ОГЛАВЛЕНИЕ - ГИПЕРССЫЛКИ НА ПРОЦЕДУРУ

1. **ЦЕЛЬ:** Описать процедуру анализа данных полногеномного секвенирования (WGS) Illumina с коротким чтением, которые будут использоваться для эпиднадзора PulseNet International (PNI) с помощью облачной платформы Terra.Bio.

2. **ОБЛАСТЬ ПРИМЕНЕНИЯ:** Данная процедура касается всех сотрудников PulseNet, использующим платформу Terra.Bio для анализа данных WGS с коротким считыванием Illumina в целях эпиднадзора в рамках сети PulseNet International. Данная СОП охватывает загрузку последовательностей и метаданных в Terra.Bio, оценку качества последовательностей, рабочие процессы сборки и генотипирования, а также загрузку последовательностей в NCBI. Филогенетические анализы рассматриваются в СОП PNID02 (Стандартная операционная процедура PulseNet International для филогенетического анализа данных WGS с использованием платформы Terra.Bio).

3. **ОПРЕДЕЛЕНИЯ/ТЕРМИНЫ:**

3.1 ANI: средняя нуклеотидная идентичность

3.2 BaseSpace: Облачная вычислительная среда Illumina для анализа, управления и хранения данных секвенирования следующего поколения, включая совместное использование данных.

3.3 Команды Bash: Bash **(**Bourne Again Shell) - это оболочка интерфейса командной строки (CLI), широко используемая в Linux и macOS. Оболочка - это компьютерная программа, которая позволяет напрямую управлять операционной системой компьютера. Команды Bash используются для управления компьютером или операционной системой без необходимости перемещаться по меню, опциям и окнам графического интерфейса пользователя.

3.4 BioProject: Коллекция биологических данных в NCBI, относящаяся к одной инициативе, исходящая от одной организации или консорциума.

3.5 Биообразец (BioSample): Предоставляемая отправителем описательная информация (метаданные) о биологических материалах, из которых получены данные, хранящиеся в NCBI.

3.6 Контиг: непрерывная консенсусная последовательность, полученная в результате сборки множества коротких, перекрывающихся фрагментов ДНК.

3.7 Покрытие: Среднее количество чтений, включающих данный нуклеотид в реконструированную последовательность.

3.8 Критические показатели качества: покрытие (после обрезки), среднее качество (Q score до обрезки), длина сборки и обилие вторичных генов (обнаружение контаминации с помощью MIDAS). Последовательности, не соответствующие минимальным пороговым значениям/допустимым диапазонам для этих показателей, определенных в данном документе, должны быть повторно секвенированы.

3.9 CSV: значения, разделенные запятыми.

3.10 Сборка DeNovo: Сборка последовательности, созданная из коротких необработанных чтений без использования эталонного генома.

3.11 FASTA: текстовый формат для представления нуклеотидных или пептидных последовательностей, в котором пары оснований или аминокислот представлены с помощью однобуквенных кодов. Последовательность в формате FASTA начинается с однострочного описания, за которым следуют строки данных о последовательности. Строка описания отличается от данных о последовательности символом "больше, чем" (">") в первом столбце.

3.12 FASTQ: текстовый формат для хранения биологической последовательности и соответствующих ей оценок качества.

3.13 GAMBIT: метод геномной аппроксимации для идентификации и отслеживания бактерий. Метод идентификации бактериальных видов, использующий алгоритм на основе k-мер для поиска по большой эталонной базе данных геномов.

3.14 Gzip: Формат файлов и программное приложение, используемое для сжатия и распаковки файлов с целью быстрой передачи данных через Интернет.

3.15 LIMS: система управления лабораторной информацией.

3.16 Mash Sketching: Mash - это набор инструментов для создания и использования эскизов MinHash, способ превращения генома в небольшую подпись, которую можно легко сравнить с другими подписями.

3.17 Метаданные: Набор данных, которые описывают и предоставляют информацию о других данных.

3.18 MIDAS: Metagenomic Intra-species Diversity Analysis System. Интегрированный вычислительный конвейер для количественной оценки численности и охвата бактериальных видов в дробовых метагеномах на основе взрывного выравнивания по панели универсальных однокопийных генов.

3.19 N50: статистика N50 обычно используется в качестве грубой оценки геномных сборок. Она представляет собой длину контига (в парах оснований), для которого половина последовательности генома собрана в контиги, превышающие или равные размеру контига N50.

3.20 NCBI: Национальный центр биотехнологической информации.

3.21 PNI: PulseNet International.

3.22 QA/QC: Обеспечение качества/контроль качества (ОК/КК).

3.23 Q score: Оценка качества для каждой отдельной позиции основания в последовательности, указывающая на точность распознавания основания. Используются оценки Phred, где Q = -10log (Вероятность ошибки). Чем выше балл качества, тем надежнее распознавание нуклеотидных оснований. Q30 означает, что вероятность неправильного распознавания в данной позиции составляет 1 к 1000.

3.24 Прочтение: Единица непрерывной последовательности ДНК (пары оснований), полученная путем секвенирования части фрагментированной целевой ДНК.

3.25 Номер SAMN: Уникальный идентификатор NCBI (номер доступа) для биообразца последовательности (метаданные).

3.26 SOP: Стандартная операционная процедура.

3.27 SRA: Архив считывания последовательностей.

3.28 Номер SRR: Уникальный идентификатор NCBI (номер доступа) для загруженных необработанных чтений последовательностей.

3.29 Terra.Bio: Облачная платформа для анализа последовательностей, разработанная Институтом Броуда Массачусетского технологического института и Гарвардским университетом и используемая компанией Theiagen Genomics (Хайлендс-Ранч, CO, США) для предоставления общей платной платформы лабораториям общественного здравоохранения для размещения, анализа и обмена данными (Libuit *et al.,* 2023).

3.30 TSV: Значения, разделенные табуляцией.

3.31 WGS: полногеномное секвенирование.

4. ОБЯЗАННОСТИ:

4.1 Сотрудники PulseNet проводят оценку качества и генотипирование последовательностей коротких чтений Illumina, созданных для международного эпиднадзора PulseNet, используя платформу Terra.Bio. Настоятельно рекомендуется делиться данными WGS с другими участниками PNI путем загрузки в NCBI.

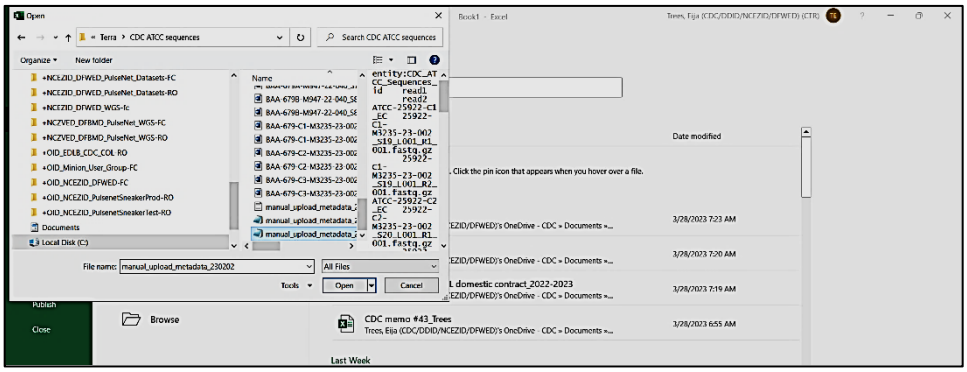
5. ПРОЦЕДУРЫ:

ПРИМЕЧАНИЕ: *Данные последовательностей могут быть импортированы в Terra одним из следующих трех способов:* ***(1)*** *Загрузка из локального сетевого хранилища*

(шаги 5.1 - 5.3), ***(2)*** *Прямая передача из облака в облако из Illumina BaseSpace (приложение PNID01-1),* ***(3)*** *Загрузка из NCBI SRA (приложение PNID01-2).*

5.1 **Подготовьте файл метаданных для последовательностей, которые будут загружены на сайт Terra** . Файл метаданных будет связывать загруженные FASTQ-файлы с соответствующими именами записей в базе данных, т.е. ключами записей.

5.1.1 С помощью Excel откройте шаблон файла метаданных в формате tsv.



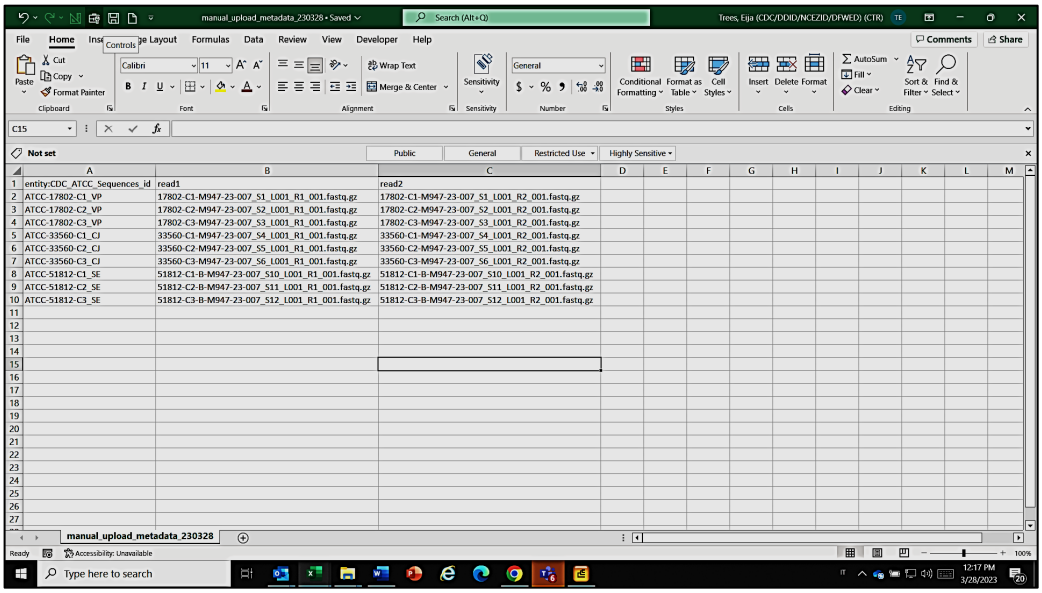
5.1.2 Необходимо заполнить как минимум следующие три столбца:

5.1.2.1 **Entity:collection_name_id**. Это название коллекции данных Terra (таблицы данных), в которую вы хотите загрузить свои последовательности, например, entity:CDC_ATCC_Sequences_id. Идентификаторы штаммов (ключи записи) называются в Terra "сущностями".

ПРИМЕЧАНИЕ: *пробелы и тире не допускаются. "entity:" и "_id" обязательны для Terra в имени столбца.*

5.1.2.2 **Read1.** Имя файла read1 fastq.gz для данной записи.

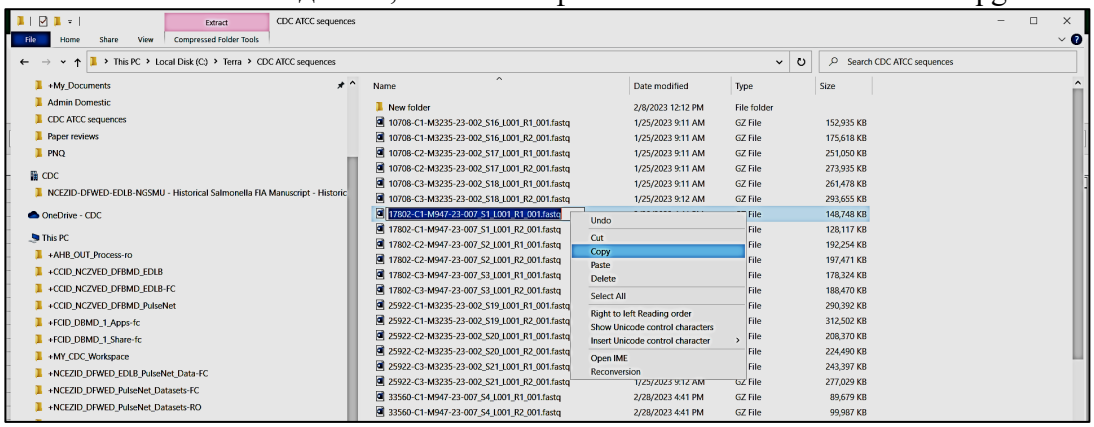
5.1.2.3 **Read2.** Имя файла read 2 fastq.gz для записи.



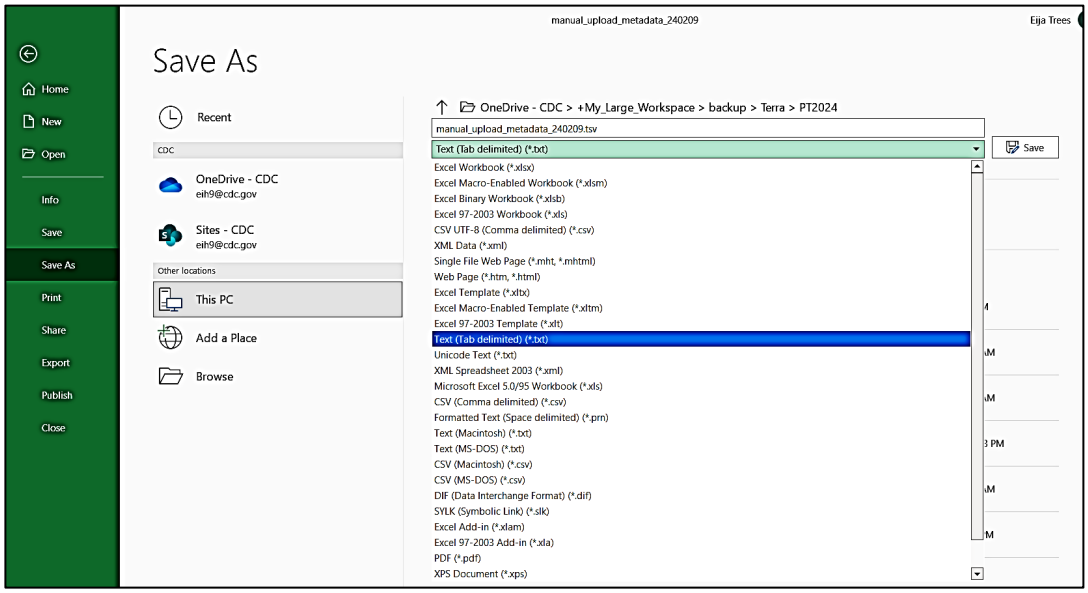
ПРИМЕЧАНИЕ: *дополнительные метаданные могут быть добавлены на данном этапе или позднее. Руководство по метаданным см. в приложении PNID01-6.*

5.1.3 Введите идентификаторы штаммов в столбце "identity" так, как вы хотите, чтобы они отображались в Terra.
ПРИМЕЧАНИЕ: *идентификатор штамма не должен совпадать с какой-либо частью имени файла fastq.gz.*

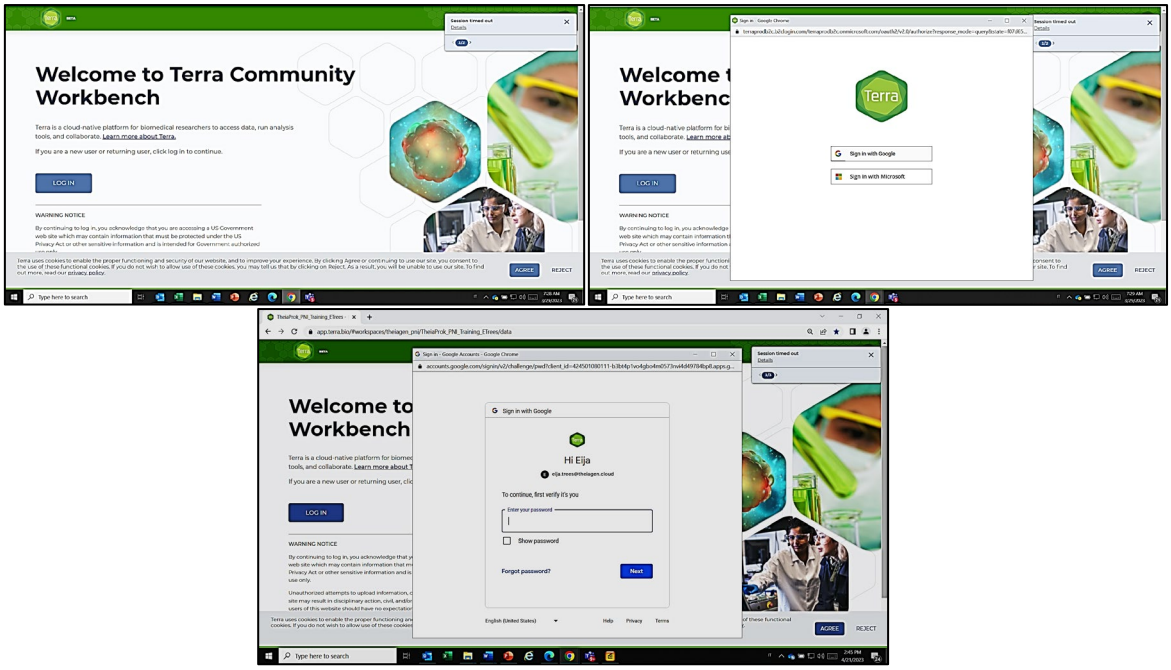
5.1.4 Скопируйте и вставьте имена файлов fastq.gz для каждого штамма в столбцы "read1" и "read2". Убедитесь, что имена файлов заканчиваются на "fastq.gz".



5.1.5 Сохраните файл в **формате tsv**: выберите "Сохранить как" и "Текст (с разделителем табуляции) (*.txt)". Убедитесь, что имя файла имеет окончание "**.tsv**".

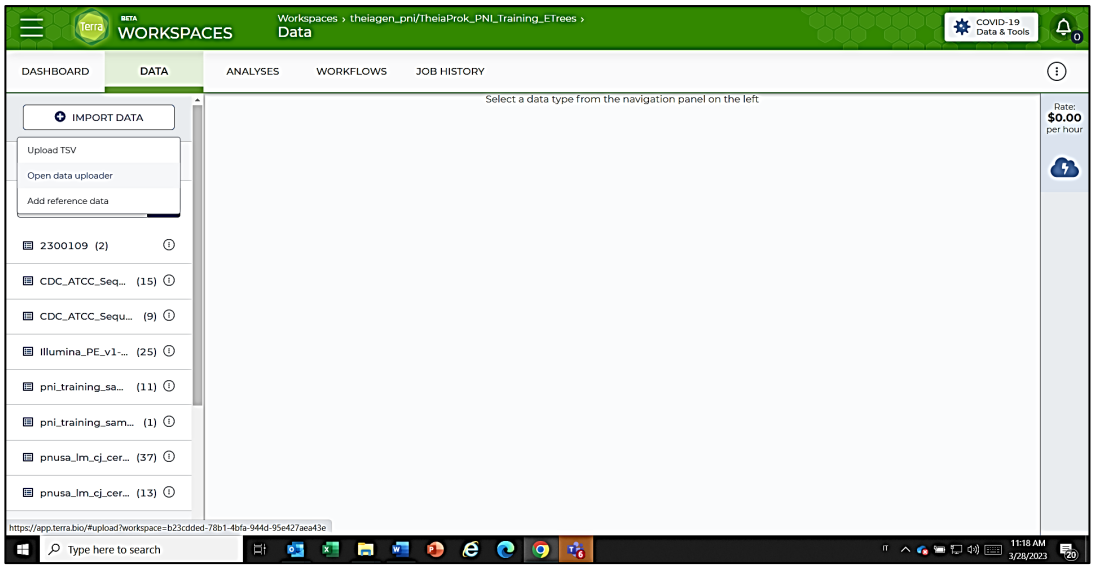


5.2 **Войдите в Terra, используя Chrome и свою учетную запись Google.**

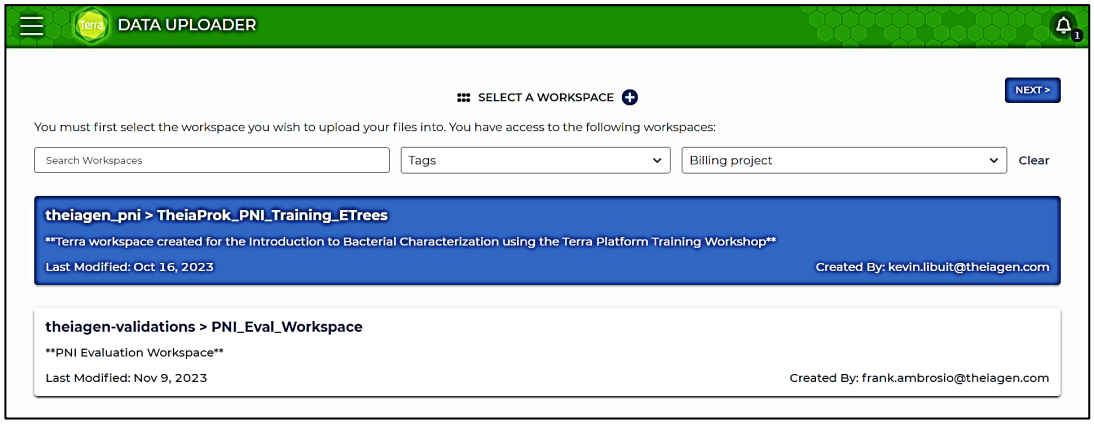


5.3 **Загрузите файлы последовательностей и метаданные в Terra**

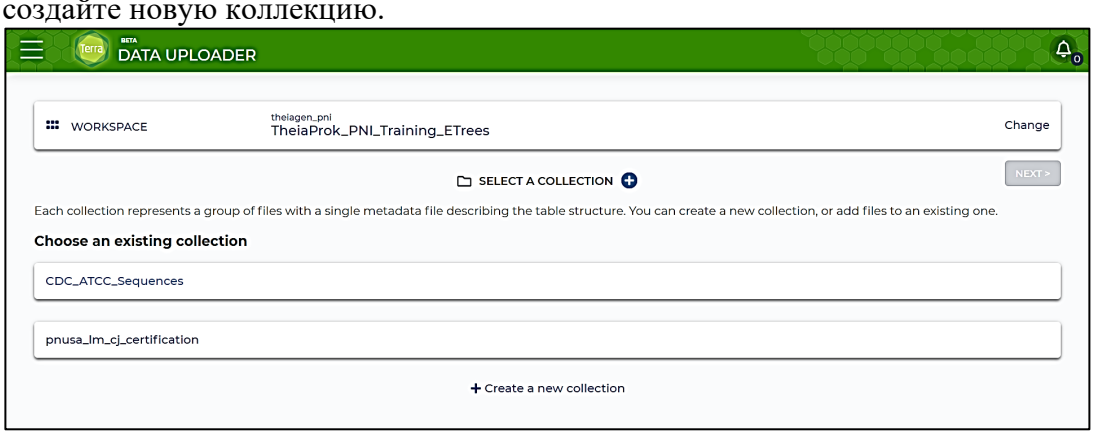
5.3.1 В разделе "Рабочее пространство Terra" выберите вкладку "Данные", нажмите "Импорт данных" и выберите "Открыть загрузчик данных" из выпадающего меню.



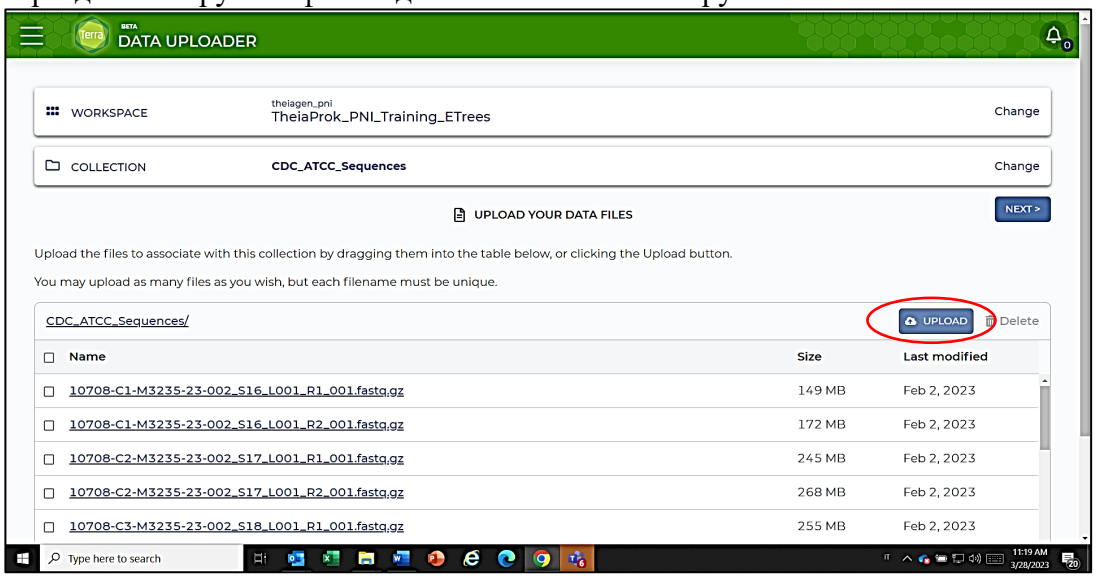
5.3.2 Откроется экран "Загрузчик данных". Если ваша учетная запись имеет доступ к нескольким рабочим пространствам, сначала вам нужно выбрать рабочее пространство, в которое вы хотите загрузить данные.



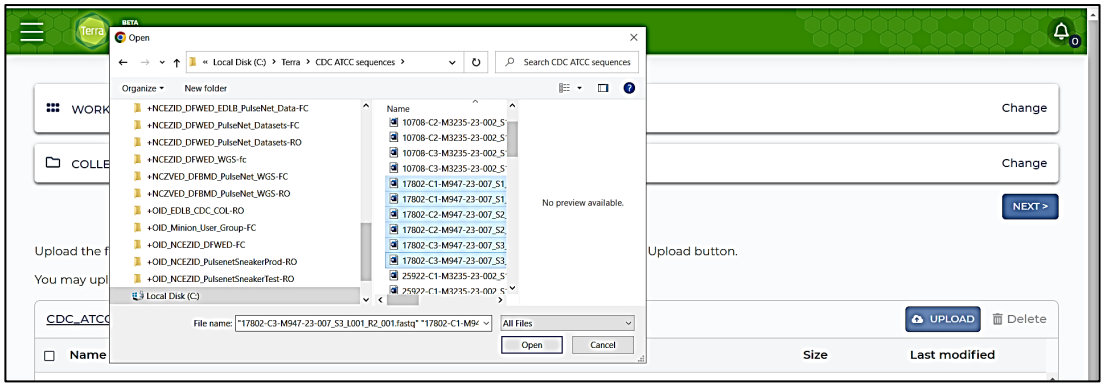
5.3.3 Выберите существующую коллекцию, щелкнув по ее названию в списке, либо создайте новую коллекцию.



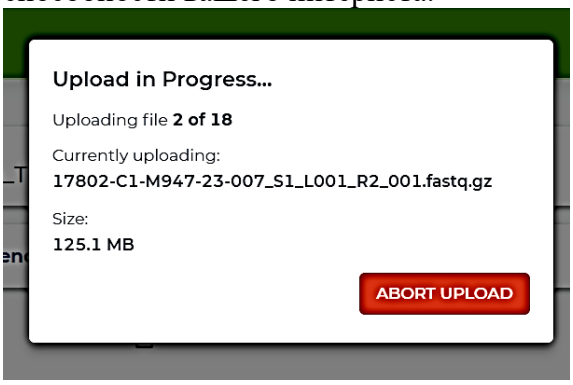
5.3.4 В разделе "Загрузить файлы данных" нажмите "Загрузить".



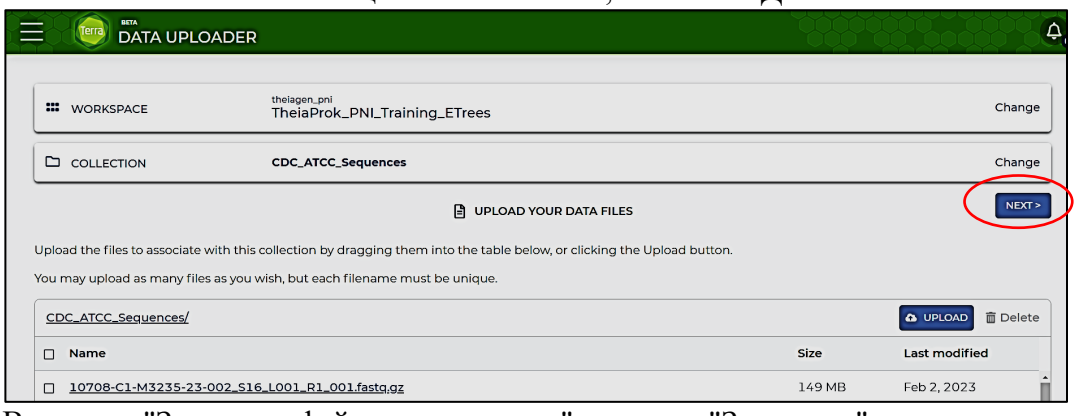
5.3.5 Перейдите в место, где сохранены файлы FASTQ, выберите файлы для загрузки и нажмите "Открыть".



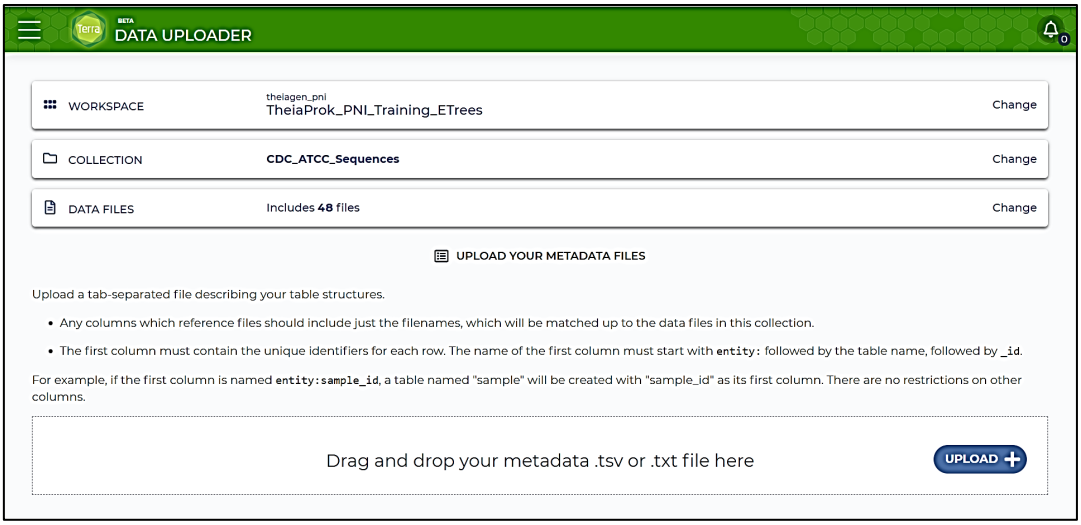
5.3.6 Появится всплывающее окно "Идет загрузка". Загрузка может занять несколько минут в зависимости от количества загружаемых файлов и пропускной способности вашего интернета.



