

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

Ръководство за потребителя на приложението

Настоящият документ и съдържанието му са собственост на Illumina, Inc. и нейните филиали („Illumina“) и са предназначени само за употреба по силата на договор от страна на клиента и във връзка с използването на продукта(ите), описан(и) в настоящия документ, и с никаква друга цел. Този документ и съдържанието му не трябва да се използват или разпространяват за никаква друга цел и/или по друг начин да бъдат съобщавани, разкривани или възпроизвеждани по какъвто и да е начин без предварителното писмено съгласие на Illumina. Illumina не предоставя посредством този документ никакъв лиценз за свой патент, търговска марка, авторско право или права по силата на общото право, нито подобни права на която и да е трета страна.

Инструкциите в този документ трябва да се следват строго и изрично от страна на квалифициран и правилно обучен персонал, за да се гарантират правилната и безопасната употреба на продукта(ите), описан(и) в настоящия документ. Цялото съдържание на този документ трябва да бъде прочетено и разбрано напълно, преди да се използва(т) такъв(такива) продукт(и).

АКО ВСИЧКИ ИНСТРУКЦИИ, СЪДЪРЖАЩИ СЕ В НАСТОЯЩИЯ ДОКУМЕНТ, НЕ БЪДАТ НАПЪЛНО ПРОЧЕТИ И ИЗРИЧНО СПАЗВАНИ, ТОВА МОЖЕ ДА ДОВЕДЕ ДО ПОВРЕДА НА ПРОДУКТА(ИТЕ), НАРАНЯВАНЕ НА ЛИЦА, ВКЛЮЧИТЕЛНО НА ПОТРЕБИТЕЛИ ИЛИ ДРУГИ ЛИЦА, И УВРЕЖДАНЕ НА ДРУГО ИМУЩЕСТВО, И ЩЕ ОТМЕНИ ВСЯКАКВА ГАРАНЦИЯ, ПРИЛОЖИМА ЗА ПРОДУКТА(ИТЕ).

ILLUMINA НЕ ПОЕМА НИКАКВА ОТГОВОРНОСТ В РЕЗУЛТАТ НА НЕПРАВИЛНАТА УПОТРЕБА НА ПРОДУКТА(ИТЕ), ОПИСАН(И) В НАСТОЯЩИЯ ДОКУМЕНТ (ВКЛЮЧИТЕЛНО ТЕХНИ ЧАСТИ ИЛИ СОФТУЕР).

© 2023 Illumina, Inc. Всички права запазени.

Всички търговски марки са собственост на Illumina, Inc. или съответните им притежатели. За конкретна информация относно търговските марки вижте www.illumina.com/company/legal.html.

Хронология на редакциите

Документ	Дата	Описание на промяната
200025238 v00	Февруари 2023 г.	Първоначално издание.

Съдържание

Хронология на редакциите	iii
Общ преглед	1
Методи за анализ	1
Създаване на планирано изпълняване	5
Настройки	8
Манифестен файл	8
Филтриране на шума (незадължително)	9
Изходни данни от анализа	9
Файлове FASTQ	11
BAM файлове	11
VCF файлове	12
Повторно нареждане на анализ на опашка	18
Техническа помощ	20

Общ преглед

DRAGEN for Illumina DNA Prep with Enrichment Dx Приложението (DRAGEN for IDPE Dx) се използва за планиране и извършване на вторичен анализ на библиотеките IDPE Dx, генерирани за секвениране на NextSeq 550Dx.

DRAGEN for IDPE Dx поддържа секвениране за анализ, когато се използва с Illumina DNA Prep with Enrichment Dx Library Prep (Подготовка на библиотеката), NextSeq 550Dx и Illumina DRAGEN Server for NextSeq 550Dx.

Методи за анализ

DRAGEN for IDPE Dx извършва размножаване, генериране на FASTQ, картографиране на разчитане, подравняване с референтен геном и извикване на малък вариант в зависимост от избраните работни потоци:

- Генериране на FASTQ
- Генериране на Germline FASTQ и VCF
- Соматично генериране на FASTQ и VCF

ЗАБЕЛЕЖКА ORA компресията е налична за използване и с трите работни процеса. DRAGEN ORA Compression е софтуер за компресия без загуби, който създава файл с разширение Original Read Archive (*.ora). Форматът ora е базиран на справка формат за компресия за файлове FASTQ и е предназначен за много бързо компресиране/декомпресиране и високо съотношение на компресия.

Генериране на FASTQ

Сглобените последователности се записват във FASTQ файлове на проба. FASTQ файловете са текстови файлове, които съдържат данни за секвениране и резултати за качество само за една проба. За всяка проба се генерират отделни FASTQ файлове за всяка лента на проточна клетка, за отчитане на секвениране. Името на пробата, както е посочено по време на настройката на цикъла, е включено в името на файла FASTQ. FASTQ файловете са основната входна информация за подравняване. Първата стъпка от генерирането на FASTQ е демултиплексиране.

Демултиплексирането присвоява клъстери, които преминават филтъра към проба чрез сравняване на всяка последователност на индексно отчитане с индексните последователности, определени за изпълняването. В тази стъпка не се вземат предвид никакви стойности на качеството. Индексните разчитания се идентифицират чрез следните стъпки:

- Пробите се номерират, започвайки от 1, в зависимост от последователността, в която са изброени за изпълняването.

- Номер на проба 0 е запазен за клъстери, които не са били зададени в проба.
- Клъстерите се задават към проба, когато индексната последователност съвпада точно или когато има единично несъвпадение на индексно разчитане.

