

ОГЛАВЛЕНИЕ - ГИПЕРССЫЛКИ НА ПРОЦЕДУРУ

- [Подготовка файла метаданных для последовательностей, которые будут загружены в Terra \(5.1\)](#)
- [Вход в Terra, используя Chrome и свою учетную запись Google \(5.2\)](#)
- [Загрузка файлов последовательностей и метаданных в Terra \(5.3\)](#)
- [Запуск рабочего процесса КК и генотипирования \(5.4\)](#)
- [Оценка метрик ККС для последовательностей \(5.5\)](#)
- [Просмотр результатов генотипирования последовательностей \(5.6\)](#)
- [Загрузка последовательностей в NCBI \(5.7\)](#)
- [Приложение PNID01-1: Импорт данных в Terra непосредственно из Illumina BaseSpace](#)
- [Приложение PNID01-2: Загрузка данных из NCBI SRA](#)
- [Приложение PNID01-3: Настройка представления таблицы данных для метрик контроля качества PulseNet](#)
- [Приложение PNID01-4а. Критические показатели качества PulseNet для рутинных последовательностей](#)
- [Приложение PNID01-4b. Шаг предварительной проверки чтения TheiaProk для исключения последовательностей низкого качества с целью экономии вычислительных ресурсов](#)
- [Приложение PNID01-5. Настройка представления таблицы данных для PulseNet для генотипирования](#)
- [Приложение PNID01-6. Загрузка дополнительных метаданных в Terra для представления в NCBI и настройка представления таблицы данных для метаданных](#)
- [Приложение PNID01-7: Обзор рабочего процесса TheiaProk для характеристики бактерий](#)

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 $Q = -10\log$ (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, СО, 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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

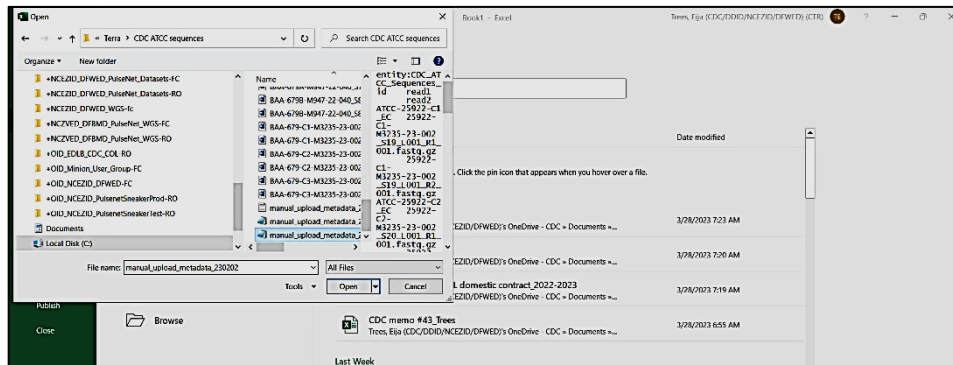
Дата вступления в силу:

Страница 4 из 67

(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.



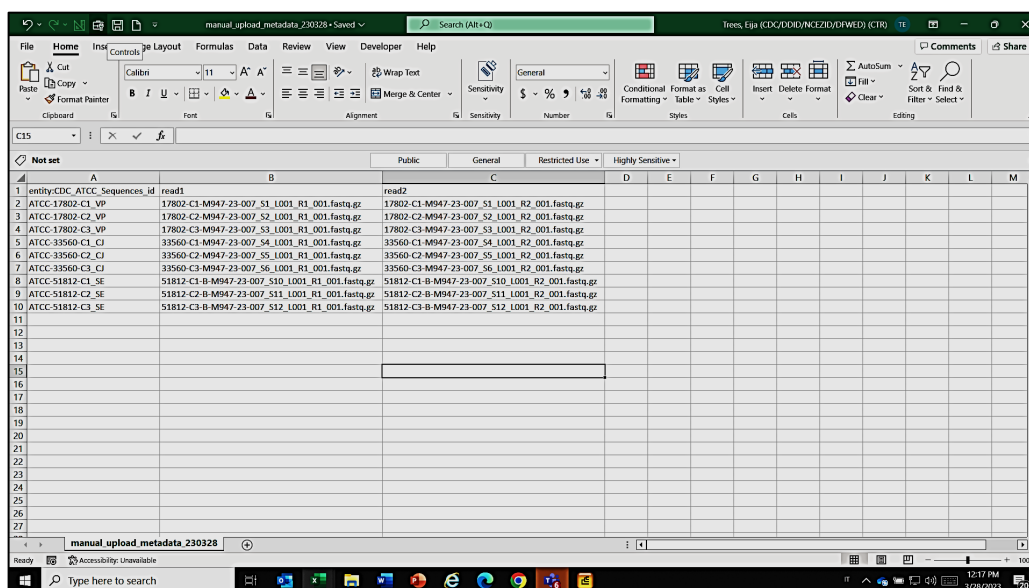
5.1.2 Необходимо заполнить как минимум следующие три столбца:

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.



МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

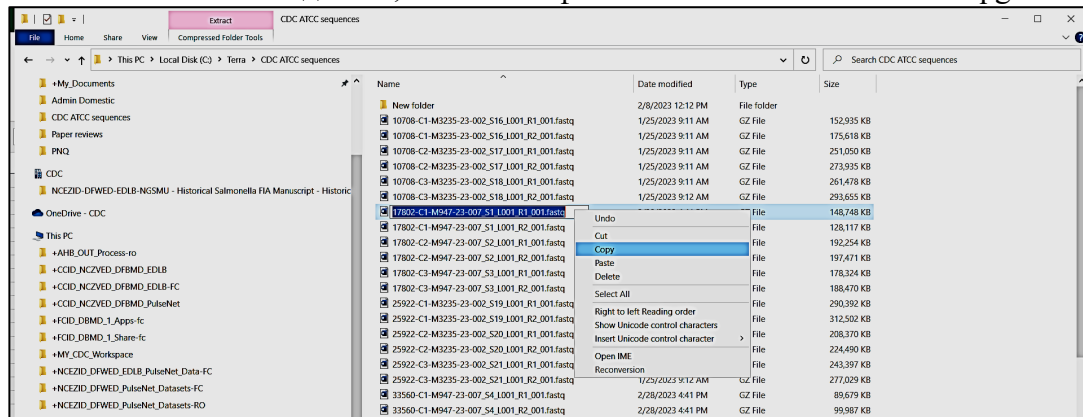
Страница 5 из 67

ПРИМЕЧАНИЕ: дополнительные метаданные могут быть добавлены на данном этапе или позднее. Руководство по метаданным см. в [приложении PNID01-6](#).

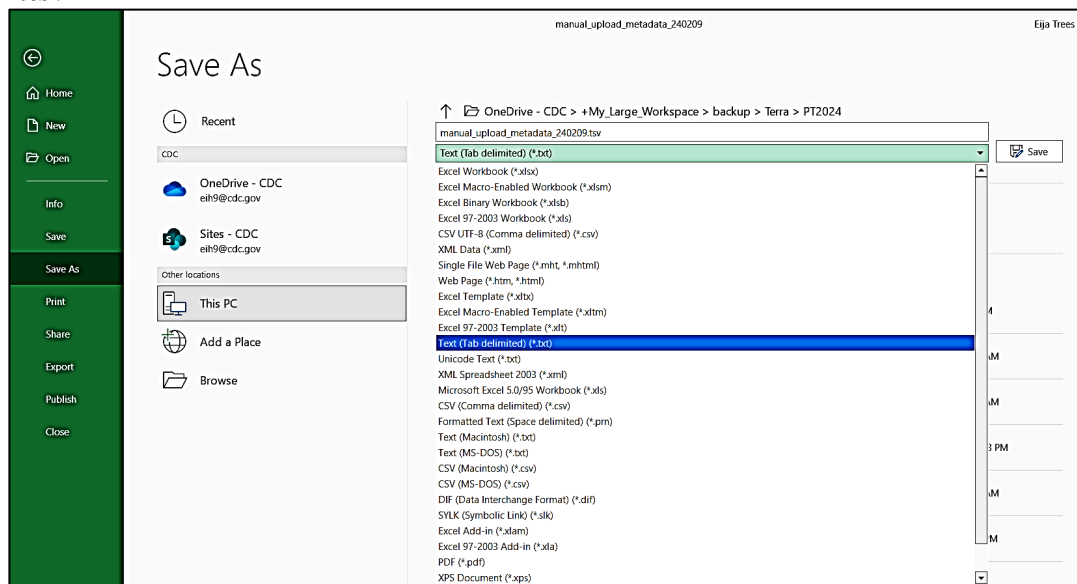
5.1.3 Введите идентификаторы штаммов в столбце "identity" так, как вы хотите, чтобы они отображались в Terra.

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.

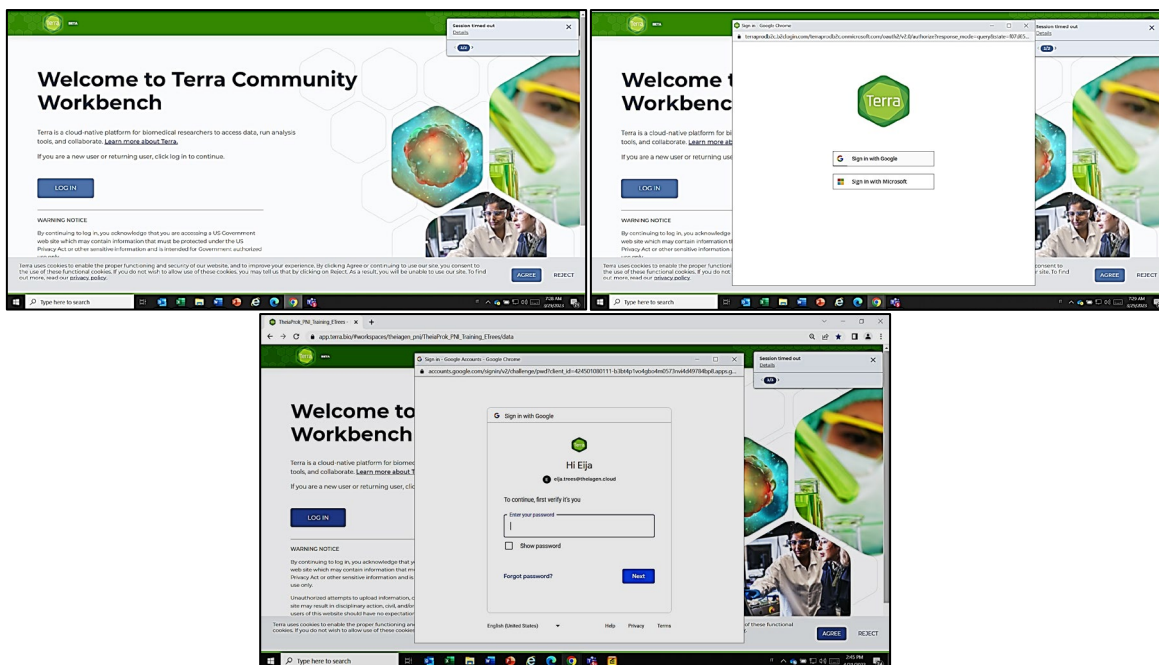
МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

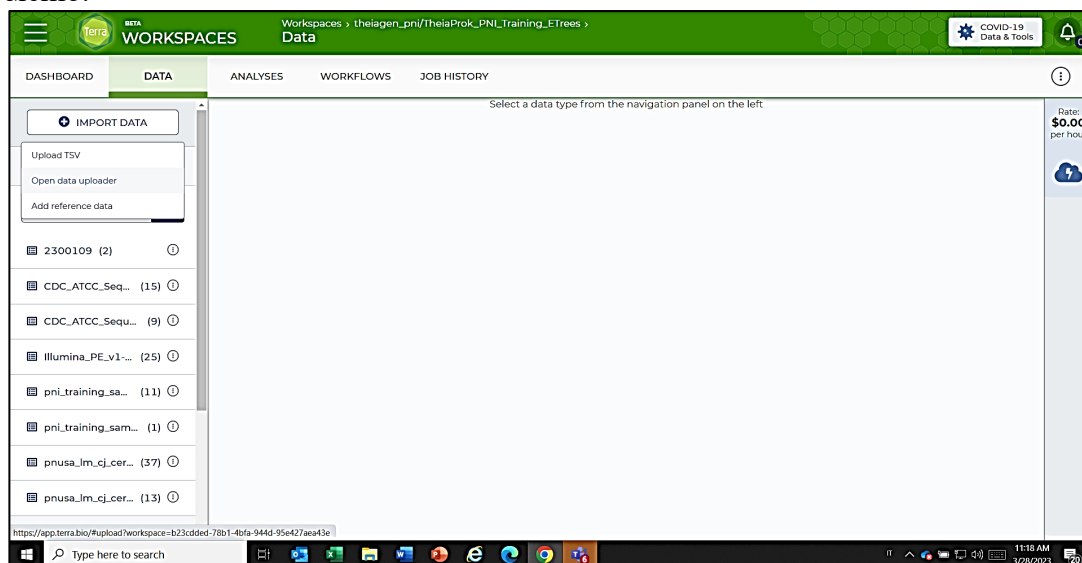
Дата вступления в силу:

Страница из 67

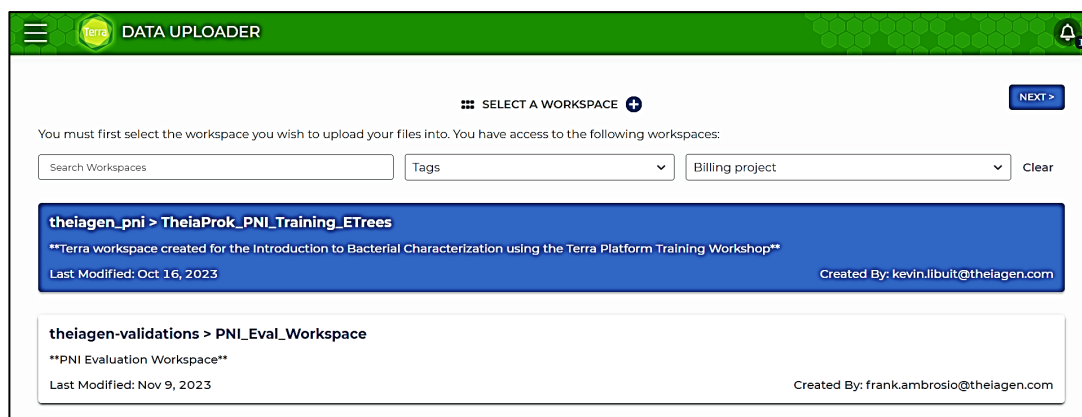


5.3 Загрузите файлы последовательностей и метаданные в Terra

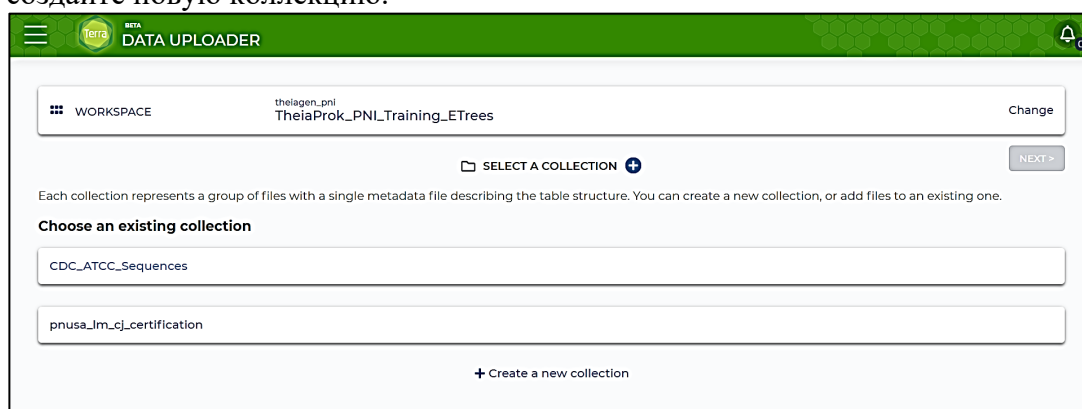
- 5.3.1 В разделе "Рабочее пространство Terra" выберите вкладку "Данные", нажмите "Импорт данных" и выберите "Открыть загрузчик данных" из выпадающего меню.



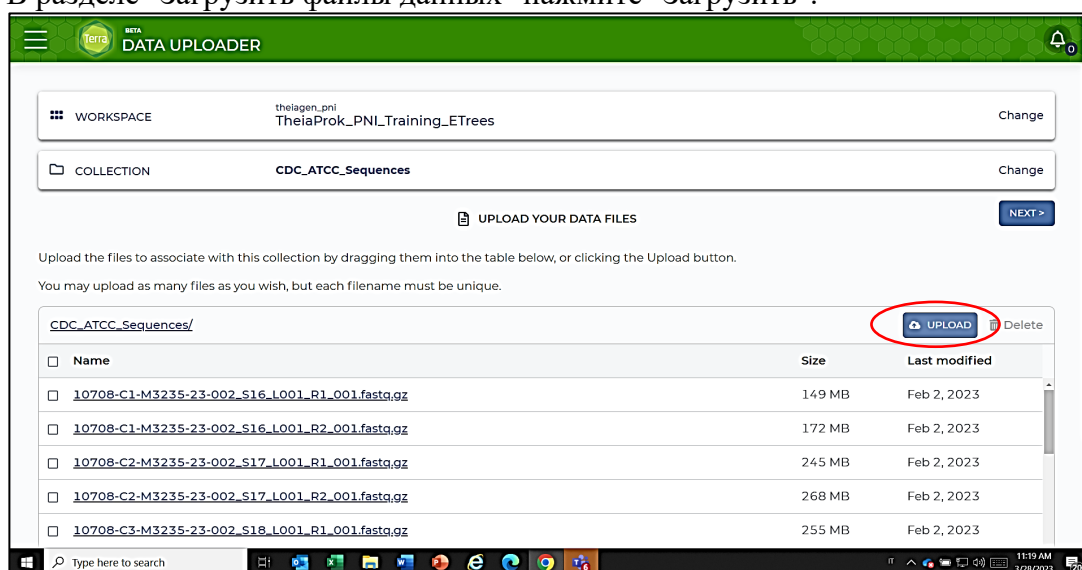
- 5.3.2 Откроется экран "Загрузчик данных". Если ваша учетная запись имеет доступ к нескольким рабочим пространствам, сначала вам нужно выбрать рабочее пространство, в которое вы хотите загрузить данные.