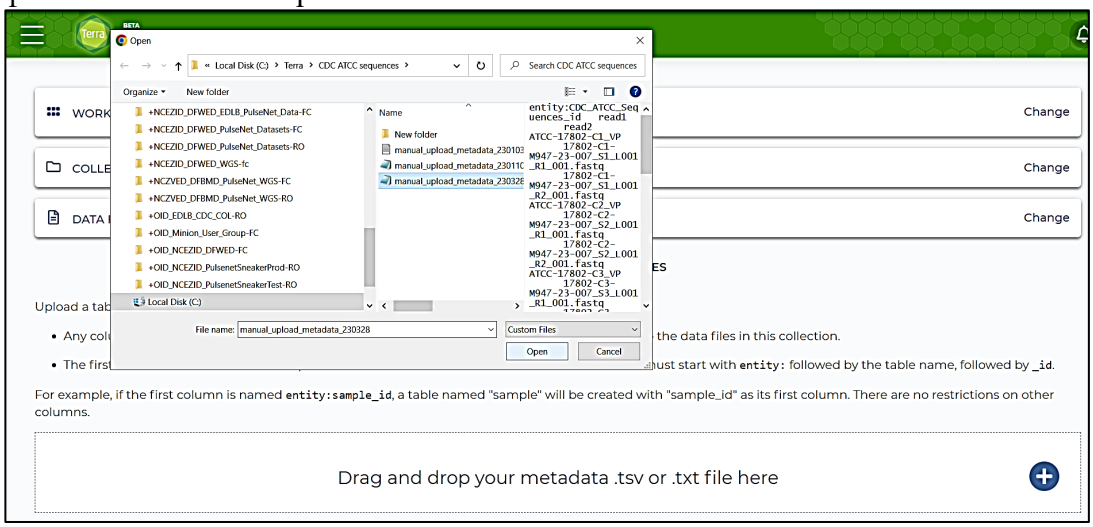
5.3.7 После того как всплывающее окно исчезнет, нажмите "Далее".



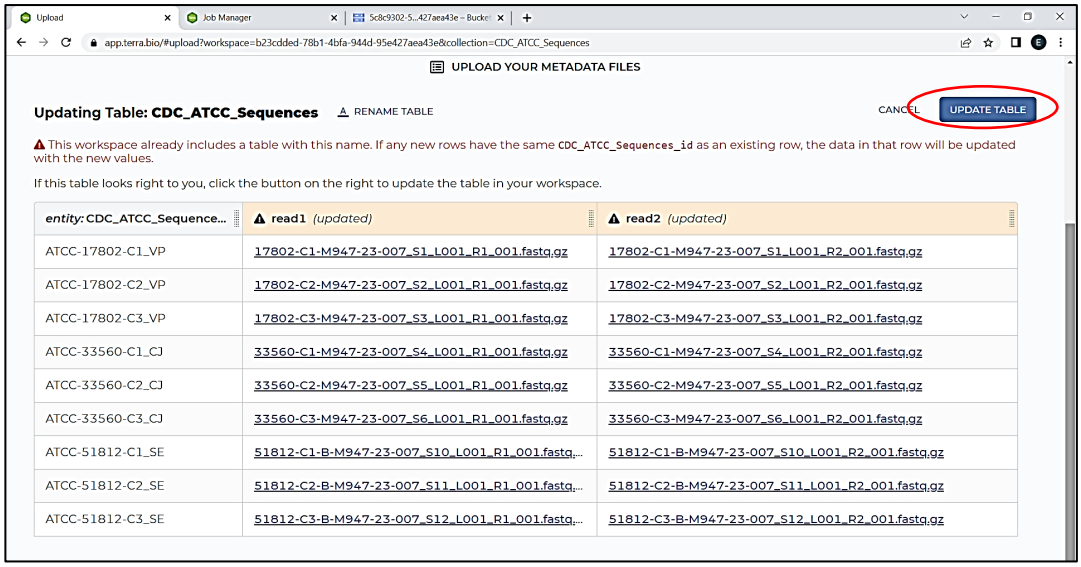
5.3.8 В разделе "Загрузить файлы метаданных" нажмите "Загрузить".



5.3.9 Перейдите в место, где сохранен файл метаданных tsv, выберите загружаемый файл и нажмите "Открыть".

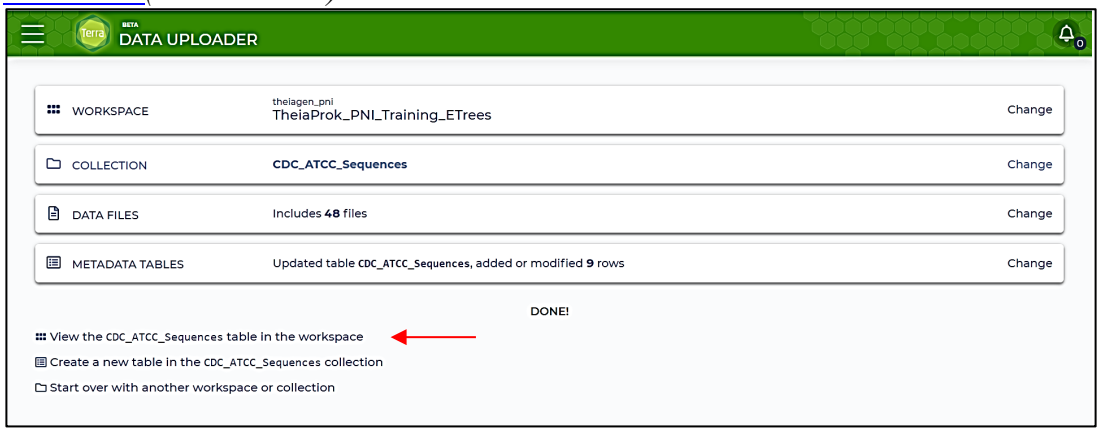


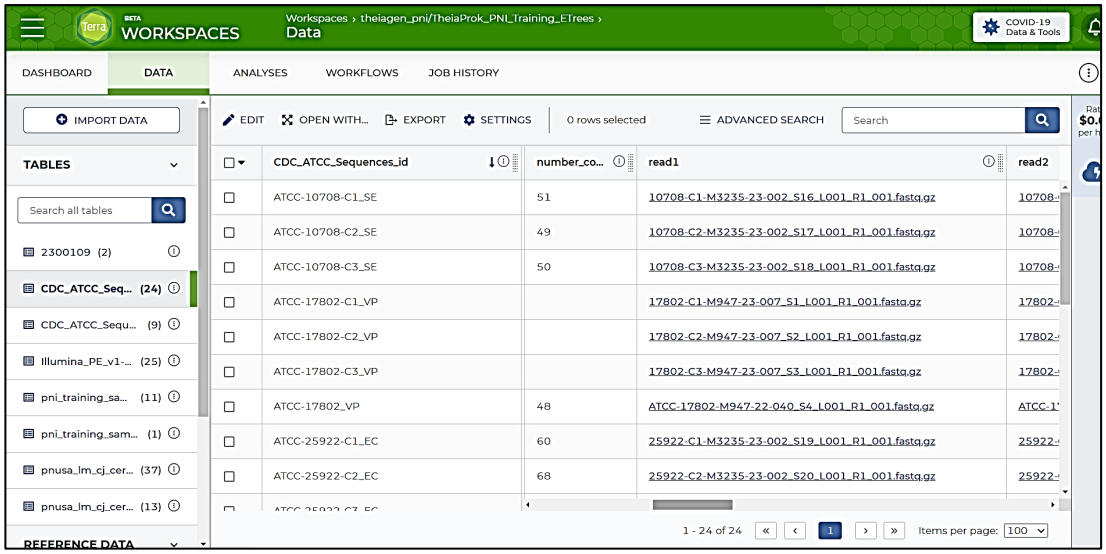
5.3.10 На следующем экране проверьте, что файлы fastq.gz связаны с правильными ключами записи, и нажмите "Обновить таблицу".



5.3.11 На экране "Загрузчика данных" появится сообщение "Выполнено". Вы можете просмотреть обновленную таблицу данных, нажав на ссылку, которая появится на экране.

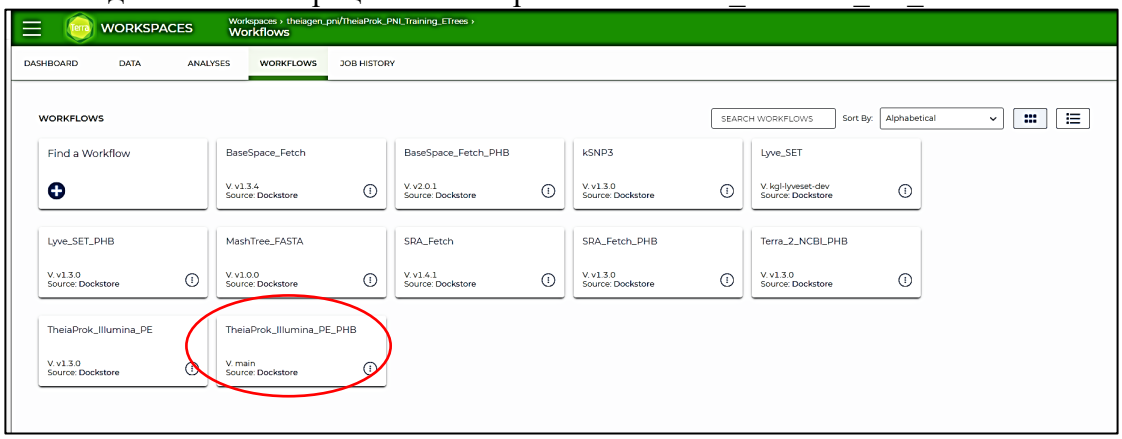
ПРИМЕЧАНИЕ: *столбцы, отображаемые в таблице данных, можно настраивать. Для удобства навигации рекомендуется создать отдельные представления для метрик QC, результатов генотипирования и метаданных. Руководство по настройке столбцов таблицы данных для эпиднадзора PulseNet см. в приложениях PNID01-3 (метрики КК), PNID01-5 (генотипирование) и PNID01-6 (метаданные).*





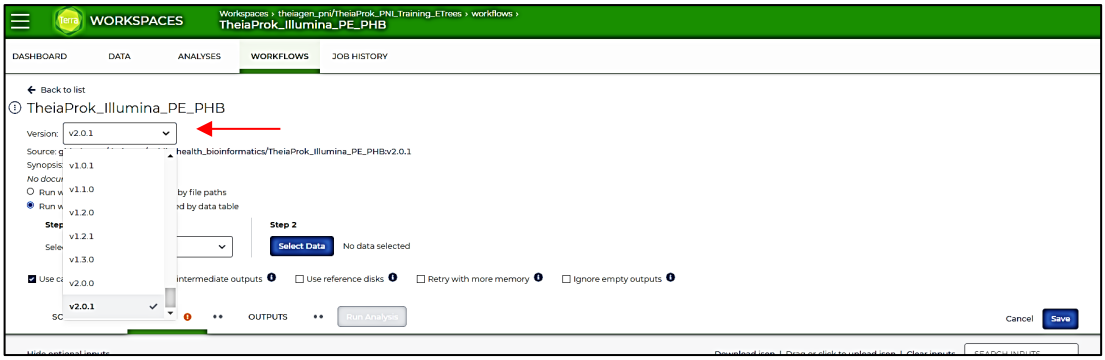
5.4 **Запустите рабочий процесс QC и генотипирования** . Рабочий процесс TheiaProk выполняет контроль качества необработанных чтений последовательностей, сборку необработанных чтений de novo с помощью Skesa, а затем контроль качества сборки и идентификацию вида. Также доступны различные анализы генотипирования, соответствующие виду.

5.4.1 На вкладке "Рабочие процессы" выберите "TheiaProk_Illumina_PE_PHB".

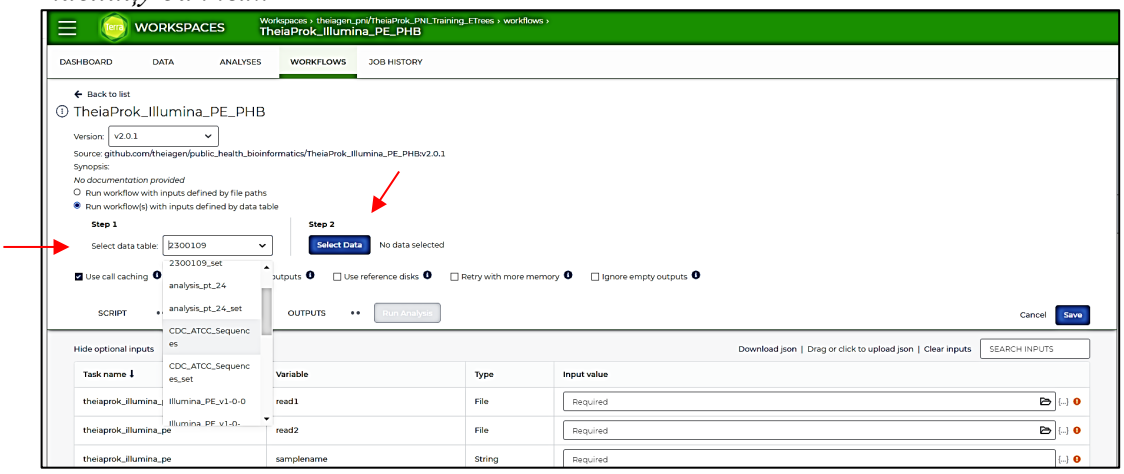


5.4.2 На экране "TheiaProk_Illumina_PE_PHB":

5.4.2.1 Выберите последнюю версию рабочего процесса TheiaProk_Illumina_PE_PHB из выпадающего меню "Версия".

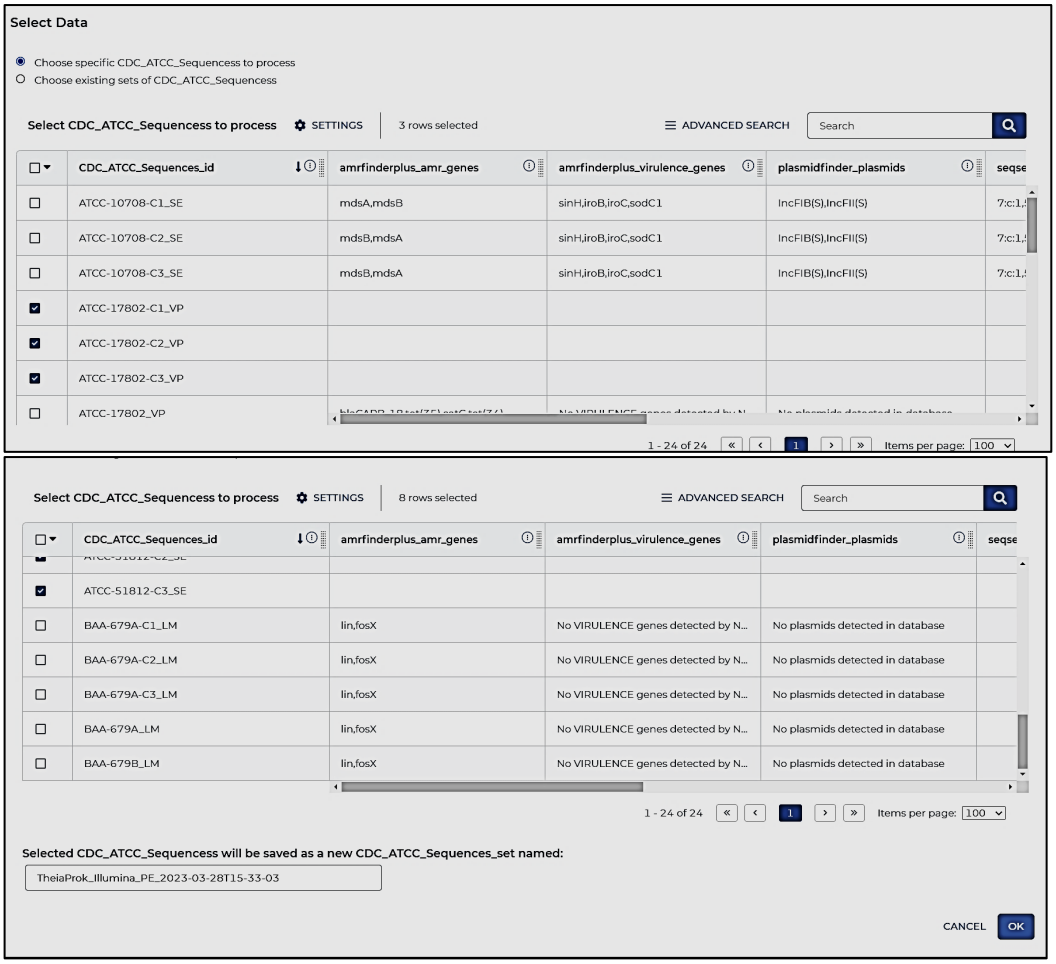


5.4.2.2 В разделе "Шаг 1" выберите "Тип корня", т. е. таблицу данных, в которой находятся образцы, например "CDC_ATCC_Sequences".
ПРИМЕЧАНИЕ: *Для большинства таблиц данных существует два варианта:* ***основная*** *таблица данных, содержащая отдельные записи образцов, и таблица данных* ***набора****, содержащая наборы образцов, используемые для филогенетических анализов, загрузки в NCBI и т.д. (например, CDC_ATCC_Sequences_set). Обязательно выберите* ***основную*** *таблицу данных.*

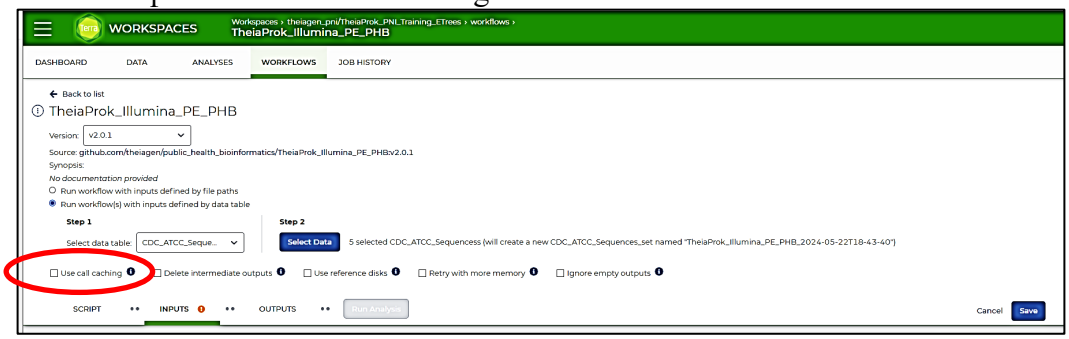


5.4.2.3 В разделе "Шаг 2" нажмите "Выбрать данные" (скриншот выше). В результате вы перейдете к таблице данных, указанной в шаге 1.

5.4.2.4 Выберите штаммы для анализа и прокрутите страницу до самого низа, чтобы нажать "OK".
ПРИМЕЧАНИЕ: *если таблица данных содержит более 100 записей и вы установили флажок "Выбрать все" напротив названия таблицы данных, будут выбраны только 100 записей.*



5.4.2.5 Снимите флажок "Use call catching".



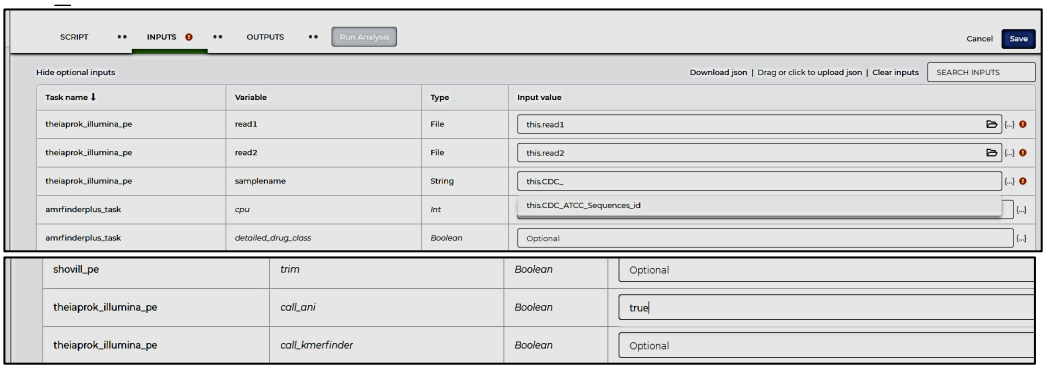
5.4.2.6 На вкладке "**Входы**" укажите следующие входные значения в столбце "Атрибут" (прокрутите список вниз):
ПРИМЕЧАНИЕ: *При заполнении столбца "Атрибут" щелчок внутри ячейки вызовет выпадающее меню атрибутов, которые вы можете выбрать, чтобы избежать опечаток (скриншот ниже).*

5.4.2.6.1 Read1: "This.read1".

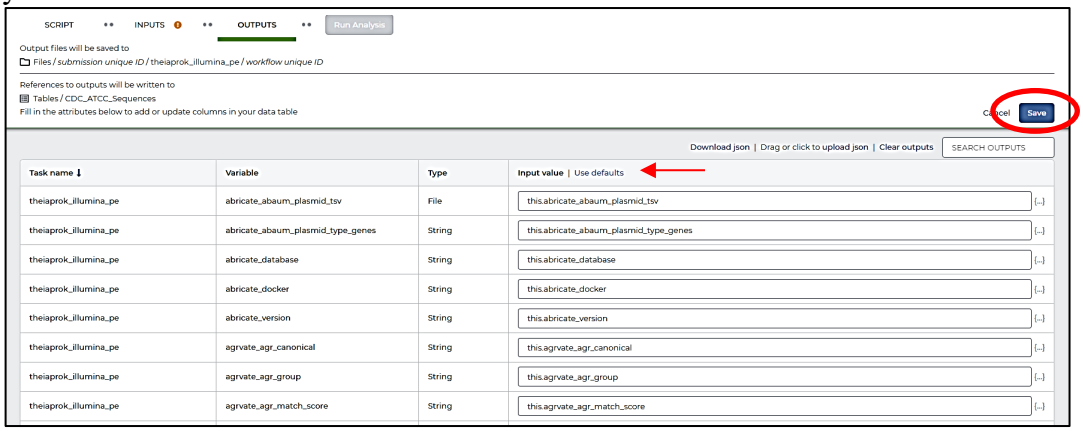
5.4.2.6.2 Read2: "This.read2".

5.4.2.6.3 Имя образца: "**This**.*data table* **name_id**", например, This.CDC_ATCC_Sequences_id.

5.4.2.6.4 Call_ani: "True".



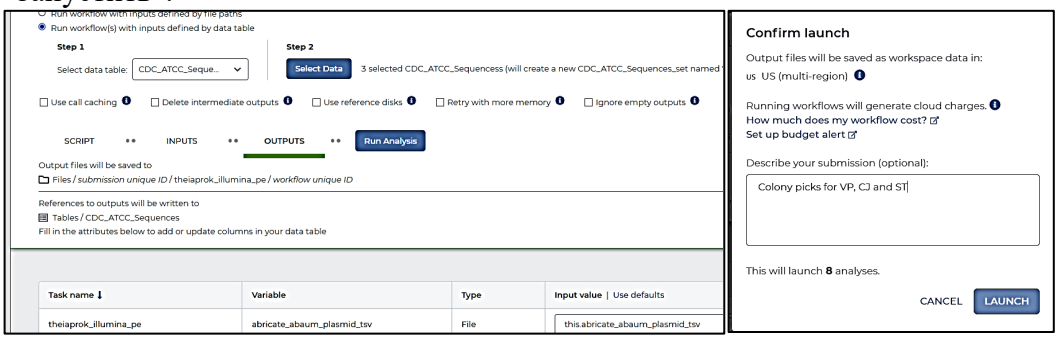
5.4.2.7 На вкладке "**Выход/Outputs**" нажмите "Использовать значения по умолчанию".



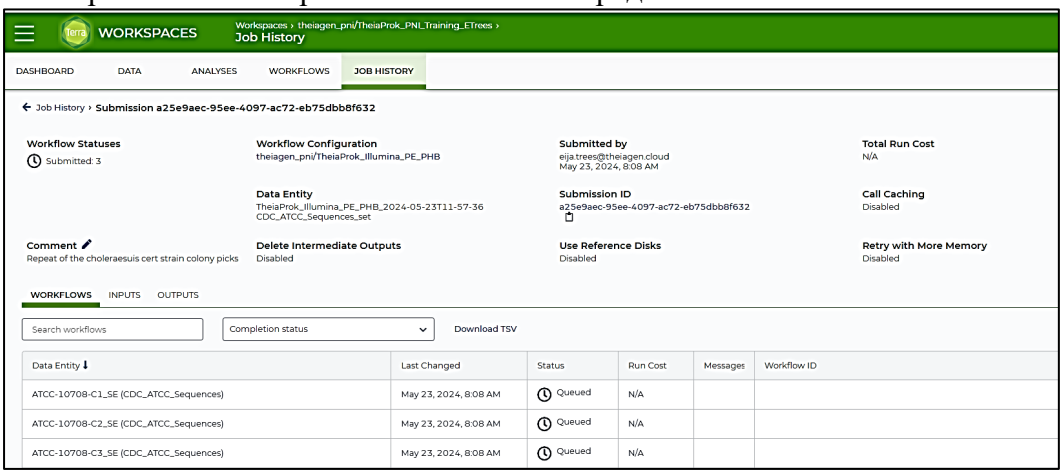
5.4.2.8 Нажмите "Сохранить" (скриншот выше).

ПРИМЕЧАНИЕ: *кнопка "Сохранить" видна только в том случае, если вы изменили входные данные по сравнению с предыдущим представлением.*

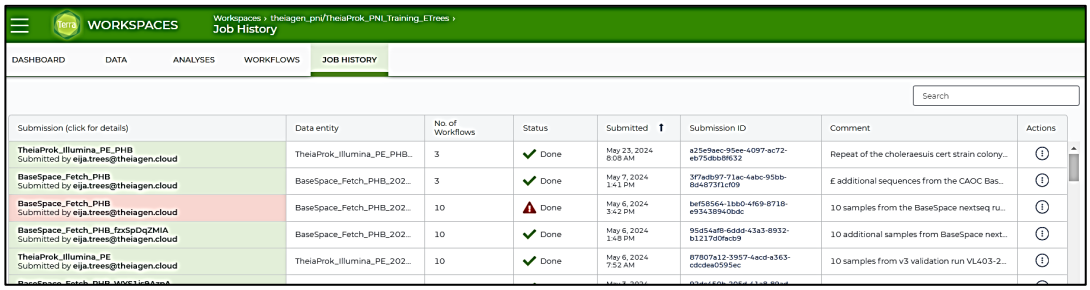
5.4.2.9 Нажмите "Запустить анализ". Появится всплывающее окно "Подтвердить запуск", в котором можно ввести дополнительное описание. Нажмите "Запустить".



5.4.3 Появится окно "Статусы рабочего процесса", где отправленные задания должны быть первоначально перечислены как "В очереди".

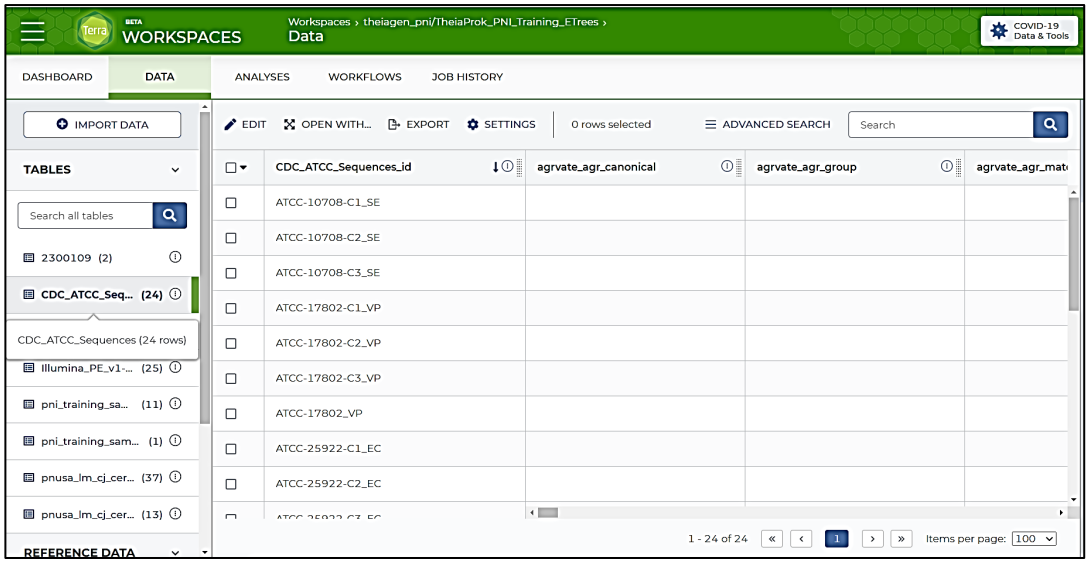


5.4.4 Перейдите на вкладку "История заданий", чтобы проверить статус отправленного задания. Успешно завершенное задание обозначается зеленой галочкой.

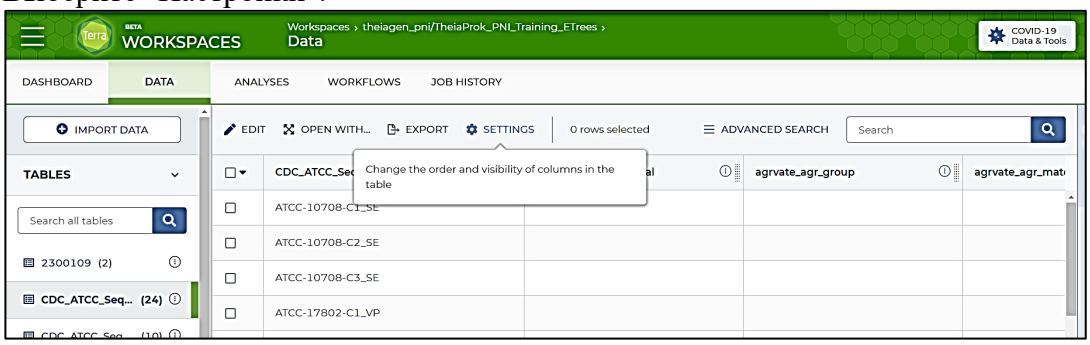


5.5 **Оцените метрики КК для последовательностей:** Показатели КК можно просмотреть либо непосредственно в таблице данных (5.5.1-5.5.3), либо экспортировать их в Excel для выбранных записей (5.5.4).

