

DRAGEN for Illumina DNA Prep with Enrichment Dx

Dokumentace k přístroji NovaSeq 6000Dx

Historie revizí

Dokument	Datum	Popis změny
200014776 v02	Září 2022	V pokynech pro vytvoření běhu byl opraven formát souboru manifestu z textu (*.txt) na BED (*.bed). V části výstupu analýzy byly konsenzuální soubory VCF opraveny na soubory VCF.
200014776 v01	Srpen 2022	Přidáno: Část Nastavení. Část Filtrování systémového šumu. Aktualizovány pokyny pro vytvoření běhu tak, aby obsahovaly podrobnější informace. Opraveny překlepy a gramatické chyby. Bylo specifikováno, že pokyny jsou určeny pro aplikaci při použití s přístrojem NovaSeq 6000Dx. Byly aktualizovány informace o obsahu výstupního souboru VCF.
200014776 v00	Březen 2022	První vydání.

Tento dokument a jeho obsah je vlastnictvím společnosti Illumina, Inc. a jejích přidružených společností (dále jen „Illumina“). Slouží výlučně zákazníkovi ke smluvním účelům v souvislosti s použitím zde popsaných produktů a k žádnému jinému účelu. Tento dokument a jeho obsah nesmí být používán ani šířen za žádným jiným účelem ani jinak sdělován, zveřejňován či rozmnožován bez předchozího písemného souhlasu společnosti Illumina. Společnost Illumina nepředává tímto dokumentem žádnou licenci na svůj patent, ochrannou známku, autorské právo či práva na základě zvykového práva ani žádná podobná práva třetích stran.

Pokyny v tomto dokumentu musí být důsledně a výslovně dodržovány kvalifikovaným a řádně proškoleným personálem, aby bylo zajištěno správné a bezpečné používání zde popsaných produktů. Veškerý obsah tohoto dokumentu musíte před použitím takových produktů beze zbytku přečíst a pochopit.

NEDODRŽENÍ POŽADAVKU NA PŘEČTENÍ CELÉHO TEXTU A NA DŮSLEDNÉ DODRŽOVÁNÍ ZDE UVEDENÝCH POKYŇŮ MŮŽE VÉST K POŠKOZENÍ PRODUKTŮ, PORANĚNÍ OSOB, AŽ UŽ UŽIVATELŮ ČI JINÝCH OSOB, A POŠKOZENÍ JINÉHO MAJETKU A POVEDE KE ZNEPLATNĚNÍ JAKÉKOLI ZÁRUKY VZTAHUJÍCÍ SE NA PRODUKT.

SPOLEČNOST ILLUMINA NA SEBE NEBERE ŽÁDNOU ODPOVĚDNOST VYPLÝVAJÍCÍ Z NESPRÁVNÉHO POUŽITÍ ZDE POPSANÝCH PRODUKTŮ (VČETNĚ DÍLŮ TĚCHTO PRODUKTŮ NEBO SOFTWARE).

© 2022 Illumina, Inc. Všechna práva vyhrazena.

Všechny ochranné známky jsou vlastnictvím společnosti Illumina, Inc. nebo jejích příslušných vlastníků. Podrobné informace o ochranných známkách naleznete na adrese www.illumina.com/company/legal.html.

Obsah

Historie revizí	ii
Přehled	1
Metody analýzy	1
Vytvoření běhu	4
Nastavení	5
Výstupy analýzy	7
Soubory FASTQ	8
Soubory BAM	8
Soubory VCF	9
Zobrazení výsledků analýzy	14
Technická pomoc	15

Přehled

Aplikace DRAGEN™ for Illumina® DNA Prep with Enrichment Dx provádí v závislosti na vybraném pracovním postupu analýzy demultiplexování, generování FASTQ, mapování čtení a zarovnání s referenčním genomem a přiřazení variant.

Metody analýzy

DRAGEN for Illumina DNA Prep with Enrichment Dx provádí v závislosti na vybraném pracovním postupu demultiplexování, generování FASTQ, mapování čtení a zarovnání s referenčním genomem:

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

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. 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

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.

