

illumina®

# TruSight Whole Genome Analysis Application

Dokumentacja produktu

ZASTRZEŻONE MATERIAŁY FIRMY ILLUMINA

Nr dokumentu: 200049931 wer. 00

Kwiecień 2024 r.

DO STOSOWANIA W DIAGNOSTYCE IN VITRO.

## Historia wersji

<b>Dokument</b>	<b>Data</b>	<b>Opis zmiany</b>
Nr dokumentu: 200049931 wer. 00	Kwiecień 2024 r.	Pierwsze wydanie.

Niniejszy dokument oraz jego treść stanowią własność firmy Illumina, Inc., a także jej podmiotów zależnych („Illumina”), i są przeznaczone wyłącznie do użytku zgodnego z umową przez klienta firmy w związku z użytkowaniem produktów opisanych w niniejszym dokumencie, z wyłączeniem innych celów. Niniejszy dokument oraz jego treść nie będą wykorzystywane ani rozpowszechniane w innych celach i/lub publikowane w inny sposób, ujawniane ani kopiowane bez pisemnej zgody firmy Illumina. Firma Illumina na podstawie niniejszego dokumentu nie przenosi żadnych licencji podlegających przepisom w zakresie patentów, znaków towarowych czy praw autorskich ani prawu powszechnemu lub prawom pokrewnym osób trzecich.

W celu zapewnienia właściwego i bezpiecznego użytkowania produktów opisanych w niniejszym dokumencie podane instrukcje powinny być ściśle przestrzegane przez wykwalifikowany i właściwie przeszkolony personel. Przed rozpoczęciem użytkowania tych produktów należy zapoznać się z całą treścią niniejszego dokumentu.

**NIEZAPOZNANIE SIĘ LUB NIEDOKŁADNE PRZESTRZEGANIE WSZYSTKICH INSTRUKCJI PODANYCH W NINIEJSZYM DOKUMENCIE MOŻE SPOWODOWAĆ USZKODZENIE PRODUKTÓW LUB OBRAŻENIA CIAŁA UŻYTKOWNIKÓW LUB INNYCH OSÓB ORAZ USZKODZENIE INNEGO MIENIA, A TAKŻE SPOWODUJE UNIEWAŻNIENIE WSZELKICH GWARANCJI DOTYCZĄCYCH PRODUKTÓW.**

**FIRMA ILLUMINA NIE PONOSI ODPOWIEDZIALNOŚCI ZA NIEWŁAŚCIWE UŻYTKOWANIE PRODUKTÓW (W TYM ICH CZĘŚCI I OPROGRAMOWANIA) OPISANYCH W NINIEJSZYM DOKUMENCIE.**

© 2024 Illumina, Inc. Wszelkie prawa zastrzeżone.

Wszystkie znaki towarowe są własnością firmy Illumina, Inc. lub ich odpowiednich właścicieli. Szczegółowe informacje na temat znaków towarowych można znaleźć pod adresem [www.illumina.com/company/legal.html](http://www.illumina.com/company/legal.html).

# Spis treści

Historia wersji .....	ii
Przegląd .....	1
Rozpoczęcie .....	1
Ustawienia .....	2
Tworzenie przebiegu .....	4
Zmiana przebiegu .....	5
Ponowne umieszczenie analizy w kolejce .....	6
Przegląd przebiegu i wyników .....	7
Podsumowanie pliku wyników .....	9
Informacje na raporcie dotyczącym kontroli jakości .....	9
Informacje o pliku rozpoznań wariantów .....	14
Pliki FASTQ .....	22
Pliki CRAM .....	23
<b>Pomoc techniczna .....</b>	<b>25</b>

# Przegląd

TruSight Whole Genome Analysis Application służy do planowania przebiegu sekwencjonowania TruSight Whole Genome i automatycznego inicjowania analizy po zakończeniu przebiegu. Analiza obejmuje demultipleksację, generowanie FASTQ, mapowanie odczytu, dopasowanie do ludzkiego genomu referencyjnego GrCh38/hg38 z funkcją wykresów oraz rozpoznawanie wariantu przy użyciu Illumina DRAGEN Server for NovaSeq 6000Dx.

Na różnych etapach analizy procedury pracy aplikacja przeprowadza kontrolę jakości (QC) zgodnie z określonymi sekwencjonowaniem, FASTQ i parametrami biblioteki próbek oraz generuje raporty z wynikami. W przypadku próbek, które spełniają wymagania wszystkich etapów kontroli jakości, aplikacja generuje pomocnicze pliki wyników do wykorzystania w aplikacjach linii zarodkowej.

TruSight Whole Genome Analysis Application wykorzystuje algorytmy rozpoznawania wariantów DRAGEN, w tym Algorytm rozpoznawania wariantów małych, Algorytm rozpoznawania poliformizmu liczby kopii i Wykrywanie ekspansji powtórzeń przy użyciu ExpansionHunter.

Aplikacja wykonuje również adnotacje o niskim, średnim lub wysokim poziomie ufności dla małych wariantów i uwzględnia tę adnotację w pliku wyjściowym.

## Rozpoczęcie

Upewnij się, że TruSight Whole Genome Analysis Application jest zainstalowany w NovaSeq 6000Dx instrument, które będzie używane do sekwencjonowania w ramach TruSight Whole Genome. Zainstalowane aplikacje można znaleźć na ekranie Aplikacje na NovaSeq 6000Dx Instrument lub Illumina Run Manager w przeglądarce na komputerze sieciowym. Aby uzyskać pomoc w zaplanowaniu instalacji, skontaktuj się z lokalnym przedstawicielem terenowym firmy Illumina.

### Wymagania dotyczące przechowywania danych

Więcej informacji na temat danych dotyczących wyników i przechowywania można znaleźć w punkcie Dokumentacja produktu NovaSeq 6000Dx (nr dokumentu 200010105) i Dokumentacja produktu DRAGEN Server for NovaSeq 6000Dx (dokument nr 200014171).

Dane wyników TruSight Whole Genome Analysis Application są wysyłane do folderu Run (Przebieg) i folderu Analysis (Analiza) w pamięci zewnętrznej. Minimalne wymagania dotyczące przechowywania danych można oszacować w stosunku do rozmiaru danych wyjściowych do każdego folderu dla pojedynczego przebiegu sekwencjonowania przedstawionego poniżej.

Konfiguracja	Folder Run (Przebieg) (GB)	Folder Analysis (Analiza) (GB)
Komórka przepływowa S2 (6 próbek)	~430	~350
Komórka przepływowa S4 (16 próbek)	~1110	~890

## Przybliżony czas analizy

Analiza rozpoczyna się automatycznie po zakończeniu przebiegu sekwencjonowania i następuje kolejno na próbkach w ramach przebiegu. Pliki danych wyników będą dostępne w magazynie zewnętrznym po zakończeniu analizy dla wszystkich próbek w przebiegu i zakończeniu kopiowania transferu do magazynu zewnętrznego. Podczas rozpoczynania przebiegu sekwencjonowania po stronie A i po stronie B w tym samym czasie sekwencjonowanie będzie wykonywane jednocześnie. Analiza tych przebiegów sekwencjonowania zostanie przeprowadzona kolejno przez TruSight Whole Genome Analysis Application po zakończeniu sekwencjonowania. Przebieg, który zakończy sekwencjonowanie i zostanie przesłany jako pierwszy, zostanie przeanalizowany jako pierwszy. Drugi przebieg sekwencjonowania zostanie przeniesiony i umieszczony w kolejce do analizy po zakończeniu pierwszej analizy. Informacje na temat określania statusu aktywnych lub nieudanych serii można znaleźć w części [Przegląd przebiegu i wyników na stronie 7](#).

