illumina®

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

Uživatelská příručka k aplikaci

VLASTNICTVÍ SPOLEČNOSTI ILLUMINA
Dokument č. 200025238 v00
Únor 2023
URČENO K DIAGNOSTICE IN VITRO.

Historie revizí

Dokument	Datum	Popis změny
200025238 v00	Únor 2023	První vydání.

Obsah

Historie revizí iii
Přehled 1
Metody analýzy 1
Vytvoření plánovaného běhu 5
Nastavení 7
Soubor manifestu 8
Filtrování šumu (volitelné) 9
Výstupy analýzy 9
Soubory FASTQ 10
Soubory BAM 11
Soubory VCF 11
Opětovné zařazení analýzy 18
Technická pomoc 19

Přehled

Aplikace DRAGEN for Illumina DNA Prep with Enrichment Dx (DRAGEN for IDPE Dx) se používá k plánování a provádění sekundární analýzy knihoven IDPE Dx generovaných pro sekvenování na NextSeq 550Dx.

DRAGEN for IDPE Dx podporuje sekvenování do analýzy při použití s přípravou Illumina DNA Prep with Enrichment Dx knihovny, NextSeq 550Dx a Server Illumina DRAGEN pro NextSeq 550Dx.

Metody analýzy

DRAGEN for IDPE Dx provádí v závislosti na vybraném pracovním postupu demultiplexování, generování FASTQ, mapování čtení, zarovnání s referenčním genomem a přiřazení malých variant:

- Generování FASTQ
- Generování Germline FASTQ a VCF
- Generování Somatic FASTQ a VCF

POZNÁMKA Komprese ORA je k dispozici pro všechny tři pracovní postupy. DRAGEN ORA Compression je plně bezztrátový kompresní software, který vytváří soubor s příponou Original Read Archive (*.ora). Formát ora je referenční kompresní formát pro soubory FASTQ a je navržen pro velmi rychlou kompresi/dekompresi a vysoký kompresní poměr.

Generování FASTQ

Sestavené sekvence jsou zapsány do souborů FASTQ za každý vzorek. Soubory FASTQ jsou textové soubory, které obsahují data sekvenování a skóre kvality pouze pro jeden vzorek. Pro každý vzorek jsou generovány samostatné soubory FASTQ podle dráhy průtokové kyvety na každé čtení sekvenování. Název vzorku, jak je specifikován během nastavení běhu, je zahrnut v názvu souboru FASTQ. Soubory FASTQ jsou primárním vstupem pro zarovnání. Prvním krokem generování FASTQ je demultiplexování. Demultiplexování přiřazuje klastry, které projdou filtrem, k určitému vzorku tím, že porovná každou sekvenci čtení indexů s indexovými sekvencemi specifikovanými pro daný běh. V tomto kroku nejsou zvažovány žádné hodnoty kvality. Čtení indexu jsou identifikována pomocí následujících kroků:

- Vzorky jsou číslovány od 1 podle pořadí, v jakém jsou uvedeny pro daný běh.
- Vzorek číslo 0 je vyhrazen pro klastry, které nebyly přiřazeny ke vzorku.
- Klastry jsou přiřazeny ke vzorku, když se indexová sekvence přesně shoduje nebo když existuje nejvýše jedna neshoda na jednotlivé čtení indexu.

Software obsahuje kompresi ORA pro kompresi souborů FASTQ. Tento formát lze volitelně povolit. Při použití formátu ORA (*.ora) je kontrolní součet md5 obsahu FASTQ po cyklu komprese a dekomprese zachován, aby byla zajištěna bezeztrátová komprese.

Mapování a zarovnání DNA

Po generování FASTQ jsou čtení mapována a zarovnána s referenčním genomem. První fází mapování je generování seedů ze čtení a následné hledání přesných shod v referenčním genomu. Tyto výsledky jsou pak upřesněny úplným zarovnáním Smith-Waterman na místech s nejvyšší hustotou shod seedů. Tento dobře zdokumentovaný algoritmus funguje tak, že porovnává každou pozici čtení se všemi kandidátními pozicemi reference. Tato porovnání odpovídají matici potenciálních zarovnání mezi čtením a referencí. Pro každou z těchto kandidátních pozic zarovnání Smith-Waterman generuje skóre, která se používají k vyhodnocení, zda nejlepší zarovnání procházející touto buňkou matice k ní dospěje shodou nebo neshodou nukleotidů (diagonální pohyb), delecí (horizontální pohyb) nebo inzercí (vertikální pohyb). Shoda mezi čtením a referencí poskytuje bonus ve skóre a neshoda nebo indel vede k penalizaci. Vybráno je zarovnání, které prochází maticí s celkově nejvyšším počtem bodů. Algoritmus je hardware akcelerován na kartách FPGA (field-programmed gate array) DRAGEN. Referenční genom použitý v aplikaci je vytvořen z UCSC hg19 FASTA s možností DRAGEN pro vytvoření alt-aware hash tabulky založené na liftoveru.

Přiřazování germinálních variant v aplikaci DRAGEN

Detekční program DRAGEN Germline Small Variant Caller přebírá mapovanou a zarovnanou DNA jako vstup a přiřazuje jednonukleotidové polymorfismy (SNP) a inserce nebo delece (indely) pomocí kombinace detekce ve sloupcích a místního *de novo* sestavení haplotypů. Chcete-li povolit DRAGEN Germline Small Variant Caller, vyberte pracovní postup pro germinální varianty.