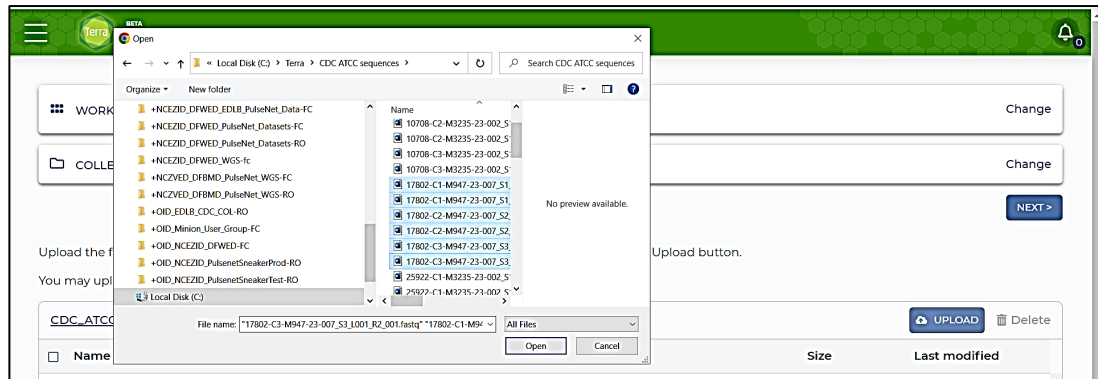
5.3.3 Выберите существующую коллекцию, щелкнув по ее названию в списке, либо создайте новую коллекцию.



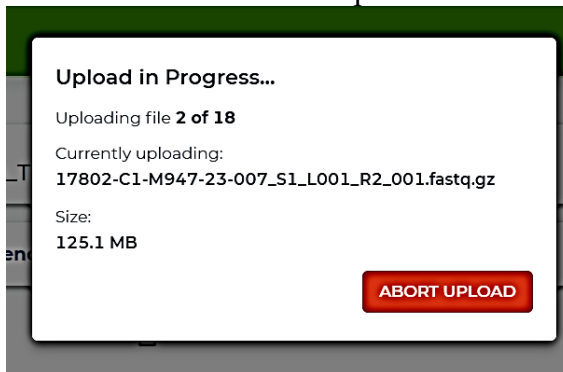
5.3.4 В разделе "Загрузить файлы данных" нажмите "Загрузить".



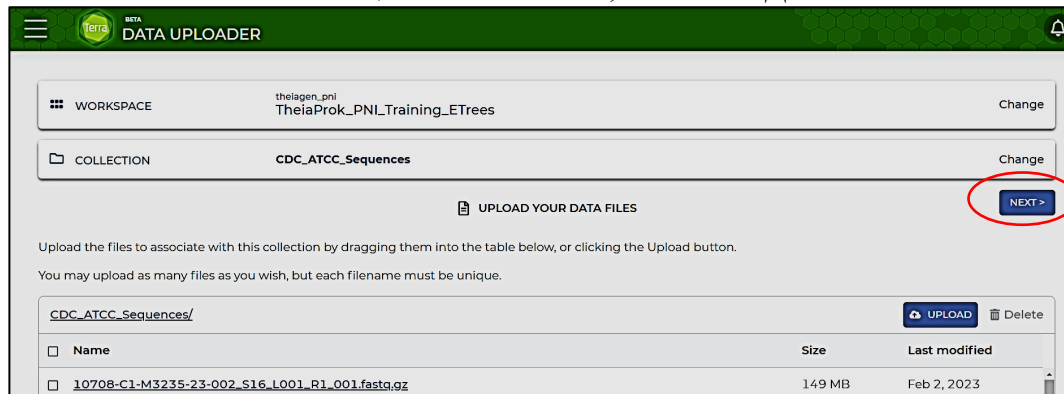
5.3.5 Перейдите в место, где сохранены файлы FASTQ, выберите файлы для загрузки и нажмите "Открыть".



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".

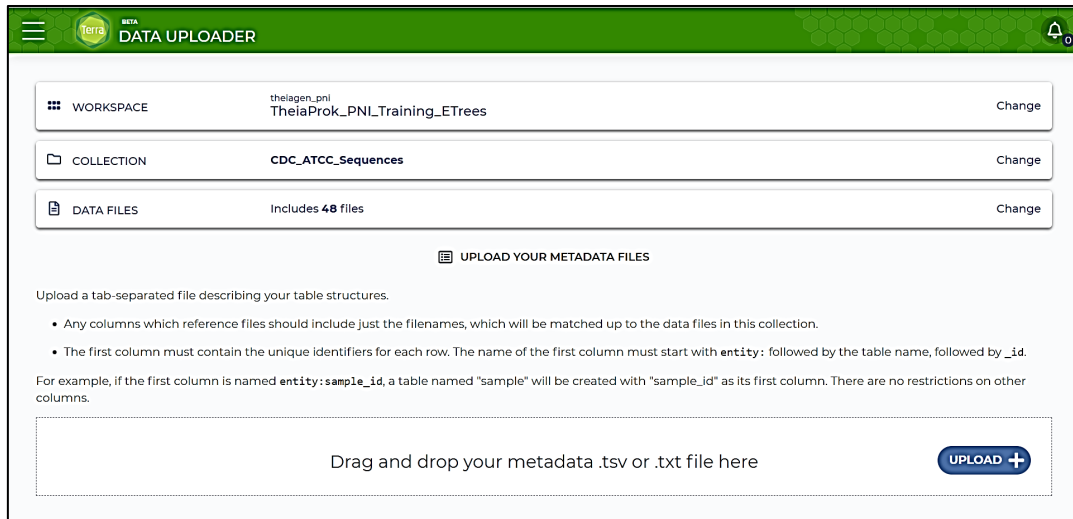
МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

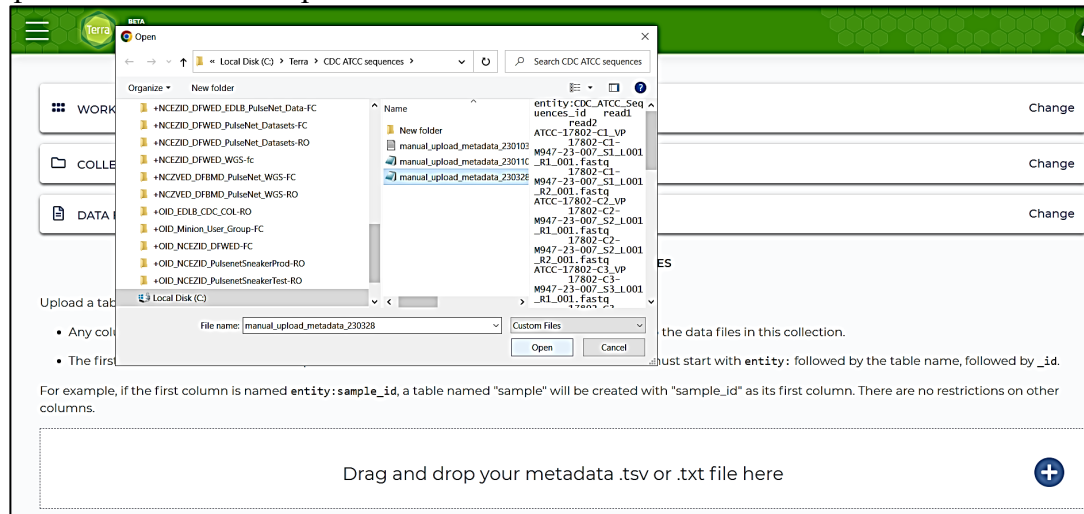
Вер. № 01

Дата вступления в силу:

Страница 9 из 67



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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 10 из 67

entity: CDC_ATCC_Sequence...	read1 (updated)	read2 (updated)
ATCC-17802-C1_VP	17802-C1-M947-23-007_S1_L001_R1_001.fastq.gz	17802-C1-M947-23-007_S1_L001_R2_001.fastq.gz
ATCC-17802-C2_VP	17802-C2-M947-23-007_S2_L001_R1_001.fastq.gz	17802-C2-M947-23-007_S2_L001_R2_001.fastq.gz
ATCC-17802-C3_VP	17802-C3-M947-23-007_S3_L001_R1_001.fastq.gz	17802-C3-M947-23-007_S3_L001_R2_001.fastq.gz
ATCC-33560-C1_CJ	33560-C1-M947-23-007_S4_L001_R1_001.fastq.gz	33560-C1-M947-23-007_S4_L001_R2_001.fastq.gz
ATCC-33560-C2_CJ	33560-C2-M947-23-007_S5_L001_R1_001.fastq.gz	33560-C2-M947-23-007_S5_L001_R2_001.fastq.gz
ATCC-33560-C3_CJ	33560-C3-M947-23-007_S6_L001_R1_001.fastq.gz	33560-C3-M947-23-007_S6_L001_R2_001.fastq.gz
ATCC-51812-C1_SE	51812-C1-B-M947-23-007_S10_L001_R1_001.fastq...	51812-C1-B-M947-23-007_S10_L001_R2_001.fastq.gz
ATCC-51812-C2_SE	51812-C2-B-M947-23-007_S11_L001_R1_001.fastq...	51812-C2-B-M947-23-007_S11_L001_R2_001.fastq.gz
ATCC-51812-C3_SE	51812-C3-B-M947-23-007_S12_L001_R1_001.fastq...	51812-C3-B-M947-23-007_S12_L001_R2_001.fastq.gz

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).

DATA UPLOADER

WORKSPACE: thelagen_pni TheiaProk_PNI_Training_ETrees [Change]

COLLECTION: CDC_ATCC_Sequences [Change]

DATA FILES: Includes 48 files [Change]

METADATA TABLES: Updated table CDC_ATCC_Sequences, added or modified 9 rows [Change]

DONE!

- View the CDC_ATCC_Sequences table in the workspace
- Create a new table in the CDC_ATCC_Sequences collection
- Start over with another workspace or collection

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

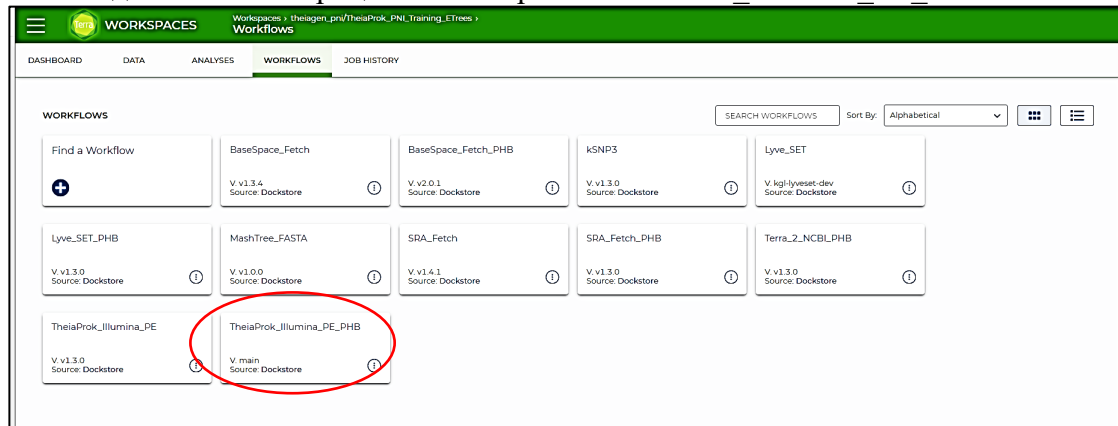
Дата вступления в силу:

Страница 11 из 67

The screenshot shows the Terra Workspaces interface. The top navigation bar includes 'Terra WORKSPACES' and 'Workspaces > theiagen_pni/TheiaProk_PNL_Training_ETrees > Data'. The main content area is divided into 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, displaying a table with columns: 'CDC_ATCC-Sequences_id', 'number_co...', 'read1', and 'read2'. The table contains 24 rows of data, including entries for ATCC-10708, ATCC-17802, and ATCC-25922. A search bar and 'ADVANCED SEARCH' options are visible at the top right of the table. The bottom right corner shows '1 - 24 of 24' items and 'Items per page: 100'.

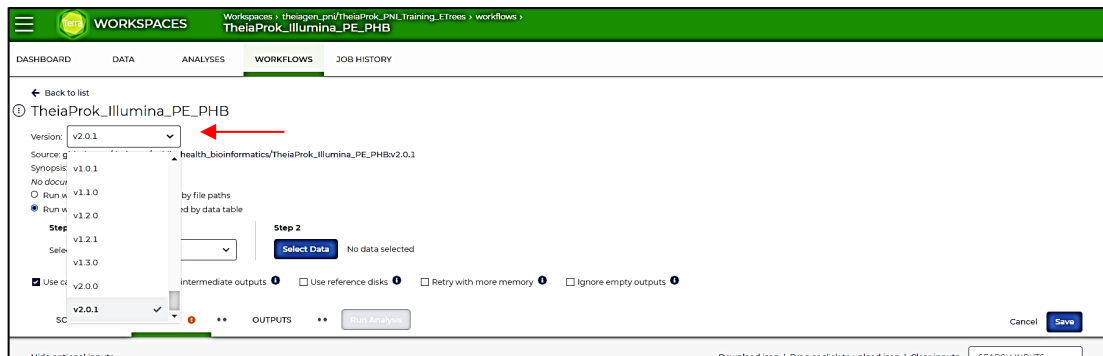
5.4 **Запустите рабочий процесс QC и генотипирования** . Рабочий процесс TheiaProk выполняет контроль качества необработанных чтений последовательностей, сборку необработанных чтений de novo с помощью Skesa, а затем контроль качества сборки и идентификацию вида. Также доступны различные анализы генотипирования,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

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



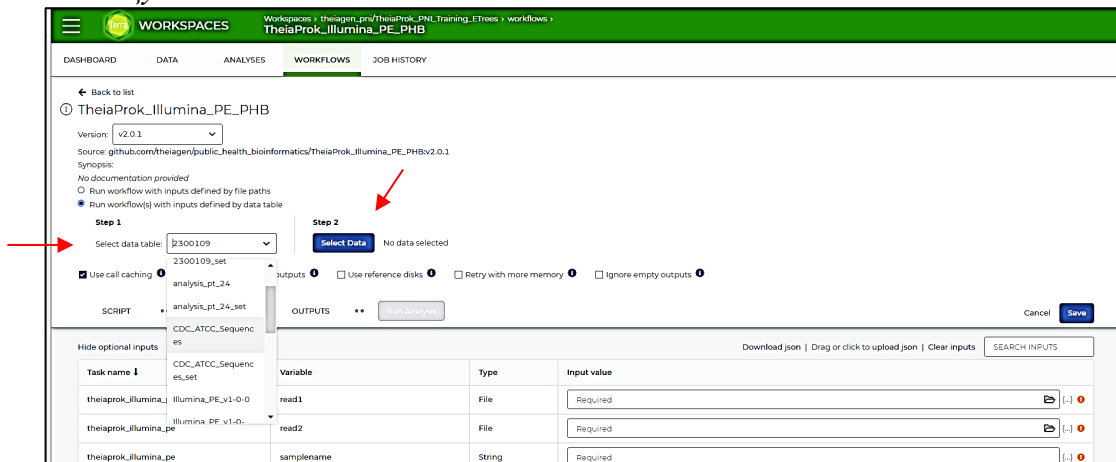
5.4.2 На экране "TheiaProk_Illumina_PE_PHB":

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 "ОК".