Przybliżony czas do uzyskania wyników analizy po zakończeniu sekwencjonowania przedstawiono poniżej dla sytuacji, gdy strona A i strona B są załadowane jednocześnie z tą samą konfiguracją.

Konfiguracja	Przebieg analizy 1 (godziny)	Przebieg analizy 2 (godziny)
Komórka przepływowa S2 (6 próbek)	~12	~24
Komórka przepływowa S4 (16 próbek)	~24	~48

## Ustawienia

Wybierz TruSight Whole Genome Analysis Application na ekranie Applications (Aplikacje), aby wyświetlić bieżącą konfigurację i zmienić uprawnienia.

### Konfiguracja

Na ekranie konfiguracji wyświetlane są następujące ustawienia aplikacji:

- **Application Name (Nazwa aplikacji)**
- **Application Version (Wersja aplikacji)**
- **DRAGEN Version (Wersja Dragen)**

- **RTA Version (Wersja RTA)**
- **Release Date (Data wersji)**
- **Organization (Organizacja)**
- **Device Identifier (Identyfikator aparatu)**
- **Production Identifier (Identyfikator produkcji)**
- **Library Prep Kits** (Zestawy do przygotowania biblioteki) – wyświetla zestaw do przygotowania biblioteki. Tego ustawienia nie można zmienić.
- **Index Adapter Kits** (Zestawy adapterów indeksujących) – wyświetla zestawy adaptera indeksującego dostępne do użycia.
- **Index Reads (Odczyty indeksów)**
- **Read Type (Typ odczytu)**
- **Index Lengths (Długości indeksu)**
- **Read Lengths** (Długości odczytu) – długości odczytu są ustawiane domyślnie po wybraniu zestawu indeksów. Tego ustawienia nie można zmienić.

## Uprawnienia

Wyznaczony administrator ma dostęp do uprawnień i może używać pól wyboru na ekranie Permissions (Uprawnienia) do zarządzania dostępem użytkowników do TruSight Whole Genome Analysis Application.

Więcej informacji na temat uprawnień i zarządzania użytkownikami można znaleźć w rozdziale Konfiguracja systemu w Dokumentacja produktu NovaSeq 6000Dx (nr dokumentu 200010105).

# Tworzenie przebiegu

Tworzenie nowych przebiegów w trybie IVD w aparacie lub poprzez uzyskanie dostępu do Illumina Run Manager (IRM) za pomocą przeglądarki na komputerze podłączonym do sieci. Aby uzyskać zdalny dostęp do aparatu, należy skorzystać z adresu i danych konta użytkownika podanych przez przedstawiciela firmy Illumina. Więcej informacji można znaleźć w Dokumentacja produktu NovaSeq 6000Dx (nr dokumentu 200010105).

Tworzenie przebiegu jest zalecaną metodą planowania przebiegu. Nie zaleca się importowania arkusza przykładowego. Foldery przykładowych plików arkuszowych, które są generowane w folderach serii i analizy, nie nadają się do importowania podczas planowania przebiegów.

## Tworzenie przebiegu

1. Na ekranie Runs (Przebiegi) wybrać opcję **Create run** (Tworzenie przebiegu).
2. Wybrać aplikację TruSight Whole Genome Analysis Application, a następnie **Next** (Dalej).
3. Na ekranie Run Settings (Ustawienia przebiegów) wprowadzić nazwę przebiegu. Nazwa przebiegu to nazwa, która identyfikuje przebieg od sekwencjonowania po analizę.
4. [Opcjonalnie] Wprowadzić opis przebiegu, aby dodatkowo ułatwić jego identyfikację. Library Prep kit (Zestaw przygotowania biblioteki) jest ustawiony domyślnie jako TruSight Whole Genome i nie można go zmienić.
5. Wybierz żądany zestaw TruSight Whole Genome indeksów z menu rozwijanego **Index Adapter Kit** (Zestaw adapterów indeksujących).  
Długość odczytu zostanie ustawiona domyślnie i nie będzie można jej zmienić. (Read (Odczyt) 1 i 2 wykorzystuje 151 cykli; Index (indeks) 1 i 2 wykorzystuje 10 cykli).
6. Wprowadzić identyfikator próbki Library (zalecany format to DX1234567-LIB), a następnie wybrać **Next** (Dalej).  
Jeśli na tym etapie nie określono identyfikatora próbki Library, planowany przebieg będzie musiał zostać wybrany przed załadowaniem materiałów eksploatacyjnych do sekwencjonowania. Jeśli na tym etapie zostanie wprowadzony nieprawidłowy identyfikator próbki Library, planowany przebieg musi zostać skorygowany przed załadowaniem materiałów eksploatacyjnych. W przypadku konieczności skorygowania przebiegu przejść do punktu protokołu [Zmiana przebiegu na stronie 5](#), gdy jest on gotowy do załadowania materiałów eksploatacyjnych.
7. Na ekranach Sample Data (Dane próbki) i Sample Settings (Ustawienia próbki) wprowadzone zostaną informacje o próbce. Dane próbki można wprowadzić ręcznie lub importując plik próbki. Identyfikator próbki musi być unikalny dla każdej próbki i może zawierać tylko znaki alfanumeryczne, podkreślenia i myślniki. Nie używać spacji. Pozycja dołka odnosi się do dołka w formacie A01 do H04 danej płytki. Informacje o sekwencji indeksowania zostaną wypełnione



automatycznie po wprowadzeniu położenia dołka płytki indeksującej. Płeć należy wprowadzić jako Male (płeć męska), Female (żeńską) lub Unknown (nieznana). Identyfikator płytki Library i identyfikator dołka Library (np. format A01) są polami wymaganymi.

- Aby ręcznie wprowadzić dane próbki, należy dodać wiersze (łącznie 6 dla S2 lub 16 dla komórki przepływu S4) i wprowadzić wymagane informacje w polach Identyfikator próbki i Pozycja studzienki. Informacje można również kopiować i wklejać z programu Excel. Wybrać opcję **Next** (Dalej). Na ekranie Sample Settings (Ustawienia próbki) wprowadzić Library Plate ID (Identyfikator płytki biblioteki), Library Well ID (Identyfikator studzienki biblioteki) i Sex (Płeć). Wybrać opcję **Next** (Dalej).
  - Aby zaimportować przykładowy plik danych, wybrać opcję **Import Samples** (Importuj próbki) i przesłać plik danych próbki. Informacje zostaną automatycznie wprowadzone do wierszy. Na tym ekranie można pobrać szablon (\*.csv). Wybrać opcję **Next** (Dalej). Na ekranie Sample Settings (Ustawienia próbki) informacje zostaną automatycznie wprowadzone do wierszy z zaimportowanego pliku danych próbki. Wybrać opcję **Next** (Dalej).
8. Na ekranie Analysis Settings (Ustawienia analizy) wprowadzić Batch Name (Nazwa partii) zarejestrowaną podczas planowania partii i przebiegu.
  9. [Opcjonalnie] Wybrać Flow Cell Type (typ komórki przepływu) S2 lub S4.
  10. Potwierdzić lub usunąć zaznaczenie pola wyboru, aby wygenerować skompresowane kwestionariusze FASTQ ORA, a następnie wybrać **Next** (Dalej).