Přiřazování germinálních variant se obvykle používá pro germinální vzorky, u kterých je známý počet ploidie dva. Přiřaditelné referenční oblasti jsou nejprve identifikovány s dostatečným pokrytím zarovnání. Rychlé skenování seřazených čtení identifikuje v těchto referenčních oblastech aktivní oblasti, které jsou soustředěny do sloupců pileup s důkazy o variantě. Aktivní oblasti jsou doplněny dostatečným kontextem, aby pokryly významný nereferenční obsah v okolí. Pokud existují důkazy indelů, aktivní oblasti dostanou další doplnění (padding).

Zarovnaná čtení jsou v každé aktivní oblasti klipována a sestavena do De Bruijnova grafu. Okraje klipovaných čtení jsou váženy počtem pozorování, přičemž referenční sekvence je páteřní. Po vyčištění a zjednodušení grafu jsou všechny cesty source-to-sink extrahovány jako kandidátní haplotypy. Každý haplotyp je algoritmem Smith-Waterman zarovnán s referenčním genomem pro identifikaci variant, které představuje. Tato sada událostí může být rozšířena detekcí na základě pozice. Pro každý pár čtení-haplotyp se pravděpodobnost $P(r|H)$ pozorování čtení – za předpokladu, že haplotyp je skutečný počáteční vzorek – odhaduje pomocí párového skrytého Markovova modelu (HMM).

Při skenování podle referenční pozice nad aktivní oblastí jsou kandidátní genotypy vytvořeny z diploidních kombinací variantních událostí (SNP nebo indely). Pro každou příhodu (včetně reference) je podmíněná pravděpodobnost P(r|e) pozorování každého překrývajícího se čtení odhadnuta jako maximální P(r|H) pro haplotypy podporující příhodu. Ty jsou kombinovány do podmíněné pravděpodobnosti P(r|e1e2) pro genotyp (pár událostí) a znásobeny tak, aby byla získána podmíněná pravděpodobnost P(R|e1e2) pozorování celého pileup čtení. Pomocí Bayesova vzorce se vypočítá posteriorní pravděpodobnost P (e1e2|R) každého diploidního genotypu a vítěz se přiřadí.

DRAGEN for IDPE Dx používá automatické filtrování. Další informace naleznete v *Poznámky k VCF souboru pro germinální pracovní postup* na straně 13.

Přiřazení somatických variant v aplikaci DRAGEN

Detekční program DRAGEN Somatic Small Variant Caller přebírá mapovanou a zarovnanou DNA jako vstup a přiřazuje SNV a indely pomocí místního *de novo* sestavení haplotypů v aktivní oblasti. Pro povolení DRAGEN Somatic Small Variant Caller vyberte aplikaci somatických variant.

Pro vzorky nádoru se obvykle používá přiřazení somatických variant. S tímto pracovním postupem DRAGEN nevytváří žádné neopodstatněné ploidy, což umožňuje detekci nízkofrekvenčních alel. U lokusů s pokrytím až 100x ve vzorku nádoru má DRAGEN detekční práh při frekvenci variantních alel 5 %. Limit se stupňuje s rostoucí hloubkou na základě jednotlivých lokusů a snižuje se na polovinu pokaždé, když se pokrytí zdvojnásobí nad 100x. Přiřaditelné referenční oblasti jsou nejprve identifikovány s dostatečným pokrytím zarovnání. Skenování seřazených čtení identifikuje v těchto referenčních oblastech aktivní oblasti, které jsou ve čteních nádoru soustředěny do sloupců pileup s důkazy o variantě. Aktivní oblasti jsou doplněny dostatečným kontextem, aby pokryly významný nereferenční obsah v okolí. Pokud existují důkazy indelů, aktivní oblasti dostanou další doplnění (padding).

Zarovnaná čtení jsou v každé aktivní oblasti klipována a sestavena do De Bruijnova grafu. Okraje klipovaných čtení jsou váženy počtem pozorování, přičemž referenční sekvence je páteřní. Po vyčištění a zjednodušení grafu jsou všechny cesty source-to-sink extrahovány jako kandidátní haplotypy. Každý haplotyp je algoritmem Smith-Waterman zarovnán s referenčním genomem pro identifikaci variant, které představuje. Pro každý pár čtení-haplotyp se pravděpodobnost P(r|H) pozorování čtení odhaduje pomocí párového skrytého Markovova modelu (HMM) – za předpokladu, že haplotyp je skutečný počáteční vzorek.

Aby bylo možné stanovit skóre limitu detekce nádoru (TLOD), DRAGEN Somatic Small Variant Caller nejprve naskenuje referenční pozici pro každou kandidátní somatickou událost a také referenční událost nad aktivní oblastí. Podmíněná pravděpodobnost P(r|e) pozorování každého překrývajícího se čtení je odhadnuta jako maximální P(r|H) pro haplotypy podporující příhodu. Ty se zkombinují do podmíněné pravděpodobnosti P(r|E) pro hypotézu události E, která zahrnuje směs referenční a kandidátské somatické alely v rozsahu možných frekvencí alel, a vynásobí se, aby se získala podmíněná pravděpodobnost P(R|E) pozorování celého pileup čtení. Odtud se vypočítá skóre TLOD jako důkaz, že ve vzorku nádoru je v daném lokusu přítomna alela ALT.

DRAGEN for IDPE Dx používá automatické filtrování. Další informace naleznete v *Poznámky k VCF souboru VCF pro somatický pracovní postup* na straně 16.

