ЗМІСТ - ГІПЕРПОСИЛАННЯ НА ПРОЦЕДУРИ

1. **МЕТА:** Описати процедуру аналізу даних повногеномного секвенування (WGS) з коротким зчитуванням Illumina, які будуть використані для спостереження PulseNet International (PNI) з використанням хмарної платформи Terra.Bio.

2. **СФЕРА ЗАСТОСУВАННЯ:** Дана процедура поширюється на весь персонал PulseNet, який використовує платформу Terra.Bio для аналізу даних секвенатора Illumina для короткого зчитування WGS в рамках діяльності з нагляду в міжнародній мережі PulseNet. Дана СОП охоплює завантаження послідовностей і метаданих до Terra.Bio, оцінку якості послідовностей, робочі процеси збірки і генотипування, а також завантаження послідовностей до NCBI. Філогенетичний аналіз охоплюється СОП PNID02 (PulseNet International Standard Operating Procedure for Phylogenetic Analysis of WGS Data Using the Terra.Bio Platform - Стандартна операційна процедура для філогенетичного аналізу даних WGS з використанням платформи Terra.Bio).

3. **ВИЗНАЧЕННЯ/ТЕРМІНИ:**
   - **3.1 ANI:** Середня нуклеотидна ідентичність.
   - **3.2 BaseSpace:** Хмарне обчислювальне середовище Illumina для аналізу, управління та зберігання даних секвенування наступного покоління, включаючи обмін даними.

3.3 Команди Bash: Bash **(**Bourne Again Shell) - це оболонка інтерфейсу командного рядка (CLI), яка широко використовується в Linux і macOS. Оболонка - це комп'ютерна програма, яка дозволяє безпосередньо керувати операційною системою комп'ютера. Команди Bash використовуються для керування комп'ютером або операційною системою без необхідності навігації по меню, опціях і вікнах у графічному інтерфейсі користувача.

3.4 Біопроєкт: Колекція біологічних даних про NCBI, пов'язаних з однією ініціативою, що походить від однієї організації або консорціуму.

3.5 Біозразки: Надана описова інформація (метадані) про біологічні матеріали, з яких отримані дані, що зберігаються в NCBI.

3.6 Contig: Суцільна консенсусна послідовність, отримана в результаті складання багатьох коротких фрагментів ДНК, що перекриваються.

3.7 Покриття: Середня кількість зчитувань, які включають певний нуклеотид у реконструйовану послідовність.

3.8 Критичні показники якості: покриття (після тримісії), середня якість (показник Q до тримісії), довжина збірки та чисельність вторинних родів (виявлення забруднення за допомогою MIDAS). Послідовності, які не відповідають мінімальним порогам/прийнятним діапазонам для цих метрик, визначених у цьому документі, слід повторно секвенувати.

3.9 CSV: Значення, розділені комами.

3.10 Збірка DeNovo: Збірка послідовностей, створена з коротких необроблених зчитувань без використання референтного геному.

3.11 FASTA: текстовий формат для представлення послідовностей нуклеотидів або пептидів, в якому пари основ або амінокислот представлені за допомогою однолітерних кодів. Послідовність у форматі FASTA починається з однорядкового опису, за яким слідують рядки даних послідовності. Рядок опису відрізняється від даних послідовності символом "більше" (">") у першому стовпчику.

3.12 FASTQ: текстовий формат для зберігання як біологічної послідовності, так і відповідних оцінок якості.

3.13 GAMBIT: Метод геномної апроксимації для ідентифікації та відстеження бактерій. Метод ідентифікації бактеріальних видів, що використовує алгоритм на основі k-мер для пошуку у великій довідковій базі даних геномів.

3.14 Gzip: Формат файлів і програмне забезпечення, що використовується для стиснення та розпакування файлів для швидкої передачі даних через Інтернет.

3.15 LIMS: Система управління лабораторною інформацією.

3.16 Mash Sketching: Mash - це набір інструментів для створення і використання скетчів MinHash, спосіб перетворення геному в невеликий підпис, який можна легко порівняти з іншими підписами.

3.17 Метадані: Набір даних, які описують та надають інформацію про інші дані.

3.18 MIDAS: Метагеномна система аналізу внутрішньовидового різноманіття. Інтегрований обчислювальний конвеєр для кількісної оцінки чисельності бактеріальних видів та покриття метагеномів на основі бласт-вирівнювання з панеллю універсальних однокопійних генів.

3.19 N50: Статистика N50 зазвичай використовується для грубої оцінки геномних збірок. Вона показує довжину контига (в базових парах), для якої половина послідовності геному зібрана в контиги, розмір яких більший або дорівнює розміру контига N50.
3.20 NCBI: Національний центр біотехнологічної інформації.
3.21 PNI: PulseNet International.
3.22 QA/QC: Забезпечення якості/контроль якості.
3.23 Оцінка якості: Оцінка якості для кожної окремої позиції бази в послідовності, що вказує на точність розпізнавання нуклеотидних основ. Використовуються оцінки Phred, де Q = -10log (ймовірність помилки). Чим вищий показник якості, тим надійнішим є показник розпізнавання. Q30 означає, що ймовірність помилкового розпізнавання в цій позиції становить 1 до 1000.
3.24 Read: Одиниця безперервної послідовності ДНК (пар нуклеотидів), отримана шляхом секвенування частини фрагментованої ДНК-мішені.
3.25 Номер SAMN: Унікальний ідентифікатор NCBI (номер доступу) для біозразка послідовності (метадані).
3.26 СОП: Стандартна операційна процедура.
3.27 SRA: Архів послідовного читання.
3.28 Номер SRR: Унікальний ідентифікатор NCBI (номер приєднання) для завантажених зчитувань сирої послідовності.
3.29 Terra.Bio: Хмарна платформа для аналізу послідовностей, розроблена Інститутом Броуда Массачусетського технологічного інституту та Гарвардським університетом і використовувана компанією Theiagen Genomics (Highlands Ranch, Колорадо, США), щоб запропонувати платну спільну платформу для лабораторій громадського здоров'я для розміщення, аналізу та обміну даними (Libuit *et al.,* 2023).
3.30 TSV: Значення, розділені табуляцією.
3.31 ПГС: **Повногеномне секвенування** всього геному.

4. ОБОВ'ЯЗКИ:

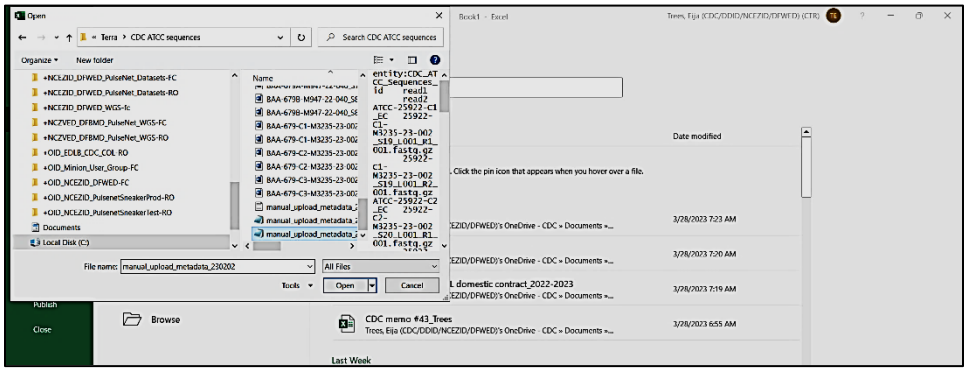
4.1 Персонал PulseNet виконує оцінку якості та генотипування коротких послідовностей зчитування Illumina, створених для спостереження PulseNet International з використанням платформи Terra.Bio. Наполегливо рекомендується ділитися даними WGS з іншими учасниками PNI шляхом завантаження в NCBI.

5. ПРОЦЕДУРА:

ПРИМІТКА: *Дані послідовностей можна імпортувати в Terra одним з трьох способів:* ***(1)*** *завантаження з локального мережевого сховища (кроки 5.1 - 5.3),* ***(2)*** *пряма передача з хмари в хмару з Illumina BaseSpace (додаток PNID01-1),* ***(3)*** *завантаження з NCBI SRA (додаток PNID01-2)*

5.1 **Підготуйте файл метаданих для послідовностей, які будуть завантажені на Terra** . Файл метаданих пов'язуватиме завантажені FASTQ-файли з відповідними іменами записів у базі даних, тобто ключами записів.

5.1.1 Використовуйте Excel, щоб відкрити шаблон файлу метаданих у форматі tsv.



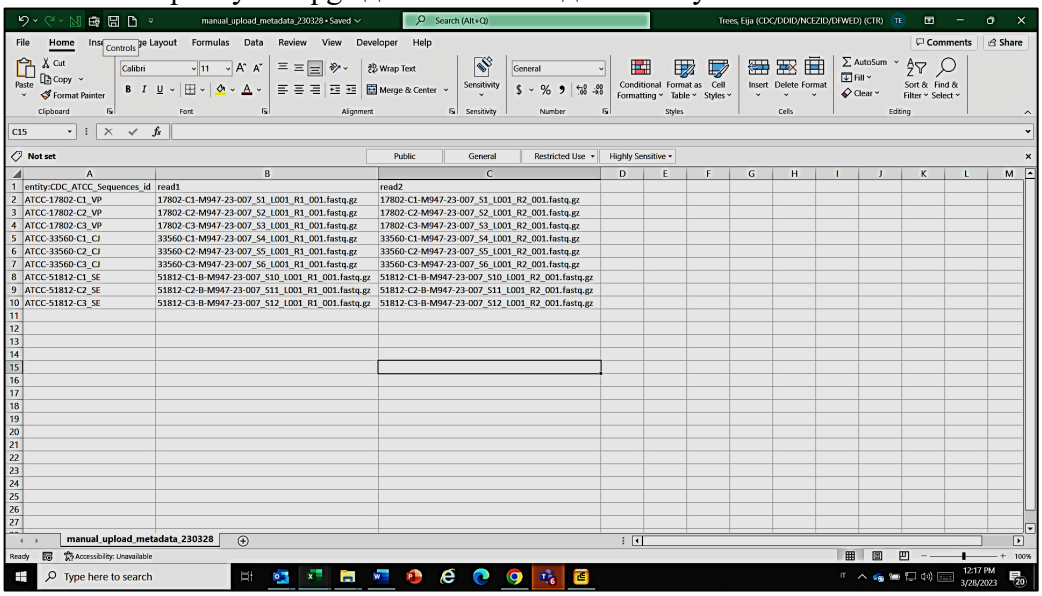
5.1.2 Наступні три колонки є обов'язковими:

5.1.2.1 **Entity:collection_name_id**. Це назва колекції даних Terra (таблиці даних), до якої ви хочете завантажити ваші послідовності, наприклад, entity:CDC_ATCC_Sequences_id. Ідентифікатори штамів (ключі для входу) в Terra називаються "сутностями".

ПРИМІТКА: *пробіли та тире не допускаються. "entity:" та "_id" є обов'язковими для Terra в назві стовпця.*

5.1.2.2 **Read1.** Ім'я файлу fastq.gz для читання 1 запису.

5.1.2.3 **Read2.** Ім'я файлу fastq.gz для читання 2 для запису.



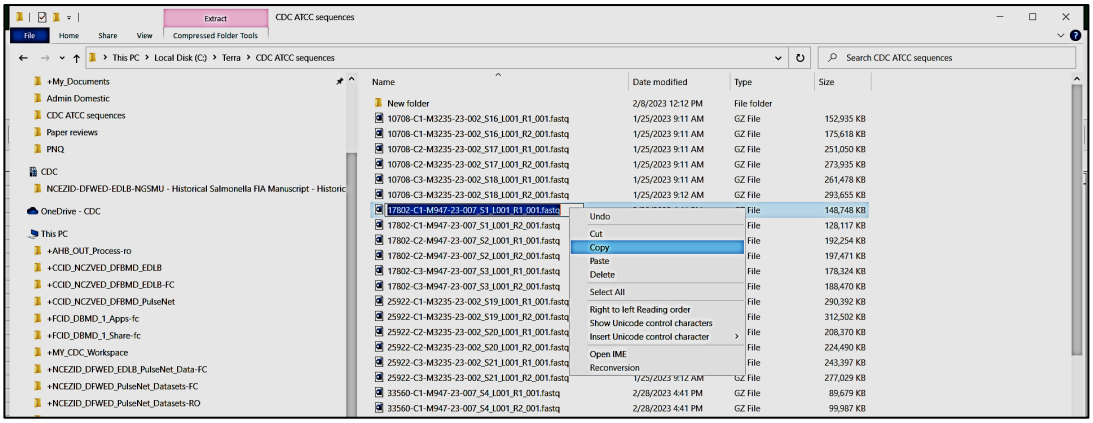
	A	B	C
1	entity:CDC_ATCC_Sequences_id	read1	read2
2	ATCC-17802-C1_VP	17802-C1-M947-23-007_S1_L001_R1_001.fastq.gz	17802-C1-M947-23-007_S1_L001_R2_001.fastq.gz
3	ATCC-17802-C2_VP	17802-C2-M947-23-007_S2_L001_R1_001.fastq.gz	17802-C2-M947-23-007_S2_L001_R2_001.fastq.gz
4	ATCC-17802-C3_VP	17802-C3-M947-23-007_S3_L001_R1_001.fastq.gz	17802-C3-M947-23-007_S3_L001_R2_001.fastq.gz
5	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
6	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
7	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
8	ATCC-51812-C1_SE	51812-C1-B-M947-23-007_S10_L001_R1_001.fastq.gz	51812-C1-B-M947-23-007_S10_L001_R2_001.fastq.gz
9	ATCC-51812-C2_SE	51812-C2-B-M947-23-007_S11_L001_R1_001.fastq.gz	51812-C2-B-M947-23-007_S11_L001_R2_001.fastq.gz
10	ATCC-51812-C3_SE	51812-C3-B-M947-23-007_S12_L001_R1_001.fastq.gz	51812-C3-B-M947-23-007_S12_L001_R2_001.fastq.gz

ПРИМІТКА: *додаткові метадані можуть бути додані на цьому етапі або пізніше. Зверніться до додатку PNID01-6 для отримання інструкцій щодо метаданих.*

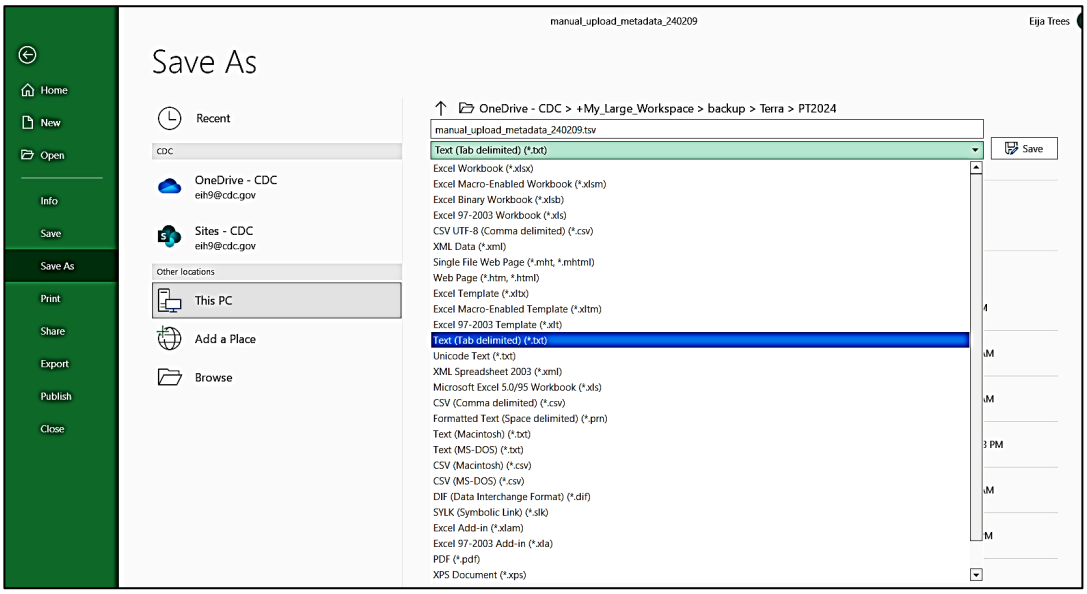
5.1.3 Введіть ідентифікатори штамів у колонці "entity/об'єкт" так, як ви хочете, щоб вони відображалися в Terra.

ПРИМІТКА: *ідентифікатор штаму не обов'язково повинен збігатися з будь-якою частиною імені файлу fastq.gz.*

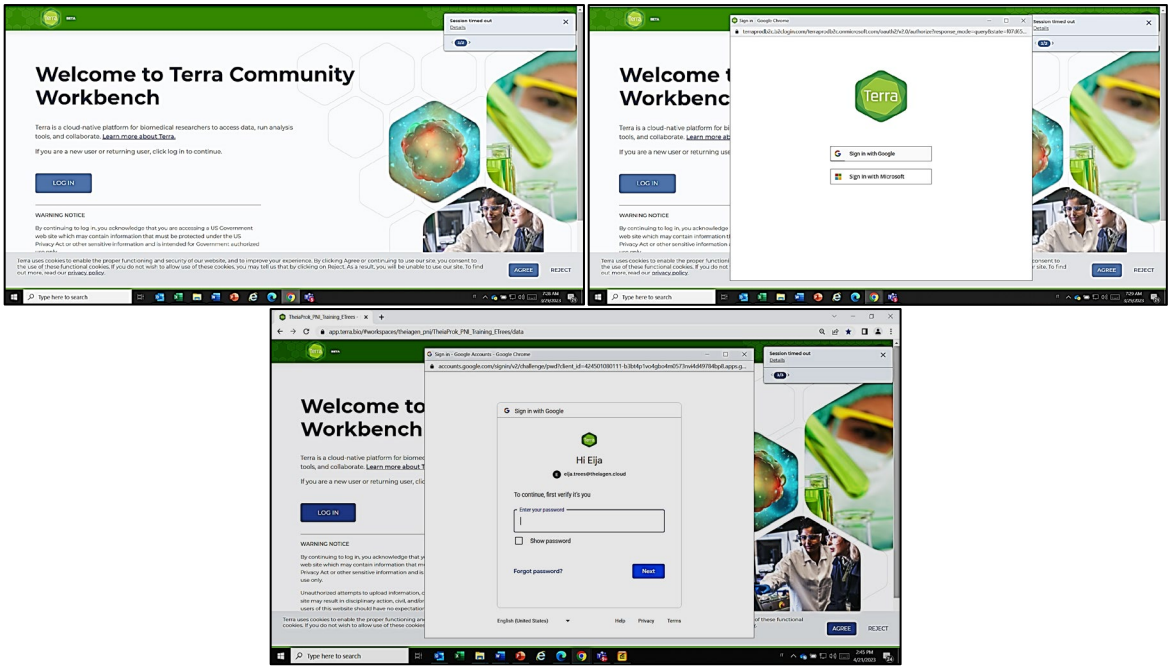
5.1.4 Скопіюйте та вставте імена файлів fastq.gz для кожного штаму в колонки "read1" та "read2". Переконайтеся, що імена файлів закінчуються на "fastq.gz".



5.1.5 Збережіть файл у **форматі tsv**: виберіть "Зберегти як" і "Текст (з розділенням табуляцією) (*.txt)". Переконайтеся, що ім'я вашого файлу закінчується на "**.tsv**".

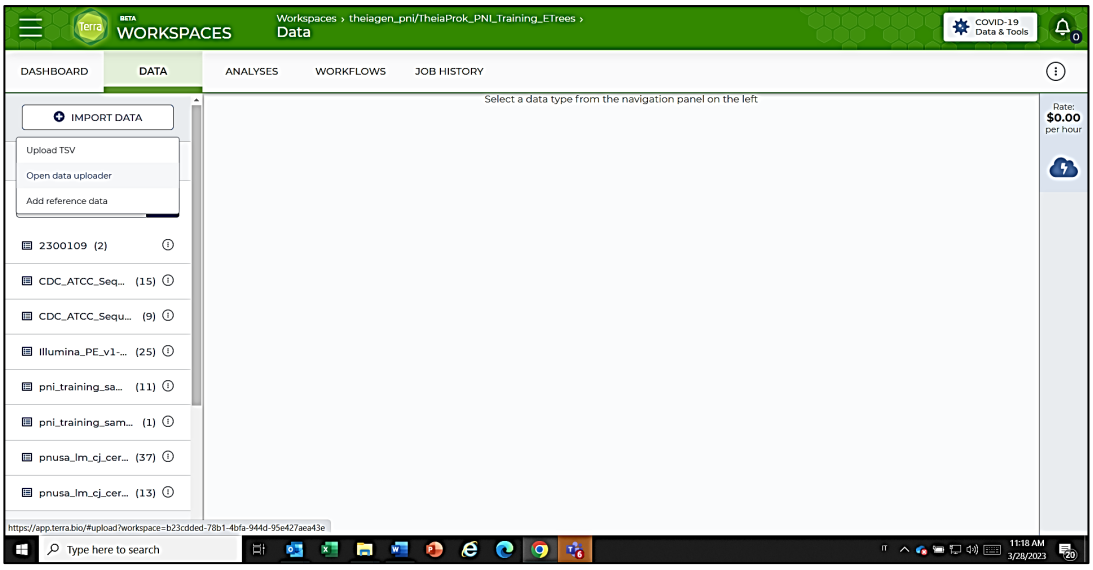


5.2 Увійдіть на Terra за допомогою Chrome та свого облікового запису Google

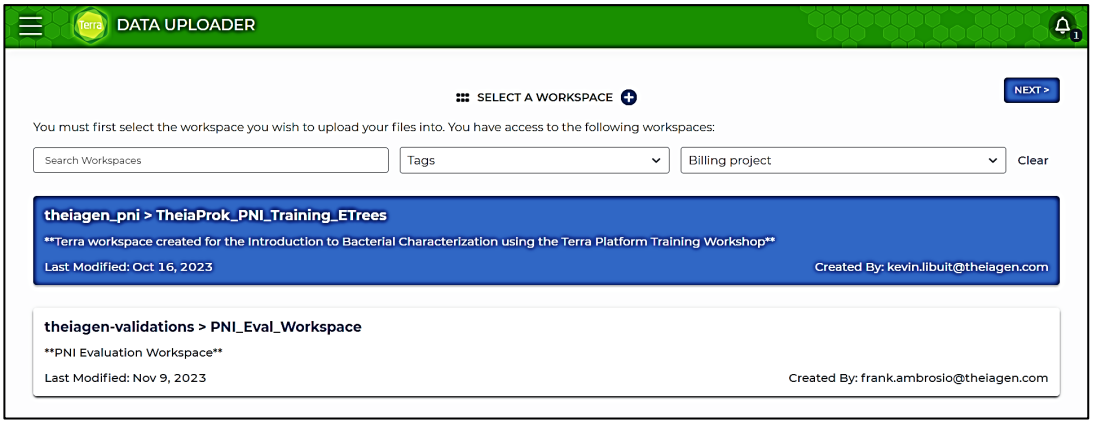


5.3 **Завантажте файли послідовностей та метадані до Terra**

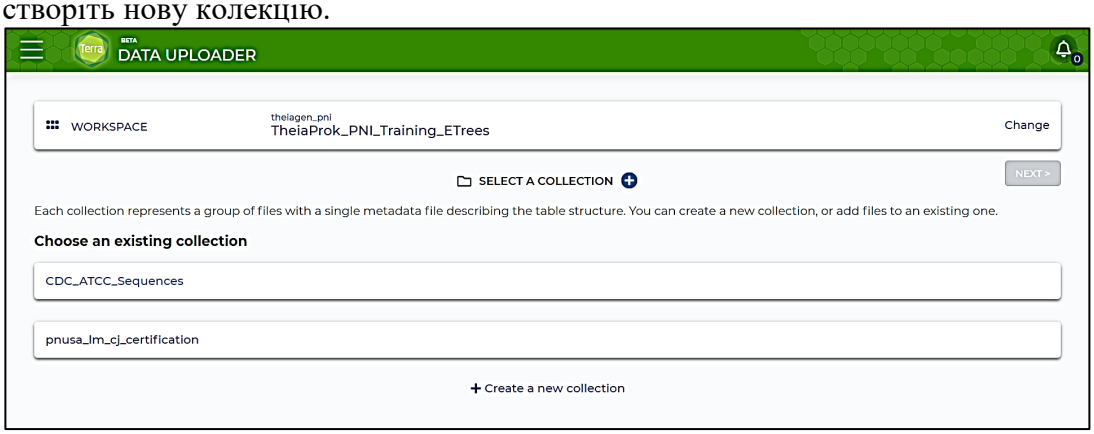
5.3.1 У розділі "Робочі простори Terra" виберіть вкладку "Дані", натисніть "Імпортувати дані" і виберіть "Відкрити завантажувач даних" у випадаючому меню.



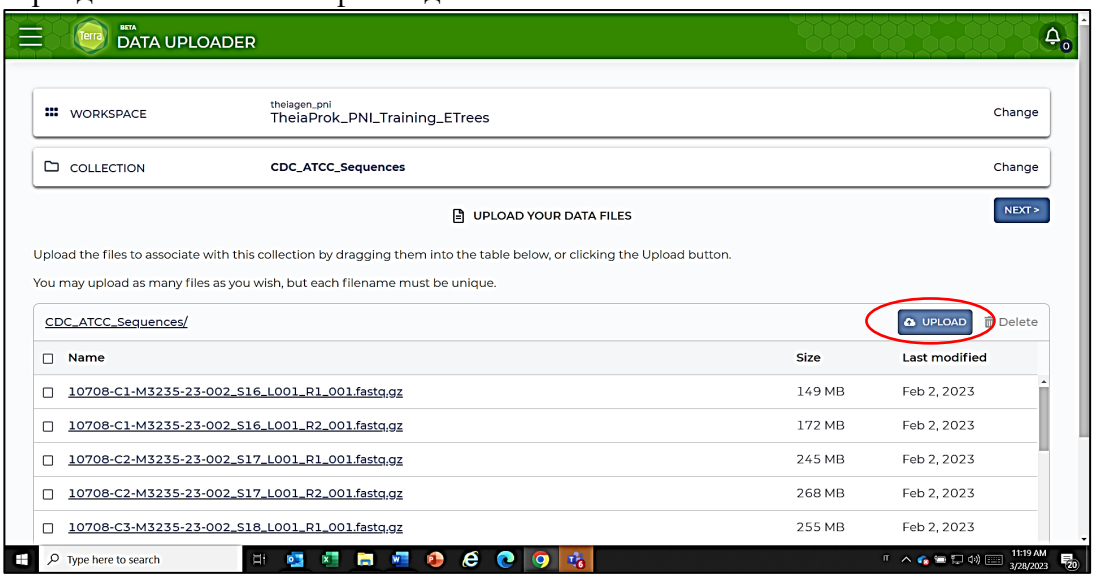
5.3.2 Відкриється вікно "Завантажувач даних". Якщо ваш обліковий запис має доступ до кількох робочих просторів, спочатку потрібно вибрати робочий простір, до якого ви бажаєте завантажити дані.



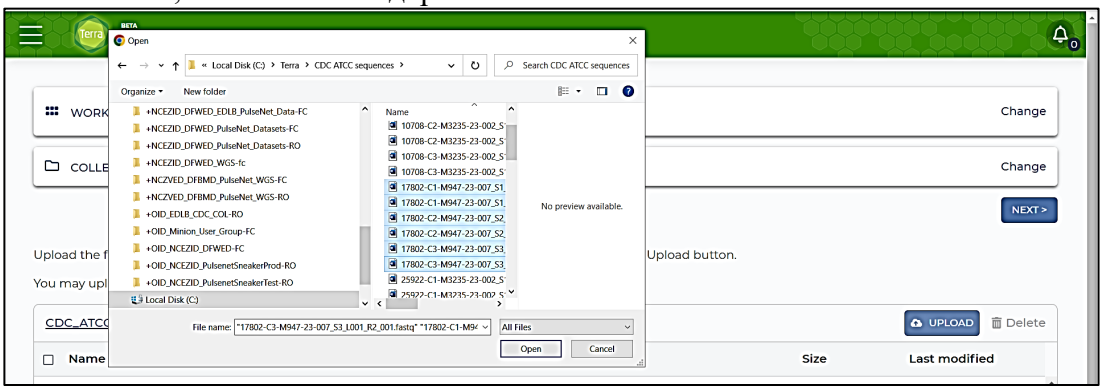
5.3.3 Або виберіть існуючу колекцію, натиснувши на назву колекції у списку, або створіть нову колекцію.



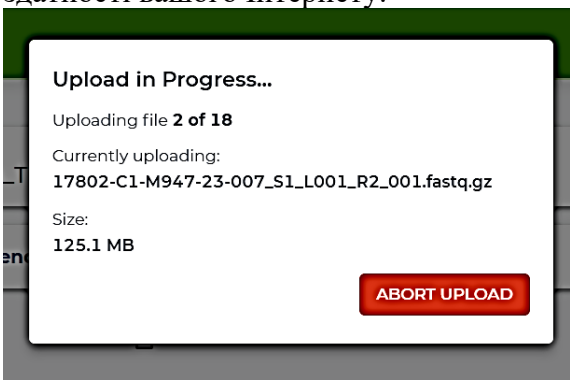
5.3.4 У розділі "Завантажити файли даних" натисніть "Завантажити".



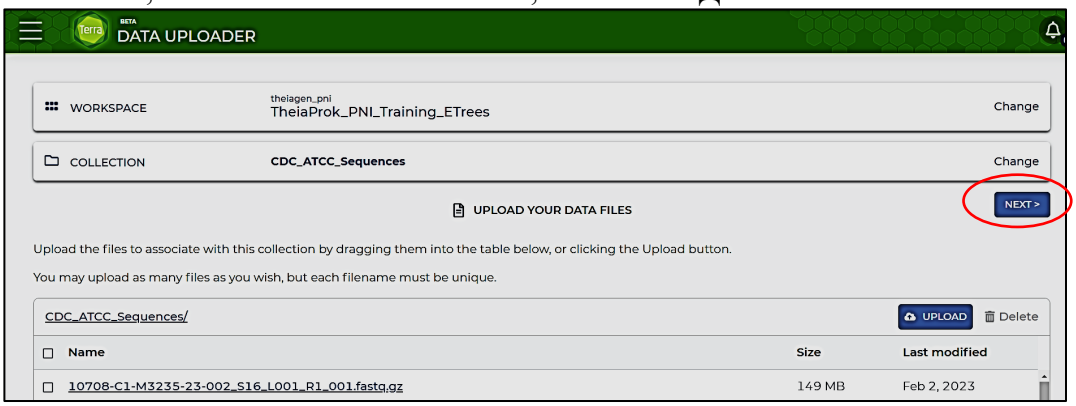
5.3.5 Перейдіть до місця, де зберігаються файли FASTQ, виберіть файли, які потрібно завантажити, і натисніть "Відкрити".



