FASTQ-Dateiformat

FASTQ ist ein textbasiertes Dateiformat mit Base-Calls und Qualitätswerten pro Read. Jeder Datensatz besteht aus vier Zeilen mit:

- dem Bezeichner
- der Sequenz
- einem Pluszeichen (+)
- den Phred-Qualitäts-Scores in einem ASCII-33-codierten Format

Der Bezeichner hat folgendes Format:

@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNumber

Beispiel:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

BAM-Dateiformat

Eine BAM-Datei (*.bam) ist die komprimierte Binärversion einer SAM-Datei, die für die Darstellung alignierter Sequenzen von bis zu 128 Mb verwendet wird. Eine ausführliche Beschreibung des SAM- und des BAM-Formats finden Sie unter samtools.github.io/hts-specs/SAMv1.pdf.

BAM-Dateien verwenden das Dateinamensformat Probenname_S#.bam, wobei # für die Probennummer steht, die sich von der Reihenfolge ableitet, in der die Proben für den Lauf aufgeführt sind.

BAM-Dateien enthalten einen Dateivorspann und einen Alignment-Bereich:

- **Dateivorspann:** Enthält Informationen über die gesamte Datei, z. B. den Namen der Probe, die Probenlänge und die Alignment-Methode. Alignments im Alignment-Bereich sind mit spezifischen Informationen im Kopfbereich verbunden.
- **Alignment-Bereich**: Enthält den Read-Namen, die Read-Sequenz, die Read-Qualität, Alignment-Informationen und anwendungsspezifische Tags. Der Read-Name enthält das Chromosom, die Startkoordinate, die Alignment-Qualität und die Übereinstimmungsdeskriptor-Zeichenfolge.

Im Alignment-Bereich sind für jeden Read bzw. jedes Read-Paar folgende Informationen aufgeführt:

- **AS:** Paired-End-Alignment-Qualität.
- **BC:** Barcode-Tag, das die dem Read zugeordnete demultiplexierte Proben-ID angibt.
- **SM:** Single-End-Alignment-Qualität.
- **XC:** Übereinstimmungsdeskriptor-Zeichenfolge.
- **XN:** Amplikon-Namens-Tag, das die dem Read zugeordnete Amplikon-ID aufzeichnet.

BAM-Indexdateien (*.bam.bai) zeigen einen Index der entsprechenden BAM-Datei.

VCF-Dateiformat

Das Variant Call Format (VCF) ist ein gängiges Dateiformat, das von der Genomik-Wissenschaftsgemeinde entwickelt wurde. Es enthält Informationen über Varianten, die an spezifischen Positionen in einem Referenzgenom gefunden wurden. VCF-Dateien haben die Dateierweiterung „.vcf“.

Der Dateivorspann der VCF-Datei enthält die Version des VCF-Dateiformats und des Varianten-Callers sowie eine Liste der im Rest der Datei verwendeten Annotationen. Im VCF-Dateivorspann sind außerdem die Referenzgenomdatei und die BAM-Datei angegeben. Die letzte Zeile im Dateivorspann enthält die Spaltenüberschriften für die Datenzeilen. Jede Datenzeile in der VCF-Datei enthält Informationen zu einer Variante.

Überschriften der VCF-Datei

Überschrift	Beschreibung
CHROM	Das Chromosom des Referenzgenoms. Chromosomen werden in derselben Reihenfolge wie in der FASTQ-Referenzdatei aufgeführt.
POS	Die Einzelbasenposition der Variante im Referenzchromosom. Bei SNPs ist diese Position die Referenzbase mit der Variante. Bei Indels oder Deletionen ist diese Position die Referenzbase unmittelbar vor der Variante.
ID	Die rs-Nummer für die Variante aus dbSNP.txt, falls vorhanden. Wenn es mehrere rs-Nummern an dieser Position gibt, wird die Liste durch Semikola getrennt. Wenn an dieser Position kein dbSNP-Eintrag existiert, wird eine Kennung für einen fehlenden Wert ('.') verwendet.
REF	Der Referenz-Genotyp. Beispielsweise wird die Deletion eines einzelnen T als Referenz-TT und alternatives T dargestellt. Eine Einzelnukleotid-Variante A bis T wird als Referenz-A und alternatives T dargestellt.
ALT	Die Allele, die sich vom Referenz-Read unterscheiden. Beispielsweise wird die Insertion eines einzelnen T als Referenz-A und alternatives AT dargestellt. Eine Einzelnukleotid-Variante A bis T wird als Referenz-A und alternatives T dargestellt.
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 in Relation zur beobachteten Fehlerrate hoch sind.