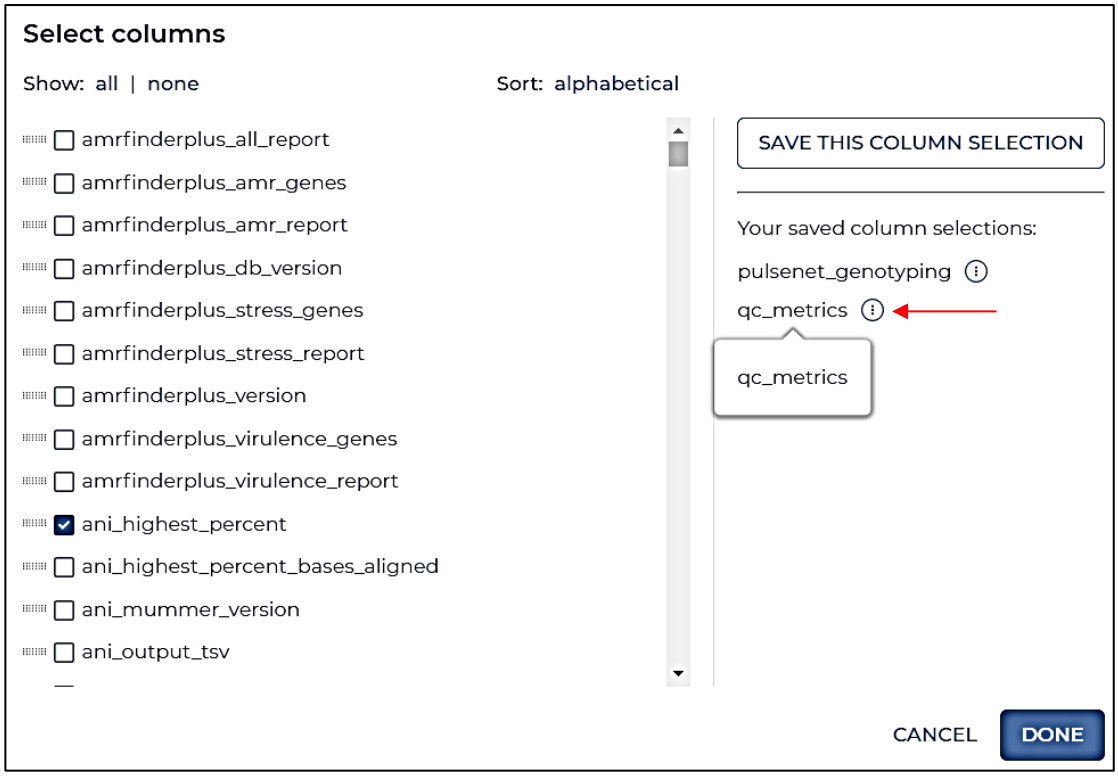
5.5.1 В разделе "Рабочее пространство Terra" выберите вкладку "Данные", затем выберите интересующую таблицу данных, например, "CDC_ATCC_Sequences".

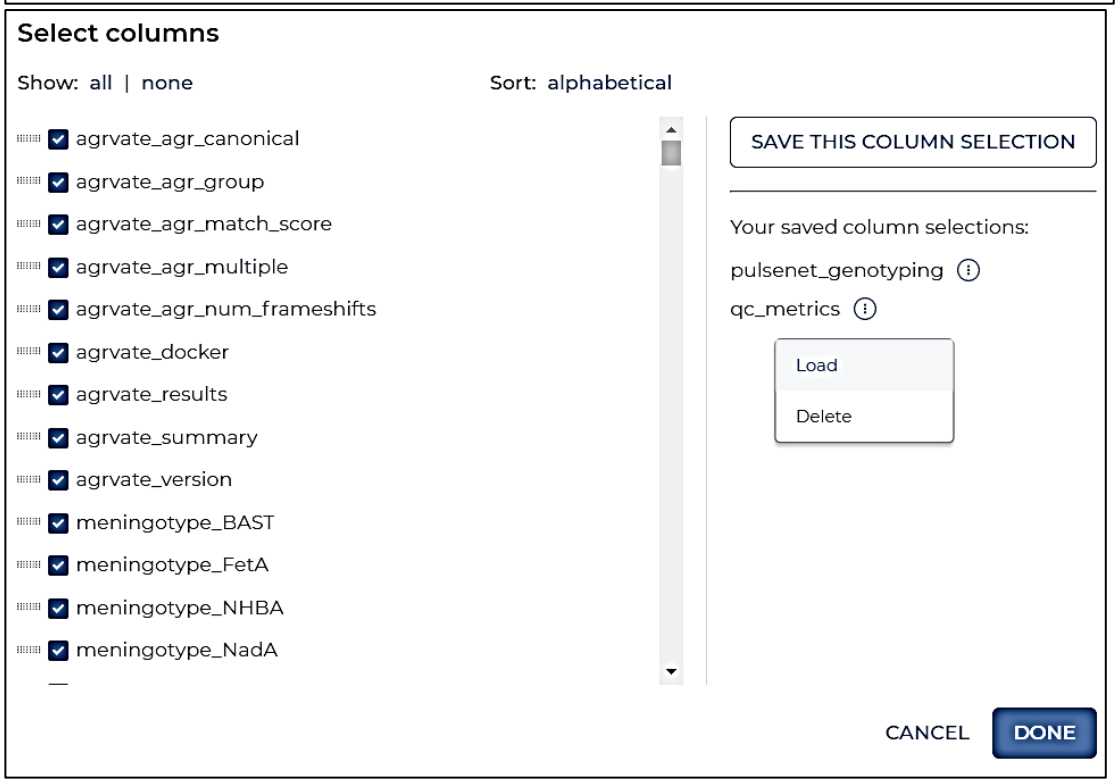


5.5.2 Выберите "Настройки".



5.5.3 На экране "Select columns" в разделе "Your saved column selections" нажмите на кружок с тремя точками рядом с "qc_metrics" и в выпадающем меню выберите "Load", а затем нажмите "Done". В результате в таблицу данных будут загружены соответствующие метрики контроля качества PulseNet. См. приложение PNID01-3, где приведены метрики QC, которые должны отображаться в таблице, а также инструкции по добавлению или удалению столбцов (метрик КК) в таблице метрик КК.

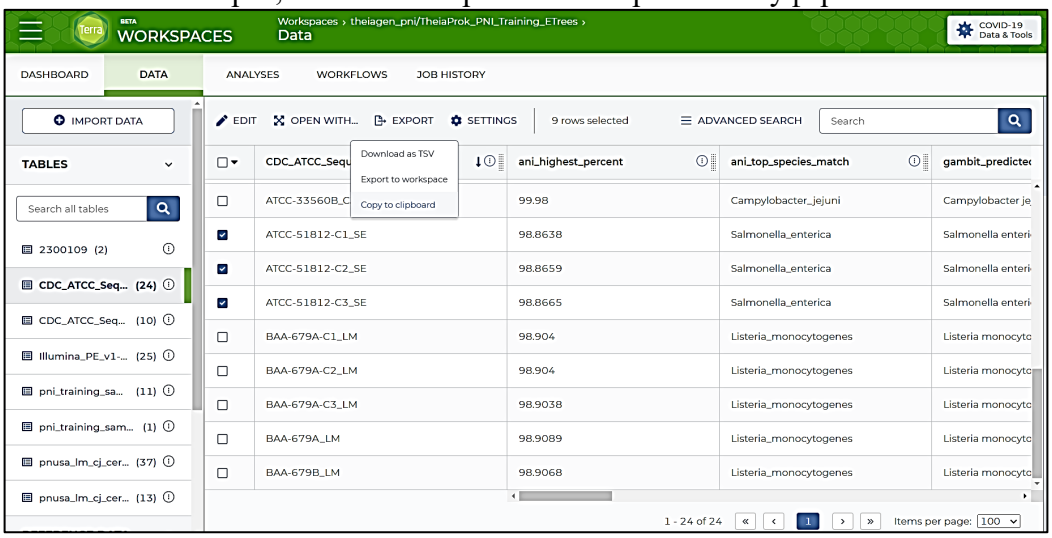




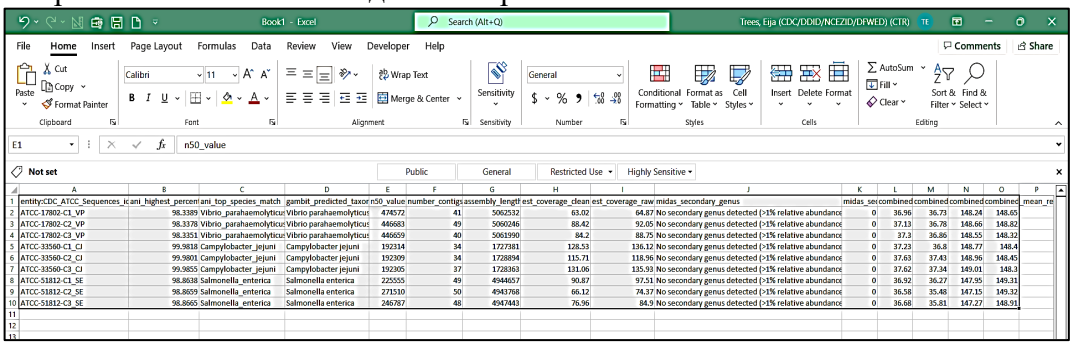
5.5.4 Экспорт показателей КК в Excel:

5.5.4.1 Выберите нужные записи.

5.5.4.2 Нажмите "Экспорт", а затем выберите "Скопировать в буфер обмена".

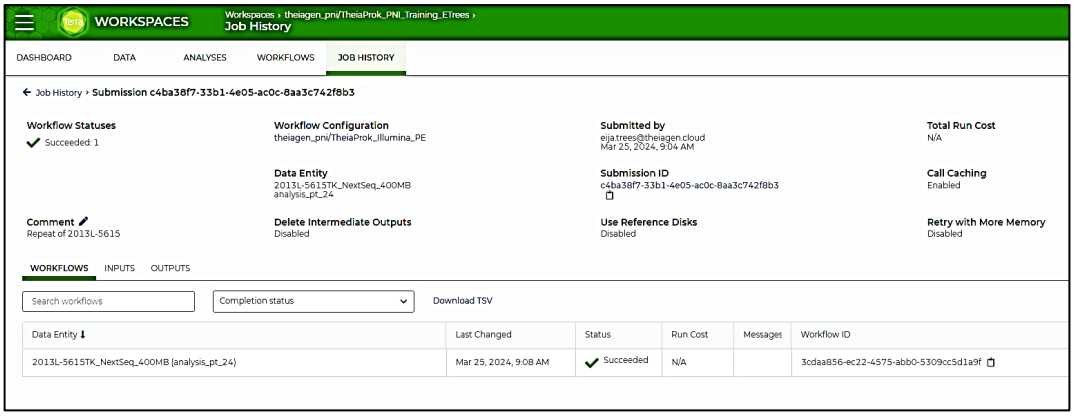


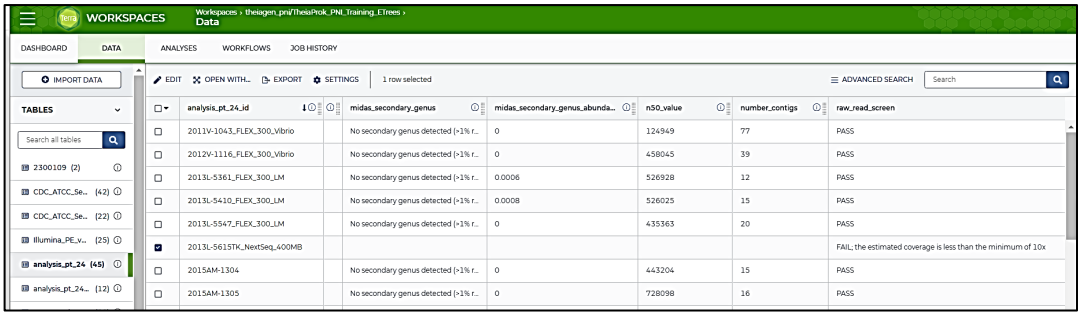
5.5.4.3 Откройте Excel и вставьте данные в рабочий лист.



5.5.5 См. приложение PNID01-4a для критических показателей качества PulseNet для приемлемых последовательностей Illumina.

5.5.6 В TheiaProk используется этап предварительной проверки считывания, на котором анализ будет остановлен для образцов, качество которых ниже определенных порогов. В этом случае на вкладке "История заданий" указано, что анализ прошел успешно, но на вкладке "Данные" нет результатов. В столбце "raw_read_screen" в метрике КК должна быть указана причина сбоя анализа последовательности. Пороговые значения, применяемые на этом этапе предварительной проверки чтения, см. в приложении PNID01-4b.



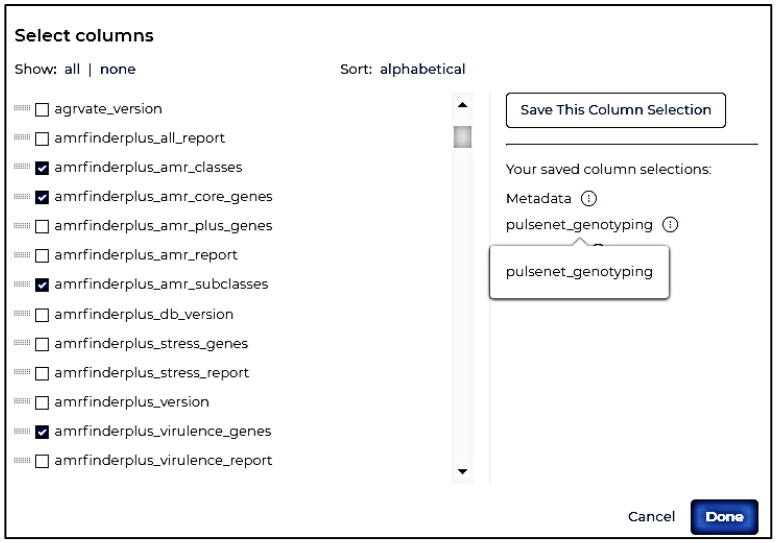


5.6 **Просмотрите результаты генотипирования последовательностей:** Результаты генотипирования можно просмотреть либо непосредственно в таблице данных (5.6.1-5.6.3), либо экспортировать их в Excel для выбранных записей (5.6.4).

5.6.1 В разделе "Рабочее пространство Terra" выберите вкладку "Данные", затем выберите интересующую вас таблицу данных, например "CDC_ATCC_Sequences".

5.6.2 На вкладке "Данные" выберите "Настройки".

5.6.3 На экране "Select columns" в разделе "Your saved column selections" нажмите на кружок с тремя точками рядом с "pulsenet_genotyping" и в выпадающем меню выберите "Load", а затем нажмите "Done". В результате в таблицу данных будут загружены анализы для генотипирования, подходящие для наблюдения PulseNet. Обратитесь к приложению PNID01-5, чтобы узнать, какие анализы для генотипирования должны появиться в таблице, а также получить инструкции по добавлению или удалению любого из столбцов (анализы для генотипирования) в таблице результатов генотипирования.



5.6.4 Экспортируйте результаты в Excel для выбранных записей; следуйте процедуре, описанной в шаге 5.5.4.

5.7 **Загрузка последовательностей в NCBI**

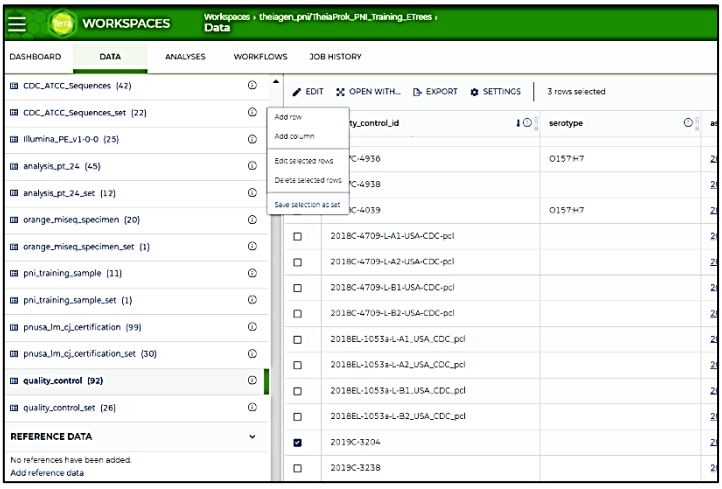
ПРИМЕЧАНИЕ: *Перед началом работы обратитесь в компанию Theiagen Genomics (support@theiagen.com) за консультацией и настройкой рабочего пространства для загрузки данных в NCBI. Процесс настройки описан в: https://theiagen.notion.site/Terra_2_NCBI-61abcedc066646b3b258f70b561e9f62.*

5.7.1 Загрузите метаданные, необходимые для отправки в NCBI: обратитесь к приложению PNID01-6 для правильного форматирования и загрузки метаданных.

5.7.2 Создайте **набор** образцов для загрузки в NCBI:

5.7.2.1 В разделе "Рабочее пространство Terra" выберите вкладку "Данные", затем выберите интересующую вас таблицу данных, например, "quality_control".

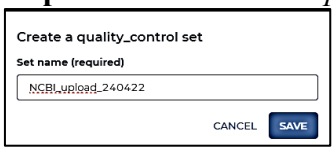
5.7.2.2 На вкладке "Данные" выберите последовательности, которые будут включены в выгрузку в NCBI.



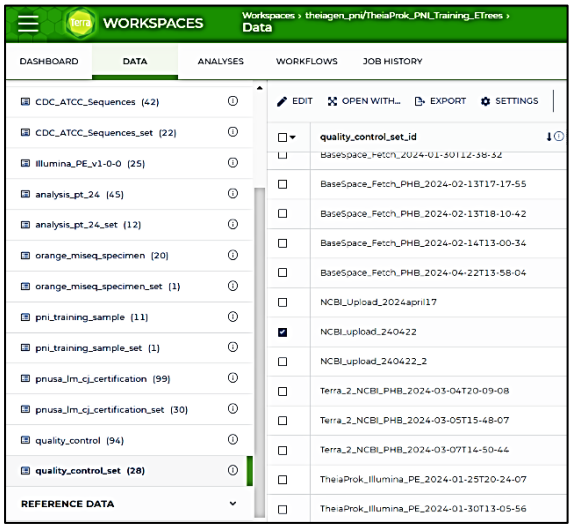
5.7.2.3 В раскрывающемся меню "Редактировать" выберите "Сохранить выборку как набор".

5.7.2.4 В появившемся всплывающем окне присвойте набору имя, например "NCBI_upload_240422", и нажмите "Сохранить".

Обратите внимание: *пробелы и тире не допускаются.*

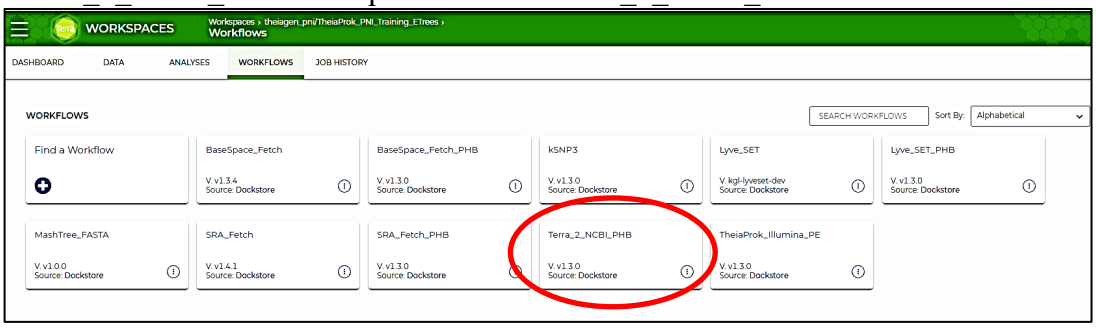


5.7.2.5 Созданный набор должен появиться в таблице данных "Набор", например, "quality_control_set".

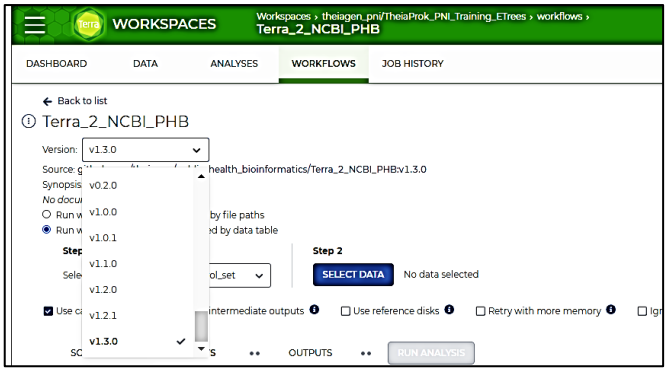


5.7.3 Настройте параметры для рабочего процесса загрузки в NCBI:

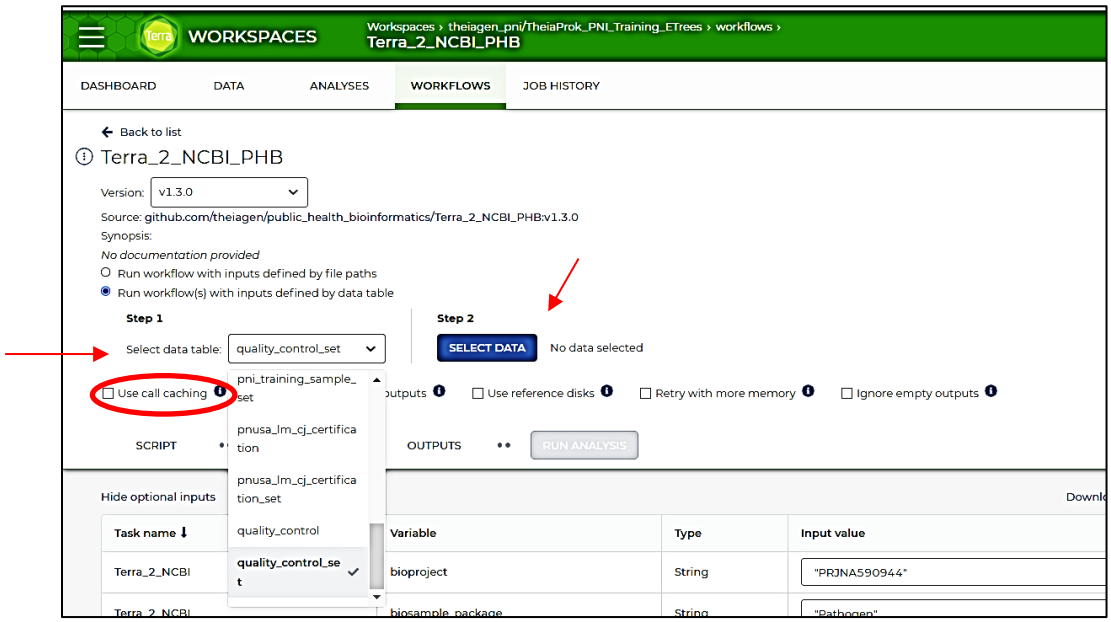
5.7.3.1 На вкладке "Рабочие процессы" выберите рабочий процесс "Terra_2_NCBI_PHB". Откроется окно "Terra_2_NCBI_PHB".



5.7.3.2 В раскрывающемся меню "Версия" выберите последнюю версию "Terra_2_NCBI_PHB".

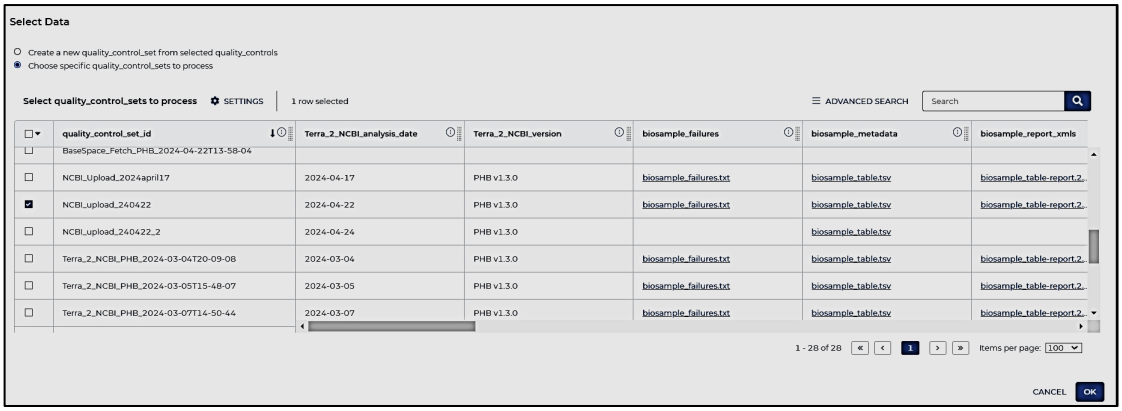


5.7.3.3 На шаге 1 в раскрывающемся меню "Выбрать тип корневой сущности" выберите таблицу данных **набора**, в которой находится набор образцов, созданный на шаге 5.7.2, например, "quality_control_set". Также снимите флажок "Use call caching".



5.7.3.4 На шаге 2 нажмите "Выбрать данные" (скриншот выше). Это приведет вас к таблице данных **набора**, выбранной на предыдущем шаге.

5.7.3.5 Выберите нужный набор образцов, например "NCBI_upload_240422", и нажмите "ОК".



5.7.3.6 На вкладке "**Входные данные**" необходимо заполнить следующие "Входные значения" для "Переменных", перечисленных ниже:

5.7.3.6.1 bioproject: введите номер биопроекта NCBI, в который вы хотите загрузить данные, в кавычках, например, "PRJNA590944".

5.7.3.6.2 biosample_package в кавычках: "Pathogen". Это шаблон/пакет метаданных, который вы используете для загрузки метаданных для последовательностей, принадлежащих к наблюдению PulseNet.

5.7.3.6.3 ncbi_config_js: введите имя файла конфигурации NCBI, созданного для вашего рабочего пространства, например workspace.ncbi_config_etrees.

5.7.3.6.4 имя_проекта в кавычках: "theiagen_pni".

5.7.3.6.5 sample_names: введите имя таблицы данных в формате: **this.data_table_names.data_table_name_id**, например, this.quality_controls.quality_control_id.
ПРИМЕЧАНИЕ: *двойной формат имени является ОБЯЗАТЕЛЬНЫМ.*

5.7.3.6.6 sra_transfer_gcp_bucket в кавычках: "gs://theiagen_sra_transfer". Это временное общедоступное хранилище Google для ваших последовательностей, к которому может получить доступ NCBI.

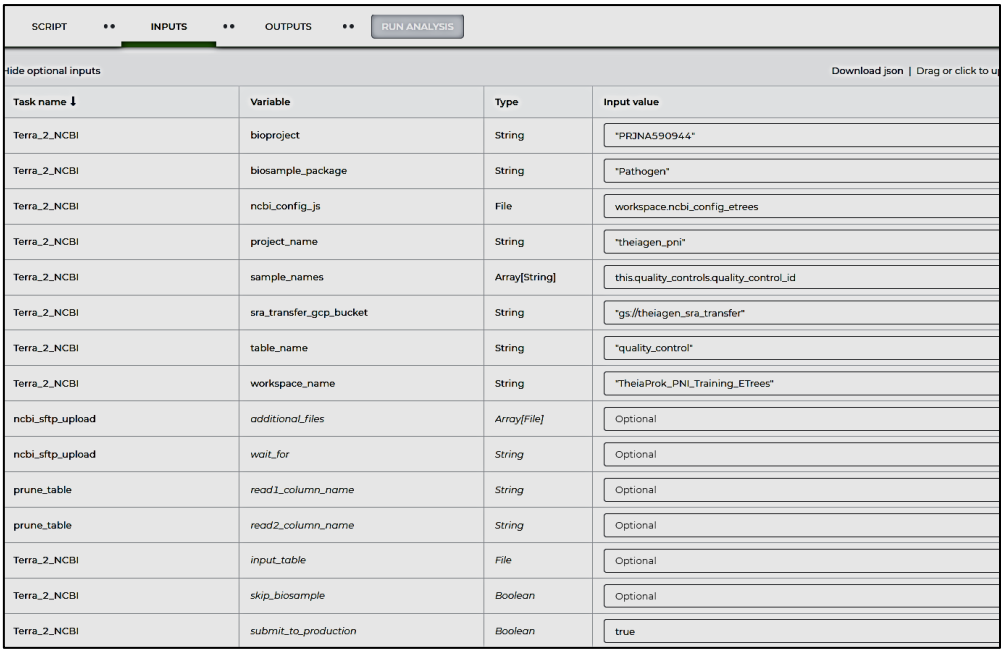
5.7.3.6.7 имя_таблицы: введите название вашей таблицы данных в кавычках, например "quality_control".

5.7.3.6.8 имя_рабочего_пространства: введите имя вашего рабочего пространства в кавычках, например, "TheiaProk_PNI_Training_ETrees".

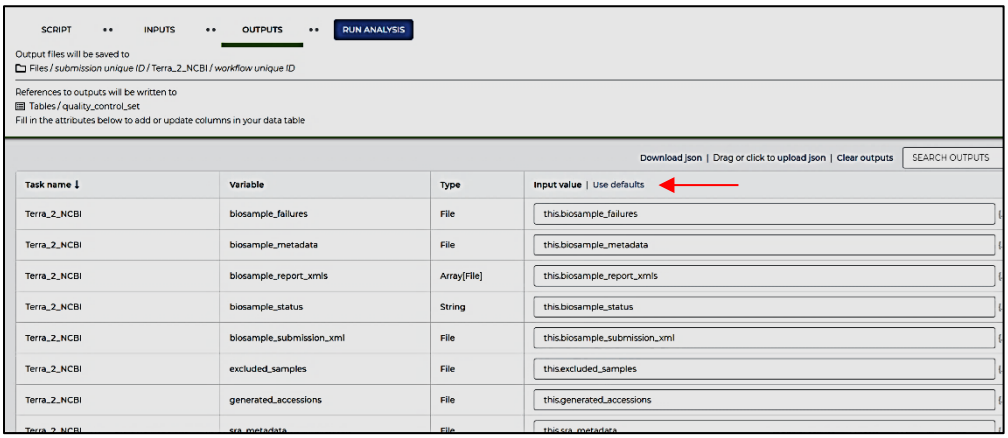
5.7.3.6.9 submit_to_production: true.

5.7.3.6.10 ОПЦИОНАЛЬНО: если вы хотите связать SRA-подачу с **существующим биообразцом** (поле "biosample_accession" уже заполнено номером SAMN на вкладке "Data"):

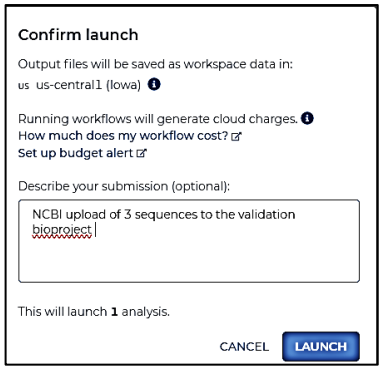
5.7.3.6.10.1 skip_biosample: true.
ПРИМЕЧАНИЕ: *для использования этой функции необходимо заполнить все необходимые поля метаданных, т. е. заполнение только поля "biosample_accession" недостаточно и не позволит пройти проверку Terra перед загрузкой.*



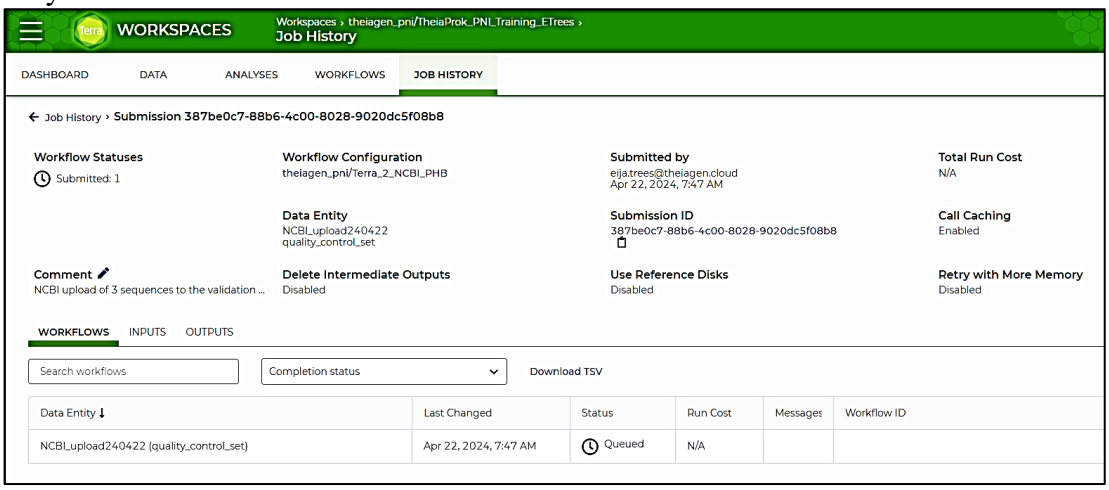
5.7.3.7 На вкладке "**Выходные данные**" нажмите "Использовать значения по умолчанию" для "Атрибутов", а затем нажмите "Сохранить". **ПРИМЕЧАНИЕ**: *кнопка "Сохранить" отображается только в том случае, если вы изменили входные данные по сравнению с предыдущим представлением.*



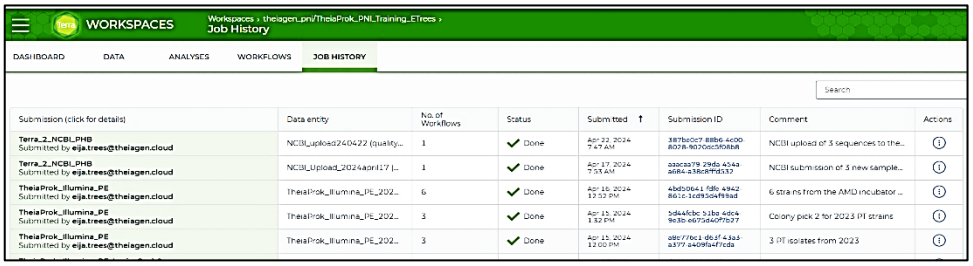
5.7.3.8 Нажмите "Запустить анализ" (скриншот выше). Появится всплывающее окно "Подтвердить запуск", в котором можно ввести дополнительное описание. Нажмите "Запустить".



