illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx

Dokumentacja produktu NovaSeq 6000Dx

ZASTRZEŻONE MATERIAŁY FIRMY ILLUMINA
Nr dokumentu: 200014776, wer. 02
Wrzesień 2022 r.

DO CELÓW DIAGNOSTYKI IN VITRO

Historia wersji

Dokument	Data	Opis zmiany
200014776 wer. 02	Wrzesień 2022 r.	Poprawiono format pliku wykazu z tekstowego (*.txt) do BED (*.bed) w instrukcji tworzenia przebiegu. Poprawiono uzgodnione pliki VCF do plików VCF w punkcie dotyczącym wyników analizy.
200014776 wer. 01	Sierpień 2022 r.	Dodano: Punkt Ustawienia. Punkt Filtrowanie szumów systematycznych Zaktualizowano instrukcje tworzenia przebiegu, aby uwzględnić więcej szczegółów. Poprawiono literówki i błędy gramatyczne. Określono, że instrukcje są przeznaczone dla aplikacji, gdy są używane z urządzeniem NovaSeq 6000Dx. Zaktualizowano informacje dotyczące zawartości pliku wyjściowego VCF.
200014776 wer. 00	Marzec 2022 r.	Pierwsze wydanie.

Spis treści

Przegląd

DRAGEN™ dla aplikacji Illumina® DNA Prep with Enrichment Dx wykonuje demultipleksowanie, generowanie FASTQ, mapowanie odczytu i dopasowanie do genomu referencyjnego oraz rozpoznawanie wariantu w zależności od wybranej procedury.

Metody analizy

DRAGEN for Illumina DNA Prep with Enrichment Dx wykonuje demultipleksowanie, generowanie FASTQ, mapowanie odczytu i dopasowanie do genomu referencyjnego w zależności od wybranej procedury:

- Generowanie pliku FASTQ
- Germline FASTQ i VCF generation
- Somatic FASTQ i VCF generation

Generowanie pliku FASTQ

Zmontowane sekwencje są zapisywane w plikach FASTQ dla każdej próbki. Pliki FASTQ to pliki tekstowe zawierające dane sekwencjonowania i wyniki jakości dla tylko jednej próbki. Dla każdej próbki generowane są oddzielne pliki FASTQ dla pasma komory przepływowej, na odczyt sekwencjonowania. Nazwa próbki określona podczas konfiguracji przebiegu jest zawarta w nazwie pliku FASTQ. Pliki FASTQ to podstawowe dane wejściowe do dopasowywania. Pierwszym etapem generowania FASTQ jest demultipleksowanie. Demultipleksowanie przypisuje klastry, które przejdą przez filtr, do próbki poprzez porównanie każdej sekwencji odczytu indeksu z sekwencjami indeksu określonymi dla danego przebiegu. Na tym etapie nie są brane pod uwagę żadne wartości dotyczące jakości. Odczyty indeksów identyfikuje się poprzez następujące czynności:

- Próbki są ponumerowane, począwszy od numeru 1, na podstawie kolejności, w jakiej je wymieniono dla przebiegu.
- Próbka numer 0 jest zarezerwowana dla klastrów, które nie zostały przypisane do próbki.
- Klastry są przypisywane do próbki, gdy sekwencja indeksu jest dokładnie zgodna lub gdy na odczyt indeksu przypada maksymalnie jedna niezgodność.

Oprogramowanie zawiera kompresję ORA do kompresowania plików FASTQ. W przypadku korzystania z formatu ORA (*.ora) suma kontrolna md5 zawartości FASTQ jest zachowywana po cyklu kompresji i dekompresji, aby zapewnić kompresję bezstratną.

Dopasowywanie i mapowanie DNA

Pierwszym etapem mapowania jest wygenerowanie punktów początkowych z odczytu, a następnie wyszukanie dokładnych dopasowań w genomie referencyjnym. Wyniki te są następnie poprawiane poprzez wykonanie pełnych dopasowań Smitha-Watermana w lokalizacjach o najwyższej gęstości

dopasowanych punktów początkowych. Ten dobrze udokumentowany algorytm działa poprzez porównanie każdej pozycji odczytu ze wszystkimi pozycjami kandydującymi odniesienia. Porównania te odpowiadają matrycy potencjalnych dopasowań między odczytem a odniesieniem. Dla każdej z tych pozycji kandydujących do dopasowania metoda Smith-Waterman generuje wyniki, które są wykorzystywane do oceny, czy najlepsze dopasowanie przechodzące przez tę komórkę matrycy dociera do niej przez dopasowanie, czy niezgodność nukleotydów (ruch ukośny), delecję (ruch poziomy) lub insercję (ruch pionowy). Dopasowanie między odczytem a odniesieniem zapewnia premię do wyniku, a niedopasowanie lub indel nakłada karę. Ścieżka o najwyższej ogólnej punktacji w matrycy to wybrane dopasowanie.

