illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx на NextSeq 550Dx

Посібник користувача застосунку

Документ № 200025238, вер. 00
Лютий 2023 р.
ВИКОРИСТОВУВАТИ ДЛЯ ДІАГНОСТИКИ IN VITRO.

Історія редакцій

Документ	Дата	Опис зміни
200025238, версія 00	Лютий 2023 р.	Початкова редакція.

Зміст

Огляд

Застосунок DRAGEN for Illumina DNA Prep with Enrichment Dx (DRAGEN for IDPE Dx) використовується для планування та проведення вторинного аналізу бібліотек IDPE Dx, створених для секвенування за допомогою NextSeq 550Dx.

Застосунок DRAGEN for IDPE Dx підтримує повний цикл від секвенування до аналізу в разі використання з набором для підготовки бібліотек Illumina DNA Prep with Enrichment Dx, системою секвенування NextSeq 550Dx та сервером Illumina DRAGEN для NextSeq 550Dx.

Методи аналізу

DRAGEN for IDPE Dx виконує демультиплексування, генерування FASTQ, зіставлення зчитувань, вирівнювання відповідно до еталонного геному та розпізнавання малих варіантів залежно від вибраного робочого процесу:

- Генерування файлів FASTQ
- Генерування файлів FASTQ та VCF зародкової лінії
- Генерування файлів Somatic FASTQ та VCF

ПРИМ. Стиснення ORA доступне для використання з усіма трьома робочими процесами. DRAGEN ORA Compression — це програмне забезпечення для повного стиснення даних без втрат, яке створює файл із розширенням Original Read Archive (*.ora). Формат ora — це еталонний формат стиснення файлів FASTQ, розроблений для дуже швидкого стиснення/декомпресії та високого коефіцієнта стиснення.

Генерування файлів FASTQ

Зібрані послідовності записуються у файли FASTQ, по окремому файлу на кожний зразок. Файли FASTQ — це текстові файли, які містять дані секвенування та показники якості лише для одного зразка. До кожного зразка створюються окремі файли FASTQ для кожної доріжки проточної кювети та для кожного зчитування секвенування. Назва зразка, вказана під час налаштування прогону, входить до назви файлу FASTQ. Файли FASTQ містять первинні вхідні дані для вирівнювання. Перший етап створення FASTQ — це демультиплексування. Демультиплексування призначає кластери, що пройшли фільтрацію, конкретному зразку шляхом порівняння кожної послідовності зчитування індексу з послідовностями індексів, указаними для прогону. На цьому етапі якісні значення не враховуються. Зчитування індексів ідентифікуються під час описаних далі етапів.

- Зразки нумеруються від 1 у порядку, у якому їх перелічено для прогону.
- Номер зразка 0 резервується для кластерів, які не було присвоєно зразку.

- Кластери присвоюються зразку в разі точної відповідності послідовності індексів або коли є одна розбіжність на зчитування індексу.

У програмне забезпечення входить стискання ORA для стиснення файлів FASTQ. Цей формат можна ввімкнути за бажанням. У разі використання формату ORA (*.ora) контрольна сума md5 вмісту FASTQ зберігається після циклу стиснення та декомпресії, що гарантує стиснення без втрат.

Зіставлення та вирівнювання ДНК

Після генерації файлів FASTQ зчитування зіставляються та вирівнюються за еталонним геномом. Першим етапом зіставлення є створення шаблонів на основі зчитування з подальшим пошуком точних збігів в еталонному геномі. Після цього отримані результати уточнюються шляхом повного вирівнювання за алгоритмом Сміта — Уотермана для ділянок із найвищою щільністю збігів шаблонів. Цей добре задокументований алгоритм працює шляхом порівняння кожної позиції зчитування з усіма потенційними позиціями еталону. Такі порівняння відповідають матриці можливих варіантів вирівнювання між зчитуванням та еталоном. Для кожної з цих потенційних позицій зіставлення алгоритм Сміта — Уотермана генерує бали, за допомогою яких можна оцінити, яким чином найкраще вирівнювання, що проходить через цю комірку матриці, досягає її: через збіг або розбіжність нуклеотидів (діагональний рух), делецію (горизонтальний рух) або інсерцію (вертикальний рух). Збіг між зчитуванням і еталоном дає бонус до бала, а невідповідність або індел накладає штраф. Шлях із найбільшим загальним балом у матриці обирається як остаточне вирівнювання. Алгоритм апаратно прискорюється на платах програмованих логічних інтегральних схем (FPGA) сервера DRAGEN. Еталонний геном, що використовується в застосунку, створений на основі файлу FASTA UCSC hg19 із використанням опції сервера DRAGEN для побудови хеш-таблиці з урахуванням альтернативних локусів (alt-aware) на основі механізму liftover.

Розпізнавання варіантів зародкової лінії за допомогою DRAGEN

Програма розпізнавання малих варіантів зародкової лінії DRAGEN приймає зіставлені й вирівняні зчитування ДНК як вхідні дані та розпізнає однонуклеотидні поліморфізми (SNP) та інсерції або делеції (індели) шляхом поєднання постовпчикового виявлення та локального складання гаплотипів *de novo*. Щоб увімкнути програму розпізнавання малих варіантів зародкової лінії DRAGEN, оберіть робочий процес для варіанту зародкової лінії.