5.7.3.9 Появится окно "Статусы рабочего процесса", в котором отправленные задания должны быть первоначально перечислены как "В очереди" или "В процессе запуска".



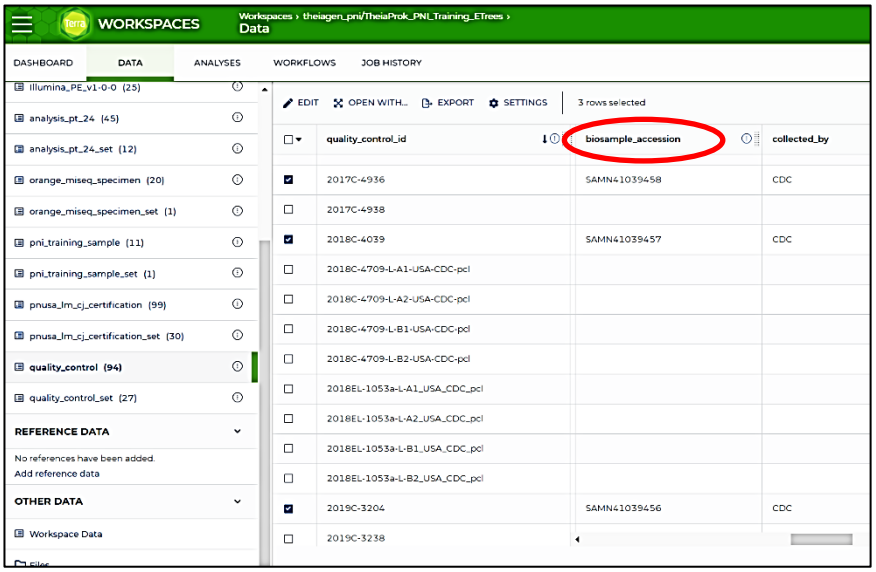
5.7.3.10Перейдите на вкладку "История заданий", чтобы проверить статус отправленного задания. Успешно завершенное задание обозначается зеленой галочкой.



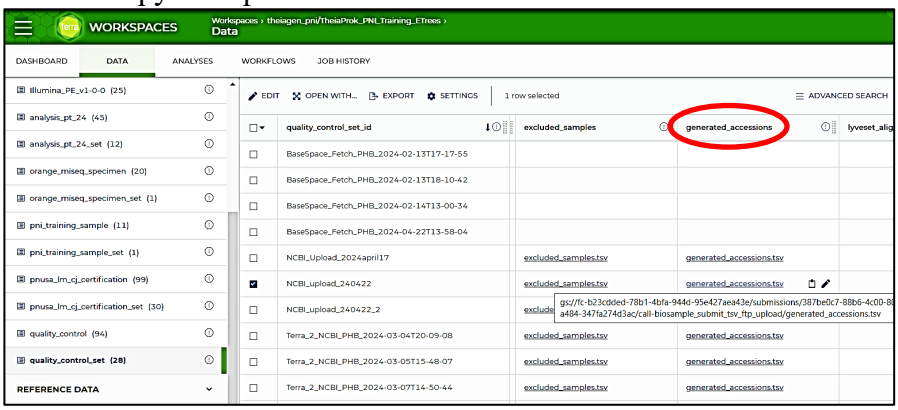
5.7.3.11Отслеживание номеров доступа NCBI:

5.7.3.11.1 Номера **доступа к биообразцам** (номера SAMN)

5.7.3.11.1.1 На вкладке "Данные" перейдите в таблицу данных, о которой идет речь, и вы должны увидеть поле "biosample_accession", заполненное номером SAMN, который присваивается каждому уникальному биообразцу.

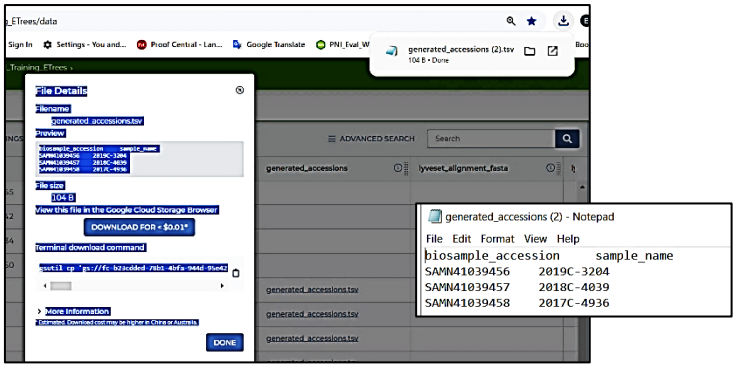


5.7.3.11.1.2 На вкладке "Данные" перейдите к рассматриваемому **набору** таблиц данных, найдите набор данных, который вы создали для загрузки в NCBI, а затем нажмите на ссылку "generated_accessions" для получения файла tsv, в котором перечислены номера биообразцов для загруженного набора последовательностей. Чтобы загрузить файл:

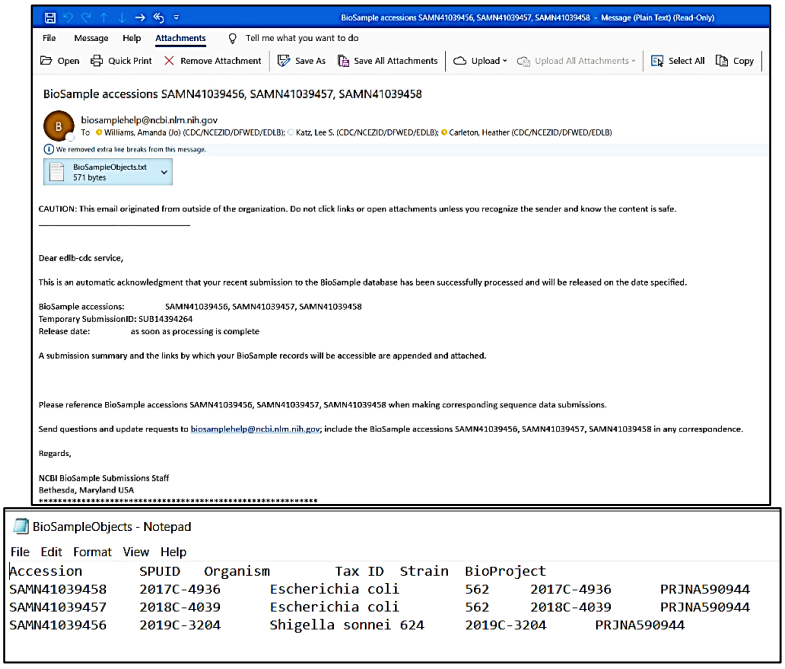


5.7.3.11.1.2.1 Нажмите "Download >$0.01*", а затем нажмите "Done".

5.7.3.11.1.2.2 Загруженный файл появится в правом верхнем углу.

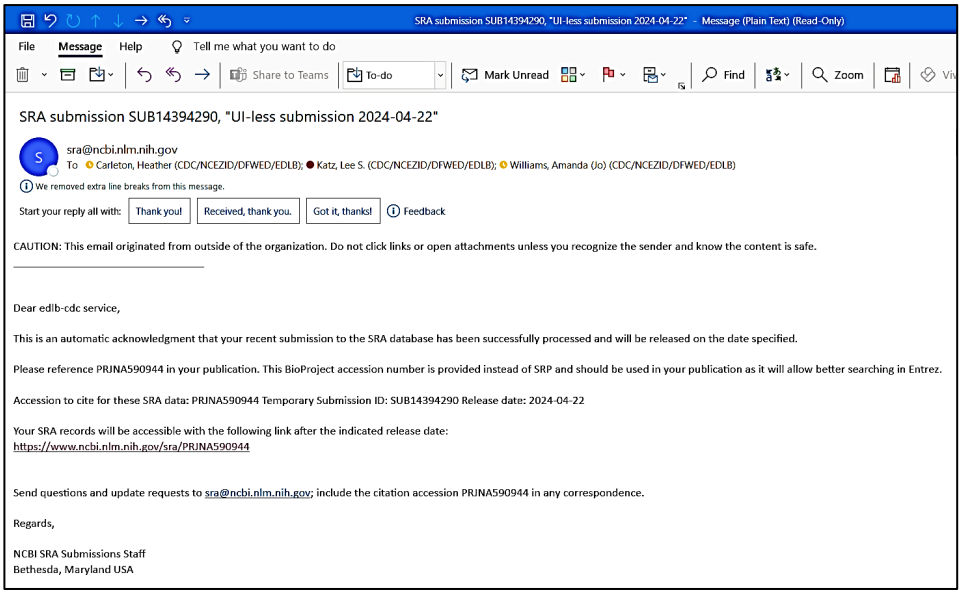


5.7.3.11.1.3 На адрес электронной почты, связанный с учетной записью NCBI, использованной для загрузки, должно прийти письмо с подтверждением, в котором будут указаны номера SAMN в теле письма и в текстовом файле "BiosampleObjects", прикрепленном к ел. почте.

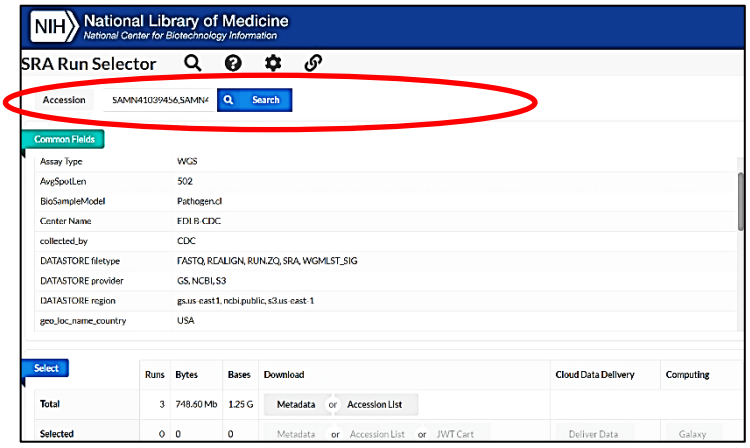


5.7.3.11.2 Номера **присоединения SRA** (номера SRR)

ПРИМЕЧАНИЕ: *На адрес электронной почты, связанный с учетной записью NCBI, использованной для загрузки, должно прийти второе письмо с подтверждением, в котором говорится, что заявка в базу данных SRA успешно обработана и будет опубликована на сайте в указанную в письме дату. Это письмо не содержит номеров доступа SRA.*

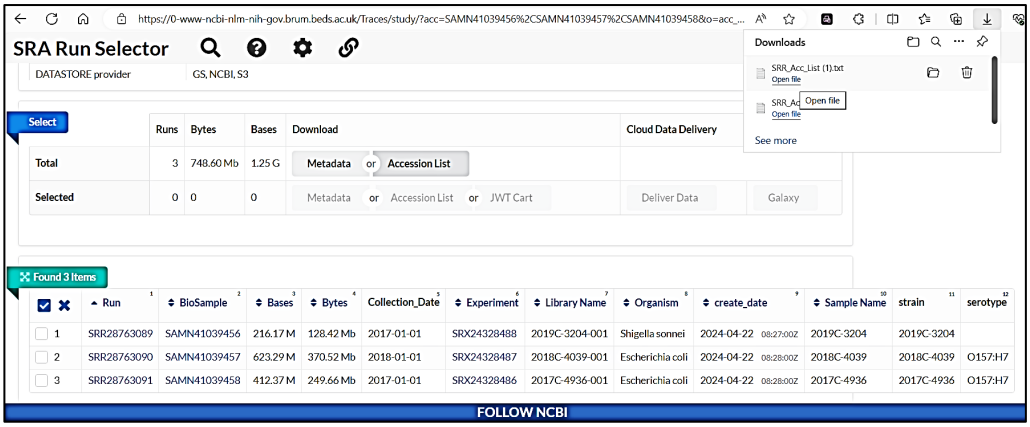


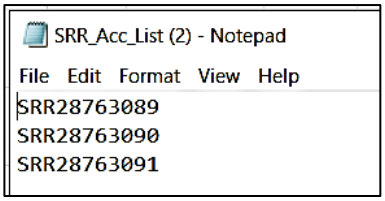
5.7.3.11.2.1 Скопируйте и вставьте номера SAMN из файла tsv с шага 5.7.3.11.1.2. или txt-файла с шага 5.7.3.11.1.3. в инструмент "NCBI Run Selector" по адресу: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/Traces/study/. Разделите числа запятыми. Нажмите на кнопку "Поиск".



5.7.3.11.2.2 В результате поиска появится таблица, содержащая всю информацию NCBI о ваших последовательностях, включая номера SRR.

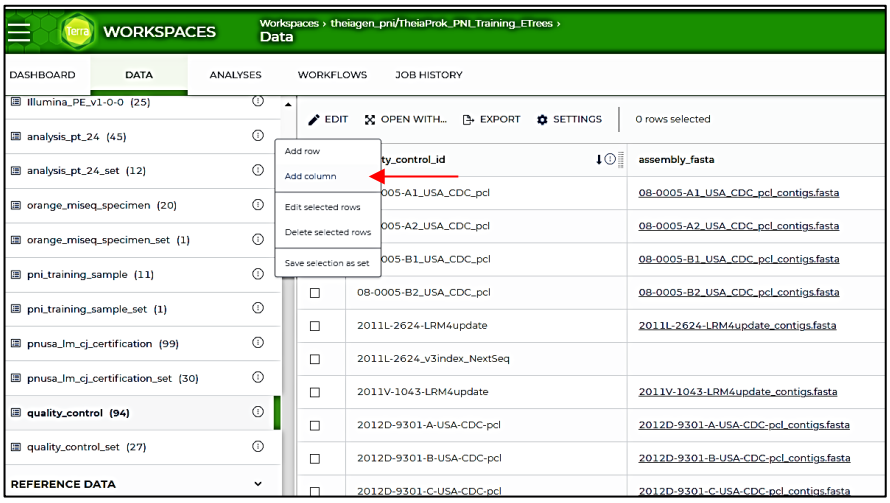
5.7.3.11.2.3 Нажмите на "Accession List", чтобы загрузить таблицу с номерами присоединения. Загруженный файл появится в правом верхнем углу.



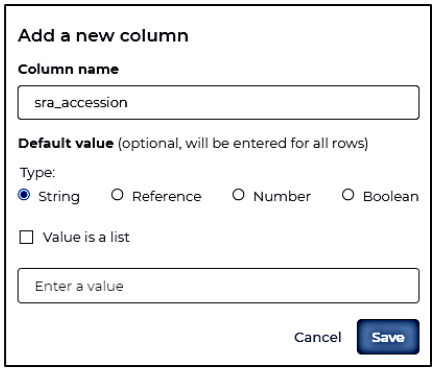


5.7.3.11.2.4 Вы можете отслеживать номера вступления в CPA либо в отдельной электронной таблице Excel, либо в системе LIMS, либо создать поле "sra_accession" в таблице данных Terra и копировать и вставлять туда номера вступления для каждой записи образца. Чтобы создать колонку sra_accession и использовать ее для отслеживания, выполните следующие действия:

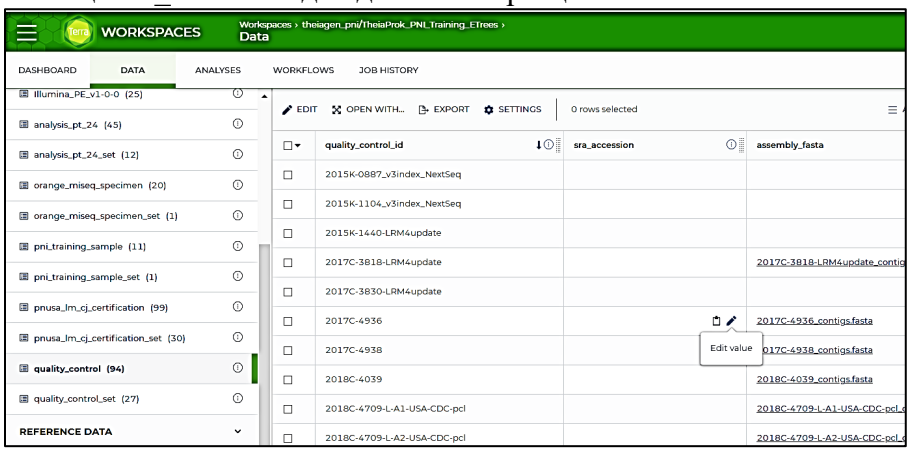
5.7.3.11.2.4.1 В раскрывающемся меню "Редактировать" выберите "Добавить колонку".



5.7.3.11.2.4.2 Во всплывающем окне "Добавить новый столбец" назовите новый столбец "sra_accession" и нажмите "Сохранить".

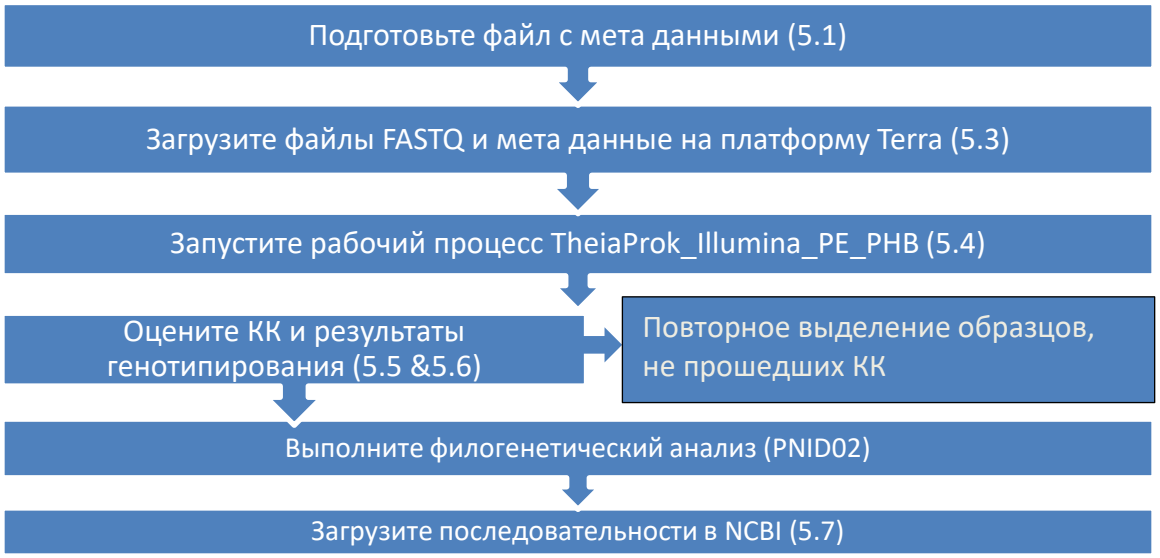


5.7.3.11.2.4.3 В таблице данных должен появиться новый столбец. Чтобы ввести номер присоединения SRR для конкретной последовательности, нажмите кнопку "Изменить значение" в столбце sra_accession для данного образца.



5.7.3.11.2.4.4 Вставьте номер присоединения SRA в поле во всплывающем окне "Изменить значение" и нажмите "Сохранить изменения".

6. БЛОК-СХЕМА:



7. СОПУТСТВУЮЩИЕ ДОКУМЕНТЫ:

7.1 **PNID02:** PulseNet International Standard Operating Procedure for Phylogenetic Analysis of WGS Data Using the Terra.Bio Platform.

8. ССЫЛКИ:

8.1 Libuit K.G., Doughty E.L., Otieno J.R., Ambrosio F., Kapsak C.J., Smith E.A., Wright S.M., Scribner M.R., Petit III R.A., Mendes C.I., Huergo M., Legacki G., Loreth C., Park D.J., Sevinsky J.R. (2023) Accelerating bioinformatics implementation in public health. Microbial Genomics 9:001051.

9. КОНТАКТЫ:

9.1 Лаборатория CDC USA PulseNet NGS: pulsenetngslab@cdc.gov
9.2 PulseNet International Quality Assurance Coordinator Eija Trees: ehyytia-trees@cdc.gov
9.3 Theiagen:
9.3.1 Общий e-mail для поддержки: support@theiagen.com
9.3.2 Мишель Скрибнер: michelle.scribner@theiagen.com
9.3.3 Фрэнк Амброзио: frank.ambrosio@theiagen.com

10. ПОПРАВКИ: Отсутствуют.

11. ПОДПИСИ:

Утверждено: _____ Date: _____
Персонал PulseNet QA/QC

Утверждено: _____ Date: _____
Технический руководитель PulseNet WGS

Утверждено: _____ Date: _____
Международный координатор PulseNet

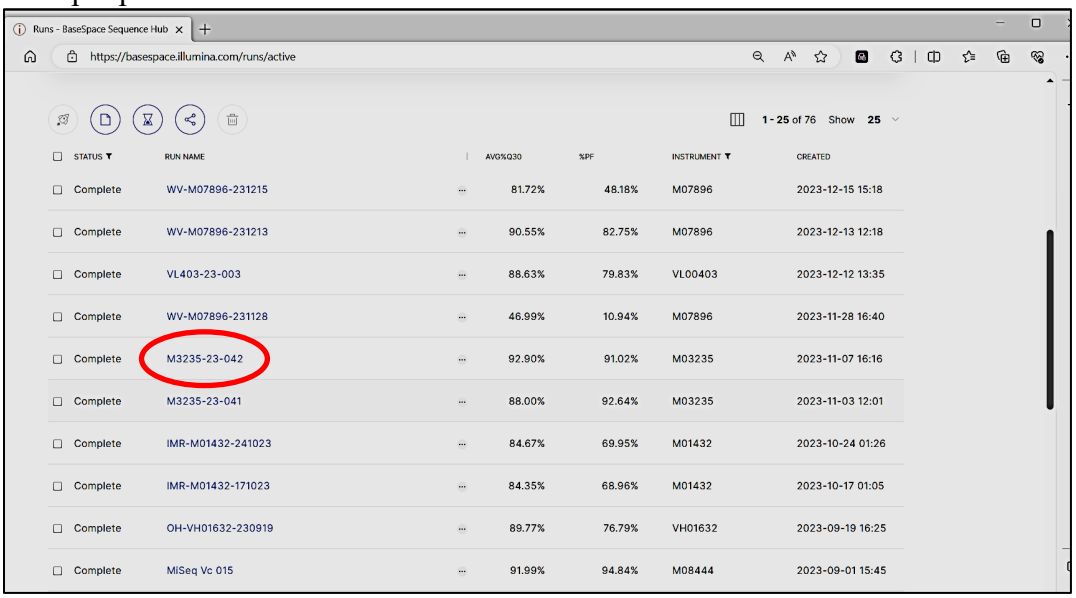
Утверждено_____ Date: _____
Руководитель группы реагирования и управления вспышками PulseNet

Утверждено_____ Date: _____
Заведующий лабораторией энтеральных болезней

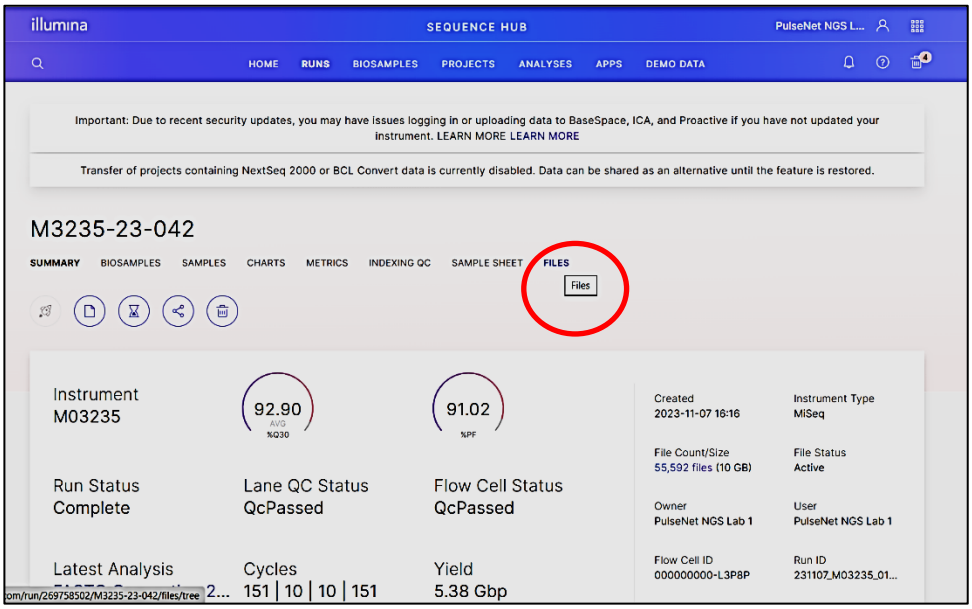
Приложение PNID01-1: Импорт данных в Terra непосредственно из Illumina BaseSpace

ПРИМЕЧАНИЕ: *Чтобы настроить рабочее пространство Terra для подключения к учетной записи Illumina BaseSpace, следуйте инструкциям, размещенным на сайте ресурсов Theiagen по адресу https://theiagen.notion.site/BaseSpace_Fetch-34978656aa2d46ba82f2059434bd9369. За дополнительной помощью обращайтесь в компанию Theiagen (контактную информацию см. в разделе "Контакты" (9)).*

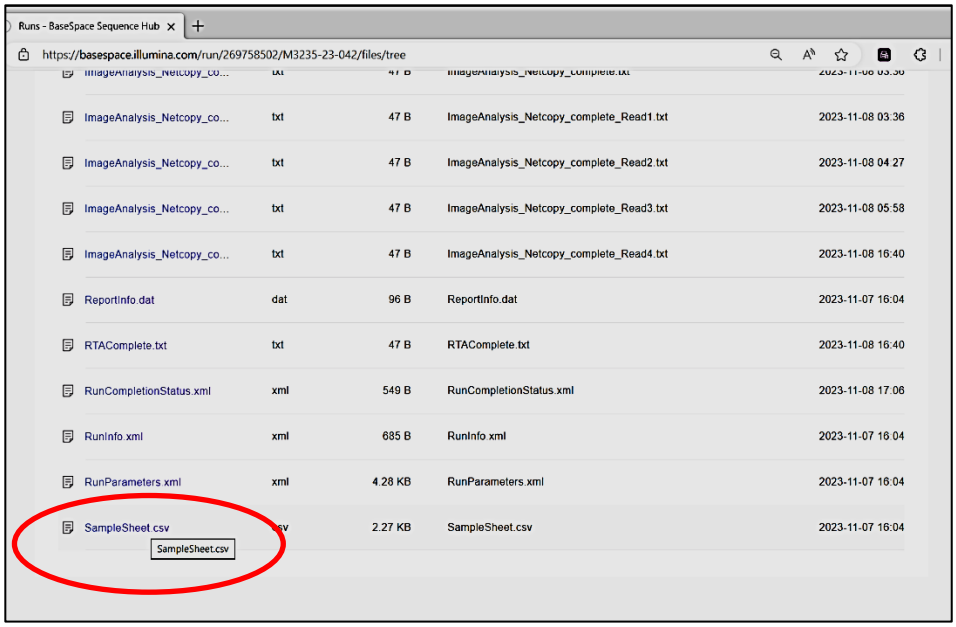
1. Войдите в свою учетную запись BaseSpace и найдите прогон, который необходимо импортировать в Terra.



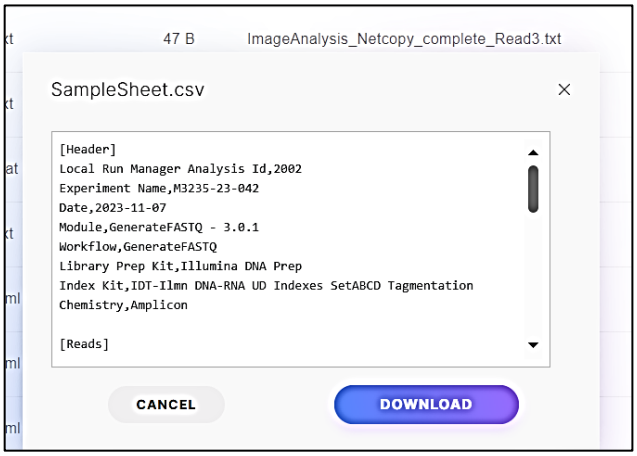
2. Загрузите таблицу образцов для прогона:
    a. Перейдите на вкладку "Файлы" и прокрутите страницу до самого низа.



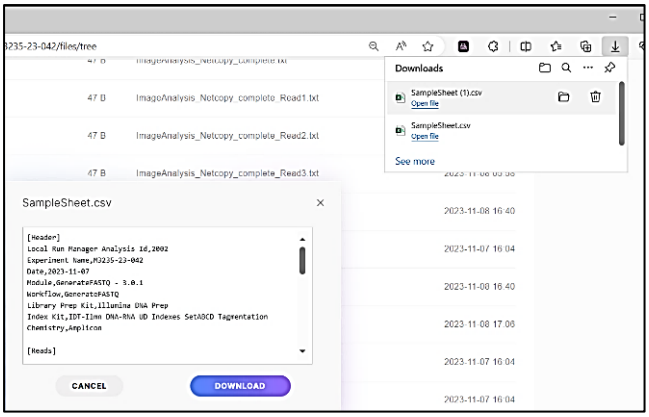
b. Нажмите на ссылку SampleSheet.csv.



