illumina®

TruSight Whole Genome Analysis Application

Product Documentation

ILLUMINA PROPRIETARY
Document # 200049931 v00
April 2024

FOR IN VITRO DIAGNOSTIC USE.

Revision History

Document	Date	Description of Change
Document # 200049931 v00	April 2024	Initial release.

The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT(S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY, AND WILL VOID ANY WARRANTY APPLICABLE TO THE PRODUCT(S).

ILLUMINA DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE IMPROPER USE OF THE PRODUCT(S) DESCRIBED HEREIN (INCLUDING PARTS THEREOF OR SOFTWARE).

Table of Contents

Overview

The TruSight Whole Genome Analysis Application is used to plan sequencing runs for TruSight Whole Genome and automatically initiate analysis after the run completes. Analysis includes demultiplexing, FASTQ generation, read mapping, alignment to the graph-enabled GrCh38/hg38 human reference genome, and variant calling using the Illumina DRAGEN Server for NovaSeq 6000Dx.

At different stages of the analysis workflow, the application performs quality control (QC) according to defined sequencing, FASTQ, and sample library metrics, and generates reports with the results. For samples that pass all QC steps, the application generates supporting output files for use in downstream germline applications.

The TruSight Whole Genome Analysis Application executes DRAGEN variant callers, including the Small Variant Caller, Copy Number Variant (CNV) Caller, and Repeat Expansion Detection with ExpansionHunter.

The application also performs annotation of low, intermediate, or high confidence tier for small variants and includes this annotation in the output file.

Getting Started

Make sure the TruSight Whole Genome Analysis Application is installed on the NovaSeq 6000Dx instrument that will be used for sequencing as part of TruSight Whole Genome. Installed applications can be found on the Applications screen on the NovaSeq 6000Dx Instrument or in Illumina Run Manager using a browser on a networked computer. For assistance to schedule installation, contact your local Illumina Field Representative.

Data Storage Requirements

Refer to the NovaSeq 6000Dx Product Documentation (document # 200010105) and DRAGEN Server for NovaSeq 6000Dx Product Documentation (document # 200014171) for general information on data output and storage.

The TruSight Whole Genome Analysis Application outputs data into a Run Folder and an Analysis Folder in external storage. The minimum storage requirements can be approximated from the size of data output into each folder for a single sequencing run shown below.

Configuration	Run Folder (GB)	Analysis Folder (GB)
S2 Flow Cell (6 samples)	~430	~350
S4 Flow Cell (16 samples)	~1110	~890

Approximate Analysis Time

Analysis begins automatically after a sequencing run is completed and occurs sequentially on samples within a run. Data output files will be available on the external storage once analysis is complete for all samples in a run and copy transfer to the external storage is complete. When starting a sequencing run on both side A and side B at the same time, sequencing will be performed concurrently. Analysis of these sequencing runs will be performed sequentially by the TruSight Whole Genome Analysis Application after sequencing completes. The run that completes sequencing and transfer first will be analyzed first. The second sequencing run will be transferred and queued for analysis after the first analysis completes. Refer to *View Run and Results* on page 6 for how to determine status of active or failed runs.

Approximate time until analysis results are available after sequencing completes is shown below for the situation when side A and side B are loaded simultaneously with the same configuration.

Configuration	Analysis Run 1 (hours)	Analysis Run 2 (hours)
S2 Flow Cell (6 samples)	~12	~24
S4 Flow Cell (16 samples)	~24	~48

Settings

Select the TruSight Whole Genome Analysis Application on the Applications screen to view current configuration and change permissions.

Configuration

The configuration screen displays the following application settings:

- **Application Name**
- **Application Version**
- **DRAGEN Version**
- **RTA Version**
- **Release Date**
- **Organization**
- **Device Identifier**
- **Production Identifier**
- **Library Prep Kits**— Displays the library prep kit. This setting cannot be changed.
- **Index Adapter Kits**— Displays the index adapter sets available for use.
- **Index Reads**
- **Read Type**

- **Index Lengths**
- **Read Lengths—** Read lengths are set by default when the index set is selected. This setting cannot be changed.

Permissions

The designated administrator has Permissions access and can use the checkboxes on the Permissions screen to manage user access for the TruSight Whole Genome Analysis Application.