Розпізнавання варіантів зародкової лінії зазвичай використовується для зразків зародкової лінії, де плоїдність дорівнює двом. Спочатку визначаються еталонні ділянки з достатнім покриттям вирівнювання, доступні для розпізнавання. У межах цих еталонних ділянок за допомогою швидкого сканування відсортованих зчитувань ідентифікуються активні ділянки, центровані на стовпцях накопичення, що свідчать про наявність варіанта. Активні ділянки доповнюються достатнім контекстом, щоб охопити значущий нееталонний вміст поблизу. За наявності ознак інделів активні ділянки отримують додаткове заповнення.

Вирівняні зчитування обрізаються в межах кожної активної ділянки та збираються у граф де Брюїна. Краї обрізаних зчитувань зважуються за кількістю спостережень, при цьому еталонна послідовність виступає основою. Після очищення та спрощення графа всі шляхи від джерела до стоку витягуються як потенційні гаплотипи. Кожен гаплотип вирівнюється відносно еталонного геному за алгоритмом Сміта — Уотермана для ідентифікації варіантів, які він представляє. Цей набір подій може бути доповнений шляхом виявлення на основі позиції. Для кожної пари «зчитування-гаплотип» ймовірність P(r|H) спостереження зчитування, за умови, що гаплотип є справжнім вихідним зразком, оцінюється за допомогою парної прихованої марковської моделі.

Шляхом сканування за еталонною позицією в межах активної ділянки формуються потенційні генотипи з диплоїдних комбінацій варіантів (SNP або інделів). Для кожної події (включно з еталоном) умовна ймовірність P(r|e) спостереження кожного зчитування з перекриванням оцінюється як максимальна P(r|H) для гаплотипів, що підтримують цю подію. Вони об'єднуються в умовну ймовірність P(r|e1e2) для генотипу (пари подій) і множаться для отримання умовної ймовірності P(R|e1e2) спостереження всього накопичення зчитувань. За формулою Байєса обчислюється апостеріорна ймовірність P(e1e2|R) кожного диплоїдного генотипу, після чого обирається найбільш імовірний варіант.

Аналіз DRAGEN for IDPE Dx застосовує автоматичне фільтрування. Докладнішу інформацію див. у розділі *Анотації файлу VCF для гермінального робочого процесу* на стор. 14.

Розпізнавання соматичних варіантів за допомогою DRAGEN

Програма розпізнавання малих соматичних варіантів DRAGEN приймає зіставлені й вирівняні зчитування ДНК як вхідні дані та розпізнає SNV та індели шляхом локального складання гаплотипів *de novo* в активній ділянці. Щоб увімкнути програму розпізнавання малих соматичних варіантів DRAGEN, оберіть застосунок для соматичних варіантів.

Розпізнавання соматичних варіантів зазвичай використовується для зразків пухлин. За цього робочого процесу DRAGEN не робить жодних припущень щодо плоїдності, що дає змогу розпізнавати низькочастотні алелі. Для локусів із глибиною покриття до 100x у пухлинному зразку DRAGEN має поріг розпізнавання за частоти алелів варіанта 5 %. Поріг розпізнавання масштабується зі збільшенням глибини для кожного локуса окремо та зменшується вдвічі щоразу, коли покриття подвоюється понад 100x. Спочатку визначаються еталонні ділянки з достатнім покриттям вирівнювання, доступні для розпізнавання. У межах цих еталонних ділянок шляхом сканування відсортованих зчитувань ідентифікуються активні ділянки, центровані на стовпцях накопичення, що свідчать про наявність варіанта у зчитуваннях пухлини. Активні ділянки доповнюються достатнім контекстом, щоб охопити значущий нееталонний вміст поблизу. За наявності ознак інделів активні ділянки отримують додаткове заповнення.

Вирівняні зчитування обрізаються в межах кожної активної ділянки та збираються у граф де Брюїна. Краї обрізаних зчитувань зважуються за кількістю спостережень, при цьому еталонна послідовність виступає основою. Після очищення та спрощення графа всі шляхи від джерела до стоку витягуються як потенційні гаплотипи. Кожен гаплотип вирівнюється відносно еталонного

геному за алгоритмом Сміта — Уотермана для ідентифікації варіантів, які він представляє. Для кожної пари «зчитування-гаплотип» ймовірність P(r|H) спостереження зчитування оцінюється за допомогою парної прихованої марковської моделі за припущення, що гаплотип є справжнім вихідним зразком.

Для визначення показника межі виявлення пухлини (TLOD) програма розпізнавання малих соматичних варіантів DRAGEN спочатку сканує активну ділянку за еталонною позицією для кожної потенційної соматичної події, а також еталонної події. Умовна ймовірність P(r|e) спостереження кожного зчитування з перекриванням оцінюється як максимальна P(r|H) для гаплотипів, що підтримують цю подію. Вони об'єднуються в умовну ймовірність P(r|E) для гіпотези події (E), яка включає суміш еталонних та потенційних соматичних алелів у діапазоні можливих частот алелів, і множаться для отримання умовної ймовірності P(R|E) спостереження всього накопичення зчитувань. На основі цього розраховується бал TLOD на доказ того, що алель ALT присутній у зразку пухлини в певному локусі.

DRAGEN for IDPE Dx застосовує автоматичне фільтрування. Докладнішу інформацію див. у розділі *Анотації VCF-файлу робочого процесу для соматичних варіантів* на стор. 16.

Створення запланованого прогону

Виконуйте наведені нижче дії, щоб налаштувати прогін через Illumina Run Manager на NextSeq 550Dx або через браузер на підключеному до мережі комп'ютері. Якщо потрібно імпортувати дані зразків, використовуйте браузер на підключеному до мережі комп'ютері. Див. Посібник з програмного забезпечення Illumina Run Manager for NextSeq 550Dx (документ № 200025239) для отримання інструкцій щодо доступу до Illumina Run Manager з комп'ютера, підключеного до мережі.