ПРИМЕЧАНИЕ: если таблица данных содержит более 100 записей и вы установили флажок "Выбрать все" напротив названия таблицы данных, будут выбраны только 100 записей.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 13 из 67

Select Data

Choose specific CDC_ATCC_Sequencess to process
 Choose existing sets of CDC_ATCC_Sequencess

Select CDC_ATCC_Sequencess to process **SETTINGS** | 3 rows selected **ADVANCED SEARCH** Search

<input type="checkbox"/>	CDC_ATCC_Sequencess_id	amrfinderplus_amr_genes	amrfinderplus_virulence_genes	plasmidfinder_plasmids	seqse
<input type="checkbox"/>	ATCC-10708-C1_SE	mdsA,mdsB	sinH,iroB,iroC,sodC1	IncFIB(S),IncFII(S)	7:c:1.1
<input type="checkbox"/>	ATCC-10708-C2_SE	mdsB,mdsA	sinH,iroB,iroC,sodC1	IncFIB(S),IncFII(S)	7:c:1.1
<input type="checkbox"/>	ATCC-10708-C3_SE	mdsB,mdsA	sinH,iroB,iroC,sodC1	IncFIB(S),IncFII(S)	7:c:1.1
<input checked="" type="checkbox"/>	ATCC-17802-C1_VP				
<input checked="" type="checkbox"/>	ATCC-17802-C2_VP				
<input checked="" type="checkbox"/>	ATCC-17802-C3_VP				
<input type="checkbox"/>	ATCC-17802_VP				

1 - 24 of 24 **1** Items per page: 100

Select CDC_ATCC_Sequencess to process **SETTINGS** | 8 rows selected **ADVANCED SEARCH** Search

<input type="checkbox"/>	CDC_ATCC_Sequencess_id	amrfinderplus_amr_genes	amrfinderplus_virulence_genes	plasmidfinder_plasmids	seqse
<input checked="" type="checkbox"/>	ATCC-51812-C3_SE				
<input type="checkbox"/>	BAA-679A-C1_LM	lin,fosX	No VIRULENCE genes detected by N...	No plasmids detected in database	
<input type="checkbox"/>	BAA-679A-C2_LM	lin,fosX	No VIRULENCE genes detected by N...	No plasmids detected in database	
<input type="checkbox"/>	BAA-679A-C3_LM	lin,fosX	No VIRULENCE genes detected by N...	No plasmids detected in database	
<input type="checkbox"/>	BAA-679A_LM	lin,fosX	No VIRULENCE genes detected by N...	No plasmids detected in database	
<input type="checkbox"/>	BAA-679B_LM	lin,fosX	No VIRULENCE genes detected by N...	No plasmids detected in database	

1 - 24 of 24 **1** Items per page: 100

Selected CDC_ATCC_Sequencess will be saved as a new CDC_ATCC_Sequencess_set named:

5.4.2.5 Снимите флажок "Use call catching".

WORKSPACES Workspaces > theiaigen_pn/TheiaProk_PNL_Training_ETrees > workflows > TheiaProk_Illumina_PE_PHB

DASHBOARD DATA ANALYSES **WORKFLOWS** JOB HISTORY

← Back to list

⊙ TheiaProk_Illumina_PE_PHB

Version: v2.0.1

Source: github.com/theiaigen/public_health_bioinformatics/theiaProk_illumina_pe_phbv2.0.1

Synopsis: No documentation provided

Run workflow with inputs defined by file paths
 Run workflow(s) with inputs defined by data table

Step 1 Select data table: CDC_ATCC_Seque... 5 selected CDC_ATCC_Sequencess (will create a new CDC_ATCC_Sequencess_set named "TheiaProk_Illumina_PE_PHB_2024-05-22T18-43-40")

Use call catching Delete intermediate outputs Use reference disks Retry with more memory Ignore empty outputs

SCRIPT **INPUTS** OUTPUTS

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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 14 из 67

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".

Task name ↓	Variable	Type	Input value
theiaprok_illumina_pe	read1	File	this.read1
theiaprok_illumina_pe	read2	File	this.read2
theiaprok_illumina_pe	samplename	String	this.CDC_
amfinderplus_task	cpu	int	this.CDC_ATCC_Sequences_id
amfinderplus_task	detailed_drug_class	Boolean	Optional
shovill_pe	trim	Boolean	Optional
theiaprok_illumina_pe	call_ani	Boolean	true
theiaprok_illumina_pe	call_kmerfinder	Boolean	Optional

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

Task name ↓	Variable	Type	Input value Use defaults
theiaprok_illumina_pe	abricate_abau..._plasmid.tsv	File	this.abricate_abau..._plasmid.tsv
theiaprok_illumina_pe	abricate_abau..._plasmid_type_genes	String	this.abricate_abau..._plasmid_type_genes
theiaprok_illumina_pe	abricate_database	String	this.abricate_database
theiaprok_illumina_pe	abricate_docker	String	this.abricate_docker
theiaprok_illumina_pe	abricate_version	String	this.abricate_version
theiaprok_illumina_pe	agrivate_agr_canonical	String	this.agrivate_agr_canonical
theiaprok_illumina_pe	agrivate_agr_group	String	this.agrivate_agr_group
theiaprok_illumina_pe	agrivate_agr_match_score	String	this.agrivate_agr_match_score

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**".

Run workflow with inputs defined by the paths
 Run workflow(s) with inputs defined by data table

Step 1: Select data table: CDC_ATCC_Sequ... [Select Data] 3 selected CDC_ATCC_Sequences (will create a new CDC_ATCC_Sequences_set named...)

Step 2: [Run Analysis]

Output files will be saved to: Files / submission unique ID / theiaprok_illumina_pe / workflow unique ID

References to outputs will be written to: Tables / CDC_ATCC_Sequences

Task name ↓	Variable	Type	Input value Use defaults
theiaprok_illumina_pe	abricate_abau..._plasmid.tsv	File	this.abricate_abau..._plasmid.tsv

Confirm launch

Output files will be saved as workspace data in: us (multi-region)

Running workflows will generate cloud charges. How much does my workflow cost? Set up budget alert

Describe your submission (optional): Colony picks for VP, C3 and ST1

This will launch 8 analyses.

[CANCEL] [LAUNCH]

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 15 из 67

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

The screenshot shows the 'Job History' page for a submission with ID 'a25e9aec-95ee-4097-ac72-eb75dbb6f632'. The page is divided into several sections:

- Workflow Statuses:** Submitted: 3
- Workflow Configuration:** theiaigen_pn/TheiaProk_Illumina_PE_PHB
- Submitted by:** eja.trees@theiaigen.cloud, May 23, 2024, 8:08 AM
- Total Run Cost:** N/A
- Data Entity:** TheiaProk_Illumina_PE_PHB_2024-05-23T11-57-36 CDC_ATCC_Sequences_set
- Submission ID:** a25e9aec-95ee-4097-ac72-eb75dbb6f632
- Call Caching:** Disabled
- Comment:** Repeat of the choleraesuis cert strain colony picks
- Delete Intermediate Outputs:** Disabled
- Use Reference Disks:** Disabled
- Retry with More Memory:** Disabled

Below these sections is a table of workflows:

Data Entity	Last Changed	Status	Run Cost	Messages	Workflow ID
ATCC-10708-C1_SE (CDC_ATCC_Sequences)	May 23, 2024, 8:08 AM	Queued	N/A		
ATCC-10708-C2_SE (CDC_ATCC_Sequences)	May 23, 2024, 8:08 AM	Queued	N/A		
ATCC-10708-C3_SE (CDC_ATCC_Sequences)	May 23, 2024, 8:08 AM	Queued	N/A		

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

The screenshot shows the 'Job History' page with a table of workflow details:

Submission (click for details)	Data entity	No. of Workflows	Status	Submitted	Submission ID	Comment	Actions
TheiaProk_Illumina_PE_PHB Submitted by eja.trees@theiaigen.cloud	TheiaProk_Illumina_PE_PHB...	3	Done	May 23, 2024 8:08 AM	a25e9aec-95ee-4097-ac72-eb75dbb6f632	Repeat of the choleraesuis cert strain colony...	
BaseSpace_Fetch_PHB Submitted by eja.trees@theiaigen.cloud	BaseSpace_Fetch_PHB_202...	3	Done	May 7, 2024 1:41 PM	37fa897-71ac-44bc-95bb-8a48731c109	£ additional sequences from the CAOC Bas...	
BaseSpace_Fetch_PHB Submitted by eja.trees@theiaigen.cloud	BaseSpace_Fetch_PHB_202...	10	Done	May 6, 2024 3:23 PM	bf58564-1bb0-4f69-871b-093c3894006	10 samples from the BaseSpace nextseq ru...	
BaseSpace_Fetch_PHB_frsgDqZMA Submitted by eja.trees@theiaigen.cloud	BaseSpace_Fetch_PHB_202...	10	Done	May 6, 2024 1:46 PM	95df44b-6d69-43a3-8932-b1217009c99	10 additional samples from BaseSpace next...	
TheiaProk_Illumina_PE Submitted by eja.trees@theiaigen.cloud	TheiaProk_Illumina_PE_202...	10	Done	May 6, 2024 7:52 AM	87807a12-3957-4acc-8363-cdc0e05959ec	10 samples from v3 validation run VL403-2...	

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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 16 из 67

The screenshot shows the Terra Workspaces interface. The top navigation bar includes 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, displaying a table with the following columns: 'CDC_ATCC_Sequences_id', 'agrvate_agr_canonical', 'agrvate_agr_group', and 'agrvate_agr_mat'. The table lists various ATCC strains such as ATCC-10708-C1_SE, ATCC-10708-C2_SE, ATCC-10708-C3_SE, ATCC-17802-C1_VP, ATCC-17802-C2_VP, ATCC-17802-C3_VP, ATCC-17802_VP, ATCC-25922-C1_EC, and ATCC-25922-C2_EC. A search bar and 'ADVANCED SEARCH' options are visible at the top right of the table area.

5.5.2 Выберите "Настройки".

This screenshot is similar to the previous one but highlights the 'SETTINGS' button in the table's header. A tooltip is displayed over the button, stating: 'Change the order and visibility of columns in the table'. The table content remains the same as in the previous screenshot.

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.

Select columns

Show: all | none Sort: alphabetical

- amrfinderplus_all_report
- amrfinderplus_amr_genes
- amrfinderplus_amr_report
- amrfinderplus_db_version
- amrfinderplus_stress_genes
- amrfinderplus_stress_report
- amrfinderplus_version
- amrfinderplus_virulence_genes
- amrfinderplus_virulence_report
- ani_highest_percent
- ani_highest_percent_bases_aligned
- ani_mummer_version
- ani_output_tsv

SAVE THIS COLUMN SELECTION

Your saved column selections:

- pulsenet_genotyping ⓘ
- qc_metrics ⓘ ←

qc_metrics

CANCEL **DONE**

Select columns

Show: all | none Sort: alphabetical

- agrvate_agr_canonical
- agrvate_agr_group
- agrvate_agr_match_score
- agrvate_agr_multiple
- agrvate_agr_num_frameshifts
- agrvate_docker
- agrvate_results
- agrvate_summary
- agrvate_version
- meningotype_BAST
- meningotype_FetA
- meningotype_NHBA
- meningotype_NadA

SAVE THIS COLUMN SELECTION

Your saved column selections:

- pulsenet_genotyping ⓘ
- qc_metrics ⓘ

Load

Delete

CANCEL **DONE**

5.5.4 Экспорт показателей КК в Excel:

5.5.4.1 Выберите нужные записи.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 18 из 67

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

The screenshot shows the Terra Bio Workspaces interface. The 'DATA' tab is active, displaying a table with 9 rows selected. The 'EXPORT' menu is open, showing options: 'Download as TSV', 'Export to workspace', and 'Copy to clipboard'. The table columns include 'CDC_ATCC_Sequ...', 'ani_highest_percent', 'ani_top_species_match', and 'gambit_predicter'. The table data includes sample IDs like 'ATCC-33560B_C', 'ATCC-51812-C1_SE', 'ATCC-51812-C2_SE', 'ATCC-51812-C3_SE', and 'BAA-679A-C1_LM'.

5.5.4.3 Откройте Excel и вставьте данные в рабочий лист.

The screenshot shows an Excel spreadsheet with the data imported from the Terra Bio Workspaces table. The spreadsheet has columns for 'Public', 'General', 'Restricted Use', and 'Highly Sensitive'. The data includes sample IDs, species names (e.g., 'Vibrio parahaemolyticus', 'Campylobacter jejuni', 'Salmonella enterica'), and various metrics like 'value', 'number', 'contig', 'assembly', 'length', 'est', 'coverage', 'clean', 'raw', 'midas', and 'secondary'. The 'midas' column contains text indicating 'No secondary genus detected (>1% relative abundance)'. The 'mean' column contains values like 148.65, 148.82, 148.55, 148.23, 148.4, 148.45, 148.3, 149.01, 149.31, 149.32, 148.91.

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](#).

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 19 из 67

The screenshot shows the 'Job History' page in Terra Workspaces. The workflow 'Submission c4ba38f7-33b1-4e05-ac0c-8aa3c742f8b3' is shown as 'Succeeded 1'. Key details include: Submitted by 'eja.trees@theiagen.cloud' on Mar 25, 2024, 9:04 AM; Data Entity '2013L-5615TK_NextSeq_400MB analysis_pt_24'; Submission ID 'c4ba38f7-33b1-4e05-ac0c-8aa3c742f8b3'; and Workflow ID '3cdaa856-ec22-4575-abb0-5309cc5d1a9f'. The workflow configuration includes 'Delete Intermediate Outputs' set to 'Disabled' and 'Use Reference Disks' set to 'Disabled'.

Data Entity	Last Changed	Status	Run Cost	Message	Workflow ID
2013L-5615TK_NextSeq_400MB (analysis_pt_24)	Mar 25, 2024, 9:08 AM	✓ Succeeded	N/A		3cdaa856-ec22-4575-abb0-5309cc5d1a9f

The screenshot shows a data table with columns: analysis_pt_24_id, midas_secondary.genus, midas_secondary.genus.abunda., n50.value, number_contigs, and raw_read_screen. The table contains 10 rows of data, with the last row showing a 'FAIL' status due to low estimated coverage.

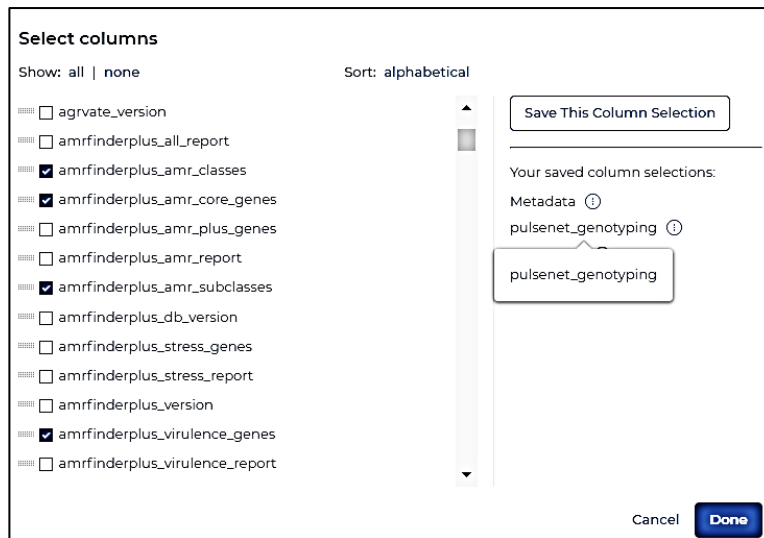
analysis_pt_24_id	midas_secondary.genus	midas_secondary.genus.abunda.	n50.value	number_contigs	raw_read_screen
2011V-1043_FLEX_300_vbribo	No secondary genus detected (>1% r...	0	124949	77	PASS
2012V-1116_FLEX_300_vbribo	No secondary genus detected (>1% r...	0	458045	39	PASS
2013L-5361_FLEX_300_LM	No secondary genus detected (>1% r...	0.0006	526928	12	PASS
2013L-5410_FLEX_300_LM	No secondary genus detected (>1% r...	0.0008	526025	15	PASS
2013L-5547_FLEX_300_LM	No secondary genus detected (>1% r...	0	435363	20	PASS
2013L-5615TK_NextSeq_400MB					FAIL: the estimated coverage is less than the minimum of 10x
2015AM-1304	No secondary genus detected (>1% r...	0	443204	15	PASS
2015AM-1305	No secondary genus detected (>1% r...	0	728098	16	PASS

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.

ty_control_id	serotype
C-4936	O157:H7
C-4938	
C-4039	O157:H7
2018C-4709-LA1-USA-CDC-pcl	
2018C-4709-LA2-USA-CDC-pcl	
2018C-4709-LB1-USA-CDC-pcl	
2018C-4709-LB2-USA-CDC-pcl	
2018EL-10533-L-A1_USA_CDC_pcl	
2018EL-10533-L-A2_USA_CDC_pcl	
2018EL-10533-L-B1_USA_CDC_pcl	
2018EL-10533-L-B2_USA_CDC_pcl	
2019C-3204	
2019C-3238	

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

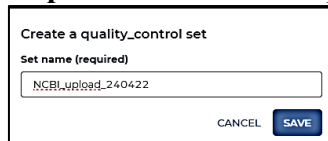
Дата вступления в силу:

Страница 21 из 67

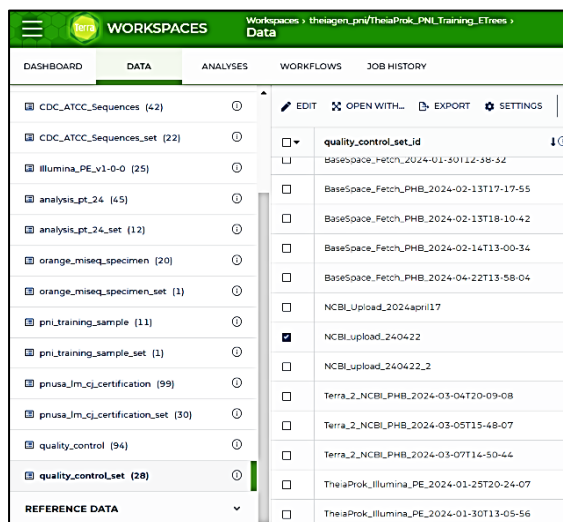
5.7.2.3 В раскрывающемся меню "Редактировать" выберите "Сохранить выборку как набор".

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.



