illumina®

DRAGEN for Illumina DNA Prep with Enrichment Dx

Produktdokumentation für NovaSeq 6000Dx

Dokument-Nr. 200014776 Version 02
September 2022
ZUR VERWENDUNG FÜR DIE IN-VITRO-DIAGNOSE

Versionsverlauf

Dokument	Datum	Beschreibung der Änderung
200014776 Version 02	September 2022	Korrigiertes Format für die Manifestdatei von Text (*.txt) zu BED (*.bed) in der Anleitung zum Erstellen eines Laufs. Korrigierte Konsens-VCF-Dateien in VCF-Dateien im Abschnitt „Analyseausgabe“.
200014776 Version 01	August 2022	Hinzugefügt: Abschnitt „Einstellungen“. Abschnitt „Systematische Rauschfilterung“. Anweisungen zur Lauferstellung für mehr Einzelheiten aktualisiert. Tipp- und Grammatikfehler korrigiert. Angegeben, dass Anweisungen für die Anwendung bestimmt sind, wenn sie mit dem NovaSeq 6000Dx Gerät benutzt wird. Informationen zum Inhalt der VCF-Ausgabedatei aktualisiert.
200014776 Version 00	März 2022	Erste Version.

Inhaltsverzeichnis

Überblick

Mit der Anwendung DRAGEN™ for Illumina® DNA Prep with Enrichment Dx können Sie Demultiplexing, FASTQ-Generierung, Read-Zuordnung und -Alignment mit einem Referenzgenom und Varianten-Calling je nach ausgewähltem Analysearbeitsablauf durchführen.

Analysemethoden

DRAGEN for Illumina DNA Prep with Enrichment Dx führt je nach ausgewähltem Arbeitsablauf Demultiplexing, FASTQ-Generierung, Read-Zuordnung und Alignment mit einem Referenzgenom durch:

- FASTQ-Generierung
- Germline FASTQ- und VCF-Generierung
- Somatic FASTQ- und VCF-Generierung

FASTQ-Generierung

Die zusammengesetzten Sequenzen werden pro Probe in FASTQ-Dateien geschrieben. FASTQ-Dateien sind Textdateien, die Sequenzierungsdaten und Qualitäts-Scores für nur eine Probe enthalten. Für jede Probe werden separate FASTQ-Dateien pro Fließzellen-Lane und Sequenzierungs-Read generiert. Der Name der Probe, wie er während der Laufeinrichtung angegeben wurde, ist im FASTQ-Dateinamen enthalten. FASTQ-Dateien sind die primären Eingabedateien für das Alignment. Der erste Schritt der FASTQ-Generierung ist das Demultiplexing. Beim Demultiplexing werden Cluster, die den Filter passieren, einer Stichprobe zugewiesen. Dazu wird jede Index-Lesesequenz mit den für den Lauf angegebenen Indexsequenzen verglichen. In diesem Schritt werden keine Qualitätswerte berücksichtigt. Index-Reads werden so identifiziert:

- Die Proben sind gemäß der Reihenfolge, in der sie für den Lauf aufgelistet sind, und mit 1 beginnend durchnummeriert.
- Die Probennummer 0 ist für Cluster reserviert, die keiner Probe zugeordnet wurden.
- Cluster werden einer Probe zugewiesen, wenn die Indexsequenz genau übereinstimmt bzw. je Index-Read maximal eine Nichtübereinstimmung festgestellt wird.

In der Software ist die ORA-Komprimierung zum Komprimieren von FASTQ-Dateien enthalten. Beim ORA-Format (*.ora) bleibt die md5-Prüfsumme des FASTQ-Inhalts nach einem Komprimierungs- und Dekomprimierungszyklus für eine verlustfreie Komprimierung erhalten.

DNA-Zuordnung und -Alignment

In der ersten Zuordnungsphase werden Seeds aus dem Read generiert und exakte Übereinstimmungen im Referenzgenom gesucht. Diese Ergebnisse werden dann durch Ausführen eines kompletten Smith-Waterman-Alignments an den Positionen mit der höchsten Dichte an Seed-Übereinstimmungen

präzisiert. Dieser gut dokumentierte Algorithmus gleicht jede Read-Position mit allen Kandidatenpositionen der Referenz ab. Dieser Abgleich entspricht einer Matrix aus möglichen Alignments zwischen Read und Referenz. Der Smith-Waterman-Algorithmus generiert für jede dieser potenziellen Alignment-Positionen Scores, anhand derer beurteilt wird, ob das beste Alignment mit einer Nukleotid-Übereinstimmung oder -Nichtübereinstimmung (diagonale Bewegung), einer Deletion (horizontale Bewegung) oder einer Insertion (vertikale Bewegung) durch diese Matrixzelle gewandert ist. Bei einer Übereinstimmung zwischen Read und Referenz wird zum Score hinzuaddiert, bei einer Nichtübereinstimmung oder einem Indel wird vom Score subtrahiert. Als Alignment wird der höchste Gesamtscore gewählt, der beim Durchwandern der Matrix erzielt wird.

