illumina®

DRAGEN TruSight Oncology 500 Analysis Software v1.1 (Local)

User Guide

ILLUMINA PROPRIETARY
Document # 200014849 v00
April 2022

For Research Use Only. Not for use in diagnostic procedures.

This document and its contents are proprietary to Illumina, Inc. and its affiliates ("Illumina"), and are intended solely for the contractual use of its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina. Illumina does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document.

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

For Research Use Only. Not for use in diagnostic procedures.

Overview

The Illumina® DRAGEN™ TruSight™ Oncology 500 Analysis Software v1.1 supports local analysis for DNA and RNA libraries generated from formalin-fixed, paraffin-embedded (FFPE) tissue samples. The TruSight Oncology 500 assay is optimized to provide high sensitivity and specificity for low-frequency somatic variants across coding exons and additional regions of biological relevance in 523 genes for DNA biomarkers. DNA biomarkers include the following:

- Single nucleotide variants (SNVs)
- Insertions
- Deletions
- Copy number variants (CNVs)
- Multinucleotide variants (MNVs)

TruSight Oncology 500 also detects immunotherapy biomarkers for tumor mutational burden (TMB) and microsatellite instability (MSI) in DNA. DNA library analysis outputs include TMB, variant call files for small and complex variants, MSI, and gene amplifications. Fusions and splice variants are detected in RNA of 55 genes and the RNA library analysis outputs include fusions and splice variant call files.

Details of the regions covered can be found in the assay manifest file, available on request from your local Illumina representative.

The DRAGEN TruSight Oncology 500 Analysis Software v1.1 allows for analysis on a single DRAGEN server or split across multiple servers.

Compatibility

The DRAGEN TruSight Oncology 500 Analysis Software v1.1 support pages on the Illumina support website provide information on compatibility with Illumina sequencing systems.

Use BCL Convert to produce FASTQ files for DRAGEN TruSight Oncology 500 Analysis Software v1.1. Using bcl2fastq does not produce the same results and is discouraged. See the DRAGEN TruSight Oncology 500 Analysis Software v1.1 support pages on the Illumina support website for settings and compatibility information for using DRAGEN TruSight Oncology 500 Analysis Software v1.1 with BCL Convert.

Additional Resources

The DRAGEN TruSight Oncology 500 Analysis Software support pages on the Illumina support website provide additional resources. These resources include software, training, compatible products, sample sheets, and the following documentation. Always check support pages for the latest versions.

Document	Description
TruSight Oncology 500 Reference Guide (document # 1000000067621)	Information on using the TruSight Oncology 500 kit.

Installation Requirements

The DRAGEN TruSight Oncology 500 Analysis Software is compatible with Illumina DRAGEN Server v3.

Hardware

- The DRAGEN TruSight Oncology 500 Analysis Software only runs on the DRAGEN server.
- Assay pipeline requires that mkfifo is enabled on the network attached storage (NAS).

Software

- By default Linux CentOS 7.3 operating system, or later, is provided.
- Before installing DRAGEN TruSight Oncology 500 Analysis Software, container engine Docker v18.09 or later is required. Use the install instructions for CentOS provided in the Docker documentation.

Storage Requirements

For optimal performance, run analysis on data stored locally on the DRAGEN server. Analysis of data stored on NAS can take longer and performance can be less reliable.

The DRAGEN server provides a NVMe SSD located in the `/staging` directory to use as the software output directory. Network-attached storage is required for long-term storage.

When running the DRAGEN TruSight Oncology 500 Analysis Software, set the `--analysisFolder` command-line option to a directory in `/staging` to make sure the DRAGEN server processes read and write data on the NVMe SSD.

Before beginning analysis, develop a strategy to copy data from the DRAGEN server to a network-attached storage. Delete output data on the DRAGEN server as soon as possible.

The following are the run and analysis output sizes for each sequencing system per 101 bp:

Sequencing System	Run Folder Output (Gb)	Analysis Output (Gb)
NovaSeq 6000 SP flow cell	85–100	250–374
NovaSeq 6000 S1 flow cell	164–200	360–665
NovaSeq 6000 S2 flow cell	290–460	890–1600
NovaSeq 6000 S4 flow cell	800–1200	2700–4100

For Research Use Only. Not for use in diagnostic procedures.

Sequencing System	Run Folder Output (Gb)	Analysis Output (Gb)
NextSeq 500/550 and 550Dx HO flow cell	32–55	82–85

If adequate disk space is not available at analysis run time, the run does not complete.