Vytvoření plánovaného běhu

Pomocí následujících kroků můžete vytvořit běh v Illumina Run Manager na přístroji NextSeq 550Dx nebo pomocí prohlížeče na počítači připojeném k síti. Pokud chcete importovat data vzorků, použijte prohlížeč na počítači připojeném k síti. Pokyny k přístupu k Illumina Run Manager ze síťového počítače naleznete v Příručka k softwaru Illumina Run Manager pro NextSeq 550Dx (dokument č. 200025239).

Existují dva různé způsoby, jak vytvořit nový plánovaný běh:

- **Import Run** (Importovat běh) – použijte vzorový list z existujícího běhu jako šablonu pro nový běh. Informace o importu běhu naleznete v Příručka k softwaru Illumina Run Manager pro NextSeq 550Dx (dokument č. 200025239).
- **Create Run** (Vytvořit běh) – zadejte parametry běhu ručně. Následující pokyny vysvětlují, jak postupovat při vytváření běhu.

POZNÁMKA Požadovaná vstupní pole v uživatelském rozhraní jsou označena hvězdičkou (*).

Aplikace

1. Na kartě Planned (Plánované) na obrazovce Runs (Běhy) vyberte možnost **Create Run** (Vytvořit běh).
2. Vyberte aplikaci DRAGEN for Illumina DNA Prep with Enrichment Dx a potom vyberte **Next** (Další).

Nastavení běhu

1. Na obrazovce Run Settings (Nastavení běhu) zadejte jedinečný název běhu. Název běhu identifikuje běh od sekvenování až po analýzu.
2. **[Volitelné]** Pro další identifikaci běhu zadejte popis běhu.
3. Vyberte sadu (sady) indexového adaptéru použitou během přípravy knihovny.
4. Zkontrolujte délku čtení a v případě potřeby ji upravte. Výchozí hodnota Read 1 (1. čtení) a Read 2 (2. čtení) je 151 cyklů. Index 1 a Index 2 mají pevnou hodnotu 10 cyklů a nelze je měnit.
5. **[Volitelné]** Zadejte ID zkumavky knihovny.
6. Vyberte možnost **Next** (Další).

Data vzorku

Data vzorku zahrnují ID vzorku, pozici jamky (pozice jamky na desce indexu) a název knihovny. Při použití indexu A a B zahrnuje pozice jamky také identifikátor desky.

Údaje o vzorcích lze zadat dvěma způsoby:

- **Import Samples** (Importovat vzorky) – použijte vzorový soubor, který je k dispozici ke stažení na obrazovce Sample Data (Data vzorku).
- **Manually** (Ručně) – na obrazovce Sample Data (Data vzorku) zadejte data vzorku přímo do tabulky.

Importování vzorků

Při plánování sekvenování pomocí prohlížeče na síťovém počítači je na obrazovce Sample Data (Data vzorku) k dispozici ke stažení šablona (`*.csv`). Vzorový soubor není k dispozici ke stažení při přístupu k Illumina Run Manager prostřednictvím softwaru operačního systému NextSeq 550Dx. Chcete-li zadat data vzorků pomocí funkce Import Samples (Importovat vzorky), proveďte následující kroky.

POZNÁMKA Než budete pokračovat, proveďte kroky nastavení běhu.

1. Chcete-li stáhnout prázdný soubor CSV, vyberte možnost **Download Template** (Stáhnout šablonu).
2. Ze vzorového souboru zadejte data vzorku a poté soubor uložte. Název knihovny je volitelný.

POZNÁMKA Při použití indexu A a B musí údaje pro sloupec B zahrnovat pozici desky i jamky (pozice jamky na desce indexu). Příklad: A-A01, A-A02, A-A03.

3. Vyberte možnost **Import Samples** (Importovat vzorky) a přejděte na vzorový soubor obsahující informace o datech vzorků z předchozího kroku.
4. Vyberte možnost **Open** (Otevřít), **Proceed** (Pokračovat) a poté **Next** (Další).

POZNÁMKA Změna ID vzorku před výběrem možnosti Next (Další) může vést k chybě. Před provedením změn dokončete nastavení běhu, abyste předešli chybám.

Ruční zadání vzorků

Pomocí tabulky na obrazovce Sample Data (Data vzorku) zadejte ručně údaje o vzorku.

1. Do pole Sample ID (ID vzorku) zadejte jedinečné ID vzorku.
2. Pro výběr příslušného indexu pro vzorky použijte možnost **Well position** (Pozice jamky, index A nebo index B) nebo **Plate - Well Position** (Deska – poloha jamky, index A a B).
   Pole Index i7, Index 1, Index i5 a Index 2 se vyplní automaticky.
3. **[Volitelné]** Zadejte název knihovny.
4. Podle potřeby přidejte řádky a opakujte kroky 1–3, dokud nebudou do tabulky přidány všechny vzorky. Můžete přidat více řádků najednou tak, že nejprve zadáte počet řádků, které chcete přidat, a poté vyberete ikonu +. Řádky můžete také odstranit tak, že vyberete políčko vedle čísla řádku a poté kliknete na ikonu koše.
5. Vyberte možnost **Next** (Další).