**UWAGA** Domyślnie TruSight Whole Genome Analysis Application generuje skompresowane ORA FASTQ. Zmiana tego ustawienia spowoduje zwiększenie rozmiaru końcowego danych wyniku.

11. Na ekranie Run Review (Przegląd przebiegu) należy przejrzeć wprowadzone informacje. Jeśli nie są wymagane żadne zmiany, wybierz opcję **Save** (Zapisz). Jeśli potrzebne są zmiany, wybierz opcję **Back** (Wstecz), aby powrócić do odpowiedniego ekranu.



## PRZESTROGA

TruSight Whole Genome został zwalidowany dla 6 próbek podczas korzystania z komory przepływu NovaSeq 6000Dx S2 i 16 próbek podczas korzystania z komory przepływu NovaSeq 6000Dx S4. Upewnij się, że wprowadzono prawidłową liczbę próbek dla wybranej konfiguracji komórki przepływu.

## Zmiana przebiegu

Jeśli zmiany są wymagane po utworzeniu przebiegu i przed załadowaniem materiałów eksploatacyjnych do sekwencjonowania, należy sprawdzić przebiegi w trybie IVD albo w aparacie, albo uzyskując dostęp Illumina Run Manager (IRM) za pomocą przeglądarki na komputerze sieciowym.

1. Wybrać **Runs** (Przebiegi).

2. Wybrać Run name (nazwa przebiegu) na karcie Planned Runs (Planowane przebiegi).
3. Wybierz **Edit** (Edytuj).
4. W razie potrzeby należy zaktualizować informacje o przebiegu lub próbce. Na przykład wprowadzić lub skorygować Library Tube ID (identyfikator próbki biblioteki) aby pasował do tego, który został użyty podczas wykonywania procedury.
5. Wybrać **Dalej** (Dalej), aby przejść do Run Review (Przegląd przebiegu).
6. Wybrać **Save** (Zapisz).
7. Wybrać opcję **Exit** (Wyjdź).

Wrócić do sekwencjonowania w trybie IVD, aby ponownie załadować materiały eksploatacyjne. Przebieg powinien teraz zostać automatycznie podświetlony.

W przypadku aktualizacji Library Tube ID (identyfikator próbki biblioteki) podczas ładowania materiałów eksploatacyjnych wrócić do opcji Run Selection (Wybór przebiegu) w oprogramowaniu sterującym i wybrać opcję **Refresh** (Odśwież) dla powiązanej kolumny A lub B. Przebieg powinien teraz zostać automatycznie podświetlony. Jeśli nie, należy wybrać opcję **Back** (Wstecz), aby powtórzyć Load consumables (Ładowanie materiałów eksploatacyjnych)

## Ponowne umieszczanie analizy w kolejce

Patrz rozdział Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132), aby określić, który typ analizy kolejki jest najbardziej odpowiedni.

### Ponownie umieść analizę w kolejce bez zmian

1. Wybrać nazwę Completed Run (Ukończony przebieg), aby wyświetlić Run Details (Szczegóły przebiegu).
2. Wybrać **Requeue Analysis** (Ponowne umieszczanie analizy w kolejce).
3. Wybrać **Requeue Analysis with no changes** (Ponownie umieść analizę w kolejce bez zmian).
4. Podać szczegóły w polu Reanalysis Reason (Powód ponownej analizy).
5. Wybrać **Requeue Analysis** (Ponowne umieszczanie analizy w kolejce).
6. Wyjdź ze strony i przejdź do strony Active Runs (Aktywne przebiegi), aby potwierdzić, że kolejka jest w toku.

### Ponownie umieść analizę w kolejce ze zmianami

1. Wybrać nazwę Completed Run (Ukończony przebieg), aby wyświetlić Run Details (Szczegóły przebiegu).
2. Wybrać **Requeue Analysis** (Ponowne umieszczanie analizy w kolejce).
3. Wybrać **Edit run settings** (Edytuj ustawienia przebiegu) i **Requeue Analysis** (Ponownie umieść analizę w kolejce).

4. Podać szczegóły w polu Reanalysis Reason (Powód ponownej analizy).
5. Wybrać **Requeue Analysis** (Ponowne umieszczanie analizy w kolejce).
6. Potwierdzić lub zaktualizować Run Settings (Ustawienia przebiegu), a następnie wybierz **Next** (Dalej).
7. W razie potrzeby skorygować informacje o próbce, ręcznie aktualizując pola lub wybierz opcję **Download Template** (Pobierz szablon), aby utworzyć plik `sampledata.csv` z bieżącymi informacjami. Poprawić informacje i usunąć istniejące wiersze na karcie Sample Data (Dane próbki) przed użyciem funkcji (Import Sample) (Importuj próbkę) w celu wypełnienia poprawionych danych próbki.
8. Na ekranie Run Review (Przegląd przebiegu) przejrzeć informacje o serii i wybrać opcję **Save** (Zapisz), aby rozpocząć ponowną analizę.
9. Wybrać **Exit** (Wyjdź) i przejść do strony Active Runs (Aktywne przebiegi), aby potwierdzić, że ponowne kolejkowanie jest w toku.  
Oryginalny folder danych przebiegu musi znajdować się w zewnętrznej lokalizacji przechowywania określonej w Run Details (Szczegóły przebiegu), aby można było przeprowadzić ponowną analizę. Jeśli ponowna analiza zakończy się niepowodzeniem, należy upewnić się, że przebieg nie został przeniesiony ani usunięty.

## Przegląd przebiegu i wyników

1. Na ekranie głównym Illumina Run Manager w trybie IVD wybierz opcję **Runs** (Przebiegi).
2. Na karcie Completed Runs (Ukończone przebiegi) wybrać Run name (Nazwę przebiegu).  
Na tej karcie zostaną również wyświetlone przebiegi, które zostały ukończone z powodu niepowodzenia sekwencjonowania, przesyłania danych lub analizy. Aktywne przebiegi i ich status są wyświetlane na karcie Active Runs (Aktywne przebiegi). Więcej informacji można znaleźć w Dokumentacja produktu NovaSeq 6000Dx (nr dokumentu 200010105).
3. Wybrać nazwę przebiegu na karcie Completed Runs (Ukończone przebiegi), aby wyświetlić szczegóły i wyniki przebiegu dla ścieżki do Analysis Output Folder (Folder wyników analizy).  
W przypadku nieudanych przebiegów należy sprawdzić status każdego kroku, a następnie zapoznać się z rozdziałem Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132).
4. Przejść do folderu analizy na dysku lokalnym i otworzyć skonsolidowany raport, aby przejrzeć wynik PASS/FAIL (POWODZENIE/NIEPOWODZENIE) dla każdego kroku kontroli jakości w następujący sposób:
  - Informacje na temat kontroli jakości przebiegu sekwencjonowania znajdują się w punkcie Podsumowanie wyniku kontroli jakości sekwencjonowania
  - Informacje na temat kontroli jakości FASTQ dla każdej próbki w przebiegu można znaleźć w Podsumowaniu wyniku kontroli jakości FASTQ

- Informacje na temat kontroli jakości biblioteki dla każdej próbki w przebiegu można znaleźć w dokumencie Podsumowanie wyniku kontroli jakości biblioteki próbek

W przypadku zaobserwowania wyniku FAIL (NIEPOWODZENIE) należy zanotować krok kontroli jakości i zapoznać się z rozdziałem Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132).