Konkretne wartości wybrane dla wyników w tym algorytmie wskazują, jak zrównoważyć, do dopasowania z wieloma możliwymi interpretacjami, możliwość indelu w przeciwieństwie do jednego lub więcej SNP lub preferencję dopasowania bez przycinania. Domyślne wartości punktacji DRAGEN są uzasadnione dla dopasowania średnich długości odczytów do całego ludzkiego genomu referencyjnego dla zastosowań związanych z wykrywaniem wariantów. Każdy zestaw parametrów punktacji Smitha-Watermana reprezentuje nieprecyzyjny model błędów mutacji genomowej i sekwencjonowania. W przypadku niektórych zastosowań bardziej odpowiednie mogą być wartości punktacji inaczej dostosowanego dopasowania.

Wykrywanie wariantu linii zarodkowej DRAGEN

DRAGEN Germline Small Variant Caller przyjmuje zmapowane i dopasowane odczyty DNA jako dane wejściowe i rozpoznaje SNP oraz indele poprzez połączenie wykrywania kolumnowego i lokalnego łączenia haplotypów *de novo*.

Rozpoznawalne obszary referencyjne są najpierw identyfikowane z wystarczającym pokryciem dopasowania. W obrębie tych obszarów odniesienia szybkie skanowanie posortowanych odczytów identyfikuje obszary aktywne, które są wyśrodkowane wokół kolumn nagromadzenia z dowodami wariantu. Te obszary aktywne są wyścielane kontekstem wystarczającym do pokrycia pobliskich znaczących treści bez odniesień. Jeśli istnieją dowody na indele, obszary aktywne otrzymują dodatkową wyściółkę.

Dopasowane odczyty są przycinane w każdym aktywnym obszarze i łączone na wykresie De Bruijn. Krawędzie przyciętych odczytów są ważone według liczby obserwacji, a sekwencja odniesienia stanowi podstawę. Po pewnym oczyszczeniu i uproszczeniu wykresu wszystkie ścieżki źródło-odbiornik są wyodrębniane jako kandydujące haplotypy. Każdy haplotyp jest dopasowywany metodą Smitha-Watermana do genomu referencyjnego w celu identyfikacji reprezentowanych przez niego wariantów. Ten zestaw zdarzeń może zostać uzupełniony przez wykrywanie oparte na położeniu. Dla każdej pary odczyt-haplotyp prawdopodobieństwo obserwacji odczytu $P(r|H)$, przy założeniu, że dany haplotyp jest rzeczywistą próbą początkową, jest szacowane przy użyciu pary ukrytego modelu Markowa (HMM).

Dzięki skanowaniu według pozycji referencyjnej na obszarze aktywnym, kandydujące genotypy są tworzone z diploidalnych kombinacji zdarzeń wariantowych (SNP lub indeli). Dla każdego zdarzenia (w tym odniesienia) warunkowe prawdopodobieństwo obserwacji $P(r|e)$ każdego zachodzącego na siebie

odczytu jest szacowane jako maksymalna wartość P(r|H) dla haplotypów uzasadniających zdarzenie. Są one łączone w warunkowe prawdopodobieństwo P(r|e1e2) dla genotypu (pary zdarzeń) i mnożone w celu uzyskania warunkowego prawdopodobieństwa P(R|e1e2) obserwacji całego nagromadzenia odczytów. Przy użyciu wzoru Bayesa oblicza się prawdopodobieństwo a posteriori P(e1e2|R) każdego genotypu diploidalnego, a zwycięzca zostaje wykryty.

W trybie gVCF stosowanym do skalowalnego rozpoznawania wariantów w wielu próbkach, DRAGEN Germline Small Variant Caller może być uruchamiany na próbkę w celu wygenerowania pliku pośredniego rozpoznania wariantu genomowego (gVCF). Następnie gVCF można wykorzystać do skutecznego wspólnego genotypowania wielu próbek, co umożliwia szybkie przyrostowe przetwarzanie próbek i skalowanie do dużych rozmiarów kohort.

Ponieważ DRAGEN Germline Small Variant Caller jest wyposażony w algorytmy, które umożliwiają skuteczne rozróżnianie skorelowanych błędów od rzeczywistych wariantów, reguły filtrowania są bardzo proste.

Wykrywanie wariantu somatycznego DRAGEN

DRAGEN Germline Small Variant Caller przyjmuje zmapowane i wyrównane odczyty DNA jako dane wejściowe i rozpoznaje SNV i indele poprzez lokalny łączenie *de novo* haplotypów w obszarze aktywnym.

Rozpoznawalne obszary referencyjne są najpierw identyfikowane z wystarczającym pokryciem dopasowania. W obrębie tych obszarów odniesienia skanowanie posortowanych odczytów identyfikuje obszary aktywne, które są wyśrodkowane wokół kolumn nagromadzenia z dowodami wariantu w odczytach nowotworu. Te obszary aktywne są wyścielane kontekstem wystarczającym do pokrycia pobliskich znaczących treści bez odniesień. Jeśli są dowody na indele, obszary aktywne otrzymują dodatkową wyściółkę.

Dopasowane odczyty są przycinane w każdym aktywnym obszarze i łączone na wykresie De Bruijn. Krawędzie przyciętych odczytów są ważone według liczby obserwacji, a sekwencja odniesienia stanowi podstawę. Po pewnym oczyszczeniu i uproszczeniu wykresu wszystkie ścieżki źródło-odbiornik są wyodrębniane jako kandydujące haplotypy. Każdy haplotyp jest dopasowywany metodą Smitha-Watermana do genomu referencyjnego w celu identyfikacji reprezentowanych przez niego wariantów. Dla każdej pary odczyt-haplotyp prawdopodobieństwo obserwacji odczytu P(r|H) szacuje się przy użyciu pary ukrytego modelu Markowa (HMM) przy założeniu, że halotyp ten jest rzeczywistą próbką początkową.

