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.О4 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

Содержание

1. Введение

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

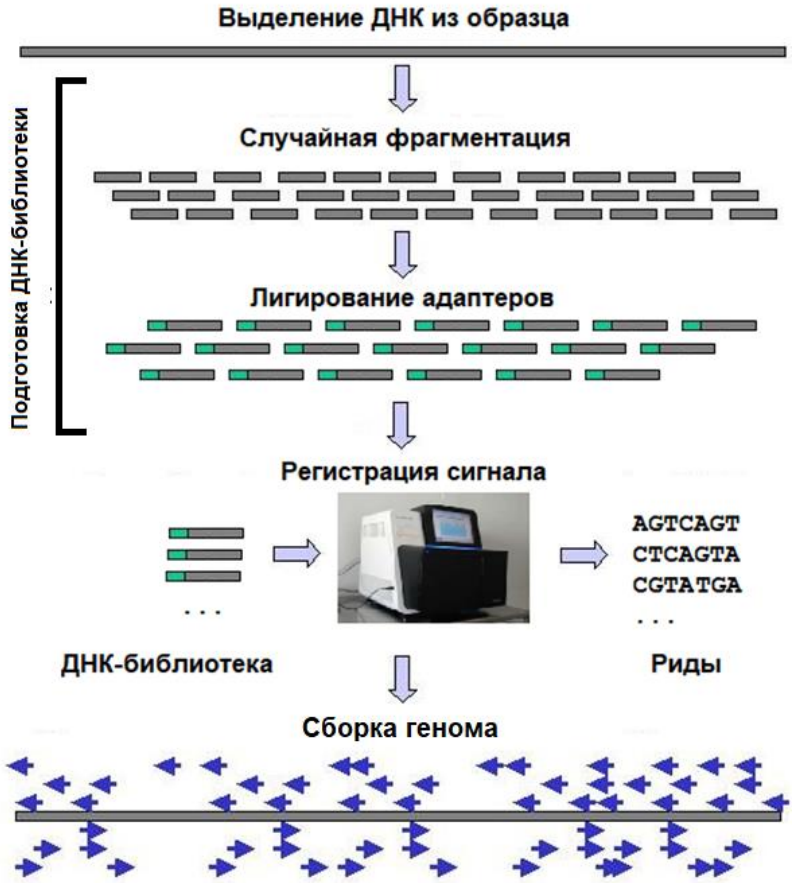


Рис. 1. Основные этапы NGS

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 ридах и искажать результаты анализа.

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

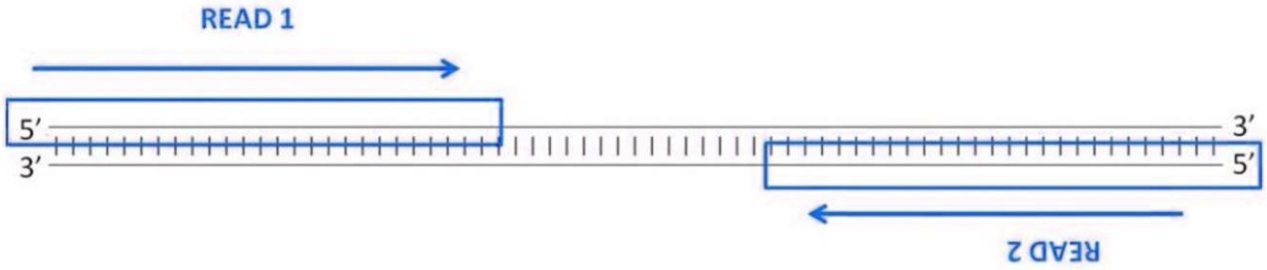


Рис. 2. Прямое (Read 1) и обратное (Read 2) прочтения последовательности

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 Джонса 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 элементах 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.

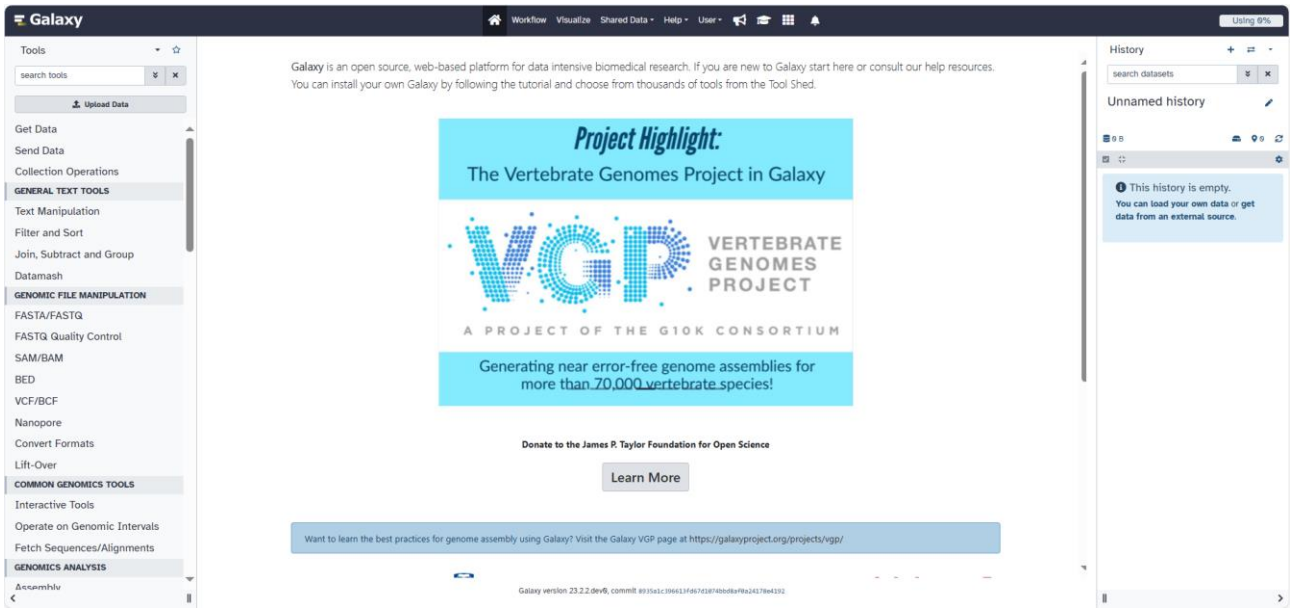


Рис. 3. Главная страница Galaxy

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.

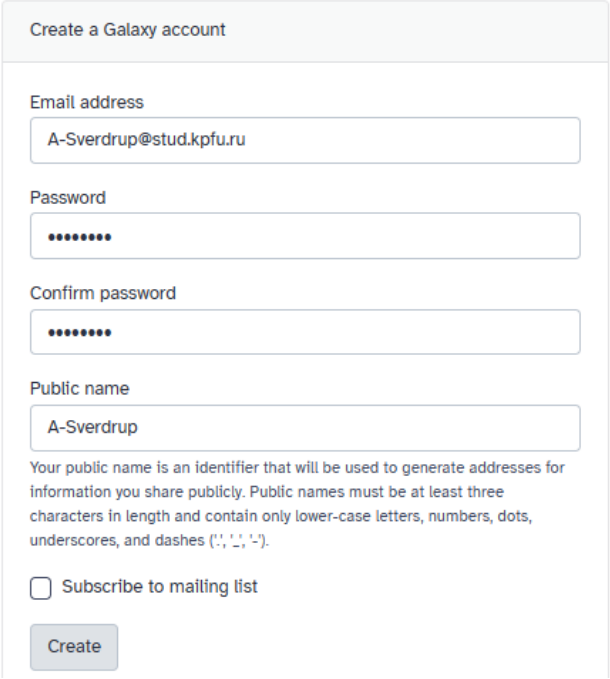


Рис. 4. Форма регистрации

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

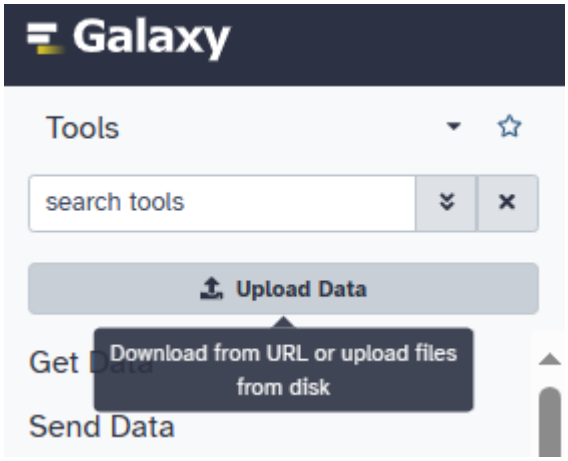


Рис. 5. Выбор команды ***Upload data***

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).

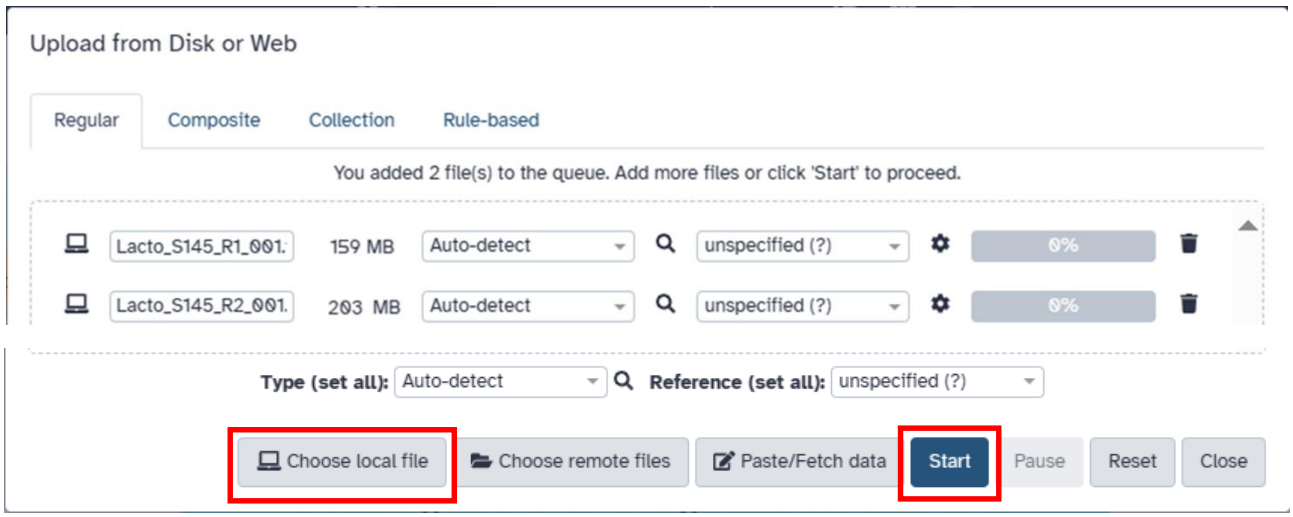


Рис. 6. Загрузка файлов на сервер

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

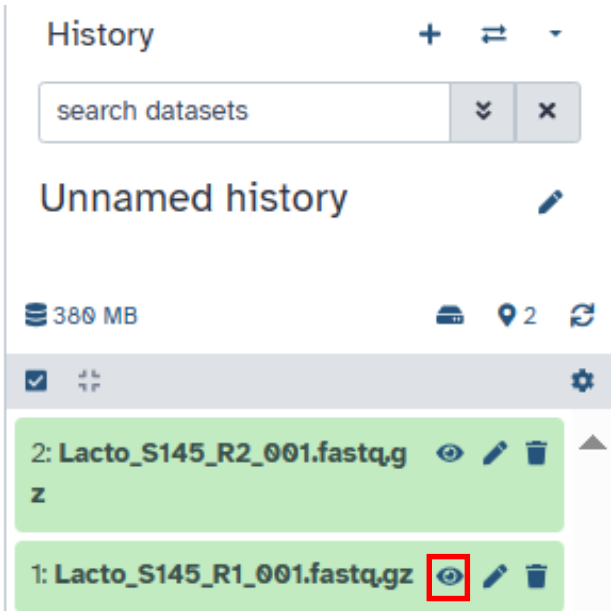


Рис. 7. Загруженные файлы на панели истории

Для просмотра файлов нажмите кнопку 👁 (рис. 7). По умолчанию выводится только начало файла – первый мегабайт данных (рис. 8). Каждая последовательность в Fastq-файле записывается в 4 строки: 1) идентификатор последовательности, начинающийся со знака «@»; 2) нуклеотидная последовательность; 3) знак «+»; 4) закодированное символами качество прочтений Phred Quality Score.

Рис. 8. Предпросмотр файла `Lacto_S145_R1_001.fastq.gz`

Качество прочтений записывается в формате Phred Quality Score – каждый символ в четвертой строке Fastq-файла кодирует значение качества от 0 до 42 (рис.9).

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Рис.9. Кодировка значений Phred Quality Score в символы fastq

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***.

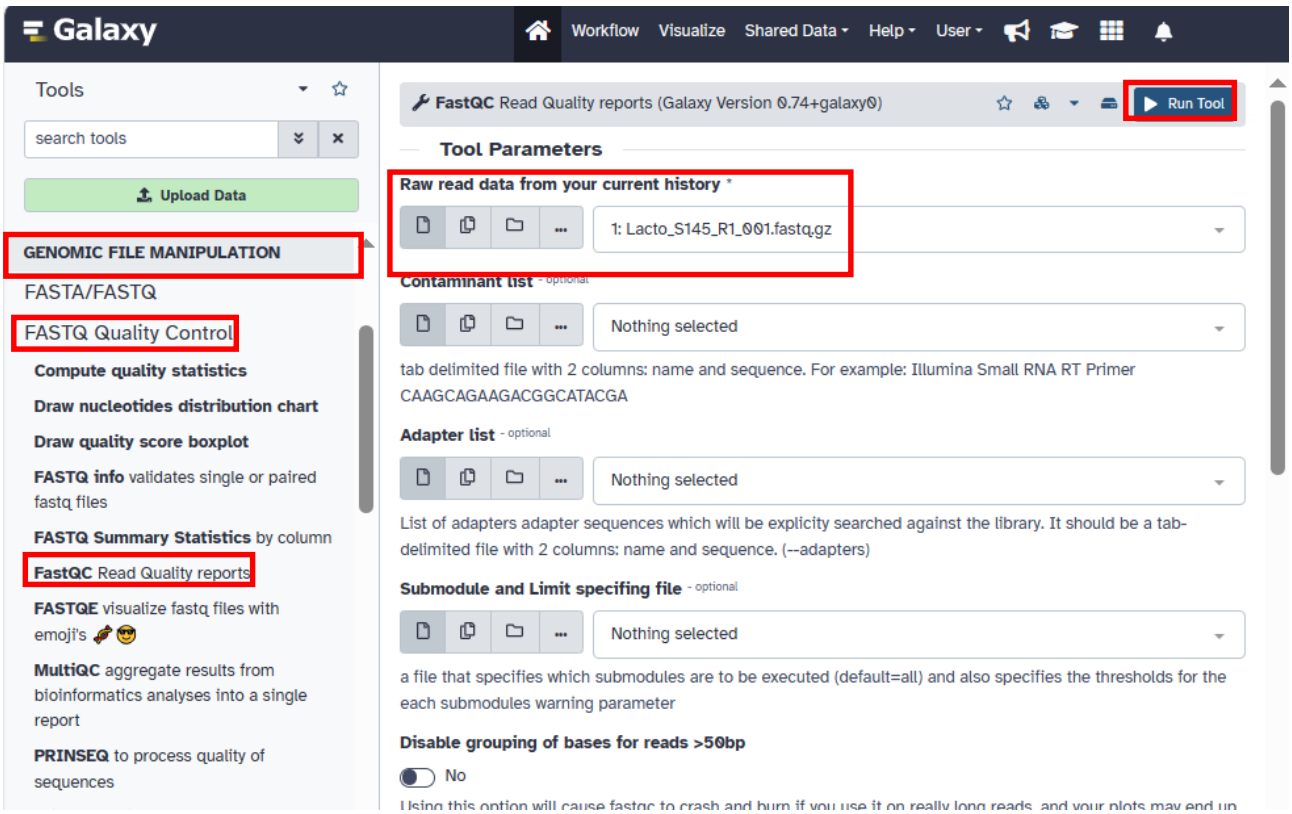


Рис. 11. FastQC в панели инструментов и параметры выполнения

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

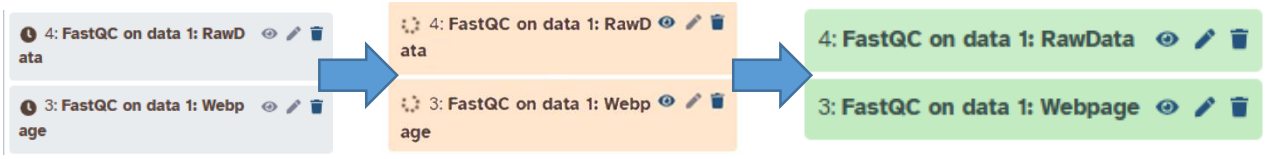


Рис. 12. Процесс выполнения: в очереди (слева), вычисление (в центре), результаты (справа)

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

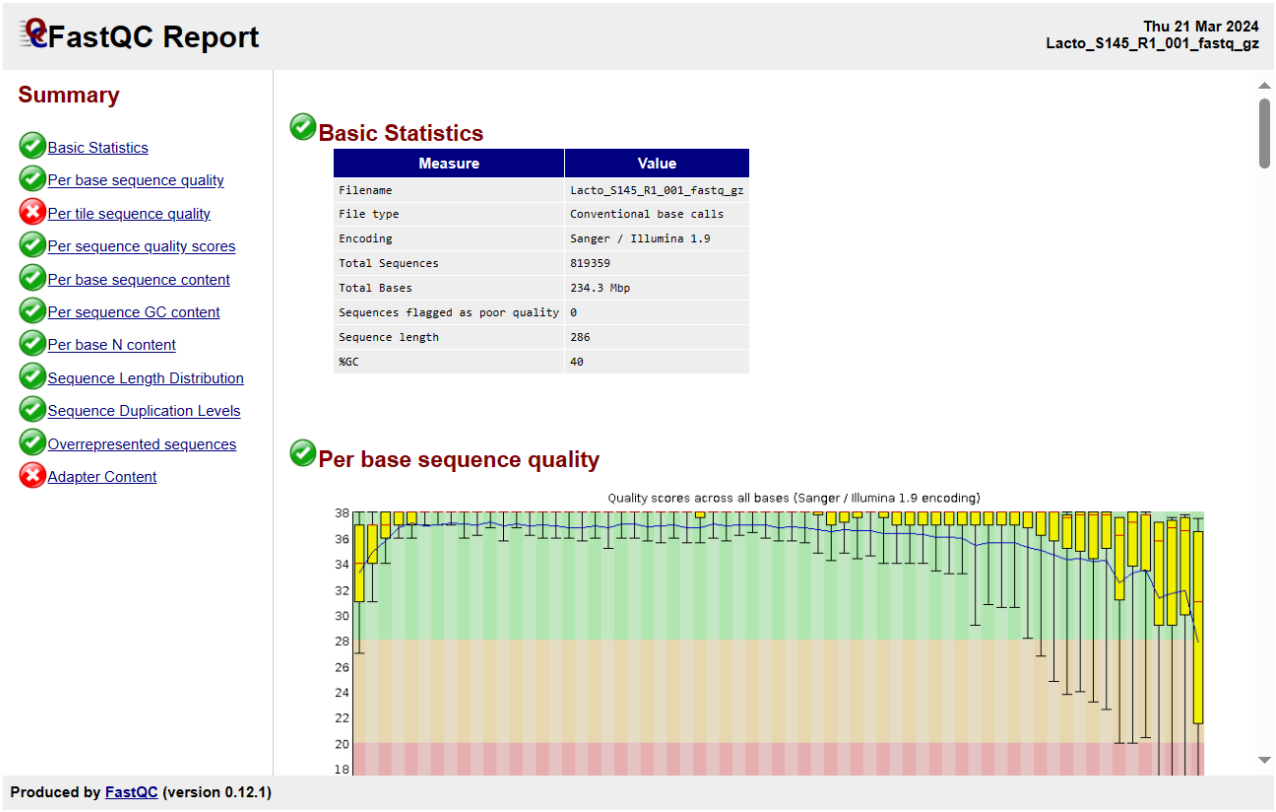


Рис. 13. Отчёт по качеству секвенирования ***FastQC*** для `Lacto_S145_R1_001.fastq.gz`

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 риде, 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).