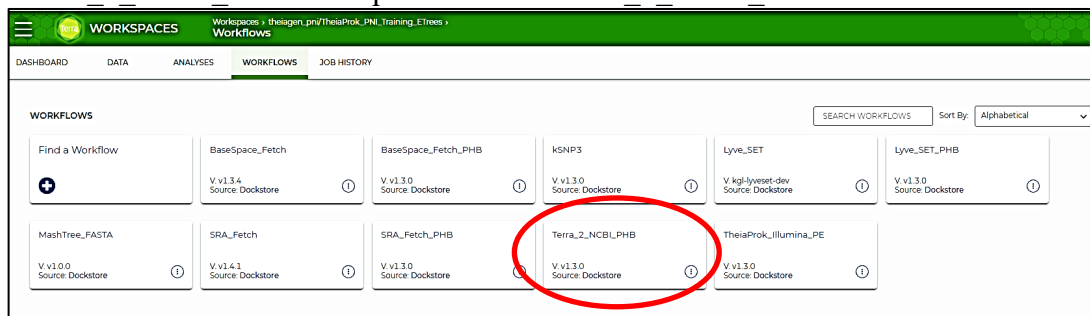
5.7.2.5 Созданный набор должен появиться в таблице данных "Набор", например, "quality_control_set".



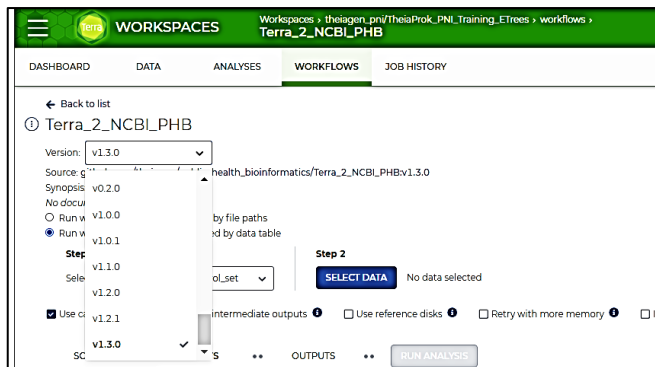
quality_control_set_id
baseSpace_Fetch_2024-01-30112-38-52
BaseSpace_Fetch_PHB_2024-02-13T17-17-55
BaseSpace_Fetch_PHB_2024-02-13T18-10-42
BaseSpace_Fetch_PHB_2024-02-14T13-00-34
BaseSpace_Fetch_PHB_2024-04-22T13-58-04
NCBI_upload_2024april17
<input checked="" type="checkbox"/> NCBI_upload_240422
NCBI_upload_240422.2
Terra_2_NCBI_PHB_2024-03-04T20-09-08
Terra_2_NCBI_PHB_2024-03-05T15-48-07
Terra_2_NCBI_PHB_2024-03-07T14-50-44
TheiaProk_Illumina_PE_2024-01-25T20-24-07
TheiaProk_Illumina_PE_2024-01-30T13-05-56

5.7.3 Настройте параметры для рабочего процесса загрузки в NCBI:

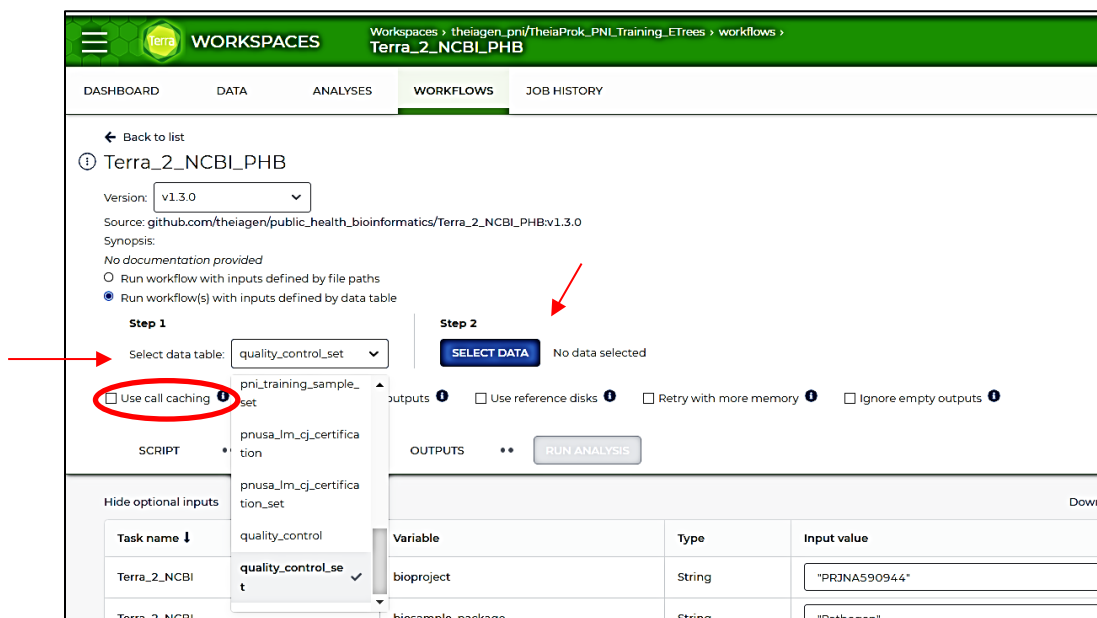
5.7.3.1 На вкладке "Рабочие процессы" выберите рабочий процесс "Terra_2_NCBI_PHB". Откроется окно "Terra_2_NCBI_PHB".



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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 23 из 67

<input type="checkbox"/>	quality_control_set_id	Terra_2_NCBI.analysis_date	Terra_2_NCBI.version	biosample_failures	biosample_metadata	biosample_report_xmls
<input type="checkbox"/>	BaseSpace_Fetch_PHB_2024-04-22T13:58-04					
<input type="checkbox"/>	NCBI_Upload_2024Apr17	2024-04-17	PHB v1.3.0	biosample_failures.txt	biosample_table.tsv	biosample_table-report.2
<input checked="" type="checkbox"/>	NCBI_Upload_240422	2024-04-22	PHB v1.3.0	biosample_failures.txt	biosample_table.tsv	biosample_table-report.2
<input type="checkbox"/>	NCBI_Upload_240422_2	2024-04-24	PHB v1.3.0		biosample_table.tsv	
<input type="checkbox"/>	Terra_2_NCBI_PHB_2024-03-04T20:09-08	2024-03-04	PHB v1.3.0	biosample_failures.txt	biosample_table.tsv	biosample_table-report.2
<input type="checkbox"/>	Terra_2_NCBI_PHB_2024-03-05T15:48-07	2024-03-05	PHB v1.3.0	biosample_failures.txt	biosample_table.tsv	biosample_table-report.2
<input type="checkbox"/>	Terra_2_NCBI_PHB_2024-03-07T14:50-44	2024-03-07	PHB v1.3.0	biosample_failures.txt	biosample_table.tsv	biosample_table-report.2

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.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 24 из 67

SCRIPT				INPUTS		OUTPUTS		RUN ANALYSIS	
Hide optional inputs								Download json Drag or click to upload json	
Task name ↓	Variable	Type	Input value						
Terra_2_NCBI	bioproject	String	*PRJNA590944*						
Terra_2_NCBI	biosample_package	String	*Pathogen*						
Terra_2_NCBI	ncbi_config_js	File	workspace ncbi_config_etrees						
Terra_2_NCBI	project_name	String	*theiagen_pni*						
Terra_2_NCBI	sample_names	Array[String]	this.quality_controls.quality_control_id						
Terra_2_NCBI	sra_transfer_gcp_bucket	String	*gs://theiagen_sra_transfer*						
Terra_2_NCBI	table_name	String	*quality_control*						
Terra_2_NCBI	workspace_name	String	*TheiaProk_PNI_Training_ETrees*						
ncbi_sftp_upload	additional_files	Array[File]	Optional						
ncbi_sftp_upload	wait_for	String	Optional						
prune_table	read1_column_name	String	Optional						
prune_table	read2_column_name	String	Optional						
Terra_2_NCBI	input_table	File	Optional						
Terra_2_NCBI	skip_biosample	Boolean	Optional						
Terra_2_NCBI	submit_to_production	Boolean	true						

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.

SCRIPT				INPUTS		OUTPUTS		RUN ANALYSIS	
Output files will be saved to								Download json Drag or click to upload json Clear outputs	
Files / submission unique ID / Terra_2_NCBI / workflow unique ID								SEARCH OUTPUTS	
References to outputs will be written to									
Tables / quality_control_set									
Fill in the attributes below to add or update columns in your data table									
Task name ↓	Variable	Type	Input value Use defaults ←						
Terra_2_NCBI	biosample_failures	File	this.biosample_failures						
Terra_2_NCBI	biosample_metadata	File	this.biosample_metadata						
Terra_2_NCBI	biosample_report_xmIs	Array[File]	this.biosample_report_xmIs						
Terra_2_NCBI	biosample_status	String	this.biosample_status						
Terra_2_NCBI	biosample_submission_xml	File	this.biosample_submission_xml						
Terra_2_NCBI	excluded_samples	File	this.excluded_samples						
Terra_2_NCBI	generated_accessions	File	this.generated_accessions						
Terra_2_NCBI	sra_metadata	File	this.sra_metadata						

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 25 из 67

Confirm launch

Output files will be saved as workspace data in:
us-us-central1 (lowa) ⓘ

Running workflows will generate cloud charges. ⓘ
How much does my workflow cost? ⓘ
Set up budget alert ⓘ

Describe your submission (optional):

NCBI upload of 3 sequences to the validation
bioproject

This will launch 1 analysis.

CANCEL
LAUNCH

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

The screenshot shows the 'Job History' page in Terra Workspaces. The submission ID is 387be0c7-88b6-4c00-8028-9020dc5f08b8. The workflow is 'NCBIUpload240422 quality_control_set'. The status is 'Submitted: 1'. The submission was made by 'eja.trees@theiagen.cloud' on Apr 22, 2024, at 7:47 AM. The total run cost is N/A. The comment is 'NCBI upload of 3 sequences to the validation...'. The 'Delete Intermediate Outputs' and 'Use Reference Disks' options are disabled. The 'Call Caching' and 'Retry with More Memory' options are enabled.

Data Entity	Last Changed	Status	Run Cost	Messages	Workflow ID
NCBIUpload240422 (quality_control_set)	Apr 22, 2024, 7:47 AM	Queued	N/A		

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

Submission (click for details)	Data entity	No. of Workflows	Status	Submitted	Submission ID	Comment	Actions
Terra_2_NCBILPHB Submitted by: eja.trees@theiagen.cloud	NCBIUpload240422 (quality...	1	✓ Done	Apr 22, 2024 7:47 AM	387be0c7-88b6-4c00-8028-9020dc5f08b8	NCBI upload of 3 sequences to the...	ⓘ
Terra_2_NCBILPHB Submitted by: eja.trees@theiagen.cloud	NCBIUpload2024apr117 [...]	1	✓ Done	Apr 17, 2024 7:51 AM	836d919-796b-4c49-8d66-418c9f95537	NCBI submission of 3 new sample...	ⓘ
TheiaProk_Illumina_PE Submitted by: eja.trees@theiagen.cloud	TheiaProk_Illumina_PE_202...	6	✓ Done	Apr 15, 2024 12:02 PM	4bb2064-f8b-4942-861c-3e85049f8ad	6 strains from the AMD incubator ...	ⓘ
TheiaProk_Illumina_PE Submitted by: eja.trees@theiagen.cloud	TheiaProk_Illumina_PE_202...	3	✓ Done	Apr 15, 2024 1:32 PM	5d44c8c-518b-46c4-9b3b-9d75d407027	Colony pick 2 for 2023 PH strains	ⓘ
TheiaProk_Illumina_PE Submitted by: eja.trees@theiagen.cloud	TheiaProk_Illumina_PE_202...	3	✓ Done	Apr 15, 2024 12:00 PM	89c7f6c-d6f7-43a3-a377-24091af7cda	3 PH isolates from 2023	ⓘ

5.7.3.11 Отслеживание номеров доступа NCBI:

5.7.3.11.1 Номера доступа к биообразцам (номера SAMN)

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 26 из 67

- 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.

The screenshot shows the Terra Workspaces interface with the 'Data' tab selected. A table of quality control data is displayed. The 'biosample_accession' column is circled in red. The table has the following columns: quality_control_id, biosample_accession, and collected_by. The data rows are as follows:

quality_control_id	biosample_accession	collected_by
<input type="checkbox"/> 2017C-4936	SAMN41039458	CDC
<input type="checkbox"/> 2017C-4938		
<input checked="" type="checkbox"/> 2018C-4039	SAMN41039457	CDC
<input type="checkbox"/> 2018C-4709-L-A1-USA-CDC-pcl		
<input type="checkbox"/> 2018C-4709-L-A2-USA-CDC-pcl		
<input type="checkbox"/> 2018C-4709-L-B1-USA-CDC-pcl		
<input type="checkbox"/> 2018C-4709-L-B2-USA-CDC-pcl		
<input type="checkbox"/> 2018EL-1053a-L-A1_USA_CDC_pcl		
<input type="checkbox"/> 2018EL-1053a-L-A2_USA_CDC_pcl		
<input type="checkbox"/> 2018EL-1053a-L-B1_USA_CDC_pcl		
<input type="checkbox"/> 2018EL-1053a-L-B2_USA_CDC_pcl		
<input checked="" type="checkbox"/> 2019C-3204	SAMN41039456	CDC
<input type="checkbox"/> 2019C-3238		

- 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:

The screenshot shows the Terra Workspaces interface with the 'Data' tab selected. A table of quality control set data is displayed. The 'generated_accessions' column is circled in red. The table has the following columns: quality_control_set_id, excluded_samples, generated_accessions, and lysaset_allp. The data rows are as follows:

quality_control_set_id	excluded_samples	generated_accessions	lysaset_allp
<input type="checkbox"/> BaseSpace_Fetch_PHB_2024-02-13T17-17-55			
<input type="checkbox"/> BaseSpace_Fetch_PHB_2024-02-13T18-10-42			
<input type="checkbox"/> BaseSpace_Fetch_PHB_2024-02-14T13-00-34			
<input type="checkbox"/> BaseSpace_Fetch_PHB_2024-04-22T13-58-04			
<input type="checkbox"/> NCBI_upload_2024april17	excluded_samples.tsv	generated_accessions.tsv	
<input checked="" type="checkbox"/> NCBI_upload_240422	excluded_samples.tsv	generated_accessions.tsv	
<input type="checkbox"/> NCBI_upload_240422_2	exclude	gs://fc-b23c05ed-78b1-4bfa-944f-95e427ae43e/submissions/387be0c7-8bb6-4c0d-8a484-347a274d3ac/call-biosample-submit_tsvftp_upload/generated_accessions.tsv	
<input type="checkbox"/> Terra_2_NCBIPHB_2024-03-04T20-09-08	excluded_samples.tsv	generated_accessions.tsv	
<input type="checkbox"/> Terra_2_NCBIPHB_2024-03-05T15-48-07	excluded_samples.tsv	generated_accessions.tsv	
<input type="checkbox"/> Terra_2_NCBIPHB_2024-03-07T14-50-44	excluded_samples.tsv	generated_accessions.tsv	

- 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.

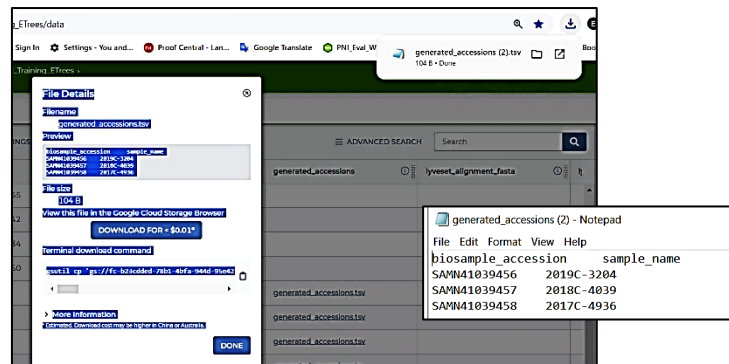
МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

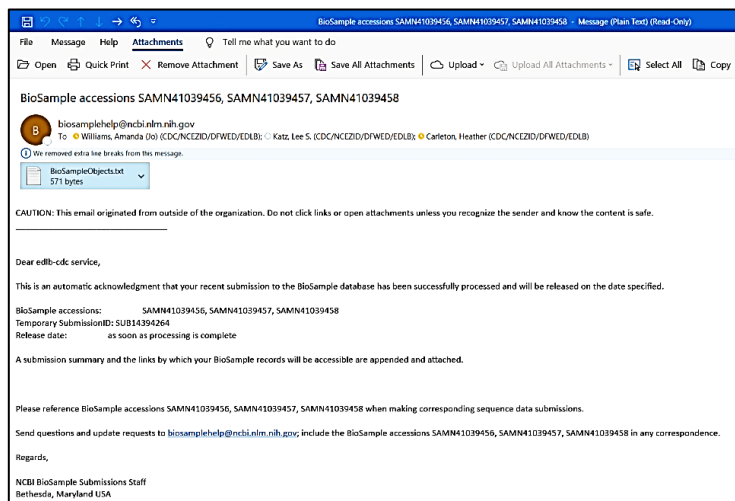
Вер. № 01

Дата вступления в силу:

Страница 27 из 67



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.



BioSampleObjects - Notepad

Accession	SUID	Organism	Tax ID	Strain	BioProject
SAMN41039458	2017C-4936	Escherichia coli	562	2017C-4936	PRJNA590944
SAMN41039457	2018C-4039	Escherichia coli	562	2018C-4039	PRJNA590944
SAMN41039456	2019C-3204	Shigella sonnei	624	2019C-3204	PRJNA590944

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.