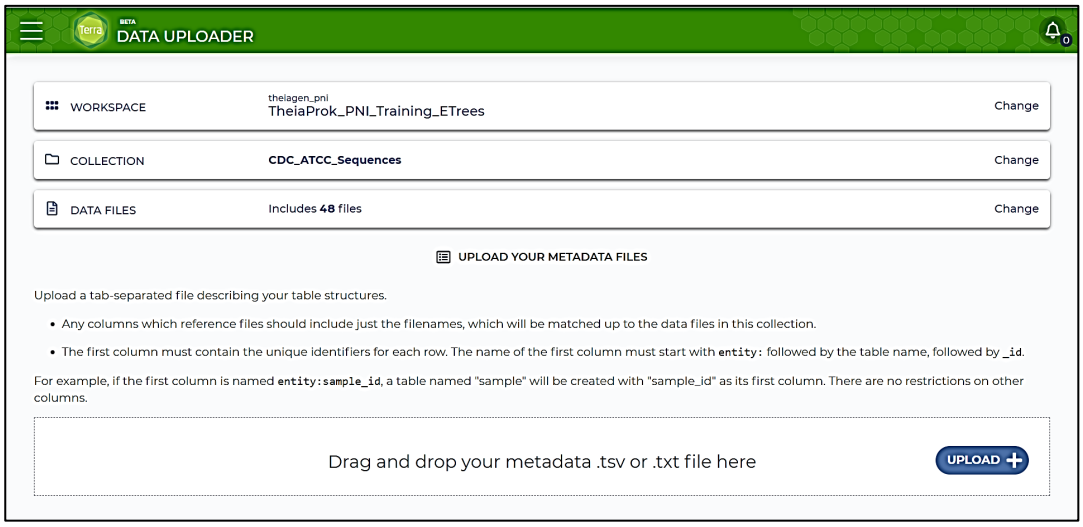
5.3.6 З'явиться спливаюче вікно "Завантаження триває". Завантаження може зайняти кілька хвилин, залежно від кількості файлів, що завантажуються, та пропускної здатності вашого Інтернету.



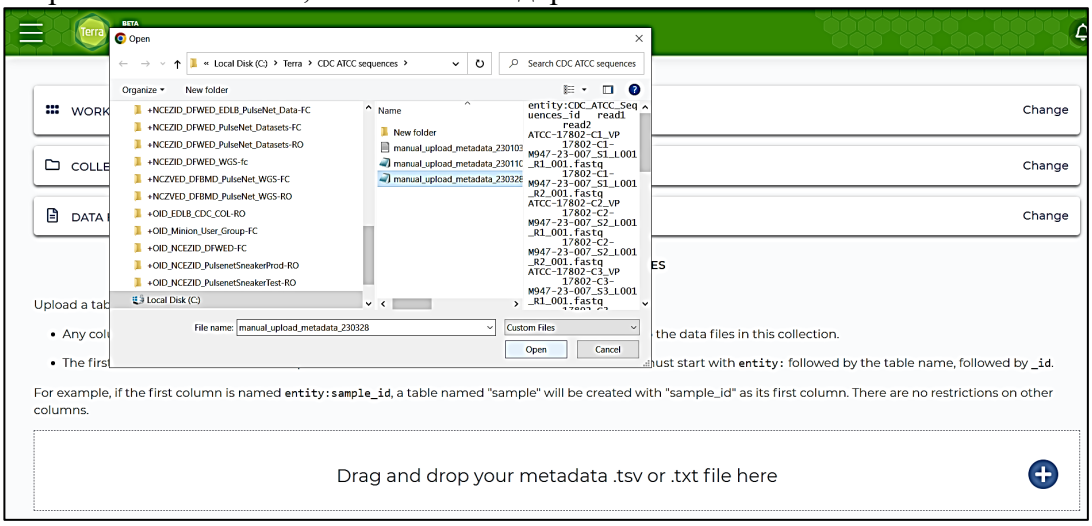
5.3.7 Після того, як спливаюче вікно зникне, натисніть "Далі".



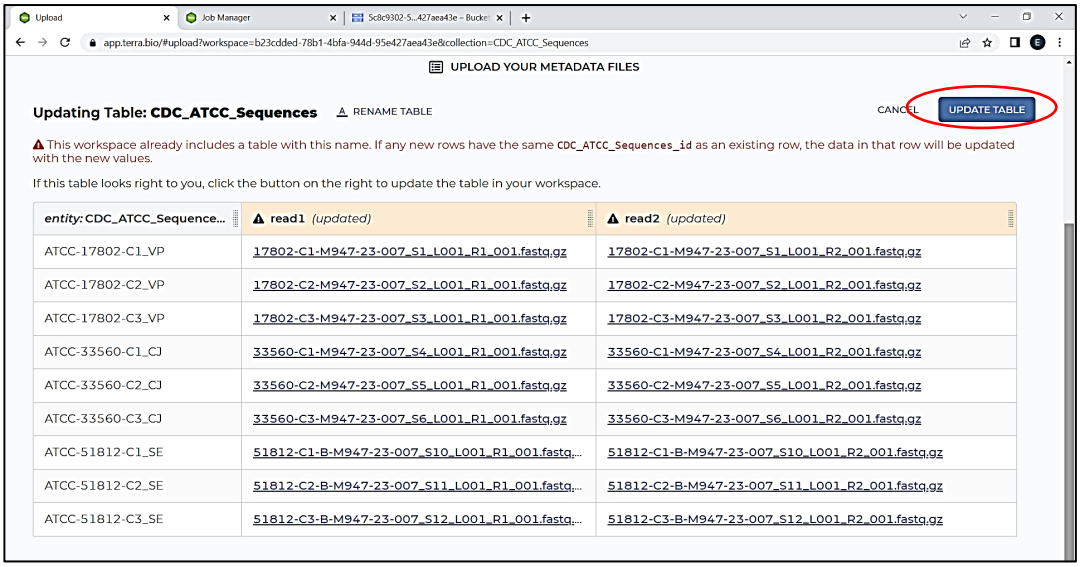
5.3.8 У розділі "Завантажити файли метаданих" натисніть "Завантажити"



5.3.9 Перейдіть до місця, де зберігається файл метаданих tsv, виберіть файл, який потрібно завантажити, і натисніть "Відкрити".

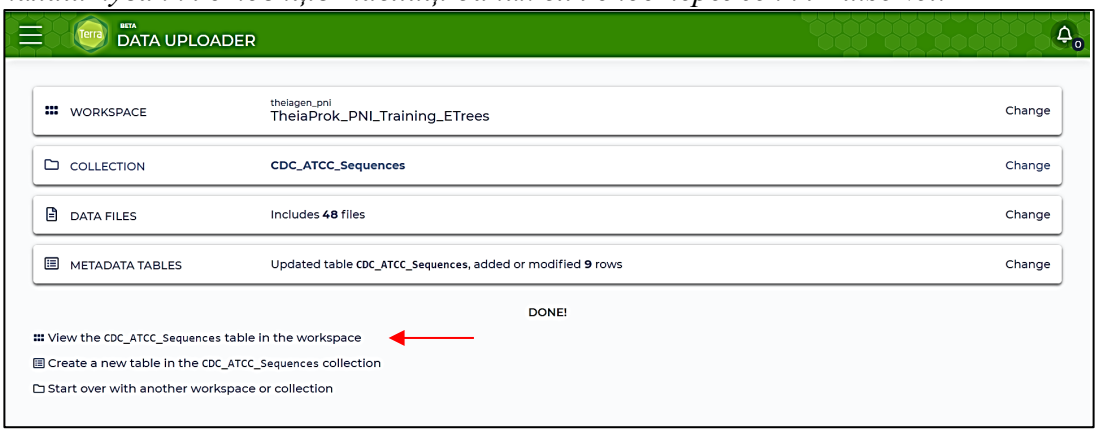


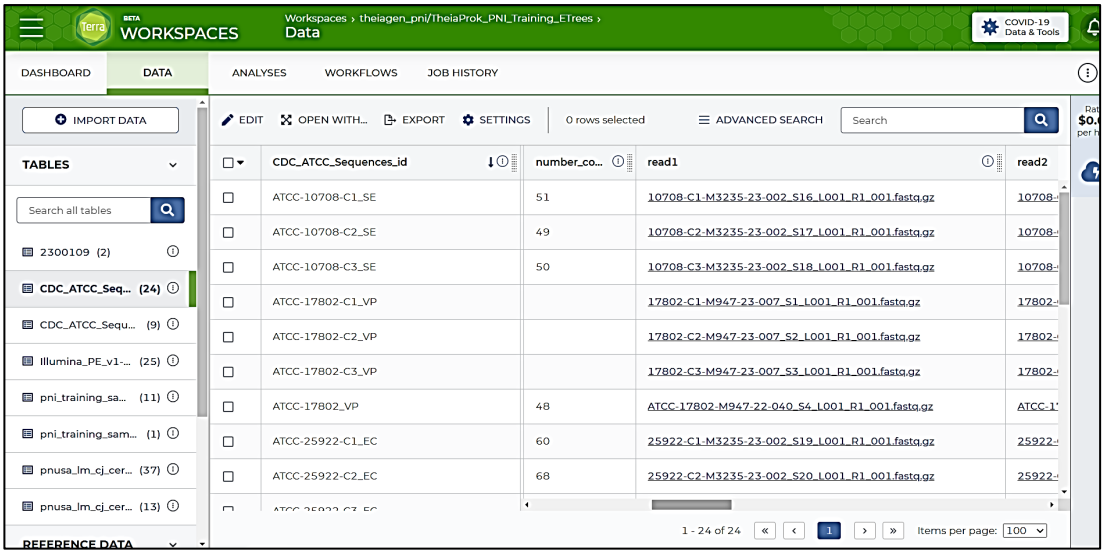
5.3.10 На наступному екрані перевірте, що файли fastq.gz пов'язані з правильними ключами входу, і натисніть "Оновити таблицю".



5.3.11 На екрані "Завантаження даних" з'явиться повідомлення "Готово". Ви можете переглянути оновлену таблицю даних, натиснувши на посилання, яке з'явиться на екрані.

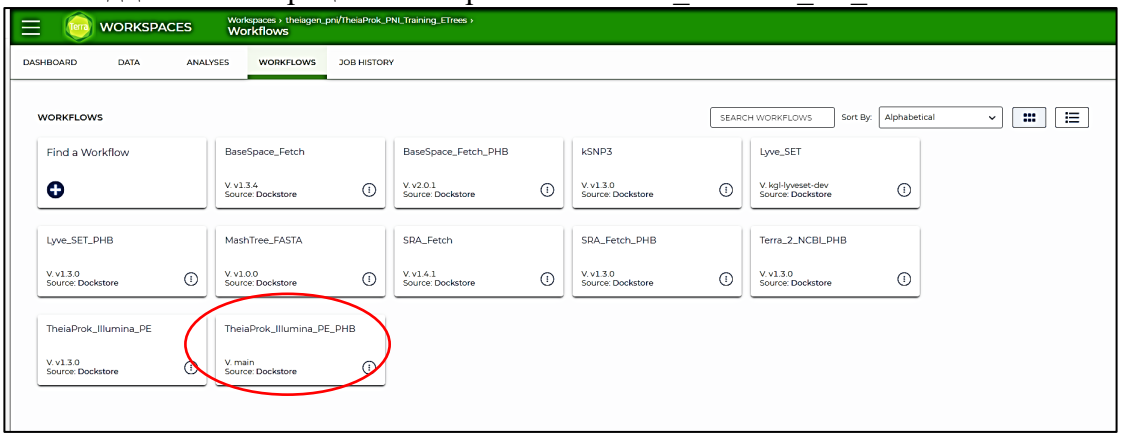
ПРИМІТКА: *Стовпці, що відображаються в таблиці даних, можна налаштувати. Для зручності навігації рекомендується створити окремі подання для показників КК, результатів генотипування та метаданих. Зверніться до додатків PNID01-3 (Показники контролю якості), PNID01-5 (Генотипування) та PNID01-6 (Метадані) для отримання інструкцій щодо налаштування стовпців таблиці даних для спостереження PulseNet.*





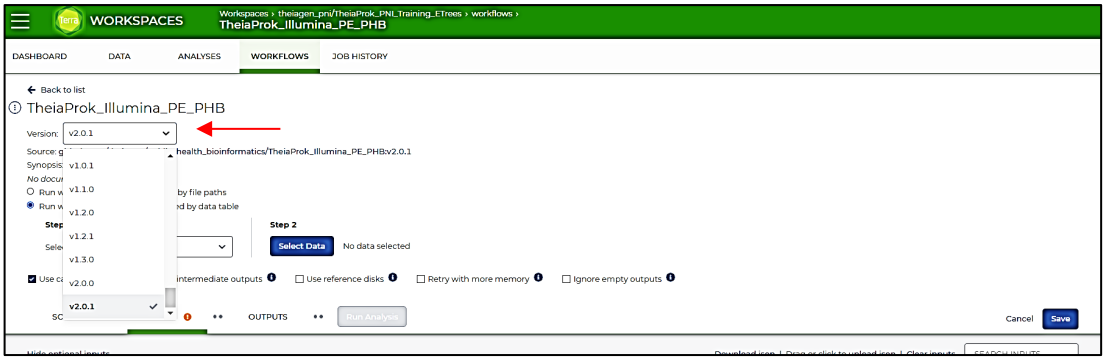
5.4 **Запустіть робочий процес контролю якості та генотипування**. Робочий процес TheiaProk виконує QC сирих зчитувань послідовностей, збирає сирі зчитування de novo за допомогою Skesa, а потім виконує QC збірки та ідентифікацію виду. Також доступні різноманітні аналізи генотипування, що підходять для даного виду.

5.4.1 На вкладці "Робочі процеси" виберіть "TheiaProk_Illumina_PE_PHB".

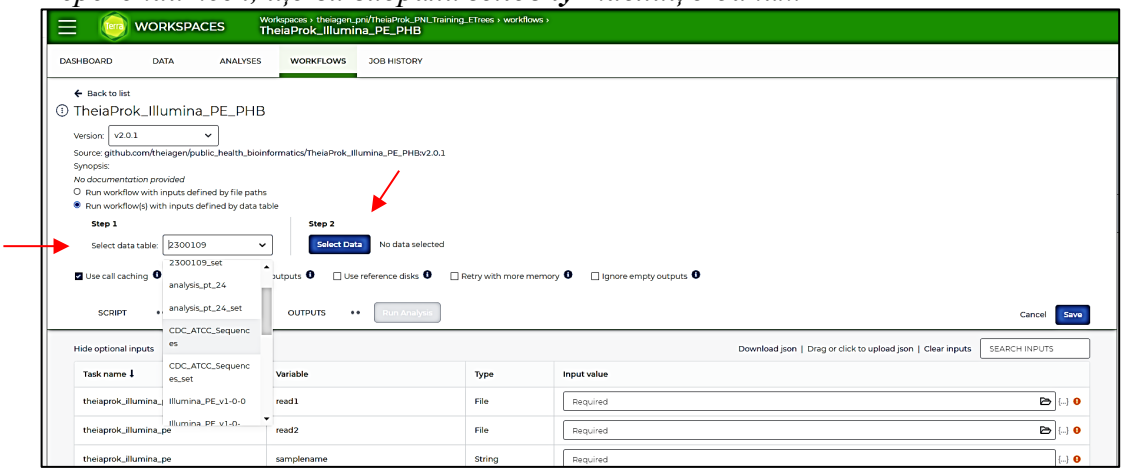


5.4.2 На екрані "TheiaProk_Illumina_PE_PHB":

5.4.2.1 Виберіть останню версію робочого процесу TheiaProk_Illumina_PE_PHB зі спадного меню "Версія".

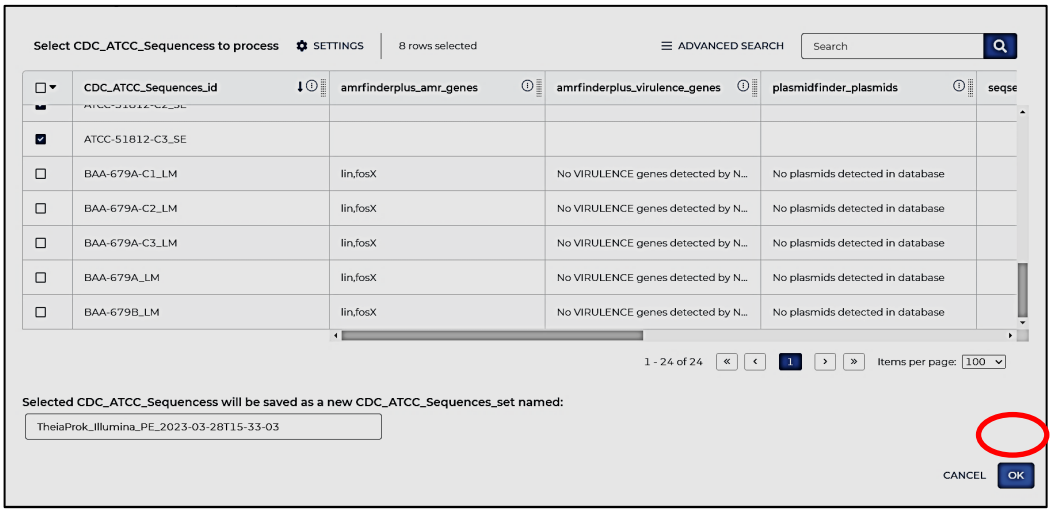


5.4.2.2 У розділі "Крок 1" виберіть "Тип кореня", тобто таблицю даних, в якій знаходяться зразки, наприклад, "CDC_ATCC_Sequences".
ПРИМІТКА: *Для більшості таблиць даних є два варіанти:* ***основна*** *таблиця даних, що містить окремі записи зразків, і таблиця даних наборів, що містить набори зразків, які використовуються для філогенетичного аналізу, завантаження з NCBI тощо (наприклад, CDC_ATCC_Sequences_set). Переконайтеся, що ви вибрали* ***головну*** *таблицю даних.*

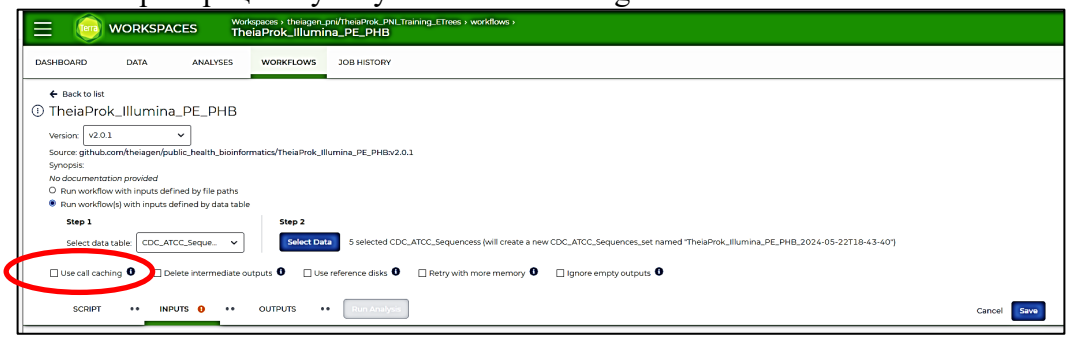


5.4.2.3 У розділі "Крок 2" натисніть "Вибрати дані" (скріншот вище). Ви потрапите до таблиці даних, вказаної на кроці 1.

5.4.2.4 Виберіть штами для аналізу і прокрутіть сторінку донизу, щоб натиснути "ОК".
ПРИМІТКА: *якщо таблиця даних містить понад 100 записів і ви встановите прапорець "Виділити все" біля назви таблиці даних, буде вибрано лише 100 записів.*



5.4.2.5 Зніміть прапорець з пункту "Use call catching".



5.4.2.6 На вкладці "**Вхідні дані**" вкажіть наступні вхідні значення в колонці "Атрибут" (прокрутіть список вниз):

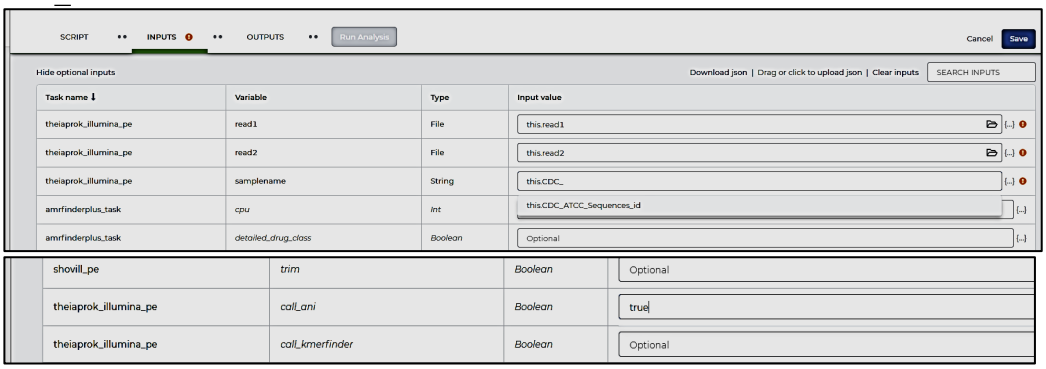
ПРИМІТКА: *Коли ви заповнюєте стовпчик "Атрибут", клацання всередині клітинки відкриває спадне меню атрибутів, які ви можете вибрати, щоб уникнути помилок (скріншот нижче).*

5.4.2.6.1 Зчитування 1: "This.read1".

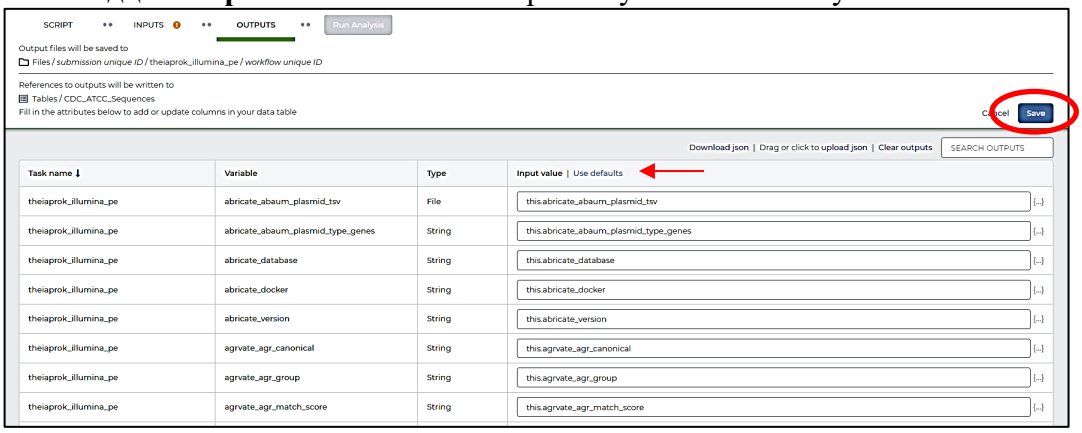
5.4.2.6.2 Read2: "This.read2".

5.4.2.6.3 Ім'я зразка: "**This**.*data table* **name_id**", наприклад, This.CDC_ATCC_Sequences_id.

5.4.2.6.4 Call_ani: "True".



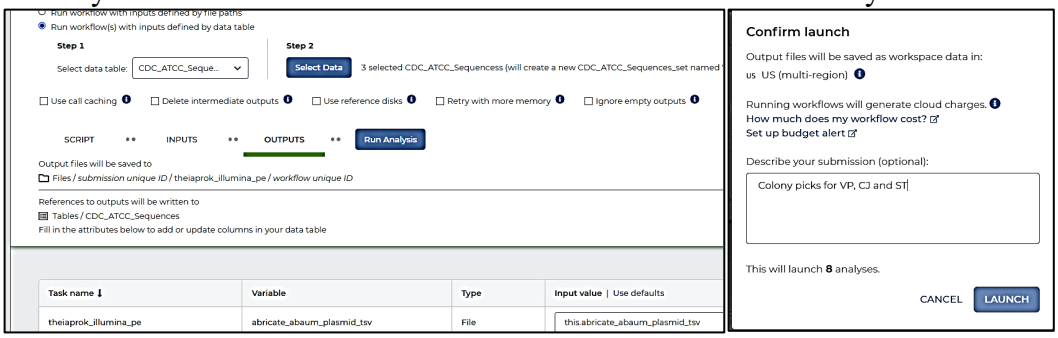
5.4.2.7 На вкладці "**Outputs**" натисніть "Використовувати за замовчуванням".



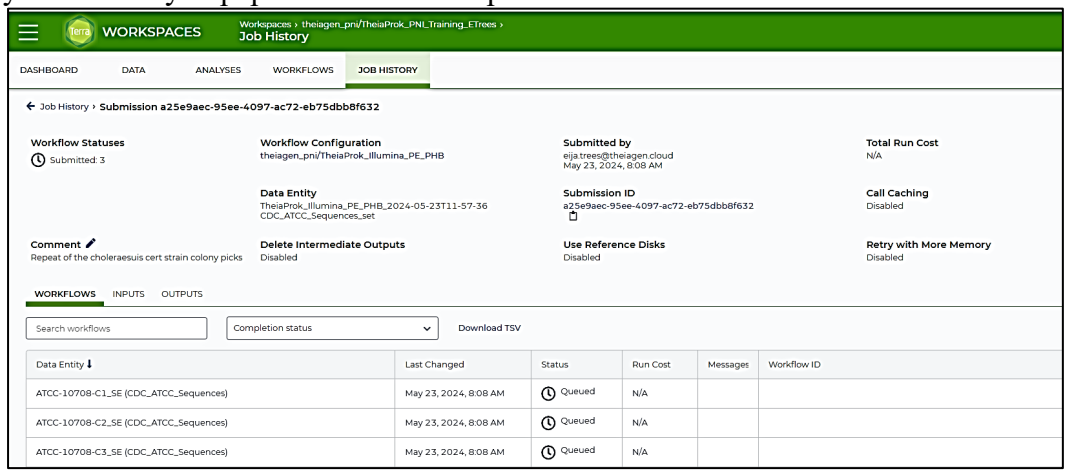
5.4.2.8 Натисніть "Зберегти" (скріншот вище).

ПРИМІТКА: *кнопка "Зберегти" відображається лише в тому випадку, якщо ви змінили вхідні дані з попереднього подання.*

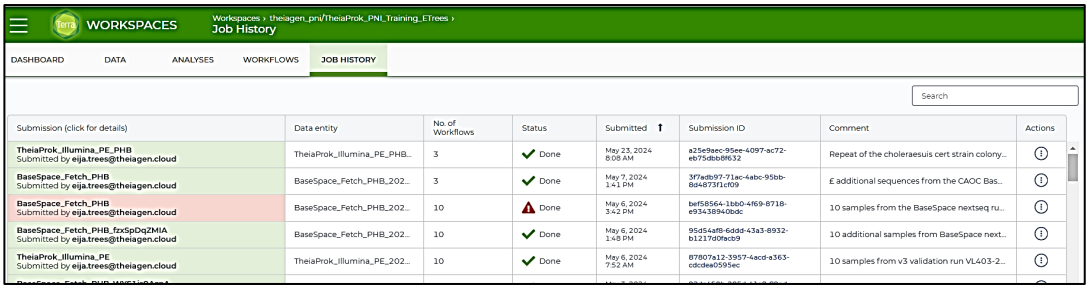
5.4.2.9 Натисніть "Запустити аналіз". З'явиться спливаюче вікно "Підтвердити запуск", в якому ви можете ввести необов'язковий опис. Натисніть "Запустити".



5.4.3 З'явиться вікно "Статуси робочого процесу", де ваші надіслані завдання повинні бути спочатку перераховані як "В черзі".

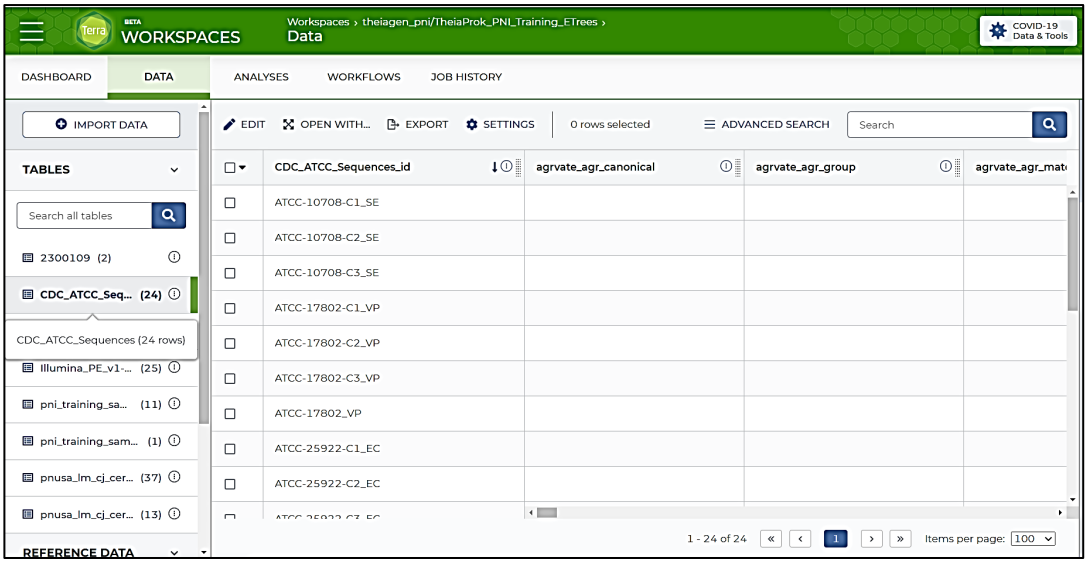


5.4.4 Перейдіть на вкладку "Історія завдань", щоб перевірити статус вашого завдання. Успішно виконане завдання позначається зеленою галочкою.