Anmerkungen in der VCF-Datei

Überschrift	Beschreibung
FILTER	Wenn alle Filter passiert werden, wird PASS in die Filterspalte geschrieben. • LowDP: Wird auf Positionen angewendet, deren Abdeckungstiefe geringer als 450-fach in jedem Pool ist. Bei Amplikonpositionen, die vom Vorwärts- und vom Rückwärtsread abgedeckt werden, entspricht dies einer 900-fachen Single-Read-Abdeckung. • LowGQ: Die Genotypisierungsqualität (GQ) liegt unter dem Cutoff. • q30: Qualitäts-Score < 30. • LowVariantFreq: Die Variantenhäufigkeit liegt unter dem angegebenen Grenzwert. • PB: Probe Pool Bias (Sondenpool-Verzerrung). Die Variante wurde nicht oder mit geringer Häufigkeit in einem oder zwei Sondenpools gefunden. • R3x6: Anzahl der an die Varianten-Calls angrenzenden Wiederholungen (Länge von einem bis drei Basenpaaren) ≥ 6. • SB: Der Strand Bias (die Strangverzerrung) liegt über dem angegebenen Grenzwert.
INFO	Die Spalte „INFO“ kann folgende Angaben enthalten: • AC: Allelanzahl in Genotypen für jedes ALT-Allel, in derselben Reihenfolge wie aufgeführt. • AF: Allelhäufigkeit für jedes ALT-Allel, in derselben Reihenfolge wie aufgeführt. • AN: Gesamtzahl der Allele in aufgerufenen Genotypen. • CD: Markierung, die angibt, dass der SNP in der codierenden Region von mindestens einem RefGene-Eintrag vorkommt. • DP: Tiefe (Anzahl der Base-Calls, die an einer Position ausgerichtet sind und beim Varianten-Calling verwendet werden). • Exon: Eine kommagetrennte Liste der Exon-Regionen, die aus dem RefGene gelesen wurden. • FC: Functional Consequence (funktionale Konsequenz). • GI: Eine kommagetrennte Liste der Gen-IDs, die aus dem RefGene gelesen wurden. • QD: Variantenzuverlässigkeit/-qualität nach Tiefe. • TI: Eine kommagetrennte Liste der Transkript-IDs, die aus dem RefGene gelesen wurden.
FORMAT	Die Spalte „Format“ enthält durch Doppelpunkt getrennte Felder. Beispiel: GT:GQ. Welche Felder in der Liste aufgeführt werden, hängt vom verwendeten Varianten-Caller ab. Folgende Felder sind verfügbar: • AD: Eintrag im Format X,Y, wobei X für die Anzahl der Referenz-Calls und Y für die Anzahl der alternativen Calls steht. • DP: Ungefähre Read-Tiefe. Reads mit MQ = 255 oder mit fehlerhaften „Mates“ werden herausgefiltert. • GQ: Genotypqualität. • GQX: Genotypqualität. GQX ist das Minimum des GQ-Werts und der QUAL-Spalte. In der Regel sind diese Werte ähnlich. Als Minimum ist GQX der konservativere Messwert für die Genotypqualität. • GT: Genotyp. 0 entspricht der Referenzbase, 1 entspricht dem ersten Eintrag in der ALT-Spalte usw. Der Schrägstrich (/) gibt an, dass keine Phasierungsinformationen vorhanden sind. • NC: Anteil der Basen, bei denen kein Call erfolgte oder deren Base-Call-Qualität unter dem Mindestwert lag. • NL: Noise Level (Rauschpegel) – geschätzte Rauschstärke beim Base-Calling an dieser Position. • PB: Probe Pool Bias (Sondenpool-Verzerrung). Werte gegen 0 weisen auf eine größere Verzerrung eines Sondenpools und auf die geringere Zuverlässigkeit eines Varianten-Calls hin. • SB: Strand Bias (Strangverzerrung) an dieser Position. Höhere negative Werte geben eine geringere Verzerrung und Werte nahe 0 eine höhere Verzerrung an. • VF: Variant frequency (Variantenhäufigkeit). Der Prozentsatz der Reads, die das alternative Allel unterstützen.
SAMPLE (PROBE)	Die Probenspalte enthält die Werte, die in der Spalte „FORMAT“ angegeben sind.

Genom-VCF-Dateien

Genom-VCF-Dateien (gVCF) sind Dateien im Format VCF v4.1, die mehrere Konventionen befolgen, um alle Bereiche im Genom in einem angemessenen kompakten Format darzustellen. gVCF-Dateien (*.genome.vcf.gz) erfassen für jede Probe alle Bereiche der Region von Interesse in einer einzigen Datei.