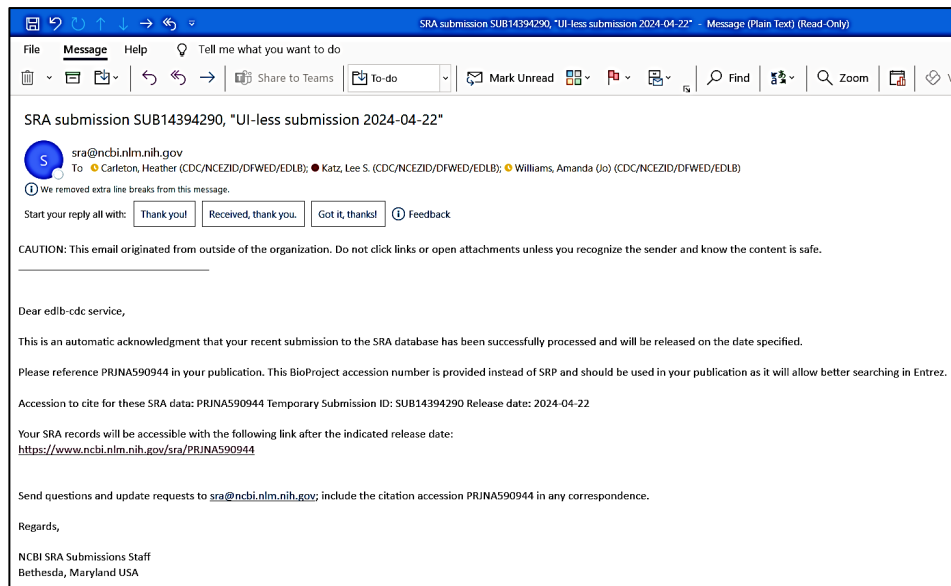
МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

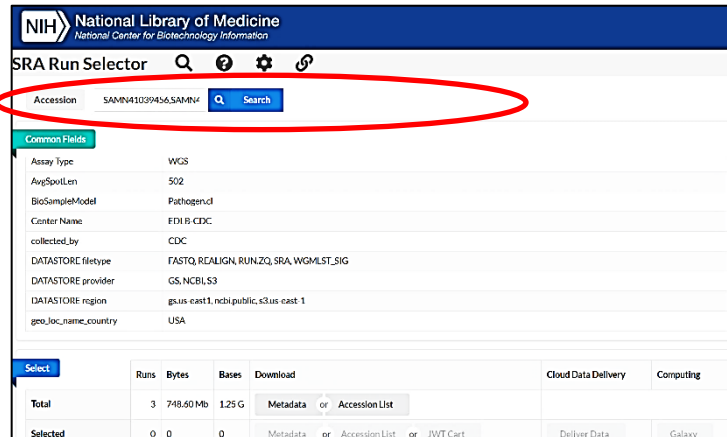
Вер. № 01

Дата вступления в силу:

Страница 28 из 67



- 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.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 29 из 67

The screenshot shows the SRA Run Selector interface. At the top, there are search and filter options. Below, a summary table shows 3 runs with a total size of 748.60 Mb and 1.25 G bases. A 'Download' button is visible, with a dropdown menu open showing options for 'SRR_Acc_List (1).txt' and 'SRR_Ac...'. Below the summary is a table of 3 items with columns for Run, BioSample, Bases, Bytes, Collection Date, Experiment, Library Name, Organism, create_date, Sample Name, strain, and serotype.

Run	BioSample	Bases	Bytes	Collection Date	Experiment	Library Name	Organism	create_date	Sample Name	strain	serotype
1	SRR28763089	SAMN41039456	216.17 M	2017-01-01	SRX24328488	2019C-3204-001	Shigella sonnei	2024-04-22 08:27:00Z	2019C-3204	2019C-3204	
2	SRR28763090	SAMN41039457	623.29 M	2018-01-01	SRX24328487	2018C-4039-001	Escherichia coli	2024-04-22 08:28:00Z	2018C-4039	2018C-4039	O157:H7
3	SRR28763091	SAMN41039458	412.37 M	2017-01-01	SRX24328486	2017C-4936-001	Escherichia coli	2024-04-22 08:28:00Z	2017C-4936	2017C-4936	O157:H7

The screenshot shows a Notepad window titled 'SRR_Acc_List (2) - Notepad'. The text inside the window is:

```

SRR28763089
SRR28763090
SRR28763091
    
```

5.7.3.11.2.4 Вы можете отслеживать номера вступления в СРА 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".

The screenshot shows the Terra Workspaces interface. A table of data is displayed with columns for 'assembly_fastq' and 'assembly_fasta'. A context menu is open over the table, and the 'Add column' option is selected, pointing to a new column header 'ty_control_id'.

assembly_fastq	assembly_fasta
005-A1_USA_CDC_pcl	08-0005-A1_USA_CDC_pcl.contigs.fasta
005-A2_USA_CDC_pcl	08-0005-A2_USA_CDC_pcl.contigs.fasta
005-B1_USA_CDC_pcl	08-0005-B1_USA_CDC_pcl.contigs.fasta
08-0005-B2_USA_CDC_pcl	08-0005-B2_USA_CDC_pcl.contigs.fasta
2011L-2624-LRM4update	2011L-2624-LRM4update.contigs.fasta
2011L-2624_v3index_NextSeq	
2011V-1043-LRM4update	2011V-1043-LRM4update.contigs.fasta
2012D-9301-A-USA-CDC-pcl	2012D-9301-A-USA-CDC-pcl.contigs.fasta
2012D-9301-B-USA-CDC-pcl	2012D-9301-B-USA-CDC-pcl.contigs.fasta
2012D-9301-C-USA-CDC-pcl	2012D-9301-C-USA-CDC-pcl.contigs.fasta

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 30 из 67

Add a new column

Column name

Default value (optional, will be entered for all rows)

Type:

String Reference Number Boolean

Value is a list

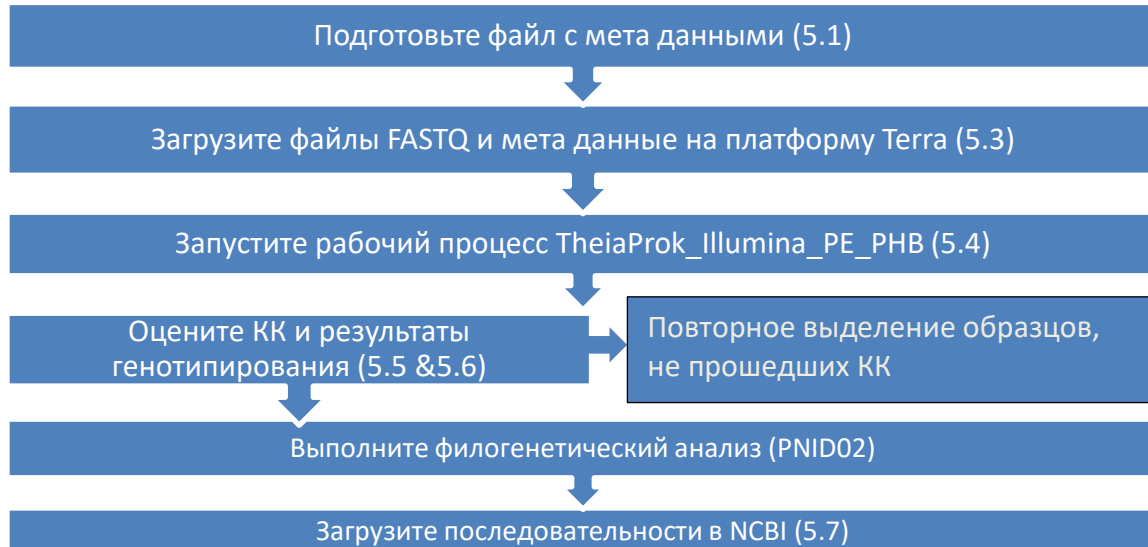
Cancel **Save**

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.

quality_control_id	sra_accession	assembly_fasta
2015K-0887.v3index_NextSeq		
2015K-1104.v3index_NextSeq		
2015K-1440-LRM4update		
2017C-3818-LRM4update		2017C-3818-LRM4update-contigs.fasta
2017C-3830-LRM4update		
2017C-4936		2017C-4936_contigs.fasta
2017C-4938		2017C-4938_contigs.fasta
2018C-4039		2018C-4039_contigs.fasta
2018C-4709-L-A1-USA-CDC-pcl		2018C-4709-L-A1-USA-CDC-pcl-contigs.fasta
2018C-4709-L-A2-USA-CDC-pcl		2018C-4709-L-A2-USA-CDC-pcl-contigs.fasta

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.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 32 из 67

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.

STATUS	RUN NAME	AVG Q30	%PF	INSTRUMENT	CREATED
Complete	WV-M07896-231215	81.72%	48.18%	M07896	2023-12-15 15:18
Complete	WV-M07896-231213	90.55%	82.75%	M07896	2023-12-13 12:18
Complete	VL403-23-003	88.63%	79.83%	VL00403	2023-12-12 13:35
Complete	WV-M07896-231128	46.99%	10.94%	M07896	2023-11-28 16:40
Complete	M3235-23-042	92.90%	91.02%	M03235	2023-11-07 16:16
Complete	M3235-23-041	88.00%	92.64%	M03235	2023-11-03 12:01
Complete	IMR-M01432-241023	84.67%	69.95%	M01432	2023-10-24 01:26
Complete	IMR-M01432-171023	84.35%	68.96%	M01432	2023-10-17 01:05
Complete	OH-VH01632-230919	89.77%	76.79%	VH01632	2023-09-19 16:25
Complete	MiSeq Vc 015	91.99%	94.84%	M08444	2023-09-01 15:45

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

M3235-23-042

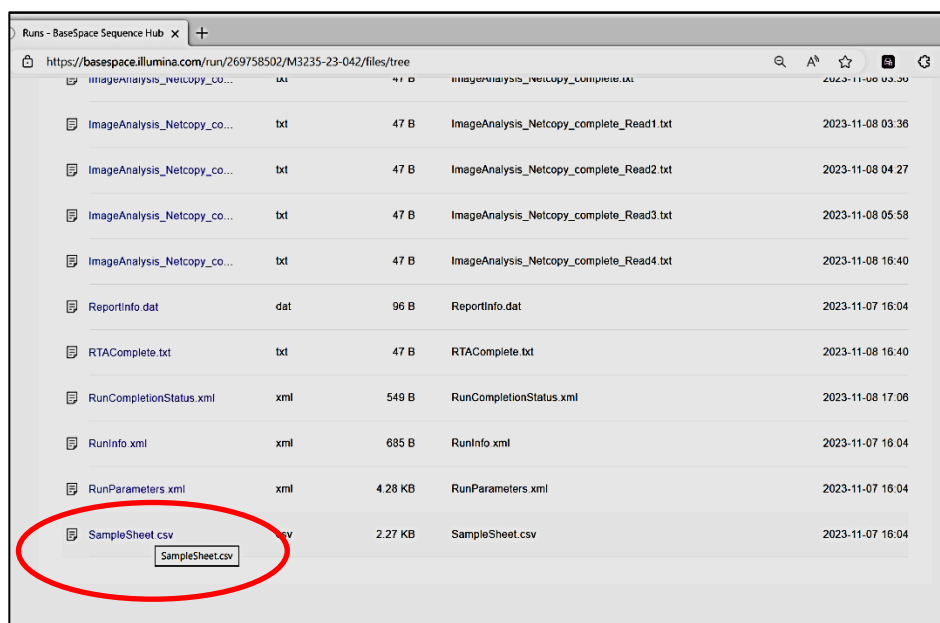
SUMMARY BIOSAMPLES SAMPLES CHARTS METRICS INDEXING QC SAMPLE SHEET **FILES**

Important: Due to recent security updates, you may have issues logging in or uploading data to BaseSpace, ICA, and Proactive if you have not updated your instrument. LEARN MORE LEARN MORE

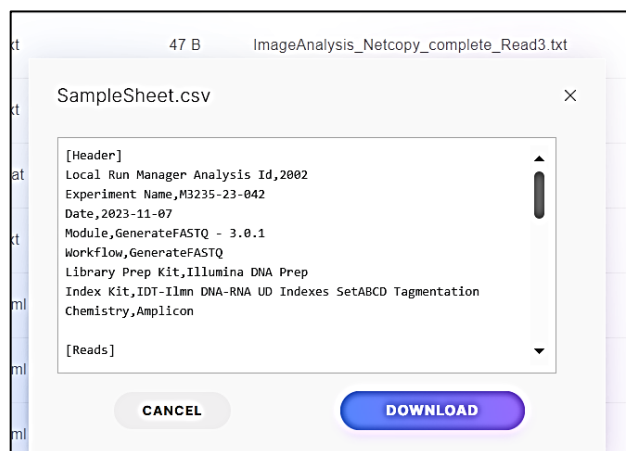
Transfer of projects containing NextSeq 2000 or BCL Convert data is currently disabled. Data can be shared as an alternative until the feature is restored.

Instrument	M03235	92.90 AVG %Q30	91.02 %PF	Created	2023-11-07 16:16	Instrument Type	MiSeq
Run Status	Complete	Lane QC Status	QcPassed	File Count/Size	55,592 files (10 GB)	File Status	Active
Latest Analysis	2023-11-07 16:16	Flow Cell Status	QcPassed	Owner	PulseNet NGS Lab 1	User	PulseNet NGS Lab 1
Cycles	151 10 10 151	Yield	5.38 Gbp	Flow Cell ID	00000000-L3P8P	Run ID	231107_M03235_01...

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".

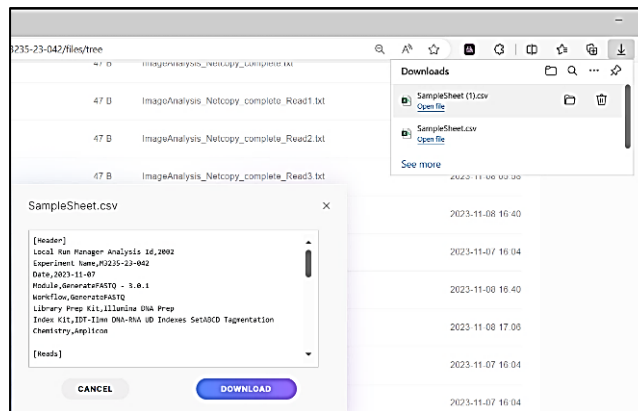
МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 35 из 67



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:

Sample_ID	Sample_Name	Description	Index_Plant	Index	I7_Index	Index	I5_Index	Index2	Sample_Project
D5480-M3235-23-042	D5480-M3235-23-042		B	A06	UDP0137	CCGGTTCU	UDP0137	TATATTCGAG	
ATCC-BAA-460-M3235-23-042	ATCC-BAA-460-M3235-23-042		B	B06	UDP0138	GGCCAAT/	UDP0138	CGGTCCGATA	
2011L-2624-M3235-23-042	2011L-2624-M3235-23-042		B	C06	UDP0139	GAATACT	UDP0139	ACAATAGAGT	
2015K-0092-M3235-23-042	2015K-0092-M3235-23-042		B	D06	UDP0140	TACGTGA/	UDP0140	CGGTATTAG	
2015K-1440-M3235-23-042	2015K-1440-M3235-23-042		B	E06	UDP0141	CTTATTGG	UDP0141	GATAACAAGT	
2013V-1178-M3235-23-042	2013V-1178-M3235-23-042		B	F06	UDP0142	ACAACAC	UDP0142	AGTTATCACA	
2014C-3598-M3235-23-042	2014C-3598-M3235-23-042		B	G06	UDP0143	GTTGGAT	UDP0143	TTCCAGGTAA	
2014C-3857-M3235-23-042	2014C-3857-M3235-23-042		B	H06	UDP0144	AATCCAAT	UDP0144	CATGTAGAGG	
2015C-3881-M3235-23-042	2015C-3881-M3235-23-042		B	A07	UDP0145	TATGATG	UDP0145	GATTGCATA	

i. entity_datatable_name_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. basebase_sample_name: скопируйте и вставьте содержимое поля SampleSheet "Sample_Name".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 36 из 67

- 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 Analysts 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-illum DNA-RNA UD Indexes SetABCD Tagmentation													
9	Chemistry	Amplicon													
10															
11	[Reads]														
12		151													
13		151													
14															
15	[Settings]														
16	adapter	CTGTCTCTTATACATCT													
17															
18	[Data]														
19	Sample_ID	Sample_Name	Descriptio	Index	Pla	Index	Pla	I7	Index	index	I5	Index	index2	Sample	Project
20	2023FD-00134	CAOC-M5870-230530E	A	A07	UDP0049	AGTGTGTG	UDP0049		CTGGTAG	CPO	230530E				
21	BE230960535	BE230960535-CAOC-M5870-230530E	A	B07	UDP0050	GACACCA	UDP0050		TCAACGT	Sal	230530E				
22	2023FD-00135	CAOC-M5870-230530E	A	C07	UDP0051	CCTGTCTG	UDP0051		ACTGTGTG	CPO	230530E				
23	2023FD-00136	CAOC-M5870-230530E	A	D07	UDP0052	TGATGTA	UDP0052		GTGCGT	CPO	230530E				
24	2023FD-00137	CAOC-M5870-230530E	A	E07	UDP0053	GGAAATG	UDP0053		AGCACAT	CPO	230530E				
25	BE231320288	BE231320288-CAOC-M5870-230530E	A	F07	UDP0054	GCATAAG	UDP0054		TTCCGTC	Ecoli	Shig230530E				
26	2023FD-00138	CAOC-M5870-230530E	A	G07	UDP0055	CTGAGGA	UDP0055		CTTAACC	CPO	230530E				
27	2023FD-00139	CAOC-M5870-230530E	A	H07	UDP0056	AACGCAC	UDP0056		GCCTCGG	CPO	230530E				
28	BE231330092	BE231330092-CAOC-M5870-230530E	A	A08	UDP0057	TCTATCT	UDP0057		CGTCGAC	Sal	230530E				
29	BE231330093	BE231330093-CAOC-M5870-230530E	A	B08	UDP0058	CTCGCTC	UDP0058		TACTAGT	Sal	230530E				
30	BE231350225	BE231350225-CAOC-M5870-230530E	A	C08	UDP0059	CTGTGTG	UDP0059		ATAGACO	Sal	230530E				