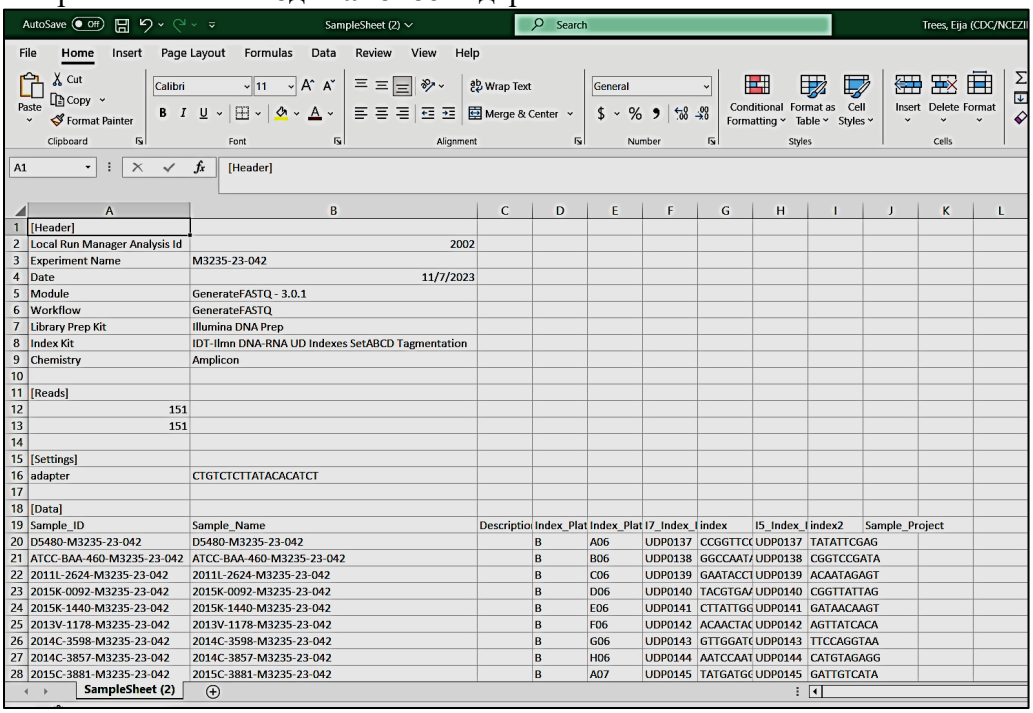
c. Во всплывающем окне "SampleSheet.csv" нажмите на кнопку "Загрузить".



d. Загруженный файл csv появится в правом верхнем углу в разделе "Загрузки".



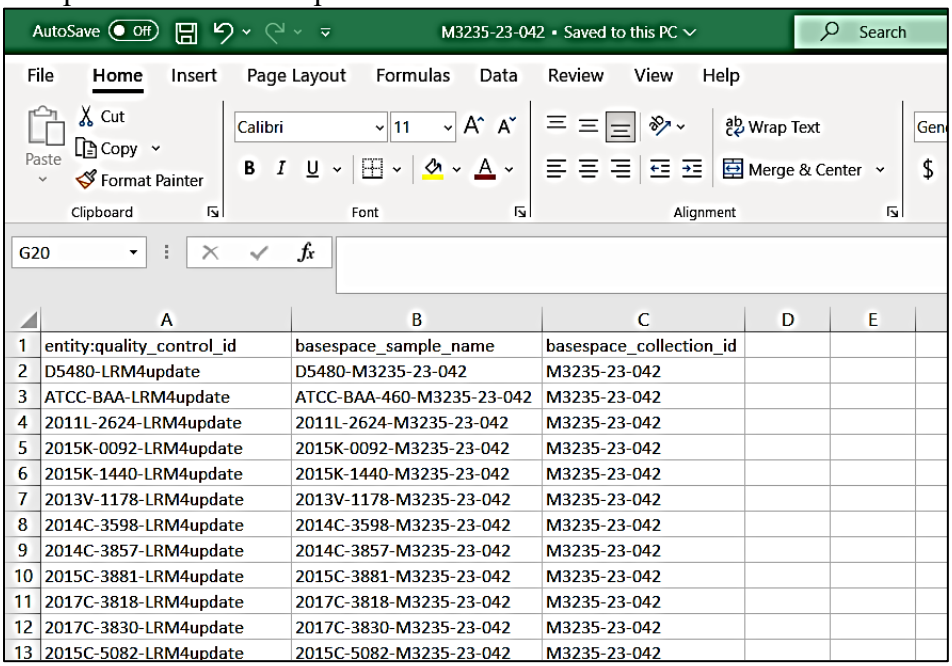
3. Откройте файл SampleSheet. Столбцы, необходимые для файла метаданных tsv, будут зависеть от столбцов и содержания столбцов, присутствующих в листе образцов.
4. Подготовьте файл метаданных tsv:
    a. Столбцы необходимы, если столбцы "Sample_Name" и "Sample_ID" в SampleSheet имеют **одинаковое** содержание:



	A	B	C	D	E	F	G	H	I	J	K	L
1	[Header]											
2	Local Run Manager Analysis Id	2002										
3	Experiment Name	M3235-23-042										
4	Date	11/7/2023										
5	Module	GenerateFASTQ - 3.0.1										
6	Workflow	GenerateFASTQ										
7	Library Prep Kit	Illumina DNA Prep										
8	Index Kit	IDT-Ilmn DNA-RNA UD Indexes SetABCD Tagmentation										
9	Chemistry	Amplicon										
10												
11	[Reads]											
12	151											
13	151											
14												
15	[Settings]											
16	adapter	CTGTCTCTTATACACATCT										
17												
18	[Data]											
19	Sample_ID	Sample_Name	Description	Index_Plat	Index_Plat	I7_Index_I	index	I5_Index_I	index2	Sample_Project		
20	D5480-M3235-23-042	D5480-M3235-23-042		B	A06	UDP0137	CCGGTTCC	UDP0137	TATATTCGAG			
21	ATCC-BAA-460-M3235-23-042	ATCC-BAA-460-M3235-23-042		B	B06	UDP0138	GGCCAATA	UDP0138	CGGTCCGATA			
22	2011L-2624-M3235-23-042	2011L-2624-M3235-23-042		B	C06	UDP0139	GAATACCT	UDP0139	ACAATAGAGT			
23	2015K-0092-M3235-23-042	2015K-0092-M3235-23-042		B	D06	UDP0140	TACGTGAA	UDP0140	CGGTTATTAG			
24	2015K-1440-M3235-23-042	2015K-1440-M3235-23-042		B	E06	UDP0141	CTTATTGG	UDP0141	GATAACAAGT			
25	2013V-1178-M3235-23-042	2013V-1178-M3235-23-042		B	F06	UDP0142	ACAACTAC	UDP0142	AGTTATCACA			
26	2014C-3598-M3235-23-042	2014C-3598-M3235-23-042		B	G06	UDP0143	GTTGGATC	UDP0143	TTCCAGGTAA			
27	2014C-3857-M3235-23-042	2014C-3857-M3235-23-042		B	H06	UDP0144	AATCCAAT	UDP0144	CATGTAGAGG			
28	2015C-3881-M3235-23-042	2015C-3881-M3235-23-042		B	A07	UDP0145	TATGATGC	UDP0145	GATTGTCATA			

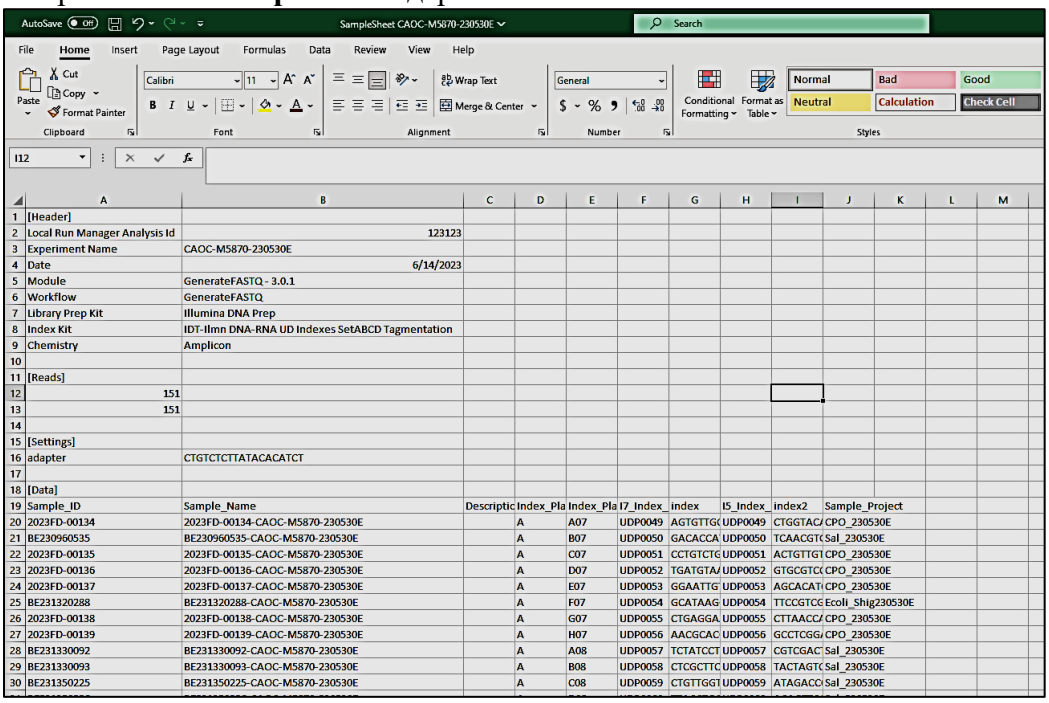
        i. **entity_datatablename_id**:
            1. Введите название таблицы данных (новой или существующей) в ячейку A1 между "entity:" и "id".
            2. Введите идентификаторы образцов в столбец A так, как вы хотите, чтобы они отображались в таблице данных Terra.
        ii. basepace_sample_name: скопируйте и вставьте содержимое поля SampleSheet "Sample_Name".

iii. basespace_collection_id: введите название Прогона/Run так, как оно отображается на BaseSpace.



	A	B	C	D	E
1	entity:quality_control_id	basespace_sample_name	basespace_collection_id		
2	D5480-LRM4update	D5480-M3235-23-042	M3235-23-042		
3	ATCC-BAA-LRM4update	ATCC-BAA-460-M3235-23-042	M3235-23-042		
4	2011L-2624-LRM4update	2011L-2624-M3235-23-042	M3235-23-042		
5	2015K-0092-LRM4update	2015K-0092-M3235-23-042	M3235-23-042		
6	2015K-1440-LRM4update	2015K-1440-M3235-23-042	M3235-23-042		
7	2013V-1178-LRM4update	2013V-1178-M3235-23-042	M3235-23-042		
8	2014C-3598-LRM4update	2014C-3598-M3235-23-042	M3235-23-042		
9	2014C-3857-LRM4update	2014C-3857-M3235-23-042	M3235-23-042		
10	2015C-3881-LRM4update	2015C-3881-M3235-23-042	M3235-23-042		
11	2017C-3818-LRM4update	2017C-3818-M3235-23-042	M3235-23-042		
12	2017C-3830-LRM4update	2017C-3830-M3235-23-042	M3235-23-042		
13	2015C-5082-LRM4update	2015C-5082-M3235-23-042	M3235-23-042		

b. Столбцы, необходимые, когда столбцы "Sample_Name" и "Sample_ID" в SampleSheet имеют **разное** содержание:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	[Header]												
2	Local Run Manager Analysis Id	123123											
3	Experiment Name	CAOC-M5870-230530E											
4	Date	6/14/2023											
5	Module	GenerateFASTQ - 3.0.1											
6	Workflow	GenerateFASTQ											
7	Library Prep Kit	Illumina DNA Prep											
8	Index Kit	IDT-Ilmn DNA-RNA UD Indexes SetABCD Tagmentation											
9	Chemistry	Amplicon											
10													
11	[Reads]												
12	151												
13	151												
14													
15	[Settings]												
16	adapter	CTGTCTCTTATACACATCT											
17													
18	[Data]												
19	Sample_ID	Sample_Name	Descriptic	Index_Pla	Index_Pla	I7_Index_	index	I5_Index_	index2	Sample_Project			
20	2023FD-00134	2023FD-00134-CAOC-M5870-230530E		A	A07	UDP0049	AGTGTTG(	UDP0049	CTGGTAC/	CPO_230530E			
21	BE230960535	BE230960535-CAOC-M5870-230530E		A	B07	UDP0050	GACACCA	UDP0050	TCAACGT(	Sal_230530E			
22	2023FD-00135	2023FD-00135-CAOC-M5870-230530E		A	C07	UDP0051	CCTGTCTC	UDP0051	ACTGTTG1	CPO_230530E			
23	2023FD-00136	2023FD-00136-CAOC-M5870-230530E		A	D07	UDP0052	TGATGTA/	UDP0052	GTGCGTC(	CPO_230530E			
24	2023FD-00137	2023FD-00137-CAOC-M5870-230530E		A	E07	UDP0053	GGAATTG	UDP0053	AGCACAT(	CPO_230530E			
25	BE231320288	BE231320288-CAOC-M5870-230530E		A	F07	UDP0054	GCATAAG	UDP0054	TTCCGTCC	Ecoli_Shig230530E			
26	2023FD-00138	2023FD-00138-CAOC-M5870-230530E		A	G07	UDP0055	CTGAGGA	UDP0055	CTTAACC/	CPO_230530E			
27	2023FD-00139	2023FD-00139-CAOC-M5870-230530E		A	H07	UDP0056	AACGCAC	UDP0056	GCCTCGG/	CPO_230530E			
28	BE231330092	BE231330092-CAOC-M5870-230530E		A	A08	UDP0057	TCTATCCT	UDP0057	CGTCGAC	Sal_230530E			
29	BE231330093	BE231330093-CAOC-M5870-230530E		A	B08	UDP0058	CTCGCTTC	UDP0058	TACTAGT(	Sal_230530E			
30	BE231350225	BE231350225-CAOC-M5870-230530E		A	C08	UDP0059	CTGTTGGT	UDP0059	ATAGACC(	Sal_230530E			

i. **entity_datatablename_id**:

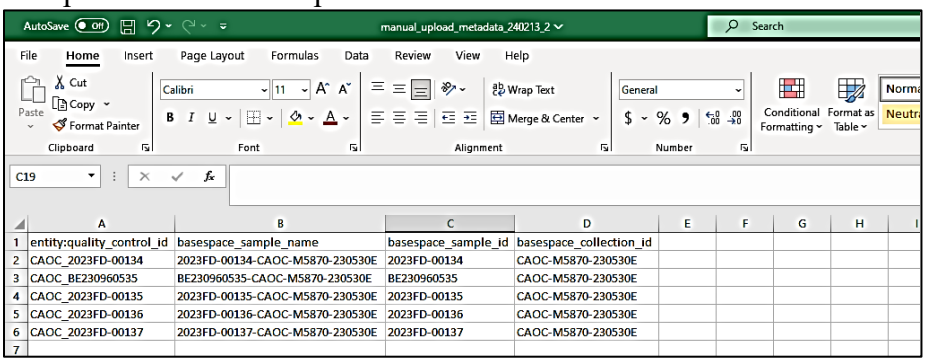
1. Введите имя таблицы данных (новой или существующей) в ячейку A1 между "entity:" и "id".

2. Введите идентификаторы образцов в столбец A так, как вы хотите, чтобы они отображались в таблице данных Terra.

ii. basepace_sample_name: скопируйте и вставьте содержимое поля SampleSheet "Sample_Name".

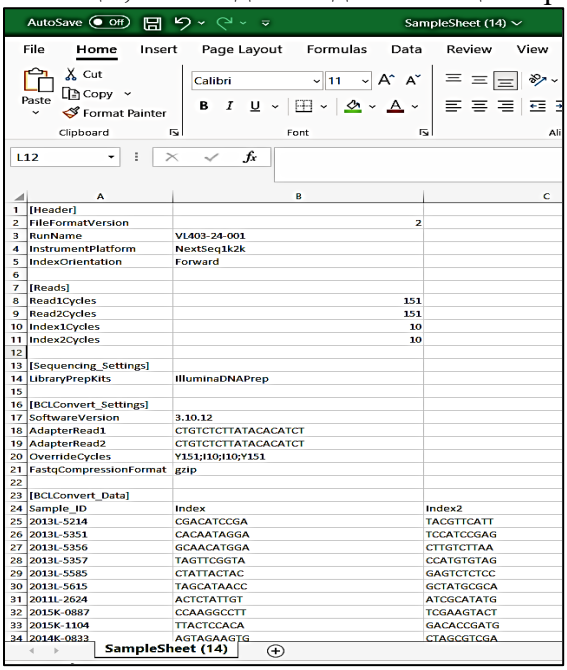
iii. basespace_sample_id: скопируйте и вставьте содержимое поля SampleSheet "Sample_ID".

iv. basespace_collection_id: введите название Run так, как оно отображается на BaseSpace.



	A	B	C	D	E	F	G	H
1	entity:quality_control_id	basespace_sample_name	basespace_sample_id	basespace_collection_id				
2	CAOC_2023FD-00134	2023FD-00134-CAOC-M5870-230530E	2023FD-00134	CAOC-M5870-230530E				
3	CAOC_BE230960535	BE230960535-CAOC-M5870-230530E	BE230960535	CAOC-M5870-230530E				
4	CAOC_2023FD-00135	2023FD-00135-CAOC-M5870-230530E	2023FD-00135	CAOC-M5870-230530E				
5	CAOC_2023FD-00136	2023FD-00136-CAOC-M5870-230530E	2023FD-00136	CAOC-M5870-230530E				
6	CAOC_2023FD-00137	2023FD-00137-CAOC-M5870-230530E	2023FD-00137	CAOC-M5870-230530E				
7								

c. Столбцы, необходимые для таблицы образцов **NextSeq**:



	A	B	C
1	[Header]		
2	FileFormatVersion	2	
3	RunName	VL403-24-001	
4	InstrumentPlatform	NextSeq1k2k	
5	IndexOrientation	Forward	
6			
7	[Reads]		
8	Read1Cycles	151	
9	Read2Cycles	151	
10	Index1Cycles	10	
11	Index2Cycles	10	
12			
13	[Sequencing_Settings]		
14	LibraryPrepKits	IlluminaDNAPrep	
15			
16	[BCLConvert_Settings]		
17	SoftwareVersion	3.10.12	
18	AdapterRead1	CTGTCTCTTATACACATCT	
19	AdapterRead2	CTGTCTCTTATACACATCT	
20	OverrideCycles	Y151;I10;I10;Y151	
21	FastqCompressionFormat	gzip	
22			
23	[BCLConvert_Data]		
24	Sample_ID	Index	Index2
25	2013L-5214	CGACATCCGA	TACGTTCATT
26	2013L-5351	CACAATAGGA	TCCATCCGAG
27	2013L-5356	GCAACATGGA	CTTGTCTTAA
28	2013L-5357	TAGTTCGGTA	CCATGTGTAG
29	2013L-5585	CTATTACTAC	GAGTCTCTCC
30	2013L-5615	TAGCATAACC	GCTATGCGCA
31	2011L-2624	ACTCTATTGT	ATCGCATATG
32	2015K-0887	CCAAGGCCTT	TCGAAGTACT
33	2015K-1104	TTACTCCACA	GACACCGATG
34	2014K-0833	AGTAGAAGTG	CTAGCGTCGA

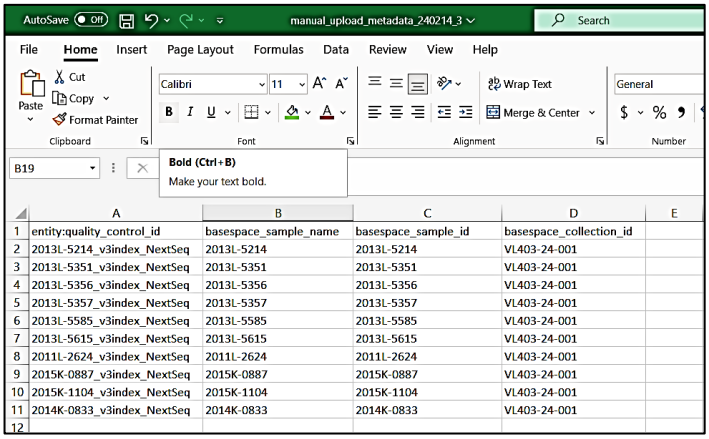
SampleSheet (14)

i. **entity_datatablename_id**:

1. Введите название таблицы данных (новой или существующей) в ячейку A1 между "entity:" и "id".
2. Введите идентификаторы образцов в столбец A так, как вы хотите, чтобы они отображались в таблице данных Terra.

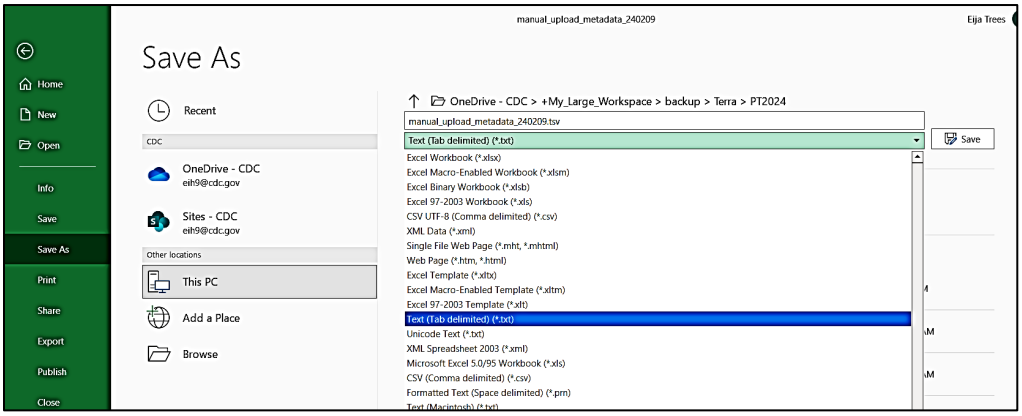
ii. basepace_sample_name: скопируйте и вставьте содержимое поля SampleSheet "Sample_ID".

iii. basespace_sample_id: скопируйте и вставьте содержимое поля SampleSheet "Sample_ID".
iv. basespace_collection_id: введите название Run так, как оно отображается на BaseSpace.

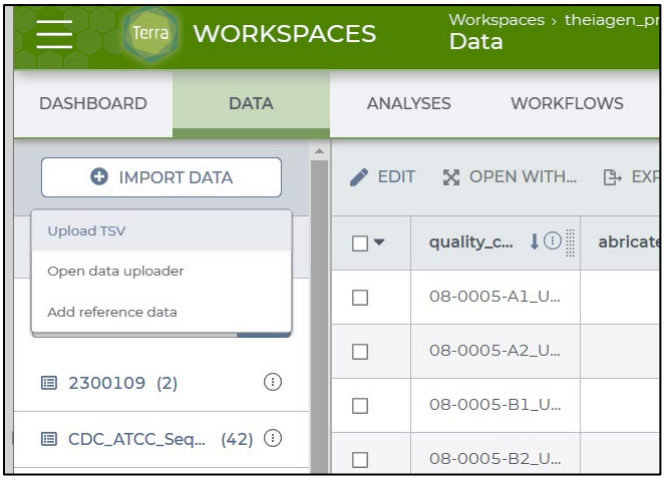


	A	B	C	D	E
1	entity:quality_control_id	basespace_sample_name	basespace_sample_id	basespace_collection_id	
2	2013L-5214_v3index_NextSeq	2013L-5214	2013L-5214	VL403-24-001	
3	2013L-5351_v3index_NextSeq	2013L-5351	2013L-5351	VL403-24-001	
4	2013L-5356_v3index_NextSeq	2013L-5356	2013L-5356	VL403-24-001	
5	2013L-5357_v3index_NextSeq	2013L-5357	2013L-5357	VL403-24-001	
6	2013L-5585_v3index_NextSeq	2013L-5585	2013L-5585	VL403-24-001	
7	2013L-5615_v3index_NextSeq	2013L-5615	2013L-5615	VL403-24-001	
8	2011L-2624_v3index_NextSeq	2011L-2624	2011L-2624	VL403-24-001	
9	2015K-0887_v3index_NextSeq	2015K-0887	2015K-0887	VL403-24-001	
10	2015K-1104_v3index_NextSeq	2015K-1104	2015K-1104	VL403-24-001	
11	2014K-0833_v3index_NextSeq	2014K-0833	2014K-0833	VL403-24-001	
12					

d. Сохраните файл в **формате tsv**: выберите "Сохранить как" и "Текст (с разделителем табуляции) (*.txt)". Убедитесь, что имя файла имеет окончание "**.tsv**".

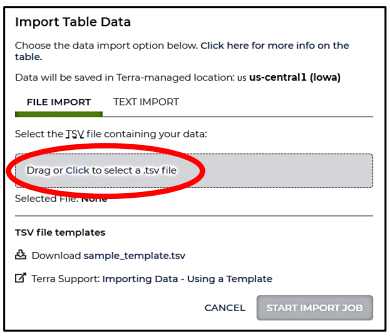


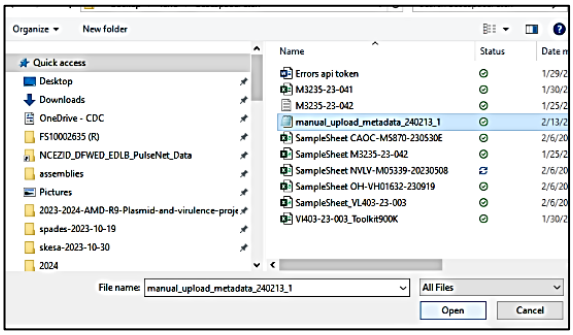
5. Импортируйте файл метаданных tsv:
   a. На вкладке "Данные" нажмите на "Импорт данных" и выберите "Загрузить tsv" из выпадающего меню.



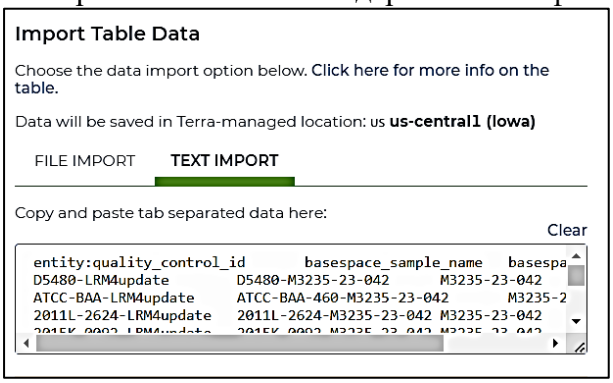
b. Во всплывающем окне "Импорт данных таблицы" вы можете импортировать метаданные двумя способами:

i. На вкладке "Импорт файлов" щелкните в центре, чтобы выбрать файл tsv, перейдите к месту, где сохранен файл tsv с метаданными, выберите файл и нажмите "Открыть".

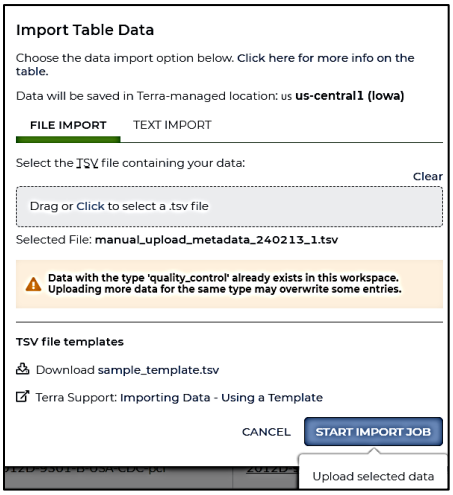




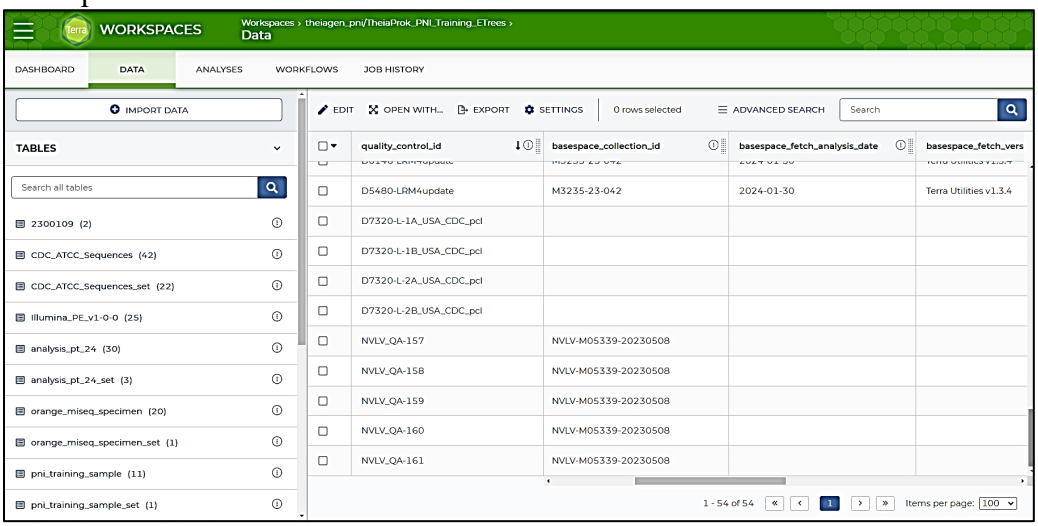
ii. Также вы можете переключиться на вкладку "Импорт текста" и скопировать и вставить содержимое tsv-файла в поле посередине.



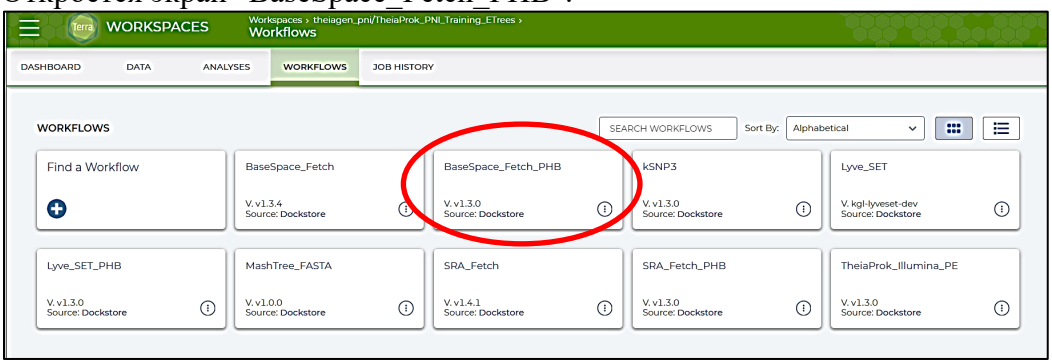
c. Во всплывающем окне "Импорт данных таблицы" вы получите предупреждение о том, что в таблице данных уже существуют данные (при импорте в существующую таблицу данных) и загрузка в нее дополнительных данных может привести к перезаписи существующих данных. Нажмите кнопку "Начать задание импорта".



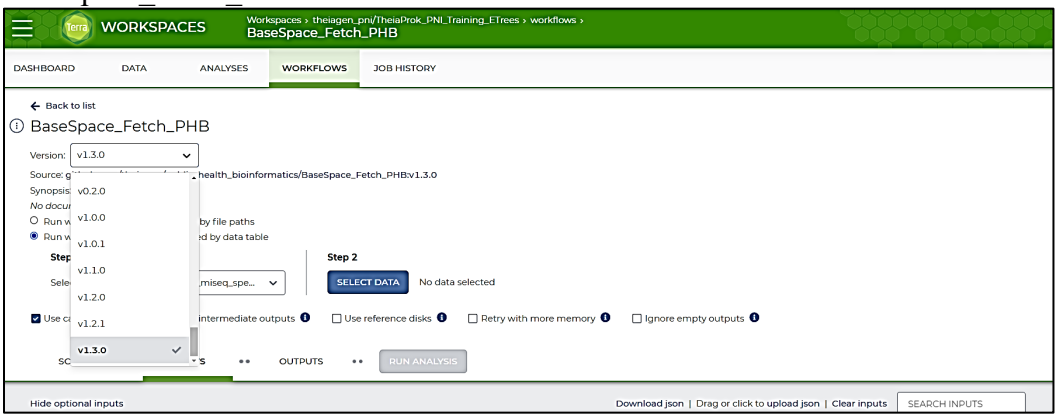
d. После завершения импорта вы должны увидеть новые записи, созданные в таблице данных для последовательностей, которые будут импортированы из BaseSpace.