Nastavení analýzy

1. Vyberte požadovaný pracovní postup analýzy:
   • Generování FASTQ
   • Generování FASTQ a VCF pro germinální pracovní postup (vyžaduje se soubor manifestu)
   • Generování FASTQ a VCF pro somatický pracovní postup (vyžaduje se soubor manifestu)
2. **[Volitelné]** **Generate ORA compressed FASTQs** (Generování ORA komprimovaných FASTQ) je ve výchozím nastavení povoleno. ORA komprese FASTQ bezztrátově komprimuje soubory FASTQ až 5× ve srovnání s fastq.gz. Pokud preferujete nekomprimovaná data (fastq.gz), zrušte zaškrtnutí políčka **Generate ORA compressed FASTQs** (Generovat ORA komprimované FASTQ).
3. U germinálních a somatických pracovních postupů je vyžadován soubor manifestu. Pomocí rozevírací nabídky **Manifest File Selection** (Výběr souboru manifestu) vyberte soubor manifestu. Manifest je soubor BED (`*.bed`), ve kterém jsou údaje odděleny pomocí tabulátoru a který specifikuje názvy a umístění cílových referenčních oblastí. Další informace naleznete v části *Soubor manifestu* na straně 8 (Soubor manifestu).
4. **[Volitelné]** Pro somatické pracovní postupy použijte rozbalovací nabídku **Noise File Selection** (Výběr souboru šumu) a vyberte soubor systematického šumu.
   Pro odfiltrování systematického šumu lze specifikovat soubor BED (`*.bed.gz`) s úrovní šumu pro dané pracoviště. Další informace naleznete v části *Filtrování šumu (volitelné)* na straně 9 (Filtrování šumu – volitelné).
5. Vyberte možnost **Next** (Další).

Běh Kontrola

1. Na obrazovce Review (Kontrola) zkontrolujte informace pro Run Settings (Nastavení běhu), Sample Data (Data vzorku) a Analysis Settings (Nastavení analýzy).
2. Vyberte možnost **Save** (Uložit).
   Běh se uloží na kartě Planned (Plánované) na obrazovce Runs (Běhy).

Nastavení

Pro zobrazení nebo změnu nastavení DRAGEN for IDPE Dx aplikace nejprve vyberte ikonu Aplikace na hlavní obrazovce. Poté vyberte aplikaci, kterou chcete zobrazit nebo změnit. Pro změnu nastavení je vyžadován účet správce.

Konfigurace

Konfigurační obrazovka zobrazuje následující nastavení aplikace:

- **Library Prep Kits** (Sady pro přípravu knihoven) – zobrazuje výchozí sadu pro přípravu knihovny pro aplikaci. Toto nastavení nelze změnit.

- **Index Adapter Kits** (Sady indexového adaptéru) – zobrazuje výchozí sadu indexového adaptéru pro aplikaci. Toto nastavení nelze změnit.
- **Read lengths** (Délky čtení) – délky čtení jsou pro aplikaci ve výchozím nastavení nastaveny na 151, ale lze je během vytváření běhu změnit.
- **Manifest and Noise Files** (Soubory manifestů a šumu) – umožňuje nahrávání a změnu nastavení pro soubory manifestů a šumu.
  - Vyberte možnost **Upload File** (Nahrát soubor), abyste nahráli soubory pro použití v analýze.
  - Vyberte přepínač **Default** (Výchozí) a nastavte jako výchozí soubor manifestu nebo šumu soubor vybraný během vytváření běhu, když je vybrána aplikace.
  - Zaškrtnutím políčka **Enabled** (Povoleno) nastavte soubor, který se má zobrazit v rozevírací nabídce během vytváření běhu.

Oprávnění

Pomocí zaškrtávacích políček na obrazovce Permissions (Oprávnění) můžete spravovat přístup uživatelů k aplikaci.

Soubor manifestu

Při použití DRAGEN for IDPE Dx je pro následující pracovní postupy vyžadován vstup souboru manifestu:

- Generování FASTQ a VCF pro germinální pracovní postup
- Generování FASTQ a VCF pro somatický pracovní postup

Soubor manifestu je textový soubor formátu BED (*.bed), ve kterém jsou údaje odděleny pomocí tabulátoru a který specifikuje názvy a umístění cílových referenčních oblastí. Hlavní částí souboru manifestu je část Regions (Regiony) a měla by obsahovat následující datové sloupce:

Sloupcová	Popis
Název	Jedinečný uživatelem specifikovaný název cíle
Chromozom	Umístění chromozomu (např. chr10, chr5 atd.)
Začátek	Index 1 pro počáteční pozici cíle
Konec	Index 1 pro koncovou pozici cíle
Délka upstream sondy	Délka upstream sondy. Pro aplikaci DRAGEN for IDPE Dx by měla být nastavena na 0.
Délka downstream sondy	Délka downstream sondy. Pro aplikaci DRAGEN for IDPE Dx by měla být nastavena na 0.

POZNÁMKA Pro analýzu je vyžadován platný formát souboru manifestu. Pokud je soubor manifestu neplatný, DRAGEN analýzu zastaví.

Filtrování šumu (volitelné)

Filtr systematického šumu je k dispozici pro přiřazování somatických variant a lze jej použít ke snížení falešně pozitivních přiřazení tím, že se zohlední hluk na pracovišti. Soubor systematického šumu se generuje tak, že se nejprve odebere přibližně 50 normálních vzorků (nejlépe specifických pro panel, přípravu knihovny a sekvenátor) a poté se součet frekvencí alel pod 30 % na každém pracovišti s dostatečným pokrytím vydělí celkovým počtem vzorků (předpokládá se, že frekvence alel nad 30 % jsou germinální varianty a nikoli šum). Po vygenerování hodnot šumu budou filtrovány somatické varianty detekované v daném místě.