Die gVCF-Datei weist No-Calls an Positionen aus, die nicht alle Filter passieren. Das Genotyp-Tag (GT) **./.** gibt einen No-Call an.

Weitere Informationen finden Sie unter sites.google.com/site/gvcftools/home/about-gvcf.

VCF-Dateien pro Pool und Konsensus-VCF-Dateien

Der Workflow für somatische Varianten erzeugt zwei Sets von Varianten-Call-Dateien.

- **Pro-Pool-VCF-Dateien:** Enthalten Varianten-Calls entweder aus dem Vorwärts- oder aus dem Rückwärts-Pool. Die Pro-Pool-Dateien werden im Ordner „VariantCallingLogs“ gespeichert.
- **Konsensus-VCF-Dateien:** Enthalten Varianten-Calls aus beiden Pools. Konsensus-Dateien werden im Alignment-Ordner gespeichert.

Zu den Pro-Pool-VCF-Dateien und den Konsensus-VCF-Dateien gehören sowohl VCF-Dateien (*.vcf) als auch gVCF-Dateien (*.genome.vcf). Es gelten folgende Namenskonventionen, wobei P# die Reihenfolge repräsentiert, in der die Proben für den Lauf aufgeführt sind:

- **Berichte für alle Stellen:** ProbenName_P#.genome.vcf
- **Berichte nur für Varianten:** ProbenName_P#.vcf

Die Software vergleicht die Pro-Pool-VCF-Dateien und kombiniert die Daten an jeder Position, um eine Konsensus-VCF-Datei für die Probe zu erzeugen.

Die Varianten-Calls aus jedem Pool werden unter Verwendung der folgenden Kriterien in Konsensus-VCF-Dateien zusammengeführt.

Kriterien	Result (Ergebnis)
Ein Referenz-Call in jedem Pool	Referenz-Call
Ein Referenz-Call in einem Pool und ein Varianten-Call im anderen Pool	Gefilterter Varianten-Call
Übereinstimmende Varianten-Calls mit ähnlicher Häufigkeit in jedem Pool	Varianten-Call
Übereinstimmende Varianten-Calls mit signifikant unterschiedlicher Häufigkeit in jedem Pool	Gefilterter Varianten-Call
Nicht übereinstimmende Varianten-Calls in jedem Pool	Gefilterter Varianten-Call

Kennzahlen aus jedem Pool werden unter Verwendung folgender Werte zusammengeführt.

Kennzahl	Wert
Tiefe	Addition der Tiefe aus beiden Pools
Variantenhäufigkeit	Gesamtzahl der Varianten geteilt durch die Gesamtabdeckungstiefe
Q-Score	Mindestwert der beiden Pools

Amplikon-Abdeckungsdatei

Für jede Manifestdatei wird eine Amplikon-Abdeckungsdatei generiert. „M#“ im Dateinamen steht für die Manifestnummer, die in der Probentabelle für den Lauf aufgeführt ist.

Jede Datei beginnt mit einem Dateivorspann, der die dem Manifest zugeordneten Proben-IDs enthält. In den drei Spalten unter dem Dateivorspann sind folgende Informationen aufgeführt:

- Target-ID, wie im Manifest aufgelistet.
- Abdeckungstiefe der Reads nach Filterung.
- Gesamtabdeckungstiefe.

Ergänzende Ausgabedateien

Die folgenden Ausgabedateien bieten ergänzende Informationen oder eine Zusammenfassung von Laufergebnissen und Analysefehlern. Zwar sind diese Dateien für die Beurteilung der Analyseergebnisse nicht erforderlich, sie können jedoch für die Fehlerbehebung verwendet werden. Die Dateien befinden sich im Alignment-Ordner, sofern nicht anders angegeben.

Dateiname	Beschreibung
AnalysisLog.txt	Verarbeitungsprotokoll, in dem jeder Schritt während der Analyse des aktuellen Laufordners beschrieben ist. Diese Datei enthält keine Fehlermeldungen. Die Datei befindet sich im Ordner „Alignment“.
AnalysisError.txt	Verarbeitungsprotokoll, in dem alle während der Analyse aufgetretenen Fehler aufgeführt sind. Wenn keine Fehler auftreten, enthält die Datei keine Einträge. Die Datei befindet sich im Ordner „Alignment“.
DemultiplexSummaryF1L1#.txt	Gibt die Demultiplexierungsergebnisse in einer Tabelle mit einer Zeile für jede Platte und einer Spalte für jede Probe an. „#“ ist der Platzhalter für Lane 1, 2, 3 oder 4 der Fließzelle. Die Datei befindet sich im Ordner „Alignment“.
AmpliconRunStatistics.xml	Enthält eine laufspezifische Zusammenfassungsstatistik. Die Datei befindet sich im Ordner „Alignment“.