- i. entity_datatable_name_id:

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 37 из 67

2. Введите идентификаторы образцов в столбец А так, как вы хотите, чтобы они отображались в таблице данных Terra.
 - ii. basespace_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	I
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									

с. Столбцы, необходимые для таблицы образцов NextSeq:

	A	B	C
1	[Header]		
2	FileFormatVersion		2
3	RunName	VI463-24-001	
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	OverreadCycles	Y15110p10p151	
21	FastqCompressionFormat	gzip	
22			
23	[BCLConvert_Data]		
24	Sample_ID	Index	Index2
25	2013L-5214	CGACATCCGA	TACGTTTCATT
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-0823	AGTAGAAGTG	CTAGCGTCGA

- 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 А так, как вы хотите, чтобы они отображались в таблице данных Terra.
- ii. basespace_sample_name: скопируйте и вставьте содержимое поля SampleSheet "Sample_ID".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

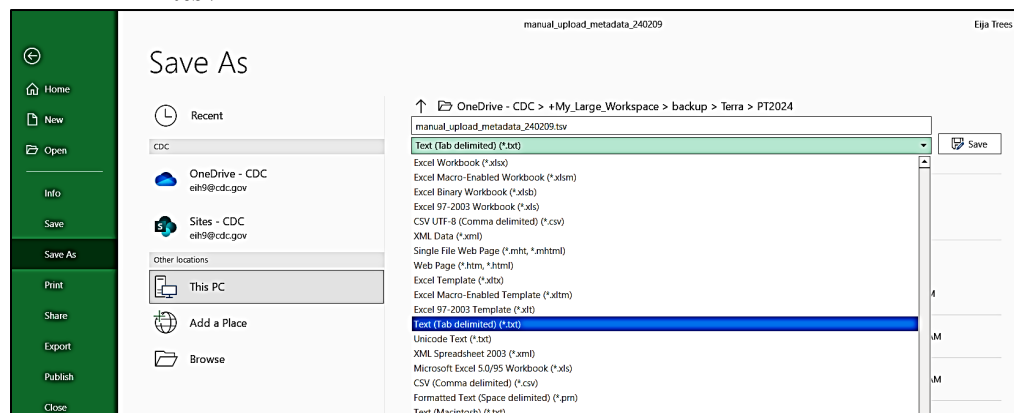
Дата вступления в силу:

Страница 38 из 67

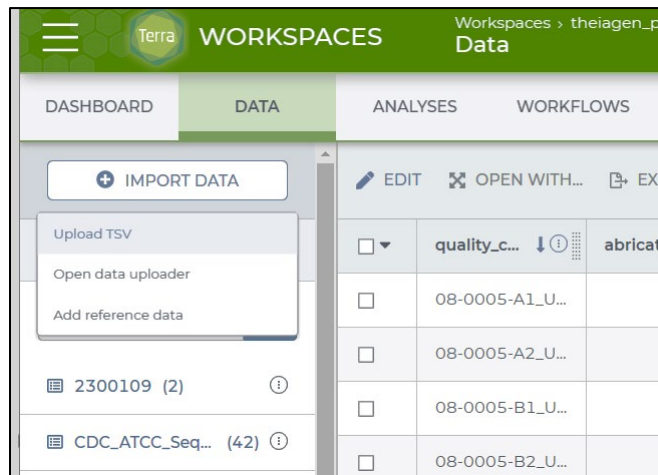
- iii. basespace_sample_id: скопируйте и вставьте содержимое поля SampleSheet "Sample_ID".
- iv. basespace_collection_id: введите название Run так, как оно отображается на BaseSpace.

	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					

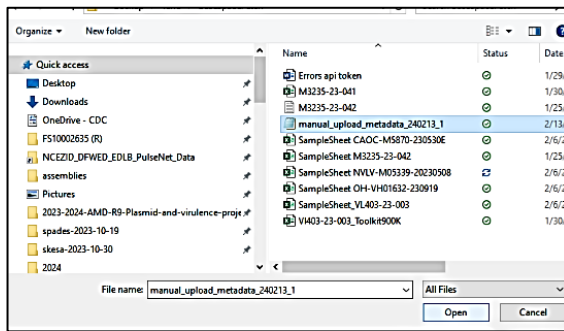
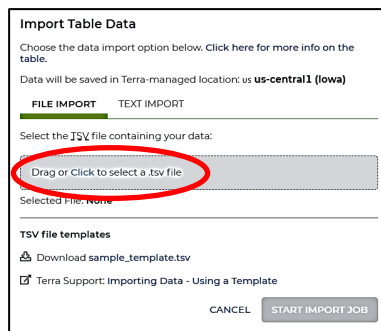
- d. Сохраните файл в **формате tsv**: выберите "Сохранить как" и "Текст (с разделителем табуляции) (*.txt)". Убедитесь, что имя файла имеет окончание **".tsv"**.



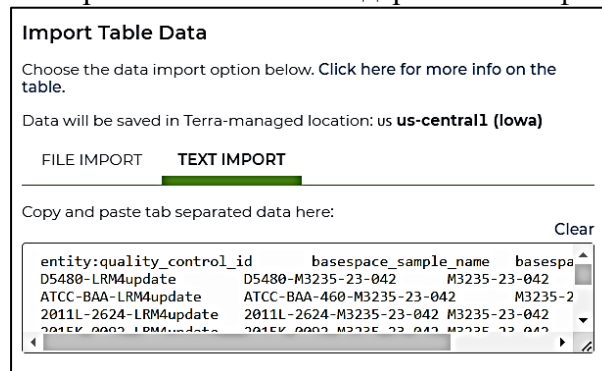
- 5. Импортируйте файл метаданных tsv:
 - a. На вкладке "Данные" нажмите на "Импорт данных" и выберите "Загрузить tsv" из выпадающего меню.



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



- 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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 40 из 67

Import Table Data

Choose the data import option below. Click here for more info on the table.

Data will be saved in Terra-managed location: us-central1 (Iowa)

FILE IMPORT TEXT IMPORT

Select the TSV file containing your data: Clear

Drag or Click to select a .tsv file

Selected File: manualUpload_metadata_240213_1.tsv

⚠ Data with the type 'quality_control' already exists in this workspace. Uploading more data for the same type may overwrite some entries.

TSV file templates

Download sample_template.tsv

Terra Support: Importing Data - Using a Template

CANCEL START IMPORT JOB

Upload selected data

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

WORKSPACES Workspaces > theiagen_pni/TheiaProk_PNI_Training_ETrees > Data

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

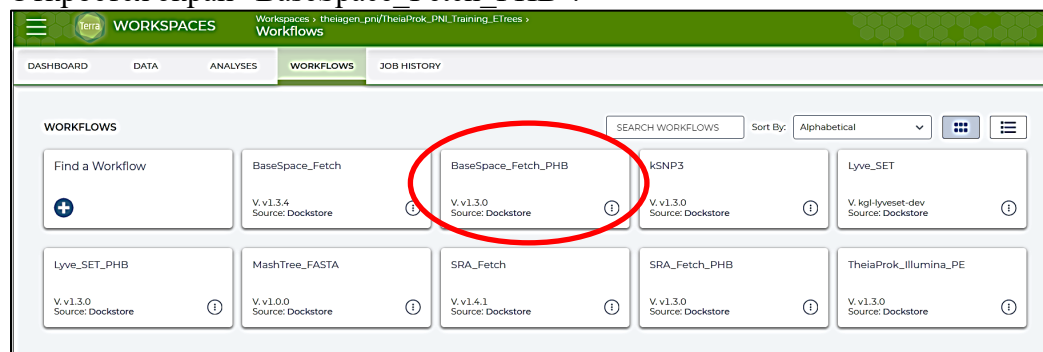
IMPORT DATA

EDIT OPEN WITH... EXPORT SETTINGS 0 rows selected ADVANCED SEARCH Search

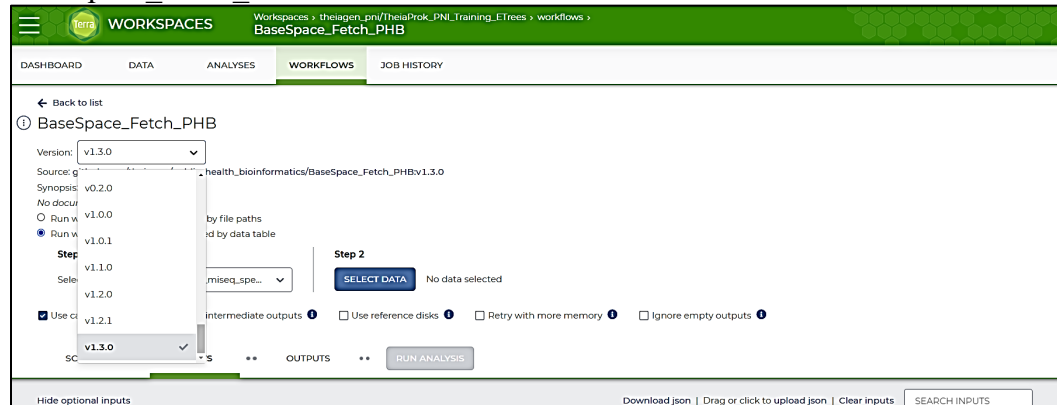
quality_control_id	basespace_collection_id	basespace_fetch_analysis_date	basespace_fetch_vers
D5480-LRM4update	M3235-23-042	2024-01-30	Terra Utilities v1.3.4
D7320-L-1A_USA_CDC_pci			
D7320-L-1B_USA_CDC_pci			
D7320-L-2A_USA_CDC_pci			
D7320-L-2B_USA_CDC_pci			
NVLV_QA-157	NVLV-M05339-20230508		
NVLV_QA-158	NVLV-M05339-20230508		
NVLV_QA-159	NVLV-M05339-20230508		
NVLV_QA-160	NVLV-M05339-20230508		
NVLV_QA-161	NVLV-M05339-20230508		

1 - 54 of 54 Items per page: 100

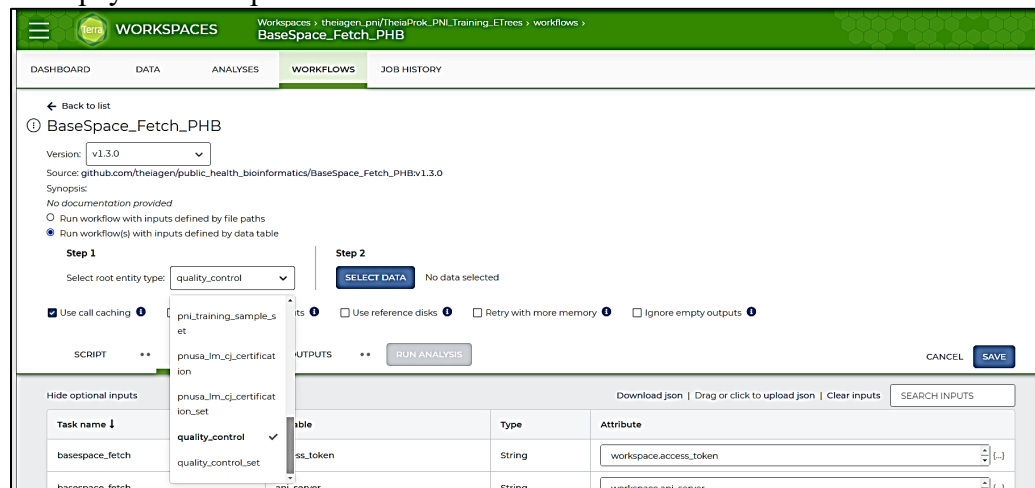
6. Запустите рабочий процесс "BaseSpace_Fetch_PHB":
- a. На вкладке "Рабочие процессы" нажмите на "BaseSpace_Fetch_PHB". Откроется экран "BaseSpace Fetch PHB".



- 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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 42 из 67

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. basespace_sample_name: "this.basespace_sample_name".
- v. sample_name: "**this.datatable_name_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.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 43 из 67

Task name ↓	Variable	Type	Attribute
basespace_fetch	access_token	String	workspace.access_token
basespace_fetch	api_server	String	workspace.api_server
basespace_fetch	basespace_collection_id	String	this.basespace_collection_id
basespace_fetch	basespace_sample_name	String	this.basespace_sample_name
basespace_fetch	sample_name	String	this.quality_control_id
basespace_fetch	basespace_sample_id	String	this.basespace_sample_id
fetch_bs	cpu	Int	Optional
fetch_bs	disk_size	Int	Optional

- 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.

Task name ↓	Variable	Type	Attribute Use defaults
basespace_fetch	basespace_fetch_analysis_date	String	this.basespace_fetch_analysis_date
basespace_fetch	basespace_fetch_version	String	this.basespace_fetch_version
basespace_fetch	read1	File	this.read1
basespace_fetch	read2	File	this.read2

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

Confirm launch

Output files will be saved as workspace data in:
 us us-central1 (lowa) ⓘ

Running workflows will generate cloud charges. ⓘ
 How much does my workflow cost? ⓘ
 Set up budget alert ⓘ

Describe your submission (optional):

MiSeq run M3235-23-042 from BaseSpace

This will launch **20** analyses.

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 44 из 67

The screenshot shows the Terra Workspaces interface for a job history entry. The submission is in a 'Queued' state. The table below lists the workflow steps:

Data Entity	Last Changed	Status	Run Cost	Messages	Workflow ID
NVLV_QA-157 (quality_control)	Feb 13, 2024, 12:25 PM	Queued	N/A		
NVLV_QA-158 (quality_control)	Feb 13, 2024, 12:25 PM	Queued	N/A		
NVLV_QA-159 (quality_control)	Feb 13, 2024, 12:25 PM	Queued	N/A		
NVLV_QA-160 (quality_control)	Feb 13, 2024, 12:25 PM	Queued	N/A		
NVLV_QA-161 (quality_control)	Feb 13, 2024, 12:25 PM	Queued	N/A		

- к. 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".

The screenshot shows the same Terra Workspaces job history entry, but now the submission is in a 'Succeeded' state. The table below lists the workflow steps with their completion status and links:

Data Entity	Last Changed	Status	Run Cost	Messages	Workflow ID	Links
NVLV_QA-157 (quality_control)	Feb 13, 2024, 12:30 PM	Succeeded	N/A		6154609d-c6dc-4d1d-a1c8-5286b71303..	📄 📁
NVLV_QA-158 (quality_control)	Feb 13, 2024, 12:29 PM	Succeeded	N/A		d3535af4-758a-490d-9be9-16696457800	📄 📁
NVLV_QA-159 (quality_control)	Feb 13, 2024, 12:29 PM	Succeeded	N/A		1e1351e2-5f82-4978-9f77-dec8a6e59472	📄 📁
NVLV_QA-160 (quality_control)	Feb 13, 2024, 12:29 PM	Succeeded	N/A		f561f71f-6ada-455c-8820-8f6fb54b4094	📄 📁
NVLV_QA-161 (quality_control)	Feb 13, 2024, 12:29 PM	Succeeded	N/A		3be09ce-a2e3-4302-af5a-01187ac5eb07	📄 📁

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 45 из 67

WORKSPACES Workspaces > theisgen_pnl/TheiaProk_PNL_Training_ETrees > Data

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

EDIT OPEN WITH... EXPORT SETTINGS 0 rows selected ADVANCED SEARCH Search

	quality_controlLid	sample_name	read1	read2
<input type="checkbox"/>	D5480-LRM4update	13235-23-042	D5480-LRM4update_R1.fastq.gz	D5480-LRM4update_R2.fastq.gz
<input type="checkbox"/>	D7320-L-1A_USA_CDC_pcl		D7320-L-1A-M3235-21-007_S12_L001_R1_001.fast...	D7320-L-1A-M3235-21-007_S12_L001_...
<input type="checkbox"/>	D7320-L-1B_USA_CDC_pcl		D7320-L-1B-M3235-21-007_S13_L001_R1_001.fast...	D7320-L-1B-M3235-21-007_S13_L001_...
<input type="checkbox"/>	D7320-L-2A_USA_CDC_pcl		D7320-L-2A-M3235-21-007_S14_L001_R1_001.fast...	D7320-L-2A-M3235-21-007_S14_L001_...
<input type="checkbox"/>	D7320-L-2B_USA_CDC_pcl		D7320-L-2B-M3235-21-007_S15_L001_R1_001.fast...	D7320-L-2B-M3235-21-007_S15_L001_...
<input type="checkbox"/>	NVLV_QA-157		NVLV_QA-157_R1.fastq.gz	NVLV_QA-157_R2.fastq.gz
<input type="checkbox"/>	NVLV_QA-158		NVLV_QA-158_R1.fastq.gz	NVLV_QA-158_R2.fastq.gz
<input type="checkbox"/>	NVLV_QA-159		NVLV_QA-159_R1.fastq.gz	NVLV_QA-159_R2.fastq.gz
<input type="checkbox"/>	NVLV_QA-160		NVLV_QA-160_R1.fastq.gz	NVLV_QA-160_R2.fastq.gz
<input type="checkbox"/>	NVLV_QA-161		NVLV_QA-161_R1.fastq.gz	NVLV_QA-161_R2.fastq.gz

1 - 54 of 54 Items per page: 100

WORKSPACES Workspaces > theisgen_pnl/TheiaProk_PNL_Training_ETrees > Data

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

EDIT OPEN WITH... EXPORT SETTINGS 0 rows selected ADVANCED SEARCH Search

	quality_controlLid	basespace_fetch_analysis_date	basespace_fetch_version	basespace_sample_name
<input type="checkbox"/>	D5480-LRM4update	2024-01-30	Terra Utilities v1.3.4	D5480-M3235-23-042
<input type="checkbox"/>	D7320-L-1A_USA_CDC_pcl			
<input type="checkbox"/>	D7320-L-1B_USA_CDC_pcl			
<input type="checkbox"/>	D7320-L-2A_USA_CDC_pcl			
<input type="checkbox"/>	D7320-L-2B_USA_CDC_pcl			
<input type="checkbox"/>	NVLV_QA-157	2024-02-13	PHB v1.3.0	QA-157
<input type="checkbox"/>	NVLV_QA-158	2024-02-13	PHB v1.3.0	QA-158
<input type="checkbox"/>	NVLV_QA-159	2024-02-13	PHB v1.3.0	QA-159
<input type="checkbox"/>	NVLV_QA-160	2024-02-13	PHB v1.3.0	QA-160
<input type="checkbox"/>	NVLV_QA-161	2024-02-13	PHB v1.3.0	QA-161

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.

	analysis...	read1	read2	sra_accession
<input type="checkbox"/>	SRR8643861	SRR8643861_1.fastq.gz	SRR8643861_2.fastq.gz	SRR8643861

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

Import Table Data

Choose the data import option below. Click here for more info on the table.