Specifické hodnoty vybrané pro skóre v tomto algoritmu ukazují, jak vyvážit, pro zarovnání s více možnými interpretacemi, možnost indelu na rozdíl od jednoho nebo více SNP nebo preferenci zarovnání bez clippingu. Výchozí hodnoty skórování DRAGEN jsou pro aplikace přiřazování variant přiměřené pro zarovnání střední délky čtení s celým lidským referenčním genomem. Každá sada parametrů skórování Smith-Waterman představuje nepřesný model genomové mutace a chyb sekvenování. Pro některé aplikace mohou být vhodnější odlišně vyladěné hodnoty skórování zarovnání.

Přiřazování germinálních variant v aplikaci DRAGEN

Detekční program DRAGEN Germline Small Variant Caller přebírá mapovanou a zarovnanou DNA jako vstup a přiřazuje SNP a indely pomocí kombinace detekce ve sloupcích a místního *de novo* sestavení haplotypů.

Přiřaditelné referenční oblasti jsou nejprve identifikovány s dostatečným pokrytím zarovnání. Rychlé skenování seřazených čtení identifikuje v těchto referenčních oblastech aktivní oblasti, které jsou vycentrovány kolem sloupců pileup s důkazy o variantě. Aktivní oblasti jsou doplněny dostatečným kontextem, aby pokryly významný nerefereční 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 páteří. 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.

V režimu gVCF používaném pro přiřazení škálovatelné vícevzorkové varianty lze spustit detekční program DRAGEN Germline Small Variant Caller pro každý vzorek a vytvořit tak přechodný soubor přiřazení genomové varianty (gVCF). gVCF lze potom použít k účinné společné genotypizaci více vzorků, což umožňuje rychlé přírůstkové zpracování vzorků a škálování na velké velikosti kohort.

Vzhledem k tomu, že detekční program DRAGEN Germline Small Variant Caller má algoritmy, které umožňují efektivní rozlišení korelovaných chyb od skutečných variant, pravidla filtrování jsou velmi jednoduchá.

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.

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 vycentrovány kolem sloupců pileup s důkazy o variantě. Aktivní oblasti jsou doplněny dostatečným kontextem, aby pokryly významný nerefereční 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 páteří. 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 TLOD, DRAGEN Somatic Small Variant Caller nejprve skenuje podle referenční pozice každou kandidátní somatickou příhodu a také i referenční událost nad aktivní oblastí. Podmínečná pravděpodobnost $P(r|e)$ pozorování každého překrývajícího se čtení je odhadnuta jako maximální $P(r|H)$ pro haplotypy podporující danou událost. 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.

Vytvoření běhu

Pomocí následujících kroků můžete vytvořit běh v Illumina Run Manager na přístroji NovaSeq 6000Dx nebo pomocí prohlížeče na počítači připojeném k síti. Data vzorků lze zadat ručně nebo importem seznamu vzorků.

Nastavení aplikace a běhu

1. 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).
3. Na obrazovce Run Settings (Nastavení běhu) zadejte název běhu. Název běhu identifikuje běh od sekvenování až po analýzu.
4. **[Volitelné]** Pro další identifikaci běhu zadejte popis běhu.
5. Ujistěte se, že vybraná Library Prep Kit (Sada pro přípravu knihoven) je sada pro přípravu knihoven Illumina DNA Prep with Enrichment Dx.
6. Vyberte požadovanou sadu indexového adaptéru.
7. Zadejte Read Length (Délka čtení).
Výchozí hodnota Read 1 (1. čtení) a Read 2 (2. čtení) je 151 cyklů.
Index 1 a Index 2 mají pevnou hodnotu 10 cyklů.
8. **[Volitelné]** Zadejte ID zkumavky knihovny.
9. Vyberte možnost **Next** (Další).

Data vzorku

Pomocí tabulky na obrazovce Sample Data (Data vzorku) zadejte ručně informace o vzorku. Případně vyberte možnost **Import Samples** (Importovat vzorky) a informace o vzorcích nahrajte. Informace o importu informací o vzorcích najdete v části [Import vzorků na straně 5](#) (Importovat vzorky).

Ruční zadání vzorků

1. Do pole Sample ID (ID vzorku) zadejte jedinečné ID vzorku.
2. Použijte možnost **Plate - Well Position** (Destička - Poloha jamky) pro výběr polohy jamky. Pole i7 Index, Index 1, i5 Index 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.
5. Vyberte možnost **Next** (Další).

Import vzorků

Když v Illumina Run Manager plánujete běh pomocí prohlížeče na počítači připojeném k síti, na obrazovce Sample Data (Data vzorku) je k dispozici ke stažení šablona (*.csv).