Софтуерът включва ORA компресия за компресиране на FASTQ файлове. Този формат може да бъде активиран по избор. При използване на ORA (*.ora) формат, md5 контролна сума на съдържанието на FASTQ се запазва след цикъл на компресия и декомпресия, за да се осигури компресия без загуби.

ДНК картографиране & подравняване

След генериране на FASTQ, разчитанията се картографират и подравняват с референтен геном. Първият етап на картографиране е да се генерират сийдове от разчитането и след това да се търсят точни съвпадения в референтния геном. След това, тези резултати се усъвършенстват чрез изпълнение на пълни подравнявания Smith-Waterman на местата с най-висока плътност на съвпаденията със сийдове. Този добре документиран алгоритъм работи чрез сравняване на всяка позиция на разчитането с всички кандидат позиции на референцията. Тези сравнения съответстват на матрица от потенциални подравнявания между разчитане и справка. За всяка от тези позиции за подравняване на кандидатите, Smith-Waterman генерира резултати, които се използват за оценка дали най-доброто подравняване, преминаващо през тази матрична клетка, достига до нея чрез нуклеотидно съвпадение или несъответствие (диагонално движение), делеция (хоризонтално движение) или инсерция (вертикално движение). Съвпадението между разчитане и справка предоставя бонус за резултата, а несъответствие или индел налагат санкция. Общата най-висока пътека за оценяване през матрицата е избраното подравняване. Алгоритъмът е хардуер, ускорен на картите DRAGEN за полево програмируем гейт масив (FPGA). Референтният геном, използван в приложението, е създаден от UCSC hg19 FASTA с опцията DRAGEN за създаване на алтернативна хеш таблица, базирана на повдигане.

DRAGEN Germline Variant Calling

DRAGEN Germline Small Variant Caller взема картографирани и подравнени ДНК разчитания като входни данни и обозначава единични нуклеотидни полиморфизми (SNPs) и инсерция или делеция (indels) чрез комбинация от колонно разпознаване и локално *de novo* сглобяване на хаплотипове. За да активирате DRAGEN Germline Small Variant Caller, изберете работния поток на варианта на герминационния работен процес.

Обозначаването на варианти за герминативна линия обикновено се използва за проби от герминативна линия, където е известно, че пloidията е две. Референтни региони, които подлежат на обозначаване, първо се идентифицират с достатъчно покритие на подравняването. В рамките на тези референтни региони, бързото сканиране на сортираните разчитания идентифицира активните области, които са центрирани върху колони с данни за вариант. Активните региони са подплатени с достатъчно контекст, за да покрият значително, нереперентно съдържание наблизо. Ако има данни за платформи, активните региони получават допълнителна подложка.

Подравнените разчитания се подрязват във всеки активен регион и се сглобяват в графика на De Bruijn. Краищата на отрязаните разчитания се претеглят от броя наблюдения, като референтната последователност е основата. След известно почистване и опростяване на графиката, всички пътища от

източник до потребител се извличат като кандидати за хаплотипове. Всеки хаплотип е Smith-Waterman, подравнен с референтния геном, за да идентифицира вариантите, които представлява. Този набор от събития може да бъде увеличен чрез откриване въз основа на позицията. За всяка разчетена хаплотипна двойка вероятността $P(r|H)$ да наблюдава разчитането, приемайки, че хаплотипът е истинската начална проба, се оценява с помощта на скрит модел Markov (HMM).

Сканиране по референтна позиция в активния регион, генотиповете на кандидатите се формират от диплоидни комбинации от вариантни събития (SNPs или indels). За всяко събитие (включително референция), условната вероятност $P(r|e)$ да се наблюдава всяко припокриващо се разчитане се оценява като максимума $P(r|H)$ за хаплотипове, подкрепящи събитието. Те се комбинират в условната вероятност $P(r|e1e2)$ за генотип (събитие двойка) и се умножават, за да се получи условната вероятност $P(R|e1e2)$ за наблюдение на цялата разчетена купчина. С помощта на формулата на Bayes се изчислява постериорната вероятност $P(e1e2|R)$ на всеки диплоиден генотип и се обозначава победителят.

DRAGEN for IDPE Dx прилага автоматично филтриране. Вижте [Анотации на VCF файл на герминативната линия на стр. 13](#) за повече информация.

DRAGEN Somatic Variant Calling

DRAGEN Somatic Small Variant Caller взема картографирани и подравнени ДНК разчитания като входни данни и извиква SNV и indels чрез локално *de novo* сглобяване на хаплотипове в активен регион. За да активирате DRAGEN Somatic Small Variant Caller, изберете приложение за соматичен вариант.

Обозначаване на соматичен вариант обикновено се използва за туморни проби. С този работен процес DRAGEN не прави никакви пloidни предположения, което позволява откриване на нискочестотни алели. За локуси с покритие до 100x в туморната проба, DRAGEN има праг на откриване при честоти на вариантни алели от 5%. Граничните скали с нарастваща дълбочина на база на местоположение намаляват наполовина всеки път, когато покритието се удвои над 100x. Референтни региони, които подлежат на обозначаване, първо се идентифицират с достатъчно покритие на подравняването. В рамките на тези референтни региони, сканиране на сортираните разчитания идентифицира активните области, които са центрирани върху колони с данни за вариант в туморните разчитания. Активните региони са подплатени с достатъчно контекст, за да покрият значително, нереперентно съдържание наблизо. Ако има данни за платформи, активните региони получават допълнителна подложка.