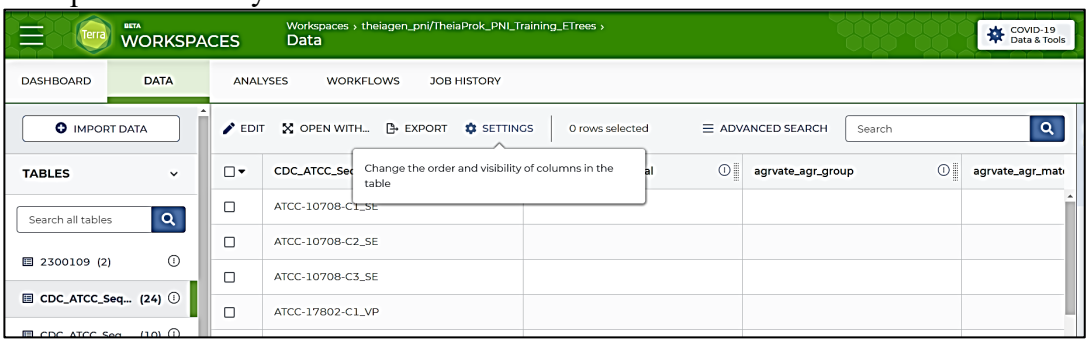


5.5 **Оцініть метрики КЯ для послідовностей:** Показники КЯ можна переглянути безпосередньо в таблиці даних (5.5.1-5.5.3) або експортувати в Excel для вибраних записів (5.5.4).

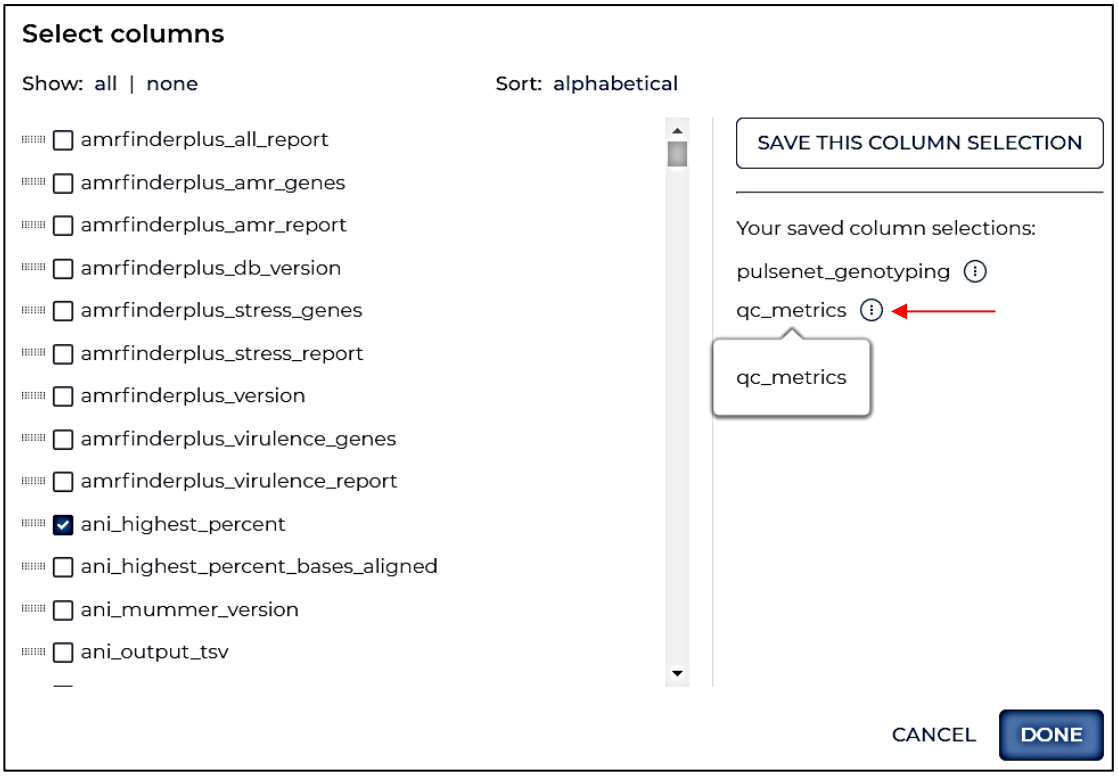
5.5.1 У "Робочих просторах Terra" виберіть вкладку "Дані", а потім виберіть таблицю даних, яка вас цікавить, наприклад, "CDC_ATCC_Sequences".

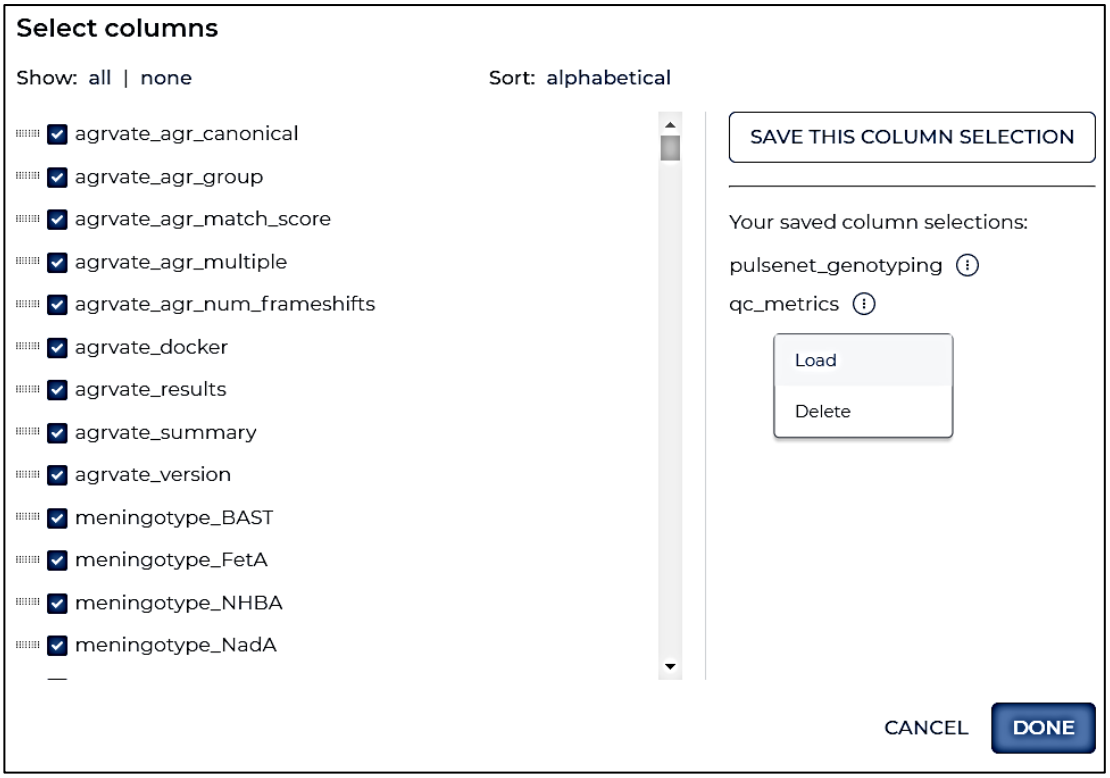


5.5.2 Виберіть "Налаштування".



5.5.3 На екрані "Вибір стовпців" у розділі "Ваші збережені вибори стовпців" натисніть на коло з 3 крапками поруч з "qc_metrics" і у випадаючому меню виберіть "Завантажити", а потім натисніть "Готово". Це призведе до завантаження відповідних метрик PulseNet QC в таблицю даних. Зверніться до Додатку PNID01-3 для отримання інформації про показники QC, які повинні з'явитися в таблиці, а також для отримання інструкцій про те, як додати або видалити будь-який з стовпців (показники QC) в таблиці показників QC.

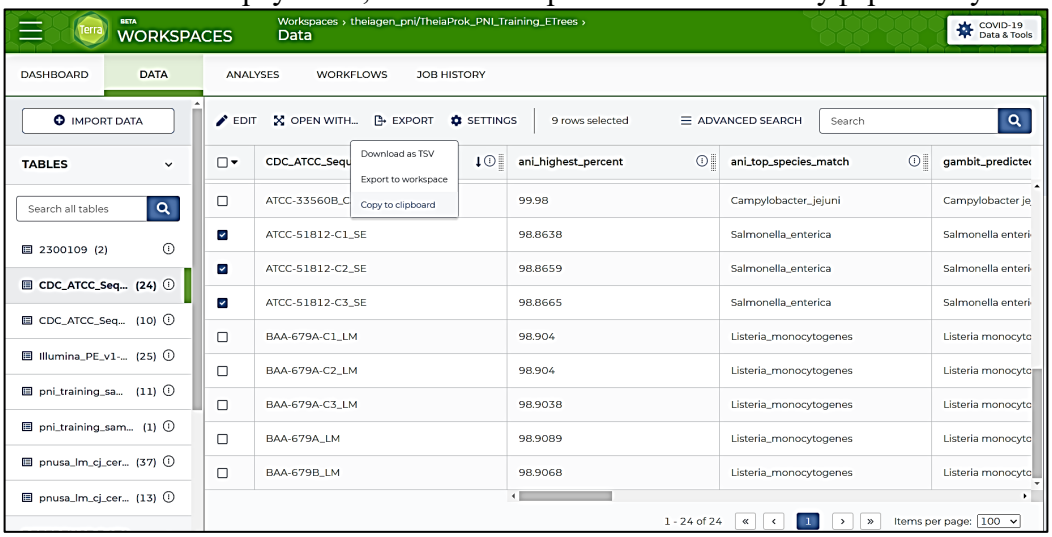

Select columns
Show: all | none
Sort: alphabetical
amrfinderplus_all_report
amrfinderplus_amr_genes
amrfinderplus_amr_report
amrfinderplus_db_version
amrfinderplus_stress_genes
amrfinderplus_stress_report
amrfinderplus_version
amrfinderplus_virulence_genes
amrfinderplus_virulence_report
ani_highest_percent
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

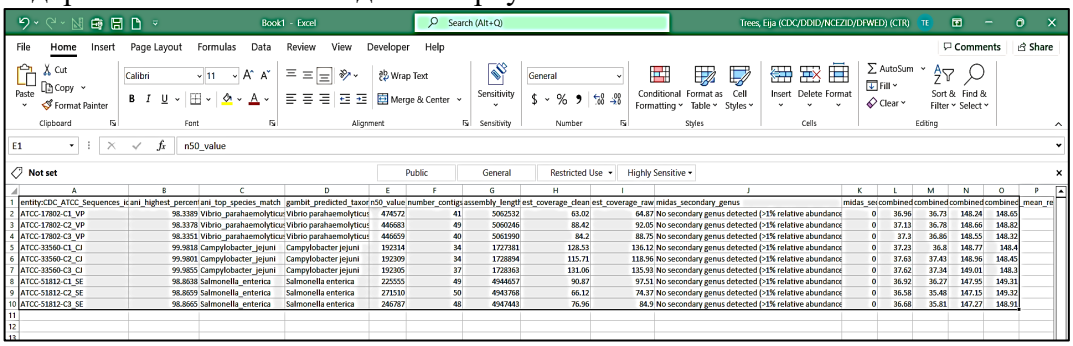
5.5.4 Експорт метрик контролю якості в Excel:

5.5.4.1 Виберіть потрібні записи.

5.5.4.2 Натисніть "Експортувати", а потім виберіть "Копіювати в буфер обміну".

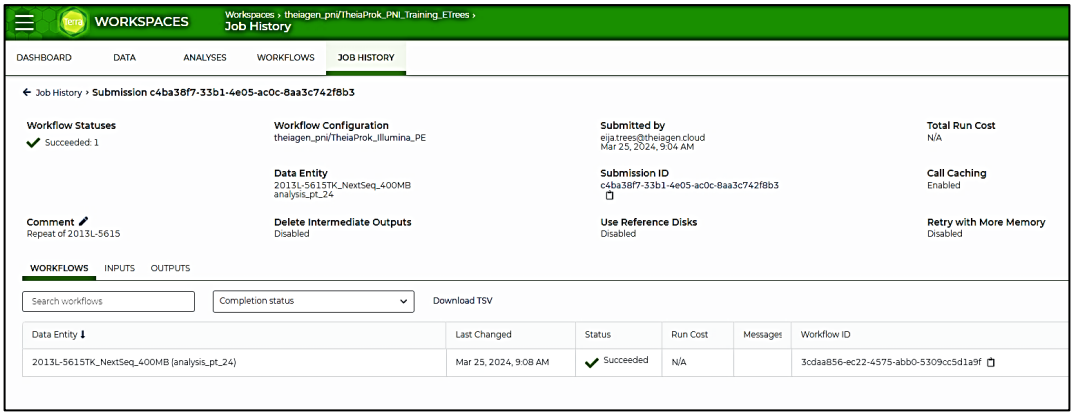


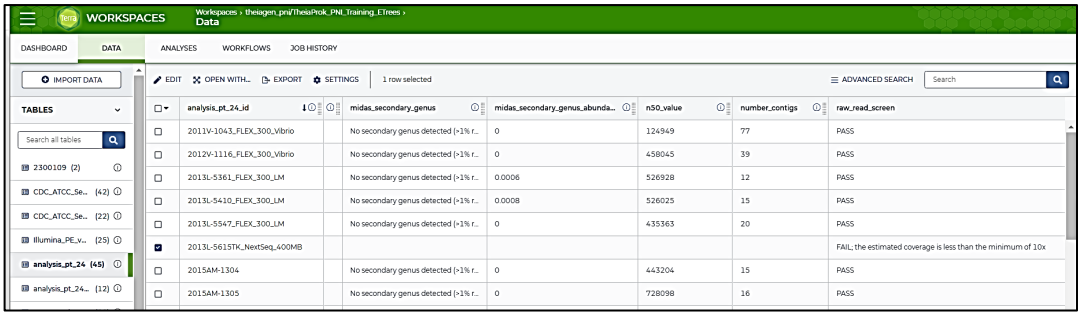
5.5.4.3 Відкрийте Excel і вставте дані на аркуш.



5.5.5 Критичні показники якості PulseNet для прийнятних послідовностей Illumina див. у Додатку PNID01-4a.

5.5.6 TheiaProk використовує етап попередньої перевірки зчитування, на якому він зупиняє аналіз для зразків, які не досягають певних порогових значень якості. У цьому випадку на вкладці "Історія завдань" буде вказано, що зразок пройшов успішно, але на вкладці "Дані" не буде результатів. У колонці "raw_read_screen" в метриці qc має бути вказана причина помилки аналізу послідовності. Порогові значення, що застосовуються на цьому етапі попереднього скринінгу зчитування, наведено в Додатку PNID01-4b.



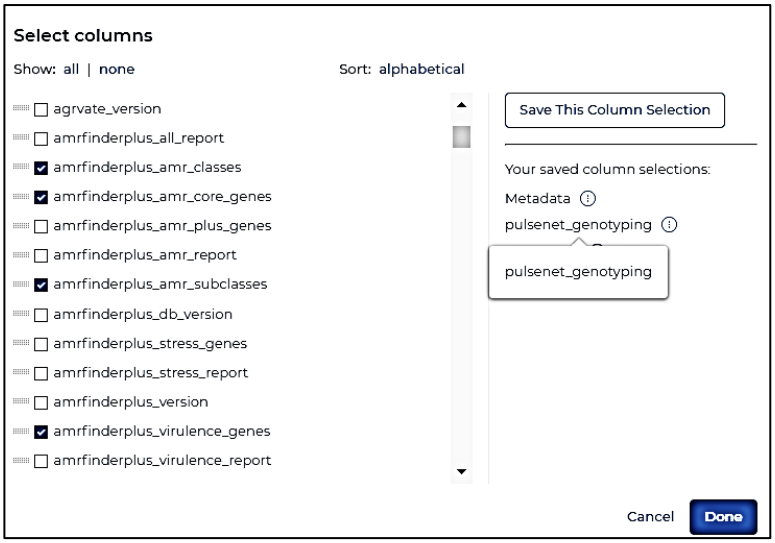


5.6 **Перегляньте результати генотипування для послідовностей:** Результати генотипування можна переглянути або безпосередньо в таблиці даних (5.6.1-5.6.3), або експортувати в Excel для вибраних записів (5.6.4).

5.6.1 У "Робочих просторах Terra" виберіть вкладку "Дані", а потім виберіть таблицю даних, яка вас цікавить, наприклад, "CDC_ATCC_Sequences".

5.6.2 На вкладці "Дані" виберіть "Налаштування".

5.6.3 На екрані "Вибір стовпчиків" у розділі "Ваш збережений вибір стовпчиків" натисніть на коло з 3 крапками поруч з "pulsenet_genotyping" і у випадаючому меню виберіть "Завантажити", а потім натисніть "Готово". Це призведе до завантаження аналізів генотипування, які підходять для спостереження PulseNet, в таблицю даних. Зверніться до Додатку PNID01-5 для отримання інформації про тести генотипування, які повинні з'явитися в таблиці, а також для отримання інструкцій, як додати або видалити будь-яку колонку (тести генотипування) в таблиці результатів генотипування.



5.6.4 Експортуйте результати в Excel для вибраних записів; виконайте процедуру, описану в кроці 5.5.4.

5.7 **Завантажуйте послідовності до NCBI**

ПРИМІТКА: *Зверніться до компанії Theiagen Genomics (support@theiagen.com) для отримання інструкцій перед початком роботи та налаштування робочого простору для завантаження даних до NCBI. Процес конфігурації описано за посиланням: https://theiagen.notion.site/Terra_2_NCBI-61abcedc066646b3b258f70b561e9f62.*

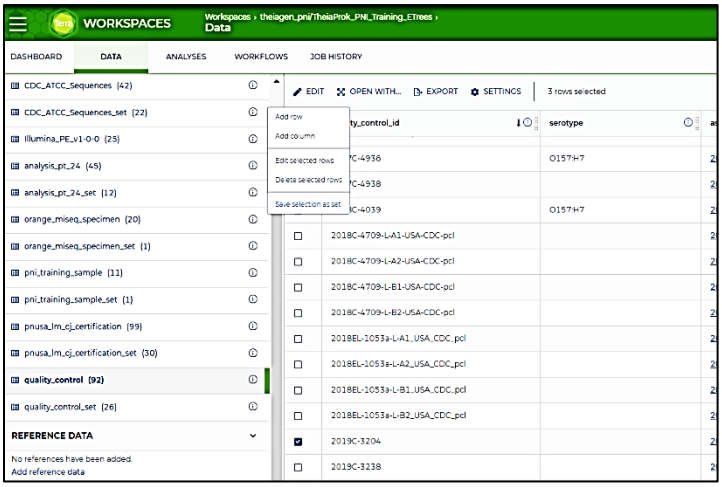
5.7.1 Завантажте метадані, необхідні для подання до NCBI: див. додаток PNID01-6 для правильного форматування та завантаження метаданих.

5.7.2 Створіть **набір** зразків для завантаження в NCBI:

5.7.2.1 У розділі "Робочі простори Terra" виберіть вкладку "Дані", а потім виберіть таблицю даних, яка вас цікавить, наприклад, "контроль якості".

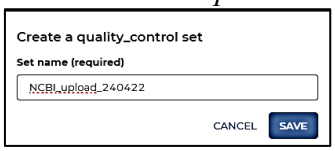
5.7.2.2 На вкладці "Дані" виберіть послідовності, які потрібно включити до завантаження в NCBI.

5.7.2.3 У випадаючому меню "Редагувати" виберіть "Зберегти виділення як задане".

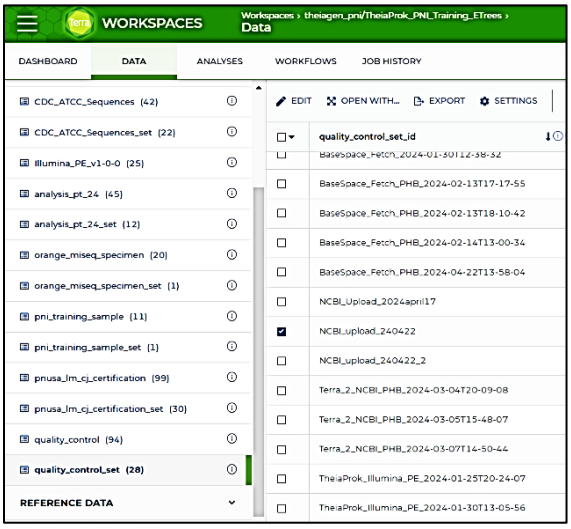


5.7.2.4 У спливаючому вікні назвіть набір, наприклад, "NCBI_upload_240422" і натисніть "Зберегти".

ПРИМІТКА: *пробіли та тире не допускаються.*

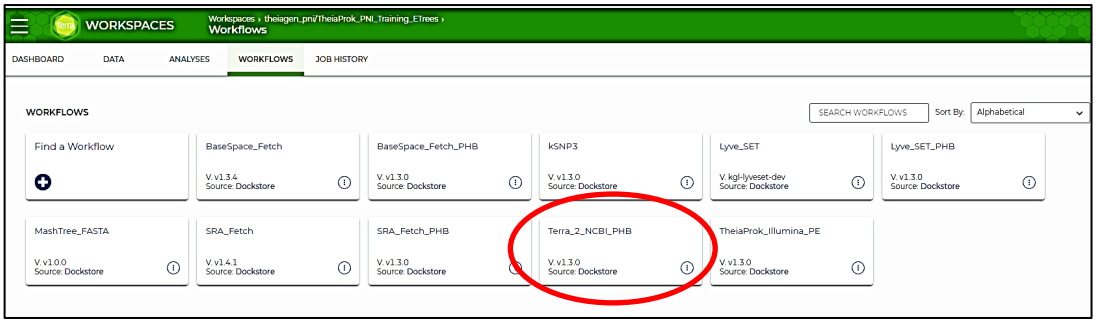


5.7.2.5 Новостворений набір тепер має з'явитися в таблиці даних "набір", наприклад, "набір_контролю_якості".

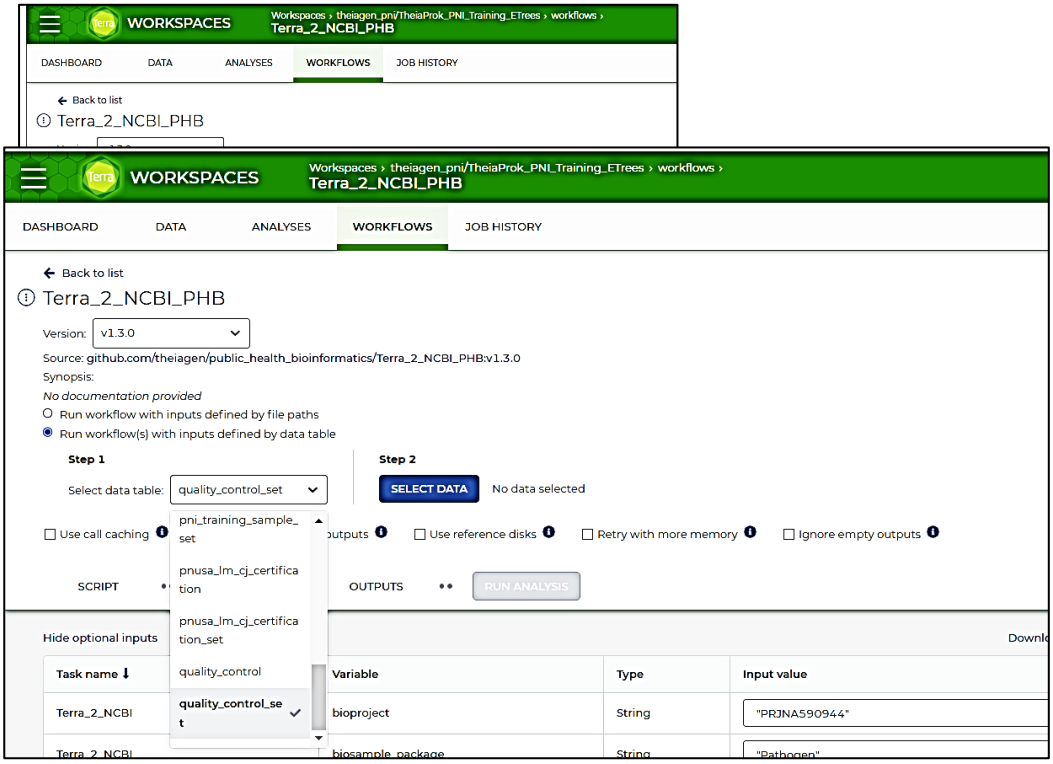


5.7.3 Налаштуйте параметри робочого процесу завантаження до NCBI:

5.7.3.1 На вкладці "Робочі процеси" виберіть робочий процес "Terra_2_NCBI_PHB". Відкриється вікно "Terra_2_NCBI_PHB".



5.7.3.2 У випадаючому меню "Версія" виберіть останню версію "Terra_2_NCBI_PHB".



5.7.3.3 На кроці 1 у випадаючому меню "Select root entity type" виберіть таблицю даних набору, в якій знаходиться зразок набору, створений на кроці 5.7.2, наприклад, "quality_control_set". Також зніміть позначку з пункту " Use call catching ".

5.7.3.4 На кроці 2 натисніть "Вибрати дані" (скріншот вище). Ви потрапите до таблиці даних, вибраної на попередньому кроці.

5.7.3.5 Виберіть потрібний набір зразків, наприклад, "NCBI_upload_240422" і натисніть "ОК".

5.7.3.6 На вкладці "**Вхідні** дані" потрібно заповнити наступні "Вхідні значення" для перелічених нижче "Змінних":

5.7.3.6.1 біопроект: введіть номер біопроекту NCBI, до якого ви бажаєте завантажити дані, у лапках, наприклад, "PRJNA590944".

5.7.3.6.2 biosample_package в лапках: "Патоген". Це шаблон/пакет метаданих, який ви використовуєте для завантаження метаданих для послідовностей, що належать до спостереження PulseNet.

5.7.3.6.3 ncbi_config_js: введіть назву файлу конфігурації NCBI, створеного для вашої робочої області, наприклад, workspace.ncbi_config_etrees.

5.7.3.6.4 назва_проекту в лапках: "theiagen_pni".

5.7.3.6.5 sample_names: введіть назву вашої таблиці даних у форматі: **this.data_table_names.data_table_name_id**, наприклад, це.якість_контролює.ідентифікатор_контролю_якості. **ПРИМІТКА:** *формат подвійного імені є ОБОВ'ЯЗКОВИМ.*

5.7.3.6.6 sra_transfer_gcp_bucket в лапках: "gs://theiagen_sra_transfer". Це тимчасове загальнодоступне сховище Google для ваших послідовностей, до якого може отримати доступ NCBI.

5.7.3.6.7 назва_таблиці: введіть назву вашої таблиці даних у лапках, наприклад, "quality_control".

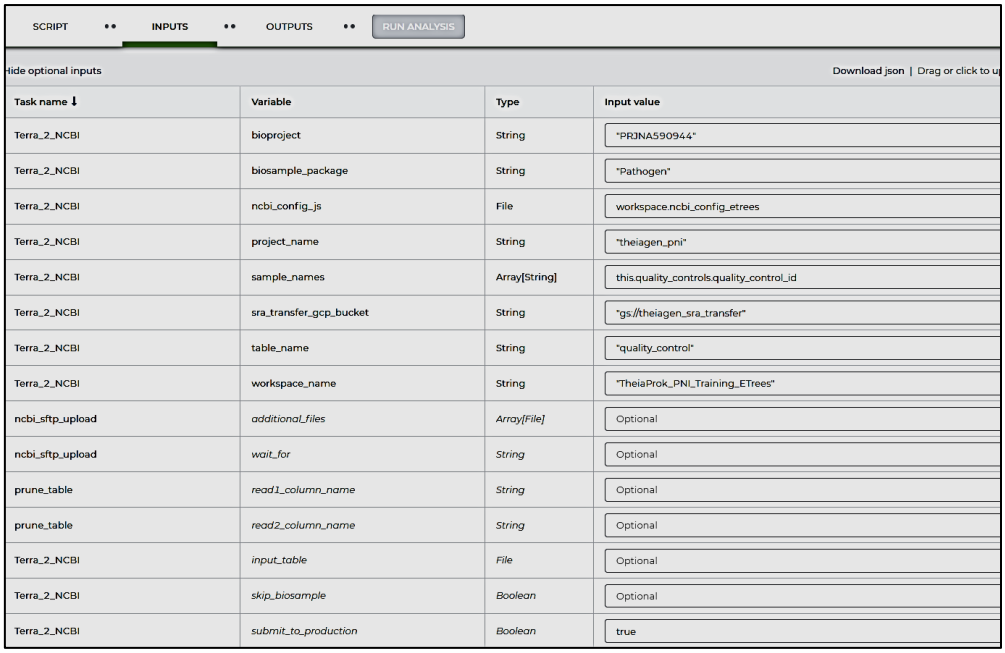
5.7.3.6.8 ім'я_робочого_простору: введіть ім'я вашого робочого простору в лапках, наприклад, "TheiaProk_PNI_Training_ETrees".

5.7.3.6.9 submit_to_production: true.

5.7.3.6.10 НЕОБОВ'ЯЗКОВО: якщо ви хочете пов'язати заявку SRA з **існуючим біозразком** (поле "biosample_accession" вже заповнене номером SAMN на вкладці "Дані"):

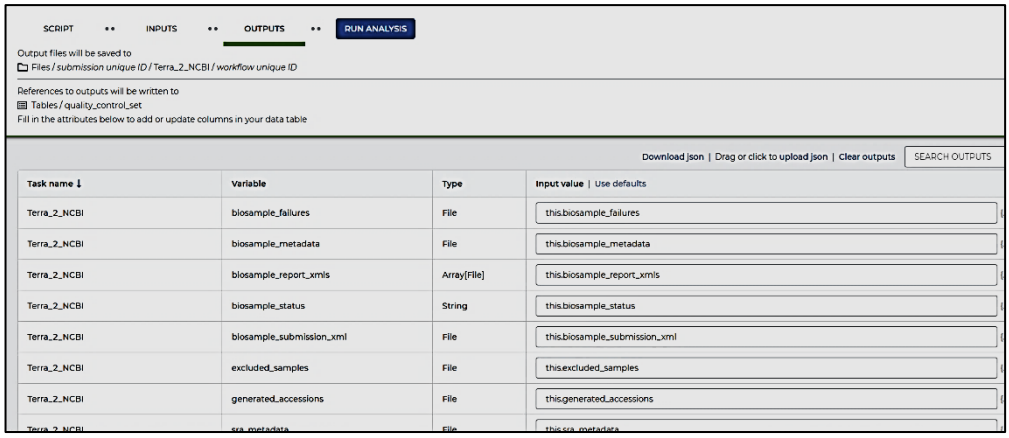
5.7.3.6.10.1 skip_biosample: true.

