data...: Contigs** из Galaxy

Откройте главную страницу RAST и нажмите » **Upload a new genome** (рис. 48).

Info: [RAST Access Problems](#)
[Click here](#) for instructions on how to resolve several of the most common problems accessing RAST or your RAST data.

Command-Line API "301 Permanently Moved" Errors
[Click here](#) for instructions on how to resolve "301 Permanently Moved" errors when using the RAST batch command-line interface.

To monitor RAST's load and view other news and statistics for RAST and the SEED, please visit "[The Daily SEED.](#)"

As of Fri Mar 22 02:50:03 2024, there are 21 jobs in the RAST queue
Job Load is Moderate

Jobs Overview

The overview below list all genomes currently processed and the progress on the annotation. To get a more detailed report on an annotation in case of questions or problems using this service, please contact: rast@mcs.anl.gov.

Progress bar color key:

- not started
- queued for computation
- in progress
- requires user input
- failed with an error
- successfully completed

Jobs you have access to :

You currently have no jobs.

[» Upload a new genome](#)

Рис. 48. Главная страница RAST

Выберите загруженный из Galaxy файл **SPAdes on data... and data...: Contigs** в поле **Sequences File** и нажмите **Use this data and go to step 2** (рис. 49).

Upload a Genome

Please note: This service is not a BLAST-like service. It is designed to annotate complete or nearly complete (>97%) assembled prokaryotic genomes, and complete phages or it cannot analyze eukaryotes, small fragments of genomes, unassembled reads, or metagenomes.

You may upload a prokaryotic genome in one or more contigs, as either a single multirecord [FASTA](#) format file or a Genbank format file.

Our pipeline will use the taxonomy identifier as a handle for the genome. Therefore if at all possible please input the numeric [taxonomy identifier](#), and genus, species and strain

Please note that RAST will only provide you with its most complete analysis, including *Subsystems*, *Metabolic Reconstruction* and *Scenarios*, if you submit all relevant contigs for

If you wish to upload multiple genomes at once, you may be interested in using the batch upload interface that is available in the [myRAST distribution](#). See [this tutorial](#) for more

Confidentiality information: Data entered into the server will not be used for any purposes or in fact integrated into the main SEED environment, it will remain on this server

If you use the results of this annotation in your work, please cite:

- *The RAST Server: Rapid Annotations using Subsystems Technology.* Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formosa K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil L *BMC Genomics*, 2008, [[PubMed entry](#)]
- *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. *Nucleic Acids Res.* 2014 [[PubMed entry](#)]
- *RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes.* Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, Stevens R, Vonstein V, Wattam AR, Xia F *Sci Rep.*, 2015, [[PubMed entry](#)]

File formats: You can either use [FASTA](#) or Genbank format.

- If in doubt about FASTA, [this service](#) allows conversion into FASTA format.
- Due to limits on identifier sizes imposed by some of the third-party bioinformatics tools that RAST uses, we limit the size of contig identifiers to 70 characters or fewer.
- If you use GenBank, you have the option of preserving the original GenBank gene calls in the options block below. By default, genes will be recalled.

File Upload:
 Sequences File

[Use this data and go to step 2](#)

Рис. 49. Работа с RAST, шаг 1 – загрузка генома

Введите NCBI **Taxonomy ID** организма (в нашем примере – 1588) и нажмите на кнопку **Fill in form based on NCBI taxonomy-ID** для автоматического заполнения данных. Нажмите **Use this data and go to step 3** (рис. 50).

В поле **Choose RAST annotation scheme** выберите **Classic Rast**. Включите флажок **Compute similarities?**. Оставьте остальные параметры по умолчанию и нажмите кнопку **Finish the upload** (рис. 51).

RAST Rapid Annotation using Subsystem Technology version

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

»Home »Your Jobs »Tutorials »Help

Upload a Genome

Review genome data

We have analyzed your upload and have computed the following information.