Існує два різні способи створення нового запланованого прогону:

- **Import Run** (Імпорт прогону) — використовуйте протокол аналізу з наявного прогону як шаблон для нового прогону. Щоб отримати інформацію про імпорт запуску, див. Посібник з програмного забезпечення Illumina Run Manager for NextSeq 550Dx (документ № 200025239).
- **Create Run** (Створити прогін) — вручну введіть параметри прогону. У наведених нижче інструкціях описано, як створити прогін.

ПРИМ. Обов'язкові поля в інтерфейсі користувача позначені знаком зірочки (*).

Застосунок

1. На екрані Runs (Прогони) у вкладці Planned (Заплановані) виберіть **Create Run** (Створити прогін).
2. Виберіть застосунок DRAGEN for Ilumina DNA Prep with Enrichment Dx, а потім натисніть **Next** (Далі).

Налаштування прогону

1. На екрані Run Settings (Параметри прогону) введіть унікальне ім'я прогону. Ім'я прогону ідентифікує прогін на всіх етапах від секвенування до аналізу.
2. **[Необов'язково]** Введіть опис прогону, який допоможе додатково його ідентифікувати.
3. Виберіть набір (набори) індексних адаптерів для використання під час підготовки бібліотеки.
4. Перегляньте значення довжини зчитування й за потреби змініть його. Зчитування 1 і зчитування 2 мають значення за замовчуванням 151 цикл. Індекс 1 та індекс 2 мають фіксоване значення 10 циклів і не можуть бути змінені.
5. **[Необов'язково]** Введіть ідентифікатор пробірки з бібліотекою.
6. Виберіть **Next** (Далі).

Дані зразка

Дані про зразок включають ідентифікатор зразка, положення лунки (індексні позиції лунок планшета) та назву бібліотеки. У разі використання індексів A і B положення лунки також включає ідентифікатор планшета.

Є два способи введення даних зразка:

- **Через імпорт зразків** — використовуйте файл шаблону, доступний для завантаження на екрані Sample Data (Дані зразка).
- **Вручну** — введіть зразки даних безпосередньо в таблицю на екрані Sample Data (Дані зразка).

Імпорт зразків

Файл шаблону (`*.csv`) можна завантажити на екрані Sample Data (Дані зразка) під час планування прогону через браузер на підключеному до мережі комп'ютері. Файл шаблону недоступний для завантаження під час доступу до Illumina Run Manager через операційну систему NextSeq 550Dx. Щоб ввести зразки даних за допомогою функції Import Samples (Імпорт зразків), виконайте наведені нижче дії.

ПРИМ. Перш ніж продовжити, виконайте дії з налаштування прогону.

1. Щоб завантажити порожній файл CSV, натисніть **Download Template** (Завантажити шаблон).
2. У файлі шаблону введіть дані зразка, а потім збережіть файл. Вводити ім'я бібліотеки необов'язково.

ПРИМ. Під час використання індексів A і B дані для стовпчика B мають включати положення як планшета, так і лунки (індексні позиції лунок планшета). Приклад: A-A01, A-A02, A-A03.

3. Виберіть **Import Samples** (Імпорт зразків) і перейдіть до файлу шаблону, що містить інформацію про дані зразків з попереднього етапу.
4. Виберіть **Open** (Відкрити), **Proceed** (Продовжити), а потім **Next** (Далі).

ПРИМ. Зміна ідентифікатора зразка до натискання кнопки Next (Далі) може призвести до помилки. Щоб уникнути помилок, завершіть налаштування запуску перед внесенням змін.

Введення зразків уручну

Використовуйте таблицю на екрані Sample Data (Дані зразка) для введення даних зразка вручну.

1. Введіть унікальний ідентифікатор зразка в поле Sample ID (Ідентифікатор зразка).

2. Використовуйте **Well Position** (Положення лунки) (індекс A або індекс B) або **Plate - Well Position** (Планшет – положення лунки) (індекс A та B), щоб вибрати відповідний індекс для зразків.
   Поля i7 Index (Індекс i7), Index 1 (Індекс 1), i5 Index (Індекс i5) та Index 2 (Індекс 2) заповнюються автоматично.
3. **[Необов'язково]** Введіть ім'я бібліотеки.
4. Додайте рядки та за потреби повторіть дії 1–3, доки всі зразки не будуть додані до таблиці. Можна додати кілька рядків одночасно: спочатку введіть потрібну кількість рядків, а потім натисніть значок «+». Ви також можете видалити рядки, установивши прапорець поруч із номером рядка, а потім натиснувши значок кошика.
5. Виберіть **Next** (Далі).

Параметри аналізу

1. Виберіть потрібний робочий процес аналізу:
   • генерування файлів FASTQ;
   • генерування файлів FASTQ та VCF робочого процесу для зародкового лінії (вимагається файл маніфесту);
   • генерування файлів FASTQ та VCF робочого процесу для соматичних варіантів (вимагається файл маніфесту).
2. **[Необов'язково]** **За замовчуванням увімкнено функцію генерування стиснених файлів FASTQ ORA**. Стиснення FASTQ ORA без втрат стискає файли FASTQ до 5 разів у порівнянні з fastq.gz. Зніміть прапорець **Generate ORA compressed FASTQs** (Генерувати файли FASTQ зі стисненням ORA), якщо перевага надається даним без стиснення (fastq.gz).
3. Для робочих процесів із зародковою лінією та соматичними варіантами потрібен файл маніфесту. Використовуйте розкривне меню **Manifest File Selection** (Вибір файлу маніфесту), щоб вибрати файл маніфесту. Маніфест — це файл BED (`*.bed`), розділений знаками табуляції, який визначає назви та розташування цільових еталонних ділянок. Для отримання додаткової інформації див. розділ *Файл маніфесту* на стор. 8.
4. **[Необов'язково]** Для робочих процесів із соматичними варіантами скористайтеся розкривним меню **Noise File Selection** (Вибір файлу шуму), щоб вибрати файл системного шуму.
   Для фільтрації системного шуму можна вказати файл BED (`*.bed.gz`) із рівнем шуму для конкретної ділянки. Для отримання додаткової інформації див. розділ *Фільтрування шуму (необов'язково)* на стор. 9.
5. Виберіть **Next** (Далі).