Aby określić wynik TLOD, DRAGEN Somatic Small Variant Caller najpierw skanuje według pozycji referencyjnej każde kandydujące zdarzenie somatyczne, jak również zdarzenie referencyjne w danym obszarze aktywnym. Prawdopodobieństwo warunkowe P(r|e) obserwacji każdego zachodzącego na siebie odczytu jest szacowane jako maksymalna wartość P(r|H) dla haplotypów potwierdzających zdarzenie. Są one łączone w warunkowe prawdopodobieństwo P(r|E) dla hipotezy zdarzenia, E, obejmujące mieszaninę alleli odniesienia i kandydata w zakresie możliwych częstotliwości alleli i

mnożone w celu uzyskania warunkowego prawdopodobieństwa P(R|E) obserwacji całego nagromadzenia odczytów. Z tego poziomu obliczany jest wynik TLOD jako dowód na obecność allelu ALT w próbce guza w danym locus.

Tworzenie przebiegu

Aby skonfigurować przebieg w Illumina Run Manager (Menedżer przebiegu Illumina) na NovaSeq 6000Dx lub za pomocą przeglądarki na komputerze sieciowym, należy wykonać poniższe czynności. Dane próbki można wprowadzić ręcznie lub importując arkusz próbki.

Ustawienia i uruchamianie aplikacji

1. Na ekranie Runs (Przebiegi) wybierz opcję **Create run** (Tworzenie przebiegu).
2. Wybierz aplikację DRAGEN for Illumina DNA Prep with Enrichment Dx, a następnie **Next** (Dalej).
3. Na ekranie Run Settings (Ustawienia przebiegu) wprowadź nazwę przebiegu. Nazwa przebiegu to nazwa, która identyfikuje przebieg od sekwencjonowania po analizę.
4. **[Opcjonalnie]** Wprowadź opis przebiegu, aby dodatkowo ułatwić jego identyfikację.
5. Upewnij się, że wybrany Library Prep Kit (Zestaw do przygotowania biblioteki) jest zestawem przygotowania bibliotekiIllumina DNA Prep with Enrichment Dx.
6. Wybierz żądany zestaw adaptera indeksującego.
7. Wprowadź Read Length (Długość odczytu).
   Wartości domyślne Read 1 (Odczyt 1) i Read 2 (Odczyt 2) to 151 cykli.
   Index 1 (Indeks 1) i Index 2 (Indeks 2) mają stałą wartość 10 cykli.
8. **[Opcjonalnie]** Wprowadź ID probówki Library.
9. Wybierz **Next** (Dalej).

Dane próbki

Użyj tabeli na ekranie Sample Data (Dane próbki), aby ręcznie wprowadzić informacje o próbce. Można również wybrać opcję **Import Samples** (Import próbek), aby przesłać informacje o próbce. Informacje na temat importowania informacji o próbkach można znaleźć w punkcie *Importowanie próbek* na stronie 6.

Ręczne wprowadzanie próbek

1. Wprowadź unikalny identyfikator próbki w polu Sample ID (Identyfikator próbki).
2. Użyj **Plate - Well Position** (Płytka – pozycja dołka), aby wybrać pozycję dołka.
   Pola i7 Index, Index 1, i5 Index i Index 2 wypełniają się automatycznie.
3. **[Opcjonalne]** Wprowadź nazwę biblioteki.
4. Dodaj wiersze i powtarzaj kroki od 1do3, według potrzeb, aż wszystkie próbki zostaną dodane do tej tabeli.
5. Wybierz **Next** (Dalej).

Importowanie próbek

Szablon (*.csv) jest dostępny do pobrania na ekranie Sample Data (Dane próbki) podczas planowania przebiegu w Illumina Run Manager (Menedżer przebiegu Illumina) przy użyciu przeglądarki na komputerze sieciowym.

1. Wybierz opcję **Download Template** (Pobierz szablon), aby pobrać pusty plik CSV.
2. Z tego pliku CSV wprowadź informacje o próbce i zapisz plik.
   Arkusz próbki w pliku CVS zawiera następujące kolumny danych: Sample ID (Identyfikator próbki), Plate — Well Position (Płytka — pozycja dołka), **opcjonalna** Library Name (Nazwa biblioteki).
3. Wybierz opcję **Import Samples** (Import próbek) i przejdź do lokalizacji pliku CSV.
4. Wybierz **Next** (Dalej).