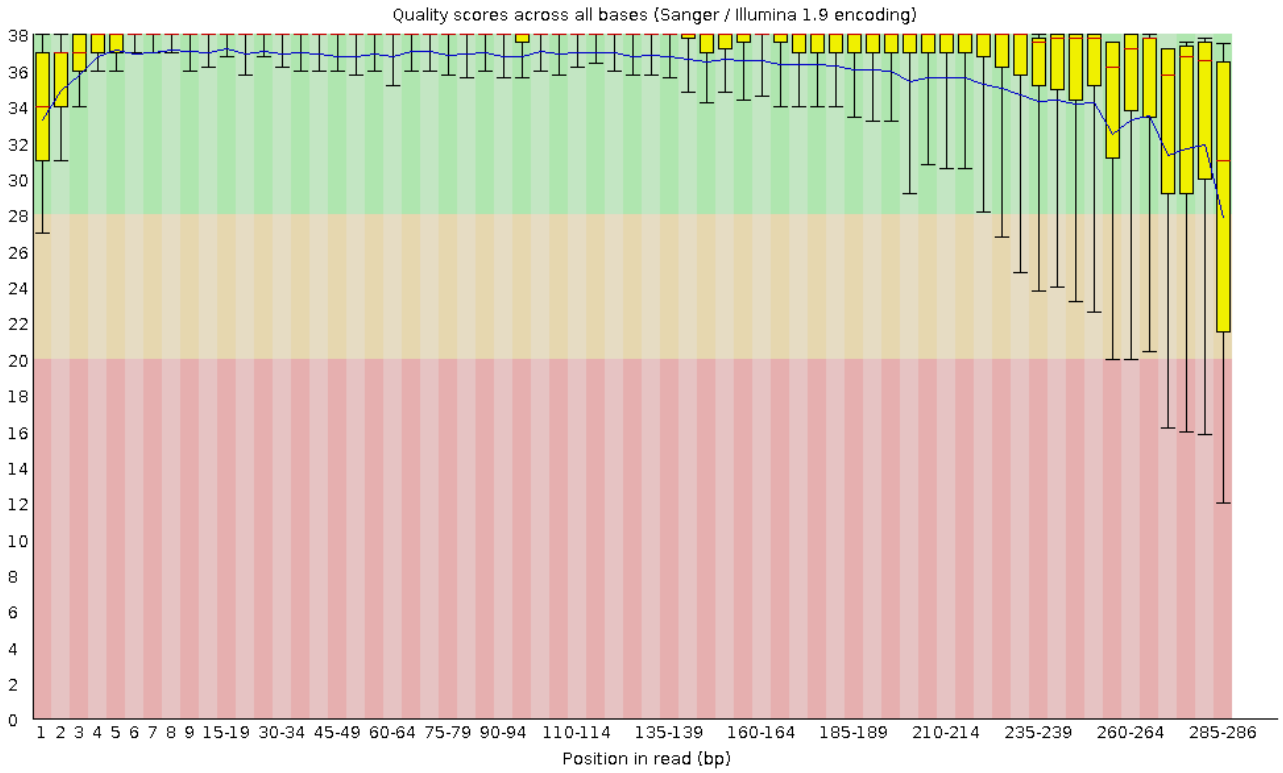


Рис. 14. ***Per base sequence quality*** для `Lacto_S145_R1_001.fastq.gz`

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.

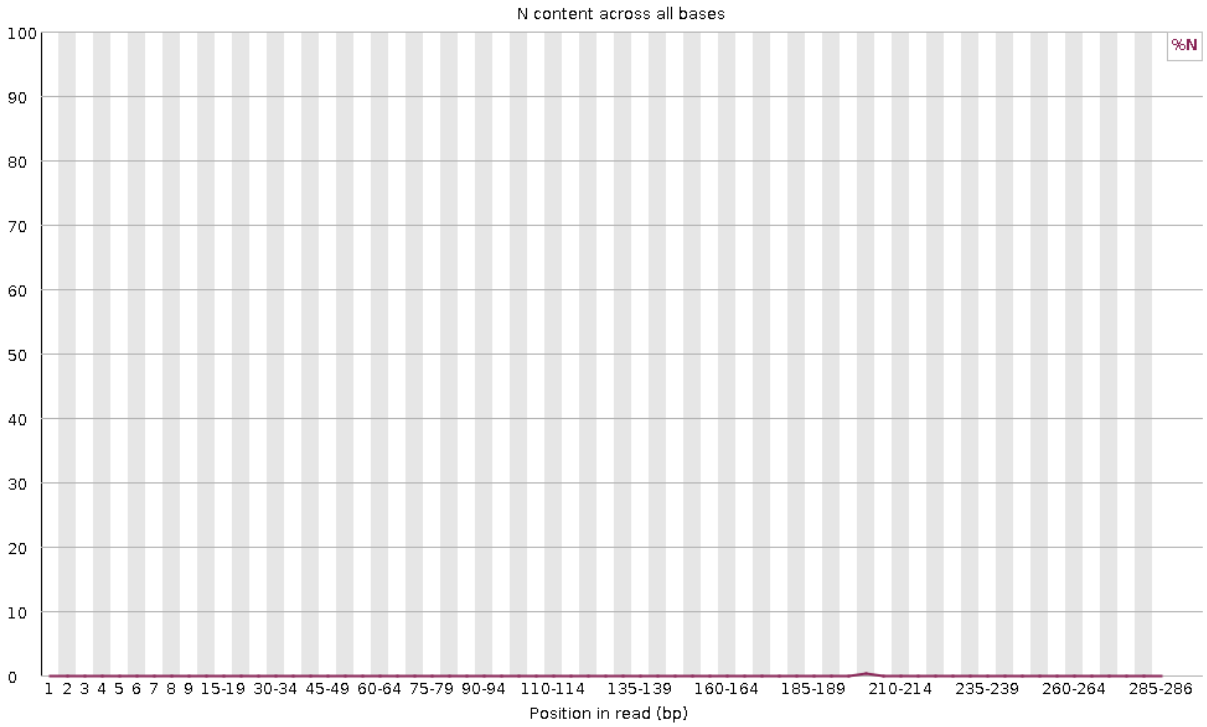


Рис. 15. ***Per base N content*** для `Lacto_S145_R1_001.fastq.gz`

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 (A, T, C, G) в каждой позиции рида в процентах.

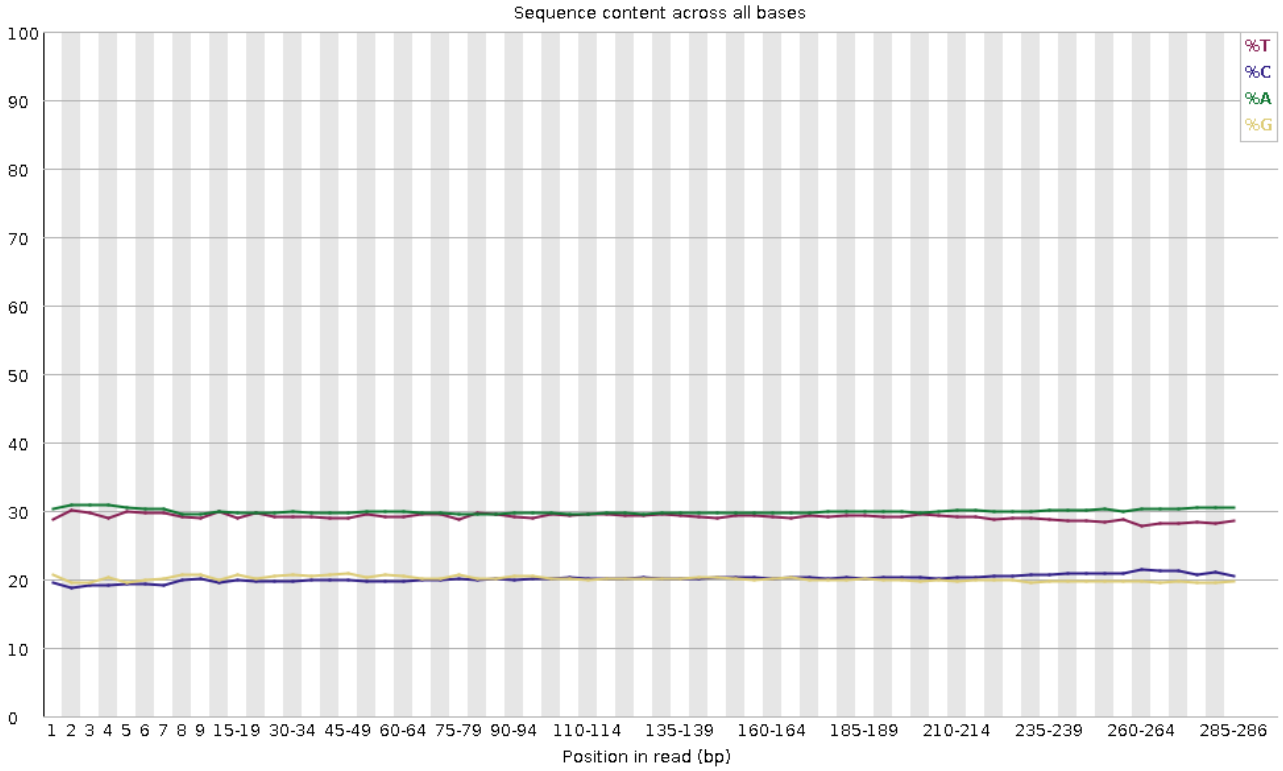


Рис. 16. ***Per base sequence content*** для `Lacto_S145_R1_001.fastq.gz`

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

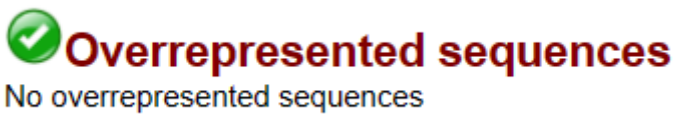


Рис. 17. ***Overrepresented sequences*** для `Lacto_S145_R1_001.fastq.gz`

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

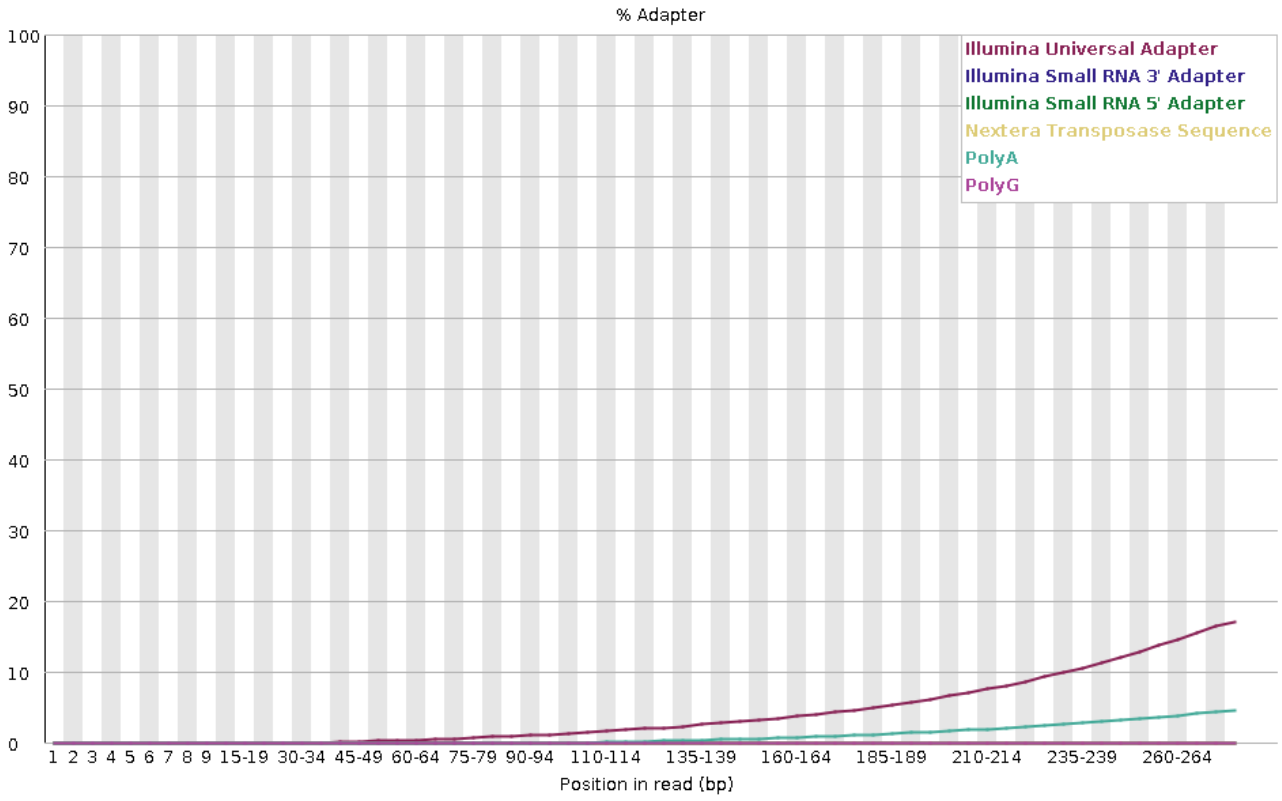


Рис. 18. ***Adapter content*** для `Lacto_S145_R1_001.fastq.gz`

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

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 ридами (`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).