ПРИМІТКА: *для використання цієї функції необхідно заповнити необхідні поля метаданих, тобто заповнення лише поля "biosample_accession" є недостатнім і не пройде перевірку попереднього завантаження в Terra.*



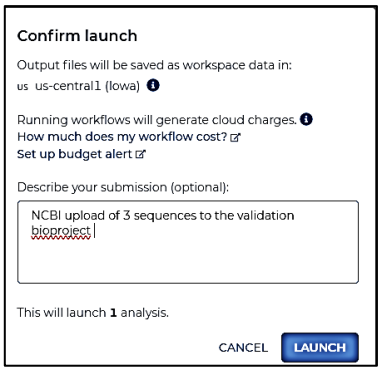
Task name	Variable	Type	Input value
Terra_2_NCBI	bioproject	String	"PRJNA590944"
Terra_2_NCBI	biosample_package	String	"Pathogen"
Terra_2_NCBI	ncbi_config_js	File	workspace.ncbi_config_etrees
Terra_2_NCBI	project_name	String	"theiagen_pni"
Terra_2_NCBI	sample_names	Array[String]	this.quality_controls.quality_control_id
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 вкладці "**Вихідні** дані" натисніть "Використовувати значення за замовчуванням" для "Атрибутів", а потім натисніть "Зберегти". **ПРИМІТКА**: *кнопка "Зберегти" відображається, тільки якщо ви змінили вхідні дані з попереднього подання.*

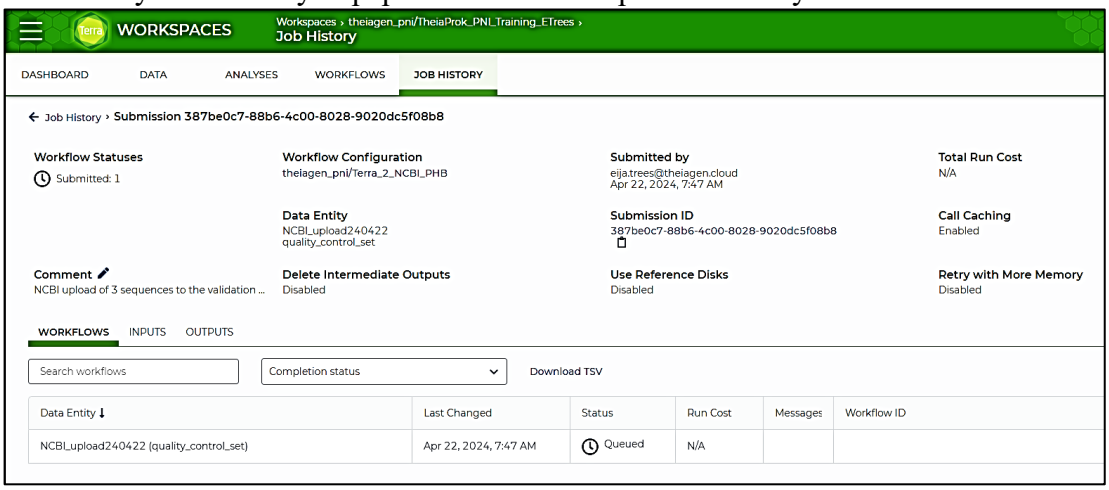


Task name	Variable	Type	Input value | Use defaults
Terra_2_NCBI	biosample_failures	File	this.biosample_failures
Terra_2_NCBI	biosample_metadata	File	this.biosample_metadata
Terra_2_NCBI	biosample_report_xmls	Array[File]	this.biosample_report_xmls
Terra_2_NCBI	biosample_status	String	this.biosample_status
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

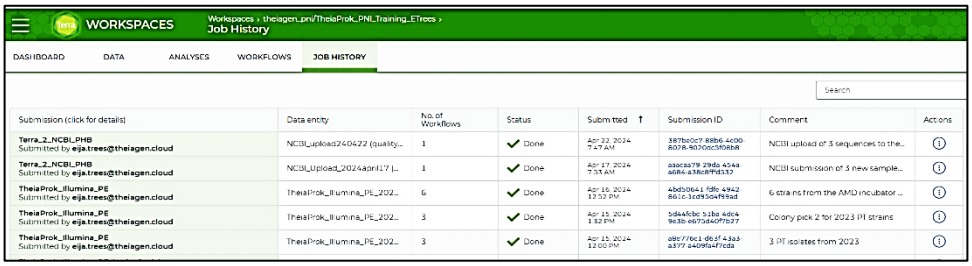
5.7.3.8 Натисніть "Запустити аналіз" (скріншот вище). З'явиться спливаюче вікно "Підтвердити запуск", в якому ви можете ввести необов'язковий опис. Натисніть "Запустити".



5.7.3.9 З'явиться вікно "Статуси робочого процесу", де ваші надіслані завдання повинні бути спочатку перераховані як "В черзі" або "Запуск".



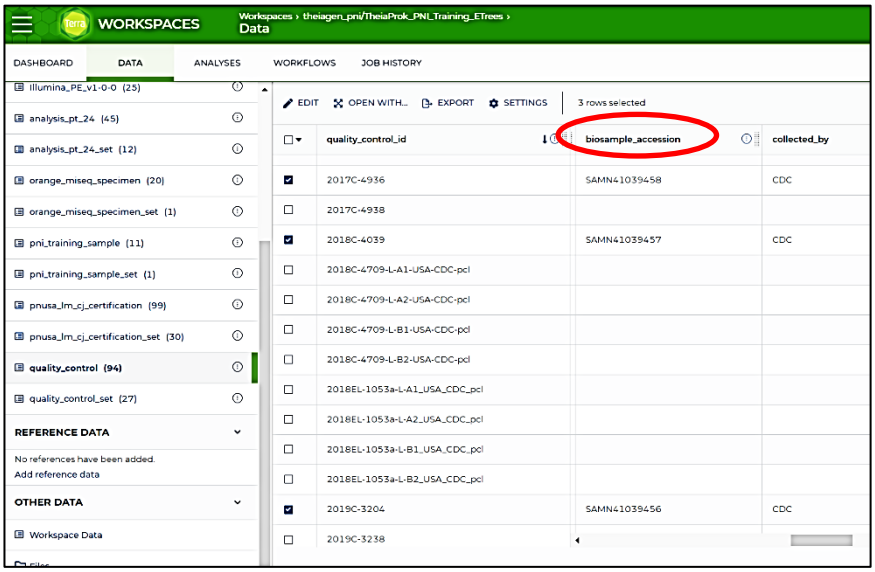
5.7.3.10 Перейдіть на вкладку "Історія завдань", щоб перевірити статус вашого завдання. Успішно виконане завдання позначається зеленою галочкою.



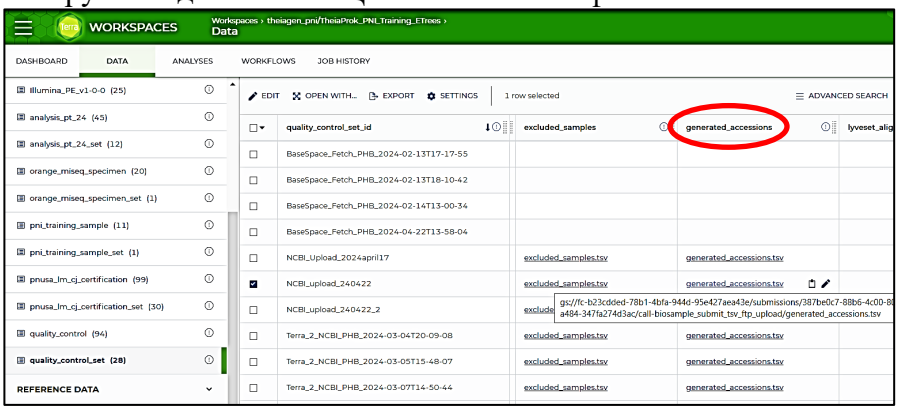
5.7.3.11Відстеження номерів доступу до NCBI:

5.7.3.11.1 Номери **приналежності біозразків** (номери SAMN)

5.7.3.11.1.1 У вкладці "Дані" перейдіть до відповідної таблиці даних, і ви побачите поле "biosample_accession", заповнене номером SAMN, який присвоюється кожному унікальному біозразку.

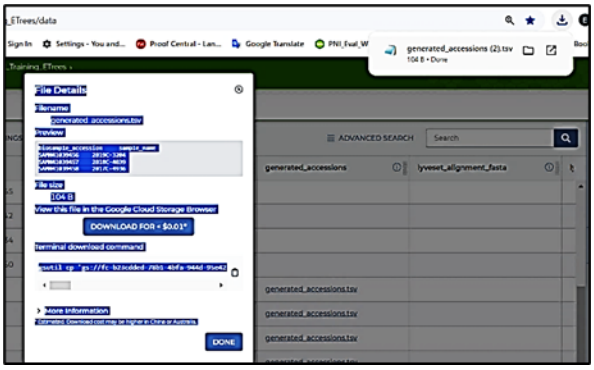


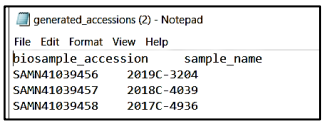
5.7.3.11.1.2 У вкладці "Дані" перейдіть до відповідного **набору** таблиць даних, знайдіть набір даних, який ви створили для завантаження в NCBI, а потім натисніть на посилання "згенеровані_доступи" для tsv-файлу, в якому перераховані номери біозразків для завантаженого набору послідовностей. Щоб завантажити файл:



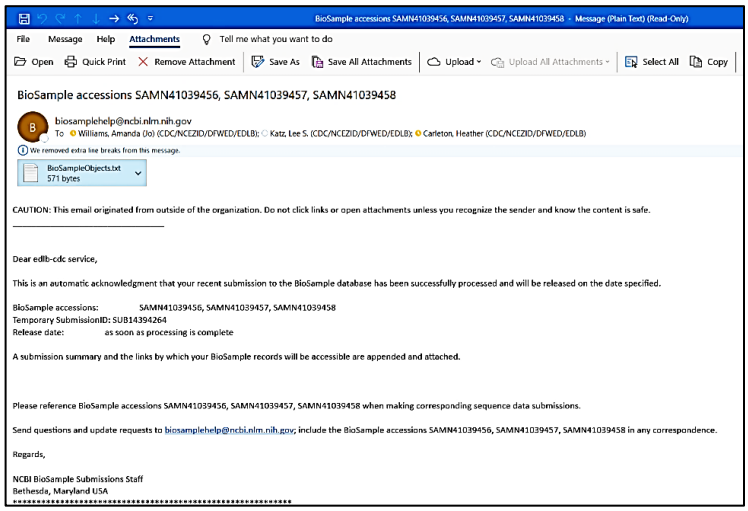
5.7.3.11.1.2.1 Натисніть "Завантажити >$0.01*", а потім натисніть "Готово".

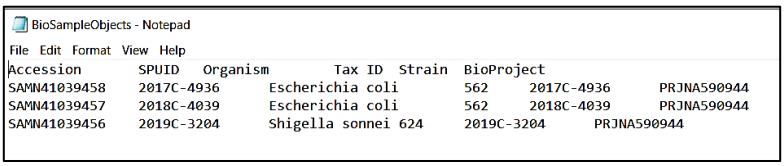
5.7.3.11.1.2.2 Завантажений файл з'явиться у верхньому правому куті.





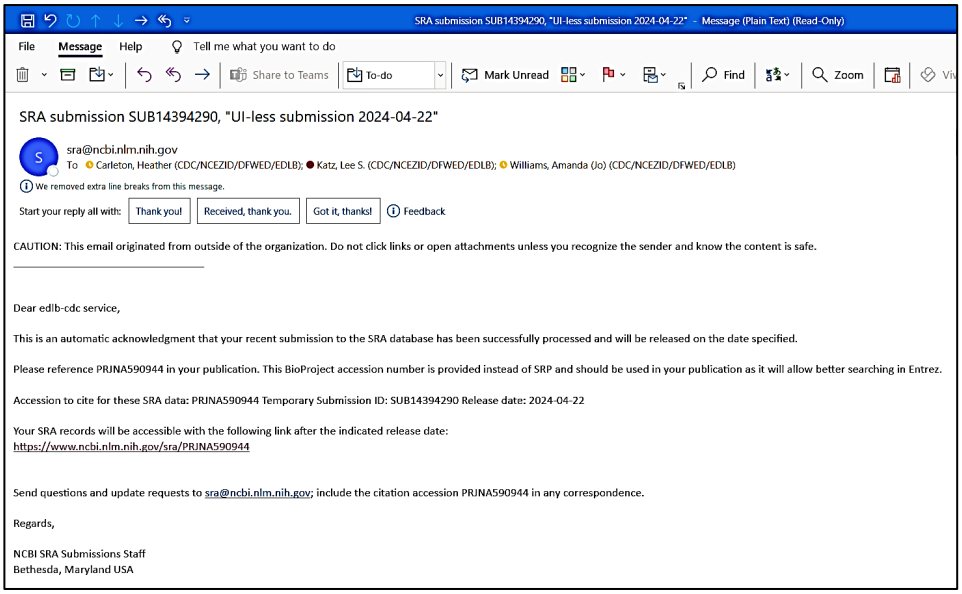
5.7.3.11.1.3 На електронну адресу, пов'язану з обліковим записом NCBI, який було використано для завантаження, має надійти лист-підтвердження з переліком номерів SAMN в тілі листа та в текстовому файлі "BiosampleObjects", прикріпленому до листа.



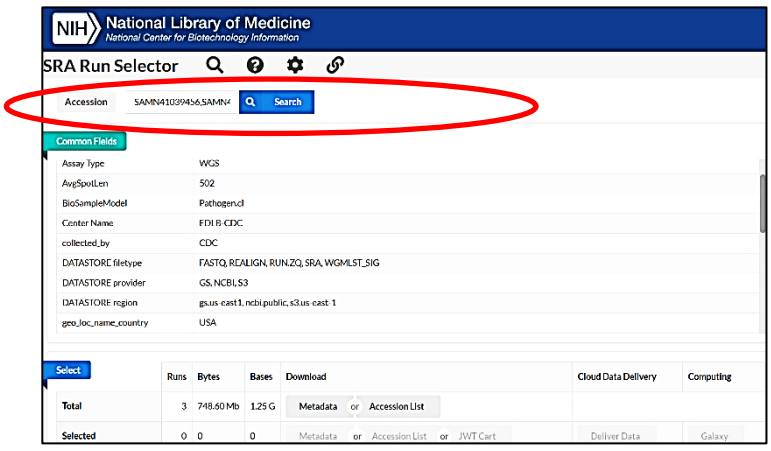


5.7.3.11.2 Номери **приєднання до SRA** (номери SRR)

ПРИМІТКА: *На електронну адресу, пов'язану з обліковим записом NCBI, який використовувався для завантаження, має надійти другий електронний лист із підтвердженням того, що подання до бази даних SRA було успішно оброблено і буде опубліковане в дату, зазначену в електронному листі. Електронний лист не містить номерів приєднання до SRA.*

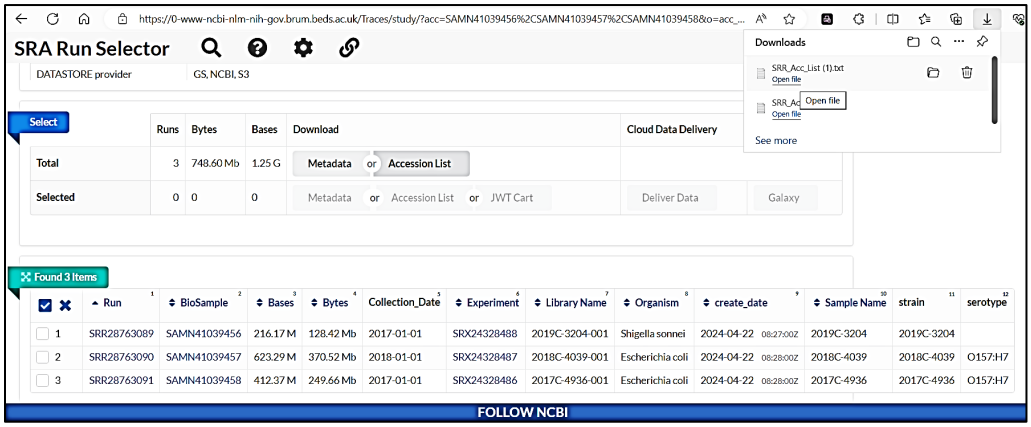


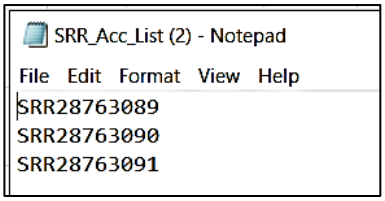
5.7.3.11.2.1 Скопіюйте та вставте номери SAMN з файлу tsv з кроку 5.7.3.11.1.2. або файлу txt з кроку 5.7.3.11.1.3. до інструменту "NCBI Run Selector" за адресою: https://0-www-ncbi-nlm-nih-gov.brum.beds.ac.uk/Traces/study/. Розділіть числа комами. Натисніть на кнопку "Пошук".



5.7.3.11.2.2 Пошук поверне таблицю, яка містить всю інформацію NCBI про ваші послідовності, включно з номерами доступу до SRR.

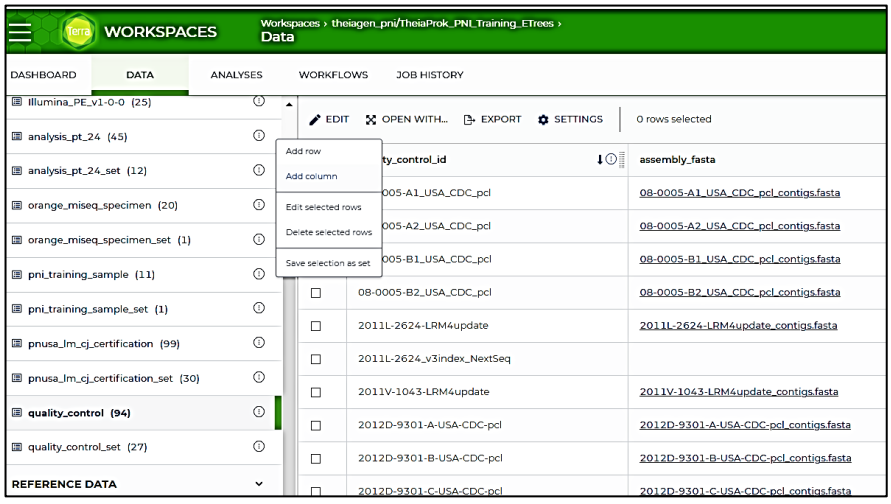
5.7.3.11.2.3 Натисніть на "Accession List", щоб завантажити таблицю з номерами приєднань. Завантажений файл з'явиться у верхньому правому куті.



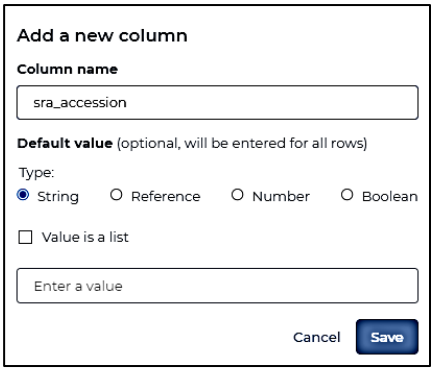


5.7.3.11.2.4 Ви можете відстежувати номери приєднання до SRA або в окремій таблиці Excel, або в системі LIMS, або ви можете створити поле "sra_accession" у вашій таблиці даних Terra і скопіювати та вставити туди номери приєднання для кожного зразка запису. Створити стовпець sra_accession і використовувати його для відстеження:

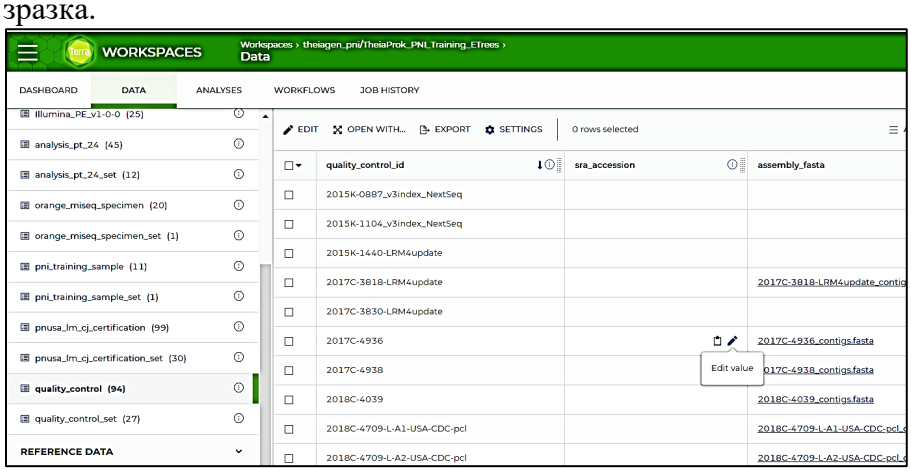
5.7.3.11.2.4.1 У випадаючому меню "Редагувати" виберіть "Додати стовпчик".



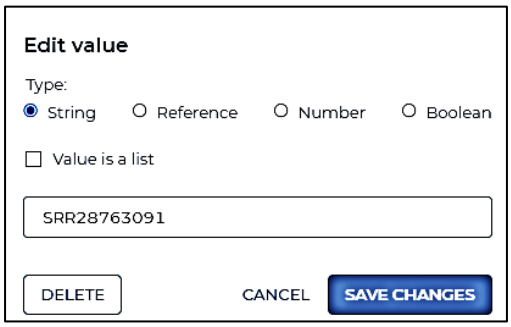
5.7.3.11.2.4.2 У спливаючому вікні "Додати новий стовпчик" назвіть новий стовпчик "sra_accession" і натисніть "Зберегти".



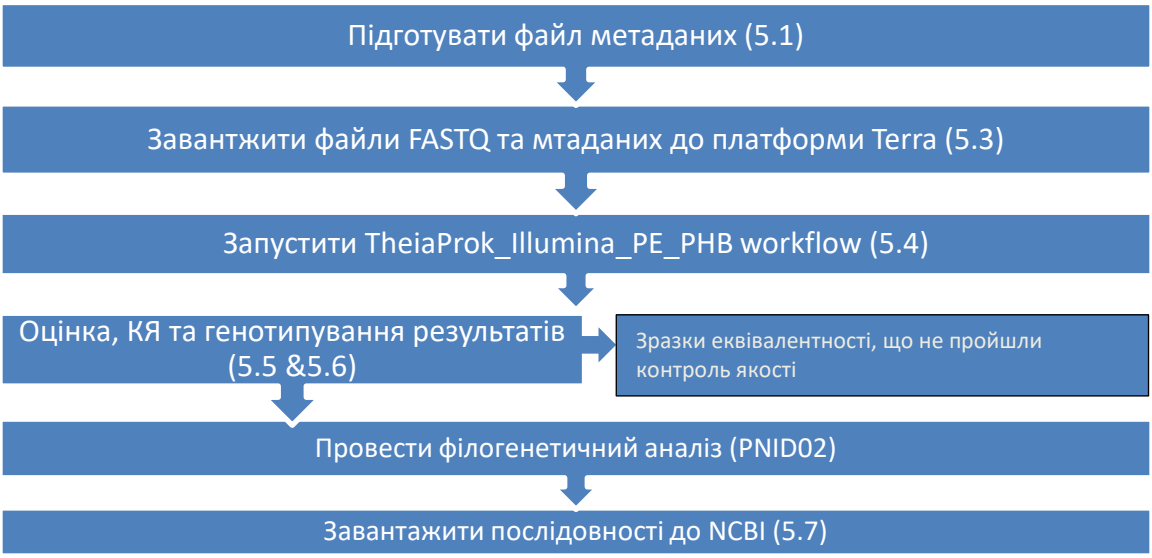
5.7.3.11.2.4.3 Новий стовпець має з'явитися в таблиці даних. Щоб ввести номер приєднання SRR для певної послідовності, натисніть "Редагувати значення" у стовпчику sra_accession для цього зразка.



5.7.3.11.2.4.4 У спливаючому вікні "Змінити значення" вставте номер приєднання до SRA у поле та натисніть "Зберегти зміни".



6. ТЕХНОЛОГІЧНА КАРТА:



7. ПОВ'ЯЗАНІ ДОКУМЕНТИ:

7.1 **PNID02:** Міжнародна стандартна операційна процедура PulseNet для філогенетичного аналізу даних WGS з використанням платформи Terra.Bio.

8. ВИКОРИСТАНА ЛІТЕРАТУРА:

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. КОНТАКТИ:

9.1 Лабораторія NGS PulseNet Центру контролю та профілактики захворювань США: pulsenetngslab@cdc.gov

9.2 Міжнародний координатор із забезпечення якості PulseNet Ейя Трієс: ehyytia-trees@cdc.gov

9.3 Theiagen:

9.3.1 Загальна електронна пошта для підтримки: support@theiagen.com

9.3.2 Мішель Скрібнер: michelle.scribner@theiagen.com

9.3.3 Френк Амбросіо: frank.ambrosio@theiagen.com

10. ЗМІНИ: Відсутні.

11. ПІДПИСИ ПРО ЗГОДУ:

Затверджено _____ Дата: _____
Персонал PulseNet ЗЯ/КЯ

Затверджено _____ Дата: _____
Технічний керівник PulseNet WGS

Затверджено _____ Дата: _____
Міжнародний координатор PulseNet

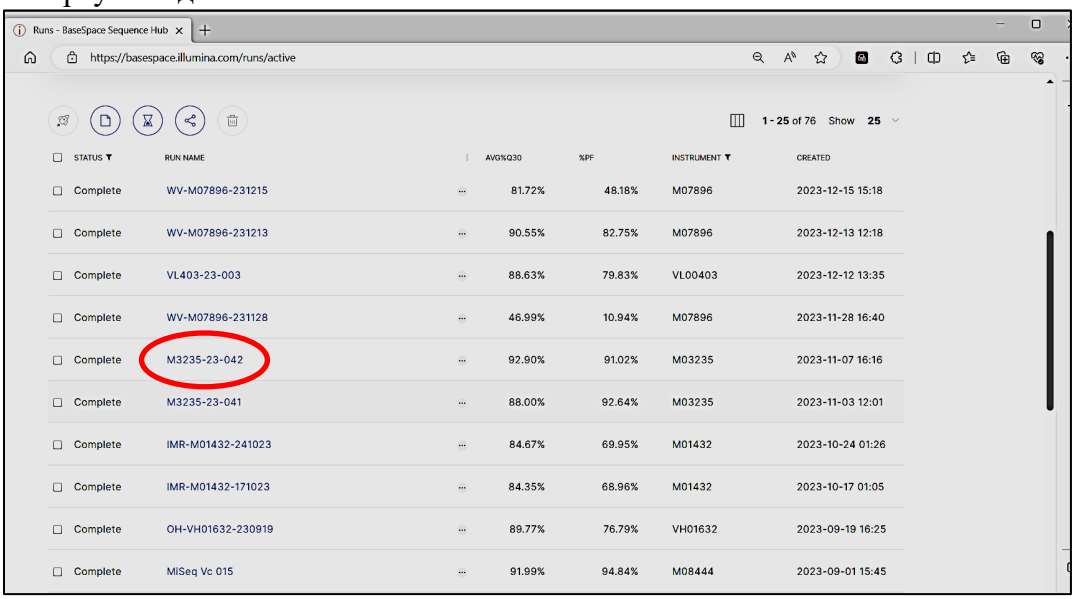
Затверджено_____ Дата: _____
Керівник групи реагування та управління спалахами PulseNet

Затверджено_____ Дата: _____
Завідувач сектору лабораторії кишкових захворювань

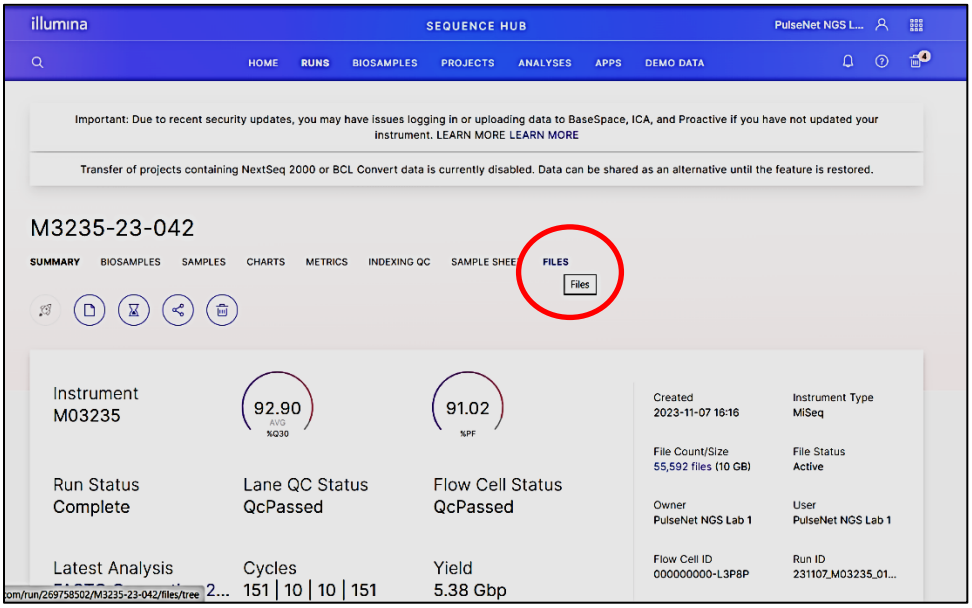
Додаток PNID01-1: Імпорт даних до Terra безпосередньо з Illumina BaseSpace

ПРИМІТКА: *Щоб налаштувати робочу область Terra для підключення до облікового запису Illumina BaseSpace, дотримуйтесь інструкцій на сайті ресурсів Theiagen за адресою https://theiagen.notion.site/BaseSpace_Fetch-34978656aa2d46ba82f2059434bd9369. Для отримання додаткової допомоги зверніться до компанії Theiagen (контактну інформацію див. у розділі Контакти (9)).*

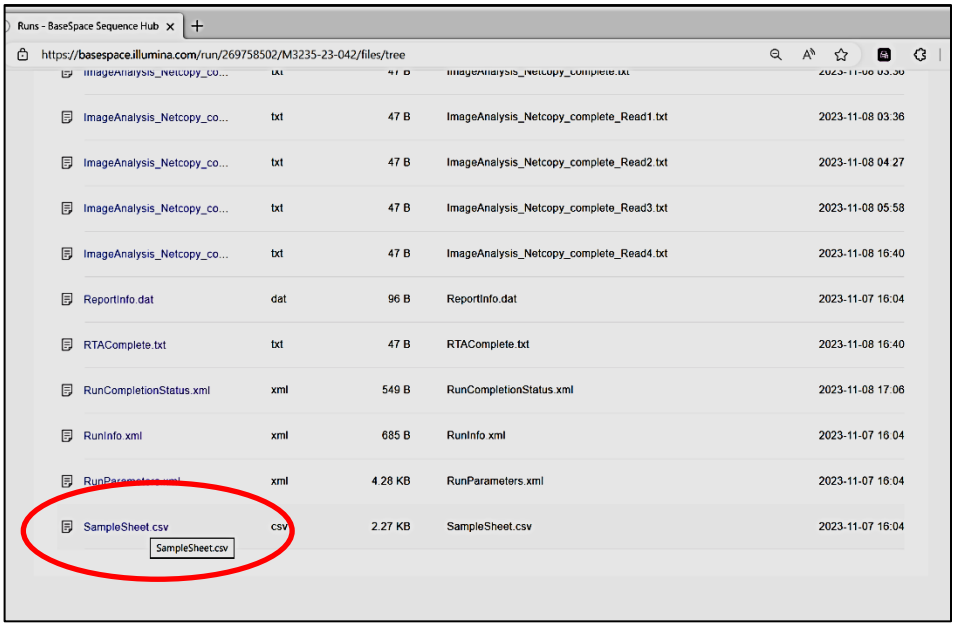
1. Увійдіть у свій обліковий запис BaseSpace і знайдіть цикл, який потрібно імпортувати до Terra.