## Podsumowanie pliku wyników

TruSight Whole Genome Analysis Application zapisuje następujące główne pliki wyników. Lokalizacja głównych plików wyników znajduje się w poniższych częściach zawierających informacje o plikach.

Testy i próbki, które nie spełniają kryteriów ważności, nie generują plików CRAM, ROH bed ani plików \*genome.vcf.

Plik wyniku	Opis
Raport skonsolidowany (*.csv)	Zawiera parametry jakości używane do sprawdzania ważności przebiegu (w tym całkowitą wydajność i Q30), parametry ważności próbki (w tym wydajności FASTQ), parametry kontroli jakości biblioteki oraz parametry tylko dla informacji (FIO) dla wszystkich próbek w przebiegu.
Przykładowy raport (*.csv)	Zawiera wyniki kontroli jakości sekwencjonowania, FASTQ i kontroli jakości biblioteki próbek. Raport zawiera również zgodność ploidalności i parametry FIO dla pojedynczej próbki, jak również dla powiązanego przebiegu sekwencjonowania.
Mały wariant i mSNV VCF (*.annotated.hard-filtered-gvcf.gz)	Zawiera informacje o rozpoznaniu wariantu dla małych wariantów (SNV, indel) i mitochondrialnych SNV.
CNV VCF (*.cnv.vcf.gz)	Zawiera informacje o rozpoznaniu wariantu dla uzysku i utraty liczby kopii.
Powtórzenia VCF (*.repeats.vcf.gz)	Zawiera informacje o rozpoznaniu wariantów dla rozszerzeń STR i SMN1.
ROH BED (*.roh.bed)	Zawiera informacje o regionach homozygotyczności.
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.
Dopasowywanie CRAM (*.cram)	Zawiera dopasowane odczyty dla danej próbki.

## Informacje na raporcie dotyczącym kontroli jakości

Raport skonsolidowany <<RunID>>\_Consolidated\_Report.csv znajduje się w katalogu TruSightWholeGenomeAnalysis\_x.x.x\_run-complete i zawiera informacje o parametrach jakości

używanych do przekazywania lub odrzucania próbek na różnych etapach analizy. Poszczególne przykładowe raporty <<Sample\_ID>>\_Sample\_Report.csv można znaleźć w folderach <Sample\_ID> w katalogu TruSightWholeGenomeAnalysis\_x.x.x\_run-complete.

Nagłówki raportu zawierają następujące informacje o przebiegu: wersja aplikacji, nazwa partii, identyfikator próbki puli biblioteki, nazwa przebiegu sekwencjonowania, identyfikator przebiegu sekwencjonowania i typ komórki przepływu. Poniższe tabele opisują informacje zawarte w skonsolidowanym raporcie. Indywidualny przykładowy raport zawiera te same informacje z wyjątkiem parametrów demultipleksu.

Tabela 1 Parametry kontroli jakości sekwencjonowania

Parametry	Specyfikacja	Opis
Nieindeksowany uzysk całkowity (GB)	Nd.	Brak specyfikacji, ponieważ przebiegi o niższym uzysku mogą spowodować przejście bibliotek próbek. Należy spodziewać się $\geq 3000$ Gpz dla S4 i $\geq 1000$ GB dla komory przepływowej S2.
Łącznie % $\geq$ Q30	$\geq 85$	Pomiar podstawowej jakości na poziomie przebiegu. Ustawiono minimalną specyfikację, ponieważ przebiegi o zbyt niskiej wartości %Q30 nie spełnią wymagań nukleotydów Q30 w ramach kontroli jakości biblioteki próbek.
Podsumowanie wyniku kontroli jakości sekwencjonowania	POWODZENIE lub NIEPOWODZENIE	W przypadku niepowodzenia kontroli jakości sekwencjonowania należy zapoznać się z sekcją Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132).

Tabela 2 Parametry demultipleksu

Parametry	Specyfikacja	Opis
Zidentyfikowano odczyty procentowe	Nd.	Całkowita frakcja odczytów przechodzących przez filtr podczas przebiegu, które zostały przypisane do próbek podczas demultipleksacji.
Odsetek CV	Nd.	Zapewnia pomiar równomierności odczytów demultipleksowanych dla każdej pary indeksów podczas przebiegu. Należy spodziewać się $< 25\%$ dla przebiegów bez niepowodzeń wyniku kontroli jakości FASTQ.

Tabela 3 Parametry kontroli jakości FASTQ

Parametry	Specyfikacja	Opis
Wydajność na próbkę (pz)	$\geq 90\ 000\ 000\ 000$	Minimalne ustawienie jest równoważne ~26-krotnemu średniemu pokryciu autosomalnemu w celu sklasyfikowania bibliotek próbek, które nie spełniają wymagań kontroli jakości w celu skrócenia czasu analizy.
Podsumowanie wyniku kontroli jakości FASTQ	POWODZENIE lub NIEPOWODZENIE	W przypadku niepowodzenia kontroli jakości FASTQ należy zapoznać się z sekcją Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132).

Tabela 4 Parametry kontroli jakości próbki

Parametry	Specyfikacja	Opis
Średnie pokrycie autosomalne	$\geq 35$	Średnie pokrycie autosomów. Ustalono minimalną specyfikację, aby zapewnić wydajność analityczną.
Odsetek autosomów z pokryciem przekraczającym 20x	$\geq 93,94$	Pomiar jednolitości pokrycia, który wykrywa problemy niezwiązane z obciążeniem GC. Ustalono minimalną specyfikację, aby zapewnić wydajność analityczną.
Znormalizowane pokrycie dla kategorii GC od 60% do 79%	$0,82 \leq x \leq 1,13$	Pomiar jednolitości pokrycia, który rozpoznaje obciążenia GC, w szczególności utratę pokrycia w obszarach genomu o zawartości nukleotydów z większą zawartością % GC i niższą zawartością % AT. Minimalne i maksymalne specyfikacje są ustawiane w celu zapewnienia wydajności analitycznej.
Znormalizowane pokrycie kategorii GC od 20% do 39%	$0,97 \leq x \leq 1,06$	Pomiar jednolitości pokrycia, który rozpoznaje obciążenia GC, w szczególności utratę pokrycia w obszarach genomu o zawartości nukleotydów z niższą zawartością % GC i wyższą zawartością % AT. Minimalne i maksymalne specyfikacje są ustawiane w celu zapewnienia wydajności analitycznej.