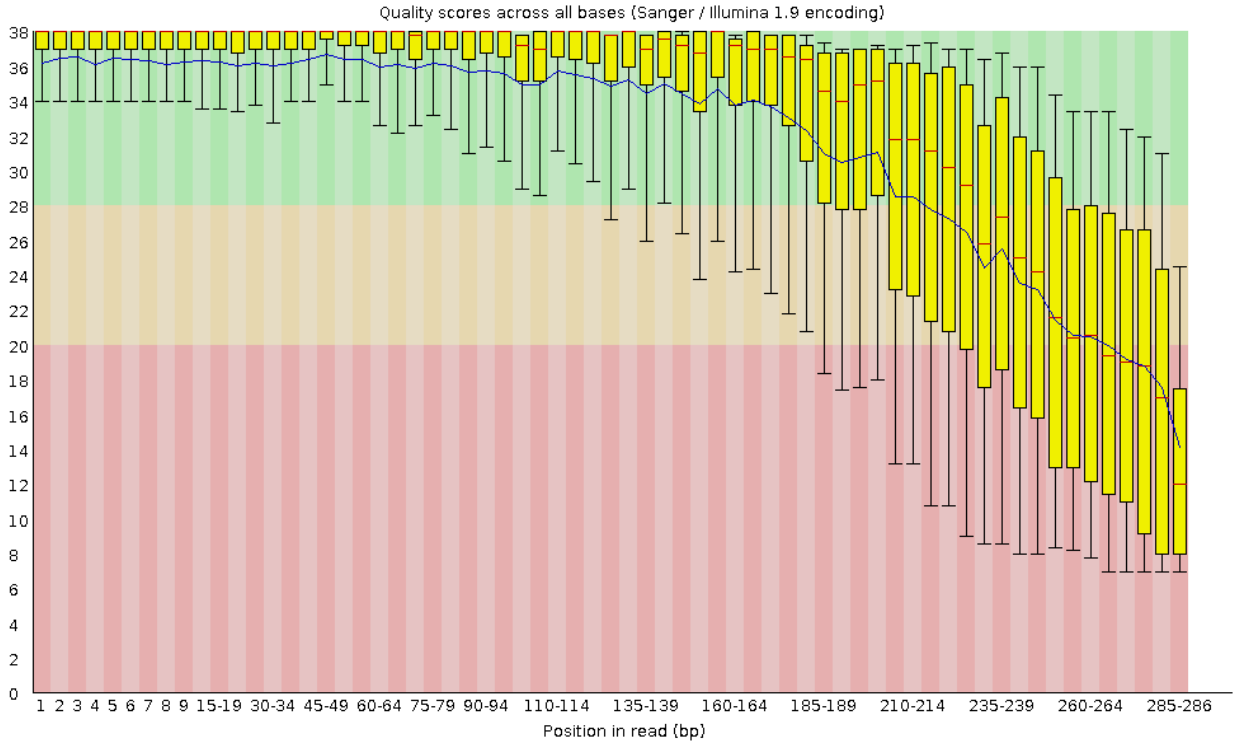


Рис. 19. ***Per base sequence quality*** для `Lacto_S145_R2_001.fastq.gz`

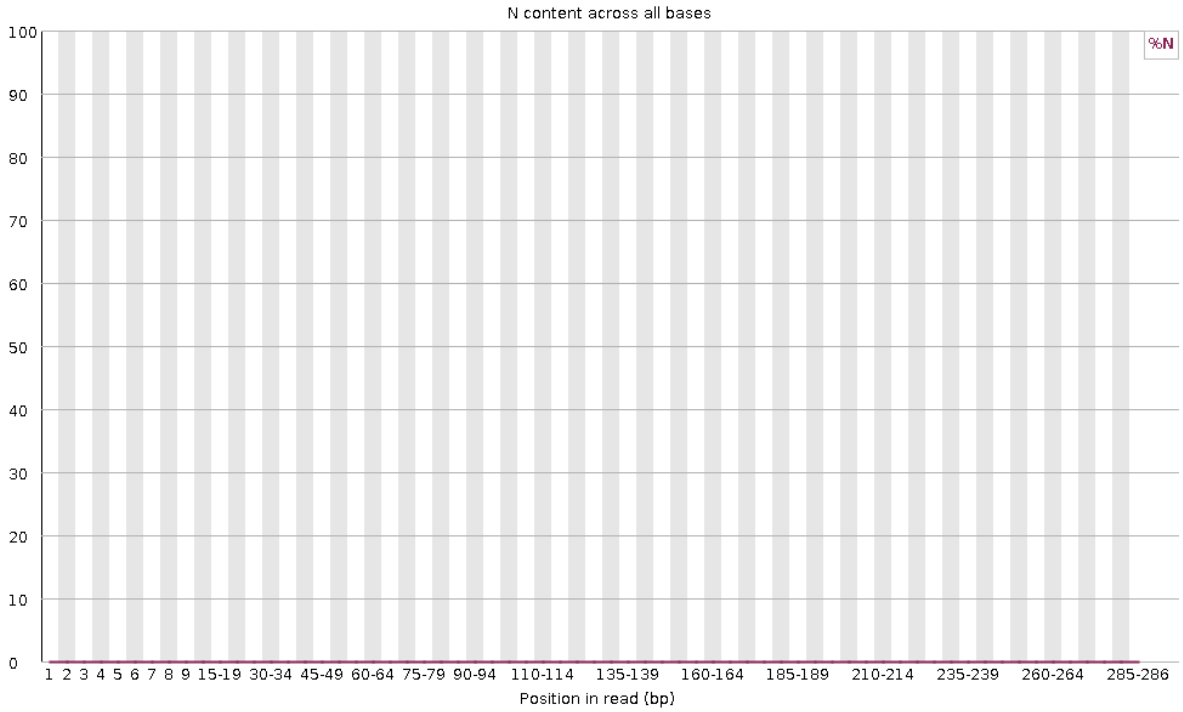


Рис. 20. ***Per base N content*** для `Lacto_S145_R2_001.fastq.gz`

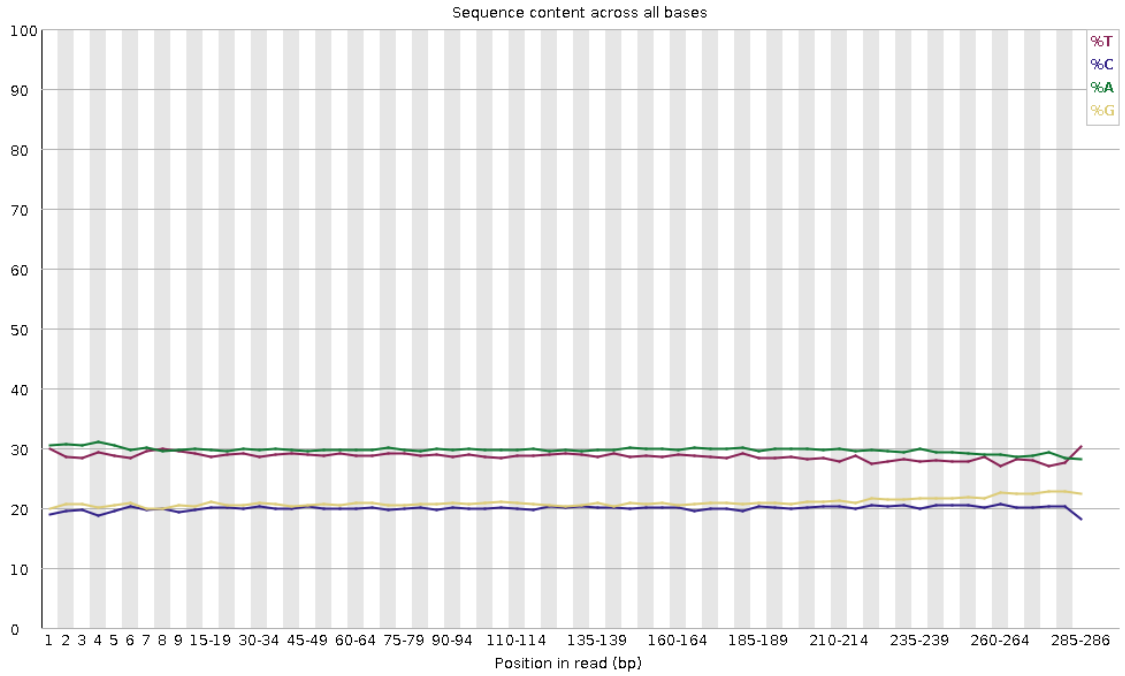


Рис. 21. ***Per base sequence content*** для `Lacto_S145_R2_001.fastq.gz`

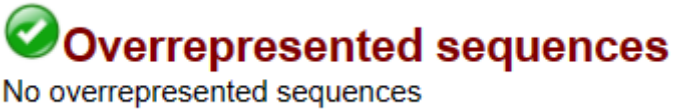


Рис. 22. ***Overrepresented sequences*** для `Lacto_S145_R2_001.fastq.gz`

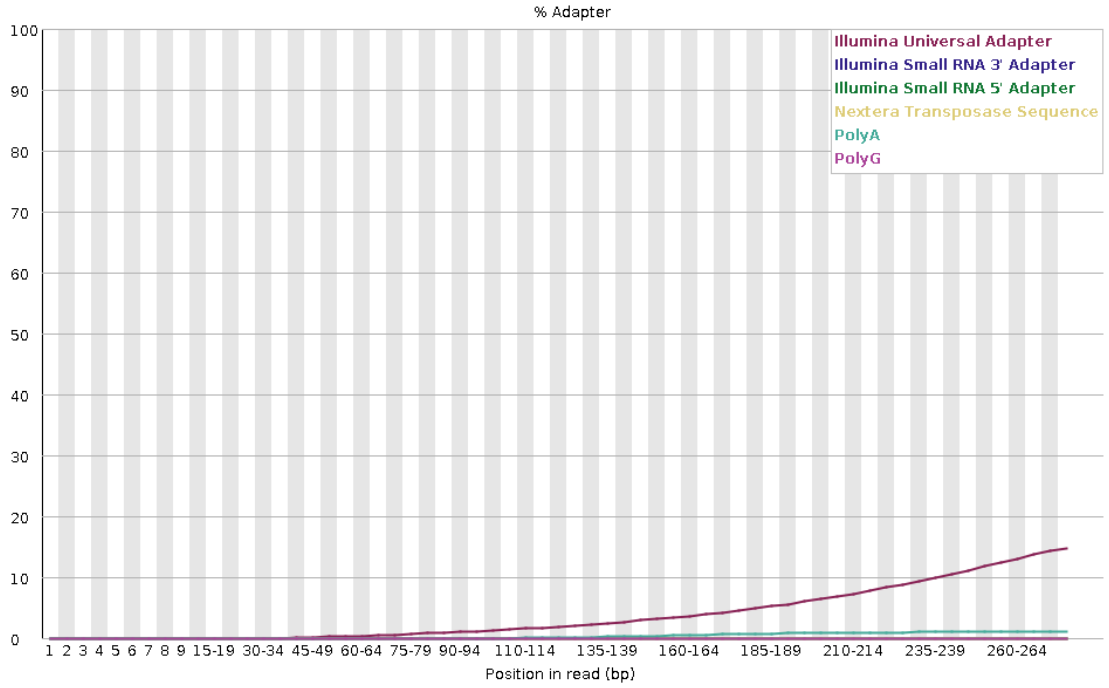


Рис. 23. ***Adapter content*** для `Lacto_S145_R2_001.fastq.gz`

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:	***Cut bases off the start of a read, if below a threshold quality (LEADING)***,
Minimum quality required to keep a base *:	***3***
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***
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***

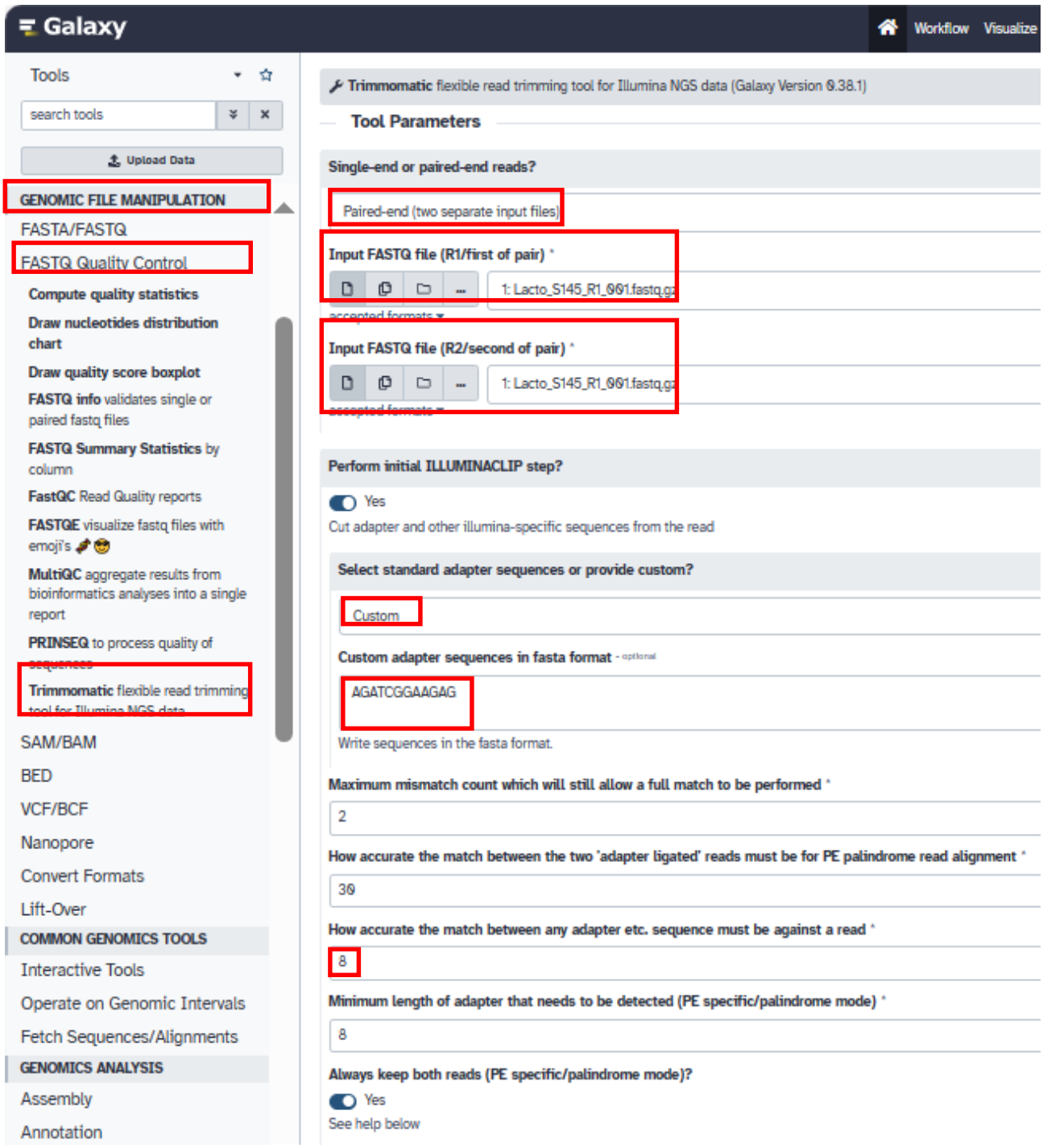


Рис. 24. ***Trimmomatic*** в панели инструментов и настройки выполнения

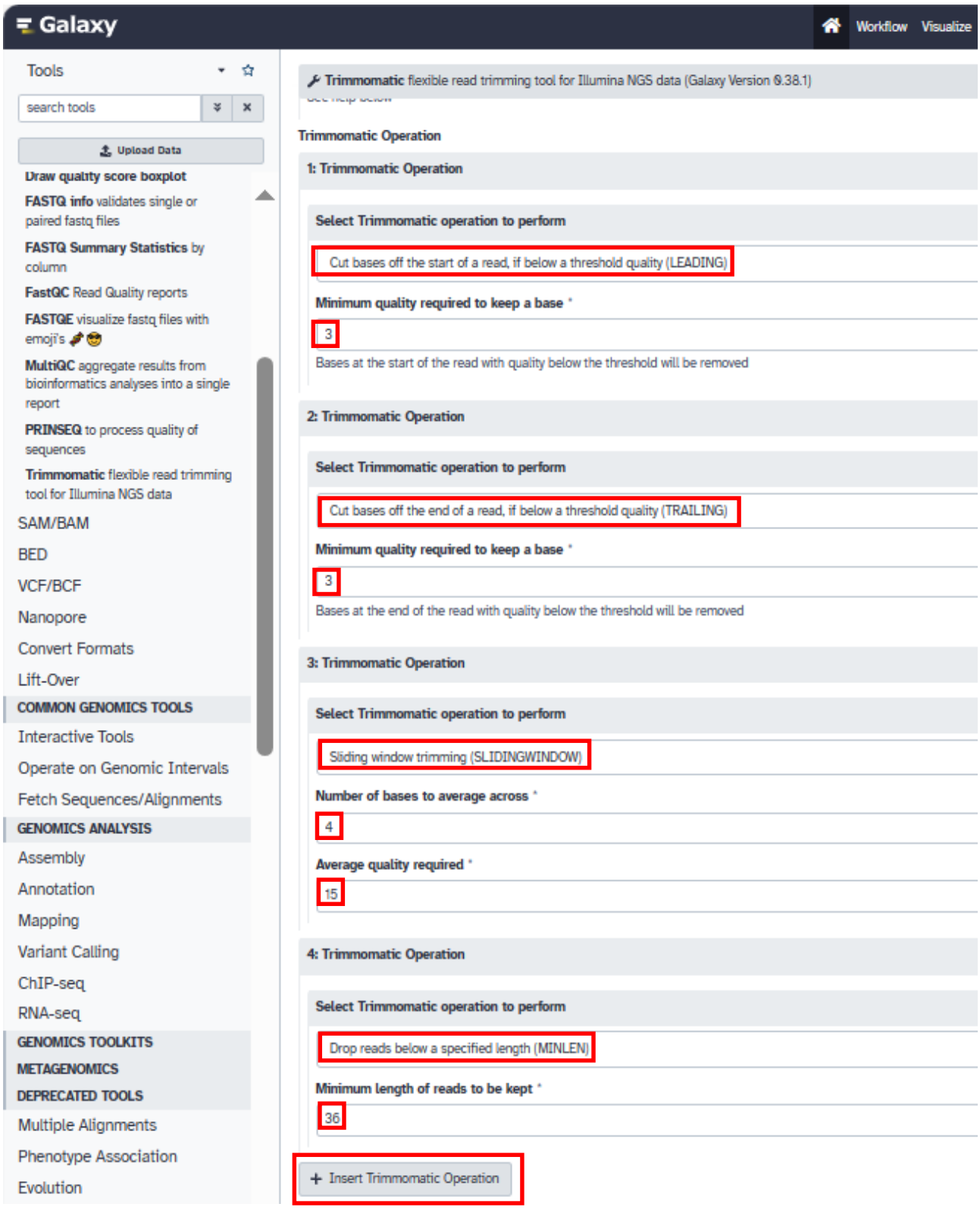


Рис. 24. Продолжение настройки ***Trimmomatic***

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

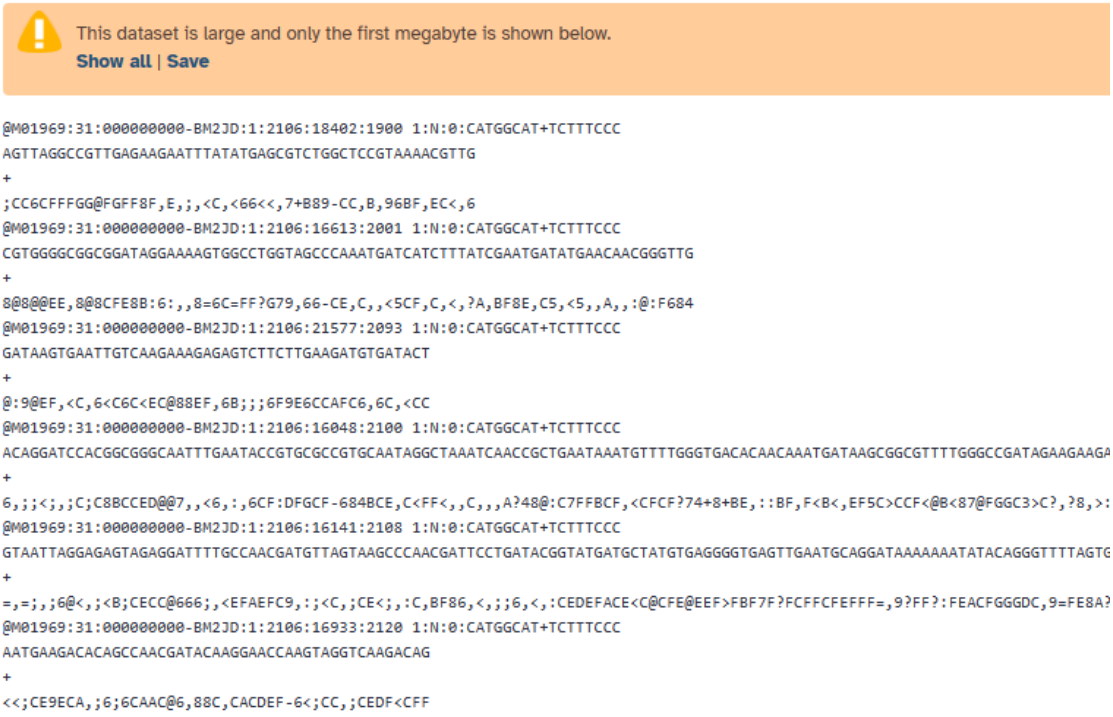


Рис. 26. Просмотр результатов тримминга

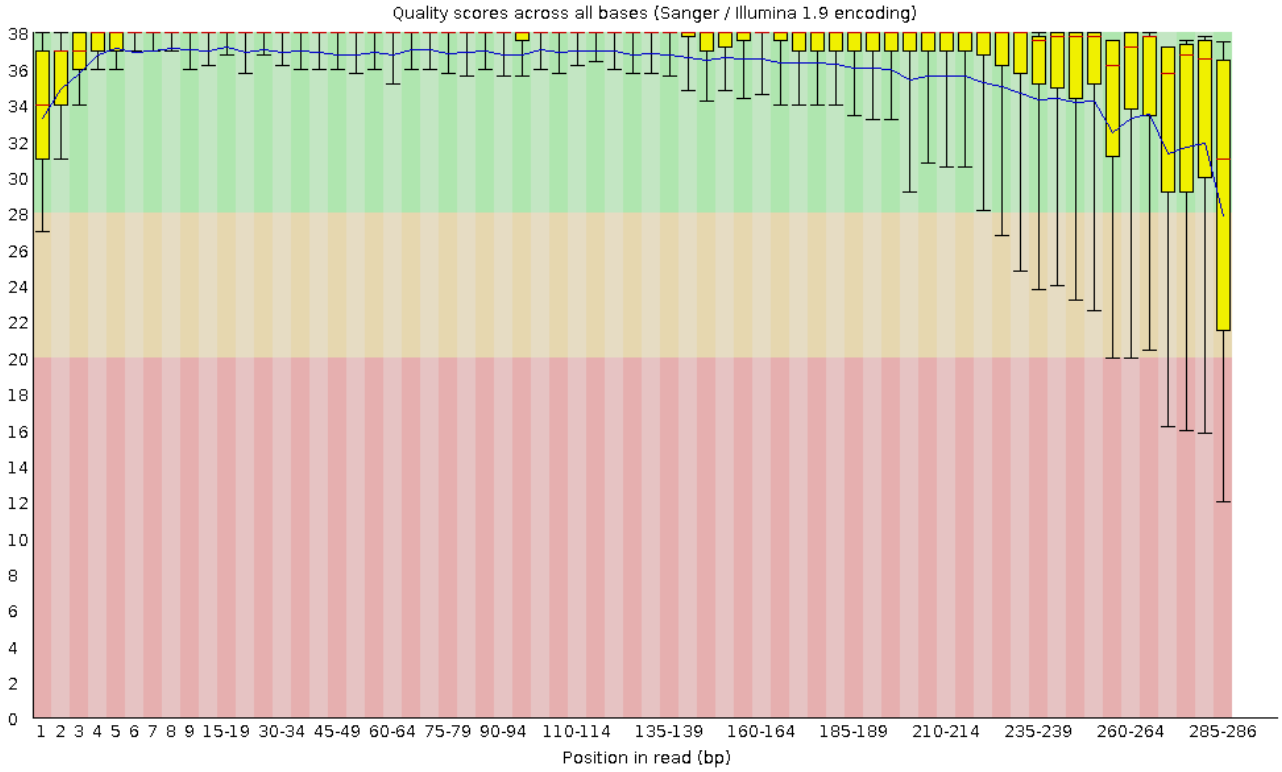
3.4. Контроль качества ридов после тримминга