Die spezifischen für Scores gewählten Werte in diesem Algorithmus weisen darauf hin, wie bei einem Alignment mit mehreren möglichen Interpretationen ein Gleichgewicht zwischen dem möglichen Vorhandensein eines Indels im Gegensatz zu einem oder mehreren SNP und der Präferenz für ein Alignment ohne Clipping erzielt werden kann. Die DRAGEN-Standardwerte für den Score sind für das Alignen von Reads moderater Länge zu einem Referenz-Humangesamtgenom für Varianten-Calling-Anwendungen angemessen. Jede Gruppe von Smith-Waterman-Score-Parametern stellt ein ungenaues Modell von Genommutationen und Sequenzierungsfehlern dar. Anders eingestellte Alignment-Score-Werte sind für einige Anwendungen möglicherweise geeigneter.

DRAGEN Calling von Keimbahnvarianten

Der DRAGEN Germline Small Variant Caller (Caller für kleine Keimbahnvarianten) nutzt kartierte und ausgerichtete DNA-Reads als Eingabe und ruft SNP sowie Indels durch eine Kombination aus spaltenweiser Erkennung und lokaler *Neu*-Zusammensetzung von Haplotypen ab.

Callfähige Referenzregionen werden zuerst mit ausreichender Alignmentabdeckung identifiziert. Innerhalb dieser Referenzregionen identifiziert ein schneller Scan der sortierten Reads aktive Regionen, die um Pile-up-Spalten mit Hinweisen auf eine Variante zentriert sind. Die aktiven Regionen werden mit genügend Kontext aufgefüllt, um wichtige, Nicht-Referenzinhalte in der Nähe abzudecken. Beim Nachweis von Indels werden die aktiven Regionen zusätzlich aufgefüllt.

Ausgerichtete Reads werden innerhalb jeder aktiven Region abgeschnitten und zu einem De Bruijn-Diagramm zusammengesetzt. Die Kanten der abgeschnittenen Reads werden nach Beobachtungszählungen gewichtet, mit der Referenzsequenz als Grundlage. Nach einiger Graphbereinigung und -vereinfachung werden alle Quelle-Senke-Pfade als Kandidaten-Haplotypen extrahiert. Jeder Haplotyp ist nach Smith-Waterman auf das Referenzgenom ausgerichtet, um die Varianten zu bestimmen, die es repräsentiert. Dieser Satz von Ereignissen kann durch einen positionsbasierten Nachweis erweitert werden. Für jedes Read-Haplotyp-Paar wird die Wahrscheinlichkeit $P(r|H)$, den Read zu beobachten, unter der Annahme, dass der Haplotyp das wahre Ausgangsmuster ist, mithilfe des Paar-Hidden-Markov-Modells (HMM) geschätzt.

Beim Scannen nach Referenzposition über der aktiven Region werden Kandidaten-Genotypen aus diploiden Kombinationen von Variantenereignissen (SNP oder Indels) gebildet. Für jedes Ereignis (einschließlich Referenz) wird die bedingte Wahrscheinlichkeit $P(r|e)$ der Beobachtung jedes überlappenden Reads als das Maximum $P(r|H)$ für Haplotypen geschätzt, die das Ereignis unterstützen.

Diese werden zu der bedingten Wahrscheinlichkeit P(r|e1e2) für einen Genotyp (Ereignispaar) kombiniert und multipliziert, um die bedingte Wahrscheinlichkeit P(R|e1e2) zu erhalten, das gesamte Read-Pile-up zu beobachten. Mit der Bayes-Formel wird die A-Posteriori-Wahrscheinlichkeit P(e1e2|R) jedes diploiden Genotyps berechnet und die Auswahl aufgerufen.

Im gVCF-Modus für skalierbares Calling von Mehrfachproben-Varianten kann der DRAGEN Germline Small Variant Caller pro Probe durchgeführt werden, um eine intermediäre genomische Varianten-Call-Datei (genomic Variant Call File, gVCF) zu generieren. Der gVCF kann dann für eine effiziente gemeinsame Genotypisierung mehrerer Proben für eine schnelle schrittweise Verarbeitung von Proben und eine Skalierung auf große Kohortengrößen genutzt werden.

Da der DRAGEN Germline Small Variant Caller über Algorithmen verfügt, mit denen korrelierte Fehler effizient von echten Varianten unterschieden werden können, sind die Filterregeln sehr einfach.

DRAGEN Calling somatischer Varianten

Der DRAGEN Somatic Small Variant Caller (Caller für kleine somatische Varianten) nutzt kartierte und ausgerichtete DNA-Reads als Eingabe und ruft SNP sowie Indels durch eine lokale *Neu*-Zusammensetzung von Haplotypen in einer aktiven Region ab.

Callfähige Referenzregionen werden zuerst mit ausreichender Alignmentabdeckung identifiziert. Innerhalb dieser Referenzregionen identifiziert ein Scan der sortierten Reads aktive Regionen, die um Pile-up-Spalten mit Hinweisen auf eine Variante in den Tumor-Reads zentriert sind. Die aktiven Regionen werden mit genügend Kontext aufgefüllt, um wichtige, Nicht-Referenzinhalte in der Nähe abzudecken. Beim Nachweis von Indels werden die aktiven Regionen zusätzlich aufgefüllt.

Ausgerichtete Reads werden innerhalb jeder aktiven Region abgeschnitten und zu einem De Bruijn-Diagramm zusammengesetzt. Die Kanten der abgeschnittenen Reads werden nach Beobachtungszählungen gewichtet, mit der Referenzsequenz als Grundlage. Nach einiger Graphbereinigung und -vereinfachung werden alle Quelle-Senke-Pfade als Kandidaten-Haplotypen extrahiert. Jeder Haplotyp ist nach Smith-Waterman auf das Referenzgenom ausgerichtet, um die Varianten zu bestimmen, die es repräsentiert. Für jedes Read-Haplotyp-Paar wird die Wahrscheinlichkeit P(r|H), den Read zu beobachten, unter der Annahme, dass der Haplotyp das wahre Ausgangsmuster ist, mithilfe des Paar-Hidden-Markov-Modells (HMM) geschätzt.