Подравнените разчитания се подрязват във всеки активен регион и се сглобяват в графика на De Bruijn. Краищата на отрязаните разчитания се претеглят от броя наблюдения, като референтната последователност е основата. След известно почистване и опростяване на графиката, всички пътища от източник до потребител се извличат като кандидати за хаплотипове. Всеки хаплотип е Smith-Waterman, подравнен с референтния геном, за да идентифицира вариантите, които представлява. За всяка разчетена хаплотипна двойка вероятността $P(r|H)$ да наблюдава разчитането се оценява с помощта на скрит модел Markov (HMM), приемайки, че хаплотипът е истинската начална проба.

За да се определи оценката за граница на откриване на тумора (TLOD), DRAGEN Somatic Small Variant Caller първо сканира по референтна позиция за всяко кандидат-соматично събитие, както и референтното събитие в активния регион. Условната вероятност $P(r|e)$ за наблюдение на всяко припокриващо се разчитане се оценява като максимума $P(r|H)$ за хаплотипове, поддържащи събитието. Те се комбинират в условната вероятност $P(r|E)$ за хипотеза за събитие, E, включваща смесица от

референтния и кандидат-соматичния алел в диапазон от възможни честоти на алели и се умножават, за да се получи условната вероятност $P(R|E)$ за наблюдение на цялата разчетена купчина. Оттам се изчислява TLOD резултат като доказателство, че алелът ALT присъства в туморната проба при даден локус.

DRAGEN for IDPE Dx прилага автоматично филтриране. Вижте [Анотации на VCF файл на соматичната линия на стр. 16](#) за повече информация.

Създаване на планирано изпълняване

Използвайте следните стъпки, за да настроите цикъл в Illumina Run Manager или в NextSeq 550Dx с браузър на мрежов компютър. Използвайте браузър на мрежов компютър, ако желаете да импортирате примерни данни. Вижте Illumina Run Manager за софтуерно ръководство NextSeq 550Dx (документ № 200025239) за инструкции относно достъпа до Illumina Run Manager от мрежов компютър.

Има два различни начина за създаване на ново планирано изпълняване:

- **Import Run** (Импортиране на изпълняване) — Използвайте примерен лист от съществуващо изпълнение като шаблон за ново изпълняване. Вижте Illumina Run Manager за софтуерно ръководство NextSeq 550Dx (документ № 200025239) за информация как да импортирате изпълняване.
- **Create Run** (Създаване на изпълняване) — Ръчно въвеждане на параметри на изпълняване. Следните инструкции описват как да създадете изпълняване.

ЗАБЕЛЕЖКА Необходимите полета за въвеждане в потребителския интерфейс са маркирани със символ звездичка (*).

Приложение

1. От раздел Planned (Планирани) на екран Runs (Изпълнявания) изберете **Create Run** (Създаване на изпълняване).
2. Изберете приложението DRAGEN for Illumina DNA Prep with Enrichment Dx и след това изберете **Next** (Напред).

Настройки на изпълняване

1. На екрана Run Settings (Настройки на изпълняване) въведете уникално име на изпълняване. Името на изпълнение идентифицира изпълняването от секвенирането до анализа.
2. **[Незадължително]** Въведете описание на изпълняването, за по-нататъшното идентифициране на изпълняването.
3. Изберете комплекта (ите) адаптери за индексирание, използвани по време на подготовката на библиотеката.
4. Прегледайте дължината на разчитане и я променете, ако е необходимо. Read 1 (Разчитане 1) и Read 2 (Разчитане 2) имат стойност по подразбиране от 151 цикъла. Index 1 (Индекс 1) и Index 2 (Индекс 2) имат фиксирана стойност от 10 цикъла и не могат да бъдат променени.
5. **[Незадължително]** Въведете идентификатор на библиотечна епруветка.
6. Изберете **Next** (Напред).

Данни за пробите

Данните за пробите включват ИД на пробата, позиция на ямката (позиция на ямката на индексната плака) и име на библиотеката. Когато се използва Index A&B (Индекс АиБ), позицията на ямките включва и идентификатор на плаки.

Има два начина за въвеждане на примерни данни:

- **Import Samples** (Импортиране на проби) — Използвайте шаблонен файл, наличен за изтегляне на екрана Sample Data (Данни за проби).
- **Manually** (Ръчно) — Въведете данните от пробите директно в таблицата на екрана Sample Data (Данни за проби).

Импортиране на проби

Когато планирате изпълняване на секвениране с браузър на мрежов компютър, на екрана Sample Data (Данни за проби) е наличен шаблонен файл (* . csv). Шаблонният файл не е наличен за изтегляне при достъп Illumina Run Manager през софтуера на операционната система на NextSeq 550Dx. За да въведете данни за проби с помощта на функцията Import Samples (Импортиране на проби), изпълнете следните стъпки.

ЗАБЕЛЕЖКА Изпълнете стъпките Run Settings (Настройка на изпълняване), преди да продължите.

1. Изберете **Download Template** (Изтегляне на шаблон), за да изтеглите празен CSV файл.
2. От шаблонния файл, въведете примерни данни и след това запазете файла. Името на библиотеката не е задължително.

ЗАБЕЛЕЖКА Когато се използва Index A&B (Индекс АиБ), данните за колона В трябва да включват както позиция на плаката, така и позиция на ямката (позиция на ямката на индексната плака). Пример: A-A01, A-A02, A-A03.

3. Изберете **Import Samples** (Импортиране на проби) и отидете до шаблонния файл, съдържащ информацията за данните за пробите от предишната стъпка.
4. Изберете **Open** (Отваряне), **Proceed** (Продължаване), а след това **Next** (Напред).