Перегляд параметрів прогону

1. На екрані Review (Огляд) перегляньте інформацію у вкладках Run Settings (Параметри прогону), Sample Data (Дані зразка) та Analysis Settings (Параметри аналізу).

2. Виберіть **Save** (Зберегти).
   Прогін зберігається у вкладці Planned (Заплановані) на екрані Runs (Прогони).

Налаштування

Щоб переглянути або змінити налаштування застосунку DRAGEN for IDPE Dx, спочатку виберіть піктограму застосунків на головному екрані. Потім виберіть застосунок, який потрібно переглянути або змінити. Для зміни налаштувань потрібен обліковий запис адміністратора.

Конфігурація

На екрані конфігурації відображаються такі налаштування застосунку:

- **Library Prep Kits** (Набори для підготовки бібліотек) — відображає набори для підготовки бібліотек, які використовуються застосунком за замовчуванням. Це налаштування не можна змінити.
- **Набори індексних адаптерів** — відображає набір індексних адаптерів за замовчуванням для програми. Це налаштування не можна змінити.
- **Довжина зчитування** — для застосунку за замовчуванням встановлено довжину зчитування 151 п. о., проте її можна змінити під час створення прогону.
- **Файли маніфесту та шуму** — завантаження та зміни налаштувань файлів маніфесту й системного шуму.
  - Виберіть **Upload File** (Завантажити файл), щоб завантажити файли для використання в аналізі.
  - Виберіть перемикач **Default** (За замовчуванням), щоб встановити файл маніфеста або шуму як такий, що буде автоматично обиратися під час створення прогону, коли можна вибрати цей застосунок.
  - Встановіть прапорець **Enabled** (Увімкнено), щоб файл відображався у розкривному меню під час створення прогону.

Дозволи

Використовуйте прапорці на екрані Permissions (Дозволи), щоб керувати доступом користувачів до застосунку.

Файл маніфесту

У разі використання застосунку DRAGEN for IDPE Dx файл маніфесту потрібен для таких робочих процесів:

- генерування файлів FASTQ та VCF робочого процесу для варіантів зародкової лінії;
- генерування файлів FASTQ та VCF робочого процесу для соматичних варіантів.

Маніфест — це файл у форматі BED (*.bed), розділений знаками табуляції, який визначає назви та розташування цільових еталонних ділянок. Основний розділ файлу маніфесту — це розділ Regions (Ділянки), який має містити такі стовпці даних:

Стовпець	Опис
Name (Назва)	Унікальна назва цілі, визначена користувачем.
Chromosome (Хромосома)	Розташування хромосом (наприклад, chr10, chr5 тощо).
Start (Початок)	1-базовий індекс для початкової позиції цільової ділянки.
Stop (Зупинка)	1-базовий індекс для кінцевої позиції цільової ділянки.
Upstream Probe Length (Довжина висхідного зонда)	Довжина зонда, розташованого вище за потоком. Для застосунку DRAGEN for IDPE Dx це значення має бути встановлено рівним 0.
Downstream Probe Length (Довжина низхідного зонда)	Довжина зонда, розташованого нижче за потоком. Для застосунку DRAGEN for IDPE Dx це значення має бути встановлено рівним 0.

ПРИМ. Для аналізу потрібен дійсний формат файлу маніфесту. DRAGEN припинить аналіз, якщо файл маніфесту недійсний.

Фільтрування шуму (необов’язково)

Фільтр системного шуму доступний для розпізнавання соматичних варіантів і може бути використаний для зменшення кількості хибнопозитивних результатів шляхом врахування шуму в конкретних локусах. Файл системного шуму генерується шляхом попереднього збору приблизно 50 нормальних зразків (бажано специфічних для панелі, методу підготовки бібліотек та секвенатора); після цього сума частот алелей нижче 30 % у кожній ділянці з достатнім покриттям ділиться на загальну кількість зразків (частоти алелей понад 30 % вважаються варіантами зародкової лінії, а не шумом). Після генерування значень шуму соматичні варіанти, виявлені в цій ділянці, будуть відфільтровані.

Фільтр можна використовувати в режимі Tumor-Normal (Пухлина — норма), але він особливо корисний для прогонів у режимі Tumor-Only (Лише пухлина), коли відповідний нормальний зразок відсутній. Файл систематичного шуму має бути у форматі BED з розширенням (*.bed.gz) і містити чотири стовпці: Chromosome (Хромосома), Start (Початок), End (Кінець) і рівні шуму для конкретної ділянки по кожному рядку. Фільтрація системного шуму є необов’язковою.