Zur Bestimmung des TLOD-Scores scannt der DRAGEN Somatic Small Variant Caller zunächst nach Referenzposition für jedes somatische Kandidatenereignis sowie das Referenzereignis über der aktiven Region. Die bedingte Wahrscheinlichkeit P(r|e) der Beobachtung jedes überlappenden Reads wird als das Maximum P(r|H) für Haplotypen geschätzt, die das Ereignis unterstützen. Diese werden zur bedingten Wahrscheinlichkeit P(r|E) für eine Ereignishypothese E kombiniert, die eine Mischung aus dem somatischen Referenz- und Kandidaten-Allel über einen Bereich möglicher Allelhäufigkeiten umfasst und multipliziert, um die bedingte Wahrscheinlichkeit P(R|E) aus der Beobachtung des gesamten Read-Pile-up zu erhalten. Daraus wird ein TLOD-Score als Nachweis dafür berechnet, dass ein ALT-Allel in der Tumorprobe an einem bestimmten Locus vorhanden ist.

Erstellen eines Laufs

Richten Sie anhand der folgenden Schritte einen Lauf im Illumina Run Manager entweder auf dem NovaSeq 6000Dx oder mit einem Browser auf einem vernetzten Computer ein. Probendaten können manuell oder durch Importieren eines Probenblatts eingegeben werden.

Anwendungs- und Laufeinstellungen

1. Wählen Sie im Bildschirm „Runs“ (Läufe) „**Create Run**“ (Erstellen eines Laufs) aus.
2. Wählen Sie die Anwendung DRAGEN for Illumina DNA Prep with Enrichment Dx und anschließend „**Next**“ (Weiter) aus.
3. Geben Sie im Bildschirm „Settings“ (Einstellungen) einen Laufnamen ein. Der Laufname ist der Name, der den Lauf von der Sequenzierung bis hin zur Analyse identifiziert.
4. **[Optional]** Geben Sie eine Laufbeschreibung ein, um den Lauf näher zu bestimmen.
5. Stellen Sie sicher, dass das ausgewählte Bibliotheksvorb.-Kit ein Illumina DNA Prep with Enrichment Dx Bibliotheksvorb.-Kit ist.
6. Wählen Sie das gewünschte Indexadapter-Kit aus.
7. Geben Sie die Read-Länge ein.
   Read 1 und Read 2 haben einen Standardwert von 151 Zyklen.
   Index 1 und Index 2 haben einen festen Wert von 10 Zyklen.
8. **[Optional]** Geben Sie eine Bibliotheksröhrchen-ID ein.
9. Wählen Sie „**Next**“ (Weiter) aus.

Probendaten

Geben Sie Probeninformationen anhand der Tabelle auf dem Bildschirm „Sample Data“ (Probendaten) manuell ein. Sie können auch „**Import Samples**“ (Importieren von Proben) auswählen, um Probeninformationen hochzuladen. Weitere Informationen zum Importieren von Probeninformationen finden Sie unter *Importieren von Proben* auf Seite 5 (Importieren von Proben).

Manuelles Eingeben der Proben

1. Geben Sie im Feld „Sample ID“ (Proben-ID) eine eindeutige Proben-ID ein.
2. Wählen Sie die Well-Position über „**Plate - Well Position**“ (Platten-Well-Position) aus.
   Die Felder „i7-Index“, „Index 1“, „i5-Index“ und „Index 2“ werden automatisch ausgefüllt.
3. **[Optional]** Geben Sie einen Bibliotheksnamen ein.
4. Fügen Sie Zeilen hinzu und wiederholen Sie die Schritte 1 bis 3 nach Bedarf, bis alle Proben zur Tabelle hinzugefügt wurden.
5. Wählen Sie „**Next**“ (Weiter) aus.

Importieren von Proben

Wenn Sie einen Lauf im Illumina Run Manager mit einem Browser auf einem vernetzten Computer planen, steht auf dem Bildschirm „Sample Data“ (Probendaten) eine Vorlage (*.csv) zum Herunterladen zur Verfügung.

1. Wählen Sie zum Herunterladen einer leeren CSV-Datei „**Download Template**“ (Herunterladen einer Vorlage) aus.
2. Geben Sie aus der CSV-Datei die Probeninformationen ein und speichern Sie die Datei.
   In der CVS-Datei des Probenblatts sind diese Datenspalten enthalten: „Sample ID“ (Proben-ID), „Plate - Well Position“ (Platten-Well-Position), „**Optional** Library Name“ (optionaler Bibliotheksname).
3. Klicken Sie auf „**Import CSV**“ (Importieren von CSV) und gehen Sie zum Speicherort der Datei mit den Probeninformationen.
4. Wählen Sie „**Next**“ (Weiter) aus.