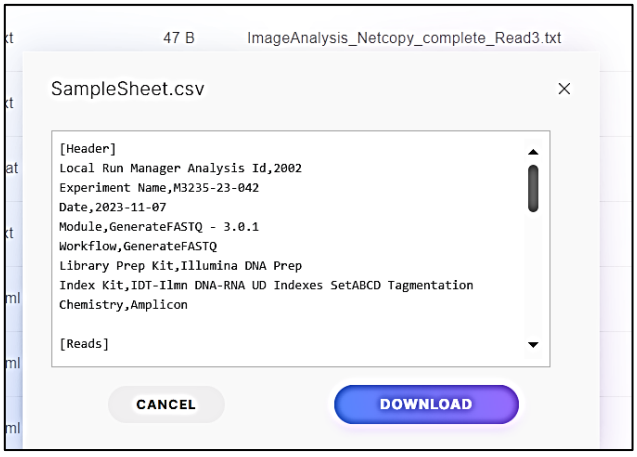
2. Завантажте Зразок для пробігу:
   a. Перейдіть на вкладку "Файли" і прокрутіть сторінку донизу.

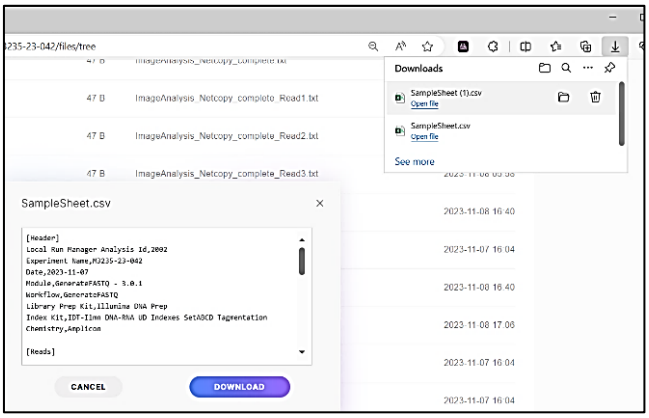


   b. Натисніть на посилання SampleSheet.csv.



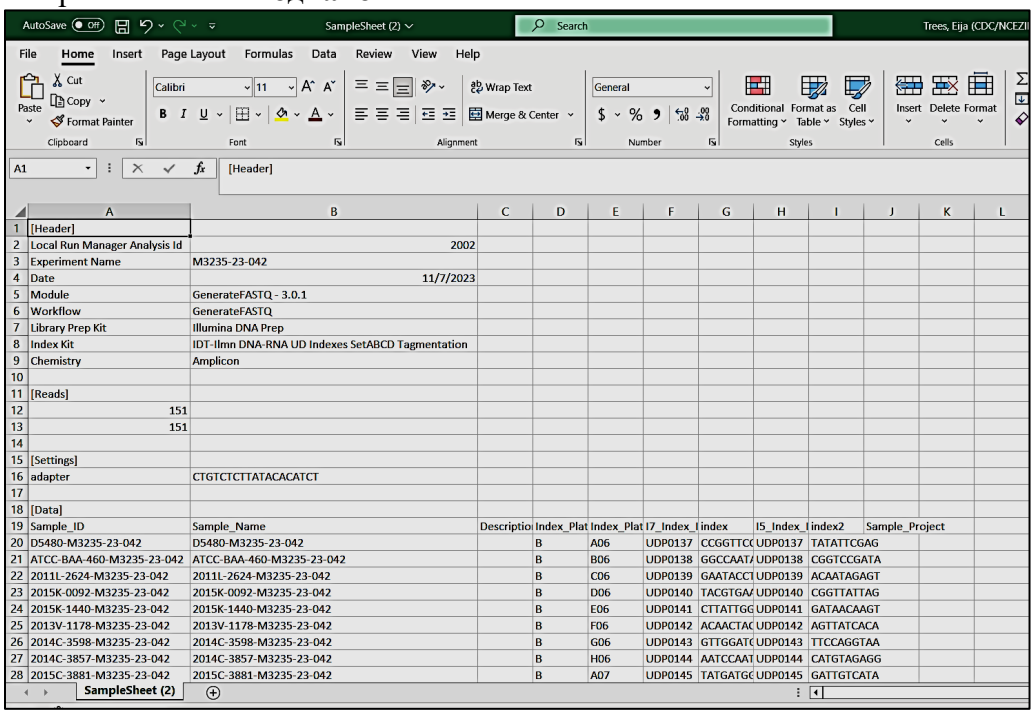
c. У спливаючому вікні "SampleSheet.csv" натисніть "Завантажити".





d. Завантажений csv-файл з'явиться у верхньому правому куті в розділі "Завантаження".

3. Відкрийте аркуш SampleSheet. Стовпці, необхідні у файлі метаданих tsv, залежатимуть від стовпців і вмісту стовпців, наявних на аркуші SampleSheet.
4. Підготуйте файл метаданих tsv:
   a. Стовпці потрібні, коли стовпці "Sample_Name" та "Sample_ID" в аркуші SampleSheet мають **однаковий** вміст:

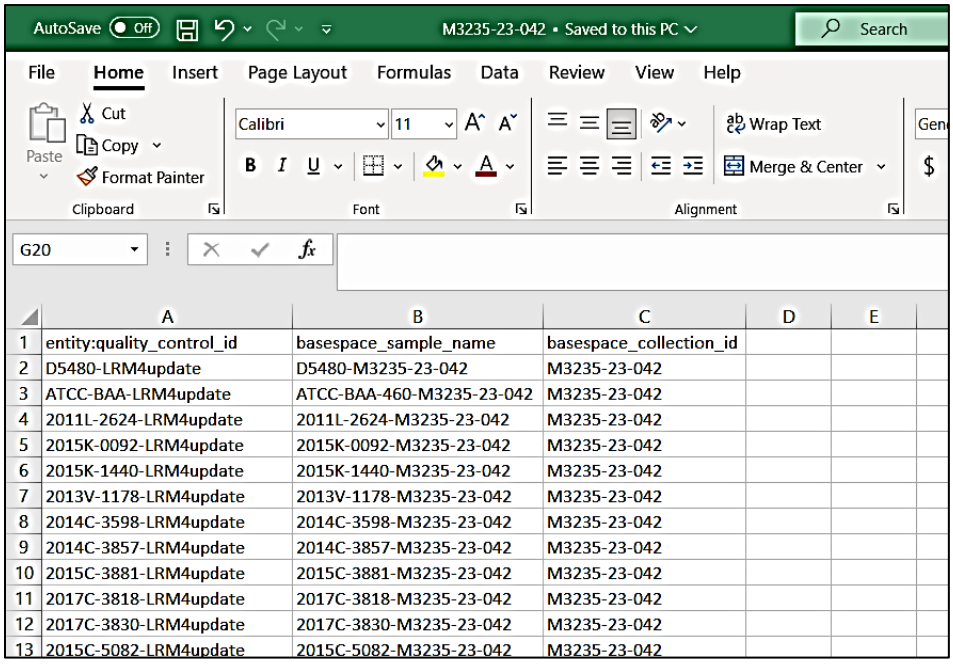


	A	B	C	D	E	F	G	H	I	J	K	L
1	[Header]											
2	Local Run Manager Analysis Id	2002										
3	Experiment Name	M3235-23-042										
4	Date	11/7/2023										
5	Module	GenerateFASTQ - 3.0.1										
6	Workflow	GenerateFASTQ										
7	Library Prep Kit	Illumina DNA Prep										
8	Index Kit	IDT-Ilmn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

i. **ідентифікатор_імені_бази_даних_об'єкта**:
   1. Введіть назву таблиці даних (нової або існуючої) в комірку A1 між "entity:" та "id".
   2. Введіть ідентифікатори зразків у стовпчик A так, як ви хочете, щоб вони з'явилися в таблиці даних Terra.

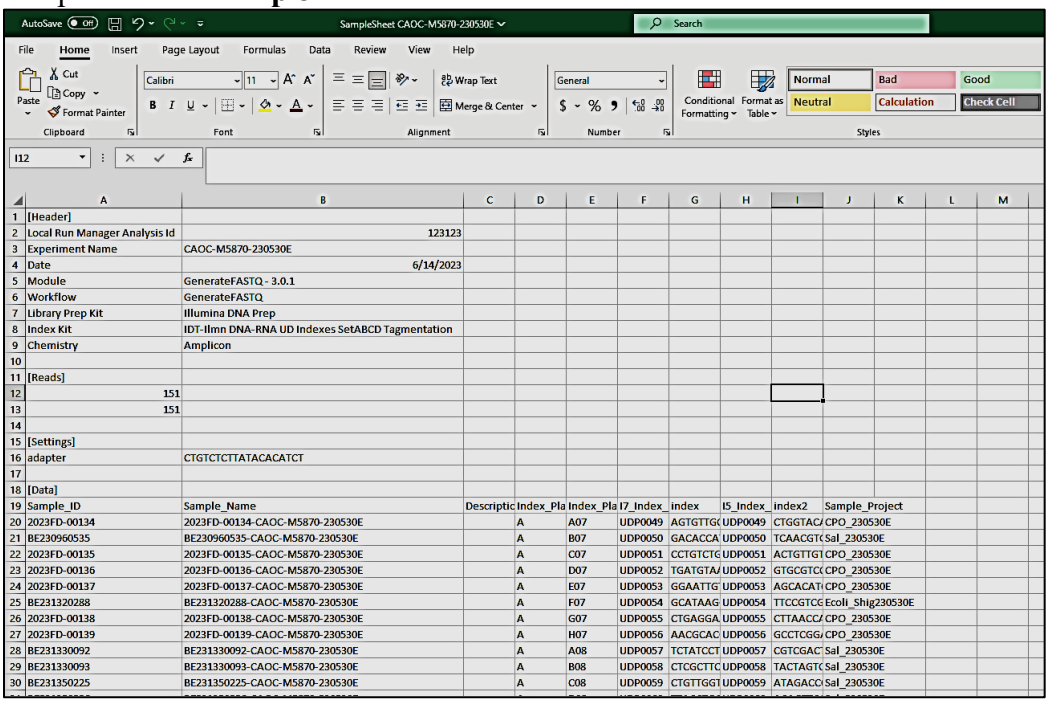
ii. basespace_sample_name: скопіюйте та вставте вміст з поля SampleSheet "Sample_Name".

iii. basespace_collection_id: введіть назву циклу так, як вона відображається у 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. Стовпці, необхідні, коли стовпці "Sample_Name" та "Sample_ID" на аркуші SampleSheet мають **різний** вміст:

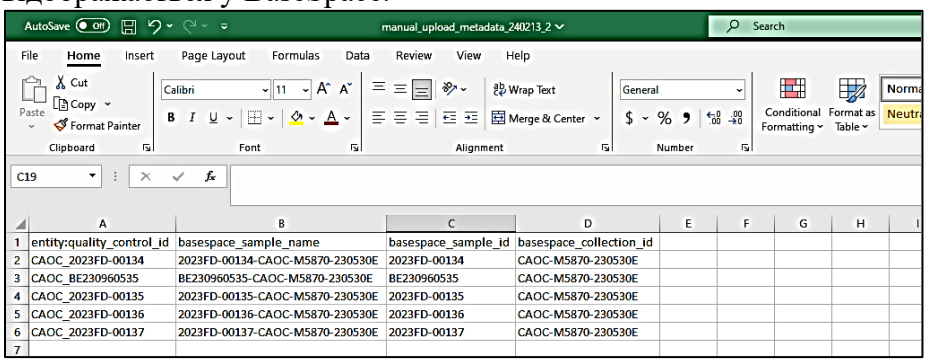


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	[Header]												
2	Local Run Manager Analysis Id	123123											
3	Experiment Name	CAOC-M5870-230530E											
4	Date	6/14/2023											
5	Module	GenerateFASTQ - 3.0.1											
6	Workflow	GenerateFASTQ											
7	Library Prep Kit	Illumina DNA Prep											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

i. **entity**_*datatablename*_**id**:

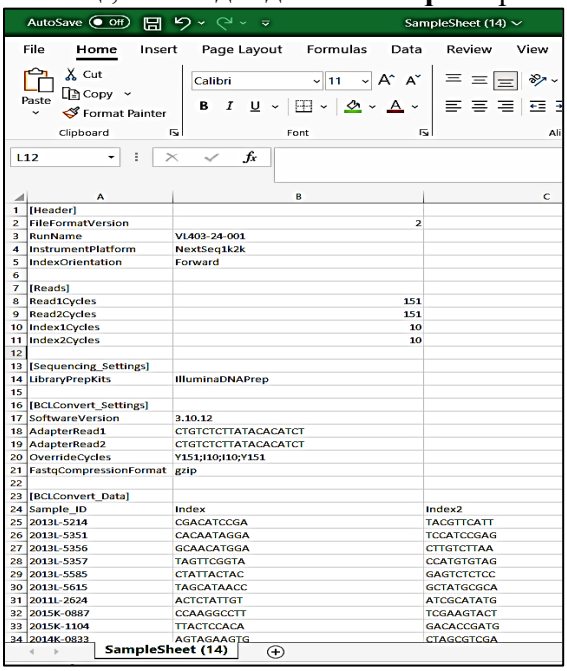
1. Введіть назву таблиці даних (нової або існуючої) в комірку A1 між "entity:" та "id".
2. Введіть ідентифікатори зразків у стовпчик A так, як ви хочете, щоб вони з'явилися в таблиці даних Terra.

ii. basespace_sample_name: скопіюйте та вставте вміст з поля SampleSheet "Sample_Name".
iii. basespace_sample_id: скопіюйте та вставте вміст з поля SampleSheet "Sample_ID".
iv. basespace_collection_id: введіть назву циклу так, як вона відображається у BaseSpace.



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

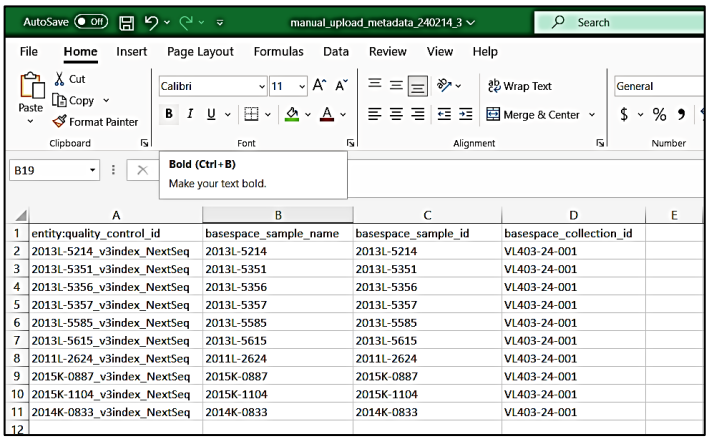
c. Стовпці, необхідні для **NextSeq** SampleSheet:



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

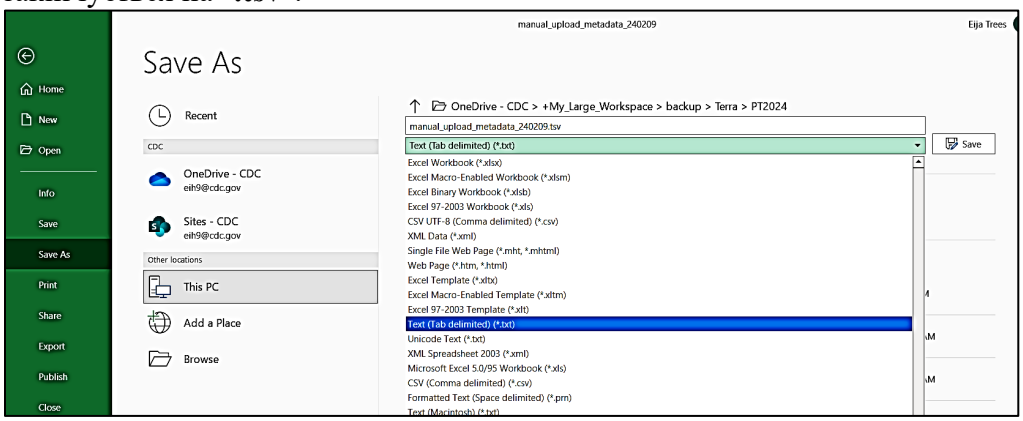
i. **entity**_*datatablename*_**id**:
    1. Введіть назву таблиці даних (нової або існуючої) в комірку A1 між "entity:" та "id".
    2. Введіть ідентифікатори зразків у стовпчик A так, як ви хочете, щоб вони з'явилися в таблиці даних Terra.
ii. basespace_sample_name: скопіюйте та вставте вміст з поля SampleSheet "Sample_ID".
iii. basespace_sample_id: скопіюйте та вставте вміст з поля SampleSheet "Sample_ID".

iv. basespace_collection_id: введіть назву циклу так, як вона відображається у BaseSpace.

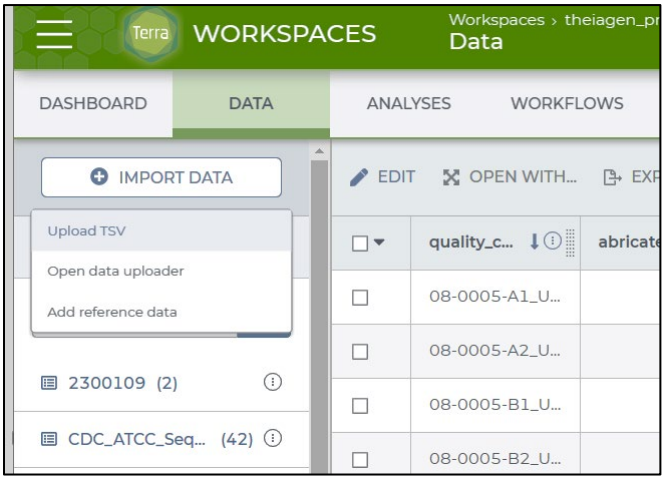


	A	B	C	D	E
1	entity:quality_control_id	basespace_sample_name	basespace_sample_id	basespace_collection_id	
2	2013L-5214_v3index_NextSeq	2013L-5214	2013L-5214	VL403-24-001	
3	2013L-5351_v3index_NextSeq	2013L-5351	2013L-5351	VL403-24-001	
4	2013L-5356_v3index_NextSeq	2013L-5356	2013L-5356	VL403-24-001	
5	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. Збережіть файл у **форматі tsv**: виберіть "Зберегти як" і "Текст (з розділенням табуляцією) (*.txt)". Переконайтеся, що ім'я вашого файлу закінчується на "**.tsv**".

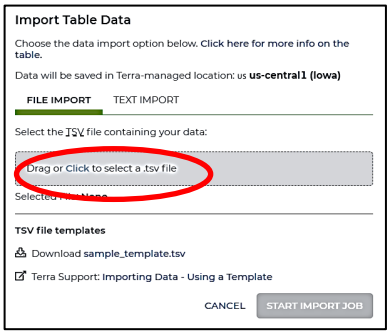


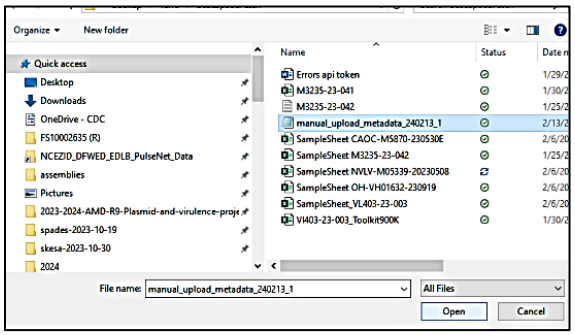
5. Імпортуйте файл метаданих tsv:
   a. На вкладці "Дані" натисніть "Імпортувати дані" і виберіть "Завантажити tsv" у випадаючому меню.



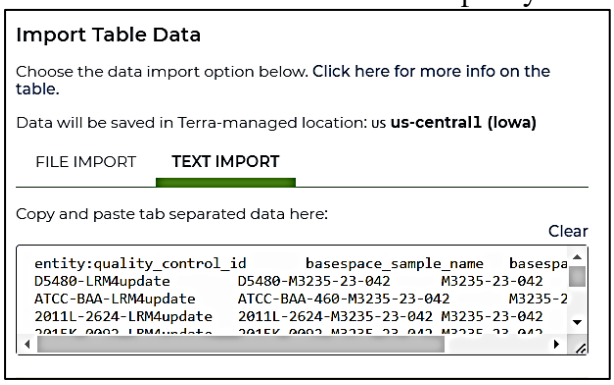
b. У спливаючому вікні "Імпортувати дані таблиці" ви можете імпортувати метадані двома способами:

   i. На вкладці "Імпорт файлів" натисніть посередині, щоб вибрати файл tsv, перейдіть до місця, де зберігається файл метаданих tsv, виберіть файл і натисніть "Відкрити".



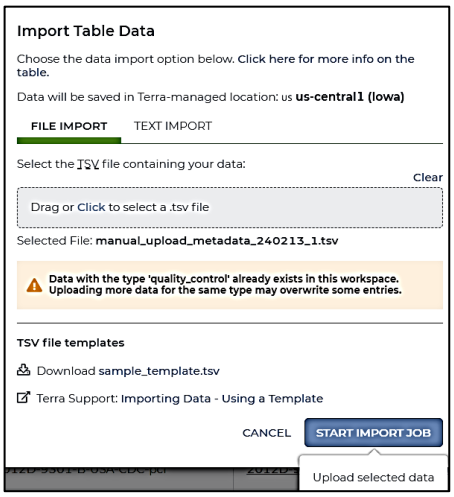


   ii. Крім того, ви можете перейти на вкладку "Імпорт тексту" і скопіювати та вставити вміст tsv-файлу в поле посередині.

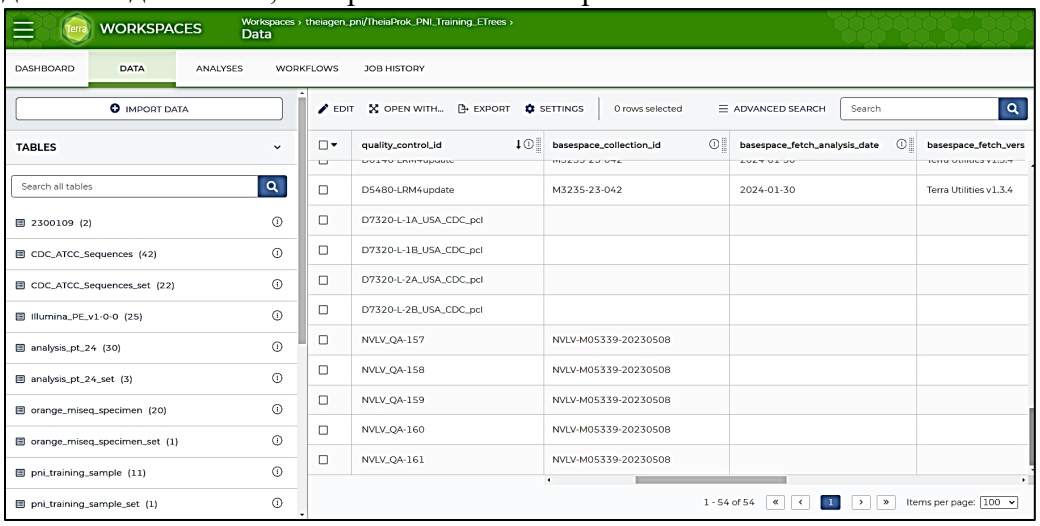


c. У спливаючому вікні "Імпортувати дані з таблиці" з'явиться попередження про те, що дані вже існують у відповідній таблиці даних (якщо імпорт

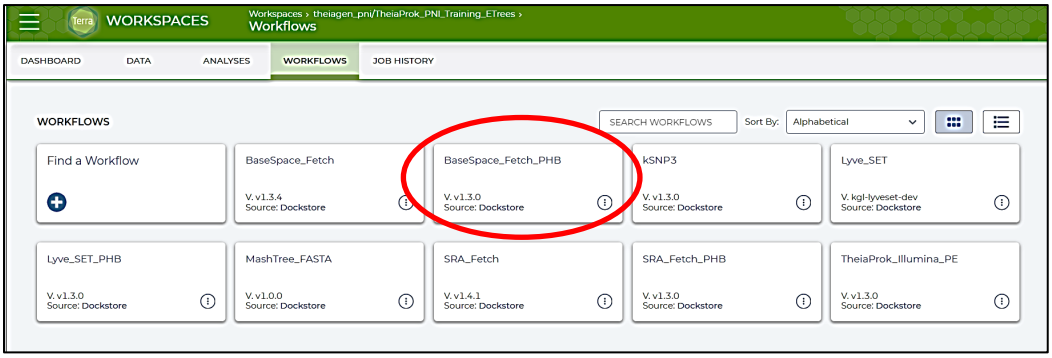
виконується в існуючу таблицю даних), і завантаження нових даних до неї може призвести до перезапису існуючих даних. Натисніть "Почати імпорт".



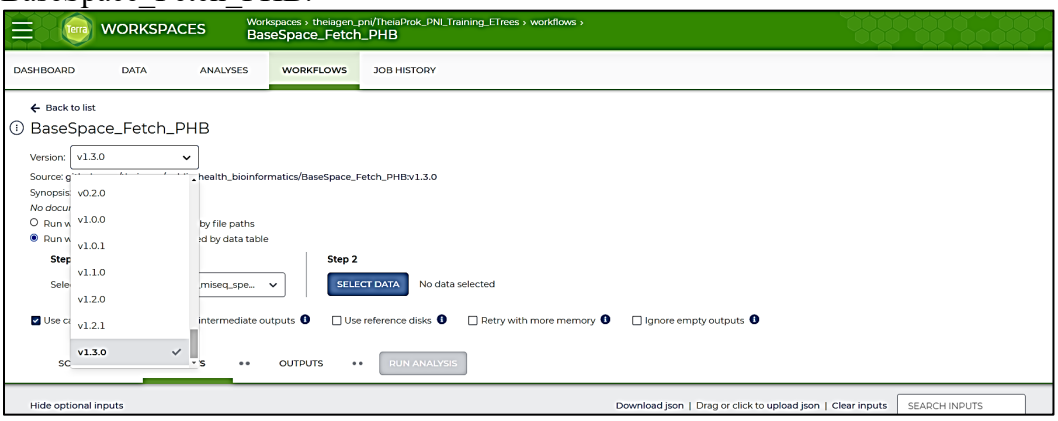
d. Після завершення імпорту ви побачите нові записи, створені в таблиці даних для послідовностей, імпортованих з BaseSpace.



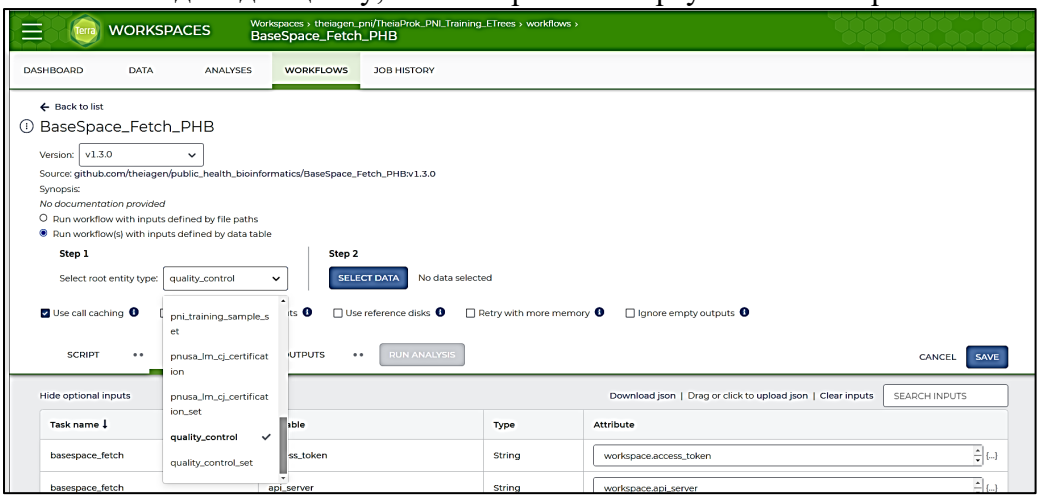
6. Запустіть робочий процес "BaseSpace_Fetch_PHB":
   a. На вкладці "Робочі процеси" натисніть "BaseSpace_Fetch_PHB". Відкриється вікно "BaseSpace_Fetch_PHB".



b. У випадаючому меню "Версія" виберіть останню версію BaseSpace_Fetch_PHB.