For more information regarding permissions and user management, refer to the System Configuration section of the NovaSeq 6000Dx Product Documentation (document # 200010105).

Run Creation

Create new runs in IVD mode either on the instrument or by accessing Illumina Run Manager (IRM) using a browser on a networked computer. To access the instrument remotely, use the address and user account information provided by your Illumina representative. Refer to NovaSeq 6000Dx Product Documentation (document # 200010105) for more information.

Create Run is the recommended method for run planning. Import Sample Sheet is not recommended. The sample sheet files output in run and analysis folders are not suitable for import during run planning.

Create Runs

1. From the Runs screen, select **Create Run**.
2. Select the TruSight Whole Genome Analysis Application, and then select **Next**.
3. On the Run Settings screen, enter a run name. The run name identifies the run from sequencing through analysis.
4. [Optional] Enter a run description to further identify the run. Library Prep kit is set by default as TruSight Whole Genome and cannot be changed.
5. Select the desired TruSight Whole Genome index set from the **Index Adapter Kit** drop-down menu. Read length will be set by default and cannot be changed. (Read 1 and 2 use 151 cycles; Index 1 and 2 use 10 cycles).
6. Enter a Library Tube ID (recommended format as DX1234567-LIB), and then select **Next**.
   If no Library Tube ID is specified at this step, the planned run will need to be selected before loading of sequencing consumables. If the incorrect Library Tube ID is entered at this step, the planned run must be corrected before loading consumables. Refer to *Run Revision* on page 5 for protocol to correct run when ready to load consumables.
7. On the Sample Data and Sample Settings screens, sample information will be entered. Sample data can be entered manually or by importing a sample data file. The sample ID must be unique for each sample and can only contain alphanumeric characters, underscores, and dashes. Do not include spaces. Well Position refers to the well in format A01 to H04 of the index plate. Index sequence information will be populated automatically when index plate Well Position is entered. Sex must be entered as Male, Female, or Unknown. Library Plate ID and Library Well ID (eg, format A01) are required fields.
   - To enter sample data manually, add rows (to a total of 6 for S2 or 16 for S4 flow cell) and enter required information into Sample ID and Well Position Fields. Information may also be copied and pasted from Excel. Select **Next**. On the Sample Settings screen, enter Library Plate ID, Library Well ID, and Sex. Select **Next**.

- To import a sample data file, select **Import Samples** and upload the sample data file. Information will be populated into rows automatically. A template (*.csv) is available for download on this screen. Select **Next**. On the Sample Settings screen, information will be populated into rows automatically from the imported sample data file. Select **Next**.

8. On the Analysis Settings screen, enter the Batch Name recorded during batch and run planning.
9. [Optional] Select the Flow Cell Type, S2 or S4.
10. Confirm or deselect the checkbox to Generate ORA compressed FASTQs, then select **Next**.

NOTE The TruSight Whole Genome Analysis Application generates ORA compressed FASTQs by default. Changing this setting will increase size of final data output.

11. On the Run Review screen, review the information entered. If no changes are needed, select **Save**. If changes are needed, select **Back** as needed to return to the appropriate screen.

⚠ CAUTION

TruSight Whole Genome has been validated for 6 samples when using the NovaSeq 6000Dx S2 flow cell, and 16 samples when using the NovaSeq 6000Dx S4 flow cell. Ensure the correct number of samples are entered for the selected flow cell configuration.

Run Revision

If changes are required after run creation and before loading consumables for sequencing, revise runs in IVD mode either on the instrument or by accessing Illumina Run Manager (IRM) using a browser on a networked computer.

1. Select **Runs**.
2. Select the Run name on the Planned Runs tab.
3. Select **Edit**.
4. Update the run or sample information as needed. For example, enter or correct the Library Tube ID to match that which was used when completing the workflow.
5. Select **Next** up to Run Review.
6. Select **Save**.
7. Select **Exit**.

Return to Sequencing in IVD mode to repeat loading of consumables. Run should now be automatically highlighted.

If updating Library Tube ID while loading consumables, return to Run Selection in Control Software and select **Refresh** for the associated column, A or B. Run should now be automatically highlighted. If not, select **Back** to repeat Load consumables.

Requeue Analysis

Refer to the Troubleshooting section in the TruSight Whole Genome Package Insert (document # 200050132) to determine which type of Requeue Analysis is most appropriate.