Filtr lze použít v režimu Tumor-Normal (Nádor-Normál), ale je zvlášť užitečný pro běhy Tumor-Only (Pouze nádor), kde není k dispozici odpovídající normál. Soubor systematického šumu musí používat soubor BED s příponou (`*.bed.gz`) a musí obsahovat čtyři sloupce: Hladiny hluku specifické pro chromozom, začátek, konec a pracoviště pro každý řádek. Filtrování systematického šumu je volitelné.

Výstupy analýzy

Aktuálně probíhající běhy se zobrazují na kartě Active (Aktivní). Dokončené cykly se zobrazují na kartě Completed (Dokončeno). DRAGEN for IDPE Dx vytvoří pro každou analýzu složku s jedinečným názvem analýzy, která je oddělena od složky obsahující data sekvenování. Složka analýzy obsahuje následující informace:

- Použitý soubor manifestu
- Verze softwaru
- ID vzorku
- Celkový počet zarovnaných čtení
- Procento zarovnaných hodnot na vzorek
- Počet přiřazených SNV na vzorek
- Počet přiřazených indelů na vzorek
- Statistika pokrytí

Výstupní soubory analýzy

Umístění složky analýzy je určeno nastavením External Storage for Analysis Results (Externí úložiště pro výsledky analýzy). Další informace o nastavení External Storage for Analysis Results (Externí úložiště pro výsledky analýzy) naleznete v Příručka k softwaru Illumina Run Manager pro NextSeq 550Dx (dokument č. 200025239).

Na obrazovce Run Details (Podrobnosti běhu) poskytuje pole External Location (Externí umístění) cestu k datům sekvenování. Jedinečný název složky analýzy je uveden v poli Analysis Output Folder (Složka výstupů analýzy) na obrazovce Run Details (Podrobnosti běhu). Přesné vygenerované soubory závisí na tom, který pracovní postup analýzy bude použit. Aplikace generuje následující výstupní soubory analýzy.

POZNÁMKA Pokud při přístupu k výstupním souborům analýzy dojde k chybě omezení maximální délky cesty k souboru, zkuste soubor přesunout na kratší úsek cesty nebo soubor otevřete jiným způsobem.

Výstupní soubor	Popis
Souhrnný výkaz variant (`*.pdf`)	Obsahuje souhrn informací o souboru, verze softwaru, informace o vzorku, statistiku úrovně čtení a SNV, inserce, delece a souhrny pokrytí. Pouze germinální a somatické pracovní postupy vytvářejí výkaz variant.
FASTQ (`*.fastq.gz` nebo `*.fastq.ora`)	Přechodné soubory obsahující přiřazení báze se skóre kvality. Soubory FASTQ jsou primárním vstupem pro krok zarovnání. Je-li vybrána komprese ORA, používá se přípona souboru `*.fastq.ora`.
Zarovnané soubory BAM (`*.bam`)	Obsahují zarovnaná čtení pro daný vzorek.
Soubory VCF genomu (`*.gvcf.gz`)	Obsahují genotyp pro každou pozici, ať už přiřazen jako varianta nebo jako reference.
Soubory VCF (`*.vcf.gz`)	Obsahují varianty přiřazené na každé pozici.
Sestava metrik běhu (`*.csv`)	Obsahuje metriky kvality běhu, včetně neindexované celkové výtěžnosti a skóre Q30.

Soubory FASTQ

FASTQ (*.fastq.gz, *.fastq.ora) je textový formát souboru obsahující přiřazení bází a hodnoty kvality na jednotlivé čtení. Každý soubor obsahuje následující informace:

- Identifikátor vzorku
- Sekvence
- Skóre kvality Phred v kódovaném formátu ASCII + 33

Identifikátor vzorku je naformátován následovně:

```
@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=AAAAAAAAAA9#:<#<;<<<????#=
```

Soubory BAM

Soubor BAM (*.bam) je komprimovaná binární verze souboru SAM (mapa zarovnání sekvence), který se používá k reprezentaci zarovnaných sekvencí do 128 Mb. Soubory BAM používají formát pojmenování souboru `SampleName_S#.bam`, přičemž # je číslo vzorku určené pořadím, ve kterém jsou uvedeny vzorky pro daný běh. V režimu více uzlů je S# nastaveno na S1 bez ohledu na pořadí vzorku.

Soubory BAM obsahují sekci záhlaví a sekci zarovnání:

- **Header** (Záhlaví) – obsahuje informace o celém souboru, jako je název vzorku, délka vzorku a metoda zarovnání. Zarovnání (Alignments) jsou v sekci zarovnání spojena s konkrétními informacemi v sekci záhlaví.
- **Alignments** (Zarovnání) – obsahuje název čtení, sekvenci čtení, kvalitu čtení, informace o zarovnání a vlastní tagy. Název čtení zahrnuje chromozom, počáteční souřadnici, kvalitu zarovnání a řetězec deskriptoru shody.

Sekce zarovnání obsahuje následující informace pro každé čtení nebo pár čtení:

- AS: Kvalita zarovnání párového konce.
- RG: Skupina čtení, která označuje počet čtení pro konkrétní vzorek.
- BC: Tag s čárovým kódem, který označuje ID demultiplexovaného vzorku spojené se čtením.
- SM: Kvalita zarovnání jednoho konce.
- XC: Shoda řetězce deskriptoru.
- XN: Značka názvu amplikonu, která zaznamenává ID amplikonu spojeného se čtením

Soubory indexu BAM (*.bam.bai) uvádějí index odpovídajícího souboru BAM.