Parametry	Specyfikacja	Opis
Średnie pokrycie mitochondrialne	≥ 500	Pokrycie chromosomu mitochondrialnego. Minimalna specyfikacja jest ustawiona w celu zapewnienia granicy wykrywalności mitochondrialnego SNV.
Odsetek nukleotydów Q30	≥ 85	Pomiar jakości nukleotydów. Ustalono minimalną specyfikację, aby zapewnić wydajność analityczną.
Szacowane zanieczyszczenie próbek	≤ 0,005	Wykrywa odczyty stanowiące zanieczyszczenie z innych próbek. Maksymalna specyfikacja jest ustawiona w celu zapewnienia granicy wykrywalności mitochondrialnej SNV (typ wariantu o najwyższej czułości na zanieczyszczenie).
Podsumowanie wyników kontroli jakości bibliotek próbek	POWODZENIE lub NIEPOWODZENIE	W przypadku niepowodzenia kontroli jakości biblioteki próbek należy zapoznać się z sekcją Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132).

Tabela 5 Parametry kontroli jakości ploidalności

Parametry	Specyfikacja	Opis
Podano ploidalność chromosomu płci	Nd.	Płeć podana przez operatora podczas Run Creation (Tworzenie przebiegu) (Female, Male, Unknown) (kobieta, mężczyzna, nieznana).
Oszacowanie ploidalności	Nd.	Ploidalność płciowa oszacowana przez DRAGEN.
Podsumowanie wyniku ploidalności	ZGODNY, NIEZGODNY lub ND	CONCORDANT (ZGODNY) wskazuje na zgodność pomiędzy podaną a szacowaną ploidalnością płci. ND oznacza płeć podaną jako Nieznana lub inną niż XX lub XY. W przypadku wyników DISCONCORDANT (NIEZGODNY) podczas tworzenia przebiegu wprowadzono nieprawidłową płeć lub mogła nastąpić zamiana próbki. Należy zapoznać się z rozdziałem Rozwiązywanie problemów w Ulotce dołączonej do opakowania TruSight Whole Genome (dokument nr 200050132).



Tabela 6 Parametry tylko dla informacji

Parametry	Opis
Mediana długości insercji	Wartość docelowa wynosi 450 pz, ale oczekuje się w zależności od przebiegu sekwencjonowania i operatora. Dopuszczalny jest zakres od około 360 do 550 pz. Konsekwentne działanie poza tym zakresem może prowadzić do częstszych przypadków niepowodzenia analizy próbki.
Odsetek mapowanych odczytów	Odsetek odczytów, które mapują genom referencyjny. Może zostać zmniejszona w odpowiedzi na skażenie gDNA inne niż ludzkie, słabą jakość próbki lub zbyt małą długość insertu, co prowadzi do problemów z mapowaniem.
Odsetek odczytów z uzupełniającymi dopasowaniami	Odsetek odczytów z mapowaniem, które dzieli się na różne lokalizacje w genomie referencyjnym.
Odsetek odczytów oznaczonych jako duplikat	Należy spodziewać się < 20%. Wartość może być podwyższona, jeśli wydajność przygotowania biblioteki jest niska lub objętość puli jest mniejsza od wymaganej, lub w odpowiedzi na problemy związane z sekwencjonowaniem.
Odsetek łagodnie przyciętych sygnałów nukleotydów – Odczyt 1	Przydatne w diagnozowaniu pierwotnej przyczyny niepowodzenia średniego pokrycia autosomalnego.
Odsetek łagodnie przyciętych sygnałów nukleotydów – Odczyt 2	Przydatne w diagnozowaniu pierwotnej przyczyny niepowodzenia średniego pokrycia autosomalnego.
Odsetek przyciętych nukleotydów – Odczyt 1	Przydatne w diagnozowaniu pierwotnej przyczyny niepowodzenia średniego pokrycia autosomalnego.
Odsetek przyciętych nukleotydów – Odczyt 2	Przydatne w diagnozowaniu pierwotnej przyczyny niepowodzenia średniego pokrycia autosomalnego.

## Informacje o pliku rozpoznań wariantów

### Pliki VCF

Pliki formatu rozpoznania wariantu (\*.vcf) zawierają informacje o wariantach znalezionych w określonych pozycjach w genomie referencyjnym i można je znaleźć w katalogu <Sample\_ID>/Analysis.

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. Ostatni wiersz nagłówka zawiera nagłówki kolumny dla wierszy danych. Każdy wiersz danych pliku VCF zawiera informację o pojedynczym wariantcie pozycji referencyjnej.

Wszystkie pliki VCF zawierają nagłówek z opisami kolumn danych wyjściowych i danych rozpoznań wariantów w kolumnach oznaczonych jako CHROM (chromosom), POS (Pozycja), ID, REF (Referencja), ALT (Alternatywa), QUAL (Jakość), FILTER (Filtr), INFO (Informacje), FORMAT, SAMPLE (Próbka). Definicje wartości kolumn mogą się różnić w zależności od rozpoznań wariantów.

### Małe warianty i mSNV VCF

Wynik jest zapisywany pod plikiem <Sample\_ID>.annotated.hard-filtered.gvcf.gz w katalogu <Sample\_ID>/Analysis.

Plik genomowy VCF (gVCF) zawiera informacje o wariantach i pozycjach uznanych za homozygotyczne dla genomu referencyjnego. W przypadku regionów homozygotycznych plik gVCF zawiera statystyki wskazujące, jak dobrze odczyty potwierdzają brak wariantów lub alternatywnych alleli. Plik gVCF zawiera sztuczny allel <NON\_REF>. Odczyty, które nie obsługują referencji lub jakichkolwiek wariantów mają przypisany allel <NON\_REF>. DRAGEN wykorzystuje te odczyty do określenia, czy można rozpoznać pozycję jako odniesienie homozygotyczne, w przeciwieństwie do pozostawiania nierozpoznanej. Wynik reprezentuje poziom pewności w skali Phreda w odniesieniu do homozygotycznego rozpoznania referencyjnego. W trybie linii zarodkowej wynik to FORMAT/GQ.

DRAGEN zapewnia filtrowanie wariantów po VCF na podstawie adnotacji obecnych w rekordach VCF. Filtrowanie wariantów opisano poniżej. Jednak ze względu na charakter algorytmów DRAGEN, które obejmują hipotezę skorelowanych błędów z rdzenia algorytmu rozpoznania wariantów, procedura poprawiła możliwości odróżniania rzeczywistych wariantów od szumu, a zatem znacznie zmniejszyła się zależność od filtrowania po VCF.

TruSight Whole Genome Analysis Application zapewnia adnotację o wyniku ufności i poziomie pewności dla małych wariantów, które można wykorzystać do dalszej poprawy wydajności. Adnotacja dotycząca poziomu zaufania nie jest filtrem jakości i nie jest bezpośrednio odzwierciedlona w statusie jakości rozpoznań wariantów. W związku z tym możliwe jest zobaczenie rozpoznań wariantów przechodzących przez filtr, które mimo wszystko są oznaczone jako obarczone niskim poziomem ufności.