Вихідні результати аналізу

Прогони в зараз виконання відображаються у вкладці Active (Активні). Завершені цикли обробки відображаються у вкладці Completed (Завершені). DRAGEN for IDPE Dx створює папку аналізу з унікальною назвою для кожного аналізу, яка є окремою від папки з даними секвенування. Папка аналізу містить таку інформацію:

- використаний файл маніфесту;
- версію ПЗ;
- ідентифікатори зразків;
- загальну кількість вирівняних зчитувань;
- відсоток вирівняних зчитувань на зразок;
- кількість SNV, визначених для кожного зразка;
- кількість інделів, визначених для кожного зразка;
- статистичні дані покриття.

Вихідні файли аналізу

Місце розташування папки аналізу визначається параметром External Storage for Analysis Results (Зовнішнє сховище для результатів аналізу). Див. Посібник з програмного забезпечення Illumina Run Manager for NextSeq 550Dx (документ № 200025239) для отримання докладнішої інформації про параметр External Storage for Analysis Results (Зовнішнє сховище для результатів аналізу).

На екрані Run Details (Відомості прогону) поле External Location (Зовнішнє розташування) містить шлях до даних секвенування. На екрані Run Details (Відомості прогону) у полі Analysis Output Folder (Папка вихідних даних аналізу) вказано унікальну назву папки аналізу. Точний перелік сформованих файлів залежить від того, який робочий процес аналізу використовується. Застосунок створює наведені далі вихідні файли аналізу.

ПРИМ. Якщо під час доступу до вихідних файлів аналізу виникає помилка обмеження максимальної довжини шляху до файлу, спробуйте перемістити файл у місце з коротшим шляхом або скористайтеся іншим методом відкриття файлу.

Вихідний файл	Опис
Підсумковий звіт про варіанти (`*.pdf`)	Містить короткий опис інформації про файли, версії програмного забезпечення, інформацію про зразки, статистику зчитувань, а також підсумки щодо однонуклеотидних варіантів (SNV), інсерцій, делецій і покриття. Файл звіту про варіанти створюється в робочому процесі тільки для генів зародкової лінії та для соматичних генів.
FASTQ (`*.fastq.gz` або `*.fastq.ora`)	Проміжні файли, що містять розпізнавання азотистих основ із відповідним показником якості. Файли FASTQ містять первинні вхідні дані для етапу вирівнювання. Коли вибрано стиснення ORA, використовується розширення файлу `*.fastq.ora`.
BAM-файли вирівнювання (`*.bam`)	Містять вирівняні зчитування для конкретного зразка.
VCF-файли геному (`*.gvcf.gz`)	Містять генотип для кожної позиції незалежно від того, чи її визначено як варіант, чи як еталон.

Вихідний файл	Опис
Файли VCF (`*.vcf.gz`)	Містять варіанти, визначені в кожній позиції.
Звіт про показники прогону (`*.csv`)	Містить показники якості щодо прогону, зокрема загальний вихід неіндексованих даних та показник Q30.

Файли FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) — це текстовий формат файлів, що містить визначення основ і значення якості для кожного зчитування. Кожен файл містить таку інформацію:

- Ідентифікатор зразка;
- послідовність;
- показники якості Phred у кодованому форматі ASCII + 33.