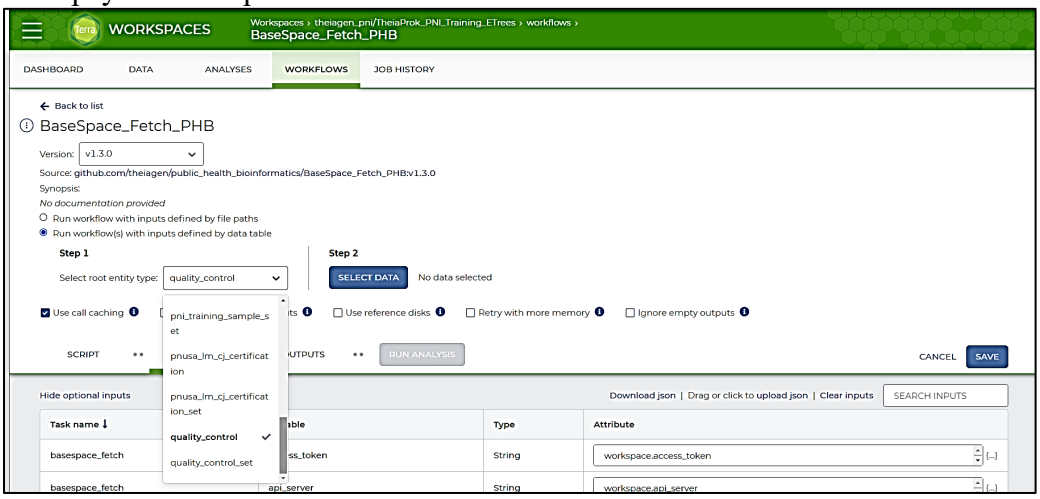
6. Запустите рабочий процесс "BaseSpace_Fetch_PHB":
   a. На вкладке "Рабочие процессы" нажмите на "BaseSpace_Fetch_PHB". Откроется экран "BaseSpace_Fetch_PHB".



b. В раскрывающемся меню "Версия" выберите последнюю версию BaseSpace_Fetch_PHB.

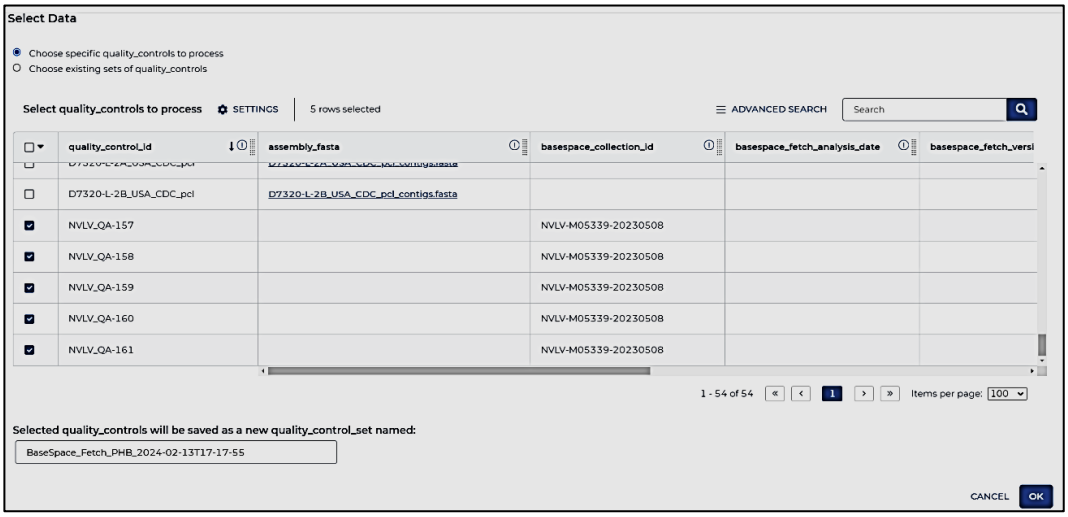


c. В разделе "Шаг 1" нажмите на раскрывающееся меню "Выбрать тип корневой сущности" и выберите таблицу данных, в которую вы загрузили (шаги 4-5) файл tsv, содержащий метаданные для прогона, подлежащего импорту из BaseSpace.

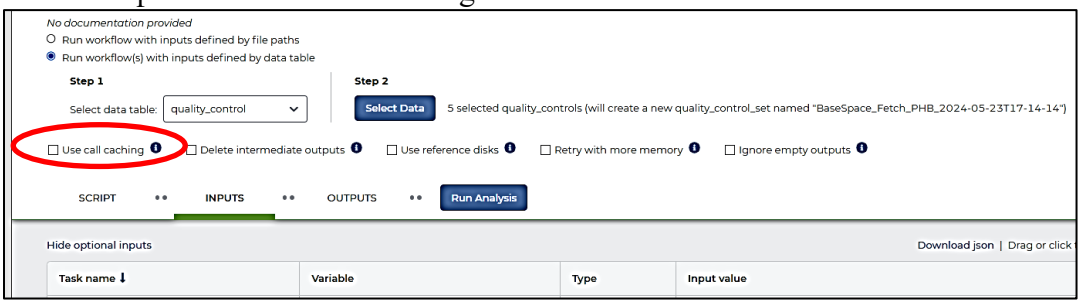


d. В разделе "Шаг 2" нажмите "Выбрать данные" (скриншот выше). Это приведет вас к экрану выбора образцов.

e. Установите флажки рядом с образцами, которые будут импортированы из BaseSpace, и нажмите "ОК". В результате вы вернетесь на экран "BaseSpace_Fetch_PHB".

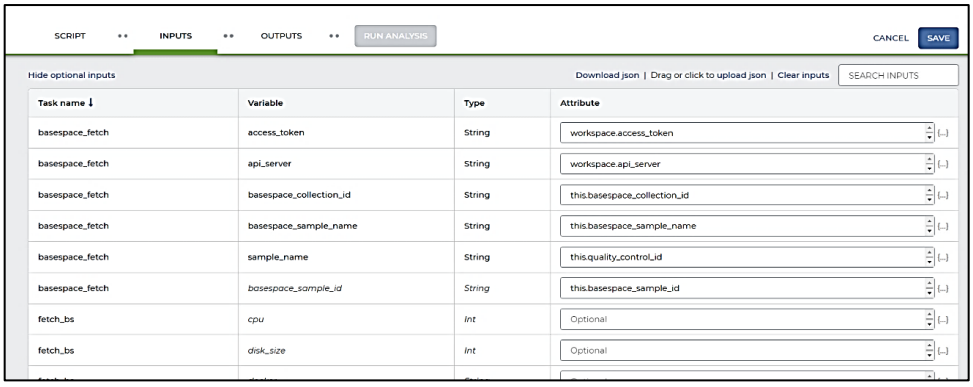


f. Снимите флажок " Use call caching ".

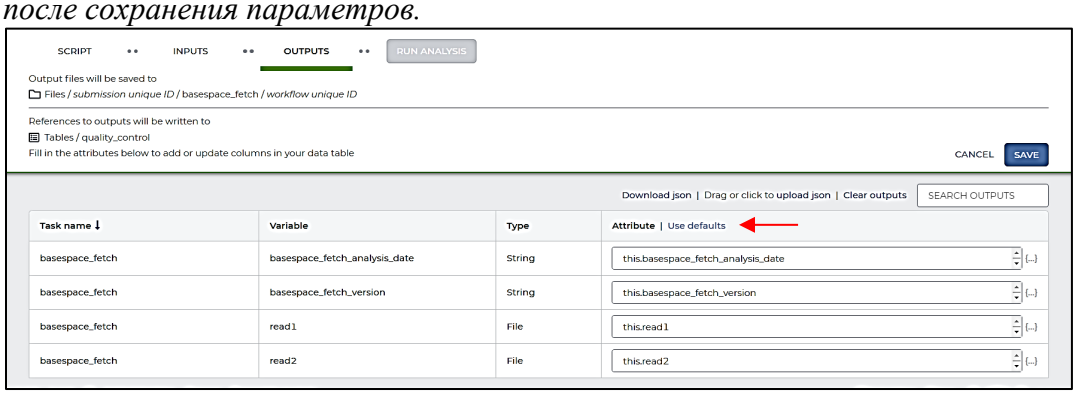


g. На вкладке "**Inputs**" определите следующие переменные в поле "Attribute":
ПРИМЕЧАНИЕ: *При заполнении столбца "Атрибут" щелчок внутри ячейки вызовет выпадающее меню атрибутов, которые вы можете выбрать, чтобы избежать опечаток*

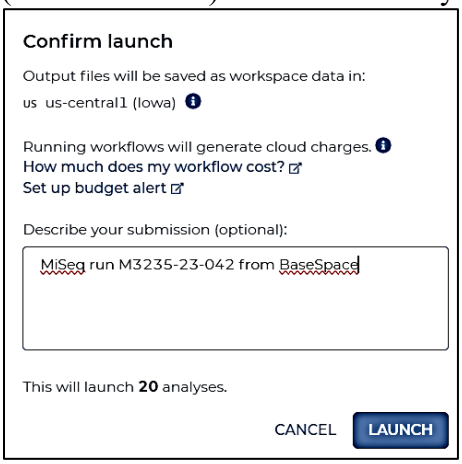
i. access_token: "workspace.access_token".
ii. api_server: "workspace.api_server".
iii. basespace_collection_id: "this.basespace_collection_id".
iv. basepace_sample_name: "this.basespace_sample_name".
v. sample_name: "**this.datatablename_id**", например, "this.quality_control_id".
vi. basespace_sample_id: "this.basespace_sample_id".
ПРИМЕЧАНИЕ: *необходимо заполнять **только** в том случае, если содержимое полей "Sample_Name" и "Sample_ID" отличается в таблице образцов прогона или вы импортируете данные из прогона NextSeq.*



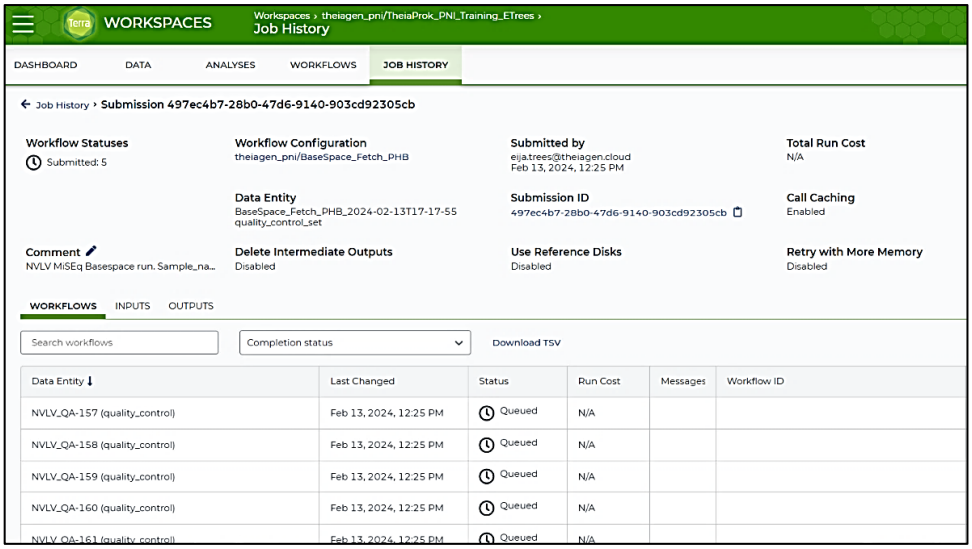
h. На вкладке "**Выходные данные**" нажмите "Использовать значения по умолчанию", затем нажмите "Сохранить", а затем "Запустить анализ". **ПРИМЕЧАНИЕ:** *Кнопка "Сохранить" видна только в том случае, если параметры (кроме идентификаторов образцов) изменились по сравнению с предыдущим заданием. Кнопка "Запустить анализ" становится выделенной после сохранения параметров.*



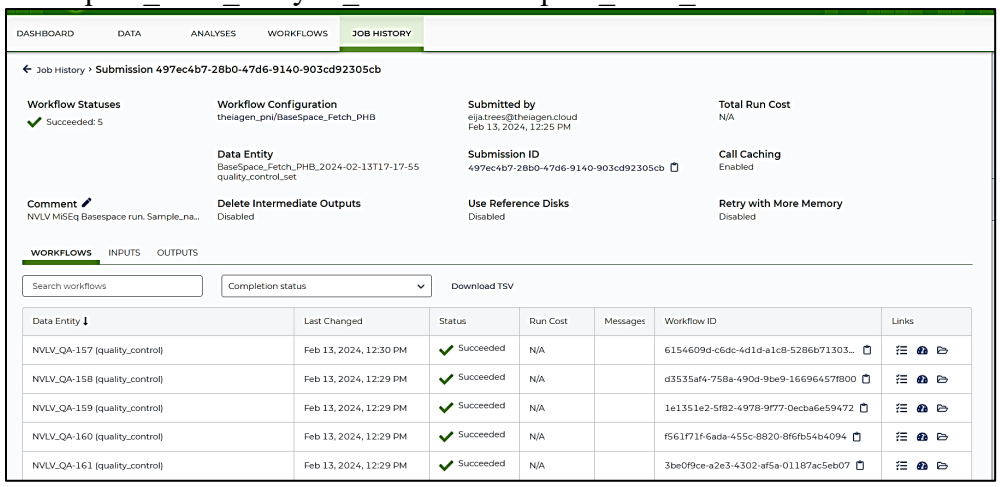
i. Во всплывающем окне "Подтверждение запуска" опишите задание (необязательно) и нажмите "Запустить".

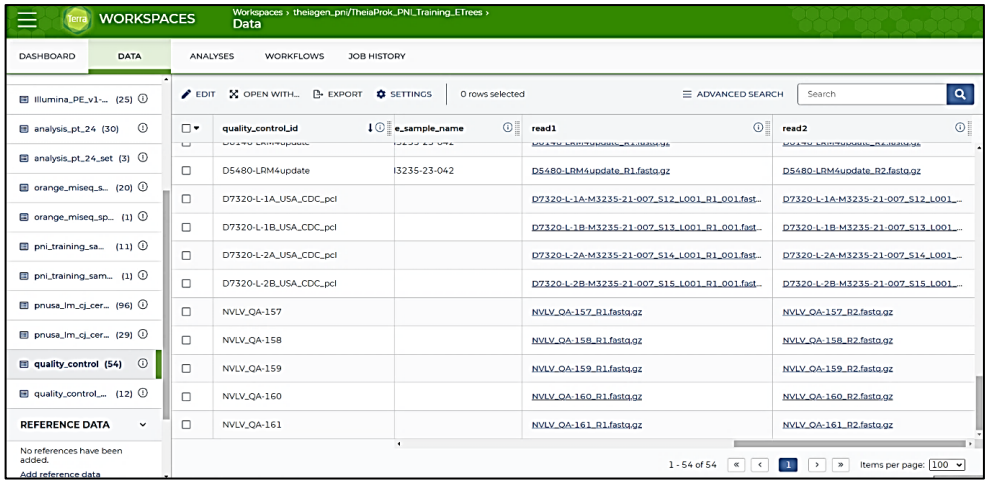


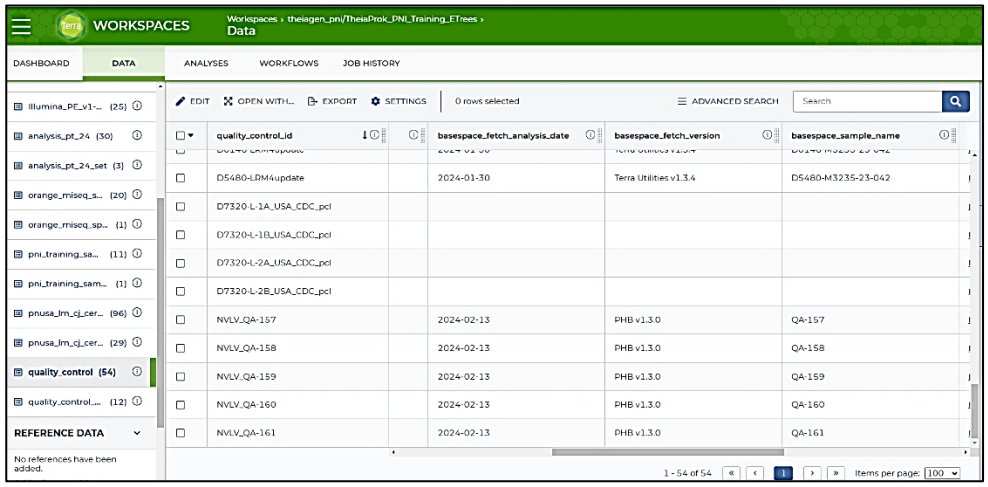
j. Откроется вкладка "История заданий", где статус отправленных заданий изначально должен быть "В очереди".



k. После завершения задания статус будет "Успешно" или "Выполнено". На вкладке "Данные" теперь должны отображаться имена файлов FASTQ в столбцах "Read1" и "Read2", а также информация в столбцах "basespace_fetch_analysis_date" и "basespace_fetch_version".

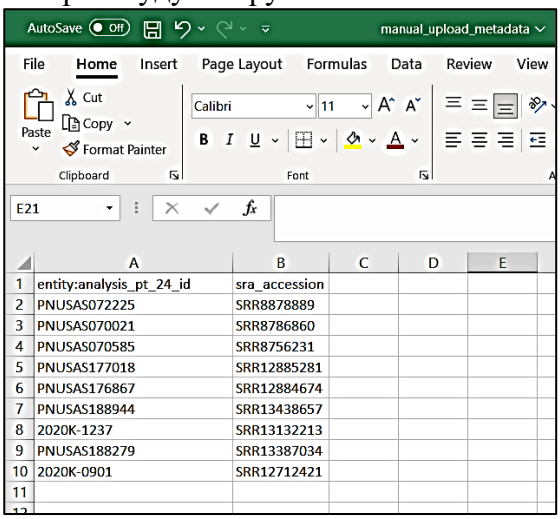



WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
quality_control_id
e_sample_name
read1
read2
D5480-LRM4update
D5480-LRM4update_R1.fastq.gz
D5480-LRM4update_R2.fastq.gz
D7320-L-1A_USA_CDC_pcl
D7320-L-1B_USA_CDC_pcl
D7320-L-2A_USA_CDC_pcl
D7320-L-2B_USA_CDC_pcl
NVLV_QA-157
NVLV_QA-158
NVLV_QA-159
NVLV_QA-160
NVLV_QA-161
REFERENCE DATA
No references have been added.
Add reference data
1 - 54 of 54
Items per page: 100


WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
quality_control_id
basespace_fetch_analysis_date
basespace_fetch_version
basespace_sample_name
D5480-LRM4update
2024-01-30
Terra Utilities v1.3.4
D5480-M3235-23-042
D7320-L-1A_USA_CDC_pcl
D7320-L-1B_USA_CDC_pcl
D7320-L-2A_USA_CDC_pcl
D7320-L-2B_USA_CDC_pcl
NVLV_QA-157
2024-02-13
PHB v1.3.0
QA-157
NVLV_QA-158
QA-158
NVLV_QA-159
QA-159
NVLV_QA-160
QA-160
NVLV_QA-161
QA-161
REFERENCE DATA
No references have been added.
1 - 54 of 54
Items per page: 100

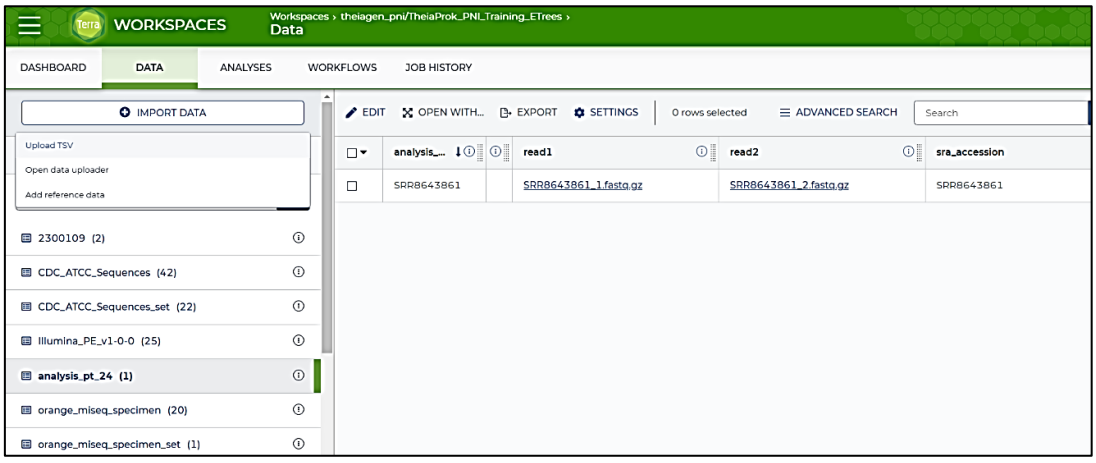
Приложение PNID01-2: Загрузка данных из NCBI SRA

1. Подготовьте файл метаданных tsv:
    a. Введите имя таблицы данных (новой или существующей) в ячейку A1 между "entity:" и "id".
    b. Введите идентификаторы образцов в столбец A так, как вы хотите, чтобы они отображались в таблице данных Terra.
    c. sra_accession: введите номера присоединения SRA для последовательностей, которые будут загружены из SRA.

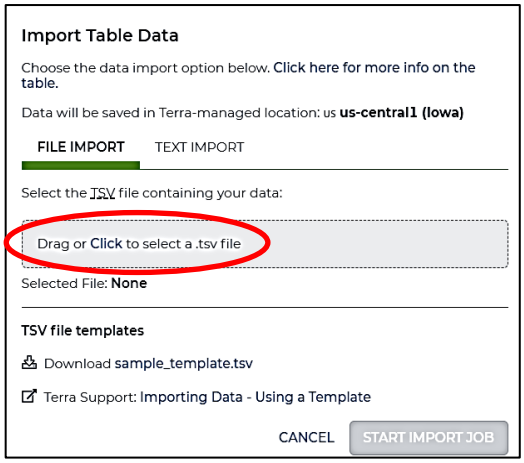


	A	B	C	D	E
1	entity:analysis_pt_24_id	sra_accession			
2	PNUSAS072225	SRR8878889			
3	PNUSAS070021	SRR8786860			
4	PNUSAS070585	SRR8756231			
5	PNUSAS177018	SRR12885281			
6	PNUSAS176867	SRR12884674			
7	PNUSAS188944	SRR13438657			
8	2020K-1237	SRR13132213			
9	PNUSAS188279	SRR13387034			
10	2020K-0901	SRR12712421			
11					

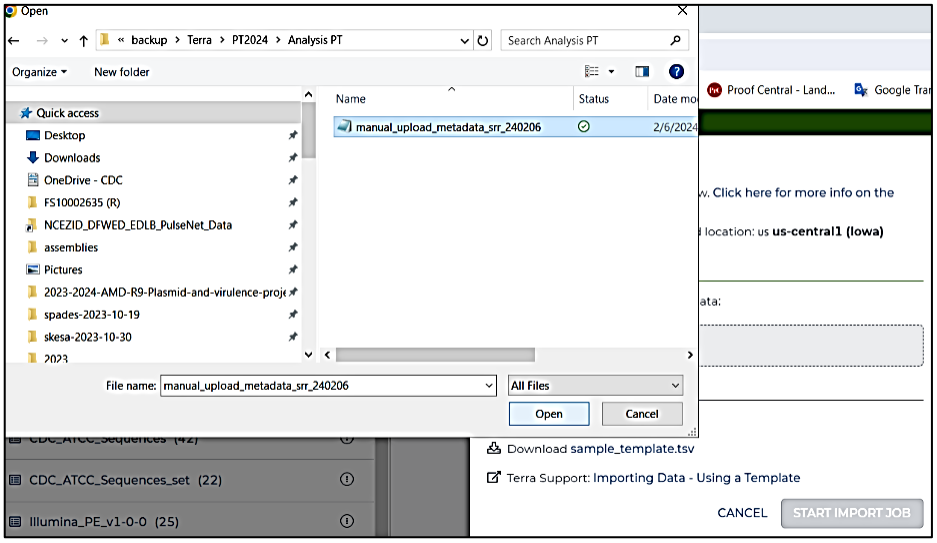
    d. Сохраните файл в **формате tsv**: выберите "Сохранить как" и "Текст (с разделителем табуляции) (*.txt)". Убедитесь, что имя файла имеет окончание "**.tsv**".
2. На вкладке "Данные" нажмите "Импорт данных" и выберите "Загрузить tsv" из выпадающего меню.



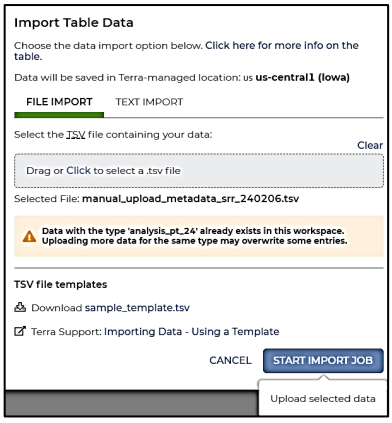
3. Во всплывающем окне "Импорт данных таблицы" на вкладке "Импорт файлов" щелкните в центре, чтобы выбрать файл tsv.



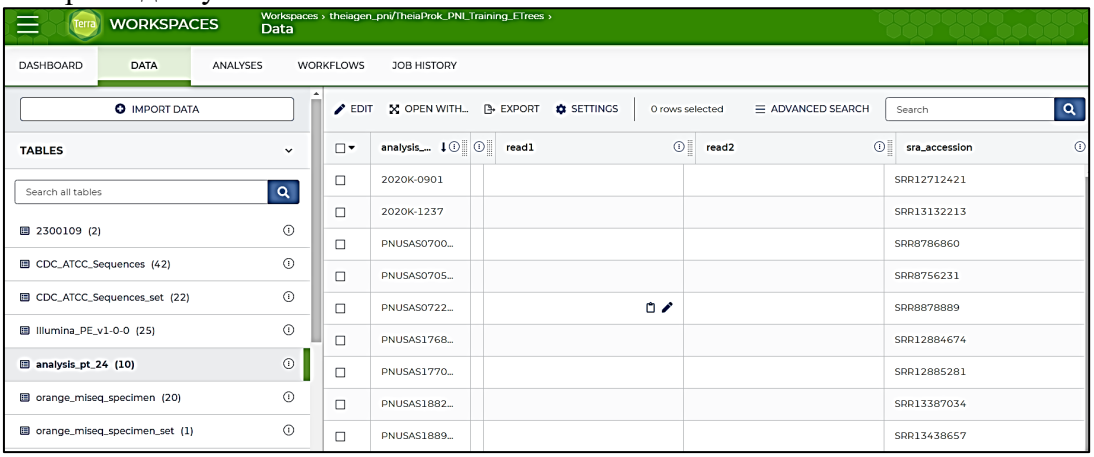
4. Перейдите в место, где сохранен файл метаданных tsv, выберите файл и нажмите "Открыть".



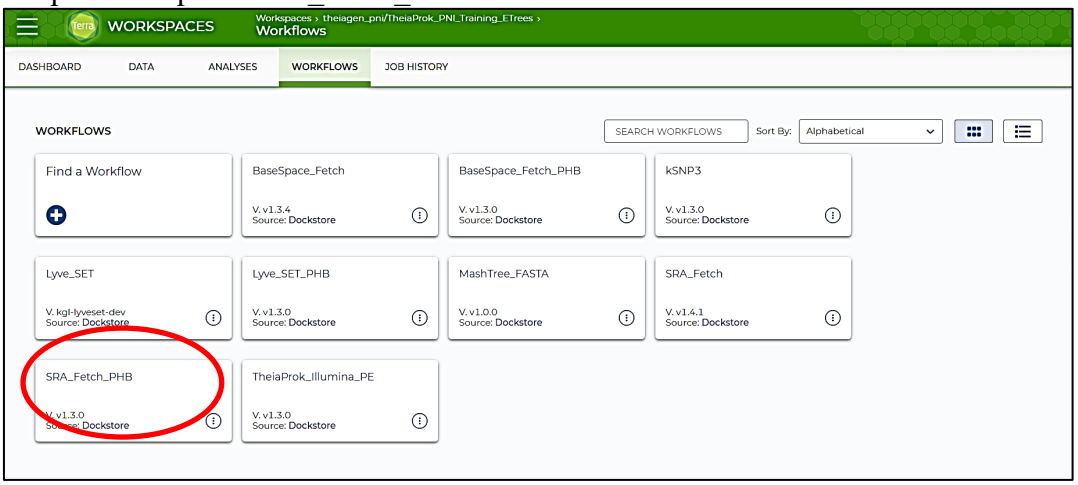
5. Во всплывающем окне "Импорт данных таблицы" появится предупреждение о том, что данные уже существуют в данной таблице данных (если импорт выполняется в существующую таблицу данных) и загрузка в нее дополнительных данных может привести к перезаписи существующих данных. Нажмите кнопку "Начать задание импорта".



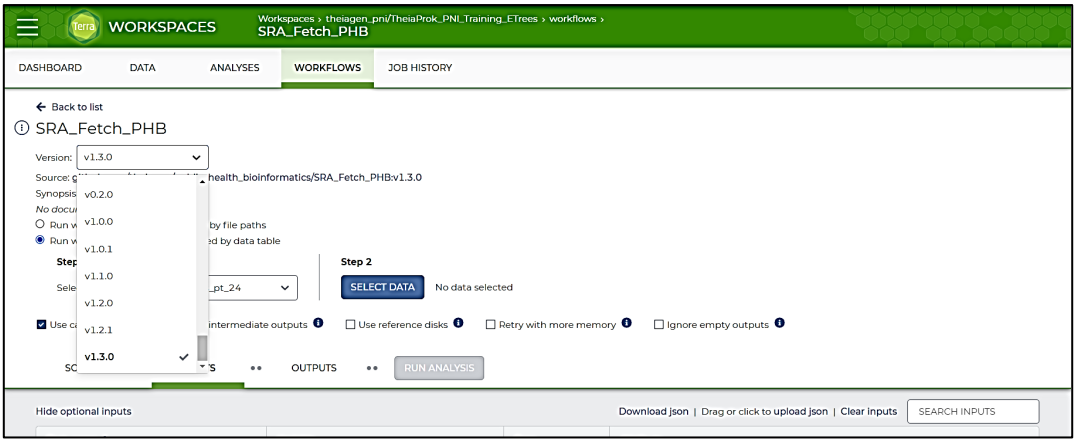
6. После завершения импорта вы должны увидеть новые записи, созданные в таблице данных для последовательностей, которые будут загружены из NCBI, вместе с их номерами доступа SRA.