Requeue analysis with no changes

1. Select the Completed Run name to view Run Details.
2. Select **Requeue Analysis**.
3. Select **Requeue Analysis with no changes**.
4. Provide details in the Reanalysis Reason field.
5. Select **Requeue Analysis**.
6. Exit the page and navigate to the Active Runs page to confirm the requeue is in progress.

Requeue analysis with changes

1. Select the Completed Run name to view Run Details.
2. Select **Requeue Analysis**.
3. Select **Edit run settings** and **Requeue Analysis**.
4. Provide details in the Reanalysis Reason field.
5. Select **Requeue Analysis**.
6. Confirm or update the Run Settings, then select **Next**.
7. Correct the sample information as necessary by manually updating the fields or select **Download Template** to create a `sampledata.csv` file with current information. Correct information and delete existing rows in the Sample Data tab before using Import Samples to populate the corrected sample data.
8. Review the information on the Run Review screen and select **Save** to start reanalysis.
9. Select **Exit** and navigate to the Active Runs page to confirm the requeue is in progress.
   The original run data folder must be present at the external storage location specified in Run Details for reanalysis to complete successfully. If reanalysis fails, make sure the run has not been moved or deleted.

View Run and Results

1. From the Illumina Run Manager main screen in IVD mode, select **Runs**.
2. From the Completed Runs tab, select the Run name.
   This tab will also display runs that have completed due to failure of sequencing, data transfer, or analysis. Active runs and their status are displayed in the Active Runs tab. Refer to NovaSeq 6000Dx Product Documentation (document # 200010105) for more information.

3. Select the Run name in the Completed Runs tab to view Run Details and Results for the path to the Analysis Output Folder.

   For failed runs, review the Status for each step and then refer to the Troubleshooting section of the TruSight Whole Genome Package Insert (document # 200050132).

4. Navigate to the analysis folder on your local drive and open the Consolidated Report to review the PASS/FAIL result for each QC step as follows:

   - For sequencing run QC, refer to Summary Sequencing QC Result
   - For FASTQ QC for each sample in the run, refer to Summary FASTQ QC Result
   - For library QC for each sample in the run, refer to Summary Sample Library QC Result

   If a FAIL result is observed, note the QC step and refer to the Troubleshooting section of the TruSight Whole Genome Package Insert (document # 200050132).

Output File Summary

The TruSight Whole Genome Analysis Application saves the following main output files. Refer to file information sections below for location of main output files.

Runs and samples which do not pass validity criteria do not produce CRAM, ROH bed, or *genome.vcf) files.

Output File	Description
Consolidated Report (*.csv)	Contains quality metrics used for run validity (including total yield and Q30), sample validity metrics (including FASTQ yield), library QC metrics, and For Information Only (FIO) metrics for all of the samples in the run.
Sample Report (*.csv)	Contains results of sequencing QC, FASTQ, and sample library QC. The report also contains ploidy concordance and FIO metrics for the individual sample as well as the associated sequencing run.
Small variant and mSNV VCF (*.annotated.hard-filtered-gvcf.gz)	Contains variant call information for small variants (SNVs, indels) and mitochondrial SNVs.
CNV VCF (*.cnv.vcf.gz)	Contains variant call information for copy number gains and losses.
Repeats VCF (*.repeats.vcf.gz)	Contains variant call information on STR expansions and SMN1.
ROH BED (*.roh.bed)	Contains information for regions of homozygosity.
FASTQ (*.fastq.gz or *.fastq.ora)	Intermediate files containing quality scored base calls. FASTQ files are the primary input for the alignment step. If ORA compression is selected, the file name reflects this.
Alignment CRAM (*.cram)	Contains aligned reads for a given sample.

QC Report Information

The Consolidated Report `<<RunID>>_Consolidated_Report.csv` is located in the `TruSightWholeGenomeAnalysis_x.x.x_run-complete` directory and contains information about quality metrics used to pass or fail samples at different stages of analysis. Individual sample reports `<<Sample_ID>>_Sample_Report.csv` may be found within the <Sample_ID> folders in the `TruSightWholeGenomeAnalysis_x.x.x_run-complete` directory.

The report headers include the following information about the run: the app version, batch name, library pool tube ID, sequencing run name, sequencing run ID, and flow cell type. The following tables describe the information included in the Consolidated Report. The individual Sample Report includes the same information except for the Demultiplex Metrics.