Soubory VCF

Soubory ve formátu přiřazení variant (*.vcf) obsahují informace o variantách nalezených na určitých pozicích v referenčním genomu.

Záhlaví souboru VCF obsahuje verzi formátu souboru VCF, verzi detekčního programu pro varianty a uvádí poznámky použité ve zbytku souboru. Záhlaví VCF také obsahuje referenční soubor genomu a soubor BAM. Poslední řádek v záhlaví obsahuje nadpisy sloupců pro datové řádky. Každý z datových řádků souboru VCF obsahuje informace o jedné variantě.

Tabulka 1 Záhlaví souboru VCF

Záhlaví	Popis
CHROM	Chromozom referenčního genomu. Chromozomy se zobrazují ve stejném pořadí jako referenční soubor FASTA.
POS	Jednobázová pozice varianty v referenčním chromozomu. Pro jednonukleotidové varianty (SNV) je tato pozice referenční bází s danou variantou. U indelů je tato pozice referenční bází bezprostředně předcházející dané variantě.
ID	Číslo rs (referenční SNP) pro SNP získané z `dbSNP.txt`, je-li to relevantní. Pokud na tomto místě existuje více čísel rs, seznam je oddělen středníky. Pokud záznam dbSNP na této pozici neexistuje, použije se značka chybějící hodnoty ('.').
REF	Název referenčního genotypu. Například delece jediného T je reprezentována jako referenční TT a alternativní T. Varianta jednoho nukleotidu A až T je reprezentována jako referenční A a alternativní T.
ALT	Alely, které se liší od referenčního čtení. Například inzerce jediného T je reprezentována jako referenční A a alternativní AT. Varianta jednoho nukleotidu A až T je reprezentována jako referenční A a alternativní T.
QUAL	Skóre kvality na stupnici Phred přiřazené detekčním programem pro varianty. Vyšší skóre ukazuje na vyšší důvěryhodnost dané varianty a nižší pravděpodobnost chyb. U skóre kvality Q je odhadovaná pravděpodobnost chyby 10-(Q/10). Například sada přiřazení Q30 má 0,1% chybovost. Řada detekčních programů pro varianty přiřazuje skóre kvality na základě svých statistických modelů, které jsou ve vztahu k pozorované chybovosti vysoké.

Tabulka 2 Poznámky k VCF souboru pro germinální pracovní postup

Záhlaví	Popis
FILTER	Pokud jsou všechny filtry úspěšné, zapíše se do sloupce filtru VYHOVUJE. Možné vstupy FILTRU zahrnují: • **DRAGENSnpHardQUAL** – používá se, pokud skóre QUAL varianty SNP nesplňuje prahovou hodnotu • **DRAGENIndelHardQUAL** – používá se, pokud skóre QUAL indelové varianty nesplňuje prahovou hodnotu • **LowDepth** – místo je filtrováno, protože hloubka pokrytí nesplňuje prahovou hodnotu • **LowGQ** – místo je filtrováno, protože kvalita genotypu nesplňuje prahovou hodnotu • **PloidyConflict** – přiřazení genotypu od detekčního programu pro varianty není konzistentní s chromozomovou ploidií • **base_quality** – místo je filtrováno, protože medián kvality bází alternativních čtení v tomto lokusu nesplňuje prahovou hodnotu • **filtered_reads** – místo je filtrováno, protože byla odfiltrována příliš velká frakce čtení • **fragment_length** – místo je filtrováno, protože absolutní rozdíl mezi mediánem délky fragmentu alternativních čtení a mediánem délky fragmentu referenčních čtení v tomto lokusu překračuje prahovou hodnotu • **low_depth** – místo je filtrováno, protože hloubka čtení je příliš nízká • **low_frac_info_reads** – místo je filtrováno, protože frakce informativních čtení je pod prahovou hodnotou • **low_normal_depth** – místo je filtrováno, protože hloubka čtení normálního vzorku je příliš nízká • **long_indel** – místo je filtrováno, protože délka indelu je příliš dlouhá • **mapping_quality** – místo je filtrováno, protože medián kvality mapování alternativních čtení v tomto lokusu nesplňuje prahovou hodnotu • **multiallelic** – místo je filtrováno, protože více než dvě alely splňují LOD nádoru • **non_homref_normal** – místo je filtrováno, protože genotyp normálního vzorku není homozygotní referencí • **no_reliable_supporting_read** – místo je filtrováno, protože neexistuje spolehlivé podpůrné somatické čtení • **panel_of_normals** – zobrazeno alespoň v jednom vzorku ve vcf panelu normálů • **read_position** – místo je filtrováno, protože medián vzdáleností mezi začátkem/koncem čtení a tímto lokusem je pod prahovou hodnotou • **RMxNRepeatRegion** – místo je filtrováno, protože celá variantní alela nebo její část je opakováním reference • **strand_artifact** – místo je filtrováno z důvodu závažného vychýlení vlákna • **str_contraction** – místo je filtrováno kvůli podezření na chybu PCR, kde je alternativní alela o jednu opakovanou jednotku menší než reference • **too_few_supporting_reads** – místo je filtrováno, protože ve vzorku nádoru je příliš málo podpůrných čtení • **weak_evidence** – skóre somatických variant nesplňuje prahovou hodnotu