Analyseeinstellungen

1. Wählen Sie den gewünschten Analysearbeitsablauf aus:
   - FASTQ-Generierung
   - Germline FASTQ- und VCF-Generierung für einen Keimbahnarbeitsablauf
   - Somatic FASTQ- und VCF-Generierung für einen somatischen Arbeitsablauf
2. **[Optional]** Markieren Sie zum Aktivieren der FASTQ-ORA-Komprimierung bei Bedarf das Kästchen „**Generate ORA compressed FASTQs**“ (Generieren von ORA-komprimierten FASTQ).
3. **[Arbeitsabläufe zur VCF-Generierung]** Wählen Sie aus dem Drop-down-Menü „**Manifest File Selection**“ (Auswahl der Manifestdatei) eine Manifestdatei aus.
   Eine Manifestdatei ist als Eingabe für den DRAGEN for Illumina DNA Prep with Enrichment Dx erforderlich. Das Manifest ist eine tabulatorgetrennte BED-Datei (*.bed), die die Namen und Positionen der Target-Referenzregionen angibt.
4. **[Arbeitsablauf zur somatischen FASTQ- und VCF-Generierung]** Wählen Sie aus dem Drop-down-Menü „**Noise File Selection**“ (Auswahl der Rauschdatei) eine Rauschdatei aus.
   Zum Herausfiltern von systematischem Rauschen kann eine BED-Datei mit positionsspezifischem Rauschpegel angegeben werden. Weitere Informationen finden Sie unter *Rauschfilterung* auf Seite 6.
5. Wählen Sie „**Next**“ (Weiter) aus.

Lauf Überprüfen

1. Überprüfen Sie auf dem Bildschirm „Review“ (Überprüfen) die Informationen, die Sie auf den Bildschirmen „Run Settings“ (Laufeinstellungen), „Sample Data“ (Probendaten) und „Analysis Settings“ (Analyseeinstellungen) eingegeben haben.
2. Wählen Sie „**Save**“ (Speichern) aus.
   Der Lauf wird im Bildschirm „Runs“ (Läufe) unter der Registerkarte „Planned“ (Geplant) gespeichert.

Einstellungen

Wählen Sie auf dem Bildschirm „Applications“ (Anwendungen) die Anwendung aus, um die aktuellen Einstellungen aufzurufen und Einstellungen zu ändern.

Konfiguration

Im Bildschirm „Configuration“ (Konfiguration) werden diese Anwendungseinstellungen angezeigt:

- **Library Prep Kits** (Bibliotheksvorb.-Kits): Zeigt das standardmäßige Bibliotheksvorb.-Kit für die Anwendung an. Diese Einstellung kann nicht geändert werden.
- **Index Adapter Kits** (Index-Adapter-Kits): Zeigt das standardmäßige Index-Adapter-Kit für die Anwendung an. Diese Einstellung kann nicht geändert werden.
- **Read-Längen**: Die Read-Längen sind für die App standardmäßig auf 151 festgelegt, können jedoch während der Lauferstellung geändert werden.
- **Manifest- und Rauschdateien**: Laden Sie die Einstellungen für Manifest- und Rauschdateien hoch und ändern Sie sie.
  - Wählen Sie „**Upload File**“ (Datei hochladen) aus, um Dateien für die Analyse hochzuladen.
  - Wählen Sie das Optionsfeld „**Default**“ (Standard) aus, um die Datei als Standardmanifest oder Rauschdatei festzulegen, die während der Lauferstellung ausgewählt wird, wenn die Anwendung ausgewählt wird.
  - Markieren Sie das Kästchen „**Enabled**“ (Aktiviert), um festzulegen, dass die Datei während der Lauferstellung im Drop-down-Menü angezeigt wird.

Berechtigungen

Verwalten Sie über die Kästchen auf dem Bildschirm „Berechtigungen“ den Benutzerzugriff für die App.

Rauschfilterung

Für den somatischen Arbeitsablauf gibt es eine systematische Rauschfilterung. Der Filter kann im Tumor-Normal-Modus verwendet werden, ist aber besonders nützlich für Nur-Tumor-Läufe, bei denen kein übereinstimmender Normalwert verfügbar ist.

Das systematische Rauschen BED sollte aus normalen Proben erzeugt werden. Es wird empfohlen, systematische Rauschdateien zu erstellen, die für die Bibliotheksvorb., das Sequenzierungssystem und das Panel spezifisch sind. Für die Erzeugung von Rauschdateien werden etwa 50 normale Proben empfohlen.

Analyseausgaben

DRAGEN for Illumina DNA Prep with Enrichment Dx speichert die folgenden Informationen im Analyseordner. Nur die Keimbahn- und somatischen Arbeitsabläufe erzeugen ein PDF.

- Verwendete Manifestdatei
- Softwareversion
- Proben-ID
- Ausgerichtete Reads insgesamt
- Prozentsatz ausgerichteter Reads pro Probe
- Anzahl aufgerufener SNV pro Probe
- Anzahl aufgerufener Indels pro Probe
- Abdeckungsstatistik

Analyse-Ausgabedateien

Die folgenden Ausgabedateien werden von der Anwendung generiert. Die genauen generierten Dateien hängen davon ab, welcher Analyseablauf angewendet wird. Ausgabedateien befinden sich im Analyseordner.