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

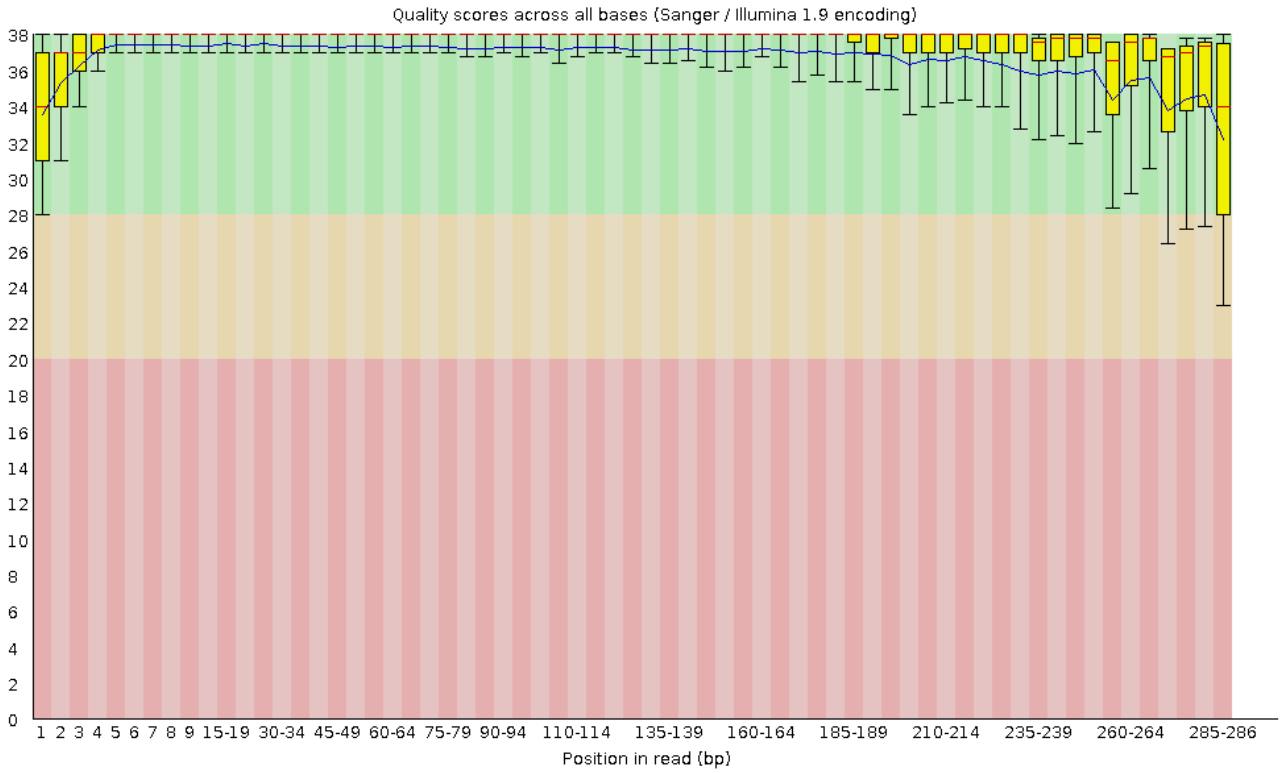
В поле ***Raw read data from your current history*** выберите с прямыми ридами ***Trimmomatic on Lacto_S145_R1_001.fastq.gz*** и нажмите ► ***Run Tool***, затем выполните программу ещё раз с обратными ридами: ***Trimmomatic on Lacto_S145_R2_001.fastq.gz: R2 paired***

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).



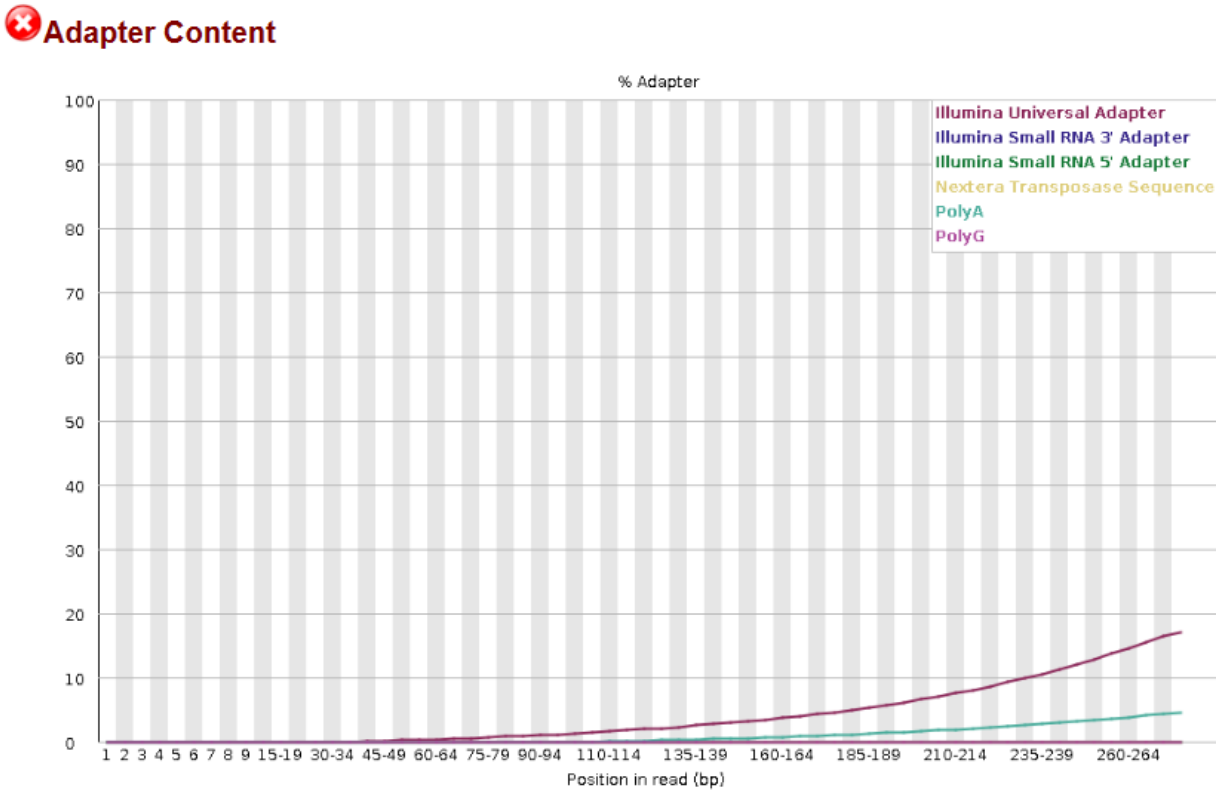
а)



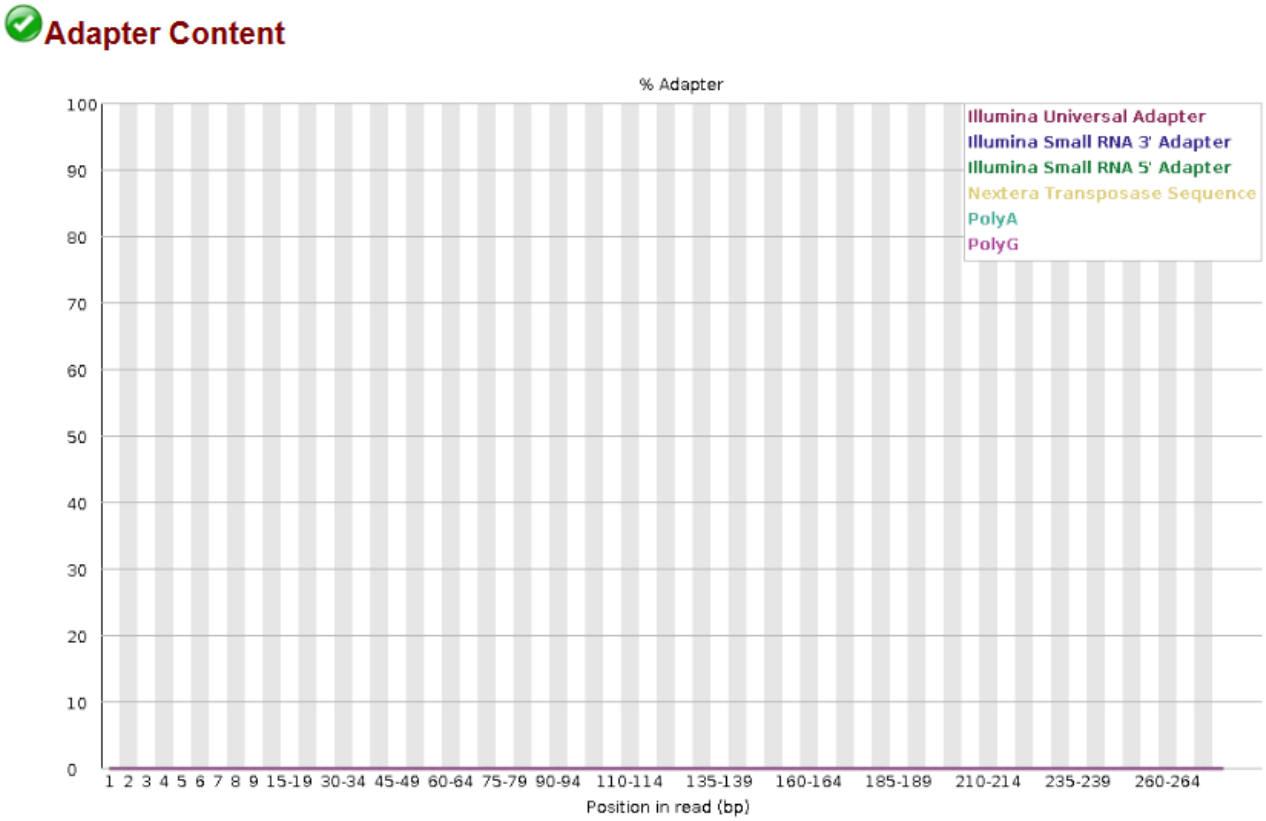
б)

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

а) до тримминга, б) после тримминга



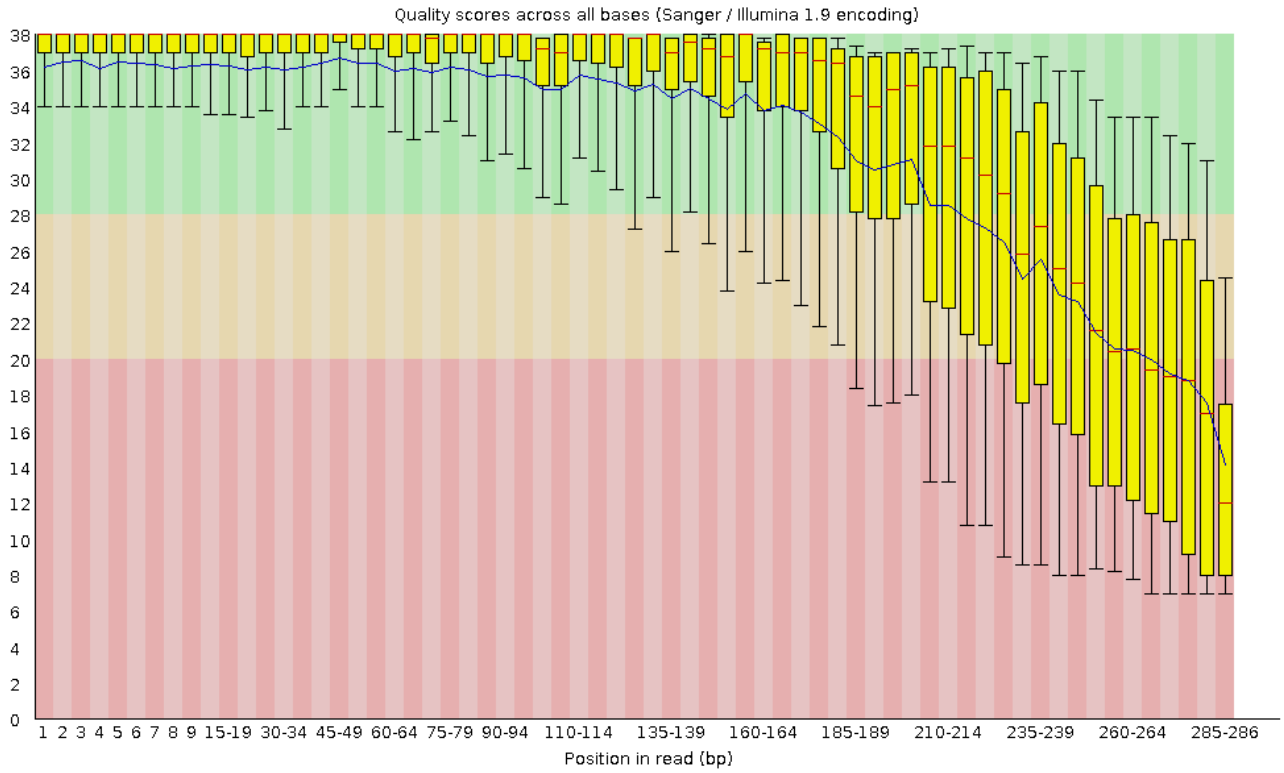
а)



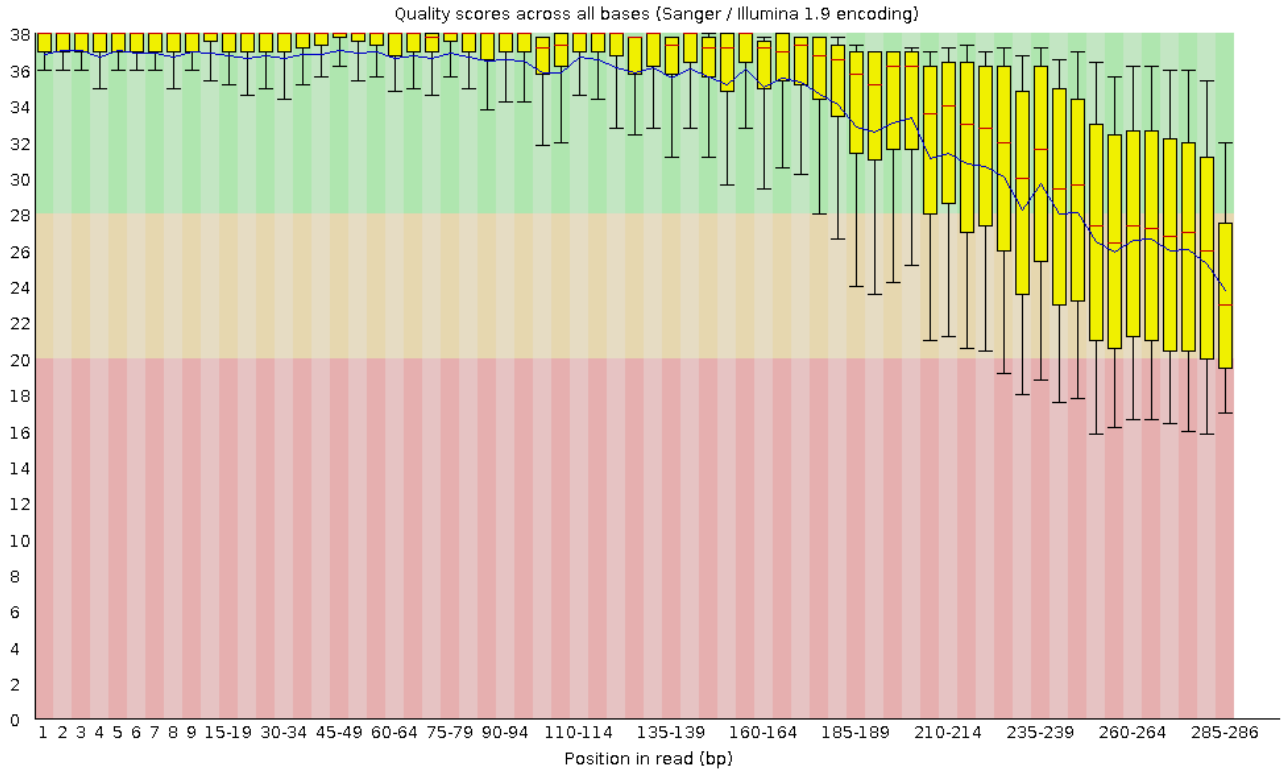
б)