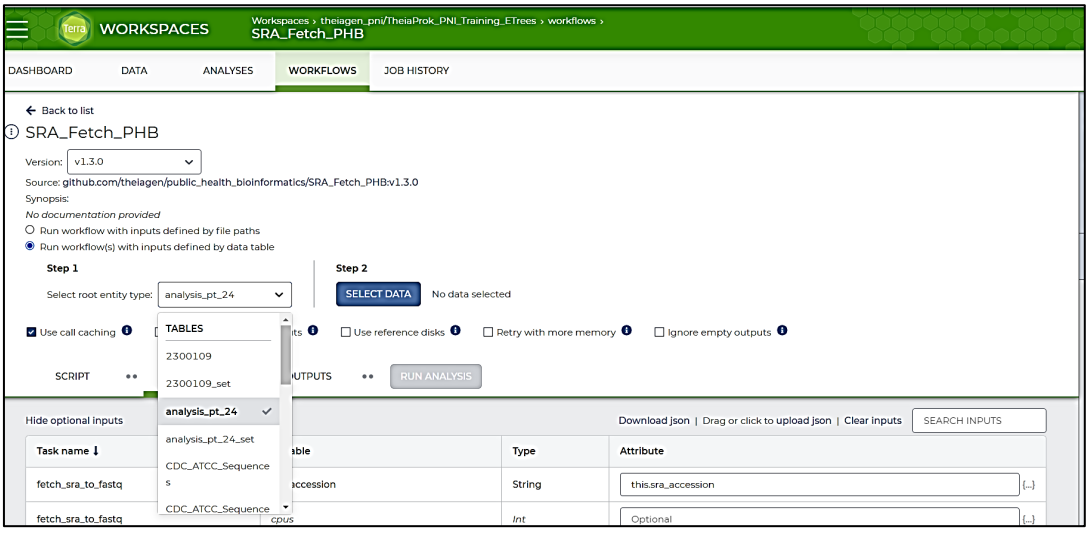
7. На вкладке "Рабочие процессы" нажмите на рабочий процесс "SRA_Fetch_PHB". Откроется экран "SRA_Fetch_PHB".



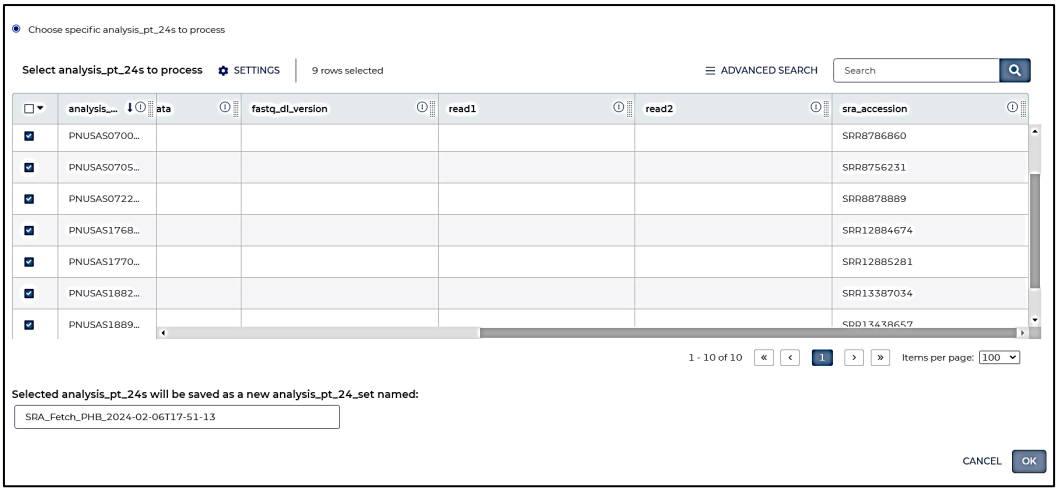
8. В раскрывающемся меню "Версия" выберите последнюю версию SRA_Fetch_PHB.



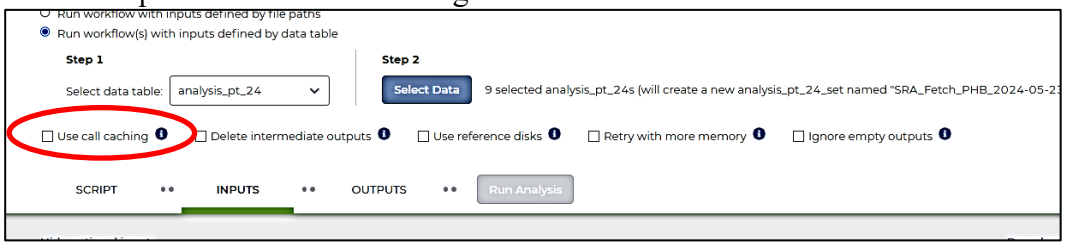
9. В разделе "Шаг 1" нажмите на раскрывающееся меню "Выбрать тип корневой сущности" и выберите таблицу данных, в которую вы импортировали (шаги 1-6) файл tsv с номерами вступлений SRA для образцов, которые необходимо загрузить из SRA.



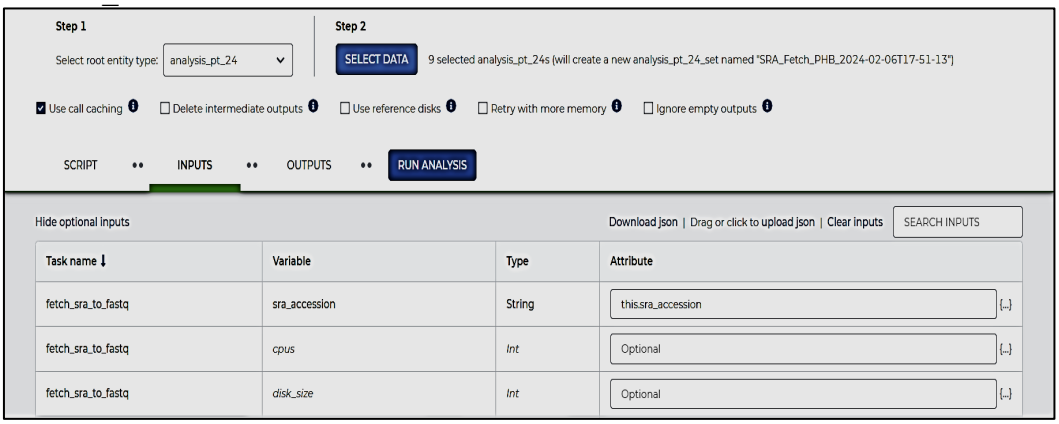
10. В разделе "Шаг 2" нажмите "Выбрать данные" (скриншот выше). Откроется экран выбора образцов.
11. Установите флажки рядом с образцами, которые будут загружены из NCBI, и нажмите "ОК". В результате вы вернетесь на экран "SRA_Fetch_PHB".



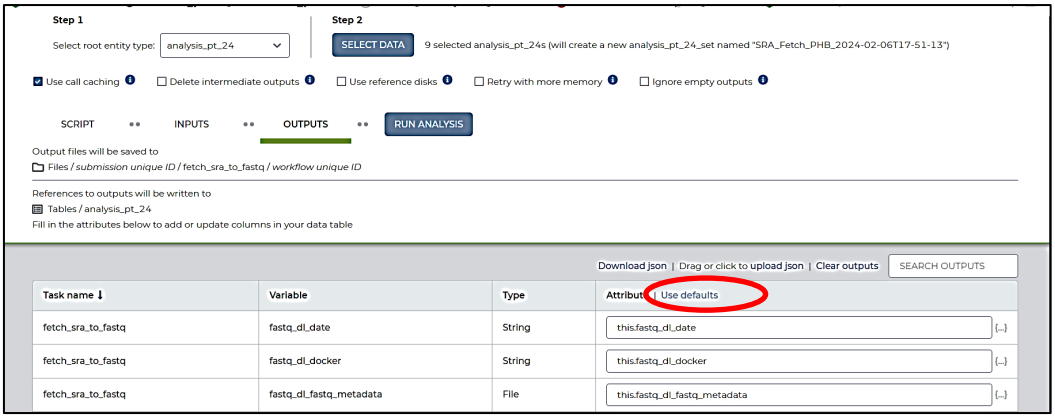
12. Снимите флажок "Use call caching".



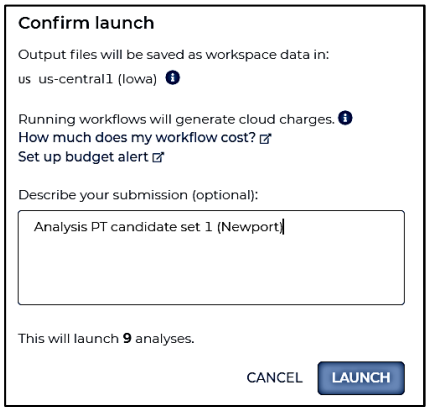
13. На вкладке "**Inputs**" определите переменную "sra_accession" в поле "Attribute": "this.sra_accession".



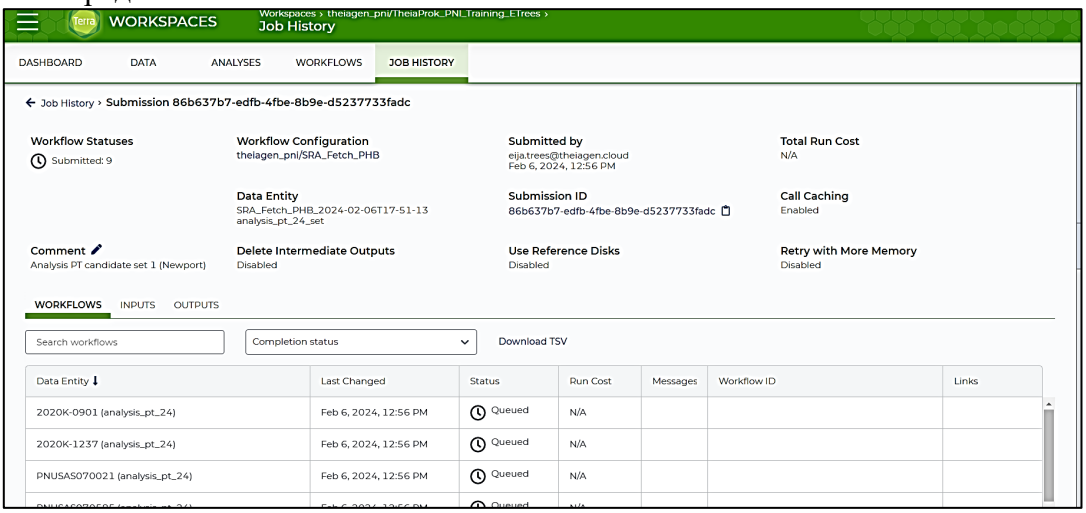
14. На вкладке "**Выходные данные**" нажмите "Использовать значения по умолчанию", затем нажмите "Сохранить", а затем "Запустить анализ". **ПРИМЕЧАНИЕ:** *Кнопка "Сохранить" видна только в том случае, если параметры (кроме идентификаторов образцов) изменились по сравнению с предыдущей отправкой задания.*



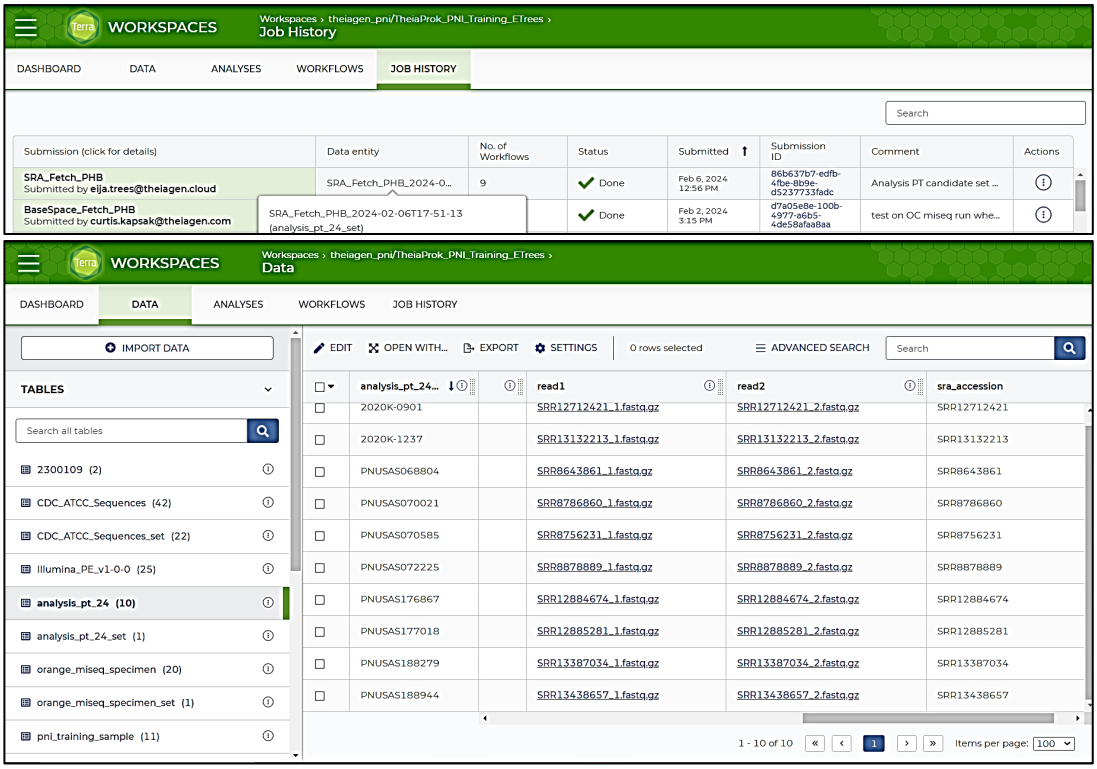
15. Во всплывающем окне "Подтверждение запуска" опишите задание (необязательно) и нажмите "Запустить".



16. Откроется вкладка "История заданий", где статус заданий изначально должен быть "В очереди".



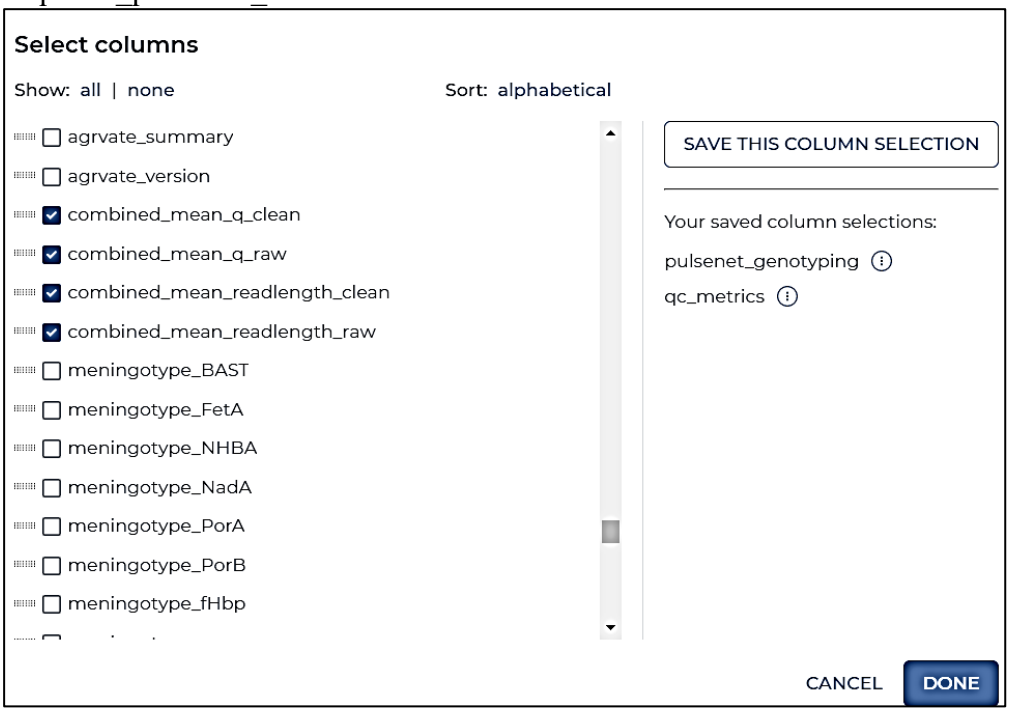
17. После завершения работы статус станет "Выполнено". На вкладке "Данные" в столбцах "Read1" и "Read2" теперь должны отображаться имена файлов FASTQ.


WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Job History
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
Search
Submission (click for details)
Data entity
No. of Workflows
Status
Submitted
Submission ID
Comment
Actions
SRA_Fetch_PHB
Submitted by eija.trees@theiagen.cloud
SRA_Fetch_PHB_2024-0...
9
Done
Feb 6, 2024 12:56 PM
86b637b7-edfb-4fbe-8b9e-d5237733fadc
Analysis PT candidate set ...
BaseSpace_Fetch_PHB
Submitted by curtis.kapsak@theiagen.com
SRA_Fetch_PHB_2024-02-06T17-51-13 (analysis_pt_24_set)
Done
Feb 2, 2024 3:15 PM
d7a05a8e-100b-4977-a6b5-4de58afaa8aa
test on OC miseq run whe...
WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
IMPORT DATA
TABLES
Search all tables
2300109 (2)
CDC_ATCC_Sequences (42)
CDC_ATCC_Sequences_set (22)
illumina_PE_v1-0-0 (25)
analysis_pt_24 (10)
analysis_pt_24_set (1)
orange_miseq_specimen (20)
orange_miseq_specimen_set (1)
pni_training_sample (11)
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
analysis_pt_24...
read1
read2
sra_accession
2020K-0901
SRR12712421_1.fastq.gz
SRR12712421_2.fastq.gz
SRR12712421
2020K-1237
SRR13132213_1.fastq.gz
SRR13132213_2.fastq.gz
SRR13132213
PNUSAS068804
SRR8643861_1.fastq.gz
SRR8643861_2.fastq.gz
SRR8643861
PNUSAS070021
SRR8786860_1.fastq.gz
SRR8786860_2.fastq.gz
SRR8786860
PNUSAS070585
SRR8756231_1.fastq.gz
SRR8756231_2.fastq.gz
SRR8756231
PNUSAS072225
SRR8878889_1.fastq.gz
SRR8878889_2.fastq.gz
SRR8878889
PNUSAS176867
SRR12884674_1.fastq.gz
SRR12884674_2.fastq.gz
SRR12884674
PNUSAS177018
SRR12885281_1.fastq.gz
SRR12885281_2.fastq.gz
SRR12885281
PNUSAS188279
SRR13387034_1.fastq.gz
SRR13387034_2.fastq.gz
SRR13387034
PNUSAS188944
SRR13438657_1.fastq.gz
SRR13438657_2.fastq.gz
SRR13438657
1 - 10 of 10
Items per page: 100

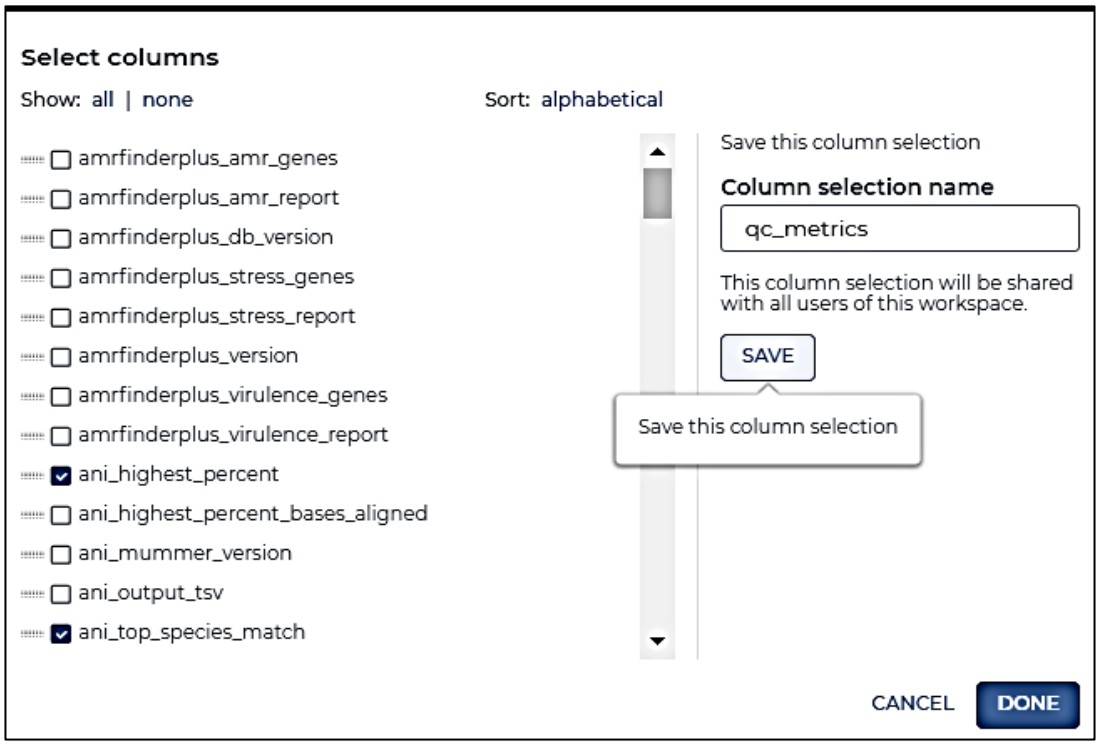
Приложение PNID01-3: Настройка представления таблицы данных для PulseNet метрики КК

ПРИМЕЧАНИЕ: *эту настройку можно выполнить только после того, как вы один раз запустили рабочий процесс TheiaProk.*

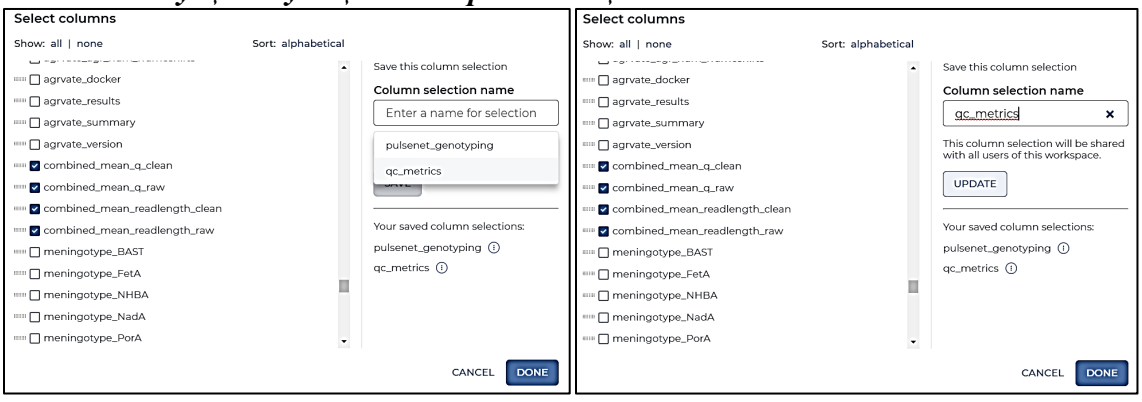
1. На вкладке "Данные" выберите интересующую вас таблицу данных, например "CDC_ATCC_Sequences", затем выберите "Настройки".
2. В разделе "Выбрать столбцы" необходимо отметить следующие метрики КК:
    a. Ani_highest_percent
    b. Ani_top_species_match
    c. Assembly_length
    d. Combined_mean_q_clean
    e. Combined_mean_q_raw
    f. Combined_mean_readlength_clean
    g. Combined_mean_readlength_raw
    h. Est_coverage_clean
    i. Est_coverage_raw
    j. Gambit_predicted_taxon
    k. Midas_secondary_genus
    l. Midas_secondary_genus_abundance
    m. N50_value
    n. Number_contigs
    o. Raw_read_screen
    p. Seqsero2_predicted_contamination



3. Нажмите "Save this column selection", вызовите выбор столбцов "qc_metrics" и нажмите "Save", а затем нажмите "Done".
   ПРИМЕЧАНИЕ: *Если вы добавляете или удаляете столбцы из существующего выбора столбцов, нажмите "Сохранить этот выбор столбцов", выберите имя из выпадающего меню и нажмите "Обновить".*
   Создание нового выбора столбцов



Изменение существующего выбора столбцов



Приложение PNID01-4a. Критические показатели качества PulseNet (Pass/Fail) для рутинных последовательностей

Организм	Среднее покрытие denovo[1]	Среднее качество (Q score)[2]	Длина сборки (МБ)	Количество вторичных видов
Listeria monocytogenes	≥ 20x	≥ 30	2.8-3.2	≤ 0.01
E. coli **(большинство серотипов)**	≥ 40x	≥ 30	4.9-6.0	≤ 0.01
Shigella spp.**/редкие *E. coli***	≥ 40x	≥ 30	4.2-4.9	≤ 0.01
Salmonella spp.	≥ 30x	≥ 30	4.4-5.7	≤ 0.01
Campylobacter spp.	≥ 20x	≥ 30	1.4-2.2	≤ 0.01
Vibrio cholerae	≥ 40x	≥ 30	3.8-4.3	≤ 0.01
Vibrio parahaemolyticus	≥ 40x	≥ 30	4.9-5.5	≤ 0.01
Vibrio vulnificus	≥ 40x	≥ 30	4.7-5.3	≤ 0.01

[1]После обрезки на основе качества (est_coverage_clean)
[2]До обрезки (combined_mean_q_raw)

Приложение PNID01-4b. Шаг предварительной сортировки чтений TheiaProk для исключения последовательностей низкого качества с целью экономии вычислительных ресурсов

Задача screen гарантирует, что количество данных о последовательностях достаточно для проведения геномного анализа. Она использует команды bash для количественной оценки чтений и пар оснований, а также mash-скетчинг для оценки размера генома и его покрытия. На каждом этапе результаты оцениваются относительно критериев "прошел/не прошел" и пороговых значений, которые могут быть определены с помощью дополнительных пользовательских данных.

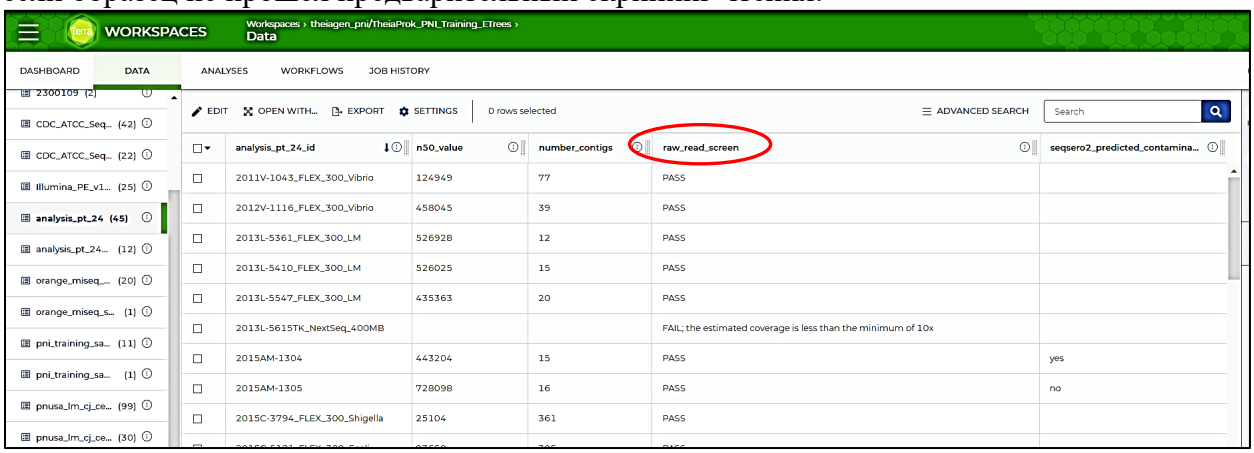
Образцы, не соответствующие этим критериям, не будут обрабатываться дальше:

1. Общее количество чтений: Образец проваливает задачу отбора чтений, если общее количество чтений меньше или равно min_reads.
2. Доля чтений по парам оснований в файлах прямых и обратных чтений: Образец провалит скрининг чтений, если в файлах reads1 или read2 содержится меньше min_proportion basepairs.
3. Количество пар оснований: Образец не пройдет проверку на чтение, если в нем меньше min_basepairs basepairs
4. Предполагаемый размер генома: Образец не пройдет проверку на чтение, если предполагаемый размер генома меньше min_genome_size или больше max_genome_size.
5. Предполагаемое покрытие генома: Образец не пройдет скрининг чтения, если предполагаемое покрытие генома меньше min_coverage.