Ідентифікатор зразка має такий формат:

```
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y
ReadNum:FilterFlag:0:SampleNumber
Приклад:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Файли BAM

Файл BAM (*.bam) — це стиснена бінарна версія файлу SAM (зіставлення вирівнювання послідовностей), яка використовується для подання вирівняних послідовностей розміром до 128 Мб. У файлах BAM використовується формат іменування файлів `SampleName_S#.bam`., де # — це номер зразка, який визначається порядком перерахування зразків для прогону. У багатовузловому режимі для S# завжди встановлюється значення S1, незалежно від черговості зразка.

Файли BAM містять розділ заголовка й розділ вирівнювання.

- **Header** (Заголовок) — містить інформацію про весь файл, зокрема ім'я зразка, довжину зразка та метод вирівнювання. Вирівнювання в розділі вирівнювання пов'язане з конкретною інформацією в розділі заголовка.
- **Alignments** (Вирівнювання) — містить назву зчитування, послідовність зчитування, якість зчитування, інформацію про вирівнювання та користувацькі мітки. Назва зчитування включає в себе хромосому, початкову координату, якість вирівнювання і рядок характеристики відповідності.

Розділ вирівнювання містить таку інформацію для кожного зчитування або пари зчитувань:

- AS: якість вирівнювання парних кінцевих фрагментів;
- RG: групу зчитувань, яка вказує кількість зчитувань для конкретного зразка;
- BC: тег із штрихкодом, який вказує на демультиплексований ідентифікатор зразка, пов'язаний зі зчитуванням;
- SM: якість вирівнювання одиничних кінцевих фрагментів;
- XC: рядок характеристики відповідності;
- XN: мітку з назвою амплікона, яка містить ідентифікатор амплікона, пов'язаний зі зчитуванням.

Файли індексів BAM (*.bam.bai) надають індекс відповідного файлу BAM.

Файли VCF

Файли формату розпізнавання варіантів (*.vcf) містять інформацію про варіанти, виявлені в конкретних позиціях еталонного геному.

Заголовок файлу VCF містить версію формату файлу VCF і версію засобу розпізнавання варіанта, а також список анотацій, що використовуються в іншій частині файлу. У заголовок VCF також включений файл еталонного геному й файл BAM. Останній рядок у заголовку містить заголовки стовпчиків для рядків даних. Кожний із рядків даних файлу VCF містить інформацію про один варіант.

Таблиця 1. Заголовки файлу VCF

Заголовок	Опис
CHROM	Хромосома еталонного геному. Хромосоми відображаються в тому ж порядку, що й довідковий файл FASTA.
POS	Одноосновні позиції варіантів у еталонній хромосомі. Для однонуклеотидних варіантів (SNV) ця позиція відповідає еталонній основі з варіантом. Для інделів ця позиція відповідає еталонній основі, що безпосередньо передує варіанту.
ID	Номер rs (еталонний SNP) для SNP, отриманий з `dbSNP.txt`, якщо застосовно. Якщо в цьому місці існує кілька номерів rs, то список розділяється крапками з комою. Якщо запис dbSNP у цій позиції відсутній, використовується маркер пропущеного значення ('.').
REF	Еталонний генотип. Наприклад, делеція окремого T представлена як еталонний TT і альтернативний T. Однонуклеотидний варіант від A до T представлений як еталонний A й альтернативний T.
ALT	Алелі, які відрізняються від зчитаного еталонного геному. Наприклад, інсерція окремого T представлена як еталонний A й альтернативний AT. Однонуклеотидний варіант від A до T представлений як еталонний A й альтернативний T.

Заголовок	Опис
QUAL	Показник якості Phred, визначений засобом розпізнавання варіантів. Більш високі показники якості варіанта свідчать про те, що він має більш високий ступінь вірогідності, а також про меншу ймовірність помилок. Для показника якості Q оцінна вірогідність помилки становить 10 –(Q/10). Наприклад, набір із розпізнаваннями Q30 має коефіцієнт помилок 0,1 %. Багато засобів розпізнавання варіантів визначають показники якості на основі своїх статистичних моделей, які є високими порівняно зі спостережуваним коефіцієнтом помилок.

Таблиця 2. Анотації файлу VCF для гермінального робочого процесу

Заголовок	Опис
FILTER (ФІЛЬТР)	Якщо всі фільтри пройдено, у стовпці filter (фільтр) записується PASS (ПРОЙДЕНО). Можливі записи у стовпчику FILTER (ФІЛЬТР): • **DRAGENSnpHardQUAL** — застосовується, якщо показник QUAL для варіанта SNP не відповідає пороговому значенню. • **DRAGENIndelHardQUAL** — застосовується, якщо показник QUAL для варіанта інделу не відповідає пороговому значенню. • **LowDepth** — позицію відфільтровано, оскільки глибина покриття не відповідає пороговому значенню. • **LowGQ** — позицію відфільтровано, оскільки якість генотипу не відповідає пороговому значенню. • **PloidyConflict** — результат визначення генотипу, отриманий модулем визначення варіантів, не узгоджується з плоїдністю хромосоми. • **base_quality** — позицію відфільтровано, оскільки медіанна якість основи альтернативних зчитувань у цьому локусі не відповідає пороговому значенню. • **filtered_reads** — позицію відфільтровано, оскільки відфільтровано надто велику частку зчитувань. • **fragment_length** — позицію відфільтровано, оскільки абсолютна різниця між медіанною довжиною фрагмента альтернативних зчитувань і медіанною довжиною фрагмента еталонних зчитувань у цьому локусі перевищує порогове значення. • **low_depth** — позицію відфільтровано, оскільки глибина зчитування надто низька. • **low_frac_info_reads** — позицію відфільтровано, оскільки частка інформативних зчитувань нижча за порогове значення. • **low_normal_depth** — позицію відфільтровано, оскільки глибина зчитування нормального зразка надто низька. • **long_indel** — позицію відфільтровано, оскільки довжина інделу надто велика. • **mapping_quality** — позицію відфільтровано, оскільки медіанна якість вирівнювання альтернативних зчитувань у цьому локусі не відповідає пороговому значенню. • **multiallelic** — позицію відфільтровано, оскільки поріг LOD пухлини проходять більше двох альтернативних алелів. • **non_homref_normal** — позицію