Záhlaví	Popis
INFORMACE	Možné záznamy INFORMACE zahrnují: • **AC** – počet alel v genotypech pro každou ALT alelu, ve stejném pořadí, v jakém jsou uvedeny • **AF** – frekvence alel pro každou alelu ALT, ve stejném pořadí, v jakém jsou uvedeny • **AN** – celkový počet alel v přiřazených genotypech • **DB** – členství v dbSNP • **FS** – p-hodnota škálovaná podle Phred měřená pomocí Fisherova exaktního testu k detekci vychýlení vlákna • **QD** – důvěryhodnost/kvalita variant podle hloubky • **R2_5P_bias** – skóre založené na „mate bias“ a vzdálenosti od 5 prime konce. • **SOR** – symetrický poměr pravděpodobností 2x2 kontingenční tabulky pro detekci vychýlení vlákna • **DP** – přibližná hloubka čtení (informativní a neinformativní); některá čtení mohla být filtrována na základě mapq atd. • **END** – poloha zastavení intervalu • **FractionInformativeReads** – frakce informativních čtení z celkového počtu čtení • **MQ** – kvalita mapování RMS • **MQRankSum** – Z-skóre z Wilcoxonova sumárního testu kvality mapování čtení Alt oproti čtení Ref • **ReadPosRankSum** – Z-skóre z Wilcoxonova sumárního testu pořadí vychýlení pozice při čtení Alt oproti čtení Ref • **SOMATIC** – alespoň jedna varianta na této pozici je somatická

Záhlaví	Popis
FORMÁT	Ve sloupci formátu jsou uvedena pole oddělená dvojtečkami. Například GT:GQ. Dostupná pole zahrnují: • **AD** – hloubky alel (počítají se pouze informativní čtení z celkových čtení) pro ref a alt alely v uvedeném pořadí • **AF** – alelové frakce pro alt alely v uvedeném pořadí • **DP** – přibližná hloubka čtení (čtení s MQ = 255 nebo se špatnými „mate“ jsou filtrována) • **F1R2** – počet čtení v párové orientaci F1R2 podporující každou alelu • **F2R1** – počet čtení v párové orientaci F2R1 podporující každou alelu • **GT** – genotyp; 0 odpovídá referenční bázi, 1 odpovídá prvnímu záznamu ve sloupci ALT atd. Přední lomítko (/) signalizuje, že nejsou k dispozici žádné informace o fázování. • **MB** – statistiky komponent vztažené na vzorek pro detekci „mate bias“ • **PS** – informace o ID fyzického fázování, kde každé jedinečné ID v daném vzorku (ale nikoli napříč vzorky) spojuje záznamy v rámci skupiny fázování • **SB** – statistiky komponent vztažené na vzorek, které tvoří Fisherův exaktní test pro detekci vychýlení vlákna • **SQ** – somatická kvalita
VZOREK	Sloupec vzorku uvádí hodnoty specifikované ve sloupci FORMÁT.

Tabulka 3 Poznámky k VCF souboru VCF pro somatický pracovní postup

Záhlaví	Popis
FILTR	Pokud jsou všechny filtry úspěšné, zapíše se do sloupce filtru VYHOVUJE. Možné vstupy FILTRU zahrnují: • **base_quality** – místo je filtrováno, protože medián kvality bází alternativních čtení v tomto lokusu nesplňuje prahovou hodnotu • **filtered_reads** – místo je filtrováno, protože byla odfiltrována příliš velká frakce čtení • **fragment_length** – místo je filtrováno, protože absolutní rozdíl mezi mediánem délky fragmentu alternativních čtení a mediánem délky fragmentu referenčních čtení v tomto lokusu překračuje prahovou hodnotu • **low_depth** – místo je filtrováno, protože hloubka čtení je příliš nízká • **low_frac_info_reads** – místo je filtrováno, protože frakce informativních čtení je pod prahovou hodnotou • **low_normal_depth** – místo je filtrováno, protože hloubka čtení normálního vzorku je příliš nízká • **long_indel** – místo je filtrováno, protože délka indelu je příliš dlouhá • **mapping_quality** – místo je filtrováno, protože medián kvality mapování alternativních čtení v tomto lokusu nesplňuje prahovou hodnotu • **multiallelic** – místo je filtrováno, protože více než dvě alely splňují LOD nádoru • **non_homref_normal** – místo je filtrováno, protože genotyp normálního vzorku není homozygotní referencí • **no_reliable_supporting_read** – místo je filtrováno, protože neexistuje spolehlivé podpůrné somatické čtení • **panel_of_normals** – zobrazeno alespoň v jednom vzorku ve vcf panelu normálů • **read_position** – místo je filtrováno, protože medián vzdáleností mezi začátkem/koncem čtení a tímto lokusem je pod prahovou hodnotou • **RMxNRepeatRegion** – místo je filtrováno, protože celá variantní alela nebo její část je opakováním reference • **strand_artifact** – místo je filtrováno z důvodu závažného vychýlení vlákna • **str_contraction** – místo je filtrováno kvůli podezření na chybu PCR, kde je alternativní alela o jednu opakovanou jednotku menší než reference • **too_few_supporting_reads** – místo je filtrováno, protože ve vzorku nádoru je příliš málo podpůrných čtení • **weak_evidence** – skóre somatických variant nesplňuje prahovou hodnotu • **systematic_noise** – místo je filtrováno na základě důkazů systémového šumu v normálech