Значения по умолчанию:
Int min_reads = 7472
Int min_basepairs = 2241820

Int min_genome_length = 100000
Int max_genome_length = 18040666
Int min_coverage = 10
Int min_proportion = 40

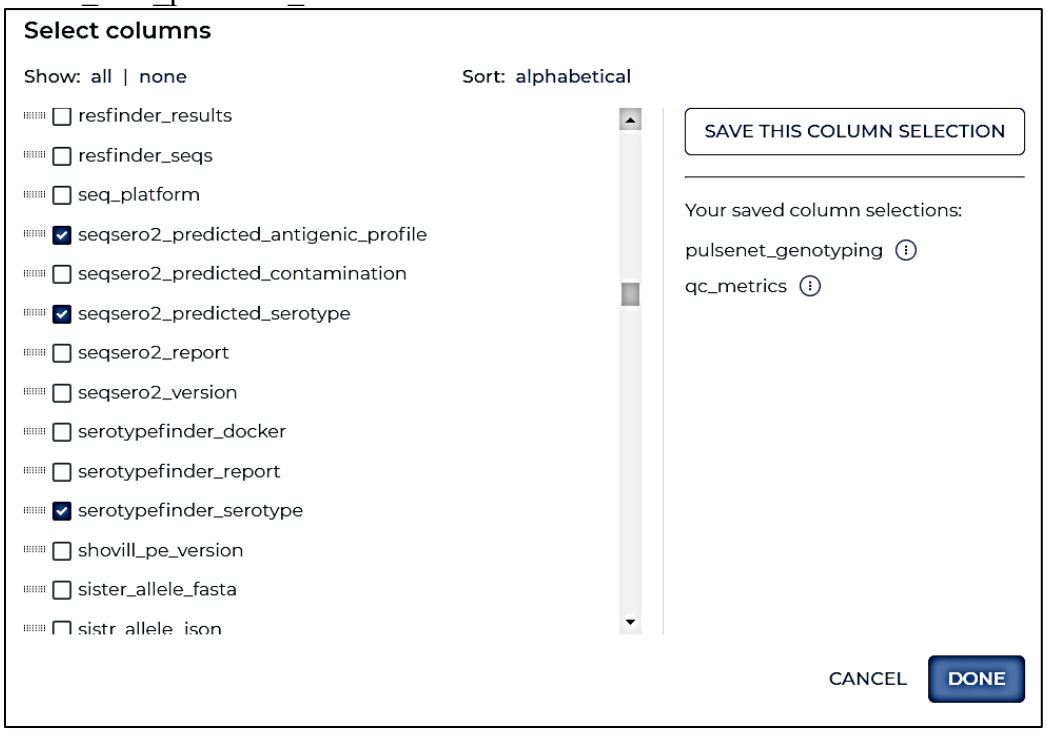
Колонка "raw_read_screen" в разделе "QC_metrics" содержит подробную информацию, если образец не прошел предварительный скрининг чтения:



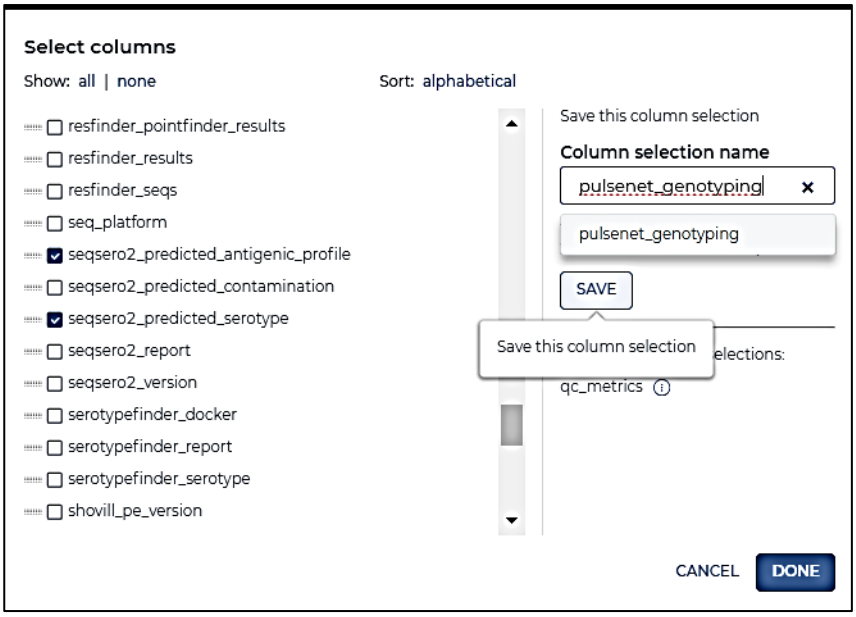
Приложение PNID01-5. Настройка представления таблицы данных для анализов PulseNet Genotyping Assays

ПРИМЕЧАНИЕ: *эту настройку можно выполнить только после того, как вы один раз запустили рабочий процесс TheiaProk.*

1. На вкладке "Данные" выберите интересующую вас таблицу данных, например "CDC_ATCC_Sequences", затем выберите "Настройки".
2. В разделе "Выбрать столбцы" необходимо отметить следующие анализы для генотипирования:
    a. Amrfinderplus_amr_classes
    b. Amrfinderplus_amr_core_genes
    c. Amrfinderplus_amr_subclasses
    d. Amrfinderplus_virulence_genes
    e. Plasmidfinder_plasmids
    f. Seqsero2_predicted_antigenic profile
    g. Seqsero2_predicted_serotype
    h. Serotypefinder_serotype
    i. Ts_mlst_predicted_st

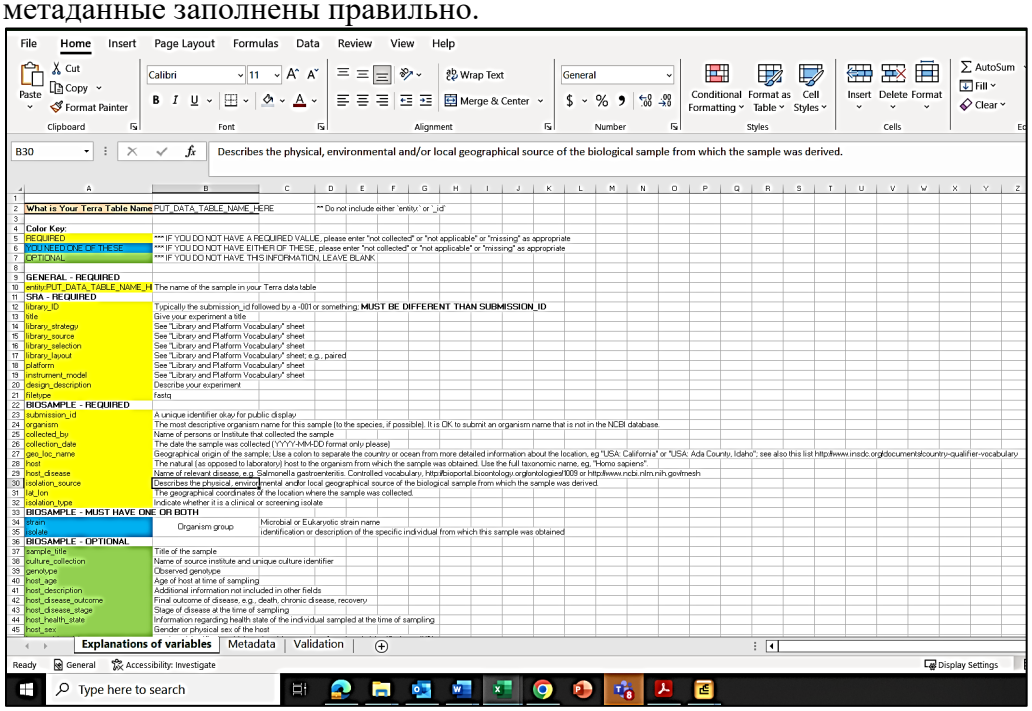


3. Нажмите "Сохранить выбор столбца", назовите выбор столбца "pulsenet_genotyping" и нажмите "Сохранить", а затем нажмите "Готово".
ПРИМЕЧАНИЕ: *Если вы добавляете или удаляете столбцы из существующего выбора столбцов, нажмите "Сохранить этот выбор столбцов", выберите имя из выпадающего меню и нажмите "Обновить".*


Select columns
Show: all | none
Sort: alphabetical
resfinder_pointfinder_results
resfinder_results
resfinder_seqs
seq_platform
seqsero2_predicted_antigenic_profile
seqsero2_predicted_contamination
seqsero2_predicted_serotype
seqsero2_report
seqsero2_version
serotypefinder_docker
serotypefinder_report
serotypefinder_serotype
shovill_pe_version
Save this column selection
Column selection name
pulsenet_genotyping
pulsenet_genotyping
SAVE
Save this column selection
elections:
qc_metrics
CANCEL
DONE

Приложение PNID01-6. Загрузка дополнительных метаданных в Terra для представления в NCBI и настройка представления таблицы данных для метаданных

- NCBI требует загрузки минимальных метаданных в NCBI для создания BioSamples для последовательностей, которые будут загружены в SRA.
- Поскольку метаданные NCBI должны быть оформлены определенным образом, используйте шаблон метаданных **Pathogen**, предоставленный компанией Theiagen, для загрузки метаданных в Terra: https://theiagen.notion.site/Terra_2_NCBI-8f014c73acc44465a3d69cf4df93adfe.
- Шаблон метаданных состоит из трех вкладок:
    - Первая вкладка, называемая "Объяснение переменных", содержит описания обязательных и необязательных полей.
    - Метаданные, которые необходимо загрузить, вводятся на второй вкладке "Метаданные".
    - На третьей вкладке "Validation" можно убедиться, что необходимые метаданные заполнены правильно.



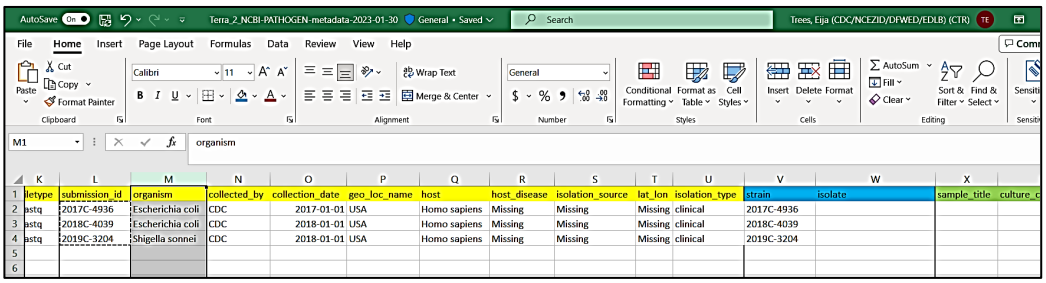
Чтобы загрузить файл метаданных патогенов в Terra:

1. Заполните обязательные и необязательные (если применимо) поля на вкладке метаданных в электронной таблице шаблона метаданных патогенов:
   ПРИМЕЧАНИЕ1: *вы можете ввести "Отсутствует" для любой требуемой информации, которой у вас нет или которую вы не хотите публично раскрывать.*
   ПРИМЕЧАНИЕ2: *приведенные ниже метаданные - это минимальные требования PulseNet USA к метаданным для загрузки в NCBI, разработанные для*

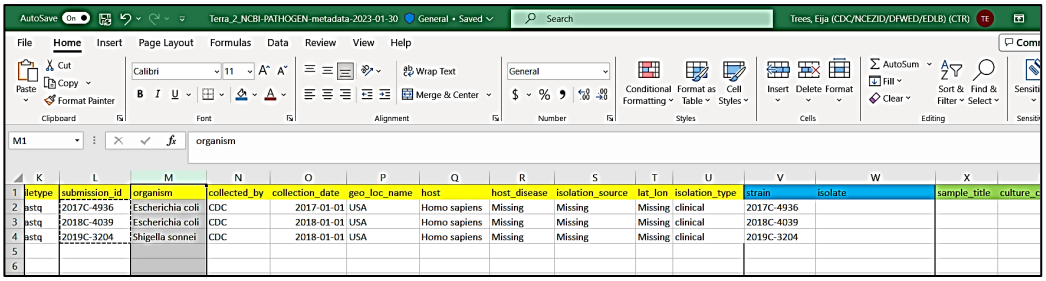
защиты конфиденциальности пациентов и целостности текущих расследований вспышек. С другой стороны, особенно в отношении неклинических образцов, предоставляется достаточно эпидемиологически полезной информации, чтобы облегчить атрибуцию.

a. В ячейке A1 введите название таблицы данных, в которую будут загружены метаданные: **entity**:**data_table_name_id**, например, entity:quality_control_id.
b. Entity: введите идентификаторы образцов, которые соответствуют идентификаторам последовательностей в Terra.
c. Submission_id: введите уникальный идентификатор образца. Это идентификатор, который будет отображаться в NCBI.
    i. Введите submission_id также в столбце "strain".
    ii. Введите "submission_id-001" в колонку "library_id".
d. Название: "PulseNet".
e. Для параметров "library_strategy", "library_source", "library_selection", "platform", "instrument_model" и "filetype" выберите правильный вариант из выпадающих меню.
f. Расположение библиотеки: "Парная".
g. Организм: род и вид.
h. Collected_by: лаборатория, предоставившая последовательность.
i. Collection_date: **год для клинических** изолятов, **год и месяц для неклинических** изолятов. Требуемый формат: ГГГГ:ММ:ДД, например, для клинического образца, выделенного в 2024 году: 2024-01-01. Для неклинического образца, выделенного в феврале 2024 года: 2024-02-01.
j. Geo_loc_name: **страна-источник для клинических** изолятов, **страна-источник и штат (или другое более подробное местоположение) для неклинических** изолятов. Используйте двоеточие для разделения страны и более подробного местоположения, например, USA:CA.
k. Isolation_source: **"Отсутствует" для клинических** изолятов, **точный источник для неклинических** изолятов, например, салат-латук, куриная грудка, мазок и т. д.
l. Isolation_type: клинический, экологический, пищевой или животный.
m. Серотип: Серотип *E. coli*.
n. Серовар: Серовар *сальмонеллы*.

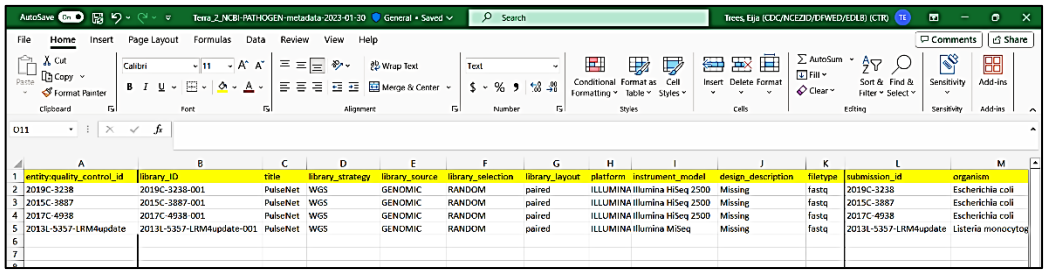
Только клинические (человеческого происхождения) последовательности

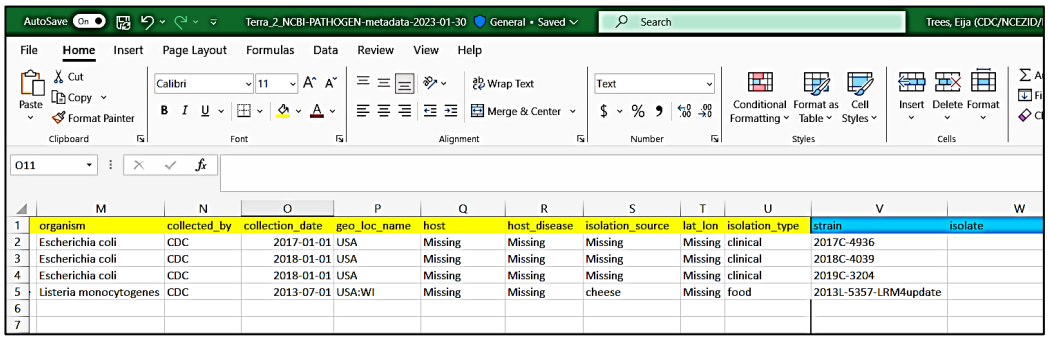


	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	letype	submission_id	organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	strain	isolate	sample_title
2	astq	2017C-4936	Escherichia coli	CDC	2017-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2017C-4936		
3	astq	2018C-4039	Escherichia coli	CDC	2018-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2018C-4039		
4	astq	2019C-3204	Shigella sonnei	CDC	2018-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2019C-3204		
5														
6														

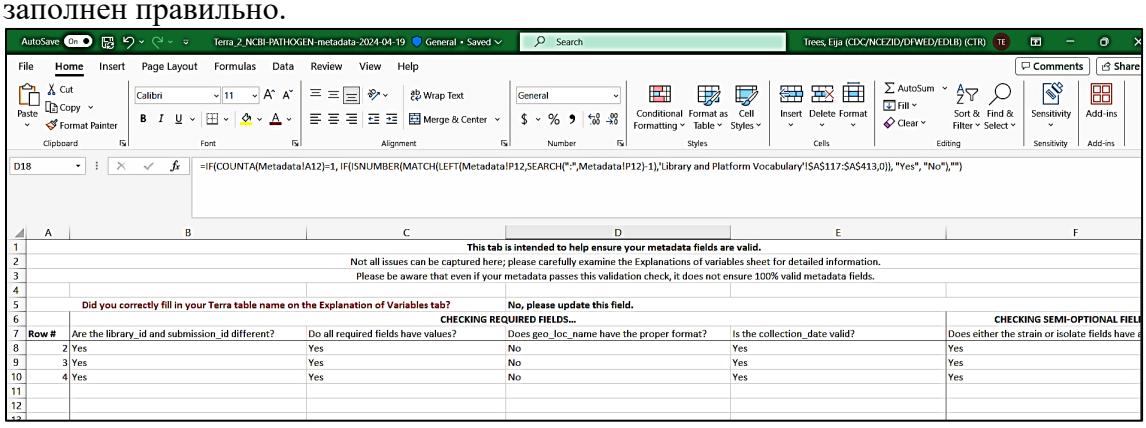


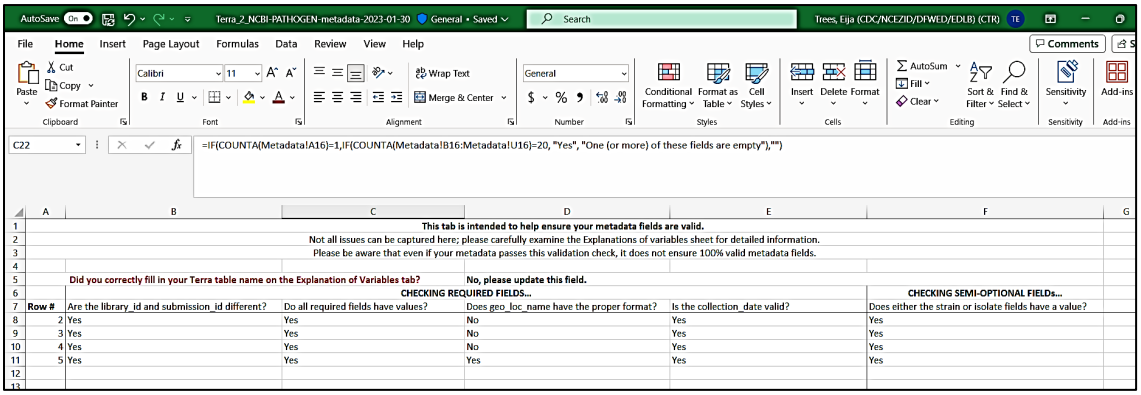
Клинические (человеческого происхождения) и неклинические последовательности





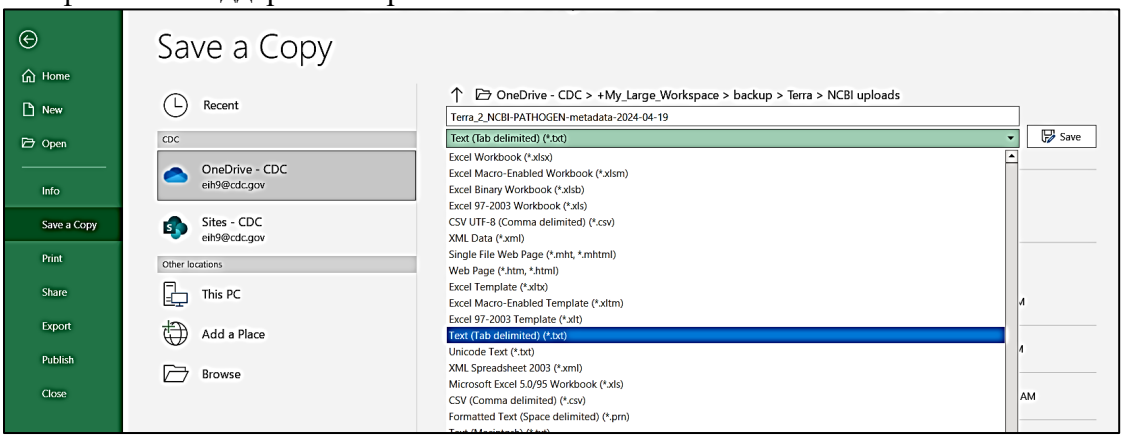
2. Перейдите на вкладку "Валидация", чтобы убедиться, что шаблон метаданных заполнен правильно.

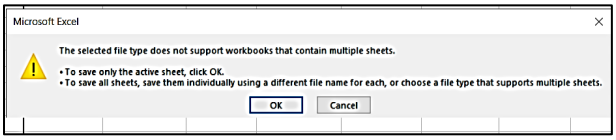




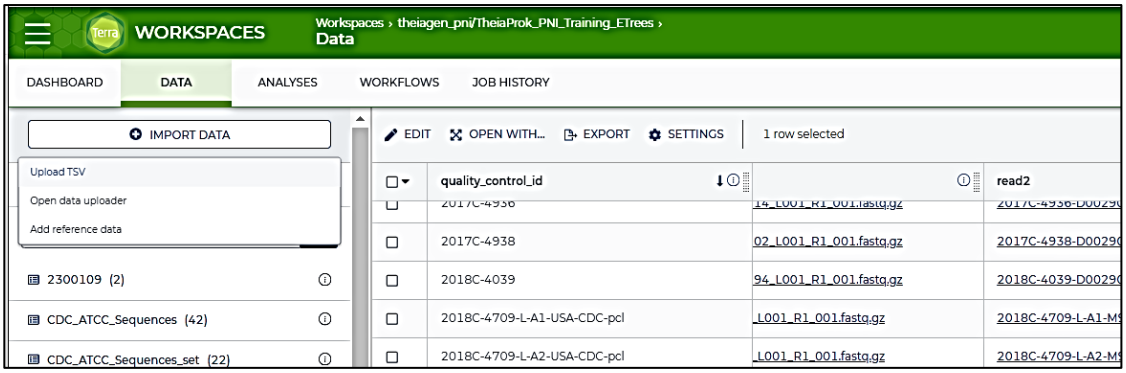
ПРИМЕЧАНИЕ: *проверка правильности формата геологического местоположения будет пройдена только в том случае, если указаны и страна-источник, и более подробное местоположение. Поэтому проверка не пройдет для клинических изолятов, для которых указана только страна-источник. Однако NCBI примет страну источника в качестве единственного местоположения.*

3. Сохраните заполненный шаблон метаданных в виде файла tsv (с разделителем табуляции). Нажмите "ОК" во всплывающем окне, сообщающем, что выбранный тип файла не поддерживает рабочие книги с несколькими листами.

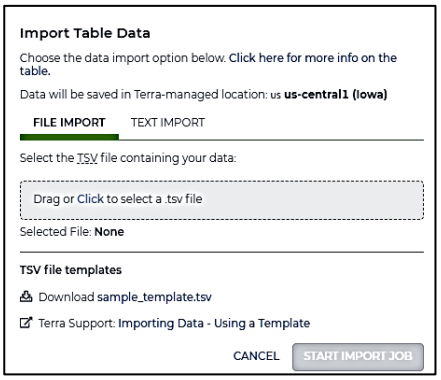




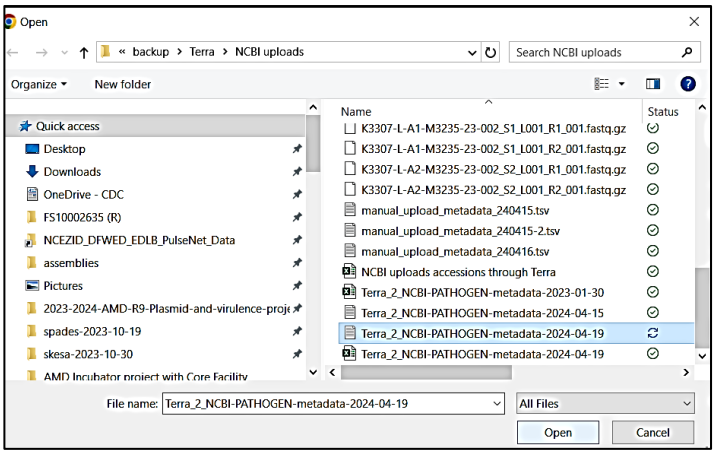
4. На вкладке "Данные" Terra Workspaces нажмите "Импорт данных" и выберите "Загрузить TSV" из выпадающего меню.



5. Во всплывающем окне "Импорт данных таблицы" на вкладке "Импорт файлов" щелкните в центре, чтобы выбрать файл tsv.

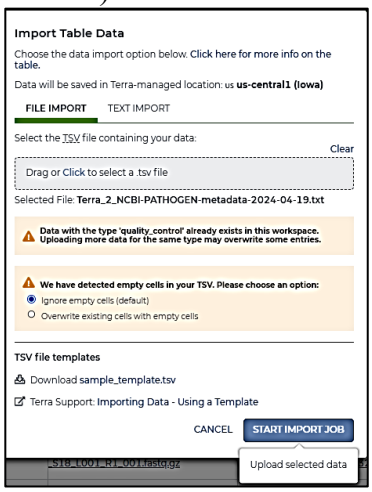


6. Перейдите в место, где сохранен файл метаданных tsv, выберите файл и нажмите "Открыть".

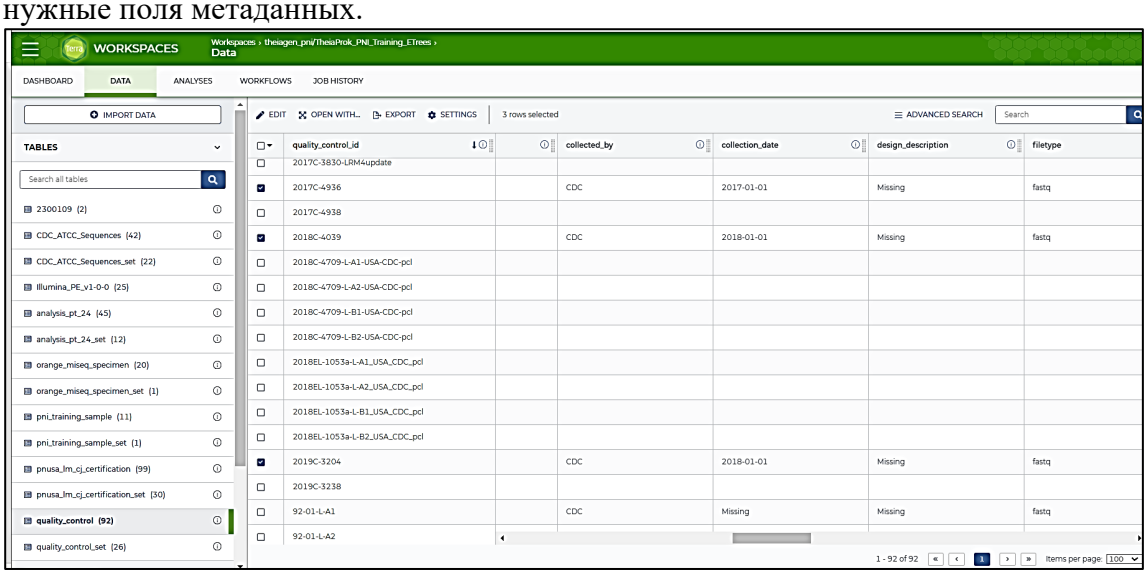


7. Во всплывающем окне "Импорт данных таблицы" появится предупреждение о том, что в таблице данных уже существуют данные и загрузка в нее дополнительных данных может привести к перезаписи существующих данных. Также будет выдано предупреждение, если в tsv-файле метаданных не для всех записей содержится

одинаковая информация (отсутствуют некоторые данные для некоторых последовательностей). Нажмите на кнопку "Начать задание импорта".



8. После завершения загрузки вы должны увидеть, что в таблице данных заполнены нужные поля метаданных.

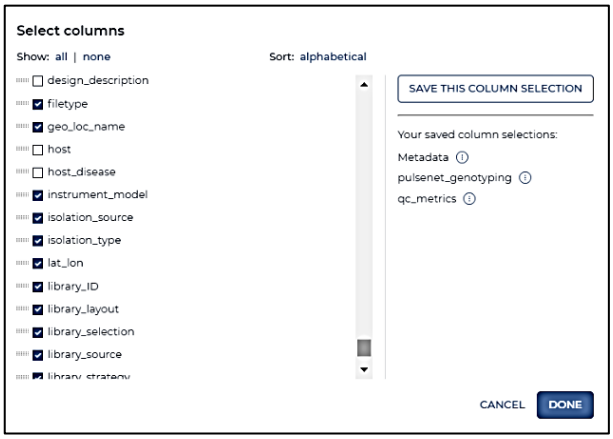


Создайте отдельное представление метаданных для таблицы данных

1. На вкладке "Данные" выберите интересующую вас таблицу данных, затем выберите "Настройки".
2. В разделе "Выбрать столбцы" отметьте все нужные столбцы метаданных.
    a. Рекомендуемые метаданные для материалов NCBI:
        i. collected_by
        ii. collection _date
        iii. filetype
        iv. geo_loc_name
        v. instrument_model
        vi. isolation_source

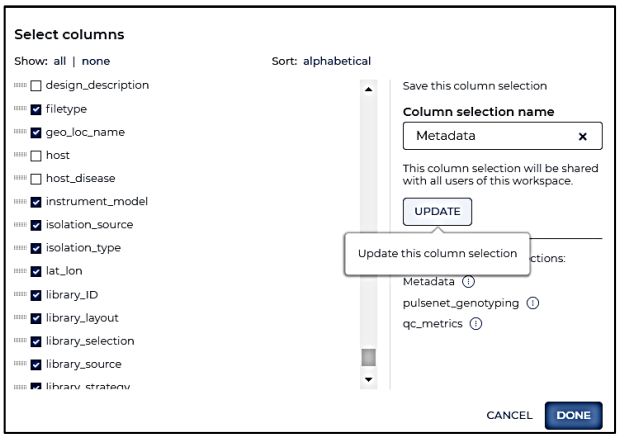
vii. isolation_type
viii. library_id
ix. library_layout
x. library_selection
xi. library_source
xii. library_strategy
xiii. organism
xiv. platform
xv. strain
xvi. submission_id
xvii. title
xviii. serotype
xix. serovar

b. Дополнительная полезная информация о последовательности
   i. read1 (имя файла R1 FASTQ)
   ii. read2 (имя файла R2 FASTQ)
   iii. assembly_fasta (местоположение сборки, созданной программой Terra)
   iv. Если данные загружаются непосредственно из Illumina BaseSpace:
      1. basespace_collection_id
      2. basespace_fetch_analysis_date
      3. basespace_fetch_version
      4. basespace_sample_id
      5. basespace_sample_name
   v. biosample_accession
   vi. sra_accession

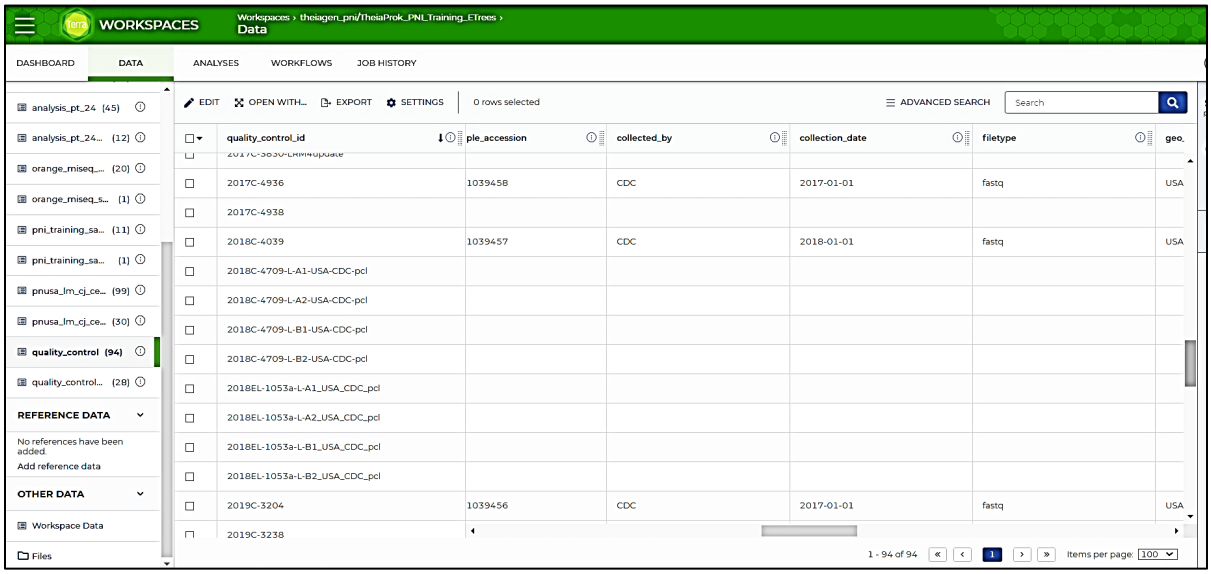


3. Нажмите кнопку "Сохранить выбор столбца".
4. Назовите выбор столбца "Метаданные" и нажмите "Сохранить" и "Готово".

ПРИМЕЧАНИЕ: *Если вы добавляете или удаляете столбцы из существующего выбора столбцов, нажмите "Сохранить этот выбор столбцов", выберите имя из выпадающего меню и нажмите "Обновить".*



5. Теперь в таблице данных должны быть видны только нужные столбцы метаданных.



Приложение PNID01-7: Обзор рабочего процесса TheiaProk для выполнения характеристики бактерий