відфільтровано, оскільки генотип нормального зразка не є гомозиготним за референтним алелем. • **no_reliable_supporting_read** — позицію відфільтровано, оскільки немає жодного надійного соматичного зчитування, що підтверджує варіант. • **panel_of_normals** — виявлено принаймні в одному зразку у VCF-файлі панелі нормальних зразків. • **read_position** — позицію відфільтровано, оскільки медіана відстаней між початком/кінцем зчитування та цим локусом нижча за порогове значення. • **RMxNRepeatRegion** — позицію відфільтровано, оскільки весь варіантний алель або його частина є повтором еталонної послідовності. • **strand_artifact** — позицію відфільтровано через виражений зсув за ланцюгами. • **str_contraction** — позицію відфільтровано через підозру на помилку ПЛР, за якої альтернативний алель на одну одиницю повтору коротший за еталонний. • **too_few_supporting_reads** — позицію відфільтровано, оскільки в пухлинному зразку надто мало зчитувань, що підтверджують варіант. • **weak_evidence** — оцінка соматичного варіанта не відповідає пороговому значенню.

Заголовок	Опис
INFO (ІНФОРМАЦІЯ)	Можливі записи у стовпчику INFO (ІНФОРМАЦІЯ): • **AC** — підрахунок алелів у генотипах для кожного алеля ALT, у тому ж порядку, що й у списку. • **AF** — частота алелів для кожного алеля ALT в тому ж порядку, що й у списку. • **AN** — загальна кількість алелів у розпізнаних генотипах. • **DB** — присутність у базі даних DbSNP. • **FS** — p-значення за шкалою Phred, отримане за допомогою точного тесту Фішера для виявлення зміщення ланцюга. • **QD** — вірогідність варіанта / якість за глибиною. • **R2_5P_bias** — показник, що базується на зміщенні спарених зчитувань та відстані від 5'-кінця. • **SOR** — симетричне відношення шансів до таблиці спряженості 2x2 для виявлення зміщення ланцюга. • **DP** — приблизна глибина зчитування (інформативні та неінформативні зчитування); деякі зчитування могли бути відфільтровані на підставі mapq тощо. • **END** — кінцева позиція інтервалу. • **FractionInformativeReads** — частка інформативних зчитувань від загальної кількості зчитувань. • **MQ** — середньоквадратична (RMS) якість вирівнювання. • **MQRankSum** — Z-оцінка за критерієм рангових сум Вілкоксона для якості вирівнювання Alt порівняно з Ref. • **ReadPosRankSum** — Z-оцінка за критерієм рангових сум Вілкоксона для зсуву позиції зчитування Alt порівняно з Ref. • **SOMATIC** — принаймні один варіант у цій позиції є соматичним.
FORMAT (ФОРМАТ)	У стовпчику формату перераховано поля, розділені двокрапкою. Наприклад, GT:GQ. Доступні поля: • **AD** — алельні глибини (з урахуванням лише інформативних зчитувань із загальної кількості зчитувань) для еталонного та альтернативних алелів у наведеному порядку. • **AF** — частки альтернативних алелів у наведеному порядку. • **DP** — приблизна глибина зчитування (зчитування з MQ=255 або з некоректними парами відфільтровано). • **F1R2** — кількість зчитувань в орієнтації пари F1R2, що підтримують кожний алель. • **F2R1** — кількість зчитувань в орієнтації пари F2R1, що підтримують кожний алель. • **GT** — генотип. 0 відповідає еталонній основі, 1 відповідає першому запису в стовпчику ALT і так далі. Скісна риска (/) указує на відсутність інформації про фазування. • **MB** — покомпонентна статистика на рівні зразка для виявлення зсуву за mate-парами. • **PS** — інформація про ідентифікатор фізичного фазування, де кожний унікальний ідентифікатор у межах певного зразка (але не між різними зразками) пов'язує записи в межах однієї групи фазування. • **SB** — покомпонентна статистика на рівні зразка, яка включає точний критерій Фішера для виявлення зсуву за ланцюгами. • **SQ** — соматична якість.
SAMPLE (ЗРАЗОК)	Стовпчик зразка містить значення, наведені у стовпчику FORMAT (ФОРМАТ).

Таблиця 3. Анотації VCF-файлу робочого процесу для соматичних варіантів

Заголовок	Опис
FILTER (ФІЛЬТР)	Якщо всі фільтри пройдено, у стовпці filter (фільтр) записується PASS (ПРОЙДЕНО). Можливі записи у стовпчику FILTER (ФІЛЬТР): • **base_quality** — позицію відфільтровано, оскільки медіанна якість основи альтернативних зчитувань у цьому локусі не відповідає пороговому значенню. • **filtered_reads** — позицію відфільтровано, оскільки відфільтровано надто велику частку зчитувань. • **fragment_length** — позицію відфільтровано, оскільки абсолютна різниця між медіанною довжиною фрагмента альтернативних зчитувань і медіанною довжиною фрагмента еталонних зчитувань у цьому локусі перевищує порогове значення. • **low_depth** — позицію відфільтровано, оскільки глибина зчитування надто низька. • **low_frac_info_reads** — позицію відфільтровано, оскільки частка інформативних зчитувань нижча за порогове значення. • **low_normal_depth** — позицію відфільтровано, оскільки глибина зчитування нормального зразка надто низька. • **long_indel** — позицію відфільтровано, оскільки довжина інделу надто велика. • **mapping_quality** — позицію відфільтровано, оскільки медіанна якість вирівнювання альтернативних зчитувань у цьому локусі не відповідає пороговому значенню. • **multiallelic** — позицію відфільтровано, оскільки поріг LOD пухлини проходять більше двох альтернативних алелів. • **non_homref_normal** — позицію відфільтровано, оскільки генотип нормального зразка не є гомозиготним за референтним алелем. • **no_reliable_supporting_read** — позицію відфільтровано, оскільки немає жодного надійного соматичного зчитування, що підтверджує варіант. • **panel_of_normals** — виявлено принаймні в одному зразку у VCF-файлі панелі нормальних зразків. • **read_position** — позицію відфільтровано, оскільки медіана відстаней між початком/кінцем зчитування та цим локусом нижча за порогове значення. • **RMxNRepeatRegion** — позицію відфільтровано, оскільки весь варіантний алель або його частина є повтором еталонної послідовності. • **strand_artifact** — позицію відфільтровано через виражений зсув за ланцюгами. • **str_contraction** — позицію відфільтровано через підозру на помилку ПЛР, за якої альтернативний алель на одну одиницю повтору коротший за еталонний. • **too_few_supporting_reads** — позицію відфільтровано, оскільки в пухлинному зразку надто мало зчитувань, що підтверджують варіант. • **weak_evidence** — оцінка соматичного варіанта не відповідає пороговому значенню. • **systematic_noise** — позицію відфільтровано на підставі ознак систематичного шуму в нормальних зразках.