Ustawienia analizy

1. Wybierz żądaną procedurę analizy:
   - Generowanie pliku FASTQ
   - Germline FASTQ i VCF generation dla procedury dla linii zarodkowej
   - Somatic FASTQ i VCF generation dla procedury somatycznej
2. **[Opcjonalnie]** W razie potrzeby zaznacz pole wyboru **Generate ORA compressed FASTQs (Generowanie ORA-kompresowanych plików FASTQ**, aby włączyć ORA kompresję FASTQ.
3. **[Procedury VCF generation]** Użyj menu rozwijanego **Manifest File Selection** (Wybór pliku wykazu), aby wybrać plik wykazu.
   Wprowadzenie pliku wykazu jest wymagane dla DRAGEN for Illumina DNA Prep with Enrichment Dx. Wykaz jest plikiem BED (*.bed) rozdzielanym tabulatorami, który określa nazwy i lokalizacje docelowych obszarów odniesienia.
4. **[Procedura Somatic FASTQ i VCF generation]** Użyj menu rozwijanego **Noise File Selection** (Wybór pliku szumu), aby wybrać plik szumu.
   Można określić plik BED z poziomem szumu charakterystycznym dla danego miejsca w celu odfiltrowania systematycznego szumu. Więcej informacji podano w punkcie *Filtrowanie szumów* na stronie 7.
5. Wybierz **Next** (Dalej).

Przebieg Przegląd

1. Na ekranie Review (Przegląd) wyświetla się informacje wprowadzone na ekranach Run Settings (Ustawienia przebiegu), Sample Data (Dane próbki) i Analysis Settings (Ustawienia analizy).
2. Wybierz **Save** (Zapisz).
   Przebieg zostanie zapisany na karcie Planned (Zaplanowane) na ekranie Runs (Przebiegi).

Ustawienia

Wybierz aplikację na ekranie Applications (Aplikacje), aby wyświetlić bieżące ustawienia i zmienić je.

Konfiguracja

Na ekranie konfiguracji wyświetlane są następujące ustawienia aplikacji:

- **Library Prep Kits** – wyświetla domyślny zestaw do przygotowania biblioteki dla danej aplikacji. Tego ustawienia nie można zmienić.
- **Index Adapter Kits**– wyświetla domyślny zestaw adapterów indeksu dla danej aplikacji. Tego ustawienia nie można zmienić.
- **Read lengths**– długości odczytu są domyślnie ustawione na 151 dla danej aplikacji, ale można je zmienić podczas tworzenia przebiegu.
- **Manifest and Noise Files**– przesyłanie i zmiana ustawień dla plików wykazu i szumu.
  - Wybierz opcję **Upload File (Prześlij plik)**, aby przesłać pliki do wykorzystania w analizie.
  - Wybierz przycisk opcji **Default (Domyślne)**, aby ustawić plik jako domyślny plik wykazu lub plik szumu wybrany podczas tworzenia przebiegu po wybraniu danej aplikacji.
  - Zaznacz pole wyboru **Enabled (Włączone)**, aby ustawić plik do wyświetlenia w menu rozwijanym podczas tworzenia przebiegu.

Uprawnienia

Użyj pól wyboru na ekranie Permissions (Uprawnienia), aby zarządzać dostępem użytkowników do aplikacji.

Filtrowanie szumów

Podczas korzystania z procedury somatycznej można filtrować szum systematyczny. Filtr ten może być używany w trybie Tumor-Normal (Guz-normalny), ale jest szczególnie przydatny w przy przebiegach Tumor-Only (Tylko guz), przy których nie jest dostępny dopasowany tryb normalny.

BED dla szumu systematycznego należy generować z próbek normalnych. Zaleca się tworzenie plików szumu systematycznego, które są specyficzne dla przygotowania biblioteki, systemu sekwencjonowania i panelu. Zaleca się użycie około 50 normalnych próbek do generowania pliku szumu.

Dane wyjściowe analizy

DRAGEN for Illumina DNA Prep with Enrichment Dx zapisuje następujące informacje w folderze analizy. Tylko procedury dla linii zarodkowej i somatycznej tworzą plik PDF.

- Użyto pliku wykazu
- Wersja oprogramowania
- Identyfikator próbki
- Łączna liczba dopasowanych odczytów
- Procent dopasowanych odczytów na próbkę
- Liczba rozpoznanych SNV na próbkę
- Liczba indeli rozpoznanych na próbkę
- Statystyki pokrycia

Pliki wyjściowe analizy

Następujące pliki wyjściowe są generowane przez tę aplikację. Dokładne wygenerowane pliki zależą od tego, którą procedurę analizy zastosowano. Pliki wyjściowe znajdują się w folderze analizy.

Plik wyjściowy	Opis
FASTQ (*.fastq.gz lub *.fastq.ora)	Pliki pośrednie zawierające wyniki jakościowe rozpoznawania nukleotydów. Pliki FASTQ to podstawowe dane wejściowe etapu dopasowywania. Jeśli wybrano kompresję ORA, odzwierciedla to nazwa pliku.
Dopasowanie plików BAM (*.bam)	Zawiera dopasowane odczyty dla danej próbki.
Pliki VCF genomu (*.gvcf.gz)	Zawierają genotyp dla każdej pozycji, rozpoznany jako wariant lub materiał referencyjny.
Pliki VCF (*.vcf.gz)	Zawierają warianty rozpoznane w każdej pozycji.
Uruchamianie raportu wskaźników (*.csv)	Zawiera wskaźniki jakości dotyczące przebiegu, w tym wynik uzysku całkowitego i Q30.