ЗАБЕЛЕЖКА Промяната на Sample ID (ИД на проба) преди избиране на Next (Напред) може да доведе до грешка. Завършете настройката на изпълняването, преди да направите промени, за да избегнете грешки.

Ръчно въвеждане на пробите

Използвайте таблицата на екрана Sample Data (Данни за проби) за ръчно въвеждане на данни за проби.

1. Въведете уникален ИД на пробата в полето Sample ID (ИД на пробата).

2. Използвайте **Well Position** (Позиция на ямките) или **Plate - Well Position** (Плака - позиция на ямките) (Index A&B) (Индекс АиБ), за да изберете съответния индекс за пробите. Полетата i7 Index, Index 1, i5 Index и Index 2 се попълват автоматично.
3. **[Незадължително]** Въведете име на библиотека.
4. Добавете редове и повторете стъпки 1–3, ако е необходимо, докато всички проби бъдат добавени към таблицата. Можете да добавите няколко реда наведнъж, като първо въведете броя на редовете, които да добавите, и след това изберете иконата +. Можете също да премахнете редовете, като изберете полето до номера на реда и след това щракнете върху иконата за боклук.
5. Изберете **Next** (Напред).

Настройки на анализа

1. Изберете желан работен поток за анализ:
 - Генериране на FASTQ
 - Генериране на FASTQ и VCF за работен процес на герминативна линия (необходим е манифестен файл)
 - Генериране на FASTQ и VCF за работен процес на соматична линия (необходим е манифестен файл)
2. **[Незадължително] Generate ORA compressed FASTQs** (Генериране на ORA компресирани FASTQ) се активира по подразбиране. Компресията FASTQ ORA без загуба компресира файловете FASTQ до 5 пъти в сравнение с fastq.gz. Премахнете отметката от **Generate ORA compressed FASTQs** (Генериране на ORA компресирани FASTQ), ако се предпочитат некомпресирани данни (fastq.gz).
3. За герминативни и соматични работни потоци е необходим манифестен файл. Използвайте падащото меню **Manifest File Selection** (Избор на манифестен файл), за да изберете манифестен файл. Манифестът е BED (*.bed) файл с разделен с табулатори, който определя имената и местоположенията на целевите референтни области. За повече информация, вижте [Манифестен файл на стр. 8](#).
4. **[Незадължително]** За соматични работни потоци използвайте падащото меню **Noise File Selection** (Избор на шумов файл), за да изберете систематичен шумов файл. За филтриране на системния шум може да се зададе файл BED (*.bed.gz) със специфично за обекта ниво на шум. За повече информация, вижте [Филтриране на шума \(незадължително\) на стр. 9](#).
5. Изберете **Next** (Напред).

Изпълняване Преглед

1. На екрана Преглед, прегледайте информацията за Run Settings (Настройки за изпълняване), Sample Data (Данни за пробите) и Analysis Settings (Настройки за анализ).
2. Изберете **Save** (Записване). Изпълняването се запазва в раздела Planned (Планирани) на екрана Runs (Изпълнявания).

Настройки

За да прегледате или промените DRAGEN for IDPE Dx настройките на приложението, първо изберете иконата Applications (Приложения) от основния екран. След това, изберете приложението, което искате да прегледате или промените. За промяна на настройките е необходим профил на администратор.

Конфигурация

Екранът за конфигуриране показва следните настройки на приложението:

- **Library Prep Kits** (Комплекти за подготовка за библиотеката) — Показва комплекта за подготовка за библиотеката по подразбиране за приложението. Тази настройка не може да бъде променена.
- **Index Adapter Kits** (Комплекти адаптери за индекс) – Показва комплекта адаптери за индекс по подразбиране за приложението. Тази настройка не може да бъде променена.
- **Read lengths** (Дължини на разчитане) — Дължините на разчитане са зададени на 151 за приложението по подразбиране, но могат да бъдат променени по време на създаване на изпълняване.
- **Manifest and Noise Files** (Манифестни файлове и файлове за шум) — Качване и промяна на настройки за манифестните файлове и файловете за шум.
 - Изберете **Upload File** (Качване на файл), за да качите файлове за използване при анализ.
 - Изберете радио бутон **Default** (По подразбиране), за да зададете файла като манифестен файл или шумов файл по подразбиране, избран по време на създаването на изпълняване, когато приложението е избрано.
 - Изберете полето **Enabled** (Активирано), за да зададете файла, който да се показва в падащото меню по време на създаването на изпълняване.

Разрешения

Използвайте квадратчетата за отметка на екрана Permissions (Разрешения), за да управлявате достъпа на потребителя за приложението.

Манифестен файл

Когато използвате DRAGEN for IDPE Dx, се изисква манифестен файл за следните работни потоци:

- Генериране на FASTQ и VCF за работен процес на герминативна линия
- Генериране на FASTQ и VCF за работен процес на соматична линия

Манифестният файл е разделен с табулатори текстов файл, който използва BED формат (*.bed), който определя имената и местоположенията на целевите референтни области. Основният раздел на манифестния файл е раздел Regions (Региони) и трябва да съдържа следните колони с данни:

Колона	Описание
Име	Уникално, определено от потребителя име за целта
Хромозома	Хромозомно местоположение (напр. chr10, chr5 и т.н.)
Начало	1 базиран индекс за началната позиция на целта
Спиране	1 базиран индекс за крайна позиция на целта
Дължина на сондата нагоре по веригата	Дължината на горната сонда. За приложението DRAGEN for IDPE Dx, това трябва да бъде настроено на 0.
Дължина на сондата надолу по веригата	Дължината на долна сонда. За приложението DRAGEN for IDPE Dx, това трябва да бъде настроено на 0.