Install DRAGEN TruSight Oncology 500 Analysis Software

Illumina recommends running Docker as a non-root user by adding the user to the Docker group. It is possible to run the DRAGEN TruSight Oncology 500 Analysis Software as root but not recommended. For information on Docker permission requirements and alternatives to running as root, see the Post-Installation steps for Linux page in the Docker documentation. You might require sudo or escalated privileges to load the docker image.

Installing the DRAGEN TruSight Oncology 500 Analysis Software requires root privileges. The installation script uninstalls any existing DRAGEN software on the server. To use a different DRAGEN pipeline, uninstall the DRAGEN TruSight Oncology 500 Analysis Software and download a DRAGEN software installation package from the Illumina DRAGEN Bio-IT Platform support page. To uninstall DRAGEN TruSight Oncology 500 Analysis Software, refer to *Uninstall DRAGEN TruSight Oncology 500 Analysis Software* on page 4.

Installation Instructions

1. Contact Illumina Customer Support to obtain the DRAGEN TruSight Oncology 500 Analysis Software installer package.
2. Install Docker 18.09 or later using the install instructions for CentOS provided in the Docker documentation.
3. Download the DRAGEN TSO 500 installation script provided in the email from Illumina. The link expires after 7 days.
4. Copy the install script to the `/staging` directory to store it in that location.
5. Use the following command to update the run script permission:

   ```
   sudo chmod +x /staging/install_DRAGEN_TSO500-<version>.run
   ```

6. Use the following command to run the installation script, which runs for approximately 10 minutes:

   ```
   sudo TMPDIR=/staging /staging/install_DRAGEN_TSO500-<version>.run
   ```

   The script removes any previously installed DRAGEN server software. During the installation process, you might be instructed to reboot or power cycle the system, which is required to complete the installation of the DRAGEN server FPGA hardware. A power cycle of the system requires the server be shut down and restarted.
7. Use the following command to install the DRAGEN server license:

For Research Use Only. Not for use in diagnostic procedures.

```
sudo /opt/edico/bin/dragen_lic -i auto
```

8. Use the following command to build the DRAGEN server hash table, which runs for approximately 20 minutes:

```
sudo /usr/local/bin/build-hashtable_DRAGEN_TSO500-<version>.sh
```

Running the System Check

Make sure the system functions properly by running the following command:

```
sudo /usr/local/bin/check_DRAGEN_TSO500-<version>.sh
```

The script checks the following functions:

- If all required services are running
- If the proper Docker image is installed
- If the Illumina DRAGEN TruSight Oncology 500 Pipeline successfully runs on a test data set.

The self-test runs for approximately 30 minutes. If the self-test prints a failure message, contact Illumina Technical Support and provide the `/staging/check_DRAGEN_TSO500_<timestamp>.tgz` output file.

When running analysis on DRAGEN server via Secure Shell Protocol (SSH), precede analysis execution commands with the no hang up command `nohup`. This command prevents analysis from terminating in the event disconnection from the DRAGEN server. All output from the analysis command are redirected to `nohup.out` in the directory.

If using MacOS, disable the ability to set environment variables automatically in Terminal settings, as this can cause the following error:

```
ERROR: locale::facet_S_create_c_locale name not valid
```

Uninstall DRAGEN TruSight Oncology 500 Analysis Software

The DRAGEN TruSight Oncology 500 Analysis Software installation includes an uninstall script called `uninstall_DRAGEN_TSO500-<VERSION>.sh`, which is installed in `/usr/local/bin`.

Executing the uninstall script removes the following assets:

- All scripts (`DRAGEN_TSO500.sh`, `test_DRAGEN_TSO500-<VERSION>.sh`, and `uninstall_DRAGEN_TSO500-<VERSION>.sh`).
- The resources found in `staging/illumina/DRAGEN_TSO500`.
- The `dragen_tso500:<VERSION>` Docker image.

To uninstall the DRAGEN TruSight Oncology 500 Analysis Software, run the following command as a root user:

```
uninstall_DRAGEN_TSO500-<version>.sh
```

Do not uninstall Docker or DRAGEN. Docker or DRAGEN can be removed by removing the associated RPM packages.