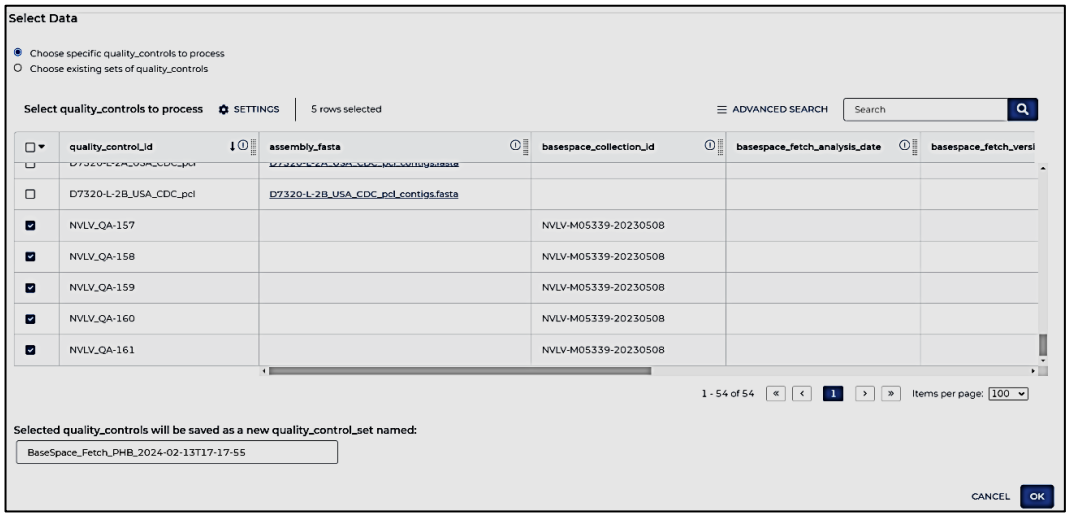


c. У розділі "Крок 1" натисніть на випадаюче меню " Select root entity type " і виберіть таблицю даних, до якої ви завантажили (кроки 4-5) tsv-файл, що містить метадані для циклу, який потрібно імпортувати з BaseSpace.

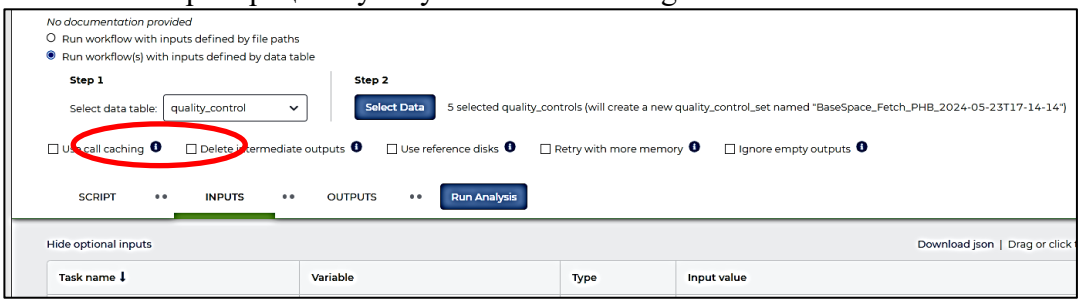


d. У розділі "Крок 2" натисніть "Вибрати дані" (скріншот вище). Ви потрапите на екран вибору вибірки.

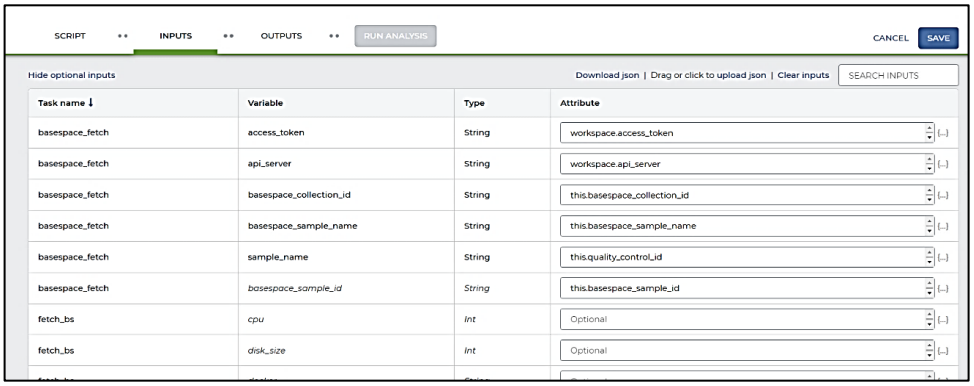
e. Встановіть прапорці навпроти зразків, які потрібно імпортувати з BaseSpace, і натисніть "ОК". Ви повернетесь до вікна "BaseSpace_Fetch_PHB".



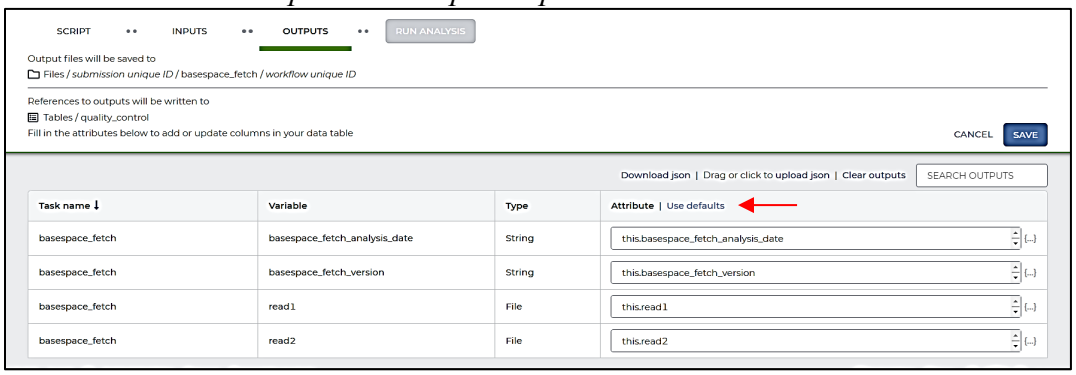
f. Зніміть прапорець з пункту "Use call catching".



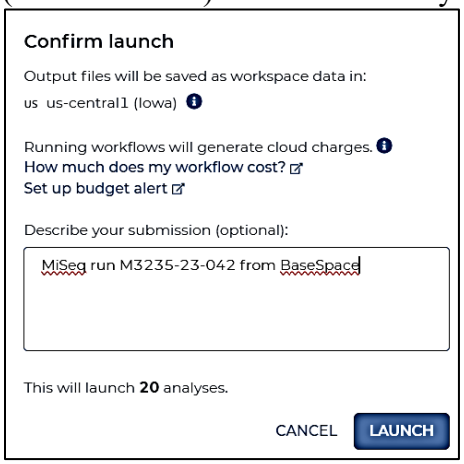
g. На вкладці "**Inputs**" визначте наступні змінні в полі "Атрибут":
ПРИМІТКА: *Коли ви заповнюєте стовпчик Атрибут, клацання всередині клітинки відкриє спадне меню атрибутів, які ви можете вибрати, щоб уникнути помилок.*
   i. access_token: "workspace.access_token".
   ii. api_server: "workspace.api_server".
   iii. basespace_collection_id: "this.basespace_collection_id".
   iv. basespace_sample_name: "this.basespace_sample_name".
   v. sample_name: "**this.dataslename_id**", наприклад, "this.quality_control_id".
   vi. basespace_sample_id: "this.basespace_sample_id".
   ПРИМІТКА: *потрібно заповнювати **лише** в **тому випадку**, якщо вміст полів "Sample_Name" та "Sample_ID" відрізняються у файлі SampleSheet циклу або ви імпортуєте дані з циклу NextSeq.*



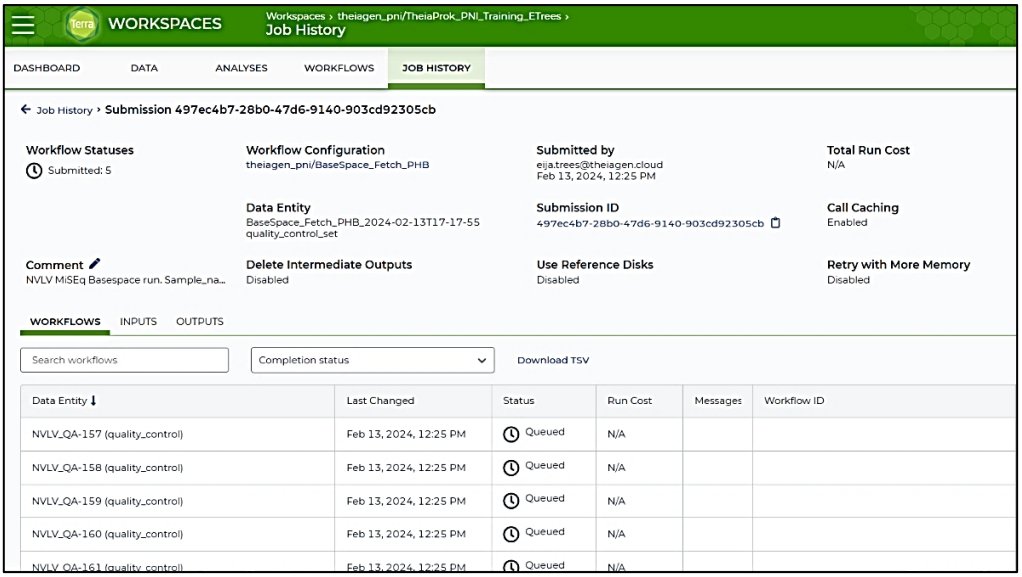
h. На вкладці "**Outputs**" натисніть "Використовувати за замовчуванням", потім "Зберегти", а потім "Запустити аналіз".
ПРИМІТКА: *Кнопка "Зберегти" відображається лише в тому випадку, якщо параметри (крім ідентифікаторів зразків) були змінені з моменту попереднього надсилання завдання. Кнопка "Виконати аналіз" стає підсвіченою після збереження параметрів.*



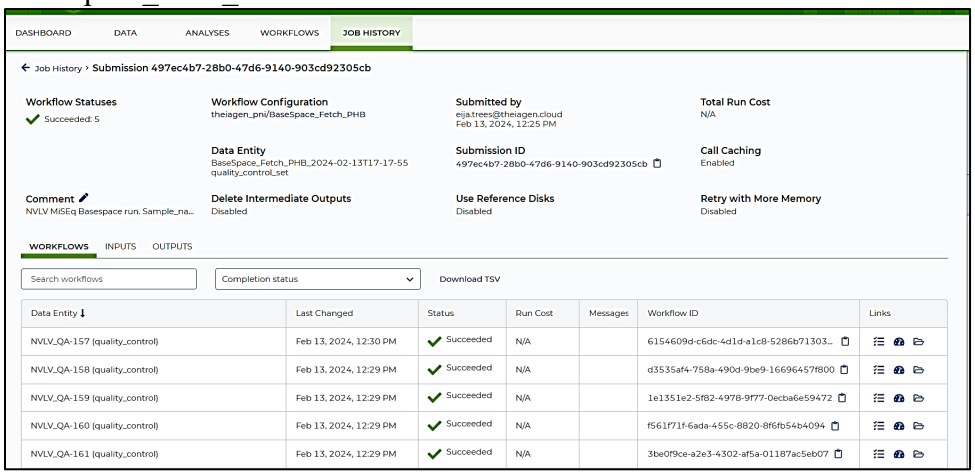
i. У спливаючому вікні "Підтвердити запуск" опишіть вашу заявку (необов'язково) і натисніть "Запустити".

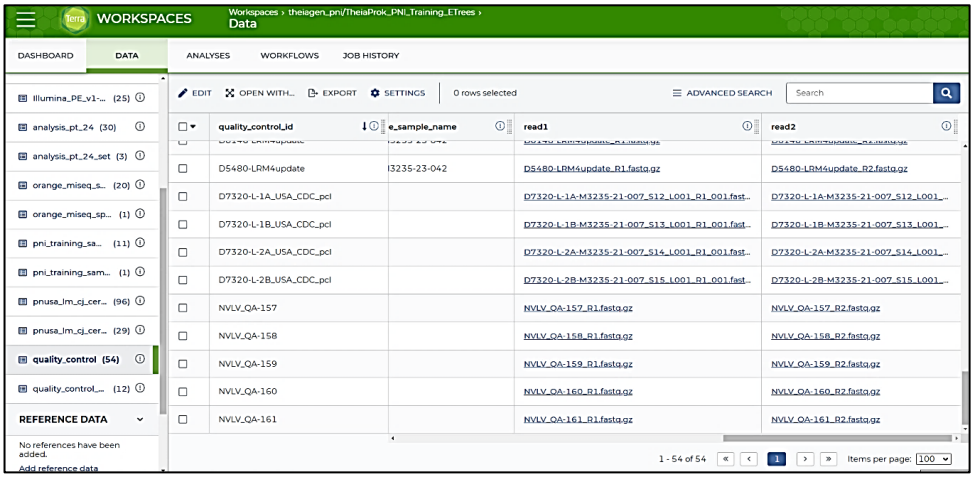


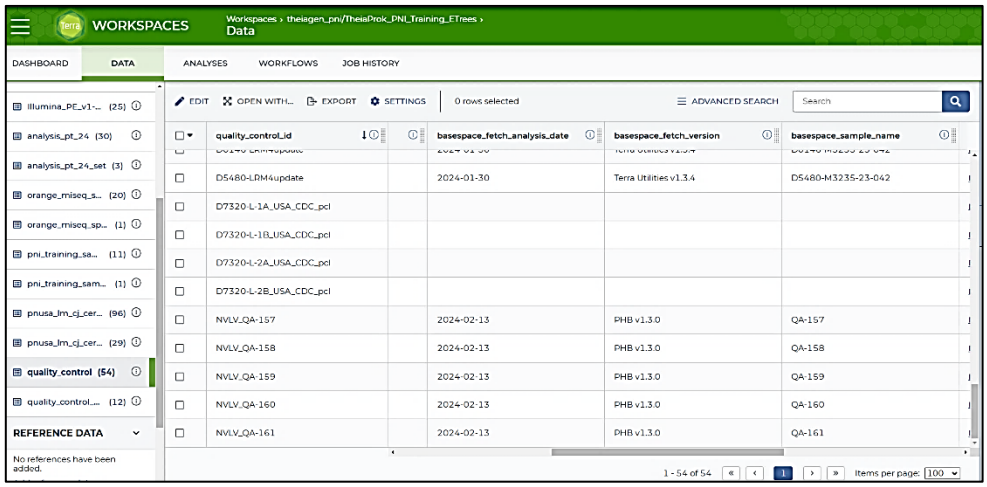
j. Відкриється вкладка "Історія робіт", де статус поданих заявок спочатку має бути "В черзі".



k. Після завершення роботи буде показано статус "Успішно" або "Виконано". На вкладці "Дані" ви побачите імена файлів FASTQ у стовпчиках “Read1” та “Read2”, а також інформацію у стовпчиках "basespace_fetch_analysis_date" і "basespace_fetch_version".

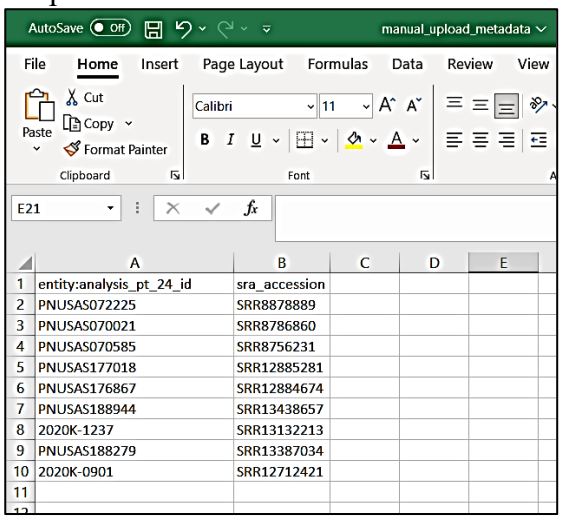



WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
quality_control_id
e_sample_name
read1
read2
D5480-LRM4update
D5480-LRM4update_R1.fastq.gz
D5480-LRM4update_R2.fastq.gz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


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

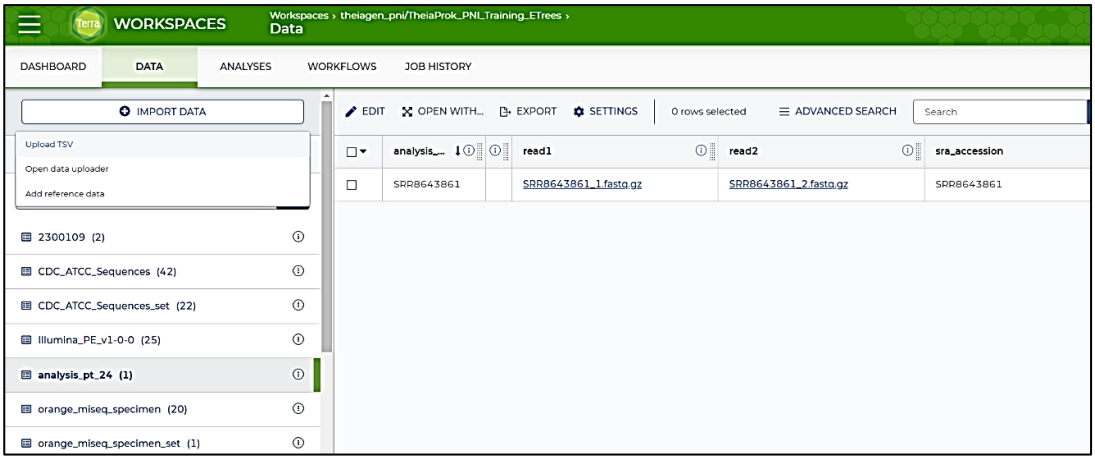
Додаток PNID01-2: Завантаження даних з NCBI SRA

1. Підготуйте файл метаданих tsv:
    a. Введіть назву таблиці даних (нової або існуючої) в комірку A1 між "entity:" та "id".
    b. Введіть ідентифікатори зразків у стовпчик A так, як ви хочете, щоб вони з'явилися в таблиці даних Terra.
    c. sra_accession: введіть номери доступу до SRA для послідовностей, які потрібно завантажити з SRA.

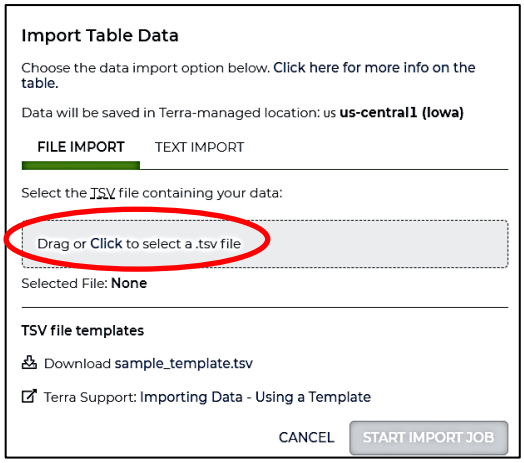


	A	B	C	D	E
1	entity:analysis_pt_24_id	sra_accession			
2	PNUSAS072225	SRR8878889			
3	PNUSAS070021	SRR8786860			
4	PNUSAS070585	SRR8756231			
5	PNUSAS177018	SRR12885281			
6	PNUSAS176867	SRR12884674			
7	PNUSAS188944	SRR13438657			
8	2020K-1237	SRR13132213			
9	PNUSAS188279	SRR13387034			
10	2020K-0901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

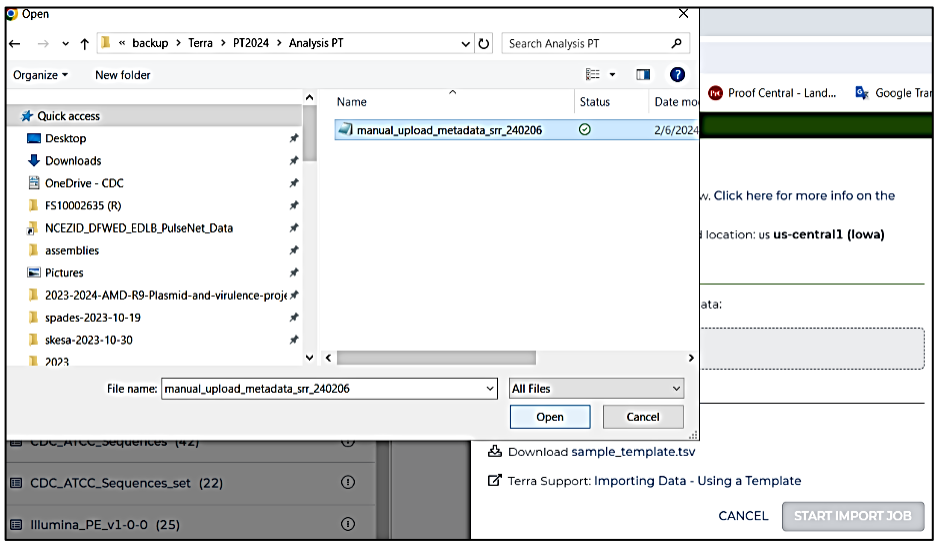
    d. Збережіть файл у **форматі tsv**: виберіть "Зберегти як" і "Текст (з розділенням табуляцією) (*.txt)". Переконайтеся, що ім'я вашого файлу закінчується на "**.tsv**".
2. На вкладці "Дані" натисніть "Імпортувати дані" і виберіть "Завантажити tsv" у випадаючому меню.



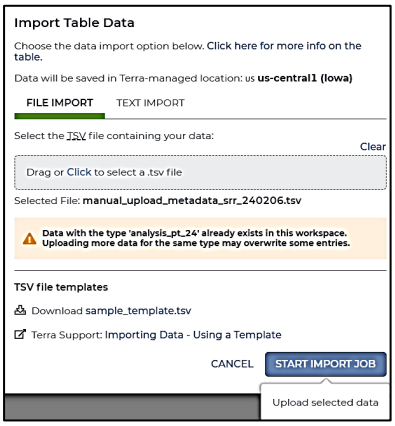
3. У спливаючому вікні "Імпорт даних таблиці" на вкладці "Імпорт файлів" клацніть посередині, щоб вибрати файл tsv.



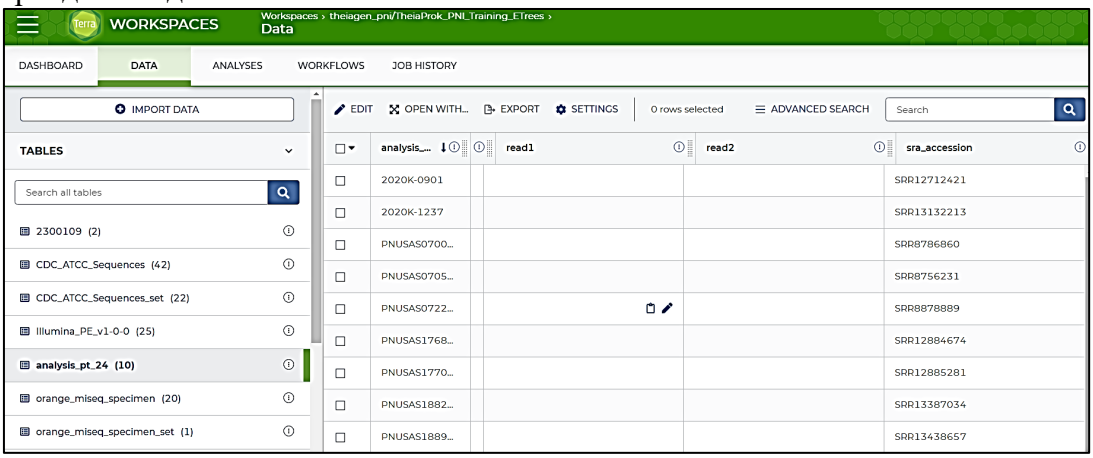
4. Перейдіть до місця, де зберігається файл метаданих tsv, виберіть його і натисніть "Відкрити".



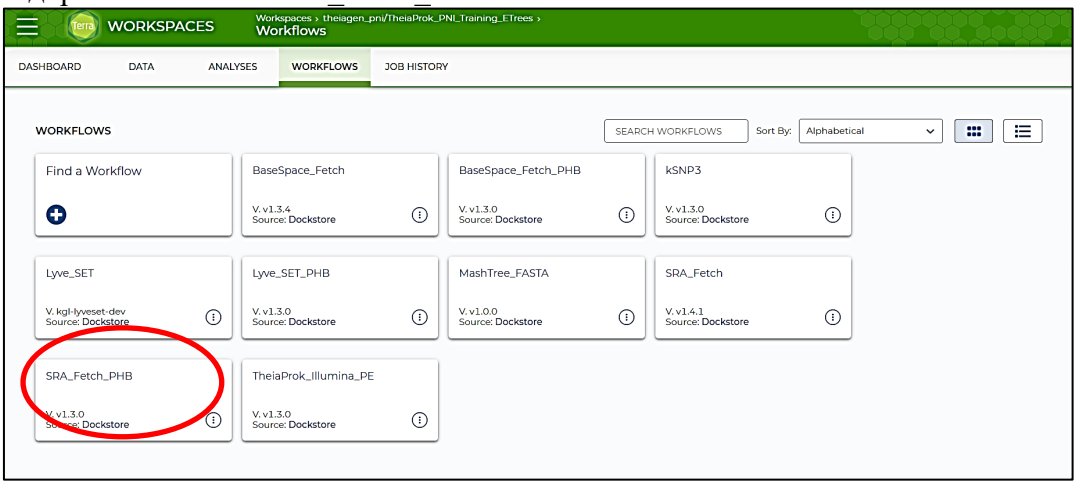
5. У спливаючому вікні "Імпортувати дані з таблиці" з'явиться попередження про те, що у відповідній таблиці даних вже існують дані (якщо імпорт виконується в існуючу таблицю даних), і завантаження нових даних до неї може призвести до перезапису існуючих даних. Натисніть "Почати імпорт".



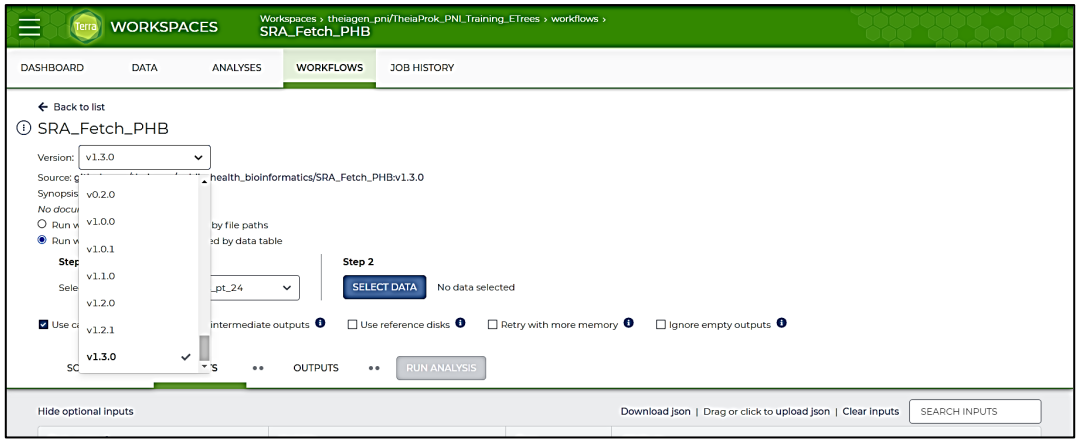
6. Після завершення імпорту ви побачите нові записи, створені в таблиці даних для послідовностей, які будуть завантажені з NCBI, разом з їхніми номерами приєднання до SRA.



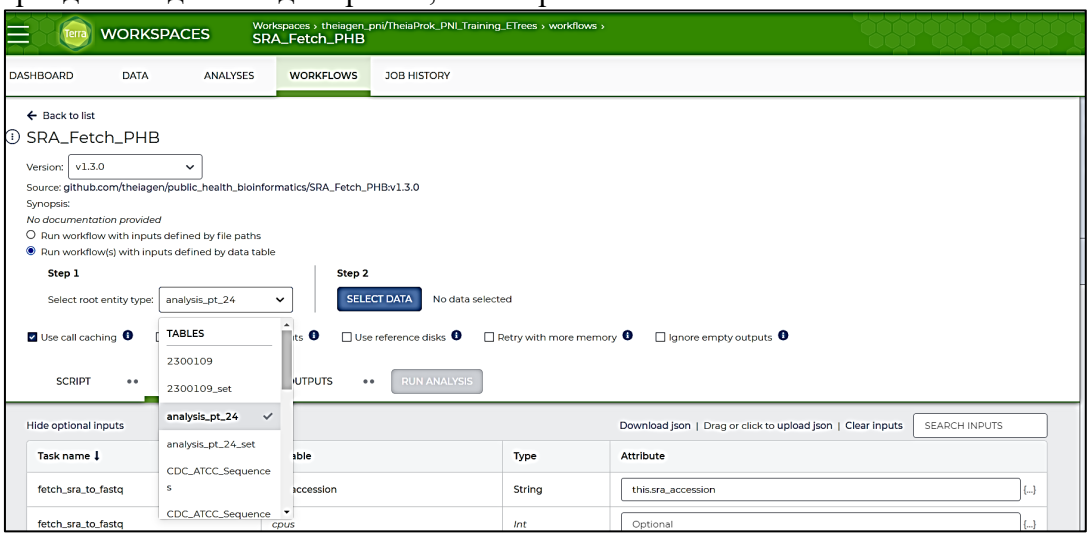
7. На вкладці "Робочі процеси" натисніть на робочий процес "SRA_Fetch_PHB". Відкриється вікно "SRA_Fetch_PHB".



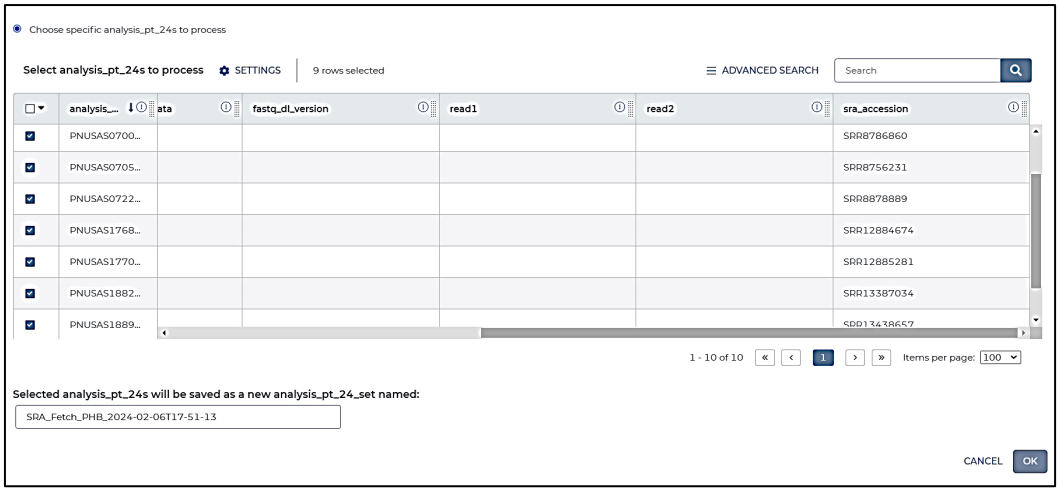
8. У випадаючому меню "Версія" виберіть останню версію SRA_Fetch_PHB.