ЗАБЕЛЕЖКА За анализ се изисква валиден формат на манифестния файл. DRAGEN ще спре анализа, ако манифестният файл е невалиден.

Филтриране на шума (незадължително)

Системният филтър за шум е наличен за соматични варианти на обозначаване и може да се използва за намаляване на фалшиви положителни обозначавания, като се отчита специфичният за обекта шум. Системният шумов файл се генерира чрез първо събиране на приблизително 50 нормални проби (за предпочитане специфични за панела, подготовката на библиотеката и секвенсера) и след това сумата от честотите на алелите под 30% във всеки център с достатъчно покритие се разделя на общия брой проби (прието е, че алелните честоти над 30% са варианти на герминативната линия, а не шум). След генериране на стойностите на шума, соматичните варианти, открити на това място, ще бъдат филтрирани.

Филтърът може да се използва в режим Tumor-Normal (Нормален туморен режим), но е особено полезен само за туморни изпълнявания, при които не е налично съответстващо нормално състояние. Системният шумов файл трябва да използва BED файл, който има разширение на файл (*.bed.gz) и трябва да включва четири колони: Хромозоми, Начало, Край и специфични за центъра нива на шум за всеки ред. Системното филтриране на шума е по избор.

Изходни данни от анализа

Изпълнявания, които в момента текат, се показват в раздел Active (Активни). Завършените цикли се показват в раздел Completed (Завършени). DRAGEN for IDPE Dx създава уникално именуван папка за анализ за всеки анализ, която е отделна от папката, съдържаща данни за секвениране. Папката за анализ включва следната информация:

- Използван манифестен файл
- Версия на софтуера
- Идентификационни номера на проби

- Общо подравнени разчитания
- Процент подравнени разчитания на проба
- Брой обозначени SNV на проба
- Брой обозначени индели на проба
- Статистически данни

Изходни файлове на анализ

Местоположението на папката за анализ се определя от настройката External Storage for Analysis Results (Външно съхранение за резултати от анализ). Вижте Illumina Run Manager за софтуерно ръководство NextSeq 550Dx (документ № 200025239) за повече информация относно настройката External Storage for Analysis Results (Външно съхранение за резултати от анализ).

На екрана Run Details (Подробности за изпълняването), полето External Location (Външно местоположение) предоставя пътя за секвениране на данни. Уникалното име на папката за анализ е предоставено в полето Analysis Output Folder (Папка за изход от анализа) на екрана Run Details (Подробности за изпълняването). Точните генерирани файлове зависят от това кой работен поток за анализ се използва. Следните изходни файлове за анализ се генерират от приложението.

ЗАБЕЛЕЖКА Ако възникне грешка в ограничението на максималната дължина на пътя на файла при достъп до изходните файлове за анализ, опитайте да преместите файла на местоположение с по-кратък път или използвайте друг метод, за да отворите файла.

Изходен файл	Описание
Обобщен отчет на варианта (*.pdf)	Съдържа резюме на информацията за файла, версиите на софтуера, информация за пробата, статистика на нивото на разчитане и SNV, инсерции, делеции и обобщения на покритието. Само герминативните и соматичните работни потоци произвеждат вариантен отчет.
FASTQ (*.fastq.gz or *.fastq.ora)	Междинни файлове, които съдържат обозначавания на бази с резултат за качество. FASTQ файловете са основната входна информация за стъпката на подравняване. Когато е избрана ORA компресия, се използва разширението на файла *.fastq.ora.
BAM файлове за подравняване (*.bam)	Съдържа подравнени разчитания за дадена проба.
Геномни VCF файлове (*.gvcf.gz)	Съдържа генотипа за всяка позиция, независимо дали е обозначена като вариант, или справка.
VCF файлове (*.vcf.gz)	Съдържа варианти, обозначени за всяка позиция.
Изпълнение на отчет за показатели (*.csv)	Съдържа показатели за качество на цикъла, включително неиндексиран общ добив и оценка Q30.

Файлове FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) е текстово-базиран файлов формат, съдържащ обозначавания на бази и стойности на качеството на едно разчитане. Всеки файл съдържа следната информация:

- Идентификатор на проба
- Последователността
- Резултатът за качество Phred се оценява в кодиран формат ASCII + 33

Идентификаторът на проба е представен, както следва:

```
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNumber
Example:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<?????#=#
```

ВAM файлове

ВAM файлът (*.bam) е компресирана двоична версия на SAM (карта за подравняване на последователност) файл, който се използва за представяне на подравнени последователности до 128 Mb. ВAM файловете използват формат за именуване на файла `SampleName_S#.bam`, където # е номерът на пробата, определен от последователността, в която пробите са посочени за изпълняването. В многовъзлов режим, S# е настроен на S1, независимо от реда на пробата.

ВAM файловете съдържат раздел за горен колонтитул и раздел за подравняване:

- **Горен колонтитул** – Съдържа информация за целия файл, като име на пробата, дължина на пробата и метод на подравняване. Подравняванията в раздела за подравнявания са свързани със специфична информация в раздела за горен колонтитул.
- **Подравнявания** – Съдържа име на разчитането, последователност на разчитането, качество на разчитането, информация за подравняване и персонализирани тагове. Името на разчитането включва хромозомата, началната координата, качеството на подравняване и низа на дескриптора за съвпадение.