Contig statistics

Statistic	As uploaded	After splitting into scaffolds
Sequence size	5060818	5060818
Number of contigs	856	856
GC content (%)	41.8	41.8
Shortest contig size	128	128
Median sequence size	1510	1510
Mean sequence size	5912.2	5912.2
Longest contig size	240910	240910
N50 value	52948	52948
L50 value	26	26

Please enter or verify the following information about this organism:

- RAST bases its genome identifiers on NCBI taxonomy-IDs.
- If you provide a valid taxonomy-ID, RAST will attempt to fill in the genome metadata for you.
- If you leave the taxonomy-ID field blank, RAST will assign a meaningless taxonomy-ID, and you will need to fill in the b
- If you plan on submitting this genome to [PATRIC](#) you will need to provide the most descriptive NCBI taxonomic groupin process for submitting your genome to PATRIC [in this document](#).
- You may search for the taxonomy-ID of your organism using the search facilities at the [NCBI taxonomy browser](#).

Genome information:

Taxonomy ID: **Fill in form based on NCBI taxonomy-ID.**

Taxonomy string: Bacteria; Terrabacteria group; Bacillota; Bacilli; Lactobacillales; Lactobacillaceae; Lentilactobacillus; Lentilactobacillus hilgardii

Domain: Bacteria Archaea Viruses

Genus:

Species:

Strain:

Genetic Code: 11 (Archaea, most Bacteria, most Virii, and some Mitochondria) 4 (Mycoplasmaea, Spiroplasmaea, Ureoplasmaea, and Fungal Mitochondria)

Use this data and go to step 3

Рис. 50. Работа с RAST, шаг 2 – данные об организме

RAST Rapid Annotation using Subsystem Technology

The NMPDR, SEED-based, prokaryotic genome annotation service. For more information about The SEED please visit theSEED.org.

» Home » Your Jobs » Tutorials » Help

Upload a Genome

Complete Upload

Please consider the following options for the RAST annotation pipeline:

RAST Annotation Settings:

Choose RAST annotation scheme	<input type="text" value="Classic RAST"/>	Choose "RASTtk" for the current modular customizable production RAST pipeline, or "Classic RAST" for the old pipeline.
Select gene caller	<input type="text" value="RAST"/>	Please select which type of gene calling you would like RAST to perform. Note that using GLIMMER-3 will disable autom
Select FIGfam version for this run	<input type="text" value="Release70"/>	Choose the version of FIGfams to be used to process this genome.
Automatically fix errors?	<input checked="" type="checkbox"/> Yes	The automatic annotation process may run into problems, such as gene candidates overlapping RNAs, or genes embedd
Fix frameshifts?	<input type="checkbox"/> Yes	If you wish for the pipeline to fix frameshifts, check this option. Otherwise frameshifts will not be corrected.
Build metabolic model?	<input type="checkbox"/> Yes	If you wish RAST to build a metabolic model for this genome, check this option.
Backfill gaps?	<input checked="" type="checkbox"/> Yes	If you wish for the pipeline to blast large gaps for missing genes, check this option.
Compute similarities?	<input checked="" type="checkbox"/> Yes	If you wish to compute similarities for the SeedViewer compare regions display, check this box.
Turn on debug?	<input type="checkbox"/> Yes	If you wish debug statements to be printed for this job, check this box.
Set verbose level	<input type="text" value="0"/>	Set this to the verbosity level of choice for error messages.
Disable replication	<input type="checkbox"/> Yes	Even if this job is identical to a previous job, run it from scratch.

Рис. 51. Работа с RAST, шаг 3 – настройка выполнения

Для просмотра аннотированного генома нажмите » **Browse annotated genome in SEED Viewer** на странице результатов RAST (рис. 52).

На странице результатов RAST представлены статистические данные по геному – общее количество кодирующих последовательностей белков (4989), количество геномных последовательностей РНК (108). На круговой диаграмме представлено распределение идентифицированных кодирующих последовательностей белков по метаболическим группам (рис. 53).

По ссылке **View closest neighbors** можно получить информацию о близкородственных видах (рис. 54). Как видно из рис. 54, как близкородственные к секвенированному геному отмечены 29 видов, большинство из которых относятся к тому же роду *Lactiplantibacillus*. Наиболее близкие виды - *Lactobacillus brevis subsp. gravesensis* ATCC 27305, *Lactobacillus buchneri* ATCC 11577 и *Lactobacillus hilgardii* ATCC 8290.



Job Details #1420836

[» Browse annotated genome in SEED Viewer](#)

» Available downloads for this job:

» [Share this genome with selected users](#)

» View [Close Strains for this job](#)

» [Back to the Jobs Overview](#)

✓ Genome Upload has been successfully completed.

Genome ID - Name:	1588.116 - Lentilactobacillus hilgardii
Job:	#1420836
User:	SverdrupAE
Date:	Fri Mar 22 06:00:23 2024
Genetic code:	11
Annotation scheme:	ClassicRAST
Preserve gene calls:	no
Automatically fix errors:	yes
Fix frameshifts:	no
Backfill gaps:	yes

✓ Rapid Propagation has been successfully completed.

✓ Quality Check has been successfully completed.

For detailed explanations of the terms used in our quality report, please refer to [our wiki](#).

Number of features:	5097
Number of warnings:	1
Number of fatal problems:	0
Convergent overlaps:	1 Warning

✓ Quality Revision has been successfully completed.

No quality revision was necessary.

✓ Similarity Computation has been successfully completed.

✓ Bidirectional Best Hit Computation has been successfully completed.

✓ Auto Assignment has been successfully completed.

✓ Computation of Pairs of Close Homologs has been successfully completed.

Рис. 52. Страница результатов RAST



Organism Overview for *Lentilactobacillus hilgardii* (1588.116)

Genome	Lentilactobacillus hilgardii (Taxonomy ID: 1588)
Domain	Bacteria
Taxonomy	Bacteria; Terrabacteria group; Bacillota; Bacilli; Lactobacillales; Lactobacillaceae; Lentilactobacillus; Lentilactobacillus hilgardii; Lentilactobacillus hilgardii
Neighbors	View closest neighbors
Size	5,060,818
GC Content	41.8
N50	52948
L50	26
Number of Contigs (with PEGs)	856
Number of Subsystems	325
Number of Coding Sequences	4989
Number of RNAs	108

For each genome we offer a wide set of information to browse, com

Browse Compare Download Annotate

Browse through the features of [Lentilactobacillus hilgardii](#) both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