Ausgabedatei	Beschreibung
FASTQ (*.fastq.gz oder *.fastq.ora)	Temporäre Dateien mit qualitativ benoteten Base-Calls. FASTQ-Dateien sind die primären Eingabedateien für den Alignment-Schritt. Wenn die ORA-Komprimierung ausgewählt ist, wird das im Dateinamen wiedergegeben.
Alignment-BAM-Dateien (*.bam)	Enthält ausgerichtete Reads für eine bestimmte Probe.
Genom-VCF-Dateien (*.gvcf.gz)	Enthält den Genotyp für jede Position, ob als Variante oder als Referenz aufgerufen.
VCF-Dateien (*.vcf.gz)	Enthält an jeder Position aufgerufene Varianten.
Laufmetrikbericht (*.csv)	Enthält Qualitätsmetriken zum Lauf, einschließlich Gesamtertrag und Q30-Score.

FASTQ-Dateien

FASTQ (*.fastq.gz, *.fastq.ora) ist ein textbasiertes Dateiformat mit den Base-Calls und Qualitätswerten pro Read. Jeder Eintrag enthält die folgenden Informationen:

- den Probenbezeichner
- die Sequenz
- ein Pluszeichen (+)
- die Phred-Qualitäts-Scores in einem ASCII-33-codierten Format