1. Chcete-li stáhnout prázdný soubor CSV, vyberte možnost **Download Template** (Stáhnout šablonu).
2. Zadejte informace o vzorku ze souboru CSV a soubor uložte.
Soubor CSV se seznamem vzorků obsahuje následující datové sloupce: Sample ID (ID vzorku), Plate - Well Position (Destička - Poloha jamky), **volitelný** Library Name (Název knihovny).
3. Vyberte možnost **Import Samples** (Importovat vzorky) a přejděte na umístění souboru CSV.
4. Vyberte možnost **Next** (Další).

Nastavení analýzy

1. Vyberte požadovaný pracovní postup analýzy:
 - Generování FASTQ
 - Generování Germline FASTQ a VCF pro germinální pracovní postup
 - Generování Somatic FASTQ a VCF pro somatický pracovní postup
2. **[Volitelné]** V případě potřeby zaškrtněte políčko **Generate ORA compressed FASTQs** (Generovat ORA komprimované FASTQ), čímž povolíte kompresi FASTQ ORA.
3. **[Pracovní postupy generování VCF]** Pomocí rozevírací nabídky **Manifest File Selection** (Výběr souboru manifestu) vyberte soubor manifestu.
Soubor manifestu je vyžadovaným vstupem pro DRAGEN for Illumina DNA Prep with Enrichment Dx. Manifest je soubor BED (*.bed), ve kterém jsou údaje odděleny pomocí tabulátoru a který specifikuje názvy a umístění cílových referenčních oblastí.
4. **[Pracovní postup generování Somatic FASTQ a VCF]** Pomocí rozevírací nabídky **Noise File Selection** (Výběr souboru šumu) vyberte soubor šumu.
Pro odfiltrování systematického šumu lze specifikovat soubor BED s úrovní šumu pro dané pracoviště. Další informace najdete v části [Filtrování šumu na straně 6](#) (Filtrování šumu).
5. Vyberte možnost **Next** (Další).

Běh Kontrola

1. Na obrazovce Review (Kontrola) zkontrolujte informace zadané na obrazovkách 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

Výběrem aplikace na obrazovce Applications (Aplikace) zobrazíte aktuální nastavení a změníte nastavení.

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

Filtrování šumu

Při použití somatického pracovního postupu je k dispozici filtrování systémového šumu. Filtr lze použít v režimu Tumor-Normal (Nádor-Normál), ale je zvláště užitečný pro běhy Tumor-Only (Pouze nádor), kde není k dispozici odpovídající normál.

BED systémového šumu by měl být generován z normálních vzorků. Doporučuje se vytvářet soubory systémového šumu, které jsou specifické pro přípravu knihovny, sekvenační systém a panel. Pro generování souborů šumu se doporučuje použít přibližně 50 normálních vzorků.

Výstupy analýzy

DRAGEN for Illumina DNA Prep with Enrichment Dx uloží následující informace do složky analýzy. Pouze germinální a somatické pracovní postupy vytvářejí PDF.

- 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

Aplikace generuje následující výstupní soubory. Přesné vygenerované soubory závisí na tom, který pracovní postup analýzy bude použit. Výstupní soubory jsou umístěny ve složce analýzy.

Výstupní soubor	Popis
FASTQ (*.fastq.gz nebo *.fastq.ora)	Přechodné soubory obsahující přiřazení báze se skóre kvality. Soubory FASTQ jsou primárním vstupem pro krok zarovnání. Pokud je vybrána komprese ORA, odrazí se to v názvu souboru.
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 celkového výtěžku 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
- Znaménko plus (+)
- Skóre kvality Phred v kódovaném formátu ASCII + 33

Identifikátor vzorku je naformátován následovně.

```
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y
ReadNum:FilterFlag:0:SampleNumber
Příklad:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAAA9#:<#<;<<<?????#=#
```

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:

- Záhlaví – obsahuje informace o celém souboru, jako je název vzorku, délka vzorku a metoda zarovnání. Zarovnání jsou v sekci zarovnání spojena s konkrétními informacemi v sekci záhlaví.
- 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: Tag názvu amplikonu, který zaznamenává ID amplikonu přidružené ke čteným souborům indexu BAM (*.bam.bai), poskytuje 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 a 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 <code>dbSNP.txt</code> , 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 inserce jediného T je reprezentována jako reference 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 souborům VCF