Záhlaví	Popis
INFORMACE	Možné záznamy INFORMACE zahrnují: • **DP** – přibližná hloubka čtení (informativní a neinformativní); některá čtení mohla být filtrována na základě mapq atd. • **END** – poloha zastavení intervalu • **FractionInformativeReads** – frakce informativních čtení z celkového počtu čtení • **MQ** – kvalita mapování RMS • **MQRankSum** – Z-skóre z Wilcoxonova sumárního testu kvality mapování čtení Alt oproti čtení Ref • **ReadPosRankSum** – Z-skóre z Wilcoxonova sumárního testu pořadí vychýlení pozice při čtení Alt oproti čtení Ref • **AQ** – skóre systémového šumu • **hotspot** – známé somatické místo, které se používá ke zvýšení důvěryhodnosti přiřazení. • **SOMATIC** – alespoň jedna varianta na této pozici je somatická
FORMÁT	Ve sloupci formátu jsou uvedena pole oddělená dvojtečkami. Například GT:GQ. Dostupná pole zahrnují: • **AD** – hloubky alel (počítají se pouze informativní čtení z celkových čtení) pro ref a alt alely v uvedeném pořadí • **AF** – alelové frakce pro alt alely v uvedeném pořadí • **DP** – přibližná hloubka čtení (čtení s MQ = 255 nebo se špatnými „mate“ jsou filtrována) • **F1R2** – počet čtení v párové orientaci F1R2 podporující každou alelu • **F2R1** – počet čtení v párové orientaci F2R1 podporující každou alelu • **GP** – posteriorní pravděpodobnosti podle stupnice Phred pro genotypy, jak jsou definovány ve specifikaci VCF • **GQ** – kvalita genotypu • **GT** – genotyp; 0 odpovídá referenční bázi, 1 odpovídá prvnímu záznamu ve sloupci ALT atd. Přední lomítko (/) signalizuje, že nejsou k dispozici žádné informace o fázování. • **MB** – statistiky komponent vztažené na vzorek pro detekci „mate bias“ • **PL** – normalizované pravděpodobnosti podle stupnice Phred pro genotypy, jak jsou definovány ve specifikaci VCF • **PRI** – předchozí pravděpodobnosti podle stupnice Phred pro genotypy • **PS** – informace o ID fyzického fázování, kde každé jedinečné ID v daném vzorku (ale nikoli napříč vzorky) spojuje záznamy v rámci skupiny fázování • **SB** – statistiky komponent vztažené na vzorek, které tvoří Fisherův exaktní test pro detekci vychýlení vlákna • **SQ** – somatická kvalita
VZOREK	Sloupec vzorku uvádí hodnoty specifikované ve sloupci FORMÁT.

Soubory VCF genomu

Soubory VCF genomu (*.gvcf.gz) se řídí souborem konvencí pro reprezentaci všech míst v genomu v přiměřeně kompaktním formátu. Soubory gVCF zahrnují všechna místa v oblasti zájmu do jednoho souboru pro každý vzorek. Soubor gVCF zobrazuje případy bez přiřazení na pozicích, které neprojdou všemi filtry. Tag genotypu (GT) ./. označuje případ bez přiřazení.

Opětovné zařazení analýzy

Analýzu lze opětovně zařadit v případě, že byla zastavena, nezdařila se nebo chcete-li znovu analyzovat běh s jiným nastavením. Pro opětovné zařazení analýzy proveďte následující kroky:

1. Na obrazovce Runs (Běhy) vyberte kartu Completed (Dokončeno) a poté vyberte název cyklu, který chcete znovu analyzovat.
   Pokud bylo Requeue Analysis (Opětovné zařazení analýzy) již provedeno, vyberte název Parent Run (nadřazeného běhu).
2. Na obrazovce Run Details (Podrobnosti běhu) vyberte za položkou Sequencing Information (Informace o sekvenování) možnost **Requeue Analysis** (Opětovné zařazení analýzy).
3. Vyberte možnost:
   - Opětovné zařazení analýzy beze změn
   - Úprava nastavení běhu a opětovné zařazení analýzy
   - Opětovné zařazení analýzy s jinou aplikací
4. Potvrďte, že umístění, kde se momentálně nacházejí data sekvenování, je uvedeno v poli **Sequencing data file path** (Cesta k souboru dat sekvenování).

POZNÁMKA Cesta k datům sekvenování by měla odpovídat cestě v nastavení External Storage for Analysis Results (Externí úložiště pro výsledky analýzy). Další informace o změně cesty externího úložiště naleznete v Příručka k softwaru Illumina Run Manager pro NextSeq 550Dx (dokument č. 200025239).

5. Zadejte důvod opětovné analýzy.
6. Vyberte možnost **Requeue Analysis** (Znovu zařadit analýzu).
7. Upravte požadované změny v částech Run Settings (Nastavení běhu), Sample Data (Data vzorku) a Analysis Settings (Nastavení analýzy).
8. Vyberte možnost **Save** (Uložit). Analýza pak bude zahájena s použitím aktuálních parametrů.

Technická pomoc

Pokud potřebujete technickou pomoc, obraťte se na technickou podporu společnosti Illumina.

Web:	www.illumina.com
E-mail:	techsupport@illumina.com

Bezpečnostní listy (SDS) – k dispozici na webu společnosti Illumina na adrese support.illumina.com/sds.html.

Dokumentace k produktu – k dispozici ke stažení z webu support.illumina.com.

Illumina
5200 Illumina Way
San Diego, Kalifornie 92122, Spojené státy americké
+1 800 809 ILMN (4566)
+1 858 202 4566 (mimo Severní Ameriku)
techsupport@illumina.com
www.illumina.com

IVD

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

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

URČENO K DIAGNOSTICE IN VITRO.

illumina®