Pliki FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) to format pliku tekstowego, który zawiera rozpoznane nukleotydy i wartości dotyczące jakości dla każdego odczytu. Każdy plik zawiera następujące informacje:

- Identyfikator próbki
- Sekwencja
- Znak plus (+)
- Wyniki jakościowe w skali Phred w formacie kodowania ASCII + 33

Identyfikator próbki jest sformatowany w następujący sposób.

```
@Aparat:IDprzebiegu:IDKomoryPrzepływowej:Pasmo:Płytka:X:Y
NrOdczytu:FlagaFiltra:0:NumerPróbki
Przykład:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Pliki BAM

Plik BAM (*.bam) jest skompresowaną, binarną wersją pliku SAM (sequence alignment map – mapa dopasowania sekwencji), który służy do przechowywania odwzorowań dopasowanych sekwencji o wielkości do 128 Mb. W odniesieniu do plików BAM stosowany jest format nazwy `SampleName_S#.bam (NazwaPróbki_S#.bam)`, gdzie „#” jest numerem próbki wynikającym z kolejności, w jakiej próbki uszeregowano do przebiegu. W trybie wielowęzłowym numer S# jest ustawiony na S1, niezależnie od kolejności próbek.

Pliki BAM zawierają sekcje nagłówka i dopasowań:

- Header (Nagłówek) – zawiera informacje dotyczące całego pliku, takie jak nazwa próbki, długość próbki i metoda dopasowania. Elementy w sekcji dopasowań są powiązane z określonymi informacjami w sekcji nagłówka.
- Alignments (Dopasowania) – zawiera nazwę odczytu, sekwencję odczytu, jakość odczytu, informacje o dopasowaniu oraz znaczniki niestandardowe. Nazwa odczytu zawiera oznaczenia chromosomu, współrzędne początkowe, jakość dopasowania oraz ciąg deskryptora dopasowań.

Sekcja dopasowań zawiera następujące informacje dla każdego odczytu lub sparowanego odczytu:

- AS: jakość dopasowania odczytu w trybie sparowanych końców.
- RG: grupa odczytów wskazująca liczbę odczytów dla danej próbki.

- BC: znacznik kodu kreskowego, który wskazuje identyfikator demultipleksowanej próbki powiązany z danym odczytem.
- SM: jakość dopasowania odczytu w trybie pojedynczego końca.
- XC: ciąg deskryptora dopasowań.
- XN: znacznik nazwy amplikonu, który rejestruje identyfikator amplikonu powiązanego z plikami indeksu BAM (*.bam.bai) odczytu, zawiera indeks odpowiadającego mu pliku BAM.

Pliki VCF

Pliki formatu rozpoznania wariantu (*.vcf) zawierają informacje o wariantach znalezionych w określonych pozycjach w genomie referencyjnym.

Nagłówek pliku VCF zawiera wersję formatu pliku VCF i wersję algorytmu do rozpoznawania wariantów. Wymienione są w nim również adnotacje używane w pozostałej części pliku. Nagłówek VCF zawiera także plik genomu referencyjnego i plik BAM. Ostatni wiersz nagłówka zawiera nagłówki kolumny dla wierszy danych. Każdy wiersz danych pliku VCF zawiera informację o pojedynczym wariancie.

Tabela 1 Nagłówki pliku VCF

Nagłówek	Opis
CHROM (Chromosom)	Chromosom genomu referencyjnego. Chromosomy pojawiają się w takiej samej kolejności jak w referencyjnym pliku FASTA.
POS (Pozycja)	Pozycja pojedynczego nukleotydu w wariancie w chromosomie referencyjnym. W przypadku wariantów pojedynczych nukleotydów (SNV) ta pozycja jest nukleotydem referencyjnym dla danego wariantu. W przypadku indeli pozycja ta jest nukleotydem referencyjnym bezpośrednio poprzedzającym dany wariant.
ID (Identyfikator)	Numer rs (odniesienia SNP) dla SNP uzyskanego z pliku `dbSNP.txt`, jeśli ma to zastosowanie. Jeśli w danej lokalizacji jest wiele numerów rs, lista jest rozdzielona średnikami. Jeśli w danej pozycji nie ma żadnego wpisu dbSNP, stosuje się znacznik brakującej wartości ('.').
REF (Referencja)	Genotyp referencyjny. Przykładowo delecja jednego nukleotydu T jest przedstawiana jako referencyjny allel TT i alternatywny T. Wariant pojedynczego nukleotydu A do T jest przedstawiany jako referencyjny allel A i alternatywny T.
ALT (Alternatywa)	Allele różniące się od odczytu referencyjnego. Na przykład: insercja jednego nukleotydu T jest przedstawiana jako referencyjny allel A i alternatywny AT. Wariant pojedynczego nukleotydu A do T jest przedstawiany jako referencyjny allel A i alternatywny T.

Nagłówek	Opis
QUAL (Jakość)	Wynik jakościowy w skali Phred przypisany przez algorytm do rozpoznawania wariantów. Wyższe wyniki wskazują wyższy poziom ufności w odniesieniu do wariantu i niższe prawdopodobieństwo błędów. W przypadku wyniku jakościowego Q szacowane prawdopodobieństwo błędu wynosi 10–(Q/10). Na przykład zestaw rozpoznań z wynikiem Q30 ma odsetek błędów równy 0,1%. Wiele algorytmów rozpoznawania wariantów przypisuje wyniki jakościowe na podstawie swoich modeli statystycznych, które są wysokie w stosunku do obserwowanego odsetka błędów.