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

а) до тримминга, б) после тримминга



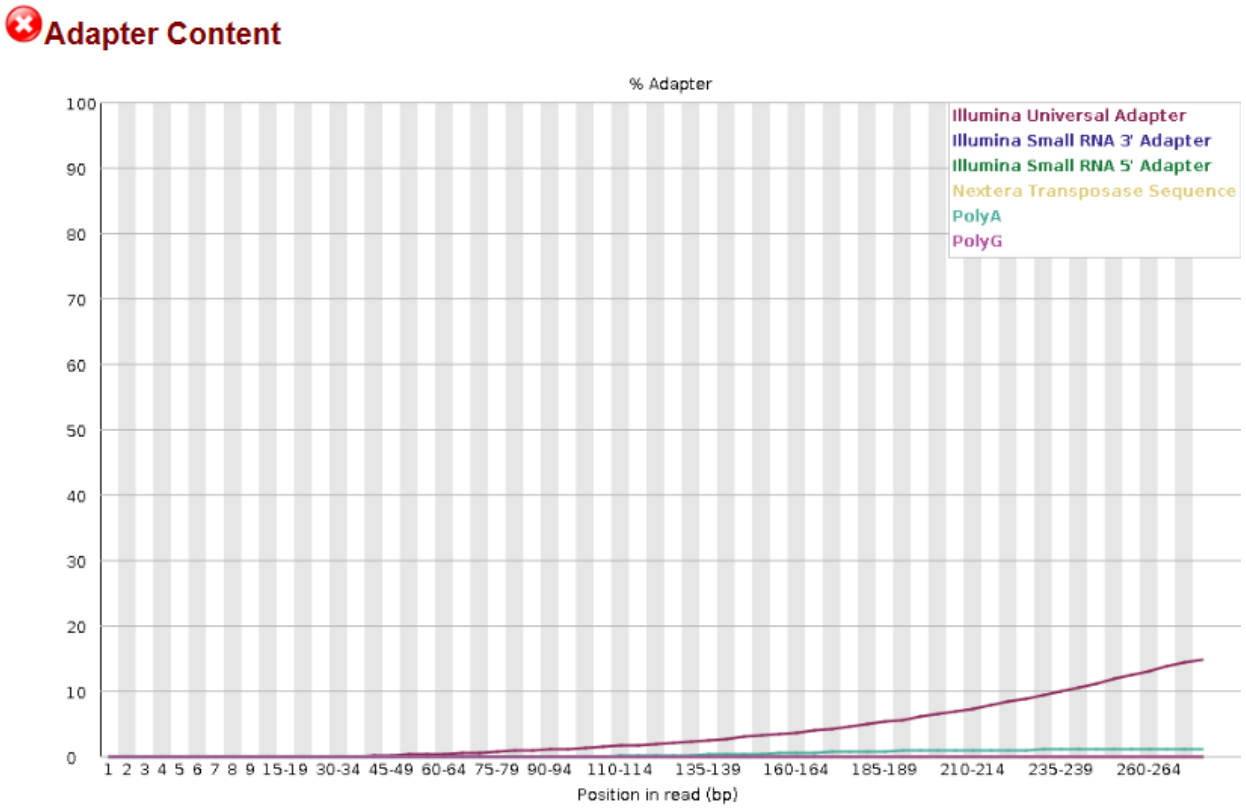
а)



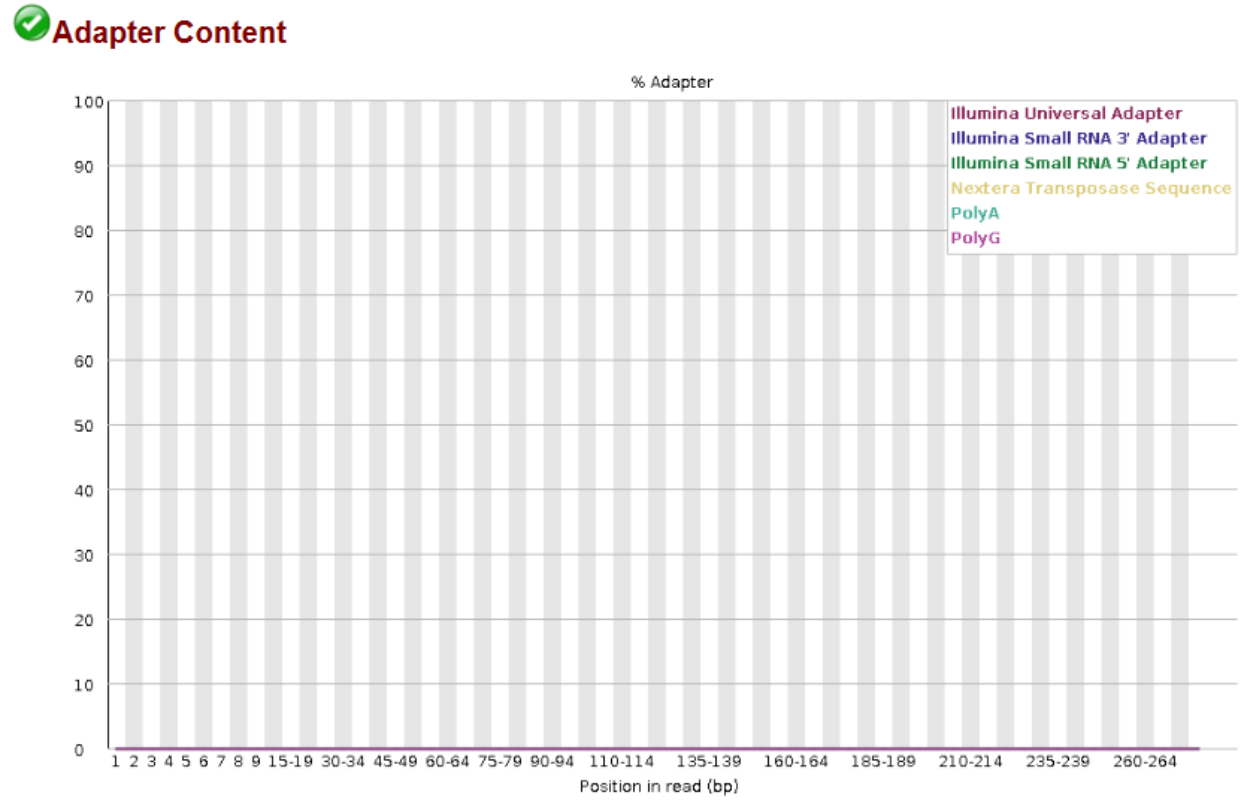
б)

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

а) до тримминга, б) после тримминга



а)



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 ридами ***Trimmomatic on Lacto_S145_R1_001.fastq.gz: R1 paired.***

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

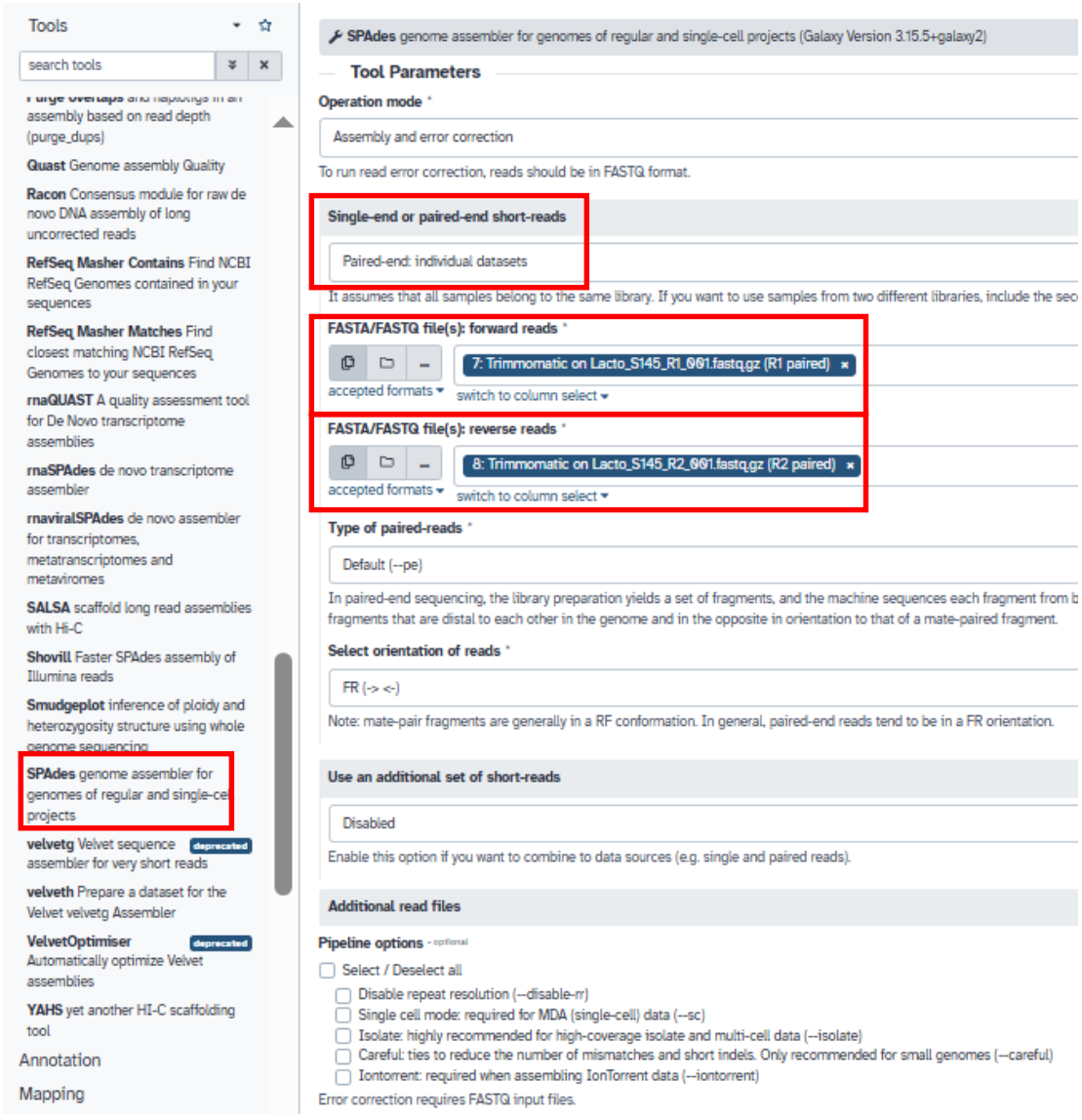


Рис. 31. Настройка SPAdes

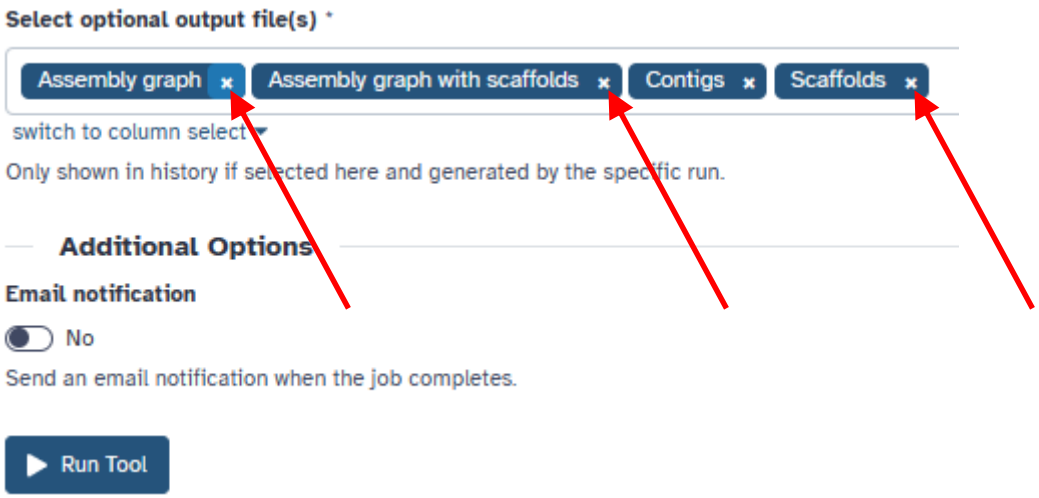


Рис. 32. Настройка SPAdes (продолжение)

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).

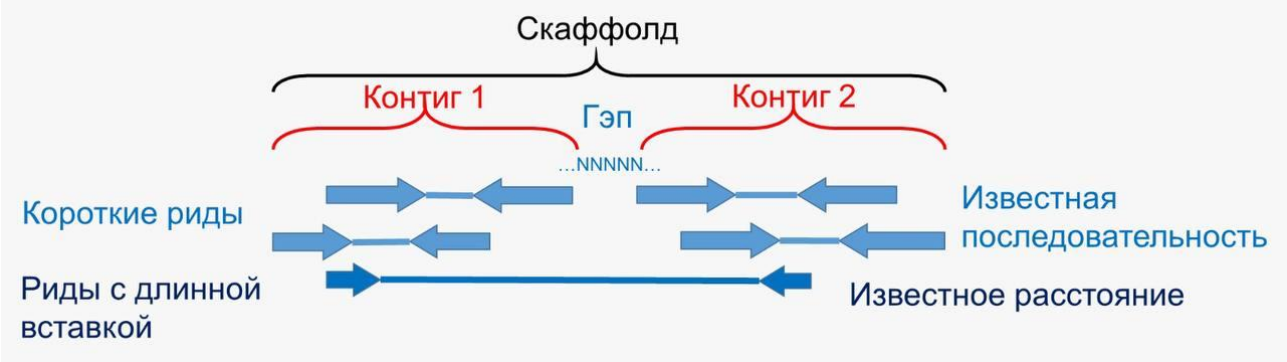


Рис. 33. Принцип сборки генома: риды – контиги – скаффолды

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).

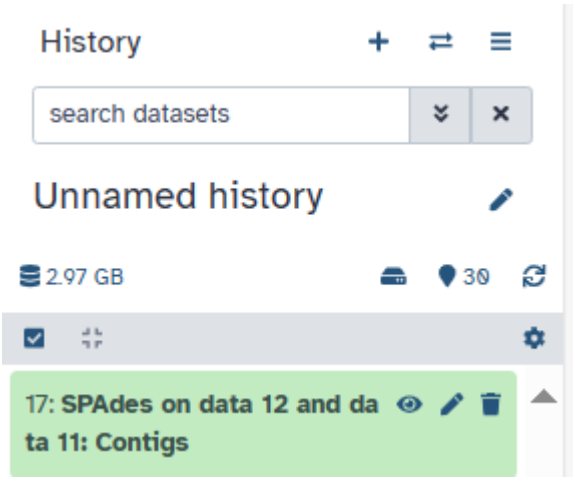


Рис. 34. Результат сборки генома на панели истории

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.

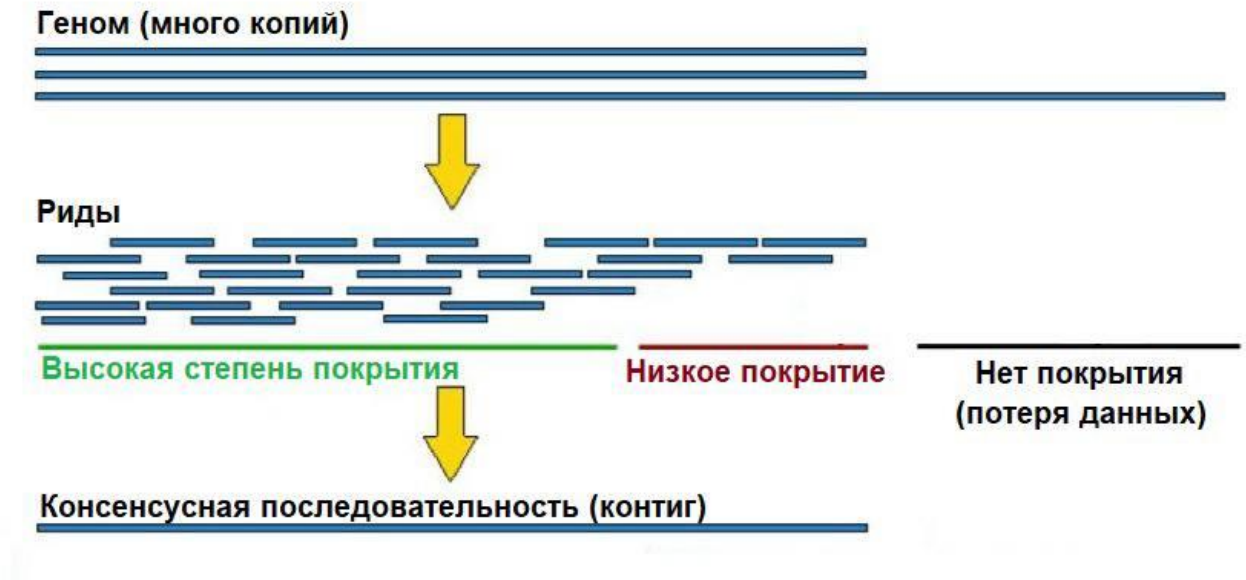


Рис. 36. Принцип сборки генома: степень покрытия контига ридами

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).