Tabela 7 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 wariacie 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)	Zawsze .
REF (Referencja)	Genotyp referencyjny. Przykładowo delekcja 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.
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 8 Adnotacje w pliku VCF

Nagłówek	Opis
FILTER (Filtr)	<ul style="list-style-type: none"> <li>• PASS (Powodzenie) Powodzenie w przypadku wszystkich filtrów.</li> <li>• <b>DRAGENSnpHardQUAL</b> – Ustaw, jeśli prawda:QUAL &lt; 10,41</li> <li>• <b>DRAGENIndelHardQUAL</b> – Ustaw, jeśli prawda:QUAL &lt; 7,83</li> <li>• <b>LowDepth</b> – Ustaw, jeśli prawda:DP &lt;= 1</li> <li>• <b>LowGQ</b> – Ustaw, jeśli prawda:GQ = 0</li> <li>• <b>PloidyConflict</b> – rozpoznanie genotypu z algorytmu rozpoznawania wariantów nie jest zgodne z ploidią chromosomu.</li> <li>• <b>base_quality</b> – miejsce filtrowane, ponieważ mediana jakości nukleotydów odczytów alternatywnych w tym locus nie spełnia progu</li> <li>• <b>filtered_reads</b> – miejsce filtrowane, ponieważ odfiltrowano zbyt dużą część odczytów</li> <li>• <b>fragment_length</b> – miejsce filtrowane, ponieważ różnica bezwzględna między medianą długości fragmentu odczytów alternatywnych a medianą długości fragmentów odczytów referencyjnych w tym locus przekracza próg</li> <li>• <b>low_af</b> – częstość alleli nie spełnia progu</li> <li>• <b>low_depth</b> – miejsce filtrowane, ponieważ głębokość odczytu jest zbyt mała</li> <li>• <b>low_frac_info_reads</b> – miejsce filtrowane, ponieważ frakcja odczytów informatywnych jest poniżej progu</li> <li>• <b>low_normal_depth</b> – miejsce filtrowane, ponieważ normalna głębokość odczytu próbek jest zbyt mała</li> <li>• <b>long_indel</b> – miejsce filtrowane, ponieważ długość indela jest zbyt duża</li> <li>• <b>mapping_quality</b> – miejsce filtrowane, ponieważ mediana jakości odwzorowania odczytów alternatywnych w tym locus nie spełnia progu</li> <li>• <b>multiallelic</b> – miejsce filtrowane, ponieważ więcej niż dwa allele alternatywne przekraczają LOD nowotworu.</li> <li>• <b>non_homref_normal</b> – miejsce filtrowane, ponieważ genotyp próbki normalnej nie jest odniesieniem homozygotycznym</li> <li>• <b>no_reliable_supporting_read</b> – miejsce filtrowane, ponieważ nie ma wiarygodnego potwierdzającego odczytu somatycznego</li> <li>• <b>panel_of_normals</b> – obserwowany w co najmniej jednej próbce w panelu normalnych vcf</li> <li>• <b>read_position</b> – miejsce filtrowane, ponieważ mediana odległości między początkiem/końcem odczytu i tym locus jest poniżej progu</li> <li>• <b>RMxNRepeatRegion</b> – miejsce filtrowane, ponieważ całość lub część allelu tego wariantu jest powtórzeniem tego odniesienia</li> <li>• <b>strand_artifact</b> – miejsce filtrowane z powodu poważnego obciążenia systematycznego nici</li> <li>• <b>str_contraction</b> – miejsce filtrowane z powodu podejrzanego błędu PCR, gdzie alternatywny allel jest o jedną jednostkę powtórzenia mniejszy niż odniesienie</li> <li>• <b>too_few_supporting_reads</b> – miejsce filtrowane, ponieważ w próbce guza jest zbyt mało pomocniczych odczytów</li> <li>• <b>weak_evidence</b> – wynik wariantu somatycznego nie spełnia progu</li> </ul>

Nagłówek	Opis
INFO (Informacje)	<ul style="list-style-type: none"> <li>• <b>DB</b> – udział w dBSNP</li> <li>• <b>FS</b> – wartość P skalowana według Phred przy użyciu dokładnego testu Fishera w celu wykrycia obciążenia systematycznego nici</li> <li>• <b>QD</b> – pewność/jakość wariantu według głębokości</li> <li>• <b>R2_5P_bias</b> – wynik oparty na obciążeniu mate i odległości od 5 głównych końców</li> <li>• <b>SOR</b> – symetryczny iloraz szans w tabeli kontyngencji 2x2 w celu wykrycia obciążenia systematycznego nici</li> <li>• <b>DP</b> – przybliżona głębokość odczytu (informatywna i nieinformatywna); niektóre odczyty mogły zostać odfiltrowane na podstawie mapq itp.</li> <li>• <b>END</b> – pozycja zatrzymania interwału</li> <li>• <b>FractionInformativeReads</b> – udział odczytów informatywnych w całkowitej liczbie odczytów</li> <li>• <b>MQ</b> – jakość mapowania RMS</li> <li>• <b>MQRankSum</b> – z-score w teście sumy rang Wilcoxon dla jakości mapowania odczytów alternatywnych w porównaniu z referencyjnymi.</li> <li>• <b>ReadPosRankSum</b> – z-score w teście sumy rang Wilcoxon dla obciążenia pozycji alternatywnych w porównaniu z referencyjnymi</li> <li>• <b>SOMATIC</b> – co najmniej jeden wariant w tej pozycji jest somatyczny</li> <li>• <b>ILENS</b> – długości Indel dla każdego wariantu ALT.</li> <li>• <b>SCORE</b> – wynik dotyczący przedziału ufności dla każdego typu wariantu obecnego w ośrodku jako (typ zmienny): (wynik dotyczący przedziału ufności).</li> <li>• <b>TIER</b> – poziom ufności dla każdego typu wariantu obecnego w ośrodku jako (typ zmienny): (poziom ufności).</li> </ul>

Nagłówek	Opis
FORMAT (Format)	<p>W kolumnie FORMAT wymienione są pola rozdzielone dwukropkami np. GT:GQ.</p> <ul style="list-style-type: none"> <li>• <b>AD</b> – głębokości alleli (zliczanie tylko informatywnych odczytów z całkowitej liczby odczytów) dla alleli referencyjnych i alternatywnych w podanej kolejności</li> <li>• <b>AF</b> – frakcje alleli dla alleli alternatywnych w podanej kolejności</li> <li>• <b>DP</b> – przybliżona głębokość odczytu (odczyty z wartością MQ = 255 lub ze złym dopasowaniem par są odfiltrowywane)</li> <li>• <b>F1R2</b> – liczba odczytów w orientacji pary F1R2 potwierdzających każdy allel</li> <li>• <b>F2R1</b> – liczba odczytów w orientacji pary F2R1 potwierdzających każdy allel</li> <li>• <b>GP</b> – prawdopodobieństwa a posteriori w skali Phred dla genotypów zgodnie ze specyfikacją VCF</li> <li>• <b>GQ</b> – jakość genotypu</li> <li>• <b>GT</b> – genotyp.</li> <li>• <b>ICNT</b> – liczby odczytów informacyjnych INDEL na podstawie referencyjnego modelu ufności.</li> <li>• <b>MB</b> – statystyki komponentów na próbkę w celu wykrycia obciążenia parowania.</li> <li>• <b>MIN_DP</b> – minimalny DP obserwowany w bloku GVCF.</li> <li>• <b>PL</b> – znormalizowane prawdopodobieństwo genotypów w skali Phreda zdefiniowane w specyfikacji VCF</li> <li>• <b>PRI</b> – prawdopodobieństwo a priori genotypów w skali Phread</li> <li>• <b>PS</b> – informacje o identyfikatorze fazowania fizycznego, gdzie każdy unikatowy identyfikator w danej próbce (ale nie w wielu próbkach) łączy rekordy w grupie fazowania</li> <li>• <b>SB</b> – statystyki komponentów na próbkę, które składają się na dokładny test Fishera w celu wykrycia obciążenia systematycznego nici.</li> <li>• <b>SPL</b> – Znormalizowane prawdopodobieństwo w skali Phred dla SNP oparte na referencyjnym modelu ufności.</li> <li>• <b>SQ</b> – jakość somatyczna</li> </ul>
SAMPLE (Próbka)	W kolumnie próbek są podane wartości wskazane w kolumnie FORMAT.