Running DRAGEN TruSight Oncology 500 Analysis Software

Start the DRAGEN TruSight Oncology 500 Analysis Software with the Bash script called `DRAGEN_TSO500.sh`, which is installed in the `/usr/local/bin` directory. The Bash script is executed on the command line and runs the software with Docker.

For arguments, see *Command-Line Options* on page 8. You can start from BCL files or from the FASTQ folder produced by BCL Convert. The following requirements apply for both methods:

- Path to the sequencing run or FASTQ folder. Copy the run or FASTQ folder to the DRAGEN server into the staging folder with a recommended organization as follows: `/staging/runs/{RunID}`. Copying the run folder onto the DRAGEN server can be done using Linux commands such as `rsync`. The sample sheet within the run folder is used unless otherwise specified through the command line.
- Path to the resource file folder. The location of this folder is pre-configured through the installation. Modifying files in this folder causes an error at run time.
- Path to the hash table. The location of this file is pre-configured through the installation.
- Analysis output folder path. This folder is created and contains output analysis files.

Sample Sheet Requirements

A DRAGEN TruSight Oncology 500 Analysis Software sample sheet is required for each analysis. The sample sheet is a comma-separated values file (*.csv) that contains information to set up and analyze a sequencing run. The sample sheet is made up of a list of samples and their index sequences. The DRAGEN TruSight Oncology 500 Analysis Software supports v1 and v2 sample sheets. See the Illumina support site for more information on the differences between the versions.

Provide the sample sheet during the run setup steps on the sequencing system. See the Illumina support site for the appropriate sample sheet template for your run.

The sample sheet is made up of a list of samples and their index sequences. Delete index IDs you do not require. Different types of sequencing runs may use different index adapters. Use the index IDs included in the DRAGEN TruSight Oncology 500 Analysis Software resource bundle.

Create a Sample Sheet

Use the following steps to create a TruSight Oncology 500 sample sheet.

1. Download the appropriate sample sheet template from the TruSight Oncology 500 pages on the Illumina support site.

2. In the Data section, enter the required parameters for each sample. The following table includes required and optional parameters.

Sample Parameter	Required	Details
Sample_ID	Yes	The Sample_ID is included in the output file names. Sample IDs are not case sensitive. Sample IDs must have the following characteristics: • Unique for the run. • No spaces. • Alphanumeric characters with underscores and dashes. If you use an underscore or dash, enter an alphanumeric character before and after the underscore or dash. Example: Sample1-T5B1_022515. • Cannot be called `all`, `default`, `none`, `unknown`, `undetermined`, `stats`, or `reports`. • It is recommended that the sample ID be based on the pair ID. Example: `<PairID>-DNA`,`<PairID>-RNA`.
Sample_Type	Yes	Enter DNA or RNA.
Sample_Name	No	Name of sample.
Sample_Plate	No	Name of the sample plate.
Sample_Well	No	Name of the sample well.
Pair_ID	Yes	Use to pair DNA and RNA samples from the same individual. Use a shared pair ID to link two samples.
Index adapter sequences	Yes	Enter TruSight Oncology 500 compatible index adapter sequences for samples. Select unique index pairs for each sample. Example indexes are listed in `SampleSheet.csv`.
Index	Yes	Index 1 sequence.
Index2	Yes	Index 2 sequence.
I7_Index_ID	No	I7 index ID.
I5_Index_ID	No	I5 index ID.
Description	No	Description of the sample.
Lane	No	Used for NovaSeq 6000 XP workflows only. Indicates which lane corresponds to a given sample. Enter a single numeric value per row. Cannot be empty if a header is present.

3. Save the sample sheet in the sequencing run folder using one of the following methods:
   - Save the sample sheet with the name `SampleSheet.csv`.
   - Name the sample sheet with the name of your choice, and specify the path to the sample sheet in the command-line options.

Command-Line Options

You can use the following command-line options with DRAGEN TruSight Oncology 500 Analysis Software. For examples, refer to Table 1.

To learn more about the input requirements, use the `--help` command-line option.