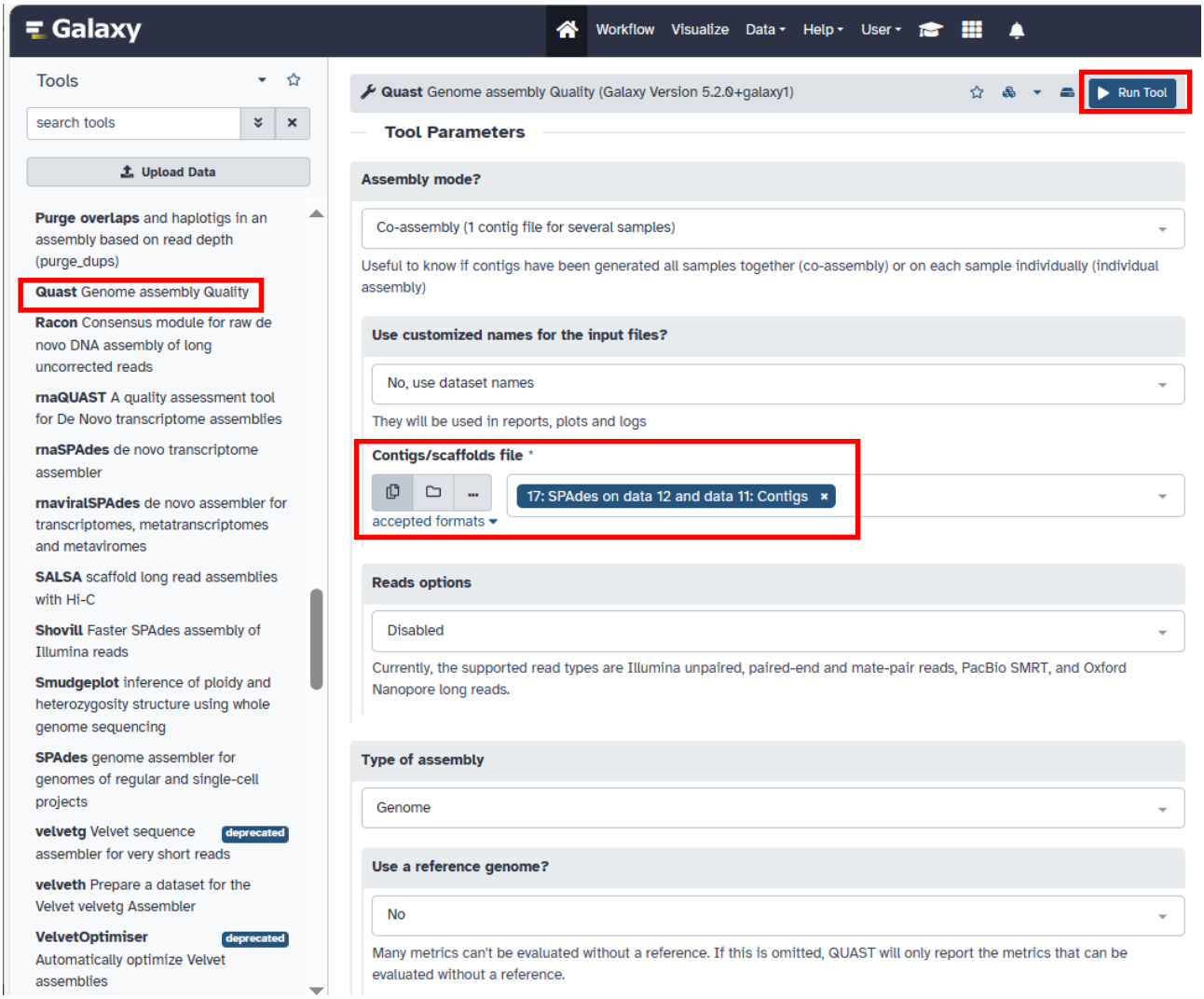


Рис. 37. ***Quast*** в панели инструментов и настройки выполнения

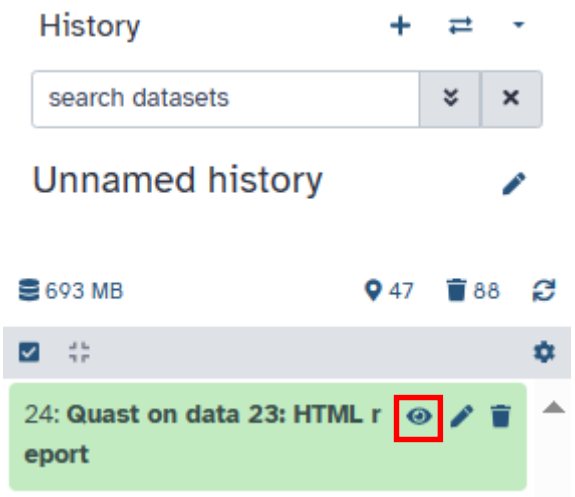


Рис. 38. Результат выполнения программы ***Quast*** в панели истории

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 Nx-значения для x = 0-100%.

auN = 68125 – площадь под графиком Nx-значений

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

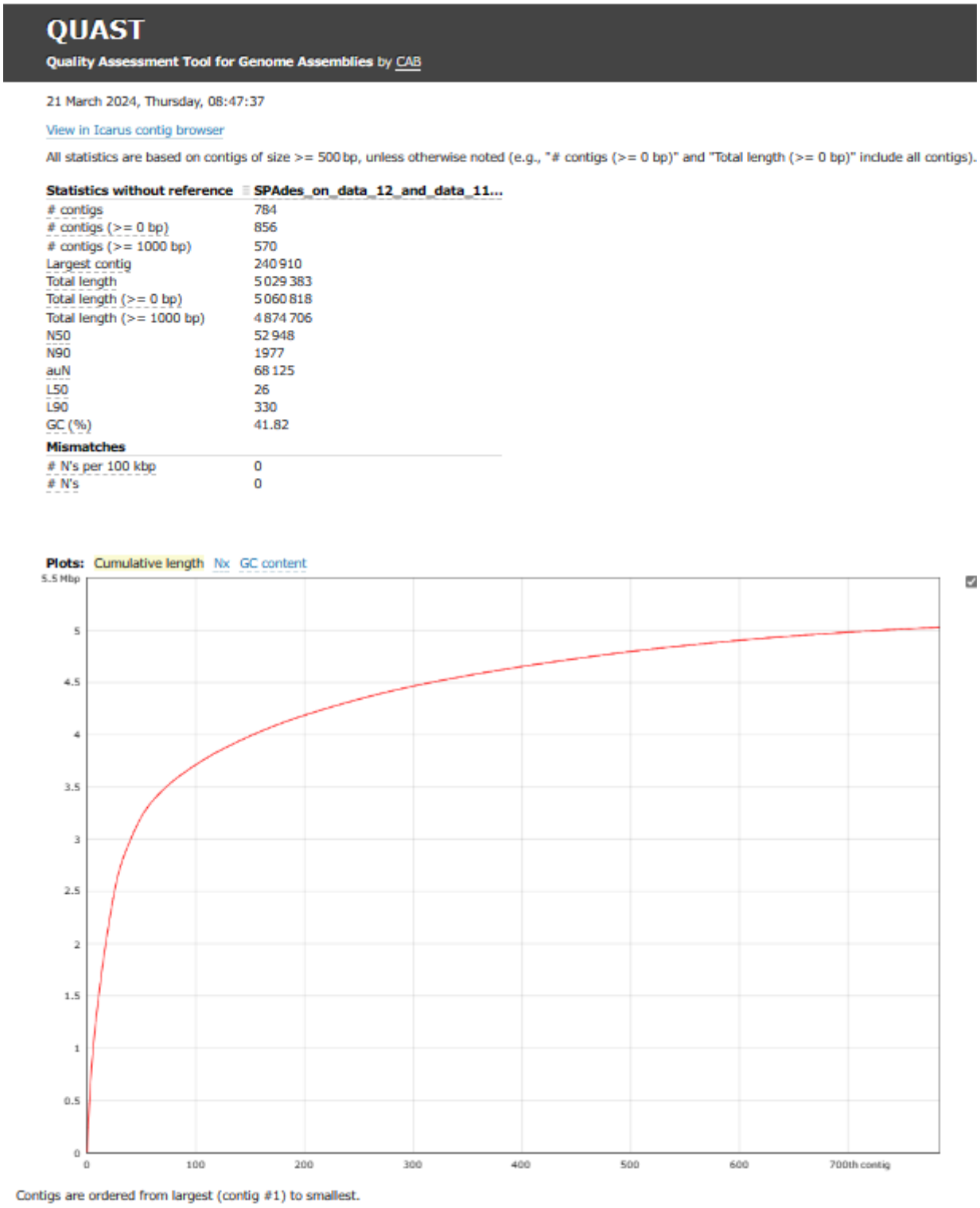


Рис. 39. Отчёт **Quast** о качестве сборки генома – таблица по статистике и кумулятивный график длины

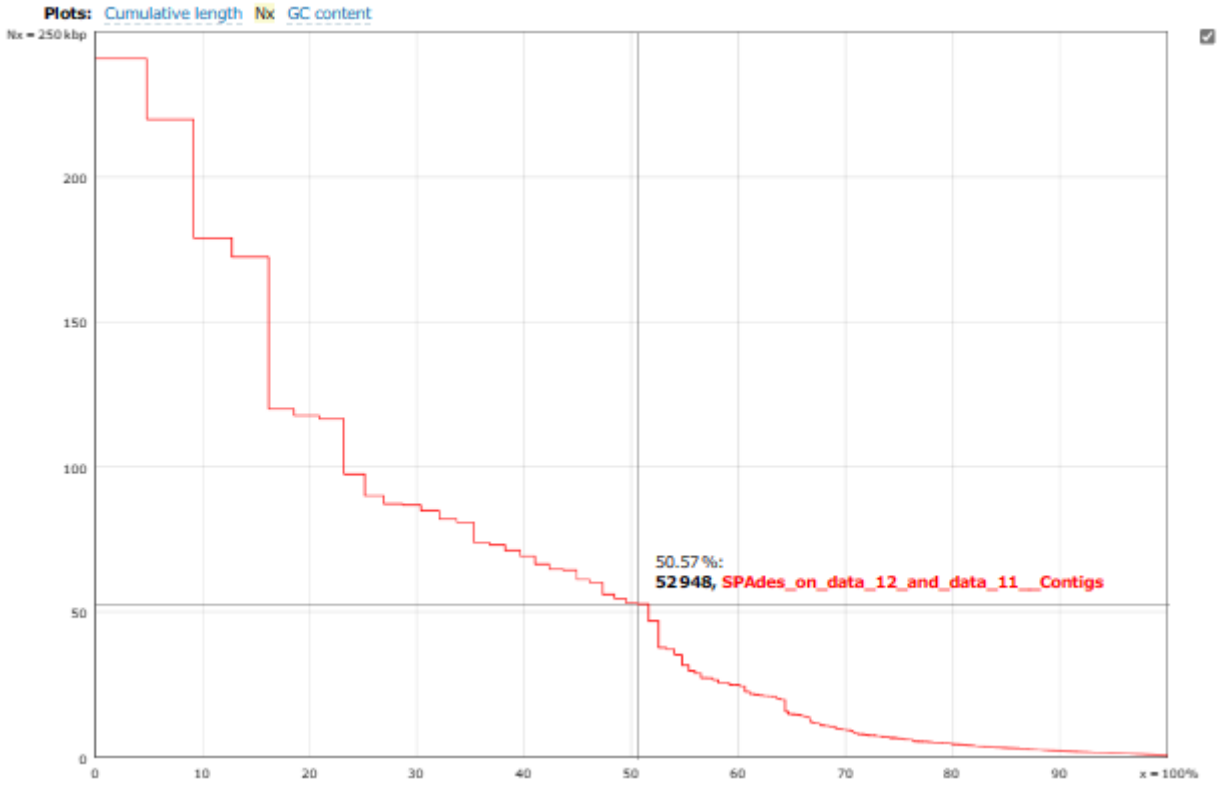


Рис. 39. (продолжение) График ***Nx***-значений, ***N50*** выделено

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).

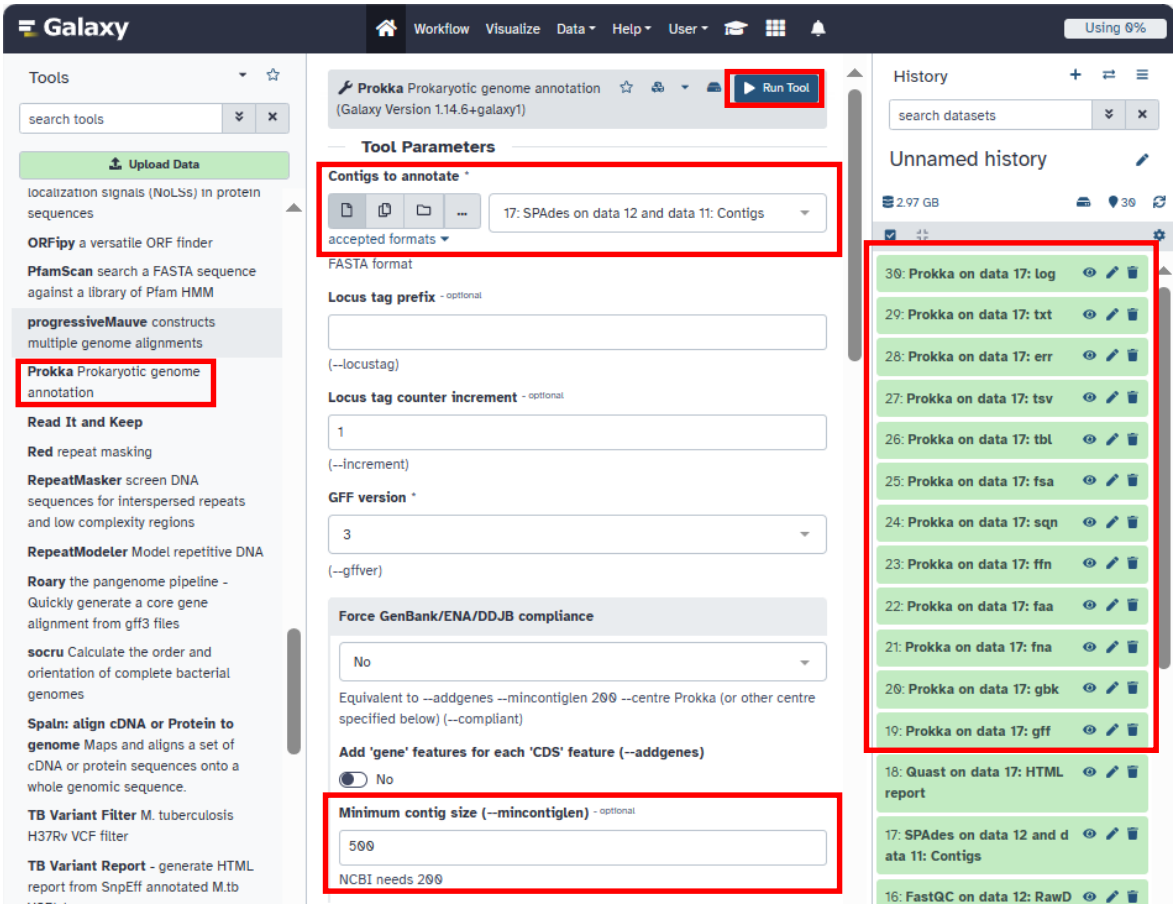
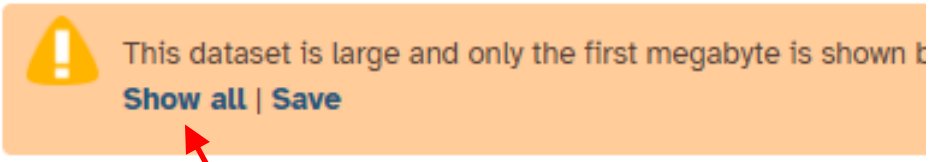


Рис. 40. ***Prokka*** в панели инструментов, настройки выполнения, результаты аннотации в панели истории