## Polimorfizm liczby kopii VCF

Etap zliczania sekwencji docelowych jest pierwszym etapem przetwarzania dla procedury CNV DRAGEN, dającym `<Sample_ID>.target.counts.gz`, a następnie wykonywana jest korekta obciążenia GC, generująca plik `*.target.counts.gc-corrected.gz`. Etap normalizacji generuje plik `*.tn.tsv.gz`. Oprogramowanie DRAGEN hosta generuje wiele plików pośrednich. `*.seg.called.merged` to plik rozpoznania końcowego zawierający zdarzenia amplifikacji i delecji.

Oprócz pliku segmentu DRAGEN emituje rozpoznania w standardowym formacie VCF. Wynik jest zapisywany w <Sample\_ID>.cnv.vcf.gz w katalogu <Sample\_ID>/Analysis.

Definicje kolumn specyficznych dla rozpoznania CNV:

Kolumna POS to pozycja początkowa wariantu. Zgodnie ze specyfikacją VCF, jeśli którykolwiek z alleli ALT jest allelem symbolicznym, takim jak <DEL>, wówczas wymagana jest podstawa wyściełająca, a POS oznacza współrzędną podstawy poprzedzającej polimorfizm. Wszystkie współrzędne w VCF są oparte na 1.

Kolumna ID służy do przedstawienia zdarzenia. Pole ID koduje typ zdarzenia i współrzędne zdarzenia.

Kolumna REF zawiera N dla wszystkich zdarzeń CNV.

Kolumna ALT określa typ zdarzenia CNV. Ponieważ VCF zawiera tylko zdarzenia CNV, używany jest tylko wpis DEL lub DUP.

Kolumna QUAL zawiera szacunkowy wynik jakości dla zdarzenia CNV, który jest używany w filtrowaniu twardym.

Kolumna FILTER (Filtr) zawiera PASS (Powodzenie), jeśli zdarzenie CNV odpowiada wymaganiom wszystkich filtrów. W przeciwnym razie kolumna zawiera nazwę filtra braku spełnienia wymagań.

Kolumna INFO zawiera informacje przedstawiające zdarzenie. Wpis REFLLEN wskazuje długość zdarzenia. Wpis SVTYPE zawsze jest CNV. Wpis END wskazuje położenie końcowe zdarzenia.

Pola FORMAT są opisane w nagłówku.

- GT – genotyp.
- SM – liniowy współczynnik kopiowania średniej segmentu
- CN – szacowany numer kopii
- BC – liczba miejsc składowania w regionie
- PE – liczba nieprawidłowo sparowanych odczytów końcowych na początku i końcu wartości granicznych

## Powtarza VCF

ExpansionHunter przeprowadza korekcję odczytów w oparciu o wykres sekwencji, które pochodzą z każdego powtórzenia celu i są wokół niego. ExpansionHunter następnie genotypuje długość powtórzenia w każdym allelu na podstawie tych wyrównań wykresu.

Więcej informacji i analiz można znaleźć w następujących dokumentach ExpansionHunter:

- Dolzhenko et al., *Detection of long repeat expansions from PCR-free whole-genome sequence data* 2017
- Dolzhenko et al., *ExpansionHunter: A sequence-graph based tool to analyze variation in short tandem repeat regions* 2019

Katalog wariantów STR TruSight Whole Genome Analysis Application zawiera specyfikacje dotyczące wielokrotnych prób wywołujących chorobę, zlokalizowanych w genach AFF2, AR, ATN1, ATXN1, ATXN10, ATXN2, ATXN3, ATXN7, ATXN8OS, C9ORF72, CACNA1A, CBL, CNBP, CSTB, DIP2B, DMPK, FMR1, FXN, GLS, HTT, JPH3, NIPA1, NOP56, NOTCH2NL, PABPN1, PHOX2B, PPP2R2B i TBP.

Wyniki powtórzonego genotypowania są generowane jako oddzielny plik VCF, który zapewnia długość każdego allelu przy każdym rozpoznanym powtórzeniu określonym w pliku katalogu specyfikacji powtórzenia. Nazwa pliku to <Sample\_ID>.repeats.gz i można go znaleźć w katalogu <Sample\_ID>/Analysis.

Niektóre kolumny są specyficzne dla algorytmu powtarzania rozpoznania:

Tabela 9 Podstawowe pola VCF

Pole	Opis
CHROM (Chromosom)	Identyfikator chromosomu
POS (Pozycja)	Położenie pierwszego nukleotydu przed regionem powtórzenia w referencyjnym przykładzie
ID (Identyfikator)	Zawsze .
REF (Referencja)	Nukleotyd referencyjny w pozycji POS
ALT (Alternatywa)	Lista powtórzonych alleli w formacie <STR <sub>n</sub> >. N to liczba powtarzających się jednostek.
QUAL (Jakość)	Zawsze .
FILTER (Filtr)	Filtr LowDepth jest stosowany, gdy całkowita głębokość locus jest poniżej 10x lub liczba odczytów obejmujących jedną lub obie końcówki jest poniżej 5.

Tabela 10 Dodatkowe pola INFO

Pole	Opis
END (KONIEC)	Położenie ostatniego nukleotydu regionu powtórzenia w odniesieniu
REF (Referencja)	Numer kopii referencyjnej
REPID	Identyfikator wariantu z katalogu wariantów
RL	Długość odniesienia w pz
RU	Jednostka powtórzenia w orientacji referencyjnej
VARID	Identyfikator wariantu z katalogu wariantów



Tabela 11 Pola GENOTYPE (na próbkę)

Pole	Opis
AD	Głębokości alleli ref i alt w podanej kolejności
ADFL	Liczba odczytów otaczających zgodnych z allelem
ADIR	Liczba odczytów w powtórzeniu się zgodnie z allelem
ADSP	Liczba odczytów zbiorczych zgodnych z allelem
DST	Wyniki (+ wykryte, - niewykryte, ? nieokreślone) testu przedstawiane według wariantu
GT	Genotyp
LC	Pokrycie locus
REPCI	Przedział ufności dla REPCN
REPCN	Liczba jednostek powtórzeń w allelu
RPL	Dawka logarytmiczna dla obecności allelu referencyjnego
SO	Typy odczytów, które obsługują allel. Wartości mogą być SPANNING, FLANKING, lub INREPEAT. Te wartości wskazują, czy odczyty są rozpiętością, bokiem, czy też są w pełni zawarte w powtórzeniu.