Разделът за подравнявания включва следната информация за всяко разчитане или двойка разчитания:

- AS: Качество на сдвоено подравняване.
- RG: Група за разчитане, която посочва броя разчитания за конкретна проба.
- BC: Баркод таг, който показва идентификатора на демултиплексираната проба, свързан с разчитането.
- SM: Качество на подравняване на единични краища.
- XC: Низ на дескриптора за съвпадение.
- XN: Таг за името на ампликон, който записва идентификатора на ампликона, свързан с разчитането

Индексните файлове BAM (*.bam.bai) предоставят индекс на съответния BAM файл.

VCF файлове

Файловете за обозначаване на варианти (*.vcf) съдържат информация относно варианти, открити на конкретни позиции в референтен геном.

Горният колонтитул на VCF файла включва версията на файловия формат VCF, версията на инструмента за обозначаване на вариант и посочва анотациите, използвани в останалата част на файла. Горният колонтитул на VCF също включва файла на референтния геном и BAM файла. Последният ред в горния колонтитул съдържа заглавията на колоните за редовете с данни. Всеки от редовете за данни на VCF файла съдържа информация за един вариант.

Таблица 1 Заглавия на VCF файл

Заглавие	Описание
CHROM	Хромозомата от референтния геном. Хромозомите се появяват в същата последователност както в референтния FASTA файл.
POS	Еднобазова позиция на варианта в референтната хромозома. За единични нуклеотидни варианти (SNV), тази позиция е референтната база с варианта. За платформите, тази позиция е референтната база, непосредствено предхождаща варианта.
ID	RS номерът (SNP за справка) за SNP, получен от <code>dbSNP.txt</code> , ако е приложимо. Ако съществуват няколко rs номера на това местоположение, списъкът се разделя с двоеточие. Ако не съществува dbSNP запис в тази позиция, се използва маркер за липсваща стойност ('.').
REF	Референтният генотип. Например делеция на единична T е представено като референтна TT и алтернативна T. А към T единичен нуклеотиден вариант е представен като референтна A и алтернативна T.
ALT	Алелите, които се различават от референтното разчитане. Например инсерция на единична T е представена като референтна A и алтернативна AT. А към T единичен нуклеотиден вариант е представен като референтна A и алтернативна T.
QUAL	Резултат за качество, определен чрез Phred скалата, назначен от инструмента за обозначаване на варианти. По-високият резултат показва по-висока доверителност във варианта и по-малка възможност за грешки. За резултат за качество на Q изчислената възможност за грешка е $10^{-Q/10}$. Например комплект от Q30 обозначавания има честота на грешки 0,1%. Много инструменти за обозначаване на варианти назначават резултати за качество на основата на техните статистически модели, които са високи във връзка с наблюдаваната честота на грешки.

Таблица 2 Анотации на VCF файл на герминативната линия

Заглавие	Описание
ФИЛТЪР	<p>Ако всички филтри преминат, в колоната за филтъра се изписва ПРЕМИНАВА.</p> <p>Възможните въвеждания на ФИЛТЪР включват:</p> <ul style="list-style-type: none"> • DRAGENSnpHardQUAL – Прилага се, ако SNP вариантът QUAL не отговаря на прага • DRAGENIndelHardQUAL — Прилага се, ако резултатът QUAL за вариант на индел не отговаря на прага • LowDepth — Филтрирано от центъра, защото дълбочината на покритие не отговаря на прага • LowGQ — Филтрирано от центъра, защото качеството на генотипа не отговаря на прага • PloidyConflict – Генотипно обозначаване, което не съответства на хромозомната пloidия • base_quality — Филтрирано от центъра, защото средното качество на базата на alt разчитания при този локус не отговаря на прага • filtered_reads — Филтрирано от центъра, защото твърде голяма част от разчитанията е филтрирана • fragment_length — Филтрирано от центъра, защото абсолютната разлика между средната дължина на фрагмента на alt разчитанията и средната дължина на фрагмента на ref разчитанията при този локус превишава прага • low_depth — Филтрирано от центъра, защото дълбочината на разчитане е твърде ниска • low_frac_info_reads — Филтрирано от центъра, защото частта от информативните разчитания е под прага • low_normal_depth — Филтрирано от центъра, защото нормалната дълбочината на разчитане на пробата е твърде ниска • long_indel — Филтрирано от центъра, защото дължината на индел е твърде дълга • mapping_quality — Филтрирано от центъра, защото средното качество на картографиране на alt разчитания при този локус не отговаря на прага • multiallelic — Филтрирано от центъра, защото повече от два alt алела преминават през туморен LOD • non_homref_normal — Филтрирано от центъра, защото нормалният генотип на пробата не е хомозиготна референция • no_reliable_supporting_read — Филтрирано от центъра, защото не съществува надеждно поддържащо соматично разчитане • panel_of_normals — Вижда се в най-малко една проба в панела на нормалностните vcf • read_position – Филтрирано от центъра, защото медианата на разстоянията между началото/края на разчитането и този локус е под прага • RMxNRepeatRegion — Филтрирано от центъра, защото целият или част от вариантния алел е повторение на референтния • strand_artifact — Филтрирано от мястото поради силно отклонение на веригата • str_contraction – Филтрирано от центъра поради подозирана PCR грешка, при която alt алелът е една повтаряща се единица по-малка от референтната • too_few_supporting_reads — Филтрирано от центъра, защото в туморната проба има твърде малко поддържащи разчитания • weak_evidence — Резултатът от соматичните варианти не отговаря на прага