Click [here](#) to get to the Genome Browser

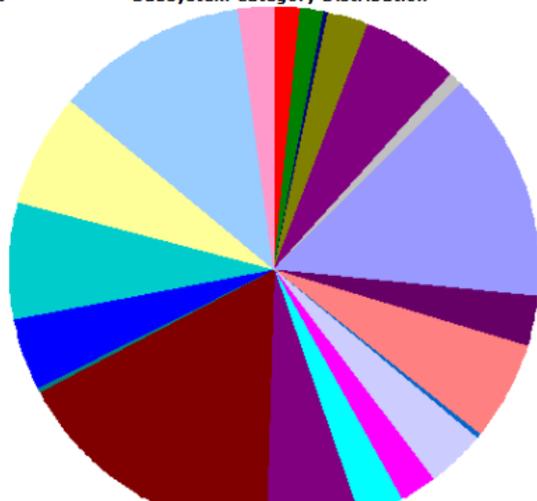
Subsystem Information

Subsystem Statistics Features in Subsystems

Subsystem Coverage



Subsystem Category Distribution



Subsystem Feature Counts

- ☐ Nucleosides and Nucleotides (189)
- ☐ Phosphorus Metabolism (86)
- ☐ Metabolism of Aromatic Compounds (11)
- ☐ Phages, Prophages, Transposable elements, Plasmids (38)
- ☐ Regulation and Cell signaling (64)
- ☐ Dormancy and Sporulation (10)
- ☐ Membrane Transport (113)
- ☐ Cell Wall and Capsule (204)
- ☐ Virulence, Disease and Defense (102)
- ☐ Motility and Chemotaxis (0)
- ☐ Photosynthesis (0)
- ☐ Carbohydrates (471)
- ☐ Secondary Metabolism (8)
- ☐ Respiration (28)
- ☐ RNA Metabolism (192)
- ☐ Amino Acids and Derivatives (548)
- ☐ Stress Response (97)
- ☐ Iron acquisition and metabolism (0)
- ☐ Cell Division and Cell Cycle (73)
- ☐ Miscellaneous (47)
- ☐ Protein Metabolism (383)
- ☐ DNA Metabolism (234)
- ☐ Nitrogen Metabolism (8)
- ☐ Potassium metabolism (9)
- ☐ Cofactors, Vitamins, Prosthetic Groups, Pigments (237)
- ☐ Fatty Acids, Lipids, and Isoprenoids (144)
- ☐ Sulfur Metabolism (14)

Рис. 53. Просмотр результатов аннотации



The SEED Viewer

SEED Viewer version 2.0

Welcome to the SEED Viewer - a read-only browser of the curated SEED data.
For more information about The SEED please visit theSEED.org.

[»Navigate](#) [»Organism](#) [»Comparative Tools](#) [»Help](#)

Closest neighbors of *Lentilactobacillus hilgardii* (1588.116)

display items per page

displaying 1 - 29 of 29

Genome ID ▲▼	Score ▲▼	Genome Name ▲▼
525310.3	541	<i>Lactobacillus brevis</i> subsp. <i>gravesensis</i> ATCC 27305
525318.3	541	<i>Lactobacillus buchneri</i> ATCC 11577
525327.3	529	<i>Lactobacillus hilgardii</i> ATCC 8290
511437.3	368	<i>Lactobacillus buchneri</i> NRRL B-30929
797515.3	367	<i>Lactobacillus parafarraginis</i> F0439
1071400.3	328	<i>Lactobacillus buchneri</i> CD034
387344.13	321	<i>Lactobacillus brevis</i> ATCC 367
387344.15	312	<i>Lactobacillus brevis</i> ATCC 367
1229281.3	280	<i>Pediococcus lolii</i> NGRI 0510Q
563194.3	270	<i>Pediococcus acidilactici</i> 7_4
862514.3	262	<i>Pediococcus acidilactici</i> DSM 20284
1080365.4	241	<i>Pediococcus acidilactici</i> MA18/5M
1036177.3	224	<i>Lactobacillus plantarum</i> subsp. <i>plantarum</i> NC8
1136177.4	213	<i>Lactobacillus pentosus</i> KCA1
220668.1	208	<i>Lactobacillus plantarum</i> WCFS1
1133596.3	208	<i>Pediococcus pentosaceus</i> IE-3
278197.10	201	<i>Pediococcus pentosaceus</i> ATCC 25745
525338.3	201	<i>Lactobacillus plantarum</i> subsp. <i>plantarum</i> ATCC 14917
220668.9	200	<i>Lactobacillus plantarum</i> WCFS1
644042.3	197	<i>Lactobacillus plantarum</i> JDM1
278197.12	193	<i>Pediococcus pentosaceus</i> ATCC 25745
889932.4	185	<i>Lactobacillus plantarum</i> subsp. <i>plantarum</i> ST-III
525366.3	162	<i>Lactobacillus vaginalis</i> ATCC 49540
525362.3	148	<i>Lactobacillus ruminis</i> ATCC 25644
525364.3	145	<i>Lactobacillus salivarius</i> ATCC 11741
362948.14	142	<i>Lactobacillus salivarius</i> UCC118
1040964.3	141	<i>Lactobacillus ruminis</i> SPM0211
299033.6	135	<i>Lactobacillus reuteri</i> F275
349123.6	133	<i>Lactobacillus reuteri</i> 100-23

displaying 1 - 29 of 29

Рис. 54. Список близкородственных видов

7. Заключение

В данном учебно-методическом пособии приведён стандартный протокол сборки и аннотации прокариотического генома, секвенированного методом NGS, с использованием веб-сервиса Galaxy.