Plik <Sample\_ID>.repeats.vcf.gz zawiera wynik SMN wraz z dowolnymi ukierunkowanymi powtórzeniami. Wynik SMN jest reprezentowany jako pojedyncze rozpoznanie SNV w pozycji wpływającej na rozłączenie w SMN1 (NM\_000344.3:c.840C/T) ze statusem Spinal Muscular Atrophy (SMA) (Rdzeniowy zanik mięśni) w następujących polach niestandardowych.

Tabela 12 Wyniki SMA w pliku Output (wynik) repeats.vcf

Pole	Opis
VARID	SMN oznacza rozpoznanie SMN.
GT	Rozpoznanie genotypu w tej pozycji przy użyciu normalnego (diploidalnego) modelu genotypu.
DST	Rozpoznanie dotyczące statusu SMA: + oznacza wykryte - oznacza niewykryte ? wskazuje nieokreślony
AD	Łączna liczba odczytów obsługujących allel C i T.
RPL	Stosunek prawdopodobieństwa Log10 między modelami, których nie dotyczy problem, a modelami, których dotyczy problem. Pozytywne wyniki wskazują, że model, którego nie dotyczy problem, jest bardziej prawdopodobny.

## ROH BED

Obszary homozygotyczności (ROH) są wykrywane jako rozpoznanie części małego wariantu. Rozpoznanie wykrywa i wysyła przebieg homozygotyczny z rozpoznaniem całego genomu w autosomalnych chromosomach ludzkich. Chromosomy płci są ignorowane, chyba że przykładowy kariotyp płci to XX, zgodnie z oceną Ploidy Estimator. Wyjście ROH umożliwia dalszym narzędziom monitorowanie i przewidywanie zgodności pomiędzy rodzicami uczestnika probandowanego.

Region jest definiowany jako kolejne rozpoznania wariantów w chromosomie bez dużej luki między tymi wariantami. Innymi słowy, regiony są podzielone według chromosomu lub dużych luk bez rozpoznania SNV. Rozmiar luki jest ustawiony na 3 Mbp.

Rozpoznanie ROH tworzy plik wyjściowy ROH o nazwie `<Sample_ID>.roh.bed` w katalogu `<Sample_ID>/Analysis`. Każdy wiersz reprezentuje jeden obszar homozygotyczności. Plik bed zawiera następujące kolumny:

```
Chromosom Start End Score #Homozygous #Heterozygous
```

Gdzie

- Wynik jest funkcją liczby wariantów homozygotycznych i heterozygotycznych, gdzie każdy wariant homozygotyczny zwiększa wynik o 0,025, a każdy wariant heterozygotyczny zmniejsza wynik o 0,975.
- Pozycje początkowe i końcowe to przedziały oparte na 0, półotwarte.
- Liczba wariantów homozygotycznych to liczba wariantów homozygotycznych w regionie.
- Liczba wariantów heterozygotycznych to liczba heterozygotycznych wariantów w regionie.

Algorytm wykrywania rozpoznania tworzy również plik parametrów o nazwie `<Sample_ID>.roh_metrics.csv`, który zawiera liczbę dużych ROH i procent SNP w dużych ROH (> 3 MB).

## Parametry szacunkowe ploidalności

Serie estymatora ploidalności wykonywane domyślnie. Estymator ploidalności wykorzystuje odczyty z mapera/alignera do obliczenia głębokości sekwencjonowania pokrycia dla każdego autosomu i allosomu w genomie ludzkim. Kariotyp płciowy próbki jest następnie szacowany przy użyciu proporcji mediany pokrycia chromosomu płci do mediany pokrycia autosomalnego. XX lub XY oraz ZGODNE, NIEZGODNE lub ND (nieokreślone) w porównaniu z dostarczonymi przykładowymi danymi są zgłaszane w skonsolidowanym raporcie. Szczegółowe wyniki, w tym każda znormalizowana zgodnie z kodowaniem mediana pokrycia, są zgłaszane w pliku `<Sample_ID>.ploidy_estimation_metrics.csv`.

## 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

Oprogramowanie generuje jeden plik FASTQ dla każdej próbki, odczytu i pasma. Na przykład dla każdej próbki w cyklu ze sparowaną końcową serią oprogramowanie generuje dwa pliki FASTQ: jeden dla odczytu 1 i jeden dla odczytu 2. Oprócz tych przykładowych plików FASTQ oprogramowanie generuje dwa pliki FASTQ na pasmo, zawierające wszystkie nieznane próbki. Pliki FASTQ dla odczytu indeksu 1 i odczytu indeksu 2 nie są generowane, ponieważ sekwencja jest uwzględniona w nagłówku każdego wpisu FASTQ. Format nazwy pliku jest utworzony z pól określonych w arkuszu próbek i wykorzystuje format nazw plików `<Sample_ID>_S#_L00#_R#_001.fastq.gz`

Pliki FASTQ są zapisywane w katalogu `<Sample_ID>/Conversion`. W katalogu FASTQ folderu analizy można znaleźć katalog dzienników z dziennikami konwersji BCL na FASTQ oraz katalog raportów zawierający różne odczytywane pliki parametrów i plik `SampleSheet.csv` używany do konwersji FASTQ. Pliki FASTQ z nieokreślonych odczytów znajdują się w katalogu `Nieokreślone/konwersja` folderu Analiza.

Identyfikator próbki jest sformatowany w następujący sposób:

```
@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=AAAAAAAAAAAA9#:<#<;<<<????#=#
```

## Pliki CRAM

Skompresowana mapa wyrównania zorientowana na odniesienie lub pliki CRAM (\*.cram) są przechowywane w katalogu `<Sample_ID>/Analysis` i zawierają nagłówki oraz rekordy dopasowywania względem genomowego pliku odniesienia używanego podczas wyrównania. Ścieżka do pliku referencyjnego jest domyślnie ustawiona na `--ht-reference hg38.fa` w pliku `<Sample_ID>/Analysis/<Sample_ID>-replay.json`.

Pliki CRAM 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 ampliconu służący do zapisu identyfikatora ampliconu powiązanego z odczytem.

Aby wyświetlić rekordy wyrównania, samtools jako `samtools view --reference <path_to_reference_folder>/hg38.fa <Sample_ID>.cram`.

Generowany jest również plik indeksu i plik sumy kontrolnej.

## Pomoc techniczna

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

**Witryna:** [www.illumina.com](http://www.illumina.com)

**Adres e-mail:** [techsupport@illumina.com](mailto:techsupport@illumina.com)

**Karty charakterystyki bezpieczeństwa (SDS)** – dostępne na stronie firmy Illumina pod adresem [support.illumina.com/sds.html](http://support.illumina.com/sds.html).

**Dokumentacja produktu** – jest dostępna do pobrania w witrynie [support.illumina.com](http://support.illumina.com).



Illumina, Inc.  
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  
The Netherlands

**Sponsor w Australii**

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

DO STOSOWANIA W DIAGNOSTYCE IN VITRO.

© 2024 Illumina, Inc. Wszelkie prawa zastrzeżone.

**illumina**<sup>®</sup>