Заглавие	Описание
ИНФОРМАЦИЯ	<p>Възможните ИНФОРМАЦИЯ записи включват:</p> <ul style="list-style-type: none"> • AC – Броят алели в генотиповете за всеки ALT алел в същата последователност, както е посочено. • AF – Честотата на алелите за всеки ALT алел в същата последователност, както е посочено. • AN – Общият брой алели в посочения генотип. • DB — Членство в DbSNP. • FS — р-стойност по Phred скалата с помощта на точния тест на Fisher за откриване на отклонения на веригата. • QD – Доверителност/качество на варианта по дълбочина. • R2_5P_bias — Оценка въз основа на mate отклонение и разстояние от 5 края на прайминг. • SOR — Симетрично съотношение на коефициентите на таблицата за непредвидени обстоятелства 2x2 за откриване на отклонения на веригата. • DP — Приблизителна дълбочина на разчитане (информативна и неинформативна); някои разчитания може да са били филтрирани въз основа на mapq и т.н. • END — Позицията на спиране на интервала. • FractionInformativeReads — Частта от информативните разчитания от общите разчитания. • MQ — RMS качество на картографиране. • MQRankSum — Z-резултат от теста на Wilcoxon на Alt спрямо качества на картографирането на Ref разчитане. • ReadPosRankSum — Z-резултат от теста на Wilcoxon за Alt спрямо отклоненията на позициите на Ref разчитане. • SOMATIC — Най-малко един вариант в тази позиция е соматичен.

Заглавие	Описание
ФОРМАТ	<p>В колоната за формат са изброени полета, разделени с двоеточие. Например GT:GQ. Наличните полета включват:</p> <ul style="list-style-type: none"> • AD — Алелни дълбочини (броене само на информативни разчитания от общите разчитания) за ref и alt алели в посочения ред. • AF — Фракции на алели за alt алели в посочения ред. • DP – Приблизителна дълбочина на разчитане (разчитания с MQ=255 или с лоши свързвания, които са филтрирани). • F1R2 — Брой разчитания в ориентацията на двойката F1R2, поддържащи всеки алел. • F2R1 — Брой разчитания в ориентацията на двойката F2R1, поддържащи всеки алел. • GT – Генотип. 0 отговаря на референтната база, 1 отговаря на първия запис в колоната ALT и така нататък. Наклонената черта (/) показва, че не е налична информация за фазиране. • MB —Статистика на компонентите за всяка проба за откриване на mate отклонение. • PS — Информация за физикалната идентификация на фазите, при която всеки уникален идентификационен номер в дадена проба (но не и между проби) свързва записи в рамките на група на фазите. • SB — Статистика на компонентите за всяка проба, която включва точния тест на Fisher за откриване на отклонения на веригата. • SQ – Соматично качество.
ПРОБА	Колоната за проба дава стойностите, специфични за колоната ФОРМАТ.

Таблица 3 Анотации на VCF файл на соматичната линия

Заглавие	Описание
ФИЛТЪР	<p>Ако всички филтри преминат, в колоната за филтъра се изписва ПРЕМИНАВА. Възможните въвеждания на ФИЛТЪР включват:</p> <ul style="list-style-type: none"> • base_quality — Филтрирано от центъра, защото средното качество на базата на alt разчитания при този локус не отговаря на прага • filtered_reads — Филтрирано от центъра, защото твърде голяма част от разчитанията са филтрирани • fragment_length — Филтрирано от центъра, защото абсолютната разлика между средната дължина на фрагмента на alt разчитанията и средната дължина на фрагмента на ref разчитанията при този локус превишава прага • low_depth — Филтрирано от центъра, защото дълбочината на разчитане е твърде ниска • low_frac_info_reads — Филтрирано от центъра, защото частта от информативните разчитания е под прага • low_normal_depth — Филтрирано от центъра, защото нормалната дълбочината на разчитане на пробата е твърде ниска • long_indel — Филтрирано от центъра, защото дължината на индел е твърде дълга • mapping_quality — Филтрирано от центъра, защото средното качество на картографиране на alt разчитания при този локус не отговаря на прага • multiallelic — Филтрирано от центъра, защото повече от два alt алела преминават през туморен LOD • non_homref_normal — Филтрирано от центъра, защото нормалният генотип на пробата не е хомозиготна референция • no_reliable_supporting_read — Филтрирано от центъра, защото не съществува надеждно поддържащо соматично разчитане • panel_of_normals — Вижда се в най-малко една проба в панела на нормалностните vcf • read_position – Филтрирано от центъра, защото медианата на разстоянията между началото/края на разчитането и този локус е под прага • RMxNRepeatRegion — Филтрирано от центъра, защото целият или част от вариантния алел е повторение на референтния • strand_artifact — Филтрирано от мястото поради силно отклонение на веригата • str_contraction – Филтрирано от центъра поради подозирана PCR грешка, при която alt алелът е една повтаряща се единица по-малка от референтната • too_few_supporting_reads — Филтрирано от центъра, защото в туморната проба има твърде малко поддържащи разчитания • weak_evidence — Резултатът от соматичните варианти не отговаря на прага • systematic_noise — Филтрирано от центъра въз основа на данни за систематичен шум в норма

Заглавие	Описание
ИНФОРМАЦИЯ	<p>Възможните ИНФОРМАЦИЯ записи включват:</p> <ul style="list-style-type: none"> • DP — Приблизителна дълбочина на разчитане (информативна и неинформативна); някои разчитания може да са били филтрирани въз основа на mapq и т.н. • END — Позицията на спиране на интервала. • FractionInformativeReads — Частта от информативните разчитания от общите разчитания. • MQ — RMS качество на картографиране. • MQRankSum — Z-резултат от теста на Wilcoxon на Alt спрямо качества на картографирането на Ref разчитане. • ReadPosRankSum — Z-резултат от теста на Wilcoxon за Alt спрямо отклоненията на позициите на Ref разчитане. • AQ — Оценка на систематичен шум. • hotspot — Известен соматичен център, използван за повишаване на доверието в обозначаването. • SOMATIC — Най-малко един вариант в тази позиция е соматичен.