Analyseordner

Im Analyseordner werden die von der Local Run Manager-Software generierten Dateien gespeichert.

Das Zusammenspiel zwischen dem Ausgabeordner und dem Analyseordner lässt sich wie folgt zusammenfassen:

- Während der Sequenzierung stellt die Echtzeitanalyse (RTA) die bei der Bildanalyse, dem Base-Calling und der Qualitätsbewertung generierten Dateien in den Ausgabeordner.
- Die Echtzeitanalyse kopiert Dateien in Echtzeit in den Analyseordner. Nachdem die Echtzeitanalyse jeder Base für jeden Zyklus einen Qualitäts-Score zugeordnet hat, legt sie in beiden Ordnern die Datei RTAComplete.txt an.
- Wenn die Datei RTAComplete.txt vorhanden ist, wird die Analyse gestartet.
- Während der Ausführung der Analyse erstellt Local Run Manager Ausgabedateien im Analyseordner und kopiert diese anschließend zurück in den Ausgabeordner.

Alignment-Ordner

Jedes Mal, wenn die Analyse erneut in die Warteschlange gestellt wird, erstellt Local Run Manager einen Alignment-Ordner mit dem Namen **Alignment_N**, wobei N eine fortlaufende Zahl ist.

Ordnerstruktur

- **Alignment:** Enthält *.bam-, *.vcf- und FASTQ-Dateien sowie analysemodulspezifische Dateien.
  - **Date and Time Stamp:** Datum und Uhrzeit der Analyse in folgendem Format: JJJJMMTT_HHMMSS

- AnalysisError.txt
- AnalysisLog.txt
- AmpliconRunStatistics.xml
- Sample1.genome.vcf.gz
- Sample1.coverage.csv
- Sample1.report.pdf
- Sample1.summary.csv
- Sample1.vcf.gz
- Sample1.bam
- FASTQ
  - Sample1
  - Stats
    - DemuxSummaryF1L1.txt
    - FastqSummaryF1L1.txt

Data

- Intensities
  - BaseCalls
    - **L001:** Enthält einen Unterordner pro Zyklus mit *.bcl-Dateien.
  - **L001:** Enthält *.locs-Dateien, eine Datei pro Platte.
- **RTA Logs:** Enthält Protokolldateien von der RTA-Software-Analyse.

InterOp: Enthält Binärdateien, die zur Übermittlung der Sequenzierungslaufkennzahlen verwendet werden.

Logs: Enthält Protokolldateien, in denen die bei der Sequenzierung durchgeführten Schritte aufgeführt sind.

RTAComplete.txt

RunInfo.xml

runParameters.xml

Technische Unterstützung

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

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

Telefonnummern des Illumina-Kundendiensts

Region	Gebührenfrei	Regional
Nordamerika	+1.800.809.4566	
Australien	+1.800.775.688	
Belgien	+32 80077160	+32 34002973
China	400.635.9898	
Deutschland	+49 8001014940	+49 8938035677
Dänemark	+45 80820183	+45 89871156
Finnland	+358 800918363	+358 974790110
Frankreich	+33 805102193	+33 170770446
Großbritannien	+44 8000126019	+44 2073057197
Hongkong	800960230	
Irland	+353 1800936608	+353 016950506
Italien	+39 800985513	+39 236003759
Japan	0800.111.5011	
Neuseeland	0800.451.650	
Niederlande	+31 8000222493	+31 207132960
Norwegen	+47 800 16836	+47 21939693
Schweden	+46 850619671	+46 200883979
Schweiz	+41 565800000	+41 800200442
Singapur	+1.800.579.2745	
Spanien	+34 911899417	+34 800300143
Taiwan	00806651752	
Österreich	+43 800006249	+43 19286540
Andere Länder	+44.1799.534000	

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

Die **Produktdokumentation** steht auf der Illumina-Website im PDF-Format zum Herunterladen zur Verfügung. Gehen Sie zu support.illumina.com, wählen Sie ein Produkt und wählen Sie anschließend **Documentation & Literature** (Dokumentation und Literatur).

Dokument-Nr. 1000000012183 v01 DEU

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

FÜR IN-VITRO-DIAGNOSTIK

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

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