Table 1 Sequencing QC Metrics

Metrics	Spec	Description
Non-Indexed Total Yield (GB)	N/A	No specification since lower yield runs may result in passing sample libraries. Expect ≥ 3000 Gbp for S4 and ≥ 1000 GB for S2 flow cell.
Total % ≥ Q30	≥ 85	Measure of base quality at the run level. Minimum specification is set since too low %Q30 runs will not pass Q30 bases in Sample Library QC.
Summary Sequencing QC Result	PASS or FAIL	For Sequencing QC failure, consult Troubleshooting section in the TruSight Whole Genome Package Insert (document # 200050132).

Table 2 Demultiplex Metrics

Metrics	Spec	Description
Percent reads identified	N/A	Total fraction of passing filter reads in the run that were assigned to samples during demultiplexing.
Percent CV	N/A	Provides a measure of the evenness of reads demultiplexed to each index pair on the run. Expect < 25% for runs without FASTQ QC Result failures.

Table 3 FASTQ QC Metrics

Metrics	Spec	Description
Yield per sample (bps)	≥ 90,000,000,000	Minimum is set to be equivalent to ~26x average autosomal coverage in order to triage sample libraries that will not pass QC to reduce analysis time.
Summary FASTQ QC Result	PASS or FAIL	For FASTQ QC failure, consult Troubleshooting section in the TruSight Whole Genome Package Insert (document # 200050132).

Table 4 Sample Library QC Metrics

Metrics	Spec	Description
Average autosomal coverage	≥ 35	Average coverage across the autosomes. Minimum specification is set to ensure analytical performance.
Percent of autosomes with coverage greater than 20X	≥ 93.94	Measure of coverage uniformity that detects issues not necessarily related to GC bias. Minimum specification is set to ensure analytical performance.
Normalized coverage at 60% to 79% GC bins	0.82 ≤ x ≤ 1.13	Measure of coverage uniformity that detects GC bias, specifically a loss of coverage in areas of the genome with higher % GC and lower % AT base composition. Minimum and maximum specifications are set to ensure analytical performance.
Normalized coverage at 20% to 39% GC bins	0.97 ≤ x ≤ 1.06	Measure of coverage uniformity that detects GC bias, specifically a loss of coverage in areas of the genome with lower % GC and higher % AT base composition. Minimum and maximum specifications are set to ensure analytical performance.
Average mitochondrial coverage	≥ 500	Coverage of the mitochondrial chromosome. Minimum specification is set to ensure mitochondrial SNV limit of detection.
Percent Q30 bases	≥ 85	Measure of base quality. Minimum specification is set to ensure analytical performance.
Estimated sample contamination	≤0.005	Detects contaminating reads from other samples. Maximum specification is set to ensure mitochondrial SNV limit of detection (the variant type with the highest sensitivity to contamination).
Summary Sample Library QC Result	PASS or FAIL	For Sample Library QC failure, consult Troubleshooting section in the TruSight Whole Genome Package Insert (document # 200050132).

Table 5 Ploidy QC Metrics

Metrics	Spec	Description
Provided sex chromosome ploidy	N/A	Sex provided by the operator during Run Creation (Female, Male, Unknown).

Metrics	Spec	Description
Ploidy estimation	N/A	Sex ploidy estimated by DRAGEN.
Summary Ploidy Result	CONCORDANT, DISCORDANT, or ND	CONCORDANT indicates agreement between provided and estimated sex ploidy. ND indicates sex provided as Unknown or estimation other than XX or XY. For DISCORDANT results, either wrong sex was entered during run creation or sample swap may have occurred. Consult the Troubleshooting section in the TruSight Whole Genome Package Insert (document # 200050132).

Table 6 For Information Only Metrics

Metrics	Description
Insert length median	Target is 450 bp but expect variation by sequencing run and operator. A range of approximately 360 to 550 bp is acceptable. Consistently operating outside of this range may lead to a higher incidence of sample failure.
Percent mapped reads	Percent of reads that map to the reference genome. May be decreased in response to non-human gDNA contamination, poor sample quality or too small insert length leading to mapping issues.
Percent reads with supplementary alignments	Percent of reads with mapping that splits across different locations in the reference genome.
Percent duplicate marked reads	Expect < 20%. May be elevated if library prep yield is low or pooling less than required volume, or in response to sequencing related issues.
Percent soft clipped bases Read 1	Useful in diagnosing root cause for failure of average autosomal coverage.
Percent soft clipped bases Read 2	Useful in diagnosing root cause for failure of average autosomal coverage.
Percent bases trimmed Read 1	Useful in diagnosing root cause for failure of average autosomal coverage.
Percent bases trimmed Read 2	Useful in diagnosing root cause for failure of average autosomal coverage.