Der Probenbezeichner ist so formatiert:

```
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNum
ber
Beispiel:
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

BAM-Dateien

Eine BAM-Datei (*.bam) ist die komprimierte Binärversion einer SAM-Datei (Sequence Alignment Map [Sequenz-Alignment-Zuordnung]), mit der ausgerichtete Sequenzen bis zu 128 MB dargestellt werden. BAM-Dateien verwenden das Dateibenennungsformat `SampleName_S#.bam`, wobei # die Probennummer ist, die durch die Reihenfolge bestimmt wird, in der die Proben für den Lauf aufgelistet werden. Im Multinode-Modus wird S# unabhängig von der Reihenfolge der Probe auf S1 festgelegt.

BAM-Dateien verfügen über einen Kopfzeilen- und einen Alignment-Abschnitt:

- Kopfzeile: Enthält Informationen über die gesamte Datei, z. B. den Namen der Probe, die Probenlänge und die Alignment-Methode. Alignments im Abschnitt „Alignments“ (Ausrichtungen) sind bestimmten Informationen im Abschnitt „Header“ (Kopfzeile) zugeordnet.
- Alignments: Enthält Read-Name, Read-Sequenz, Read-Qualität, Informationen zu den Alignments und benutzerdefinierte Tags. Der Read-Name enthält das Chromosom, die Startkoordinate, die Alignment-Qualität und den Match-Deskriptor-String.

Im Abschnitt „Alignments“ (Ausrichtungen) sind diese Informationen für jeden Read oder jedes Read-Paar enthalten:

- AS: Paired-End-Alignment-Qualität.
- RG: Read-Gruppe, die die Anzahl der Reads für eine bestimmte Probe angibt.
- BC: Barcode-Tag, der die demultiplexierte Proben-ID angibt, die dem Read zugeordnet ist.

- SM: Single-End-Alignment-Qualität.
- XC: Match-Deskriptor-String.
- XN: Amplikon-Namens-Tag, das die mit den gelesenen BAM-Indexdateien (*.bam.bai) zugewiesene Amplikon-ID aufzeichnet, stellt einen Index der entsprechenden BAM-Datei bereit.

VCF-Dateien

Varianten-Call-Format-Dateien (*.vcf) enthalten Informationen über Varianten, die an spezifischen Positionen in einem Referenzgenom gefunden wurden.

In der Kopfzeile der VCF-Datei sind die VCF-Dateiformatversion und die Varianten-Caller-Version enthalten und die Anmerkungen angegeben, die im Rest der Datei verwendet werden. In der VCF-Kopfzeile sind auch die Referenzgenomdatei und die BAM-Datei enthalten. In der letzten Zeile in der Kopfzeile sind die Spaltenüberschriften für die Datenzeilen enthalten. In allen Datenzeilen der VCF-Datei sind Informationen zu einer Variante enthalten.

Tabelle 1 Kopfzeilen in VCF-Dateien

Kopfzeile	Beschreibung
CHROM	Das Chromosom des Referenzgenoms. Chromosomen erscheinen in der gleichen Reihenfolge wie die Referenz-FASTA-Datei.
POS	Die Single-Base-Position der Variante im Referenzchromosom. Für Einzelnukleotidvarianten (Single Nucleotide Variants, SNV) ist diese Position die Referenz-Base mit der Variante. Für Indels ist diese Position die Referenzbasis unmittelbar vor der Variante.
ID	Die rs-Nummer (Referenz-SNP) für den SNP, erhalten aus `dbSNP.txt`, falls zutreffend. Wenn es mehrere rs-Nummern an dieser Position gibt, wird die Liste durch Strichpunkte getrennt. Wenn an dieser Position kein dbSNP-Eintrag existiert, wird eine Kennung für einen fehlenden Wert ('.') verwendet.
REF	Der Referenzgenotyp. Beispielsweise wird die Deletion eines einzelnen T als Referenz TT und alternatives T dargestellt. Die Einzelnukleotidvarianten A bis T werden als Referenz A und alternatives T dargestellt.
ALT	Die Allele, die sich vom Referenz-Read unterscheiden. Beispielsweise wird eine Insertion eines einzelnen T als Referenz A und alternatives AT dargestellt. Die Einzelnukleotidvarianten A bis T werden als Referenz A und alternatives T dargestellt.

Kopfzeile	Beschreibung
QUAL	Ein vom Varianten-Caller zugewiesener Phred-skalierter Qualitäts-Score. Höhere Scores weisen auf eine höhere Zuverlässigkeit der Variante und eine geringere Fehlerwahrscheinlichkeit hin. Bei einem Qualitäts-Score von Q beträgt die geschätzte Fehlerwahrscheinlichkeit 10-(Q/10). Beispielsweise hat die Reihe der Q30-Calls eine Fehlerrate von 0,1 %. Viele Varianten-Caller weisen Qualitäts-Scores auf Basis ihrer statistischen Modelle zu, die im Verhältnis zur beobachteten Fehlerrate hoch sind.

Tabelle 2 Anmerkungen in VCF-Dateien

Kopfzeile	Beschreibung
FILTER	Wenn alle Filter erfolgreich waren, wird in die Filterspalte „PASS“ (ERFOLGREICH) geschrieben. Mögliche FILTER-Einträge im Keimbahnarbeitsablauf sind: • **DRAGENSnpHardQUAL**: Wird angewendet, wenn der QUAL-Score der SNP-Variante den Schwellenwert nicht erreicht. • **DRAGENIndelHardQUAL**: Wird angewendet, wenn der QUAL-Score der Indel-Variante den Schwellenwert nicht erreicht. • **LowDepth**: Position gefiltert, da die Abdeckungstiefe den Schwellenwert nicht erreicht. • **LowGQ**: Position gefiltert, da die Qualität des Genotyps den Schwellenwert nicht erreicht. • **PloidyConflict**: Genotyp-Call vom Varianten-Caller stimmt nicht mit Chromosomen-Ploidie überein. • **base_quality**: Position gefiltert, da die mittlere Basisqualität von Alt-Reads an diesem Locus den Schwellenwert nicht erreicht. • **filtered_reads**: Position gefiltert, da ein zu großer Teil der Reads herausgefiltert wurde. • **fragment_length**: Position gefiltert, da der absolute Unterschied zwischen der mittleren Fragmentlänge von Alt-Reads und der mittleren Fragmentlänge von Ref-Reads an diesem Locus den Schwellenwert überschreitet. • **low_depth**: Position gefiltert, da die Read-Tiefe zu gering ist. • **low_frac_info_reads**: Position gefiltert, da der Anteil informativer Reads unter dem Schwellenwert liegt. • **low_normal_depth**: Position gefiltert, da die Read-Tiefe normaler Proben zu gering