Tabela 2 Adnotacje w pliku VCF

Nagłówek	Opis
FILTER (Filtr)	Jeśli wariant przeszedł przez wszystkie filtry, w kolumnie filtru wpisywana jest informacja PASS (Powodzenie). Możliwe wpisy FILTER procedury linii zarodkowej obejmują: • **DRAGENSnpHardQUAL** – stosowany, jeśli wynik QUAL wariantu SNP nie spełnia progu • **DRAGENIndelHardQUAL** – stosowany, jeśli wynik QUAL wariantu indel nie spełnia progu • **LowDepth** – miejsce filtrowane, ponieważ głębokość pokrycia nie spełnia progu • **LowGQ** – miejsce filtrowane, ponieważ jakość genotypu nie spełnia progu • **PloidyConflict** – rozpoznanie genotypu z algorytmu rozpoznawania wariantów nie jest zgodne z ploidią chromosomu. • **base_quality** – miejsce filtrowane, ponieważ mediana jakości nukletotydów odczytów alternatywnych w tym locus nie spełnia progu • **filtered_reads** – miejsce filtrowane, ponieważ odfiltrowano zbyt dużą część odczytów • **fragment_length** – miejsce filtrowane, ponieważ różnica bezwzględna między medianą długości fragmentu odczytów alternatywnych a medianą długości fragmentów odczytów referencyjnych w tym locus przekracza próg • **low_depth** – miejsce filtrowane, ponieważ głębokość odczytu jest zbyt mała • **low_frac_info_reads** – miejsce filtrowane, ponieważ frakcja odczytów informatywnych jest poniżej progu • **low_normal_depth** – miejsce filtrowane, ponieważ normalna głębokość odczytu próbek jest zbyt mała • **long_indel** – miejsce filtrowane, ponieważ długość indela jest zbyt duża • **mapping_quality** – miejsce filtrowane, ponieważ mediana jakości odwzorowania odczytów alternatywnych w tym locus nie spełnia progu • **multiallelic** – miejsce filtrowane, ponieważ więcej niż dwa allele alternatywne przekraczają LOD nowotworu. • **non_homref_normal** – miejsce filtrowane, ponieważ genotyp próbki normalnej nie jest odniesieniem homozygotycznym • **no_reliable_supporting_read** – miejsce filtrowane, ponieważ nie ma wiarygodnego potwierdzającego odczytu somatycznego • **panel_of_normals** – obserwowany w co najmniej jednej próbce w panelu normalnych vcf • **read_position** – miejsce filtrowane, ponieważ mediana odległości między początkiem/końcem odczytu i tym locus jest poniżej progu • **RMxNRepeatRegion** – miejsce filtrowane, ponieważ całość lub część allelu tego wariantu jest powtórzeniem tego odniesienia • **strand_artifact** – miejsce filtrowane z powodu poważnego obciążenia systematycznego nici

Nagłówek	Opis
FILTER (ciąg dalszy)	• **str_contraction** – miejsce filtrowane z powodu podejrzewanego błędu PCR, gdzie alternatywny allel jest o jedną jednostkę powtórzenia mniejszy niż odniesienie • **too_few_supporting_reads** – miejsce filtrowane, ponieważ w próbce guza jest zbyt mało pomocniczych odczytów • **weak_evidence** – wynik wariantu somatycznego nie spełnia progu Możliwe wpisy w polu FILTER (Filtr) dla procedury somatycznej: • **base_quality** – miejsce filtrowane, ponieważ mediana jakości nukletotydów odczytów alternatywnych w tym locus nie spełnia progu • **filtered_reads** – miejsce filtrowane, ponieważ odfiltrowano zbyt dużą frakcję odczytów • **fragment_length** – miejsce filtrowane, ponieważ różnica bezwzględna między medianą długości fragmentu odczytów alternatywnych a medianą długości fragmentów odczytów referencyjnych w tym locus przekracza próg • **low_depth** – miejsce filtrowane, ponieważ głębokość odczytu jest zbyt mała • **low_frac_info_reads** – miejsce filtrowane, ponieważ frakcja odczytów informatywnych jest poniżej progu • **low_normal_depth** – miejsce filtrowane, ponieważ normalna głębokość odczytu próbek jest zbyt mała • **long_indel** – miejsce filtrowane, ponieważ długość indela jest zbyt duża • **mapping_quality** – miejsce filtrowane, ponieważ mediana jakości odwzorowania odczytów alternatywnych w tym locus nie spełnia progu • **multiallelic** – miejsce filtrowane, ponieważ więcej niż dwa allele alternatywne przekraczają LOD nowotwóru. • **non_homref_normal** – miejsce filtrowane, ponieważ genotyp próbki normalnej nie jest odniesieniem homozygotycznym • **no_reliable_supporting_read** – miejsce filtrowane, ponieważ nie ma wiarygodnego potwierdzającego odczytu somatycznego • **panel_of_normals** – obserwowany w co najmniej jednej próbce w panelu normalnych vcf • **read_position** – miejsce filtrowane, ponieważ mediana odległości między początkiem/końcem odczytu i tym locus jest poniżej progu • **RMxNRepeatRegion** – miejsce filtrowane, ponieważ całość lub część allelu tego wariantu jest powtórzeniem tego odniesienia • **strand_artifact** – miejsce filtrowane z powodu poważnego obciążenia systematycznego nici • **str_contraction** – miejsce filtrowane z powodu podejrzewanego błędu PCR, gdzie alternatywny allel jest o jedną jednostkę powtórzenia mniejszy niż odniesienie • **too_few_supporting_reads** – miejsce filtrowane, ponieważ w próbce guza jest zbyt mało pomocniczych odczytów • **weak_evidence** – wynik wariantu somatycznego nie spełnia progu

Nagłówek	Opis
FILTER (ciąg dalszy) INFO (Informacje)	• **systematic_noise** – miejsce filtrowane na podstawie dowodów systematycznego szumu w wariantach normalnych Możliwe wpisy w polu INFO (Informacje) dla procedury linii zarodkowej obejmują: • **AC** – liczba alleli w genotypach dla każdego allelu ALT, w takiej samej kolejności, jak na liście • **AF** – częstość alleli dla każdego allelu ALT, w takiej samej kolejności, jak na liście • **AN** – łączna liczba alleli w rozpoznanych genotypach • **DB** – udział w dBSNP • **FS** – wartość P skalowana według Phred przy użyciu dokładnego testu Fishera w celu wykrycia obciążenia systematycznego nici • **QD** – pewność/jakość wariantu według głębokości • **R2_5P_bias** – wynik oparty na obciążeniu mate i odległości od 5 głównych końców • **SOR** – symetryczny iloraz szans w tabeli kontyngencji 2x2 w celu wykrycia obciążenia systematycznego nici • **DP** – przybliżona głębokość odczytu (informatywna i nieinformatywna); niektóre odczyty mogły zostać odfiltrowane na podstawie mapq itp. • **END** – pozycja zatrzymania interwału • **FractionInformativeReads** – udział odczytów informatywnych w całkowitej liczbie odczytów • **MQ** – jakość mapowania RMS • **MQRankSum** – Z-score w teście sumy rang Wilcoxona dla jakości mapowania odczytów alternatywnych w porównaniu z referencyjnymi • **ReadPosRankSum** – Z-score w teście sumy rang Wilcoxona dla obciążenia pozycji alternatywnych w porównaniu z referencyjnymi • **SOMATIC** – co najmniej jeden wariant w tej pozycji jest somatyczny Możliwe wpisy w polu INFO (Informacja) dla procedury somatycznej: • **DP** – przybliżona głębokość odczytu (informatywna i nieinformatywna); niektóre odczyty mogły zostać odfiltrowane na podstawie mapq itp. • **END** – pozycja zatrzymania interwału • **FractionInformativeReads** – udział odczytów informatywnych w całkowitej liczbie odczytów • **MQ** – jakość mapowania RMS • **MQRankSum** – Z-score w teście sumy rang Wilcoxona dla jakości mapowania odczytów alternatywnych w porównaniu z referencyjnymi • **ReadPosRankSum** – Z-score w teście sumy rang Wilcoxona dla obciążenia pozycji alternatywnych w porównaniu z referencyjnymi • **AQ** – wynik szumu systematycznego • **hotspot** – znane miejsce somatyczne, używane do zwiększania wiarygodności rozpoznania • **SOMATIC** – co najmniej jeden wariant w tej pozycji jest somatyczny

Nagłówek	Opis
FORMAT (Format)	W kolumnie formatu wymienione są pola rozdzielone dwukropkami. Przykładowo: GT:GQ Dostępne pola dla procedury linii zarodkowej obejmują: • **AD**– głębokości alleli (zliczanie tylko informatywnych odczytów z całkowitej liczby odczytów) dla alleli referencyjnych i alternatywnych w podanej kolejności • **AF**– frakcje alleli dla alleli alternatywnych w podanej kolejności • **DP** – przybliżona głębokość odczytu (odczyty z wartością MQ = 255 lub ze złym dopasowaniem par są odfiltrowywane) • **F1R2**– liczba odczytów w orientacji pary F1R2 potwierdzających każdy allel • **F2R1**– liczba odczytów w orientacji pary F2R1 potwierdzających każdy allel • **GP**– prawdopodobieństwa a posteriori w skali Phred dla genotypów zgodnie ze specyfikacją VCF • **GQ** – jakość genotypu • **GT** – genotyp. 