Variant Call File Information

VCF Files

Variant call format (*.vcf) files contain information about variants found at specific positions in a reference genome and can be found in the `<Sample_ID>/Analysis` directory.

The VCF file header includes the VCF file format version and the variant caller version and lists the annotations used in the remainder of the file. The last line in the header contains the column headings for the data lines. Each of the VCF file data lines contains information about a single reference position.

All VCF files contain a header with descriptions of output columns, and variant call data in columns labeled as CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, SAMPLE. Definitions of column values can vary among variant callers.

Small Variant and mSNV VCF

Output is saved under the `<Sample_ID>.annotated.hard-filtered.gvcf.gz` file in the `<Sample_ID>/Analysis` directory.

A genomic VCF (gVCF) file contains information on variants and positions determined to be homozygous to the reference genome. For homozygous regions, the gVCF file includes statistics that indicate how well reads support the absence of variants or alternative alleles. The gVCF file includes an artificial <NON_REF> allele. Reads that do not support the reference or any variants are assigned the <NON_REF> allele. DRAGEN uses these reads to determine if the position can be called as a homozygous reference, as opposed to remaining uncalled. The resulting score represents the Phred-scaled level of confidence in a homozygous reference call. In germline mode, the score is FORMAT/GQ.

DRAGEN provides post-VCF variant filtering based on annotations present in the VCF records. Variant hard filtering is described below. However, due to the nature of DRAGEN's algorithms, which incorporate the hypothesis of correlated errors from within the core of variant caller, the pipeline has improved capabilities in distinguishing the true variants from noise, and therefore the dependency on post-VCF filtering is substantially reduced.

The TruSight Whole Genome Analysis Application provides annotation of confidence score and confidence tier for small variants that can be used to further improve performance. Confidence tier annotation is not a quality filter and as such is not directly reflected in the quality status of the variant calls. Therefore, it is possible to see passing variant calls which are nonetheless annotated as low confidence.

Table 7 VCF File Headings

Heading	Description
CHROM	The chromosome of the reference genome. Chromosomes appear in the same order as the reference FASTA file.

Heading	Description
POS	The single-base position of the variant in the reference chromosome. For single nucleotide variants (SNVs), this position is the reference base with the variant. For indels, this position is the reference base immediately preceding the variant.
ID	Always .
REF	The reference genotype. For example, a deletion of a single T is represented as reference TT and alternate T. An A to T single nucleotide variant is represented as reference A and alternate T.
ALT	The alleles that differ from the reference read. For example, an insertion of a single T is represented as reference A and alternate AT. An A to T single nucleotide variant is represented as reference A and alternate T.
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant and lower probability of errors. For a quality score of Q, the estimated probability of an error is 10-(Q/10). For example, the set of Q30 calls has a 0.1% error rate. Many variant callers assign quality scores based on their statistical models, which are high in relation to the error rate observed.