Data will be saved in Terra-managed location: **us-central1 (Iowa)**

FILE IMPORT TEXT IMPORT

Select the **TSV** file containing your data:

Drag or Click to select a .tsv file

Selected File: **None**

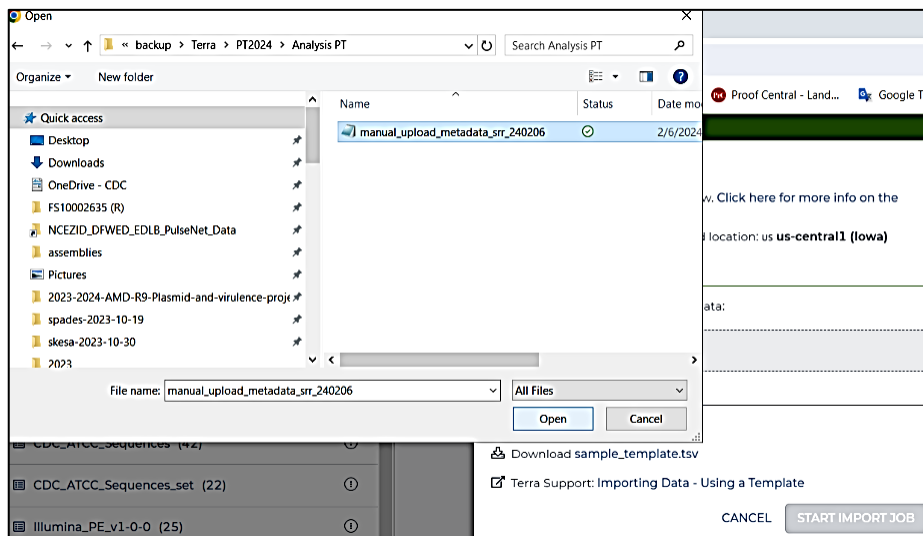
TSV file templates

Download sample_template.tsv

Terra Support: Importing Data - Using a Template

CANCEL START IMPORT JOB

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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 48 из 67

Import Table Data
Choose the data import option below. Click here for more info on the table.
Data will be saved in Terra-managed location: us-central1 (Iowa)
FILE IMPORT TEXT IMPORT
Select the TSV file containing your data: Clear
Drag or Click to select a .tsv file
Selected File: manual_upload_metadata_srr_240206.tsv
Warning: Data with the type 'analysis_pt_24' already exists in this workspace. Uploading more data for the same type may overwrite some entries.
TSV file templates
Download sample_template.tsv
 Terra Support: Importing Data - Using a Template
CANCEL START IMPORT JOB
Upload selected data

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.

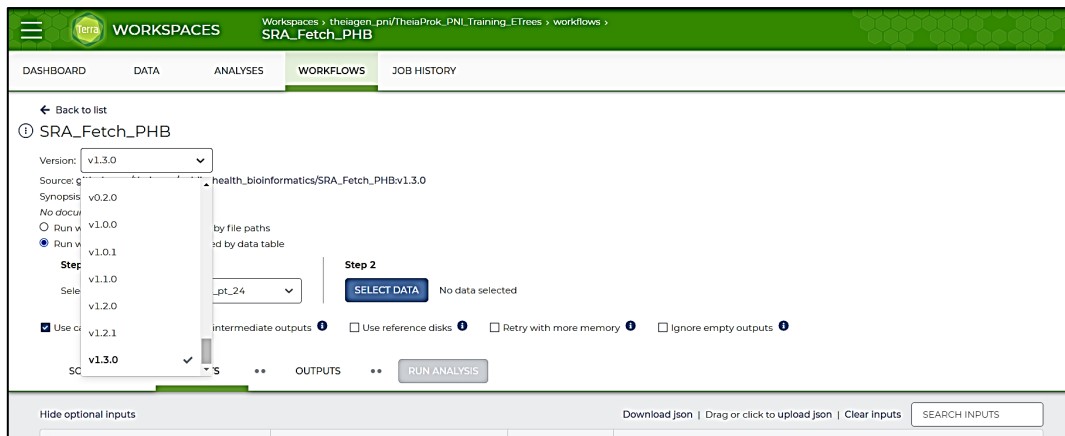
analysis_pt_24	read1	read2	sra_accession
2020K-0901			SRR12712421
2020K-1237			SRR13132213
PNUSAS0700...			SRR8786860
PNUSAS0705...			SRR8756231
PNUSAS0722...			SRR8878889
PNUSAS1768...			SRR12884674
PNUSAS1770...			SRR12885281
PNUSAS1882...			SRR13387034
PNUSAS1889...			SRR13438657

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

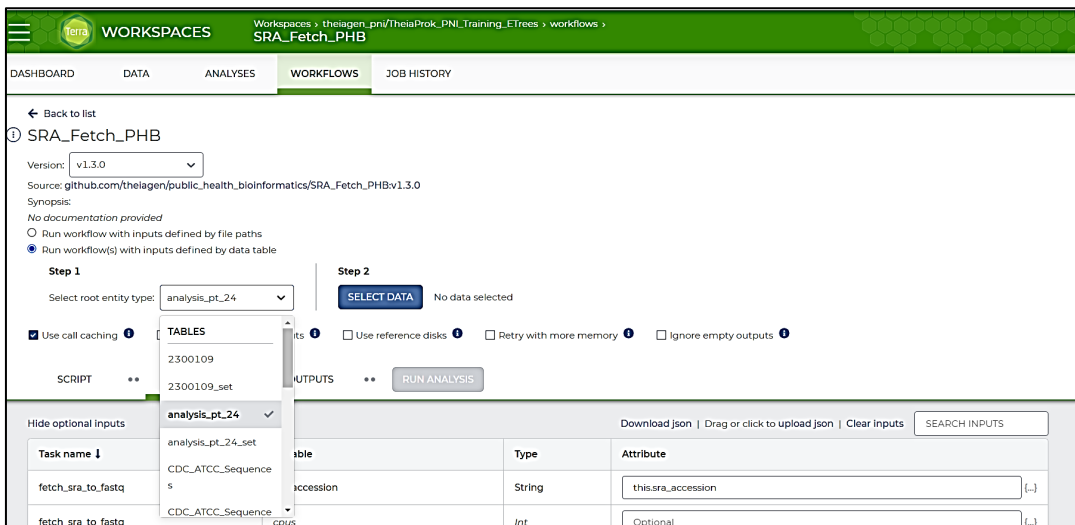
WORKFLOWS
SEARCH WORKFLOWS Sort By: Alphabetical

- Find a Workflow (+)
- BaseSpace_Fetch (V. v1.3.4 Source: Dockstore)
- BaseSpace_Fetch_PHB (V. v1.3.0 Source: Dockstore)
- kSNP3 (V. v1.3.0 Source: Dockstore)
- Lyve_SET (V. kgl-lyveset-dev Source: Dockstore)
- Lyve_SET_PHB (V. v1.3.0 Source: Dockstore)
- MashTree_FASTA (V. v1.0.0 Source: Dockstore)
- SRA_Fetch (V. v1.4.1 Source: Dockstore)
- SRA_Fetch_PHB (V. v1.3.0 Source: Dockstore)**
- TheiaProk_Illumina_PE (V. v1.3.0 Source: Dockstore)

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



- 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.



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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 50 из 67

Choose specific analysis_pt_24s to process

Select analysis_pt_24s to process **SETTINGS** 9 rows selected **ADVANCED SEARCH** Search

<input type="checkbox"/>	analysis...	data	fastq_dl_version	read1	read2	sra_accession
<input checked="" type="checkbox"/>	PNUSAS0700...					SRR8786860
<input checked="" type="checkbox"/>	PNUSAS0705...					SRR8756231
<input checked="" type="checkbox"/>	PNUSAS0722...					SRR8878889
<input checked="" type="checkbox"/>	PNUSAS1768...					SRR12884674
<input checked="" type="checkbox"/>	PNUSAS1770...					SRR12885281
<input checked="" type="checkbox"/>	PNUSAS1882...					SRR13387034
<input checked="" type="checkbox"/>	PNUSAS1889...					<DD13424657

1 - 10 of 10 **1** Items per page: 100

Selected analysis_pt_24s will be saved as a new analysis_pt_24_set named:

SRA_Fetch_PHB_2024-02-06T17:51-13

CANCEL **OK**

12. Снимите флажок "Use call caching".

Run workflow with inputs defined by file paths

Run workflow(s) with inputs defined by data table

Step 1 Select data table: analysis_pt_24 **Select Data** 9 selected analysis_pt_24s (will create a new analysis_pt_24_set named "SRA_Fetch_PHB_2024-05-2...")

Use call caching **i** Delete intermediate outputs **i** Use reference disks **i** Retry with more memory **i** Ignore empty outputs **i**

SCRIPT **INPUTS** OUTPUTS **Run Analysis**

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

Step 1 Select root entity type: analysis_pt_24 **SELECT DATA** 9 selected analysis_pt_24s (will create a new analysis_pt_24_set named "SRA_Fetch_PHB_2024-02-06T17:51-13")

Use call caching **i** Delete intermediate outputs **i** Use reference disks **i** Retry with more memory **i** Ignore empty outputs **i**

SCRIPT **INPUTS** OUTPUTS **RUN ANALYSIS**

Hide optional inputs **Download json** | **Drag or click to upload json** | **Clear inputs** **SEARCH INPUTS**

Task name ↓	Variable	Type	Attribute
fetch_sra_to_fastq	sra_accession	String	this.sra_accession [-]
fetch_sra_to_fastq	cpus	Int	Optional [-]
fetch_sra_to_fastq	disk_size	Int	Optional [-]

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.

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 51 из 67

Step 1
Select root entity type: analysis_pt_24

Step 2
SELECT DATA 9 selected analysis_pt_24s (will create a new analysis_pt_24_set named "SRA_Fetch_PHB_2024-02-06T17-51-13")

Use call caching Delete intermediate outputs Use reference disks Retry with more memory Ignore empty outputs

SCRIPT ** INPUTS ** **OUTPUTS** ** **RUN ANALYSIS**

Output files will be saved to
Files / submission unique ID / fetch_sra_to_fastq / workflow unique ID

References to outputs will be written to
Tables / analysis_pt_24

Fill in the attributes below to add or update columns in your data table

Task name ↓	Variable	Type	Attribute	Use defaults
fetch_sra_to_fastq	fastq_dl_date	String	this.fastq_dl_date	<input type="checkbox"/>
fetch_sra_to_fastq	fastq_dl_docker	String	this.fastq_dl_docker	<input type="checkbox"/>
fetch_sra_to_fastq	fastq_dl_fastq_metadata	File	this.fastq_dl_fastq_metadata	<input type="checkbox"/>

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

Confirm launch

Output files will be saved as workspace data in:
us us-central1 (lowa)

Running workflows will generate cloud charges.
How much does my workflow cost?
Set up budget alert

Describe your submission (optional):
Analysis PT candidate set 1 (Newport)

This will launch 9 analyses.

CANCEL LAUNCH

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

WORKSPACES
Job History

DASHBOARD DATA ANALYSES WORKFLOWS **JOB HISTORY**

← Job History › Submission 86b637b7-edfb-4fbe-8b9e-d5237733fadc

Workflow Statuses Submitted: 9	Workflow Configuration theiagen_pni/SRA_Fetch_PHB	Submitted by eija.trees@theiagen.cloud Feb 6, 2024, 12:56 PM	Total Run Cost N/A
Comment Analysis PT candidate set 1 (Newport)	Data Entity SRA_Fetch_PHB_2024-02-06T17-51-13 analysis_pt_24_set	Submission ID 86b637b7-edfb-4fbe-8b9e-d5237733fadc	Call Caching Enabled
	Delete Intermediate Outputs Disabled	Use Reference Disks Disabled	Retry with More Memory Disabled

WORKFLOWS INPUTS OUTPUTS

Search workflows Completion status Download TSV

Data Entity ↓	Last Changed	Status	Run Cost	Messages	Workflow ID	Links
2020K-0901 (analysis_pt_24)	Feb 6, 2024, 12:56 PM	⌚ Queued	N/A			
2020K-1237 (analysis_pt_24)	Feb 6, 2024, 12:56 PM	⌚ Queued	N/A			
PNUSA5070021 (analysis_pt_24)	Feb 6, 2024, 12:56 PM	⌚ Queued	N/A			

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 52 из 67

WORKSPACES Workspaces > thelagen_pnl/TheiaProk_PNI_Training_ETrees > Job History

DASHBOARD DATA ANALYSES WORKFLOWS **JOB HISTORY**

Submission (click for details) Data entity No. of Workflows Status Submitted Submission ID Comment Actions

SRA_Fetch_PHB Submitted by elja.trees@thelagen.cloud	SRA_Fetch_PHB_2024-0...	9	Done	Feb 6, 2024 12:56 PM	86b637b7-edfb-4be-8b9e-d5237733fadc	Analysis PT candidate set ...	
BaseSpace_Fetch_PHB Submitted by curtis.kapsak@thelagen.com	SRA_Fetch_PHB_2024-02-06T17:51:13 (analysis_pt_24_set)		Done	Feb 2, 2024 3:15 PM	d7a05e8e-100b-4977-ae6b5-4de5baf9a8a	test on OC miseq run whe...	

WORKSPACES Workspaces > thelagen_pnl/TheiaProk_PNI_Training_ETrees > Data

DASHBOARD **DATA** ANALYSES WORKFLOWS JOB HISTORY

IMPORT DATA

EDIT OPEN WITH... EXPORT SETTINGS 0 rows selected ADVANCED SEARCH

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)

	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

Select columns

Show: all | none Sort: alphabetical

agrvate_summary

agrvate_version

combined_mean_q_clean

combined_mean_q_raw

combined_mean_readlength_clean

combined_mean_readlength_raw

meningotype_BAST

meningotype_FetA

meningotype_NHBA

meningotype_NadA

meningotype_PorA

meningotype_PorB

meningotype_fHbp

...

SAVE THIS COLUMN SELECTION

Your saved column selections:

pulsenet_genotyping ⓘ

qc_metrics ⓘ

CANCEL DONE

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

The screenshot shows the 'Select columns' dialog box. On the left, there is a list of columns with checkboxes. The columns 'ani_highest_percent' and 'ani_top_species_match' are selected. On the right, there is a 'Save this column selection' section with a text input field containing 'qc_metrics'. Below the input field, there is a 'SAVE' button. A tooltip points to the 'SAVE' button with the text 'Save this column selection'. At the bottom right, there are 'CANCEL' and 'DONE' buttons.

Изменение существующего выбора столбцов

The first screenshot shows the 'Select columns' dialog box with the 'qc_metrics' selection. The 'Column selection name' field contains 'qc_metrics'. The 'UPDATE' button is visible. The second screenshot shows the same dialog box after the selection has been updated. The 'Column selection name' field now contains 'qc_metrics' with a dropdown arrow and a close button. The 'UPDATE' button is still visible. Both screenshots show the list of columns on the left and the 'CANCEL' and 'DONE' buttons at the bottom.