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
TTTCTGGCCAATAAGCAGATATTTCTGAGTGAACTAAAACTTTTAAAATCTGTTTTAGAC
TTAGGCTCAACAAAACATGTCTTATCAATCGGTTGTGGCAGCGGGTTATTTGAAGATACC
CTCCACCGCCAATATGGCCTGCCACTTTTTGATGGTGTTGAACCATCTACCGATATGGCA
CAGATTGCAAAAAAGCGGGGATTAAACGTTCGCGTTGGAAAAGCAGAAAGTATTAATCTG
CCCGCAAGACAGTATGATACCATTTACTTTAATGGCAGCTCCAGTTATATTAAGGACTTA
TTGGCAGCTTATCAAAATTGTATCAAAGCACTCAAGTCCGGCGGCCATTTTATTTTAATT
GACGTGCCAAAAGAAAGTGCTTATGGATTAATGTATATGCTTGCCAAAATTCAGGGTAAT
TATCAATCAGAAGAATTGGCAAATGTTCTTCCGAAATGGCCATACCCAATTGAACTGGTT
GATTCTGCATACTGGCATACCACCCTAGAAAAAAAGCACCTTTTAGAAAACGAGCTTCAG
CTACGACATTTAAGGTTCAAACAAACCTTGGCAGCAAACCCACTCTACACAAACGATAGT
GTTGAAGAACCTGAAGATGGCTTTTCAAAGGGTGGTTATGTCGCAATAATTGCCGAAAAA
CCTTAA
>JBGHFDFB_00002 Peptide methionine sulfoxide reductase MsrA
ATGGAAGATACAGCAATTTTTGCCGGTGGGTGTTTCTGGTGTATGGTTAAACCGTTTGAT
ACCATGCCCGGTATTAATAAAGTTATTTCCGGATATACGGGTGGTCATGTAGCAAATCCG
AGCTACGAACAGGTTAGCAGCCATACAACGGGACATACCGAGGCCGTTAAAATCAATTTT
GATCCAGATATTATTAGTTATAAGCAACTGGTTCAAATTTATTGGCAGCAAACAGATCCA
ACAGATGCAATGGGACAGTTTCAGGATCGAGGAGATAACTATCGTCCAGTGATCTTTGTG
AAAGATGAAGCGCAACGAAAAATTGCGGAGGCTTCCAAAAAAGCTTTAGAGGAAACCGGT
ATGTTTGACCAGCCGATTGTCACTCAAATTGAAGAGGCCAAACCCTTTTATCCAGCAGAA
GAGGAACATCAACAGTTTTATAAAAAGAATCCATTTCGCTTCCAAATGGAGGAAGCTGGT
GGACGTGAAAAATTTGTGAAAGAACACTGGCAGTCAAAACAAAAATAA
>JBGHFDFB_00003 Peptide methionine sulfoxide reductase MsrB
ATGGTTGAGAAAAAAGATTTAAAACATCGATTAACACCAGAACAATATGCGGTTACCCAA
GAGGCCGCTACGGAGGCACCCTTTAGTGGTCAATATGATCATTTTTATCAAGAGGGTATT
TATGTTGACGTCGTAAGTGGTGAACCTTTGTTTTCATCTAAGGACAAATATGACGCTGGG
TGTGGATGGCCATCATTTACCAAACCAATTGAGCAAACAAGTATTCATAAGAATTTGGAT
ACTTCTCACGGGATGATTAGAGAGGAAGTTCAAAGTAATGATGCTCATTCTCATCTAGGG
CATGTGTTTGGCGATGGGCCAAAAGAAGCAGGCGGCCAACGATATTGTATCAATTCCGCG
GCCTTGAAGTTCATTCCCACAGACGAGCTTGAAGAGGCAGGGTACGGTCAGTATAAACCA
ATGTTTAACTGA
```

Рис. 41. Фрагмент файла в формате ***ffn***

5. Определение таксономии исследуемого генома

5.1. Извлечение последовательности *16S рРНК*

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).

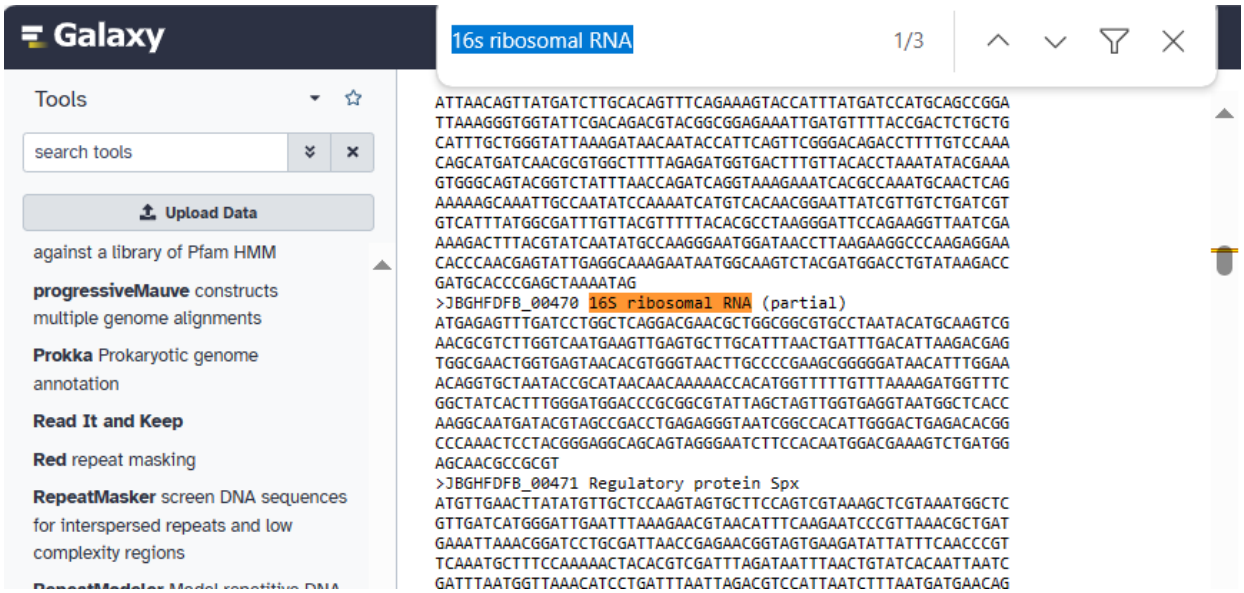


Рис. 42. Поиск последовательностей *16S ribosomal RNA*

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

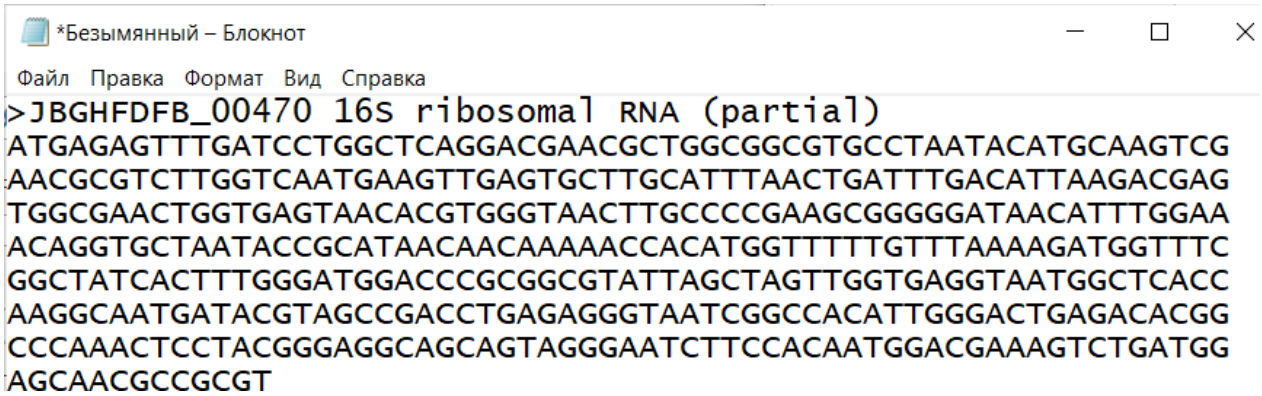