ist. • **long_indel**: Position gefiltert, da die Indel-Länge zu lang ist. • **mapping_quality**: Position gefiltert, da die mediane Zuordnungsqualität von Alt-Reads an diesem Locus den Schwellenwert nicht erreicht. • **multiallelic**: Position gefiltert, da mehr als zwei alt-Allele die Tumor-LOD passieren. • **non_homref_normal**: Position gefiltert, da der normale Probengenotyp keine homozygote Referenz ist. • **no_reliable_supporting_read**: Position gefiltert, da kein zuverlässiger unterstützender somatischer Read vorhanden ist. • **panel_of_normals**: In mindestens einer Probe der Normalpanel-VCF festgestellt. • **read_position**: Position gefiltert, da der Median der Reichweiten zwischen Start/Ende des Reads und diesem Locus unter dem Schwellenwert liegt. • **RMxNRepeatRegion**: Position gefiltert, da das Varianten-Allel oder ein Teil davon eine Wiederholung der Referenz ist. • **strand_artifact**: Position gefiltert aufgrund starker Strang-Verzerrung.

Kopfzeile	Beschreibung
FILTER (Fortsetzung)	• **str_contraction**: Position gefiltert aufgrund eines vermuteten PCR-Fehlers, bei dem das alt-Allel eine Wiederholungseinheit weniger als die Referenz ist. • **too_few_supporting_reads**: Position gefiltert, da zu wenige unterstützende Reads in der Tumorprobe vorhanden sind. • **weak_evidence**: Score der somatischen Variante erreicht Schwellenwert nicht. Mögliche FILTER-Einträge im somatischen Arbeitsablauf sind: • **base_quality**: Position gefiltert, da die mittlere Basisqualität von Alt-Reads an diesem Locus den Schwellenwert nicht erreicht. • **filtered_reads**: Position gefiltert, da ein zu großer Teil der Reads herausgefiltert wurde. • **fragment_length**: Position gefiltert, da der absolute Unterschied zwischen der mittleren Fragmentlänge von Alt-Reads und der mittleren Fragmentlänge von Ref-Reads an diesem Locus den Schwellenwert überschreitet. • **low_depth**: Position gefiltert, da die Read-Tiefe zu gering ist. • **low_frac_info_reads**: Position gefiltert, da der Anteil informativer Reads unter dem Schwellenwert liegt. • **low_normal_depth**: Position gefiltert, da die Read-Tiefe normaler Proben zu gering ist. • **long_indel**: Position gefiltert, da die Indel-Länge zu lang ist. • **mapping_quality**: Position gefiltert, da die mediane Zuordnungsqualität von Alt-Reads an diesem Locus den Schwellenwert nicht erreicht. • **multiallelic**: Position gefiltert, da mehr als zwei alt-Allele die Tumor-LOD passieren. • **non_homref_normal**: Position gefiltert, da der normale Probengenotyp keine homozygote Referenz ist. • **no_reliable_supporting_read**: Position gefiltert, da kein zuverlässiger unterstützender somatischer Read vorhanden ist. • **panel_of_normals**: In mindestens einer Probe der Normalpanel-VCF festgestellt. • **read_position**: Position gefiltert, da der Median der Reichweiten zwischen Start/Ende des Reads und diesem Locus unter dem Schwellenwert liegt. • **RMxNRepeatRegion**: Position gefiltert, da das Varianten-Allel oder ein Teil davon eine Wiederholung der Referenz ist. • **strand_artifact**: Position gefiltert aufgrund starker Strang-Verzerrung. • **str_contraction**: Position gefiltert aufgrund eines vermuteten PCR-Fehlers, bei dem das alt-Allel eine Wiederholungseinheit weniger als die Referenz ist. • **too_few_supporting_reads**: Position gefiltert, da zu wenige unterstützende Reads in der Tumorprobe vorhanden sind. • **weak_evidence**: Score der somatischen Variante erreicht Schwellenwert nicht. • **systematic_noise**: Position gefiltert aufgrund von Anzeichen systematischen Rauschens bei Normalen.

Kopfzeile	Beschreibung
INFO	Mögliche INFO-Einträge im Keimbahnarbeitsablauf sind: • **AC**: Allelanzahl in Genotypen für jedes ALT-Allel, in derselben Reihenfolge wie angegeben. • **AF**: Allelhäufigkeit für jedes ALT-Allel, in derselben Reihenfolge wie angegeben. • **AN**: Gesamtzahl der Allele in aufgerufenen Genotypen. • **DB**: dbSNP-Mitgliedschaft. • **FS**: Phred-skalierter p-Wert anhand des exakten Fisher-Tests zur Erkennung von Strang-Verzerrungen. • **QD**: Variantenzuverlässigkeit/-qualität nach Tiefe. • **R2_5P_bias**: Score basierend auf Mate-Verzerrung und Reichweite von 5 Prime Ends. • **SOR**: Symmetrisches Chancenverhältnis der 2x2-Kontingenztabelle zur Erkennung von Strang-Verzerrungen. • **DP**: Ungefähre Read-Tiefe (informativ und nicht informativ); einige Reads wurden möglicherweise basierend auf Mapq usw. gefiltert. • **END**: Endposition des Intervalls. • **FractionInformativeReads**: Der Anteil informativer Reads an den Gesamt-Reads. • **MQ**: RMS-Zuordnungsqualität. • **MQRankSum**: Z-Score aus dem Wilcoxon-Rangsummentest von Alt- vs. Ref-Read-Zuordnungsqualitäten. • **ReadPosRankSum**: Z-Score aus dem Wilcoxon-Rangsummentest von Alt- vs. Ref-Read-Positionsverzerrungen. • **SOMATIC**: Mindestens eine Variante an dieser Position ist somatisch. Mögliche INFO-Einträge im somatischen Arbeitsablauf sind: • **DP**: Ungefähre Read-Tiefe (informativ und nicht informativ); einige Reads wurden möglicherweise basierend auf Mapq usw. gefiltert. • **END**: Endposition des Intervalls. • **FractionInformativeReads**: Der Anteil informativer Reads an den Gesamt-Reads. • **MQ**: RMS-Zuordnungsqualität. • **MQRankSum**: Z-Score aus dem Wilcoxon-Rangsummentest von Alt- vs. Ref-Read-Zuordnungsqualitäten. • **ReadPosRankSum**: Z-Score aus dem Wilcoxon-Rangsummentest von Alt- vs. Ref-Read-Positionsverzerrungen. • **AQ**: Score für systematisches Rauschen. • **hotspot**: Bekannte somatische Position, mit der die Zuverlässigkeit des Calls erhöht wird. • **SOMATIC**: Mindestens eine Variante an dieser Position ist somatisch.