Záhlaví	Popis
FILTER (FILTER)	<p>Pokud jsou všechny filtry úspěšné, zapíše se do sloupce filtru PASS (VYHOVUJE).</p> <p>Možné vstupy FILTER (FILTER) v germinálním pracovním postupu zahrnují:</p> <ul style="list-style-type: none"> • 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

Záhlaví	Popis
FILTER (FILTER) (pokračování)	<ul style="list-style-type: none"> • 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 nespĺňuje prahovou hodnotu <p>Možné vstupy FILTER (FILTER) v somatickém pracovním postupu zahrnují:</p> <ul style="list-style-type: none"> • base_quality – místo je filtrováno, protože medián kvality bází alternativních čtení v tomto lokusu nespĺňuje 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 nespĺňuje 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 nespĺňuje prahovou hodnotu • systematic_noise – místo je filtrováno na základě důkazů systémového šumu v normálech

Záhlaví	Popis
INFO (INFORMACE)	<p>Možné vstupy INFO v germinálním pracovním postupu zahrnují:</p> <ul style="list-style-type: none"> • 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 <p>Možné vstupy INFO v somatickém pracovním postupu zahrnují:</p> <ul style="list-style-type: none"> • 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

Záhlaví	Popis
FORMAT (FORMÁT)	<p data-bbox="368 279 1383 352">Ve sloupci formátu jsou uvedena pole oddělená dvojtečkami. Například GT:GQ.</p> <p data-bbox="368 359 1383 390">Dostupná pole germinálního pracovního postupu zahrnují:</p> <ul data-bbox="368 396 1383 1234" style="list-style-type: none"> • 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 <p data-bbox="368 1241 1383 1272">Dostupná pole somatického pracovního postupu zahrnují:</p> <ul data-bbox="368 1278 1383 1883" style="list-style-type: none"> • 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

Záhlaví	Popis
SAMPLE (VZOREK)	Sloupec vzorku uvádí hodnoty specifikované ve sloupci FORMAT (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.

Zobrazení výsledků analýzy

Aktuálně probíhající běhy se zobrazují na kartě Active (Aktivní). Dokončené běhy se zobrazují na kartě Completed (Dokončeno). Další informace o zobrazení výsledků viz [Dokumentace k přístroji NovaSeq 6000Dx \(dokument č. 200010105\)](#).

Technická pomoc

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

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

Telefonní čísla na technickou podporu společnosti Illumina

Oblast	Bezplatná linka	Mezinárodní
Austrálie	+61 1800 775 688	
Rakousko	+43 800 006249	+43 1 9286540
Belgie	+32 800 77 160	+32 3 400 29 73
Kanada	+1 800 809 4566	
Čína		+86 400 066 5835
Dánsko	+45 80 82 01 83	+45 89 87 11 56
Finsko	+358 800 918 363	+358 9 7479 0110
Francie	+33 8 05 10 21 93	+33 1 70 77 04 46
Německo	+49 800 101 4940	+49 89 3803 5677
Hongkong, Čína	+852 800 960 230	
Indie	+91 8006500375	
Indonésie		0078036510048
Irsko	+353 1800 936608	+353 1 695 0506
Itálie	+39 800 985513	+39 236003759
Japonsko	+81 0800 111 5011	
Malajsie	+60 1800 80 6789	
Nizozemsko	+31 800 022 2493	+31 20 713 2960
Nový Zéland	+64 800 451 650	
Norsko	+47 800 16 836	+47 21 93 96 93
Filipíny	+63 180016510798	
Singapur	1 800 5792 745	
Jižní Korea	+82 80 234 5300	
Španělsko	+34 800 300 143	+34 911 899 417

Oblast	Bezplatná linka	Mezinárodní
Švédsko	+46 2 00883979	+46 8 50619671
Švýcarsko	+41 800 200 442	+41 56 580 00 00
Tchaj-wan, Čína	+886 8 06651752	
Thajsko	+66 1800 011 304	
Spojené království	+44 800 012 6019	+44 20 7305 7197
USA	+1 800 809 4566	+1 858 202 4566
Vietnam	+84 1206 5263	

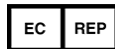
Bezpečnostní listy (SDS) – k dispozici na webu společnosti Illumina na adrese support.illumina.com/sds.html.

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



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

CE



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

Australský sponzor

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

URČENO K DIAGNOSTICE IN VITRO

© 2022 Illumina, Inc. Všechna práva vyhrazena.

illumina[®]