0 odpowiada nukleotydowi referencyjnemu, 1 odpowiada pierwszemu wpisowi w kolumnie ALT itd. Ukośnik (/) wskazuje brak informacji o fazowaniu. • **MB**– statystyki komponentów na próbkę w celu wykrycia obciążenia parowania • **PL**– znormalizowane prawdopodobieństwo genotypów w skali Phreda zdefiniowane w specyfikacji VCF • **PRI**– prawdopodobieństwo a priori genotypów w skali Phread • **PS**– informacje o identyfikatorze fazowania fizycznego, gdzie każdy unikatowy identyfikator w danej próbce (ale nie w wielu próbkach) łączy rekordy w grupie fazowania • **SB**– statystyki komponentów na próbkę, które składają się na dokładny test Fishera w celu wykrycia obciążenia systematycznego nici • **SQ**– jakość somatyczna Pola dostępne dla procedury somatycznej to: • **AD**– głębokości alleli (zliczanie tylko informatywnych odczytów z całkowitej liczby odczytów) dla alleli referencyjnych i alternatywnych w podanej kolejności • **AF**– frakcje alleli dla alleli alternatywnych w podanej kolejności • **DP** – przybliżona głębokość odczytu (odczyty z wartością MQ = 255 lub ze złym dopasowaniem par są odfiltrowywane) • **F1R2**– liczba odczytów w orientacji pary F1R2 potwierdzających każdy allel • **F2R1**– liczba odczytów w orientacji pary F2R1 potwierdzających każdy allel • **GT** – genotyp. 0 odpowiada nukleotydowi referencyjnemu, 1 odpowiada pierwszemu wpisowi w kolumnie ALT itd. Ukośnik (/) wskazuje brak informacji o fazowaniu. • **MB**– statystyki komponentów na próbkę w celu wykrycia obciążenia parowania

Nagłówek	Opis
FORMAT (ciąg dalszy)	• **PS**– informacje o identyfikatorze fazowania fizycznego, gdzie każdy unikatowy identyfikator w danej próbce (ale nie w wielu próbkach) łączy rekordy w grupie fazowania • **SB**– statystyki komponentów na próbkę, które składają się na dokładny test Fishera w celu wykrycia obciążenia systematycznego nici • **SQ**– jakość somatyczna
SAMPLE (Próbka)	W kolumnie próbek są podane wartości wskazane w kolumnie FORMAT.

Pliki VCF genomu

Pliki VCF genomu (*.gvcf.gz) są zgodne ze zbiorem konwencji dotyczących przedstawienia wszystkich miejsc w genomie w dość kompaktowym formacie. Pliki gVCF zawierają wszystkie miejsca w obszarze docelowym w jednym pliku dla każdej próbki. W pliku gVCF w pozycjach, które nie przeszły wszystkich filtrów, przedstawione są nierozpoznane nukleotydy. Znacznik genotypu (GT) „./.” wskazuje nierozpoznany nukleotyd.

Wyświetlanie wyników analizy

Aktualnie trwające przebiegi wyświetlane są na karcie Active (Aktywne). Przebiegi ukończone wyświetlane są na karcie Completed (Ukończone). Więcej informacji na temat przeglądania wyników podano w Dokumentacja produktu NovaSeq 6000Dx (nr dokumentu 200010105).

Pomoc techniczna

W celu uzyskania pomocy technicznej należy skontaktować się z działem pomocy technicznej firmy Illumina.

Witryna: www.illumina.com
Adres e-mail: techsupport@illumina.com

Numery telefonów do działu pomocy technicznej firmy Illumina

Region	Bezpłatne	Międzynarodowy
Australia	+61 1800 775 688	
Austria	+43 800 006249	+43 1 9286540
Belgia	+32 800 77 160	+32 3 400 29 73
Kanada	+1 800 809 4566	
Chiny		+86 400 066 5835
Dania	+45 80 82 01 83	+45 89 87 11 56
Finlandia	+358 800 918 363	+358 9 7479 0110
Francja	+33 8 05 10 21 93	+33 1 70 77 04 46
Niemcy	+49 800 101 4940	+49 89 3803 5677
Hongkong, Chiny	+852 800 960 230	
Indie	+91 8006500375	
Indonezja		0078036510048
Irlandia	+353 1800 936608	+353 1 695 0506
Włochy	+39 800 985513	+39 236003759
Japonia	+81 0800 111 5011	
Malezja	+60 1800 80 6789	
Holandia	+31 800 022 2493	+31 20 713 2960
Nowa Zelandia	+64 800 451 650	
Norwegia	+47 800 16 836	+47 21 93 96 93
Filipiny	+63 180016510798	
Singapur	1 800 5792 745	
Korea Południowa	+82 80 234 5300	

Region	Bezpłatne	Międzynarodowy
Hiszpania	+34 800 300 143	+34 911 899 417
Szwecja	+46 2 00883979	+46 8 50619671
Szwajcaria	+41 800 200 442	+41 56 580 00 00
Tajwan, Chiny	+886 8 06651752	
Tajlandia	+66 1800 011 304	
Wielka Brytania	+44 800 012 6019	+44 20 7305 7197
Stany Zjednoczone	+1 800 809 4566	+1 858 202 4566
Wietnam	+84 1206 5263	

Karty charakterystyki – dostępne na stronie firmy Illumina pod adresem support.illumina.com/sds.html.

Dokumentacja produktu jest dostępna do pobrania w witrynie support.illumina.com.

Illumina
5200 Illumina Way
San Diego, California 92122, USA
+1 800 809 ILMN (4566)
+1 858 202 4566 (poza Ameryką Północną)
techsupport@illumina.com
www.illumina.com

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

Sponsor australijski
Illumina Australia Pty Ltd
Nursing Association Building
Level 3, 535 Elizabeth Street
Melbourne, VIC 3000
Australia

DO CELÓW DIAGNOSTYKI IN VITRO