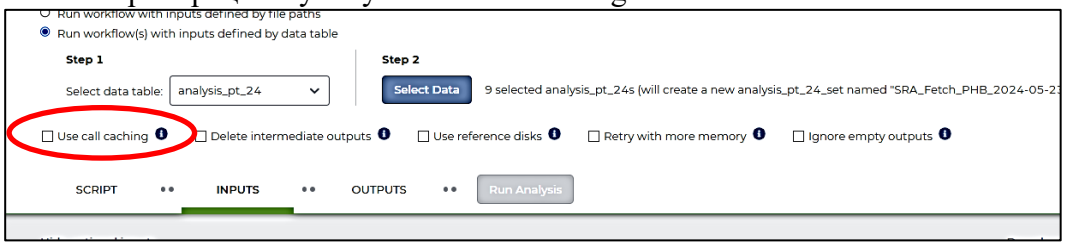
9. У розділі "Крок 1" натисніть на випадаюче меню "Select root entity type" і виберіть таблицю даних, до якої ви імпортували (кроки 1-6) tsv-файл, що містить номери приєднання до SRA для зразків, які потрібно завантажити з SRA.



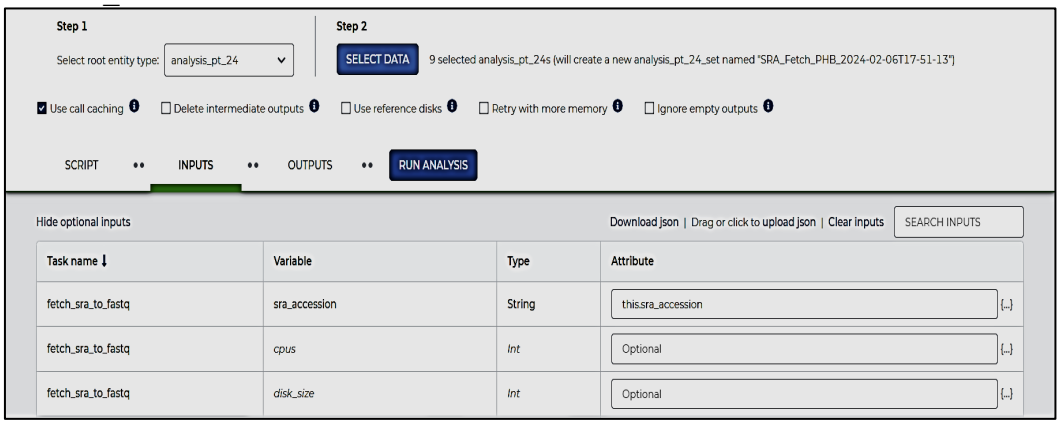
10. У розділі "Крок 2" натисніть "Вибрати дані" (скріншот вище). Ви потрапите на екран вибору вибірки.
11. Встановіть прапорці навпроти зразків, які потрібно завантажити з NCBI, і натисніть "ОК". Ви повернетесь на екран "SRA_Fetch_PHB".



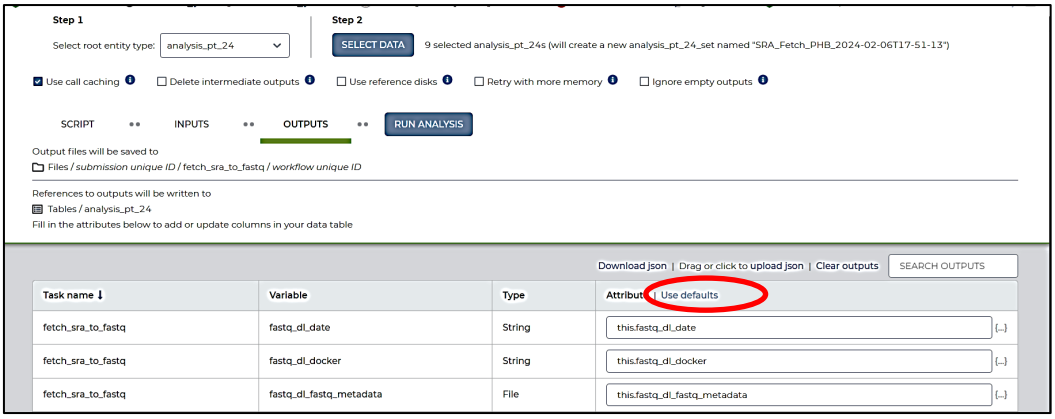
12. Зніміть прапорець з пункту " Use call catching ".



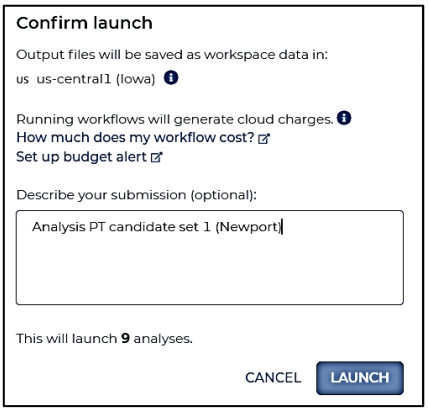
13. На вкладці "**Inputs**" у полі "Attribute" визначте змінну "sra_accession": "this.sra_accession".



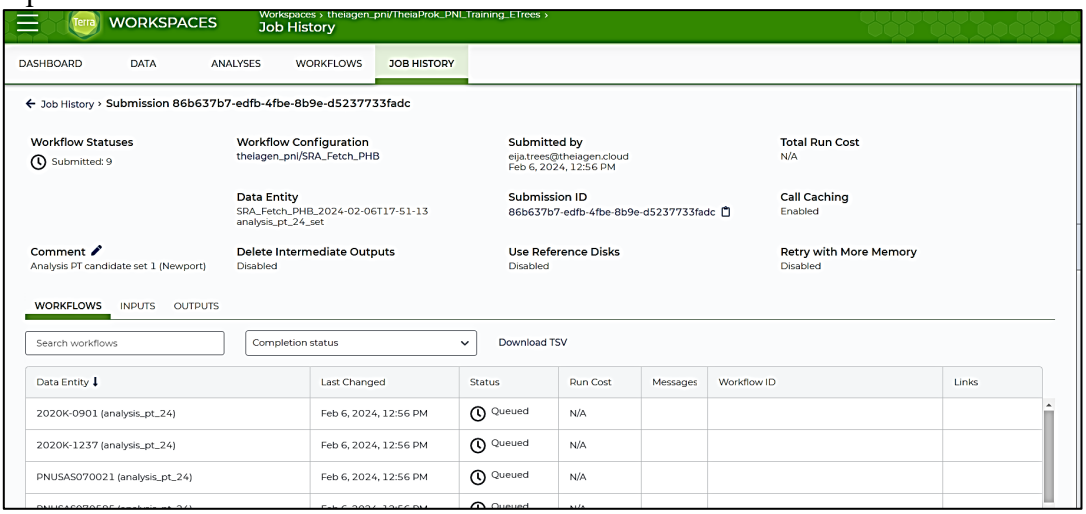
14. На вкладці "Результати" натисніть "Використовувати значення за замовчуванням", потім натисніть "Зберегти", а потім "Запустити аналіз". **ПРИМІТКА:** *Кнопка "Зберегти" відображається лише в тому випадку, якщо параметри (крім ідентифікаторів зразків) змінилися з моменту попереднього подання завдання.*



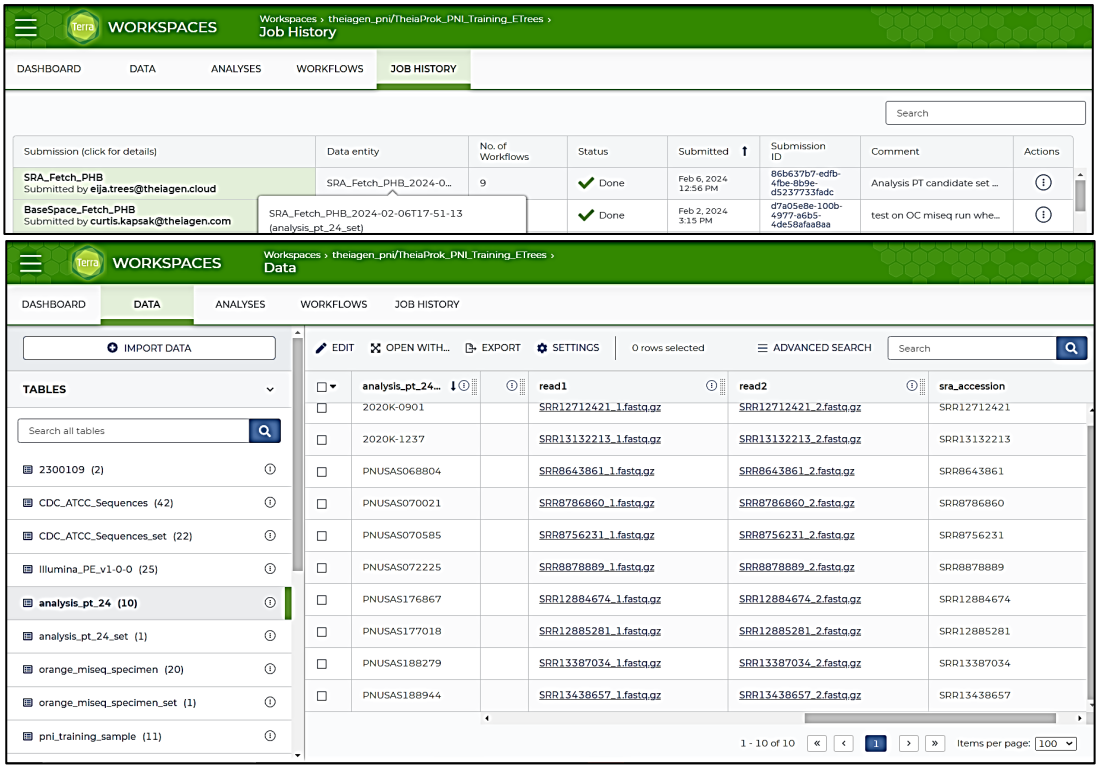
15. У спливаючому вікні "Підтвердити запуск" опишіть вашу заявку (необов'язково) і натисніть "Запустити".



16. Відкриється вкладка "Історія робіт", де статус поданих заявок спочатку має бути "В черзі".



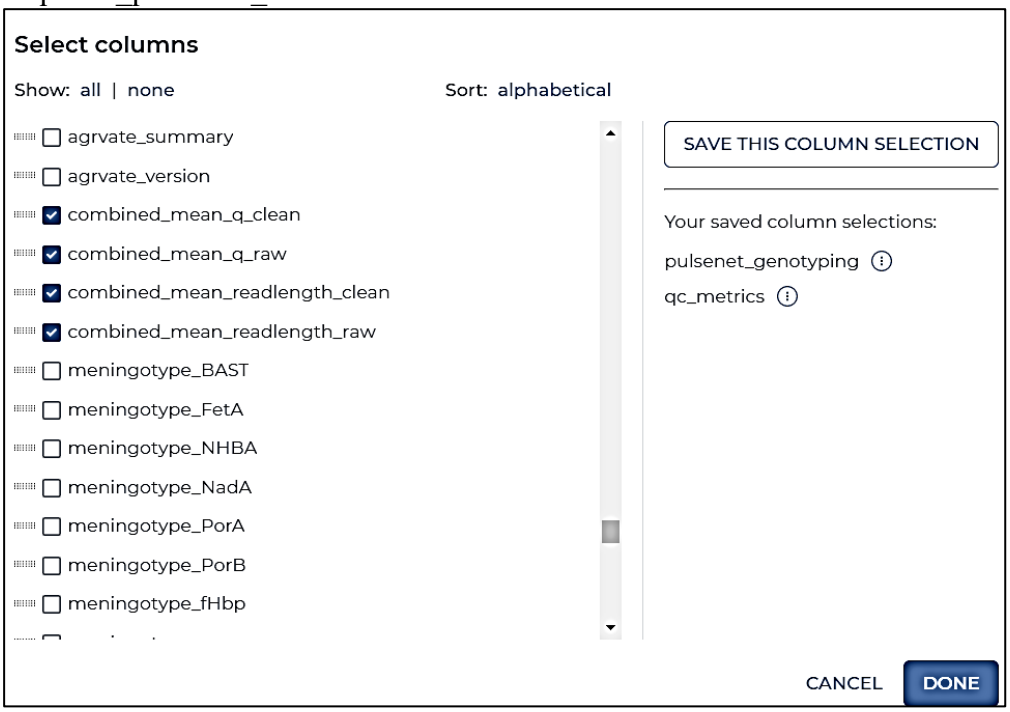
17. Після завершення роботи статус буде "Виконано". На вкладці "Дані" ви побачите імена файлів FASTQ у стовпчиках “Read1” і “Read2”.


WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Job History
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
Search
Submission (click for details)
Data entity
No. of Workflows
Status
Submitted
Submission ID
Comment
Actions
SRA_Fetch_PHB
Submitted by eija.trees@theiagen.cloud
SRA_Fetch_PHB_2024-0...
9
Done
Feb 6, 2024 12:56 PM
86b637b7-edfb-4fbe-8b9e-d5237733fadc
Analysis PT candidate set ...
BaseSpace_Fetch_PHB
Submitted by curtis.kapsak@theiagen.com
SRA_Fetch_PHB_2024-02-06T17-51-13 (analysis_pt_24_set)
Done
Feb 2, 2024 3:15 PM
d7a05e8e-100b-4977-a6b5-4de58afaa8aa
test on OC miseq run whe...
WORKSPACES
Workspaces › theiagen_pni/TheiaProk_PNI_Training_ETrees ›
Data
DASHBOARD
DATA
ANALYSES
WORKFLOWS
JOB HISTORY
IMPORT DATA
TABLES
Search all tables
2300109 (2)
CDC_ATCC_Sequences (42)
CDC_ATCC_Sequences_set (22)
Illumina_PE_v1-0-0 (25)
analysis_pt_24 (10)
analysis_pt_24_set (1)
orange_miseq_specimen (20)
orange_miseq_specimen_set (1)
pni_training_sample (11)
EDIT
OPEN WITH...
EXPORT
SETTINGS
0 rows selected
ADVANCED SEARCH
Search
analysis_pt_24...
read1
read2
sra_accession
2020K-0901
SRR12712421_1.fastq.gz
SRR12712421_2.fastq.gz
SRR12712421
2020K-1237
SRR13132213_1.fastq.gz
SRR13132213_2.fastq.gz
SRR13132213
PNUSAS068804
SRR8643861_1.fastq.gz
SRR8643861_2.fastq.gz
SRR8643861
PNUSAS070021
SRR8786860_1.fastq.gz
SRR8786860_2.fastq.gz
SRR8786860
PNUSAS070585
SRR8756231_1.fastq.gz
SRR8756231_2.fastq.gz
SRR8756231
PNUSAS072225
SRR8878889_1.fastq.gz
SRR8878889_2.fastq.gz
SRR8878889
PNUSAS176867
SRR12884674_1.fastq.gz
SRR12884674_2.fastq.gz
SRR12884674
PNUSAS177018
SRR12885281_1.fastq.gz
SRR12885281_2.fastq.gz
SRR12885281
PNUSAS188279
SRR13387034_1.fastq.gz
SRR13387034_2.fastq.gz
SRR13387034
PNUSAS188944
SRR13438657_1.fastq.gz
SRR13438657_2.fastq.gz
SRR13438657
1 - 10 of 10
Items per page: 100

Додаток PNID01-3: Налаштування подання таблиці даних для показників контролю якості PulseNet

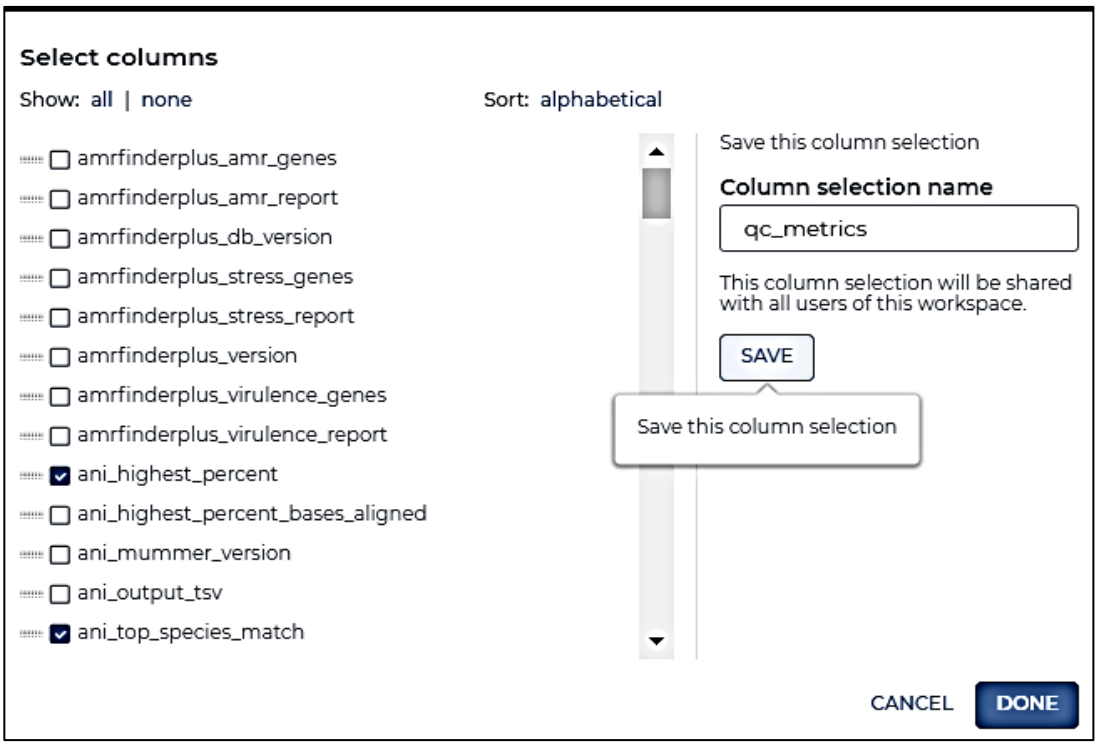
ПРИМІТКА: *Дане налаштування можна виконати лише після того, як ви один раз запустили робочий процес TheiaProk.*

1. На вкладці "Дані" виберіть таблицю даних, яка вас цікавить, наприклад, "CDC_ATCC_Sequences", а потім виберіть "Налаштування".
2. У розділі "Вибрати стовпці" слід позначити наступні метрики контролю якості:
    a. Ani_highest_percent
    b. Ani_top_species_match
    c. Assembly_length
    d. Combined_mean_q_clean
    e. Combined_mean_q_raw
    f. Combined_mean_readlength_clean
    g. Combined_mean_readlength_raw
    h. Est_coverage_clean
    i. Est_coverage_raw
    j. Gambit_predicted_taxon
    k. Midas_secondary_genus
    l. Midas_secondary_genus_abundance
    m. N50_value
    n. Number_contigs
    o. Raw_read_screen
    p. Seqsero2_predicted_contamination

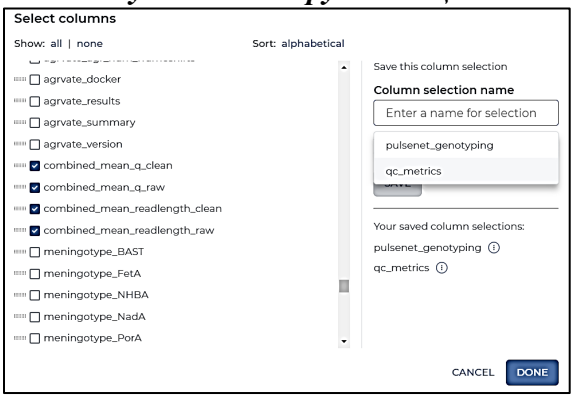


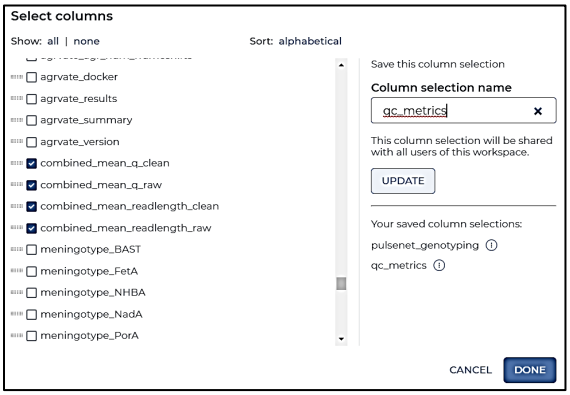
3. Натисніть "Зберегти цей вибір стовпця", назвіть вибір стовпця "qc_metrics" і натисніть "Зберегти", а потім натисніть "Готово".
   ПРИМІТКА: *Якщо ви додаєте або видаляєте стовпці з існуючого виділення стовпців, натисніть "Зберегти це виділення стовпців", виберіть назву зі спадного меню і натисніть "Оновити".*

Створення нового виділення стовпців



Зміна існуючого відбору стовпців





Додаток PNID01-4a. Критичні показники якості PulseNet (проходження/непроходження) для подачі рутинних послідовностей

Організм	Середнє охоплення denovo[1]	Середня якість (Q score)[2]	Довжина збірки (МБ)	Чисельність вторинних видів
Listeria monocytogenes	≥ 20x	≥ 30	2.8-3.2	≤ 0.01
E. coli **(most serotypes)**	≥ 40x	≥ 30	4.9-6.0	≤ 0.01
Shigella spp./**Rare *E. coli***	≥ 40x	≥ 30	4.2-4.9	≤ 0.01
Salmonella spp.	≥ 30x	≥ 30	4.4-5.7	≤ 0.01
Campylobacter spp.	≥ 20x	≥ 30	1.4-2.2	≤ 0.01
Vibrio cholerae	≥ 40x	≥ 30	3.8-4.3	≤ 0.01
Vibrio parahaemolyticus	≥ 40x	≥ 30	4.9-5.5	≤ 0.01
Vibrio vulnificus	≥ 40x	≥ 30	4.7-5.3	≤ 0.01

[1]Після обрізання на основі якості (est_coverage_clean)
[2]Перед обрізанням (combined_mean_q_raw)

Додаток PNID01-4b. Етап попереднього скринінгу зчитування TheiaProk для виключення неякісних послідовностей з метою економії обчислювальних ресурсів

Скринінгове завдання гарантує, що кількість даних про послідовності достатня для проведення геномного аналізу. Воно використовує команди bash для кількісного підрахунку зчитувань і базових пар, а також mash-схемування для оцінки розміру геному та його покриття. На кожному кроці результати оцінюються відповідно до критеріїв "пройшов/не пройшов" і порогових значень, які можуть бути визначені додатковими даними користувача.

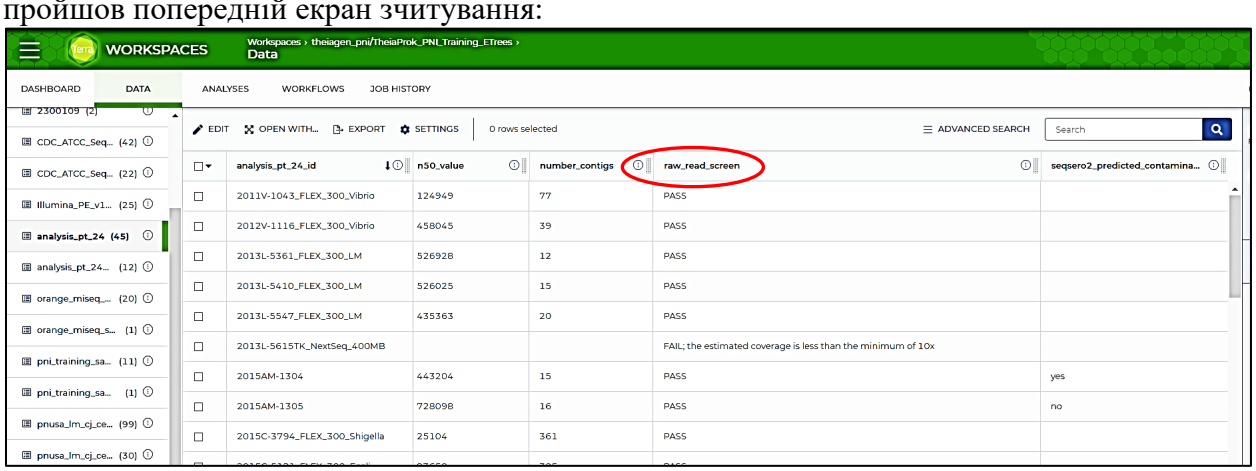
Зразки, які не відповідають цим критеріям, не будуть далі оброблятися робочим процесом:

1. Загальна кількість зчитувань: Зразок не пройде перевірку на читання, якщо його загальна кількість прочитань менша або дорівнює min_reads.
2. Пропорція базових пар, прочитаних у файлах прямого і зворотного зчитування: Зразок не пройде перевірку на зчитування, якщо у файлах reads1 або read2 буде менше, ніж min_proportion базових пар.
3. Кількість базових пар: Зразок не пройде перевірку на читання, якщо кількість базових пар менша за min_basepairs
4. Розрахунковий розмір геному: Зразок не пройде скринінг зчитування, якщо передбачуваний розмір геному менший за min_genome_size або більший за max_genome_size.
5. Розрахункове покриття геному: Зразок не пройде скринінг зчитування, якщо оцінене покриття геному менше, ніж min_coverage.