Заголовок	Опис
INFO (ІНФОРМАЦІЯ)	Можливі записи у стовпчику INFO (ІНФОРМАЦІЯ): • **DP** — приблизна глибина зчитування (інформативні та неінформативні зчитування); деякі зчитування могли бути відфільтровані на підставі mapq тощо. • **END** — кінцева позиція інтервалу. • **FractionInformativeReads** — частка інформативних зчитувань від загальної кількості зчитувань. • **MQ** — середньоквадратична (RMS) якість вирівнювання. • **MQRankSum** — Z-оцінка за критерієм рангових сум Вілкоксона для якості вирівнювання Alt порівняно з Ref. • **ReadPosRankSum** — Z-оцінка за критерієм рангових сум Вілкоксона для зсуву позиції зчитування Alt порівняно з Ref. • **AQ** — оцінка систематичного шуму. • **hotspot** — відома соматична позиція, яку використовують для підвищення впевненості у визначенні. • **SOMATIC** — принаймні один варіант у цій позиції є соматичним.
FORMAT (ФОРМАТ)	У стовпчику формату перераховано поля, розділені двокрапкою. Наприклад, GT:GQ. Доступні поля: • **AD** — алельні глибини (з урахуванням лише інформативних зчитувань із загальної кількості зчитувань) для еталонного та альтернативних алелів у наведеному порядку. • **AF** — частки альтернативних алелів у наведеному порядку. • **DP** — приблизна глибина зчитування (зчитування з MQ=255 або з некоректними парами відфільтровано). • **F1R2** — кількість зчитувань в орієнтації пари F1R2, що підтримують кожний алель. • **F2R1** — кількість зчитувань в орієнтації пари F2R1, що підтримують кожний алель. • **GP** — апостеріорні ймовірності генотипів, масштабовані за шкалою Фреда, як визначено у нормативах VCF. • **GQ** — якість генотипу. • **GT** — генотип. 0 відповідає еталонній основі, 1 відповідає першому запису в стовпчику ALT і так далі. Скісна риска (/) указує на відсутність інформації про фазування. • **MB** — покомпонентна статистика на рівні зразка для виявлення зсуву за mate-парами. • **PL** — нормалізовані ймовірності генотипів, масштабовані за шкалою Фреда, як визначено у нормативах VCF. • **PRI** — апріорні ймовірності генотипів, масштабовані за шкалою Фреда. • **PS** — інформація про ідентифікатор фізичного фазування, де кожний унікальний ідентифікатор у межах певного зразка (але не між різними зразками) пов'язує записи в межах однієї групи фазування. • **SB** — покомпонентна статистика на рівні зразка, яка включає точний критерій Фішера для виявлення зсуву за ланцюгами. • **SQ** — соматична якість.
SAMPLE (ЗРАЗОК)	Стовпчик зразка містить значення, наведені у стовпчику FORMAT (ФОРМАТ).

VCF-файли геному

VCF-файли геному (*.gvcf.gz) відповідають набору правил для представлення всіх ділянок геному в досить компактному форматі. Файли gVCF включають усі ділянки в досліджуваній області в один файл для кожного зразка. Файл gVCF відображає невизначення в позиціях, які не проходять усі фільтри. Тег генотипу (GT) ./. означає невизначення.

Повернення аналізу в чергу

Можливо, вам доведеться повернути аналіз у чергу, якщо його буде зупинено, під час його виконання станеться помилка або ви захочете повторно проаналізувати прогін з іншими налаштуваннями. Щоб повторно поставити аналіз у чергу, виконайте такі дії:

1. На екрані Run (Прогони) виберіть вкладку Completed (Завершені), а потім виберіть назву прогону для повторного аналізу.
   Якщо раніше вже виконували операцію Requeue Analysis (Повернути аналіз у чергу), виберіть назву прогону з Parent Run (Батьківський прогін).
2. На екрані Run Details (Відомості прогону), після Sequencing Information (Інформація про секвенування), виберіть **Requeue Analysis** (Повернути аналіз у чергу).
3. Виберіть варіант:
   - повернути аналіз у чергу без змін;
   - редагувати налаштування прогону та повернути аналіз у чергу;
   - повернути аналіз у чергу з іншим застосунком.
4. Переконайтеся, що місце розташування, де зараз містяться дані секвенування, вказано в полі **Sequencing data file path** (Шлях до файлу даних секвенування).

ПРИМ. Шлях до даних секвенування має збігатися зі шляхом, вказаним у налаштуваннях External Storage for Analysis Results (Зовнішнє сховище для результатів аналізу). Щоб отримати інформацію про зміну шляху до зовнішнього сховища, див. Посібник з програмного забезпечення Illumina Run Manager for NextSeq 550Dx (документ № 200025239).

5. Введіть причину повторного аналізу.
6. Натисніть **Requeue Analysis** (Повернути аналіз у чергу).
7. Внесіть бажані зміни в Run Settings (Параметри прогону), Sample Data (Дані зразків) та Analysis Settings (Параметри аналізу).
8. Виберіть **Save** (Зберегти). В аналізі буде використано поточні параметри аналізу.

Технічна допомога

Для отримання технічної допомоги зв’яжіться зі службою технічної підтримки компанії Illumina.

Вебсайт:	www.illumina.com
Електронна пошта:	techsupport@illumina.com

Паспорти безпеки продукції (SDS) доступні на вебсайті Illumina за адресою support.illumina.com/sds.html.

Документація на продукт доступна для завантаження на вебсайті support.illumina.com.

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A. (США)
+1 800 809.ILMN (4566)
+1 858 202 4566 (за межами Північної Америки)
techsupport@illumina.com
www.illumina.com

IVD

EC REP

Illumina Netherlands B.V.
Steenoven 19
5626 DK Eindhoven
The Netherlands

Австралійський спонсор
Illumina Australia Pty Ltd
Nursing Association Building
Level 3, 535 Elizabeth Street
Melbourne, VIC 3000
Австралія

ВИКОРИСТОВУВАТИ ДЛЯ ДІАГНОСТИКИ IN VITRO.

illumina®