Table 8 VCF File Annotations

Heading	Description
FILTER	• PASS: All filters passed. • **DRAGENSnpHardQUAL**—Set if true:QUAL < 10.41 • **DRAGENIndelHardQUAL**—Set if true:QUAL < 7.83 • **LowDepth**—Set if true:DP <= 1 • **LowGQ**—Set if true:GQ = 0 • **PloidyConflict**—Genotype call from variant caller not consistent with chromosome ploidy • **base_quality**—Site filtered because median base quality of alt reads at this locus does not meet threshold • **filtered_reads**—Site filtered because too large a fraction of reads has been filtered out • **fragment_length**—Site filtered because absolute difference between the median fragment length of alt reads and median fragment length of ref reads at this locus exceeds threshold • **low_af**—Allele frequency does not meet threshold • **low_depth**—Site filtered because the read depth is too low • **low_frac_info_reads**—Site filtered because the fraction of informative reads is below threshold • **low_normal_depth**—Site filtered because the normal sample read depth is too low • **long_indel**—Site filtered because the indel length is too long • **mapping_quality**—Site filtered because median mapping quality of alt reads at this locus does not meet threshold • **multiallelic**—Site filtered because more than two alt alleles pass tumor LOD • **non_homref_normal**—Site filtered because the normal sample genotype is not homozygous reference • **no_reliable_supporting_read**—Site filtered because no reliable supporting somatic read exists • **panel_of_normals**—Seen in at least one sample in the panel of normals vcf • **read_position**—Site filtered because median of distances between start/end of read and this locus is below threshold • **RMxNRepeatRegion**—Site filtered because all or part of the variant allele is a repeat of the reference • **strand_artifact**—Site filtered because of severe strand bias • **str_contraction**—Site filtered due to suspected PCR error where the alt allele is one repeat unit less than the reference • **too_few_supporting_reads**—Site filtered because there are too few supporting reads in the tumor sample • **weak_evidence**—Somatic variant score does not meet threshold

Heading	Description
INFO	• **DB**—dbSNP Membership. • **FS**—Phred-scaled p-value using Fisher's exact test to detect strand bias. • **QD**—Variant Confidence/Quality by Depth. • **R2_5P_bias**—Score based on mate bias and distance from 5 prime end. • **SOR**—Symmetric Odds Ratio of 2x2 contingency table to detect strand bias. • **DP**—Approximate read depth (informative and non-informative); some reads may have been filtered based on mapq etc. • **END**—Stop position of the interval. • **FractionInformativeReads**—The fraction of informative reads out of the total reads. • **MQ**—RMS Mapping Quality. • **MQRankSum**—Z-score from Wilcoxon rank sum test of Alt vs. Ref read mapping qualities. • **ReadPosRankSum**—Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias. • **SOMATIC**—At least one variant at this position is somatic. • **ILENS**—Indel lengths for each ALT variant. • **SCORE**—Confidence score for each variant type present at the site as (variant type):(confidence score). • **TIER**—Confidence tier for each variant type present at the site as (variant type):(confidence tier).

Heading	Description
FORMAT	The FORMAT column lists fields separated by colons, for example GT:GQ. • **AD**—Allelic depths (counting only informative reads out of the total reads) for the ref and alt alleles in the order listed. • **AF**—Allele fractions for alt alleles in the order listed. • **DP**—Approximate read depth (reads with MQ=255 or with bad mates are filtered). • **F1R2**—Count of reads in F1R2 pair orientation supporting each allele. • **F2R1**—Count of reads in F2R1 pair orientation supporting each allele. • **GP**—Phred-scaled posterior probabilities for genotypes as defined in the VCF specification. • **GQ** —Genotype quality. • **GT**—Genotype. • **ICNT**—Counts of INDEL informative reads based on the reference confidence model. • **MB**—Per-sample component statistics to detect mate bias. • **MIN_DP**—Minimum DP observed within the GVCF block. • **PL**—Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification. • **PRI**—Phred-scaled prior probabilities for genotypes. • **PS**—Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group. • **SB**—Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias. • **SPL**—Normalized, Phred-scaled likelihoods for SNPs based on the reference confidence model. • **SQ**—Somatic quality.
SAMPLE	The sample column gives the values specified in the FORMAT column.

Copy Number Variant VCF

The target counts stage is the first processing stage for the DRAGEN CNV pipeline, producing `<Sample_ID>.target.counts.gz`, then GC Bias Correction is performed, generating a `*.target.counts.gc-corrected.gz` file. Normalization stage produces `*.tn.tsv.gz` file. The DRAGEN Host Software generates many intermediate files. `*.seg.called.merged` is the final call file that contains the amplification and deletion events.

In addition to the segment file, DRAGEN emits the calls in the standard VCF format. Output is saved in `<Sample_ID>.cnv.vcf.gz` in the `<Sample_ID>/Analysis` directory.

Definitions of columns specific to CNV caller:

The POS column is the start position of the variant. According to the VCF specification, if any of the ALT alleles is a symbolic allele, such as `<DEL>`, then the padding base is required and POS denotes the coordinate of the base preceding the polymorphism. All coordinates in the VCF are 1-based.

The `ID` column is used to represent the event. The `ID` field encodes the event type and coordinates of the event.