Рис. 43. Создание и сохранение текстового файла с последовательностью *16S рРНК*

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

5.2. Определение таксономии по последовательности *16S рРНК* 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].

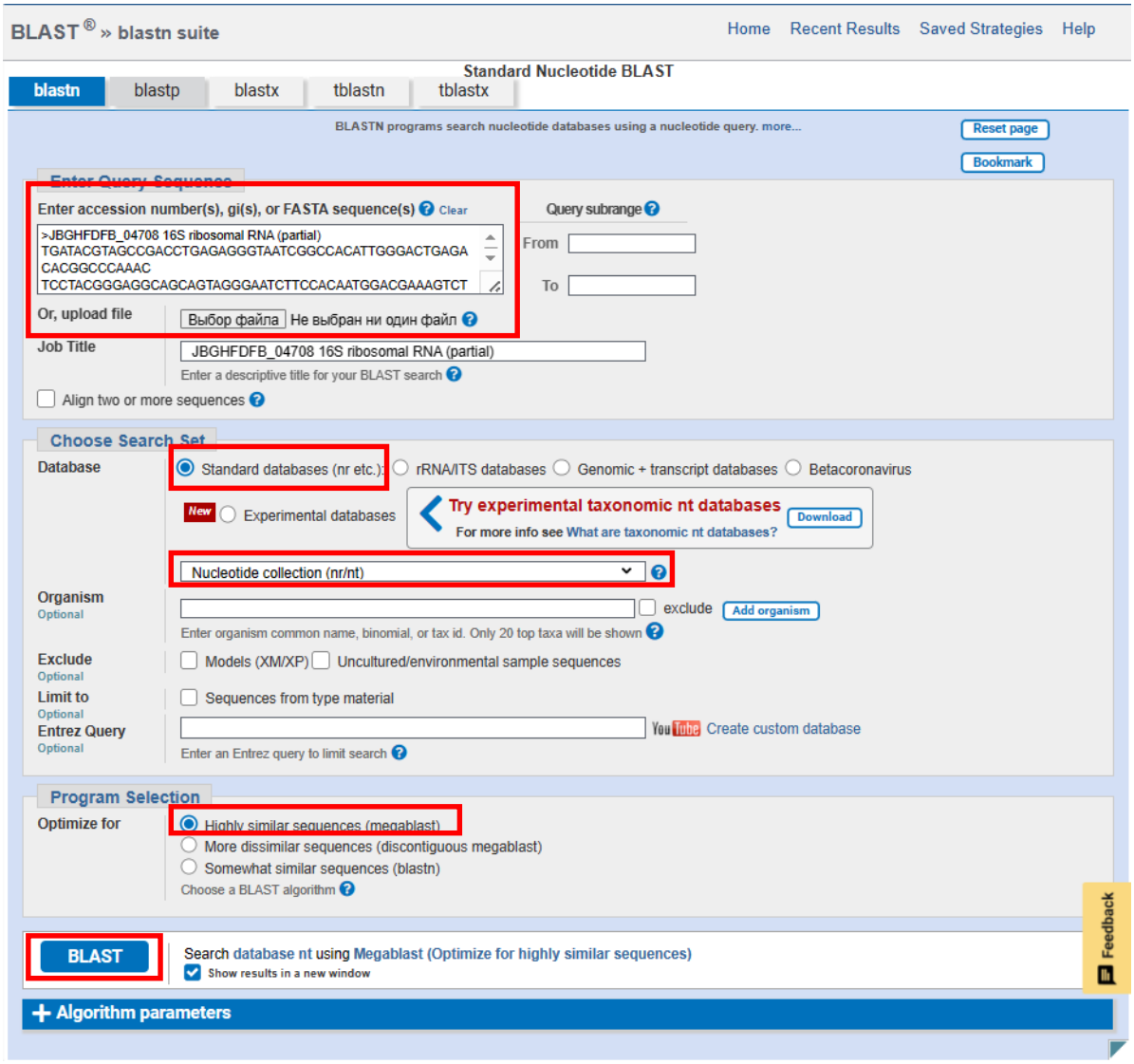


Рис. 44. Настройка выполнения blastn

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).

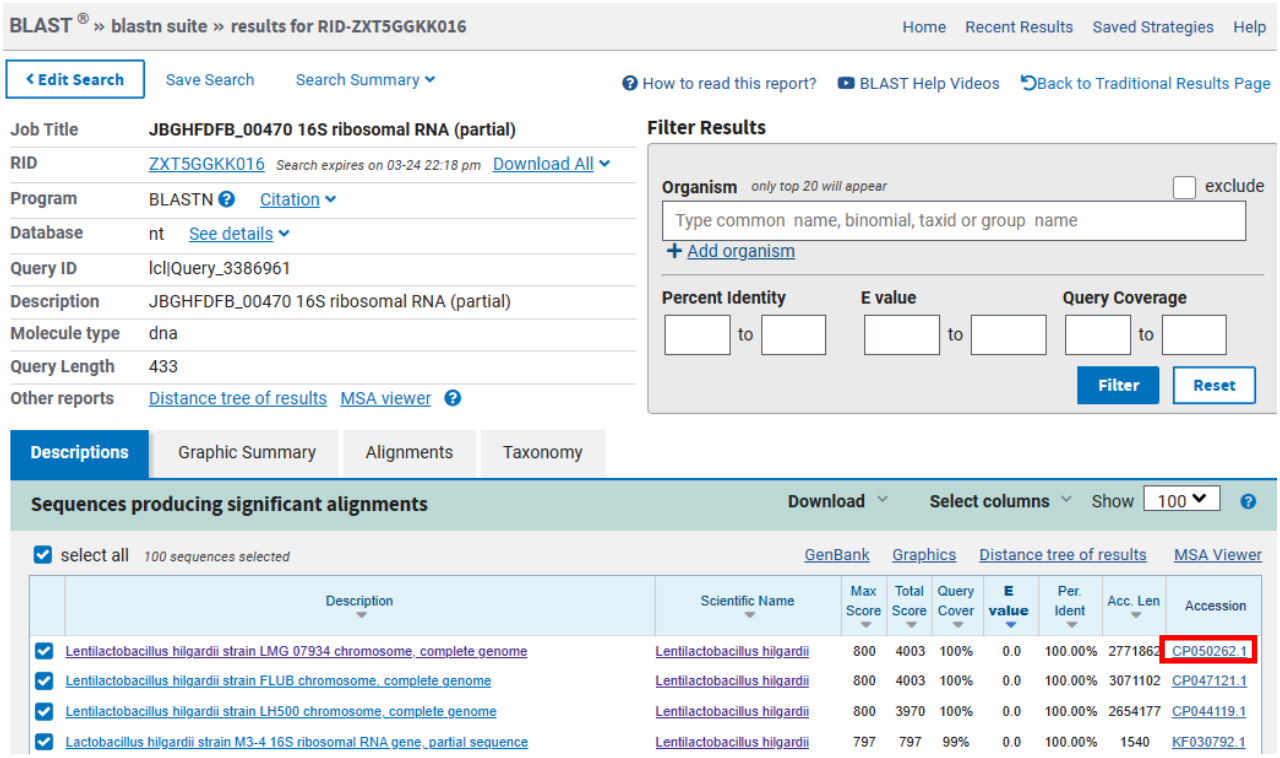
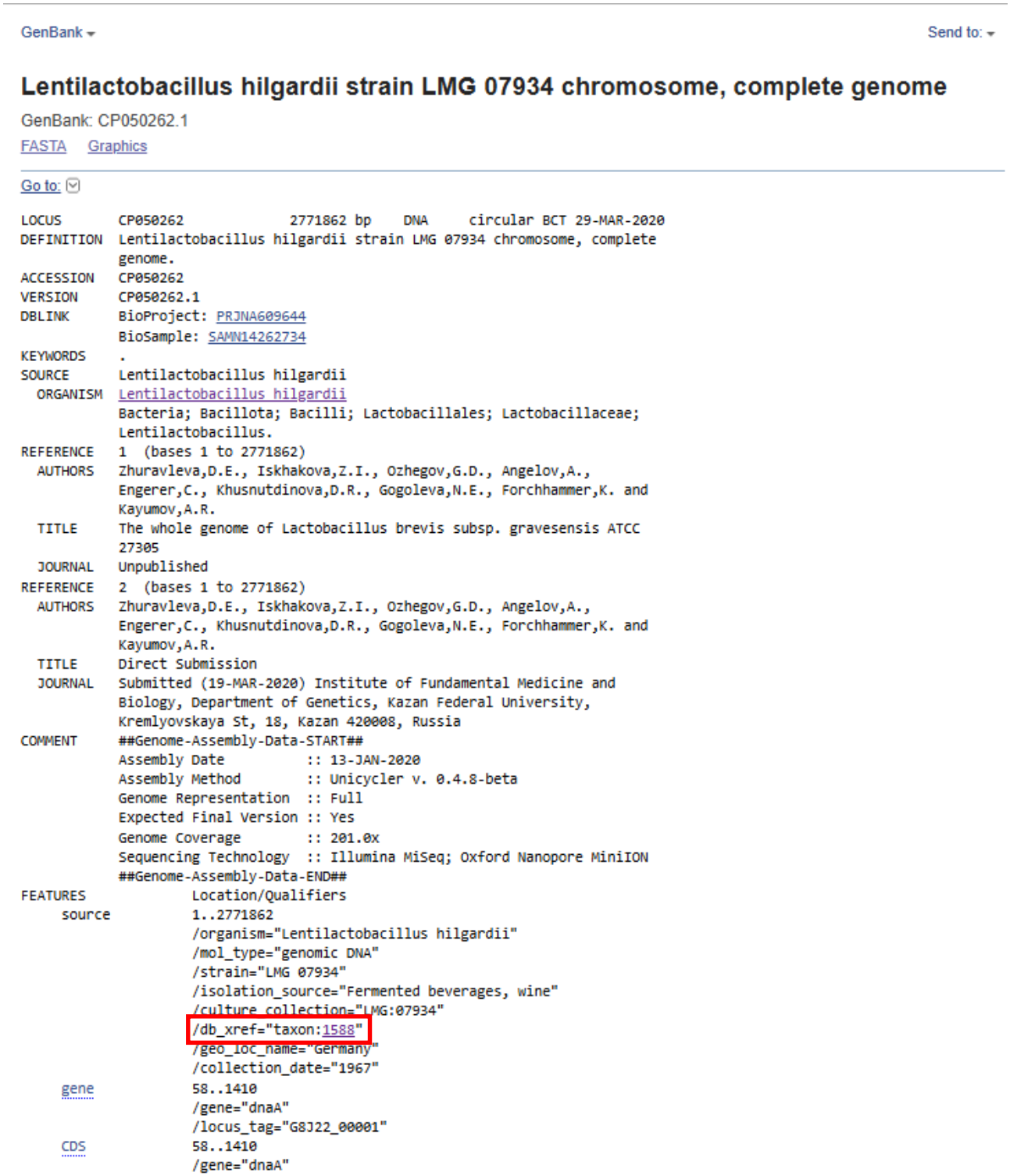


Рис. 45. Результаты поиска ***blastn*** гомологичных последовательностей к *JBGHFDFB_00470 16S ribosomal RNA*

Как видно из рис. 45, последовательность по гену *16S pPHK* при E-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 Taxonomy ID нажмите ссылку в поле Accession, чтобы перейти к аннотации известной последовательности в БД GenBank (рис. 45). Taxonomy ID приведено в поле **/db_xref**="taxon:1588" (рис. 46).



GenBank ▾ Send to: ▾

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 рРНК Lentilactibacillus 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.

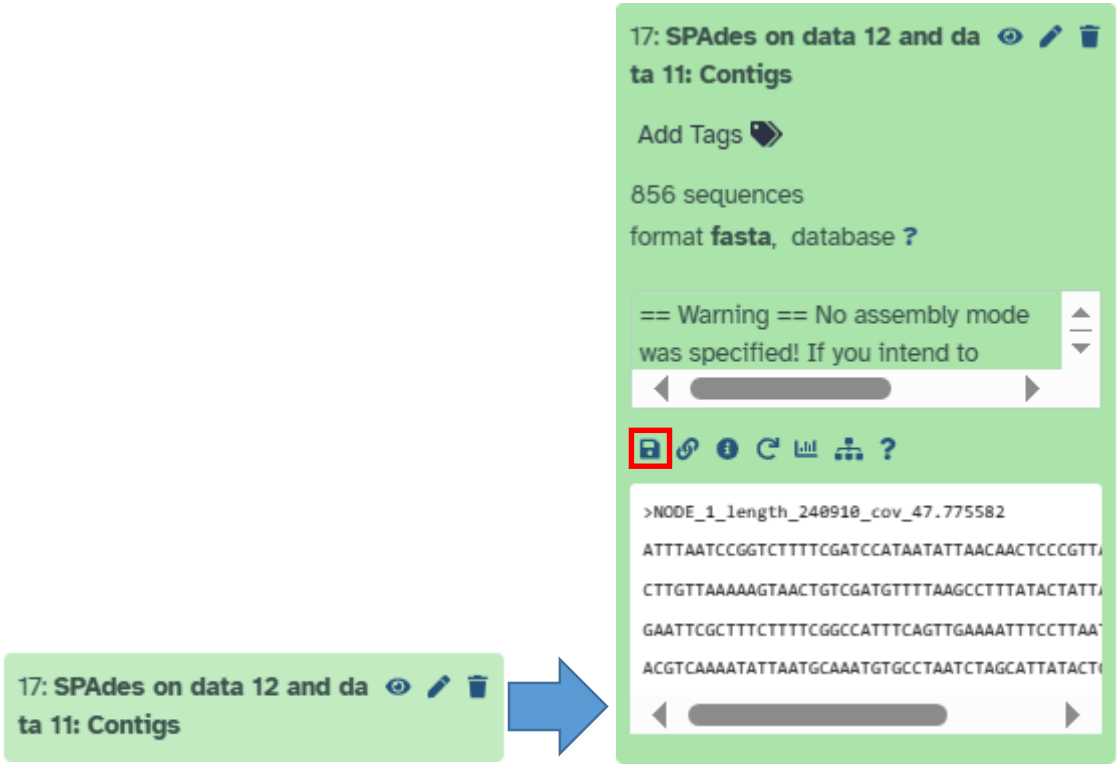


Рис. 47. Скачивание файла **SPAdes on data… and data…: Contigs** из Galaxy

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

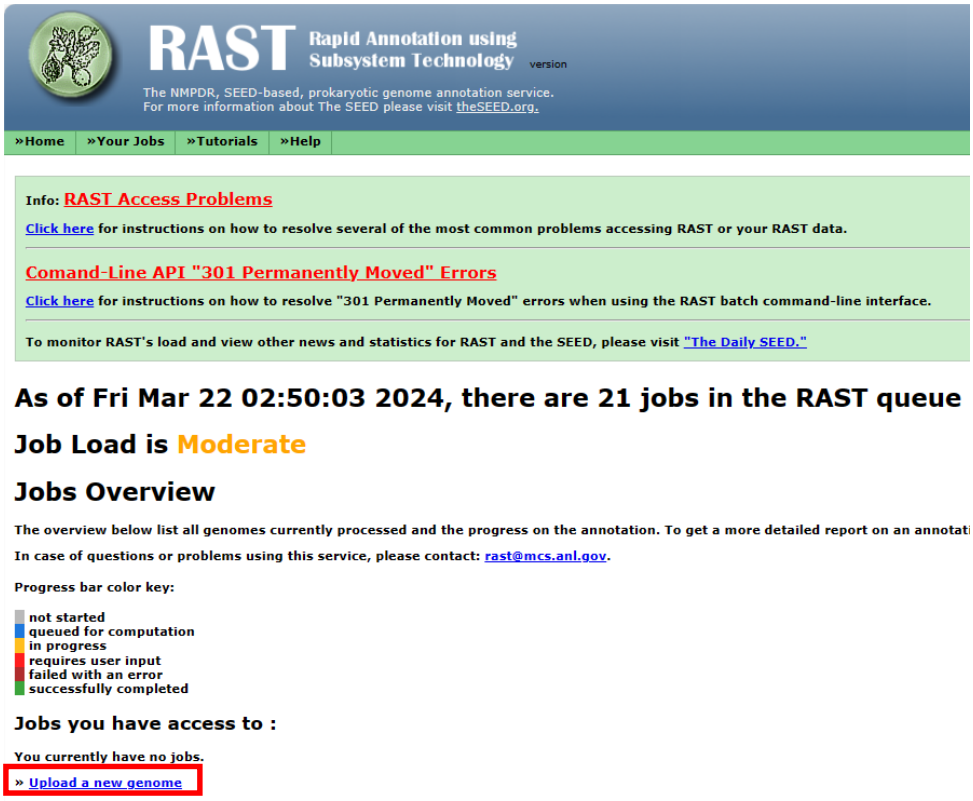


Рис. 48. Главная страница RAST

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).

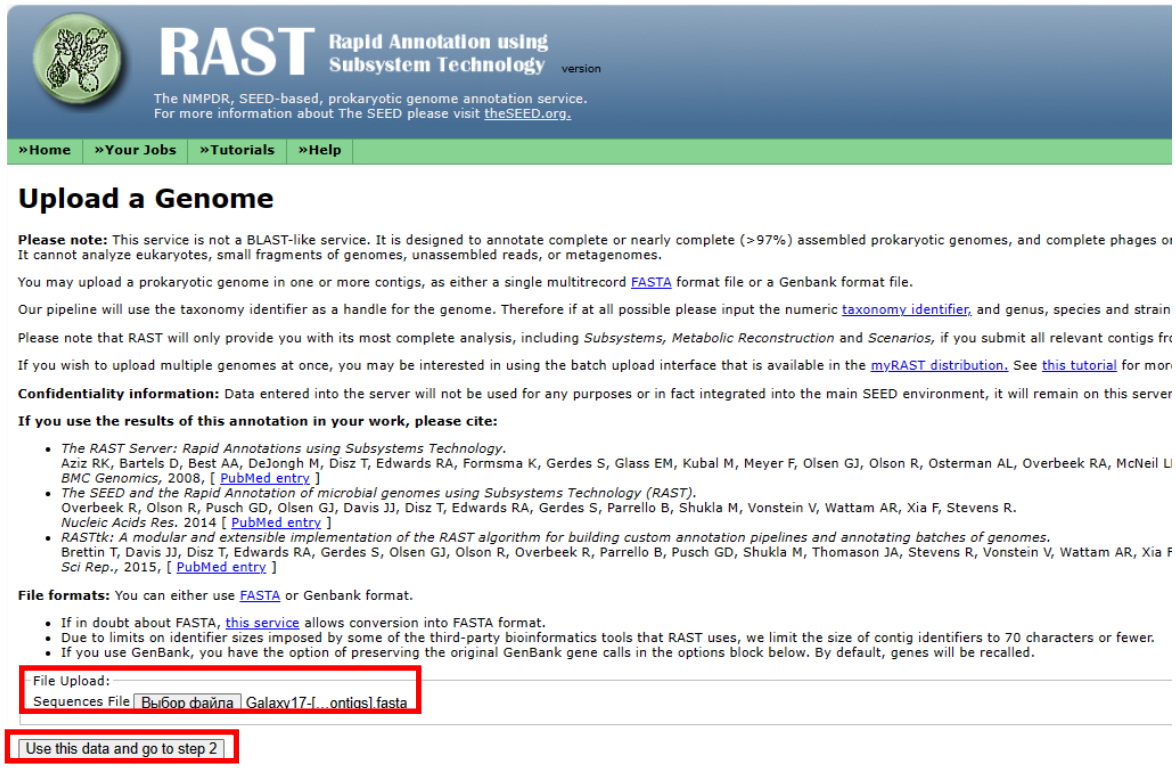


Рис. 49. Работа с RAST, шаг 1 – загрузка генома

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).

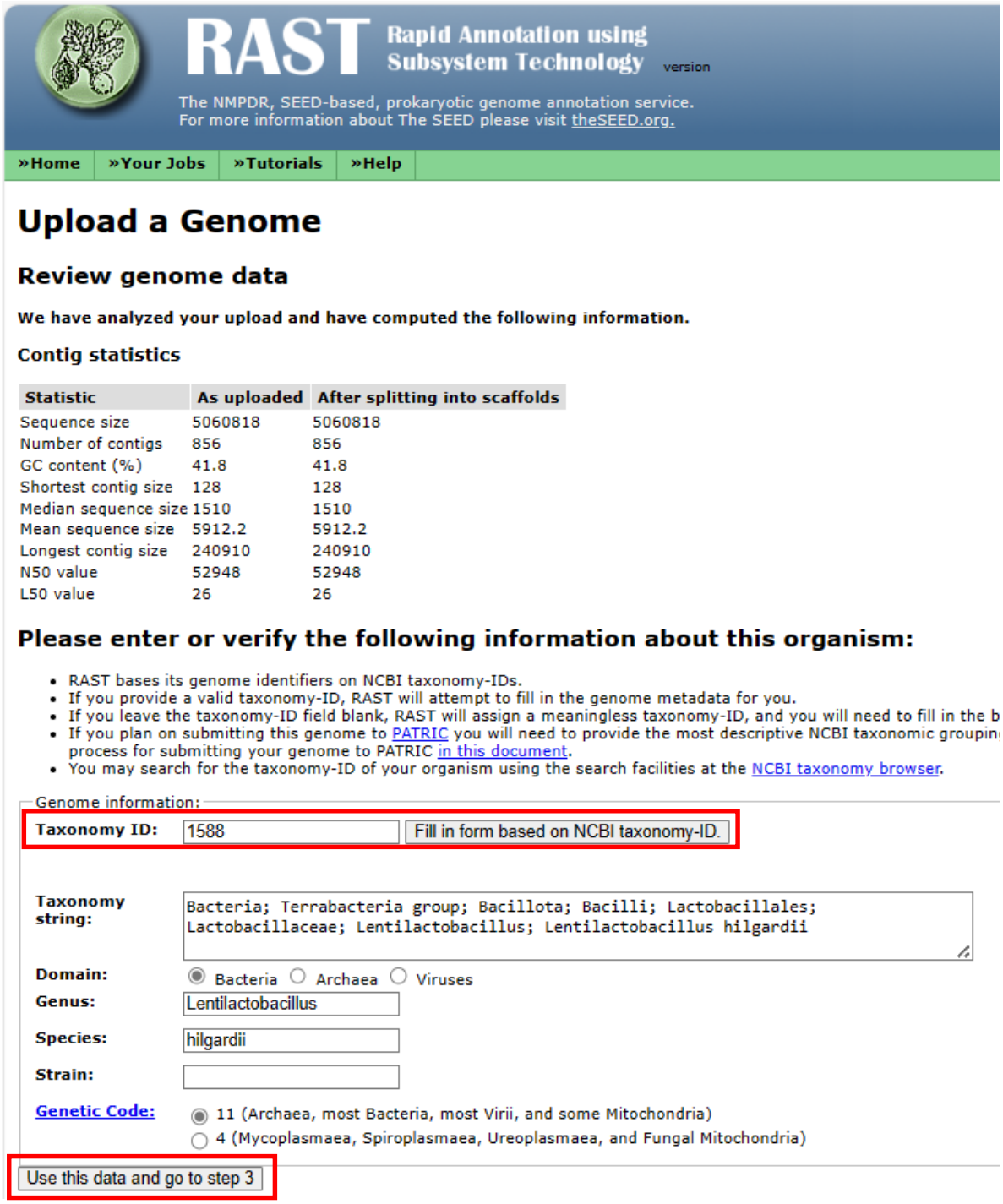


Рис. 50. Работа с RAST, шаг 2 – данные об организме

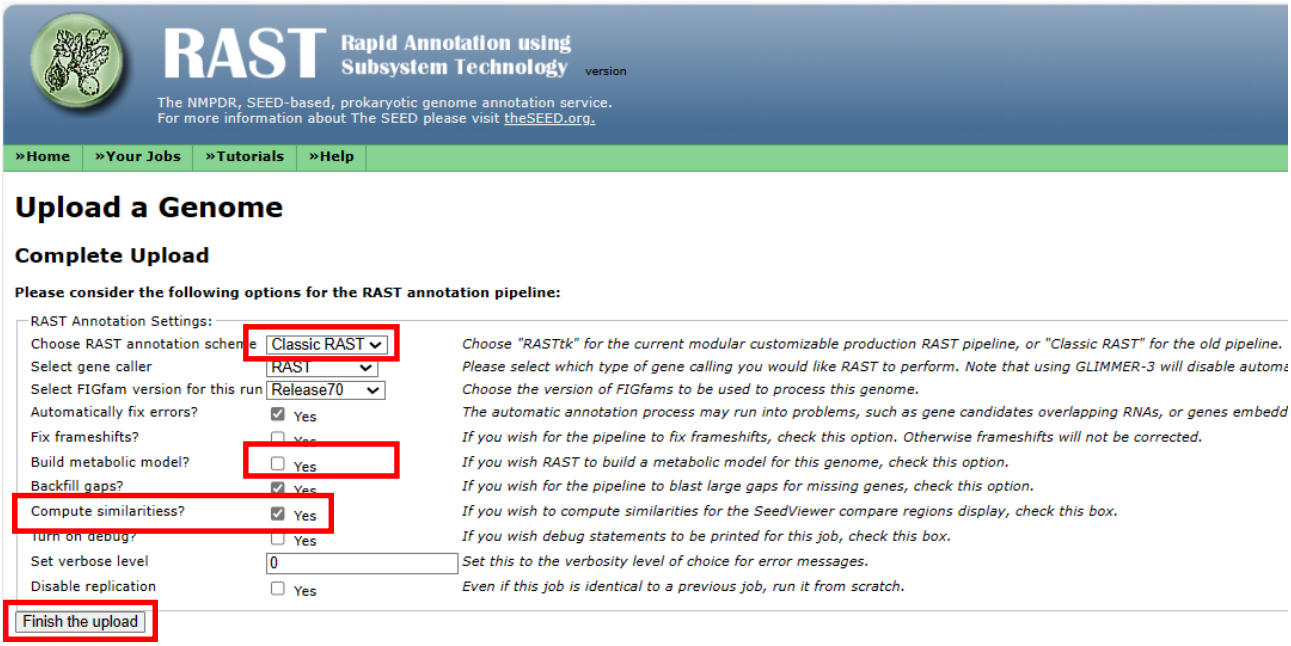


Рис. 51. Работа с RAST, шаг 3 – настройка выполнения

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*.

Рис. 52. Страница результатов RAST

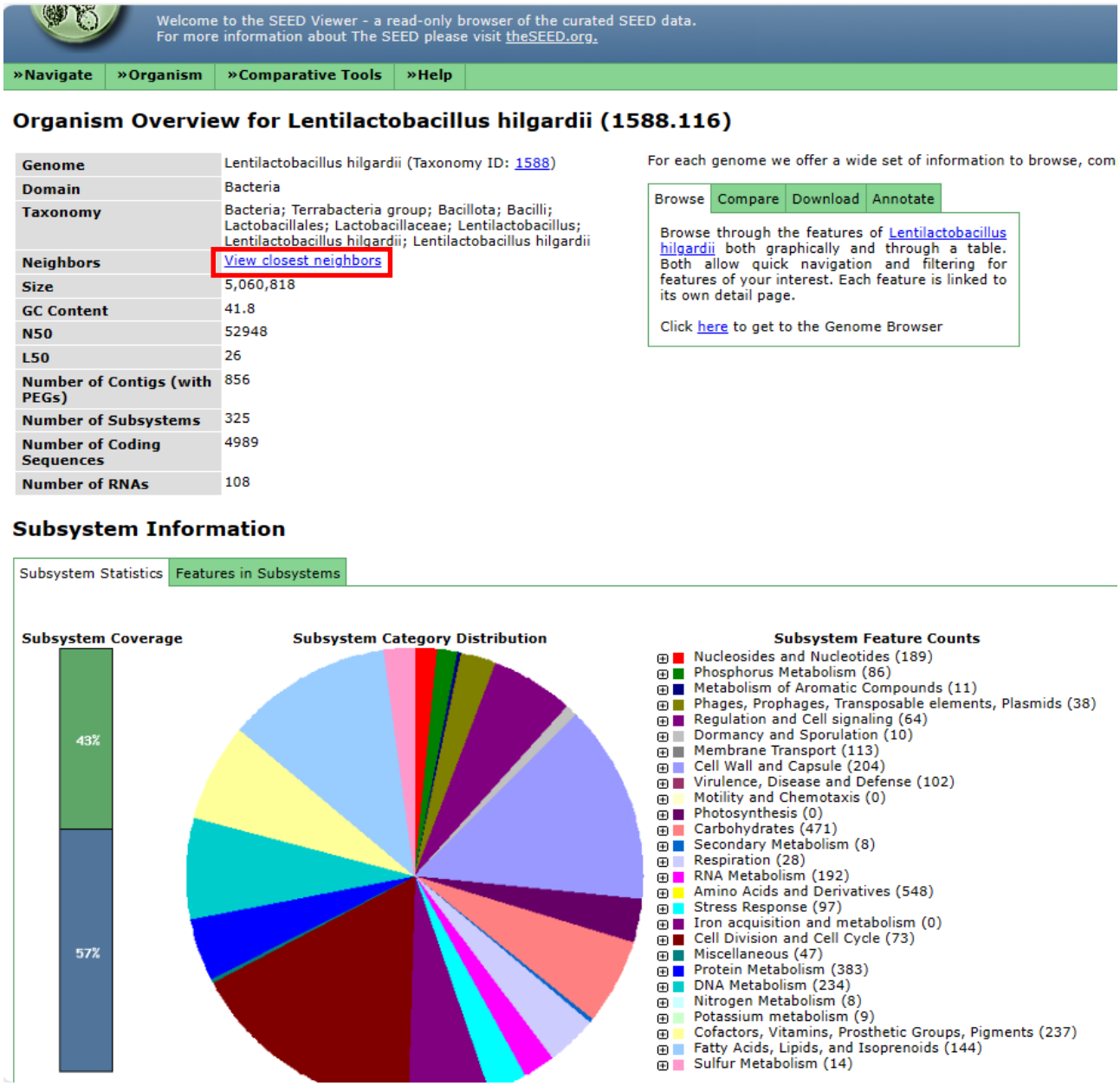


Рис. 53. Просмотр результатов аннотации

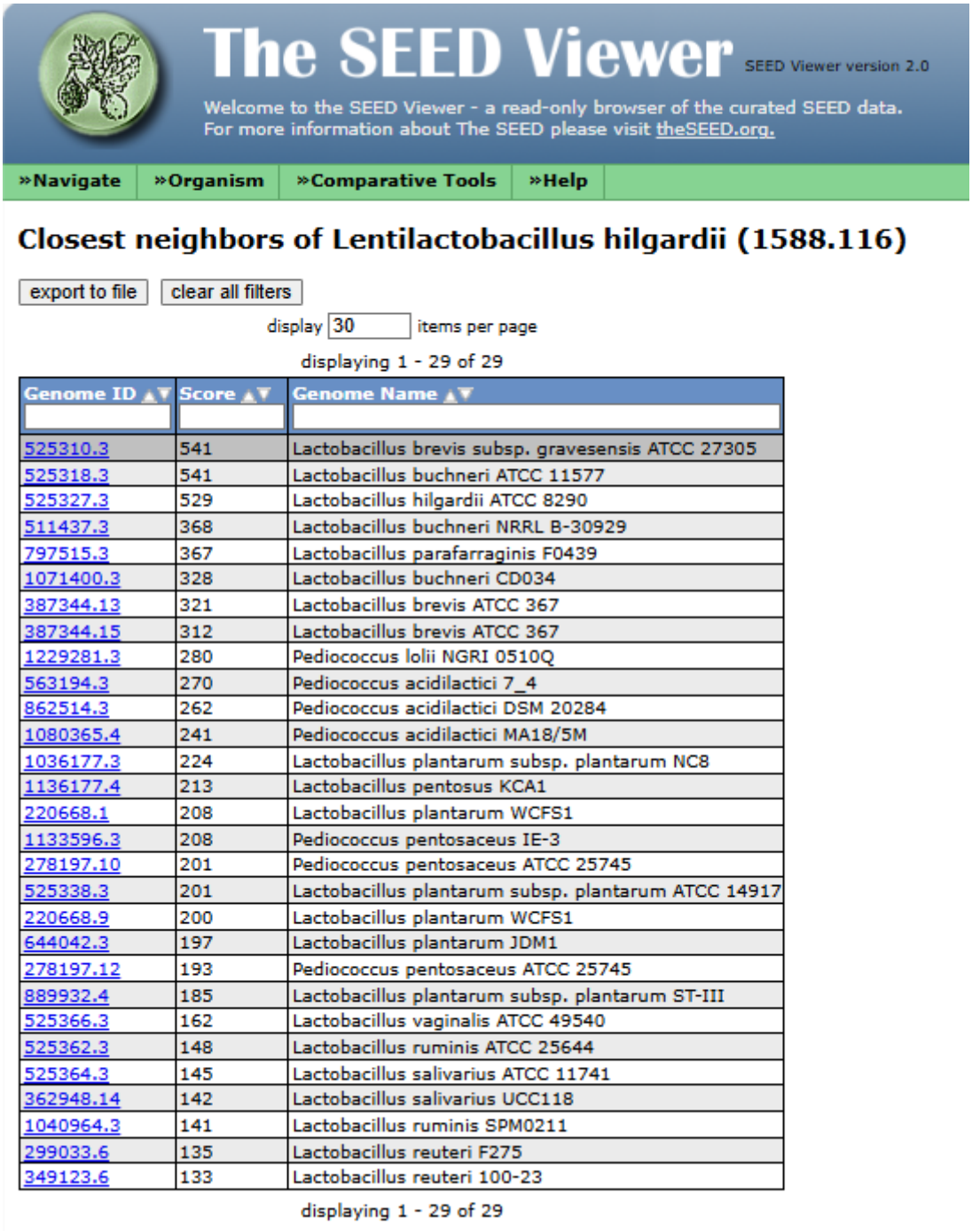


Рис. 54. Список близкородственных видов

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
- QUAST 5.2.0 manual – https://quast.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.О4 Спецпрактикум по прикладным методам в биологии»
06.03.01 Биология (бакалавр)