Заглавие	Описание
ФОРМАТ	<p>В колоната за формат са изброени полета, разделени с двоеточие. Например GT:GQ. Наличните полета включват:</p> <ul style="list-style-type: none"> • AD — Алелни дълбочини (броене само на информативни разчитания от общите разчитания) за ref и alt алели в посочения ред. • AF — Фракции на алели за alt алели в посочения ред. • DP – Приблизителна дълбочина на разчитане (разчитания с MQ=255 или с лоши свързвания, които са филтрирани). • F1R2 — Брой разчитания в ориентацията на двойката F1R2, поддържащи всеки алел. • F2R1 — Брой разчитания в ориентацията на двойката F2R1, поддържащи всеки алел. • GP — Вероятности за генотипове с постериорна Phred скала, както е определено в спецификацията на VCF. • GQ — Качество на генотипа. • GT – Генотип. 0 отговаря на референтната база, 1 отговаря на първия запис в колоната ALT и така нататък. Наклонената черта (/) показва, че не е налична информация за фазиране. • MB — Статистика на компонентите за всяка проба за откриване на mate отклонение. • PL — Нормализирани вероятности по Phred скалата за генотипове, както е определено в спецификацията на VCF. • PRI — Предварителни вероятности по Phred скалата за генотипове. • PS — Информация за физикалната идентификация на фазите, при която всеки уникален идентификационен номер в дадена проба (но не и между проби) свързва записи в рамките на група на фазите. • SB — Статистика на компонентите за всяка проба, която включва точния тест на Fisher за откриване на отклонения на веригата. • SQ – Соматично качество.
ПРОБА	Колоната за проба дава стойностите, специфични за колоната ФОРМАТ.

Геномни VCF файлове

Геномните VCF (*.gvcf.gz) файлове следват набор от конвенции за представяне на всички места в генома в приемливо компактен формат. gVCF файловете включват всички места в рамките на района на интерес в един файл за всяка проба. gVCF файлът показва липсата на обозначавания на позиции, които не преминават всички филтри. Таг за генотип (GT) ./. показва, че няма обозначаване.

Повторно нареждане на анализ на опашка

Може да наредите повторно анализ на опашка при спиране на анализа, ако анализът е бил неуспешен или ако искате да анализирате отново изпълняване с различни настройки. За да заявите анализ, изпълнете следните стъпки:

1. От екрана Run (Изпълняване) изберете раздела Completed (Завършено) и след това изберете името на изпълнението, за да анализирате повторно.
Ако преди това е извършено повторно нареждане на анализ на опашка, изберете името на изпълняването на родителското изпълняване.
2. От екрана Run Details (Подробности за изпълняването), след Sequencing Information (Информация за секвениране), изберете **Requeue Analysis** (Повторно нареждане на анализ на опашка).
3. Изберете опция:
 - Повторно нареждане на анализ на опашка без промени
 - Редактиране на настройките на повторно нареждане на анализ на опашка
 - Повторно нареждане на анализ на опашка с различно приложение
4. Потвърдете, че местоположението, където понастоящем се намират данните за секвениране, е предоставено в полето за **Sequencing data file path** (Път до файла с данни за секвениране).

ЗАБЕЛЕЖКА Пътят до данните за секвениране трябва да съответства на пътя в настройката External Storage for Analysis Results (Външно съхранение за резултати от анализ). Вижте Illumina Run Manager за софтуерно ръководство NextSeq 550Dx (документ № 200025239) за информация относно промяната на пътя за външно съхранение.

5. Въведете причина за повторен анализ.
6. Изберете **Requeue Analysis** (Повторно нареждане на анализ на опашка).
7. Редактирайте желаните промени в Run Settings (Настройки за изпълняване), Sample Data (Данни за пробите) и Analysis Settings (Настройки за анализ).
8. Изберете **Save** (Записване). Анализът започва с използване на текущите параметри на анализа.

Техническа помощ

За техническа помощ се свържете с Illumina отдела за техническа поддръжка.

Уеб сайт: www.illumina.com

Имейл: techsupport@illumina.com

Информационни листове за безопасност (ИЛБ) – налични на уеб сайта на Illumina на адрес support.illumina.com/sds.html.

Документация на продукта – налична за изтегляне от support.illumina.com.



Illumina

5200 Illumina Way

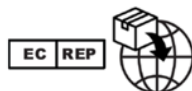
San Diego, California 92122, САЩ

+1.800.809.ILMN (4566)

+1.858.202.4566 (извън Северна Америка)

techsupport@illumina.com

www.illumina.com



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

Спонсор в Австралия

Illumina Australia Pty Ltd

Nursing Association Building

Level 3, 535 Elizabeth Street

Melbourne, VIC 3000

Австралия

ЗА IN VITRO ДИАГНОСТИЧНА УПОТРЕБА.

© 2023 Illumina, Inc. Всички права запазени.

illumina[®]