Kopfzeile	Beschreibung
FORMAT	In der Formatspalte werden durch Doppelpunkte getrennte Felder angegeben. Beispiel: GT:GQ. Im Keimbahnarbeitsablauf sind diese Felder enthalten: • **AD**: Alleltiefen (wobei nur informative Reads von den Gesamt-Reads gezählt werden) für die ref- und alt-Allele in der angegebenen Reihenfolge. • **AF**: Allelfraktionen für Alt-Allele in der angegebenen Reihenfolge. • **DP**: Ungefähre Read-Tiefe (Reads mit MQ = 255 oder schlechten Mates werden gefiltert). • **F1R2**: Anzahl der Reads in F1R2-Paarorientierung, die jedes Allel unterstützen. • **F2R1**: Anzahl der Reads in F2R1-Paarorientierung, die jedes Allel unterstützen. • **GP**: Phred-skalierte Posterior-Wahrscheinlichkeiten für Genotypen gemäß der Definition in der VCF-Spezifikation. • **GQ**: Genotypqualität. • **GT**: Genotyp. 0 entspricht der Referenzbasis, 1 entspricht dem ersten Eintrag in der Spalte „ALT“ und so weiter. Der Schrägstrich (/) gibt an, dass keine Phaseninformationen verfügbar sind. • **MB**: Komponentenstatistiken pro Probe zur Erkennung von Mate-Verzerrungen. • **PL**: Normalisierte, Phred-skalierte Wahrscheinlichkeiten für Genotypen gemäß der Definition in der VCF-Spezifikation. • **PRI**: Phred-skalierte Prior-Wahrscheinlichkeiten für Genotypen. • **PS**: Physische Phasen-ID-Informationen, bei denen jede eindeutige ID innerhalb einer bestimmten Probe (aber nicht über Proben hinweg) Datensätze innerhalb einer Phasengruppe verbindet. • **SB**: Komponentenstatistiken pro Probe mit dem exakten Fisher-Tests zur Erkennung von Strang-Verzerrungen. • **SQ**: Somatische Qualität. Im somatischen Arbeitsablauf sind diese Felder enthalten: • **AD**: Alleltiefen (wobei nur informative Reads von den Gesamt-Reads gezählt werden) für die ref- und alt-Allele in der angegebenen Reihenfolge. • **AF**: Allelfraktionen für Alt-Allele in der angegebenen Reihenfolge. • **DP**: Ungefähre Read-Tiefe (Reads mit MQ = 255 oder schlechten Mates werden gefiltert). • **F1R2**: Anzahl der Reads in F1R2-Paarorientierung, die jedes Allel unterstützen. • **F2R1**: Anzahl der Reads in F2R1-Paarorientierung, die jedes Allel unterstützen. • **GT**: Genotyp. 0 entspricht der Referenzbasis, 1 entspricht dem ersten Eintrag in der Spalte „ALT“ und so weiter. Der Schrägstrich (/) gibt an, dass keine Phaseninformationen verfügbar sind.

Kopfzeile	Beschreibung
FORMAT (Fortsetzung)	• **MB**: Komponentenstatistiken pro Probe zur Erkennung von Mate-Verzerrungen. • **PS**: Physische Phasen-ID-Informationen, bei denen jede eindeutige ID innerhalb einer bestimmten Probe (aber nicht über Proben hinweg) Datensätze innerhalb einer Phasengruppe verbindet. • **SB**: Komponentenstatistiken pro Probe mit dem exakten Fisher-Tests zur Erkennung von Strang-Verzerrungen. • **SQ**: Somatische Qualität.
SAMPLE	Die Probenspalte enthält die Werte, die in der Spalte „FORMAT“ angegeben sind.

Genom-VCF-Dateien

Genom-VCF-Dateien (*.gvcf.gz) folgen einer Reihe von Konventionen zur Darstellung aller Positionen innerhalb des Genoms in einem angemessen kompakten Format. Die gVCF-Dateien enthalten alle Positionen innerhalb der interessierenden Region in einer einzigen Datei für jede Probe. Die gVCF-Datei zeigt No-Calls an Positionen, die nicht alle Filter bestehen. Ein Genotyp(GT)-Tag von ./. weist auf einen No-Call hin.

Aufrufen von Analyseergebnissen

Laufende Läufe werden in der Registerkarte „Active“ (Aktiv) angezeigt. Abgeschlossene Läufe werden in der Registerkarte „Completed“ (Abgeschlossen) angezeigt. Weitere Informationen zum Aufrufen von Ergebnissen finden Sie in der NovaSeq 6000Dx Produktdokumentation (Dokument-Nr. 200010105).

Technische Unterstützung

Wenn Sie technische Unterstützung benötigen, wenden Sie sich an den technischen Support von Illumina.

Website: www.illumina.com
E-Mail: techsupport@illumina.com

Telefonnummern des technischen Supports von Illumina

Region	Gebührenfrei	International
Australien	+61 1800 775 688	
Österreich	+43 800 006249	+43 1 9286540
Belgien	+32 800 77 160	+32 3 400 29 73
Kanada	+1 800 809 4566	
China		+86 400 066 5835
Dänemark	+45 80 82 01 83	+45 89 87 11 56
Finnland	+358 800 918 363	+358 9 7479 0110
Frankreich	+33 8 05 10 21 93	+33 1 70 77 04 46
Deutschland	+49 800 101 4940	+49 89 3803 5677
Hongkong, China	+852 800 960 230	
Indien	+91 8006500375	
Indonesien		0078036510048
Irland	+353 1800 936608	+353 1 695 0506
Italien	+39 800 985513	+39 236003759
Japan	+81 0800 111 5011	
Malaysia	+60 1800 80 6789	
Niederlande	+31 800 022 2493	+31 20 713 2960
Neuseeland	+64 800 451 650	
Norwegen	+47 800 16 836	+47 21 93 96 93
Philippinen	+63 180016510798	
Singapur	1 800 5792 745	
Südkorea	+82 80 234 5300	

Region	Gebührenfrei	International
Spanien	+34 800 300 143	+34 911 899 417
Schweden	+46 2 00883979	+46 8 50619671
Schweiz	+41 800 200 442	+41 56 580 00 00
Taiwan, China	+886 8 06651752	
Thailand	+66 1800 011 304	
Vereinigtes Königreich	+44 800 012 6019	+44 20 7305 7197
USA	+1 800 809 4566	+1 858 202 4566
Vietnam	+84 1206 5263	

Sicherheitsdatenblätter (SDS, Safety Data Sheets) sind auf der Illumina-Website unter support.illumina.com/sds.html verfügbar.

Die **Produktdokumentation** steht unter support.illumina.com zum Herunterladen zur Verfügung.

Illumina
5200 Illumina Way
92122 San Diego, Kalifornien, USA
+1.800.809.ILMN (4566)
+1.858.202.4566 (außerhalb von Nordamerika)
techsupport@illumina.com
www.illumina.com

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

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

ZUR VERWENDUNG FÜR DIE IN-VITRO-DIAGNOSE