Значення за замовчуванням:
Int min_reads = 7472

Int min_basepairs = 2241820
Int min_genome_length = 100000
Int max_genome_length = 18040666
Int min_coverage = 10
Int min_proportion = 40

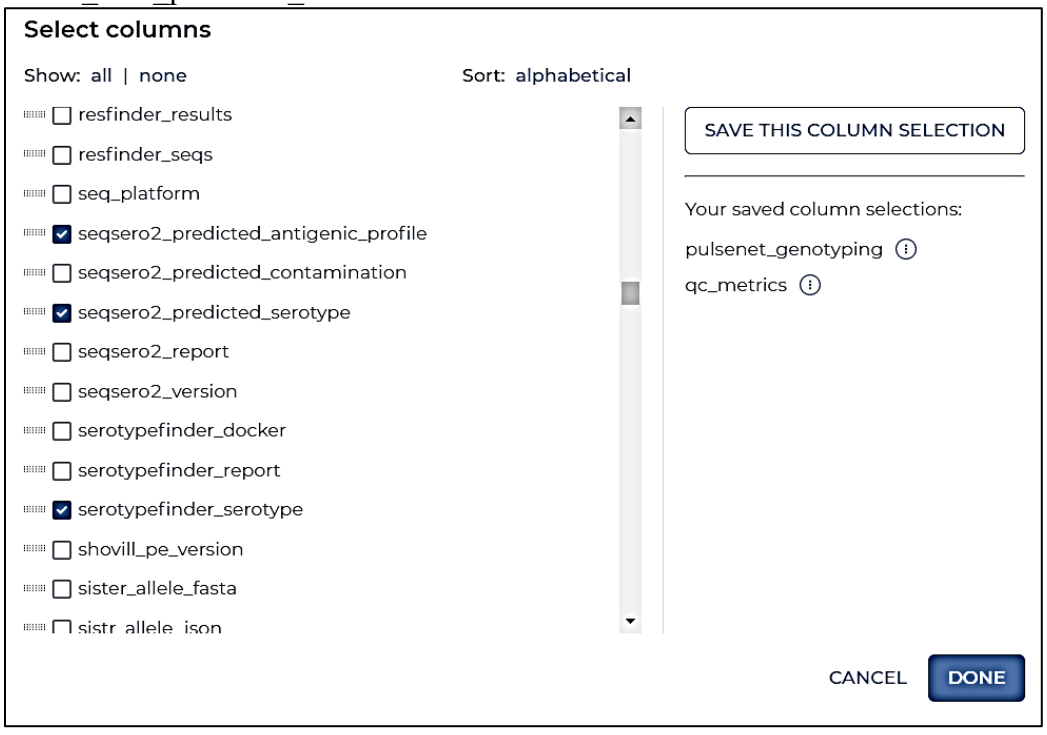
У колонці "raw_read_screen" у розділі "QC_metrics" буде показано, якщо зразок не пройшов попередній екран зчитування:



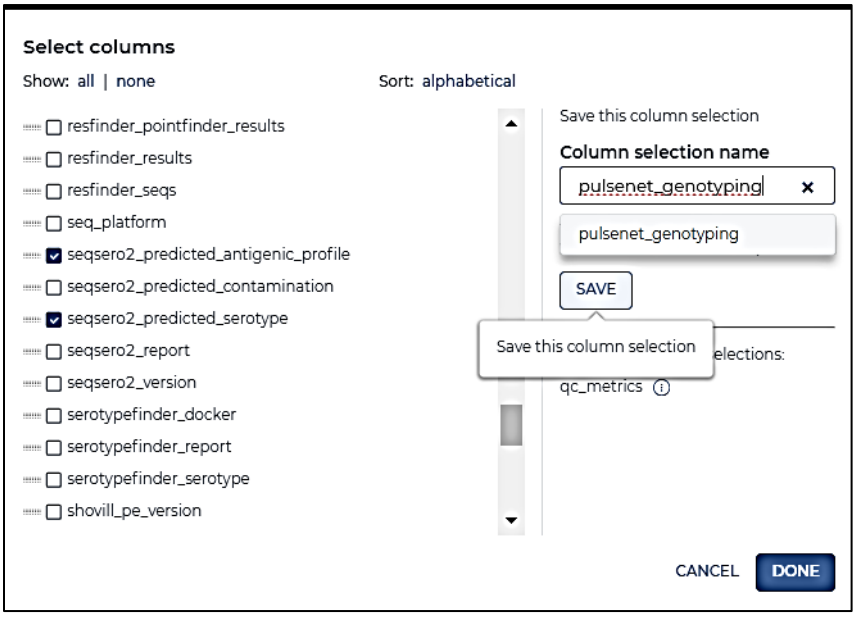
Додаток PNID01-5. Налаштування подання таблиці даних для аналізів генотипування PulseNet

ПРИМІТКА: *це налаштування можна виконати лише після того, як ви один раз запустили робочий процес TheiaProk.*

1. На вкладці "Дані" виберіть таблицю даних, яка вас цікавить, наприклад, "CDC_ATCC_Sequences", а потім виберіть "Налаштування".
2. У розділі "Вибрати стовпчики" слід позначити наступні генотипувальні аналізи:
    a. Amrfinderplus_amr_classes
    b. Amrfinderplus_amr_core_genes
    c. Amrfinderplus_amr_subclasses
    d. Amrfinderplus_virulence_genes
    e. Plasmidfinder_plasmids
    f. Seqsero2_predicted_antigenic profile
    g. Seqsero2_predicted_serotype
    h. Serotypefinder_serotype
    i. Ts_mlst_predicted_st

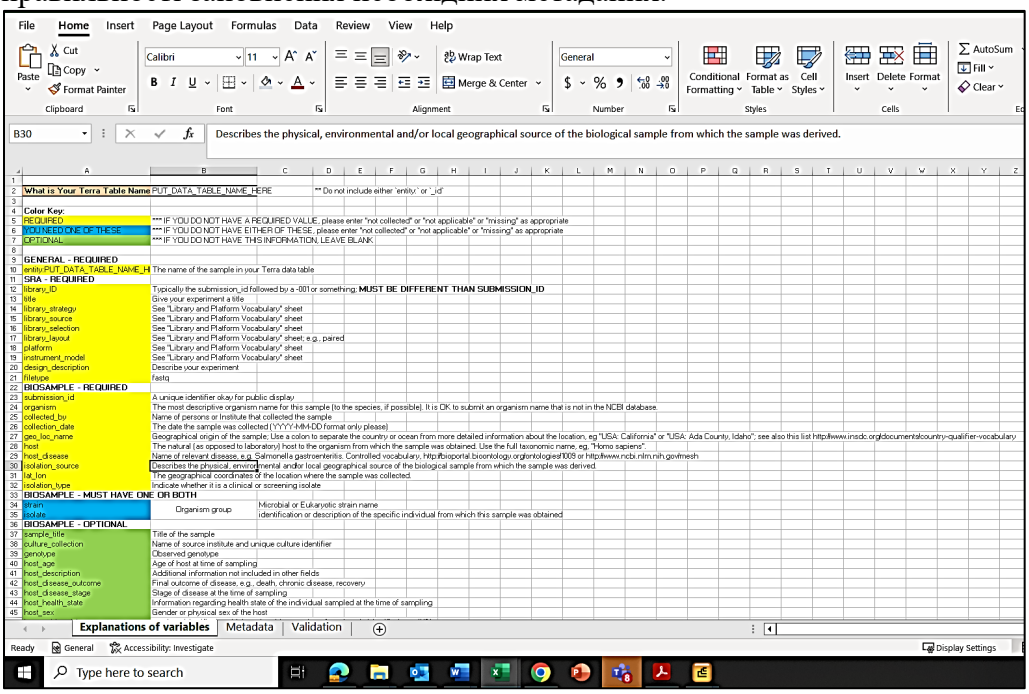


3. Натисніть "Зберегти цей вибір стовпця", назвіть вибір стовпця "pulsenet_genotyping" і натисніть "Зберегти", а потім натисніть "Готово".
ПРИМІТКА: *Якщо ви додаєте або видаляєте стовпці з існуючого виділення стовпців, натисніть "Зберегти це виділення стовпців", виберіть назву зі спадного меню і натисніть "Оновити".*


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

Додаток PNID01-6. Завантаження додаткових метаданих до Terra для подання до NCBI та налаштування вигляду таблиці даних для метаданих

- NCBI вимагає завантаження мінімальних метаданих для створення біозразків для послідовностей, які будуть завантажені в SRA.
- Оскільки метадані NCBI мають бути відформатовані певним чином, будь ласка, використовуйте шаблон метаданих **патогенів**, наданий компанією Theiagen, для завантаження метаданих до Terra: https://theiagen.notion.site/Terra_2_NCBI-8f014c73acc44465a3d69cf4df93adfe.
- Шаблон метаданих має три вкладки:
    - Перша вкладка, яка називається "Пояснення змінних", містить опис обов'язкових і необов'язкових полів.
    - Метадані для завантаження вводяться на другій вкладці "Метадані".
    - Третя вкладка "Валідація" може бути використана для перевірки правильності заповнення необхідних метаданих.



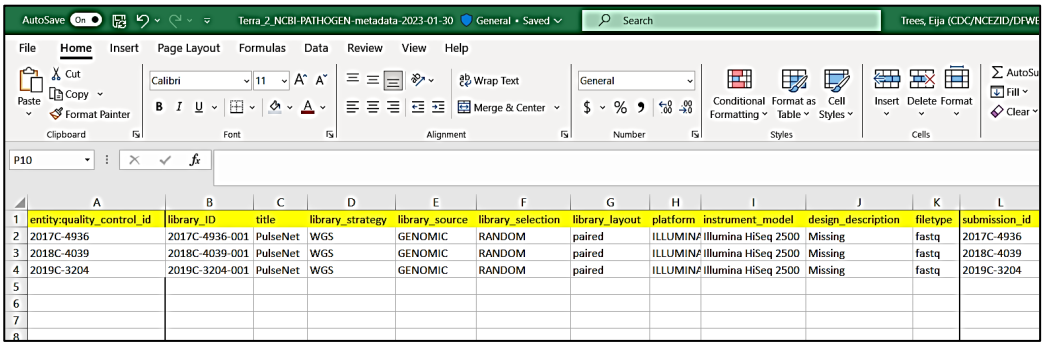
Щоб завантажити файл метаданих патогену до Terra:

1. Заповніть обов'язкові та необов'язкові (за необхідності) поля на вкладці метаданих в електронній таблиці шаблону метаданих патогену:
   ПРИМІТКА 1: *ви можете ввести "Відсутня" для будь-якої необхідної інформації, якої у вас немає або яку ви не хочете публічно розголошувати.*
   ПРИМІТКА2: *Наведені нижче метадані є мінімальними вимогами PulseNet USA до метаданих для завантаження в NCBI, розробленими з метою захисту конфіденційності пацієнтів і цілісності поточних досліджень спалахів. З іншого*

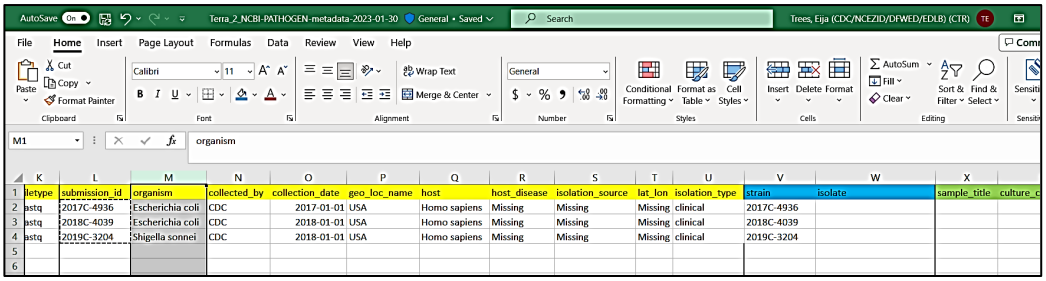
боку, надається достатньо епідеміологічно корисної інформації, особливо щодо неклінічних зразків, щоб полегшити визначення атрибутів.

a. У клітинку A1 введіть назву таблиці даних, до якої будуть завантажені метадані: **entity**:**data_table_name_id**, наприклад, entity:quality_control_id.
b. Entity: введіть ідентифікатори зразків, які відповідають ідентифікаторам послідовностей у Terra.
c. Submission_id: введіть унікальний ідентифікатор для зразка. Це ідентифікатор, який буде відображатися на сайті NCBI.
    i. Введіть submission_id також у колонці "штам".
    ii. Введіть "submission_id-001" у стовпчик "library_id".
d. Назва: "PulseNet".
e. Для "library_strategy", "library_source", "library_selection", "platform", "instrument_model" і "filetype" виберіть потрібну опцію зі спадного меню.
f. Планування бібліотеки: "Парна".
g. Організм: рід і вид.
h. Collected_by: лабораторія, що надала послідовність.
i. Дата_збору: **рік для клінічних** ізолятів, **рік і місяць для неклінічних** ізолятів. Обов'язковий формат: ГГГГ:ММ:ДД, наприклад, для клінічного зразка, виділеного в 2024 році: 2024-01-01. Для неклінічного зразка, виділеного в лютому 2024 року: 2024-02-01.
j. Geo_loc_name: **країна-джерело для клінічних** ізолятів, **країна-джерело та штат (або інше більш детальне місцезнаходження) для неклінічних** ізолятів. Використовуйте двокрапку, щоб відокремити країну та більш детальну локацію, наприклад, США:Каліфорнія.
k. Джерело_ізоляції: **"Відсутнє" для клінічних** ізолятів, **точне джерело для неклінічних** ізолятів, наприклад, салат, куряча грудка, мазок тощо.
l. Тип ізоляту: клінічний, з довкілля, харчовий або тваринного походження.
m. Серотип: Серотип *кишкової палички*.
n. Серовар: Серовар *сальмонели*.

Тільки клінічні (людського походження) послідовності

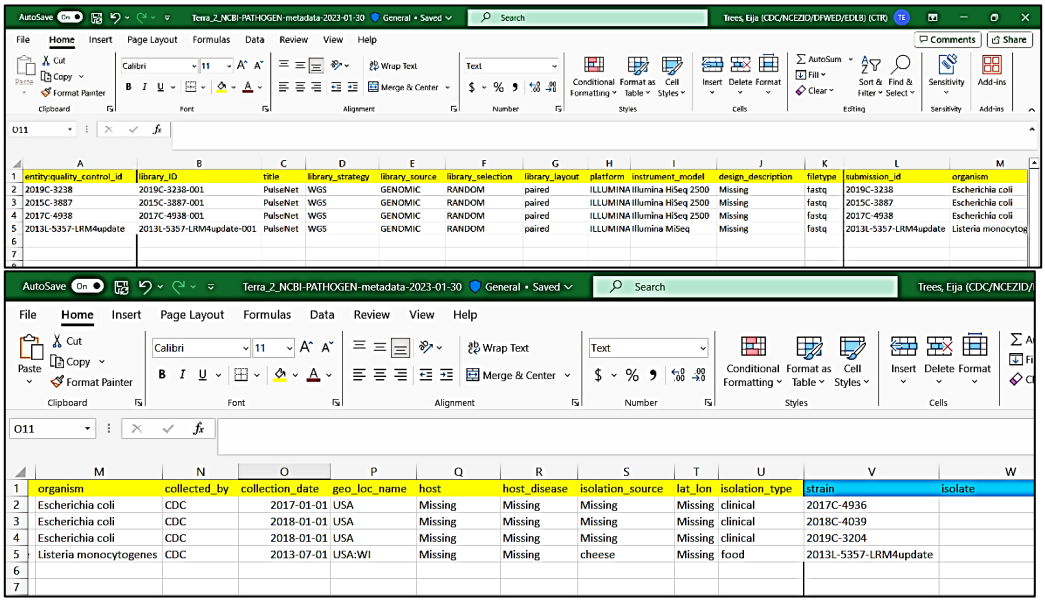


	A	B	C	D	E	F	G	H	I	J	K	L
1	entity:quality_control_id	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	instrument_model	design_description	filetype	submission_id
2	2017C-4936	2017C-4936-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2017C-4936
3	2018C-4039	2018C-4039-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2018C-4039
4	2019C-3204	2019C-3204-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2019C-3204
5												
6												
7												
8												



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

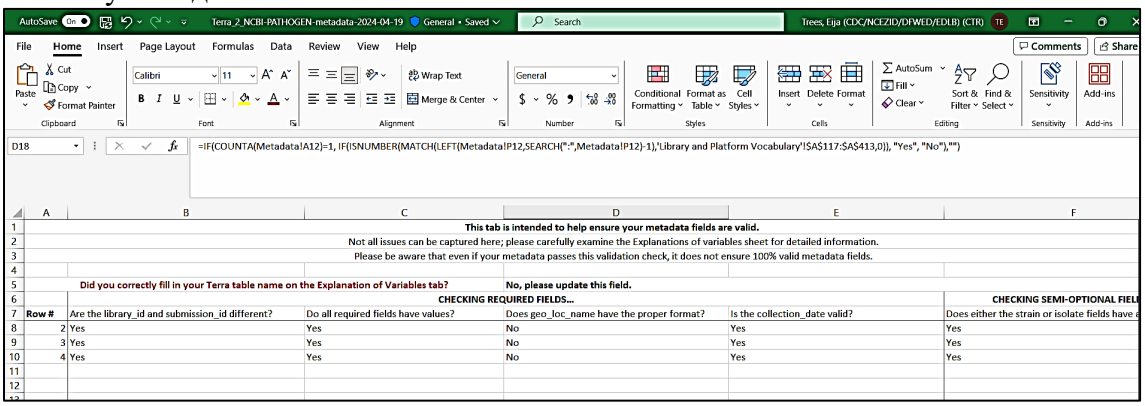
Клінічні (людського походження) та неклінічні послідовності



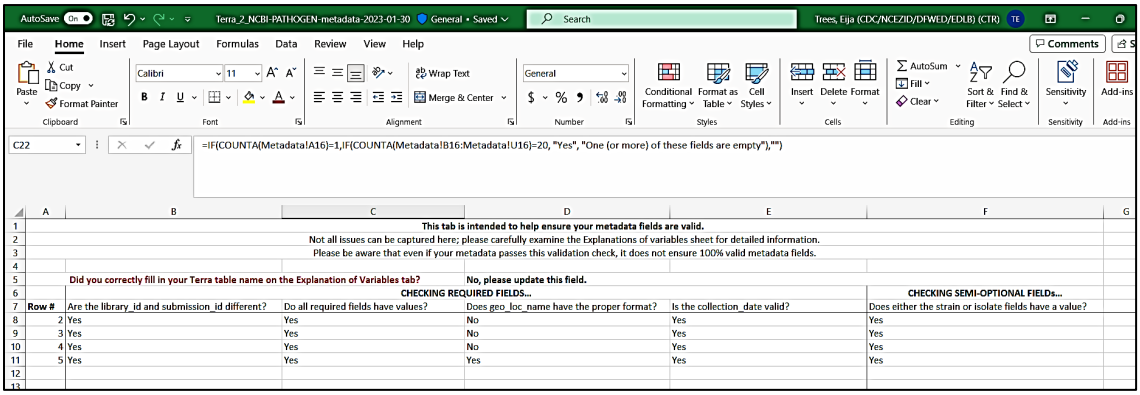
	entity:quality_control_id	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	instrument_model	design_description	filetype	submission_id	organism
1	entity:quality_control_id	library_ID	title	library_strategy	library_source	library_selection	library_layout	platform	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-4938	2017C-4938-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina HiSeq 2500	Missing	fastq	2017C-4938	Escherichia coli
5	2013L-5357-LRM4update	2013L-5357-LRM4update-001	PulseNet	WGS	GENOMIC	RANDOM	paired	ILLUMINA	Illumina MiSeq	Missing	fastq	2013L-5357-LRM4update	Listeria monocytog

	organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	strain	isolate
1	organism	collected_by	collection_date	geo_loc_name	host	host_disease	isolation_source	lat_lon	isolation_type	strain	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. Перейдіть на вкладку "Валідація", щоб підтвердити правильність заповнення шаблону метаданих.

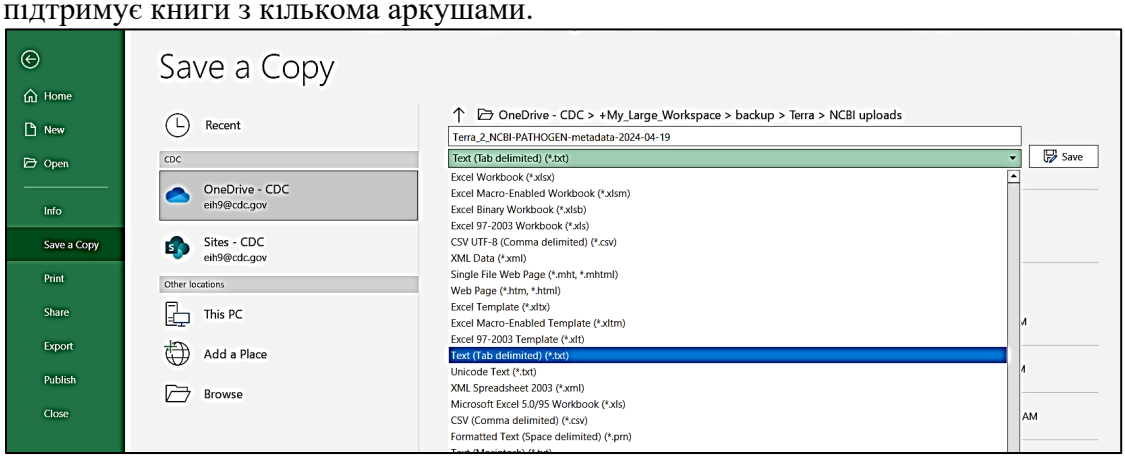


	A	B	C	D	E	F
1		This tab is intended to help ensure your metadata fields are valid.				
2		Not all issues can be captured here; please carefully examine the Explanations of variables sheet for detailed information.				
3		Please be aware that even if your metadata passes this validation check, it does not ensure 100% valid metadata fields.				
4						
5		Did you correctly fill in your Terra table name on the Explanation of Variables tab?		No, please update this field.		
6				CHECKING REQUIRED FIELDS...		CHECKING SEMI-OPTIONAL FIEL
7	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
8	2	Yes	Yes	No	Yes	Yes
9	3	Yes	Yes	No	Yes	Yes
10	4	Yes	Yes	No	Yes	Yes
11						
12						

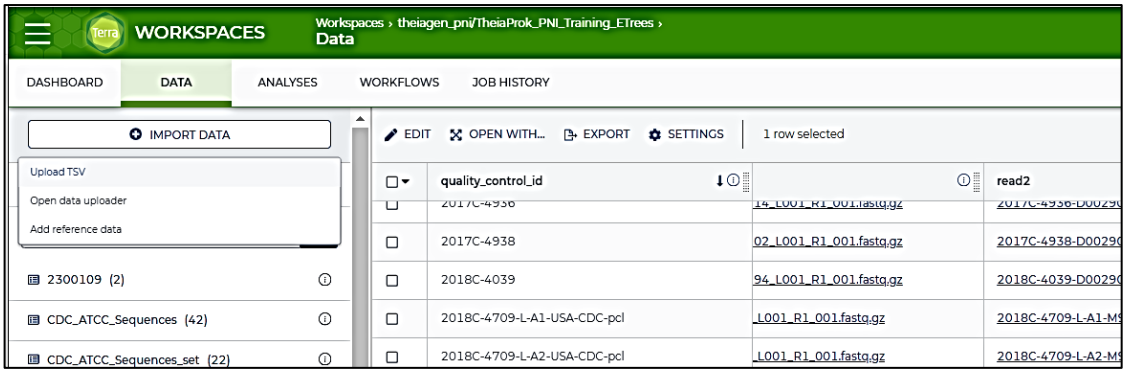


ПРИМІТКА: *валідація належного формату географічного розташування пройде тільки в тому випадку, якщо вказана країна-джерело і більш детальне місцезнаходження. Таким чином, валідація не пройде для клінічних ізолятів, для яких вказана лише країна походження. Однак NCBI прийме країну походження як єдине географічне місцезнаходження.*

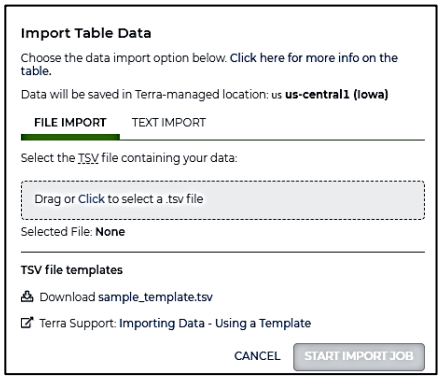
3. Збережіть заповнений шаблон метаданих у форматі tsv (з табуляцією). Натисніть "ОК" у спливаючому вікні, в якому зазначено, що вибраний тип файлу не підтримує книги з кількома аркушами.



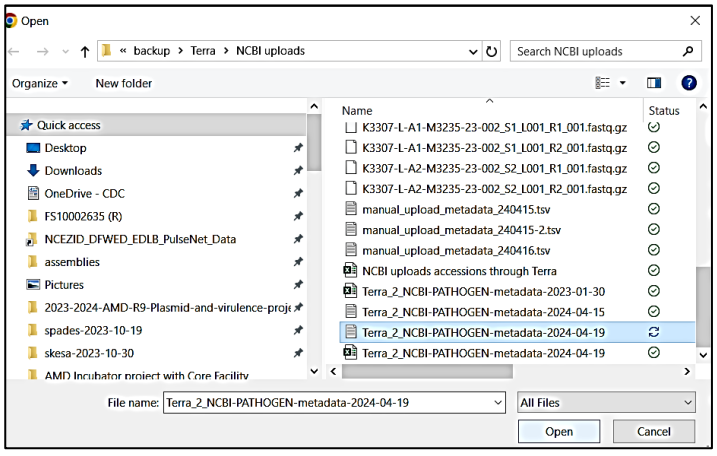
4. У вкладці "Дані" робочого простору Terra натисніть "Імпортувати дані" і виберіть "Завантажити TSV" у випадаючому меню.



5. У спливаючому вікні "Імпорт даних таблиці" на вкладці "Імпорт файлів" клацніть посередині, щоб вибрати файл tsv.

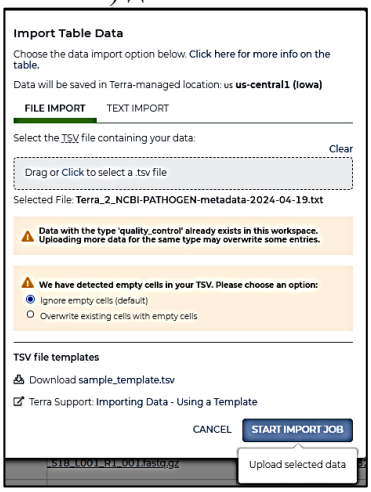


6. Перейдіть до місця, де зберігається файл метаданих tsv, виберіть його і натисніть "Відкрити".

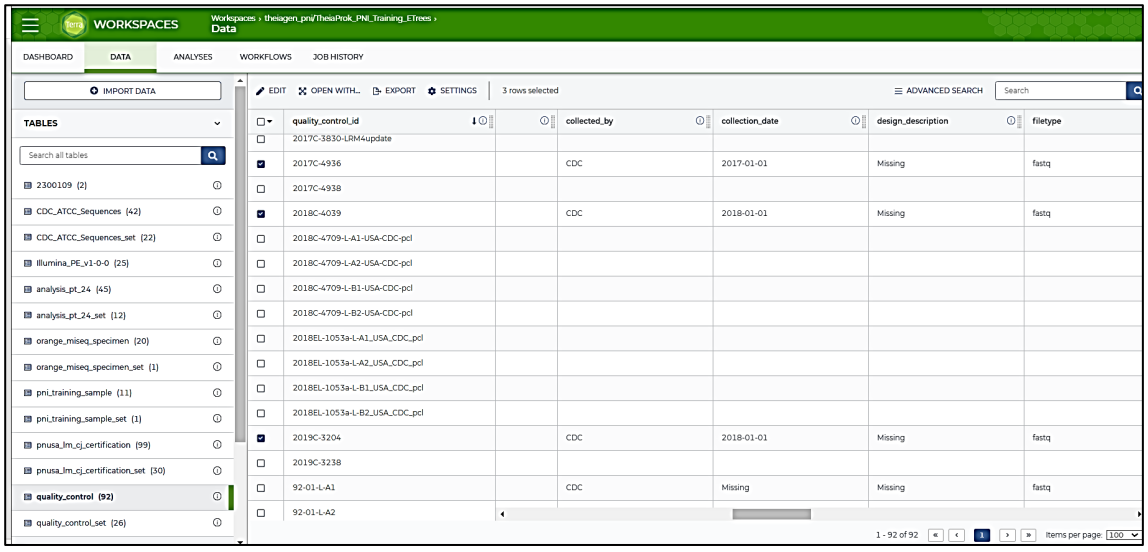


7. У спливаючому вікні "Імпортувати дані таблиці" ви побачите попередження про те, що дані вже існують у відповідній таблиці даних, і завантаження нових даних до неї може призвести до перезапису наявних даних. Також з'явиться попередження,

якщо файл метаданих tsv не містить однакової інформації (деякі дані відсутні для деяких послідовностей) для всіх записів. Натисніть на кнопку "Почати імпорт".



8. Після завершення завантаження ви побачите бажані поля метаданих, заповнені в таблиці даних.

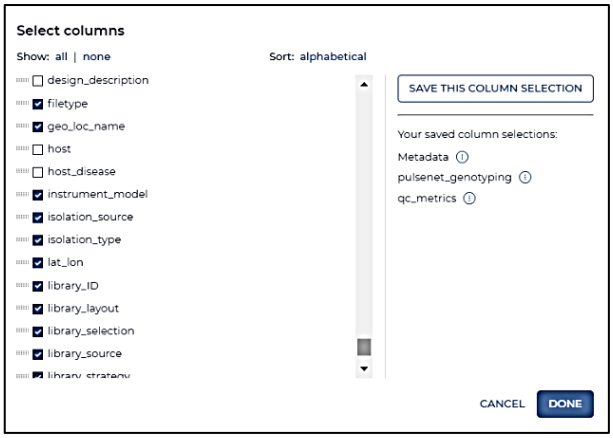


Створіть окреме подання метаданих для таблиці даних

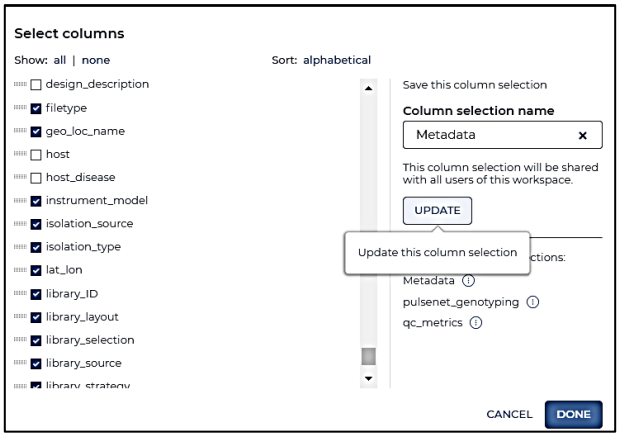
1. На вкладці "Дані" виберіть таблицю даних, яка вас цікавить, а потім виберіть "Налаштування".
2. У розділі "Вибрати стовпці" позначте всі потрібні стовпці метаданих.
    a. Рекомендовані метадані для подачі в NCBI:
        i. collected_by
        ii. collection _date
        iii. filetype
        iv. 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

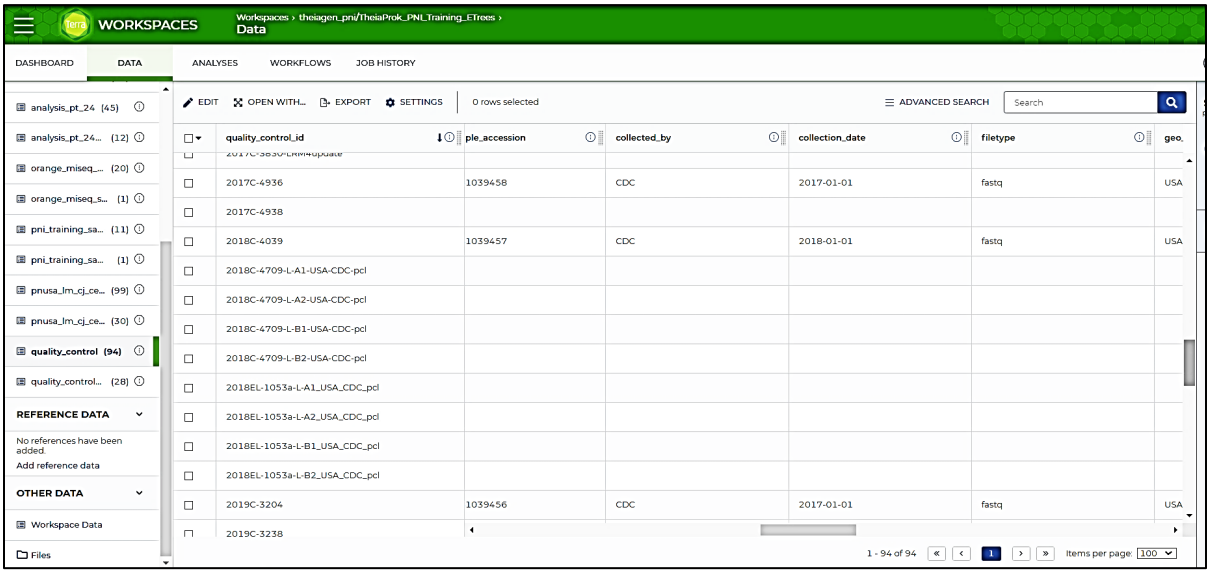
b. Додаткова корисна інформація про послідовність
   i. read1 (ім'я файлу R1 FASTQ)
   ii. read2 (ім'я файлу R2 FASTQ)
   iii. assembly_fasta (місце розташування збірки, згенерованої Terra)
   iv. Якщо ви завантажуєте дані безпосередньо з Illumina BaseSpace:
      1. basespace_collection_id
      2. basespace_fetch_analysis_date
      3. basespace_fetch_version
      4. basespace_sample_id
      5. basespace_sample_name
   v. biosample_accession
   vi. sra_accession



3. Натисніть "Зберегти цей вибір стовпця".
4. Назвіть вибір стовпця "Метадані" і натисніть "Зберегти" та "Готово".
   ПРИМІТКА: *Якщо ви додаєте або видаляєте стовпці з існуючого виділення стовпців, натисніть "Зберегти це виділення стовпців", виберіть назву зі спадного меню і натисніть "Оновити".*



5. Тепер у таблиці даних мають бути видимими лише потрібні стовпці метаданих



Додаток PNID01-7: Огляд робочого процесу TheiaProk для визначення бактеріальних характеристик