Данный протокол включает в себя первичную оценку качества ридов (визуальная оценка полученных графиков в программе FastQC), проведение тримминга (удаление ошибок секвенирования, фильтрация ридов низкого качества, обрезка адаптеров), повторную оценку качества после тримминга, сборку генома с использованием программы-сборщика SPAdes, контроль качества сборки полученного генома, аннотацию генома в программах Prokka и RAST.

Описанный протокол универсален и может быть использован для сборки и аннотации экспериментально секвенированных геномов любых видов прокариот.

8. Список литературы

1. Illumina (2011) An Introduction to Next-Generation Sequencing Technology // Illumina Inc, 12 p. 2011. Архивная версия Waback Machine (11.04.2013): https://web.archive.org/web/20130411045647/http://www.illumina.com/documents/products/Illumina_Sequencing_Introduction.pdf
2. The Galaxy Community (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, *Nucleic Acids Research*, Volume 50, Issue W1, P.W345–W351
3. Lazarus, R.; Taylor, J.; Qiu, W.; Nekrutenko, A. (2008). "Toward the commoditization of translational genomic research: Design and implementation features of the Galaxy genomic workbench". *Summit on Translational Bioinformatics*. P.56–60.
4. Schatz, M. C. (2010). "The missing graphical user interface for genomics". *Genome Biology*. V.11(8), P.128–201.
5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* V.215, P.403-410.

9. Список электронных источников

- Next-Generation Sequencing: Principles, Applications, & Advantages
<https://www.excedr.com/blog/next-generation-sequencing-101>
 - Galaxy – <https://www.galaxy.org/>
 - Get Galaxy – <https://galaxyproject.org/admin/get-galaxy/>
 - Galaxy Australia Training: Assembly using Spades <https://galaxy-au-training.github.io/tutorials/modules/spades/>
 - Galaxy Training!: An Introduction to Genome Assembly
<https://training.galaxyproject.org/training-material/topics/assembly/tutorials/general-introduction/tutorial.html>
 - Руководство по программе FastQC 0.12.0
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
 - Документация Illumina – Adapter and Kmer Sequence Files –
https://support.illumina.com/content/dam/illumina-support/help/Illumina_DRAGEN_Bio_IT_Platform_v3_7_1000000141465/Content/SW/Informatics/Dragen/FastQC_Adapter_Kmer_files_fDG.htm
 - QUILT 5.2.0 manual – <https://quilt.sourceforge.net/docs/manual.html#sec3.2>
 - RAST: Rapid Annotation using Subsystem Technology – <https://rast.nmpdr.org/>
 - BLAST: Basic Local Alignment Search Tool / National Center for Biotechnology information (NCBI) – <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
 - Справочная информация BLAST: <https://blast.ncbi.nlm.nih.gov/doc/blast-help/>
-

А.Э. СВЕРДРУП, Л.Л. ФРОЛОВА, И.И. ЗАДОРИНА

**СБОРКА И АННОТАЦИЯ ПРОКАРИОТИЧЕСКОГО ГЕНОМА
С ИСПОЛЬЗОВАНИЕМ WEB-СЕРВИСА GALAXY**

Учебно-методическое пособие по дисциплине
«Б1.В.04 Спецпрактикум по прикладным методам в биологии»
06.03.01 Биология (бакалавр)