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

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

¹После обрезки на основе качества (est_coverage_clean)

²До обрезки (combined_mean_q_raw)

Приложение PNID01-4б. 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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 56 из 67

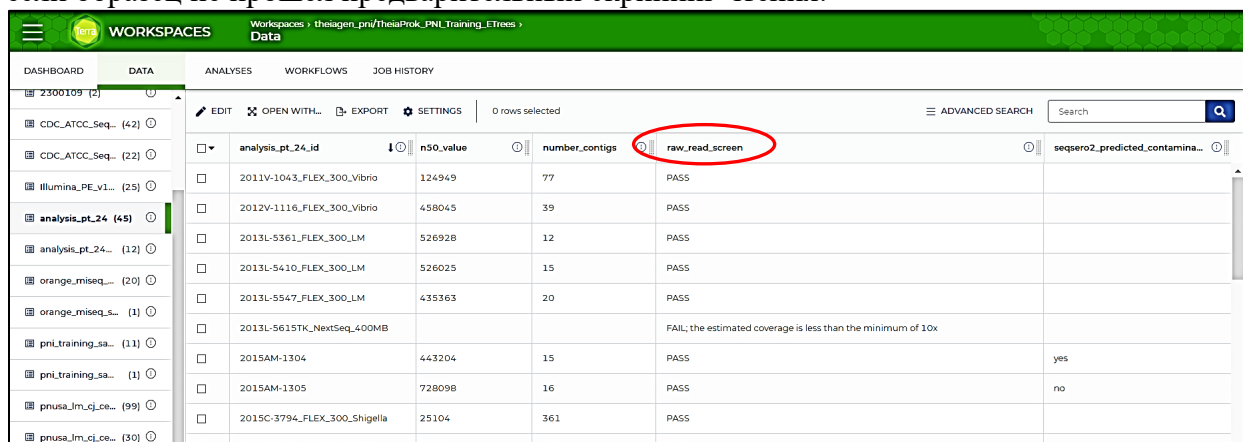
Int min_genome_length = 100000

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:



analysis_pt_24_id	n50_value	number_contigs	raw_read_screen	seqsere2_predicted_contamina...
2011V-1043_FLEX_300_Vibrio	124949	77	PASS	
2012V-1116_FLEX_300_Vibrio	458045	39	PASS	
2013L-5361_FLEX_300_LM	526928	12	PASS	
2013L-5410_FLEX_300_LM	526025	15	PASS	
2013L-5547_FLEX_300_LM	435363	20	PASS	
2013L-5615TK_NextSeq_400MB			FAIL; the estimated coverage is less than the minimum of 10x	
2015AM-1304	443204	15	PASS	yes
2015AM-1305	728098	16	PASS	no
2015C-3794_FLEX_300_Shigella	25104	361	PASS	

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

Select columns

Show: all | none Sort: alphabetical

- 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
- sister_allele_fasta
- sistr_allele_ison

SAVE THIS COLUMN SELECTION

Your saved column selections:

pulsenet_genotyping ⓘ

qc_metrics ⓘ

CANCEL DONE

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".

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 58 из 67

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

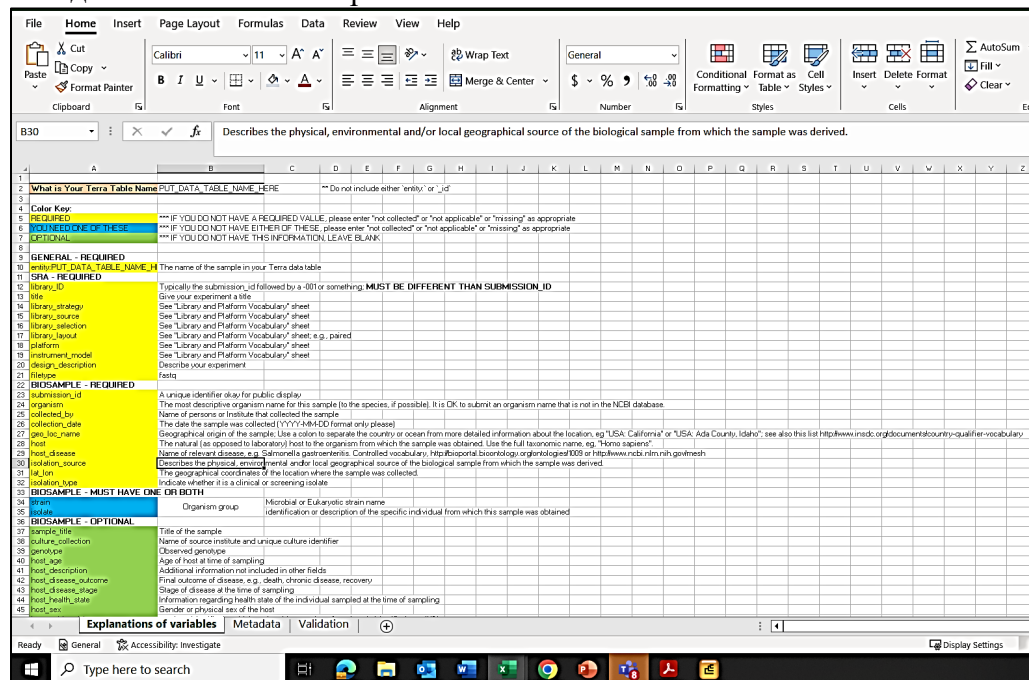
✕

Save this column selection selections:

ⓘ

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 60 из 67

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

letype	submission_id	organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	strain	isolate	sample_title	culture_c
astq	2017C-4936	Escherichia coli	CDC	2017-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2017C-4936			
astq	2018C-4039	Escherichia coli	CDC	2018-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2018C-4039			
astq	2019C-3204	Shigella sonnei	CDC	2018-01-01	USA	Homo sapiens	Missing	Missing	Missing	clinical	2019C-3204			

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 1 из 67

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

Клинические (человеческого происхождения) и неклинические последовательности

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	enthy-quality_control_id	library_id	title	library_strategy	library_source	library_selection	library_layout	platform	instrument_model	design_description	filetype	submission_id	organism
2	2019C-3238	2019C-3238-001	PulseNet_WGS	GENOMIC	RANDOM	paired	ILLUMINA	illumina	HiSeq 2500	Missing	fastq	2019C-3238	Escherichia coli
3	2015C-3887	2015C-3887-001	PulseNet_WGS	GENOMIC	RANDOM	paired	ILLUMINA	illumina	HiSeq 2500	Missing	fastq	2015C-3887	Escherichia coli
4	2017C-4936	2017C-4936-001	PulseNet_WGS	GENOMIC	RANDOM	paired	ILLUMINA	illumina	HiSeq 2500	Missing	fastq	2017C-4936	Escherichia coli
5	2013L-5357-LRM4update	2013L-5357-LRM4update-001	PulseNet_WGS	GENOMIC	RANDOM	paired	ILLUMINA	illumina	MiSeq	Missing	fastq	2013L-5357-LRM4update	Listeria monocytogenes

	M	N	O	P	Q	R	S	T	U	V	W
1	organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	istrain	isolate
2	Escherichia coli	CDC	2017-01-01	USA	Missing	Missing	Missing	Missing	clinical	2017C-4936	
3	Escherichia coli	CDC	2018-01-01	USA	Missing	Missing	Missing	Missing	clinical	2018C-4039	
4	Escherichia coli	CDC	2018-01-01	USA	Missing	Missing	Missing	Missing	clinical	2019C-3204	
5	Listeria monocytogenes	CDC	2013-07-01	USA:WI	Missing	Missing	cheese	Missing	food	2013L-5357-LRM4update	

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

Row #	Are the library_id and submission_id different?	Do all required fields have values?	Does geo_loc_name have the proper format?	Is the collection_date valid?	Does either the strain or isolate fields have...
2	Yes	Yes	No	Yes	Yes
3	Yes	Yes	No	Yes	Yes
4	Yes	Yes	No	Yes	Yes

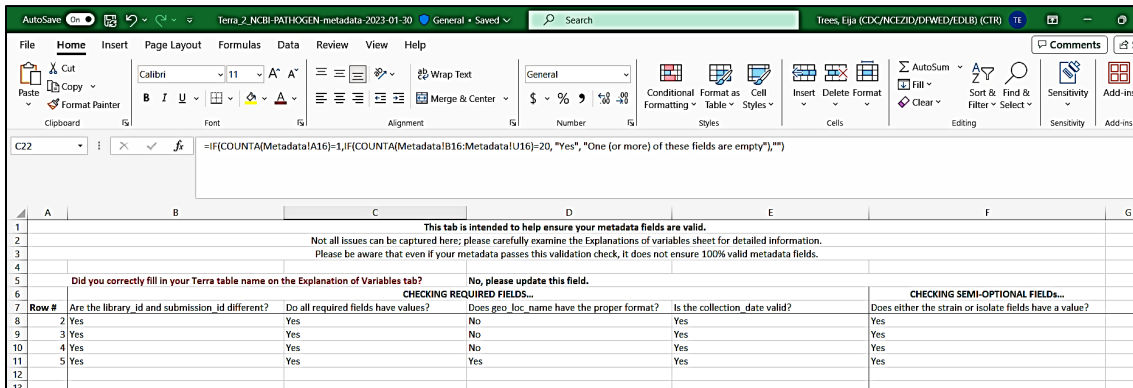
МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

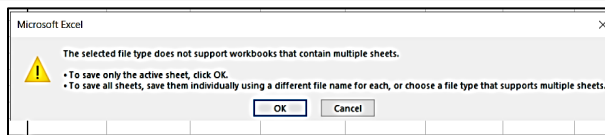
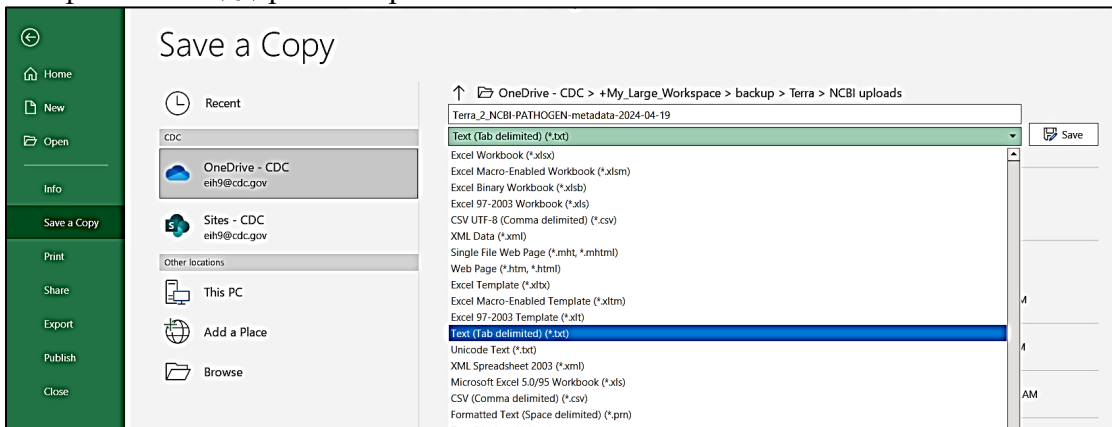
Дата вступления в силу:

Страница 62 из 67



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.

- 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.



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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 63 из 67

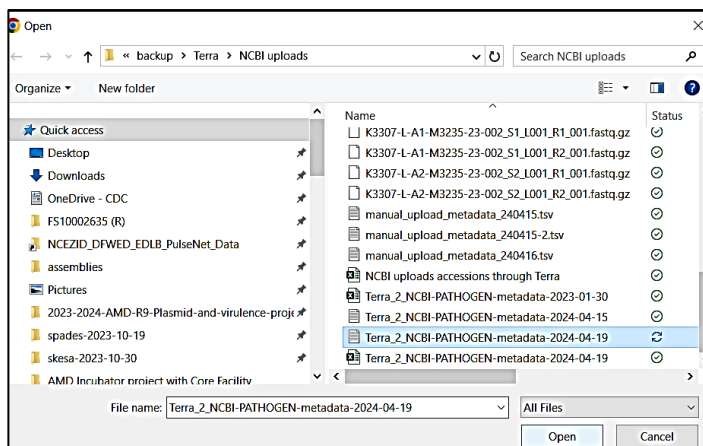
The screenshot shows the Terra Workspaces interface. The top navigation bar includes 'WORKSPACES' and 'Data'. Below it are tabs for 'DASHBOARD', 'DATA', 'ANALYSES', 'WORKFLOWS', and 'JOB HISTORY'. The 'DATA' tab is active, showing an 'IMPORT DATA' button and a table with 1 row selected. The table has columns for 'quality_control_id' and 'read2'. The table content is as follows:

quality_control_id	read2
2017C-4938	2017C-4938-L002R1
2017C-4938	2017C-4938-D002R1
2018C-4039	2018C-4039-D002R1
2018C-4709-L-A1-USA-CDC-pcl	2018C-4709-L-A1-M3235-23-002-S1-L001-R1-001.fastq.gz
2018C-4709-L-A2-USA-CDC-pcl	2018C-4709-L-A2-M3235-23-002-S2-L001-R1-001.fastq.gz

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

The screenshot shows the 'Import Table Data' dialog box. It has a title 'Import Table Data' and a subtitle 'Choose the data import option below. Click here for more info on the table.' Below this, it says 'Data will be saved in Terra-managed location: us-central1 (lowa)'. There are two tabs: 'FILE IMPORT' (selected) and 'TEXT IMPORT'. Under 'FILE IMPORT', it says 'Select the TSV file containing your data:' and there is a dashed box with the text 'Drag or Click to select a .tsv file'. Below this, it says 'Selected File: None'. There are also 'TSV file templates' with options to 'Download sample_template.tsv' and 'Terra Support: Importing Data - Using a Template'. At the bottom, there are 'CANCEL' and 'START IMPORT JOB' buttons.

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 64 из 67

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

Import Table Data
Choose the data import option below. Click here for more info on the table.
Data will be saved in Terra-managed location: us-us-central1(iowa)

FILE IMPORT TEXT IMPORT

Select the TSV file containing your data: Clear

Drag or Click to select a .tsv file

Selected File: Terra_2_NCBIPATHOGEN-metadata-2024-04-19.txt

⚠ Data with the type 'quality_control' already exists in this workspace. Uploading more data for the same type may overwrite some entries.

⚠ We have detected empty cells in your TSV. Please choose an option:

- Ignore empty cells (default)
- Overwrite existing cells with empty cells

TSV file templates

Download sample_template.tsv

Terra Support: Importing Data - Using a Template

CANCEL START IMPORT JOB

Upload selected data

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

quality_controlId	collected_by	collection_date	design_description	filetype
2017C-3830-LRM4update				
2017C-4936	CDC	2017-01-01	Missing	fastq
2017C-4938				
2018C-4039	CDC	2018-01-01	Missing	fastq
2018C-4709-L-A1-USA-CDC-pcl				
2018C-4709-L-A2-USA-CDC-pcl				
2018C-4709-L-B1-USA-CDC-pcl				
2018C-4709-L-B2-USA-CDC-pcl				
2018EL-1053a-L-A1-USA_CDC_pcl				
2018EL-1053a-L-A2-USA_CDC_pcl				
2018EL-1053a-L-B1-USA_CDC_pcl				
2018EL-1053a-L-B2-USA_CDC_pcl				
2019C-3204	CDC	2018-01-01	Missing	fastq
2019C-3238				
92-01-L-A1	CDC	Missing	Missing	fastq
92-01-L-A2				

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

The screenshot shows a 'Select columns' dialog box. On the left, there is a list of columns with checkboxes. The 'Sort' dropdown is set to 'alphabetical'. A 'SAVE THIS COLUMN SELECTION' button is visible. On the right, under 'Your saved column selections:', the following items are listed: Metadata, pulsenet_genotyping, and qc_metrics. At the bottom right, there are 'CANCEL' and 'DONE' buttons.

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

МЕЖДУНАРОДНАЯ СТАНДАРТНАЯ ОПЕРАЦИОННАЯ ПРОЦЕДУРА PULSENET ДЛЯ АНАЛИЗА ДАННЫХ ILLUMINA SHORT READ WGS С ИСПОЛЬЗОВАНИЕМ ПЛАТФОРМЫ TERRA.BIO

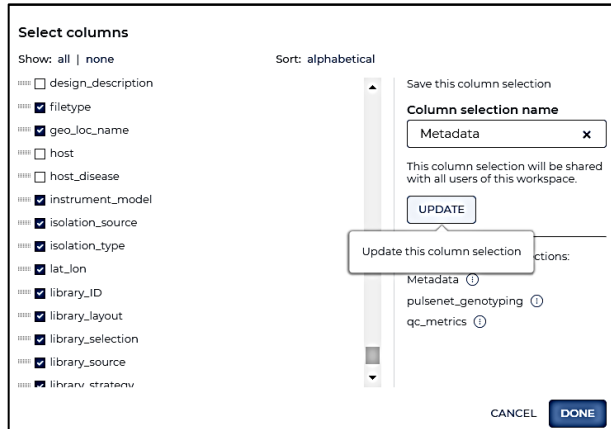
Док. № PNID01

Вер. № 01

Дата вступления в силу:

Страница 66 из 67

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.

quality_control_id	sample_accession	collected_by	collection_date	filetype	geo
2017C-4936	1039458	CDC	2017-01-01	fastq	USA
2017C-4938					
2018C-4039	1039457	CDC	2018-01-01	fastq	USA
2018C-4709-L-A1-USA-CDC-pcl					
2018C-4709-L-A2-USA-CDC-pcl					
2018C-4709-L-B1-USA-CDC-pcl					
2018C-4709-L-B2-USA-CDC-pcl					
2018EL-1053a-L-A1_USA_CDC_pcl					
2018EL-1053a-L-A2_USA_CDC_pcl					
2018EL-1053a-L-B1_USA_CDC_pcl					
2018EL-1053a-L-B2_USA_CDC_pcl					
2019C-3204	1039456	CDC	2017-01-01	fastq	USA
2019C-3238					

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