The `REF` column contains an N for all CNV events.

The `ALT` column specifies the type of CNV event. Because the VCF contains only CNV events, only the `DEL` or `DUP` entry is used.

The `QUAL` column contains an estimated quality score for the CNV event, which is used in hard filtering.

The `FILTER` column contains `PASS` if the CNV event passes all filters, otherwise the column contains the name of the failed filter.

The `INFO` column contains information representing the event. The `REFLEN` entry indicates the length of the event. The `SVTYPE` entry is always CNV. The `END` entry indicates the end position of the event.

The FORMAT fields are described in the header.

- `GT`—Genotype
- `SM`—Linear copy ratio of the segment mean
- `CN`—Estimated copy number
- `BC`—Number of bins in the region
- `PE`—Number of improperly paired end reads at start and stop breakpoints

Repeats VCF

ExpansionHunter performs a sequence-graph based realignment of reads that originate inside and around each target repeat. ExpansionHunter then genotypes the length of the repeat in each allele based on these graph alignments.

More information and analysis are available in the following ExpansionHunter papers:

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

The TruSight Whole Genome Analysis Application STR variant catalog contains specifications on disease-causing repeats located in AFF2, AR, ATN1, ATXN1, ATXN10, ATXN2, ATXN3, ATXN7, ATXN8OS, C9ORF72, CACNA1A, CBL, CNBP, CSTB, DIP2B, DMPK, FMR1, FXN, GLS, HTT, JPH3, NIPA1, NOP56, NOTCH2NL, PABPN1, PHOX2B, PPP2R2B, and TBP genes.

The results of repeat genotyping are output as a separate VCF file, which provides the length of each allele at each callable repeat defined in the repeat-specification catalog file. The file name is `<Sample_ID>.repeats.gz` and can be found in the `<Sample_ID>/Analysis` directory.

Some columns are specific to repeat expansion caller:

Table 9 Core VCF Fields

Field	Description
CHROM	Chromosome identifier
POS	Position of the first base before the repeat region in the reference
ID	Always .
REF	The reference base at position POS
ALT	List of repeat alleles in format `<STRn>` . N is the number of repeat units.
QUAL	Always .
FILTER	LowDepth filter is applied when the overall locus depth is below 10x or the number of reads that span one or both breakends is below 5.

Table 10 Additional INFO Fields

Field	Description
END	Position of the last base of the repeat region in the reference
REF	Reference copy number
REPID	Variant ID from the variant catalog
RL	Reference length in bp
RU	Repeat unit in the reference orientation
VARID	Variant ID from the variant catalog

Table 11 GENOTYPE (Per Sample) Fields

Field	Description
AD	Allelic depths for the ref and alt alleles in the order listed
ADFL	Number of flanking reads consistent with the allele
ADIR	Number of in-repeat reads consistent with the allele
ADSP	Number of spanning reads consistent with the allele
DST	Results (+ detected, – undetected, ? undetermined) of the test represented by the variant
GT	Genotype
LC	Locus Coverage
REPCI	Confidence interval for REPCN

Field	Description
REPCN	Number of repeat units spanned by the allele
RPL	Log-Likelihood ration for the presence of the reference allele
SO	Type of reads that support the allele. Values can be `SPANNING`, `FLANKING`, or `INREPEAT`. These values indicate if the reads span, flank, or are fully contained in the repeat.

The `<Sample_ID>.repeats.vcf.gz` file includes SMN output along with any targeted repeats. SMN output is represented as a single SNV call at the splice-affecting position in SMN1 (NM_000344.3:c.840C/T) with Spinal Muscular Atrophy (SMA) status in the following custom fields.

Table 12 SMA Results in repeats.vcf Output File

Field	Description
VARID	SMN marks the SMN call.
GT	Genotype call at this position using a normal (diploid) genotype model.
DST	SMA status call: + indicates detected - indicates undetected ? indicates undetermined
AD	Total read counts that support the C and T allele.
RPL	Log10 likelihood ratio between the unaffected and affected models. Positive scores indicate the unaffected model is more likely.

ROH BED

Regions of homozygosity (ROH) are detected as part of the small variant caller. The caller detects and outputs the runs of homozygosity from whole genome calls on autosomal human chromosomes. Sex chromosomes are ignored unless the sample sex karyotype is XX, as determined by the Ploidy Estimator. ROH output allows downstream tools to screen for and predict consanguinity between the parents of the proband subject.

A region is defined as consecutive variant calls on the chromosome with no large gap in between these variants. In other words, regions are broken by chromosome or by large gaps with no SNV calls. The gap size is set to 3 Mbases.

The ROH caller produces an ROH output file named `<Sample_ID>.roh.bed` in the `<Sample_ID>/Analysis` directory. Each row represents one region of homozygosity. The bed file contains the following columns:

```
Chromosome Start End Score #Homozygous #Heterozygous
```

Where

- Score is a function of the number of homozygous and heterozygous variants, where each homozygous variant increases the score by 0.025, and each heterozygous variant reduces the score by 0.975.
- Start and end positions are a 0-based, half-open interval.
- #Homozygous is number of homozygous variants in the region.
- #Heterozygous is number of heterozygous variants in the region.

The caller also produces a metrics file named `<Sample_ID>.roh_metrics.csv` that lists the number of large ROH and percentage of SNPs in large ROH (> 3 MB).

Ploidy Estimation Metrics

The Ploidy Estimator runs by default. The Ploidy Estimator uses reads from the mapper/aligner to calculate the sequencing depth of coverage for each autosome and allosome in the human genome. The sex karyotype of the sample is then estimated using the ratios of the median sex chromosome coverages to the median autosomal coverage. XX or XY, and CONCORDANT, DISCORDANT, or ND (Not Determined) compared to the sample data provided are reported in the consolidated report. The detailed results, including each normalized per-contig median coverage, is reported in the `<Sample_ID>.ploidy_estimation_metrics.csv` file.

FASTQ Files

FASTQ (*.fastq.gz, *.fastq.ora) is a text-based file format containing base calls and quality values per read. Each file contains the following information:

- The sample identifier
- The sequence
- A plus sign (+)
- The Phred quality scores in an ASCII + 33 encoded format

The software generates one FASTQ file for every sample, read, and lane. For example, for each sample in a paired-end run, the software generates two FASTQ files: one for Read 1 and one for Read 2. In addition to these sample FASTQ files, the software generates two FASTQ files per lane containing all unknown samples. FASTQ files for Index Read 1 and Index Read 2 are not generated because the sequence is included in the header of each FASTQ entry. The file name format is constructed from fields specified in the sample sheet and use the file naming format of `<Sample_ID>_S#_L00#_R#_001.fastq.gz`

FASTQ files are saved in `<Sample_ID>/Conversion` directory. In the FASTQ directory of the analysis folder, one can find the Logs directory with BCL-to-FASTQ conversion logs, and the Reports directory which contains various read metrics files, and `SampleSheet.csv` used for FASTQ conversion. FASTQ files from undetermined reads are found in the `Undetermined/Conversion` directory of the Analysis folder.

The sample identifier is formatted as follows:

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

CRAM Files

Compressed Reference-oriented Alignment Map or CRAM files (*.cram) are stored in the `<Sample_ID>/Analysis` directory and contain headers and alignment records relative to the genomic reference file used during alignment. The path to the reference file is listed in the `<Sample_ID>/Analysis/<Sample_ID>-replay.json` file, as an `--ht-reference` parameter, by default set to `hg38.fa`.

CRAM files contain a header section and an alignment section:

- **Header**—Contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.
- **Alignments**—Contains read name, read sequence, read quality, alignment information, and custom tags. The read name includes the chromosome, start coordinate, alignment quality, and the match descriptor string.

The alignments section includes the following information for each read or read pair:

- AS: Paired-end alignment quality.
- RG: Read group, which indicates the number of reads for a specific sample.
- BC: Barcode tag, which indicates the demultiplexed sample ID associated with the read.
- SM: Single-end alignment quality.
- XC: Match descriptor string.
- XN: Amplicon name tag, which records the amplicon ID associated with the read

To view alignment records, `samtools` can be used as `samtools view --reference <path_to_reference_folder>/hg38.fa <Sample_ID>.cram`.

An index file and checksum file are also generated.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Website:	www.illumina.com
Email:	techsupport@illumina.com

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download from support.illumina.com.

Illumina, Inc.
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com

CE

IVD

EC REP

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

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

FOR IN VITRO DIAGNOSTIC USE.

illumina®