Option	Required	Description
--help	No	Displays a help screen with available options.
--analysisFolder	Yes	Provide the full path to the local analysis folder. Folder must have sufficient space and must be on an NVMe SSD drive. • NovaSeq require a minimum of 3000 GB free. • NextSeq require a minimum of 500 GB free.
--resourcesFolder	Yes	Provide the full path to the local resources folder.
--runFolder	Yes	Required when `--fastqFolder` is not specified. Provide the full path to the local run folder.
--fastqFolder	Yes	Required when `--runFolder` is not specified. Provide the full path to the local FASTQ folder. Analysis starts at this location.
--user	No	Optional for docker. Specify the user ID to be used within the Docker container.
--remove	No	Optional for docker. Passes the `--rm` option to remove the subsequent container after execution.
--version	No	Displays the version of the software.
--sampleSheet	No	Provide the full path, including file name, if not provided as `SampleSheet.csv` in the run folder.
--sampleOrPairIDs	No	Provide the comma-delimited sample or pair IDs that should be processed on this node. For example, `Pair_1,Pair_2,Sample_1`.
--demultiplexOnly	No	Demultiplex to generate FASTQ only without additional analysis.

Option	Required	Description
--gather	No	Follow this option with any directories whose results should be gathered into a single Results folder.
--hashtableFolder	No	Provide the full path to the local DRAGEN hash table.

Starting from Instrument Run Folders

Refer to *Command-Line Options* on page 8 for additional commands.

Use full paths when specifying the file paths in the command line. Avoid special characters such as &, *, #, and spaces.

To run DRAGEN TruSight Oncology 500 Analysis Software using the NovaSeq 6000 Sequencing System, add the `--isNovaSeq` to any of the following commands:

1. Wait for any running DRAGEN TruSight Oncology 500 Analysis Software containers to complete before launching a new analysis. Run the following command to generate a list of running containers:

   ```
   sudo docker ps
   ```

2. Select from one of the following options:
   - Start from BCL files in the run folder with the sample sheet included in the run folder.

```
DRAGEN_TSO500.sh \
--resourcesFolder /staging/illumina/DRAGEN_TSO500/resources \
--hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable \
--runFolder /staging/{RunFolderName} \
--analysisFolder /staging/{AnalysisFolderName}
```

   - Start from BCL files in the run folder specifying a different sample sheet.

```
DRAGEN_TSO500.sh \
--resourcesFolder /staging/illumina/DRAGEN_TSO500/resources \
--hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable \
--runFolder /staging/{RunFolderName} \
--analysisFolder /staging/{AnalysisFolderName} \
--sampleSheet /staging/{SampleSheetName}.csv
```

   - Start from BCL files in the run folder specifying a different sample sheet and demultiplexing only.

```
DRAGEN_TSO500.sh \
--resourcesFolder /staging/illumina/DRAGEN_TSO500/resources \
--hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable \
--runFolder /staging/{RunFolderName} \
```

```
   --analysisFolder /staging/{AnalysisFolderName} \
   --sampleSheet /staging/{SampleSheetName}.csv \
   --demultiplexOnly
```

- Start from FASTQ folder with the sample sheet included in the FASTQ folder.

```
   DRAGEN_TSO500.sh \
   --resourcesFolder /staging/illumina/DRAGEN_TSO500/resources \
   --hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable \
   --fastqFolder /staging/{FastqFolderName} \
   --analysisFolder /staging/{AnalysisFolderName}
```

- Start from FASTQ folder with sample sheet included in the FASTQ folder and subset of samples or pairs.

```
   DRAGEN_TSO500.sh \
   --resourcesFolder /staging/illumina/DRAGEN_TSO500/resources \
   --hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable \
   --fastqFolder /staging/{FastqFolderName} \
   --analysisFolder /staging/{AnalysisFolderName} \
   --samplePairIDs “Pair_1, Pair2”
```

Starting From BCL Files

If starting from BCL (*.bcl) files, DRAGEN TruSight Oncology 500 Analysis Software requires the run folder to contain certain files and folders. These inputs are required for Docker.

The run folder contains data from the sequencing run. If you start with the run folder, make sure that the folder contains the following files:

Folder/File	Description
Config folder	Configuration files.
Data folder	*.bcl files.
Images folder	[Optional] Raw sequencing image files.
InterOp folder	Interop metric files.
Logs folder	[Optional] Sequencing system log files.
RTALogs folder	Real-Time Analysis (RTA) log files.
RunInfo.xml file	Run information.
RunParameters.xml file	Run parameters.

Folder/File	Description
SampleSheet.csv file	Sample information. If you want to use a sample sheet that is not in the run folder or a sample sheet named something other than SampleSheet.csv, provide the full path.

Starting From FASTQ Files

The following inputs are required for running the DRAGEN TruSight Oncology 500 Analysis Software using FASTQ (*.fastq) files. The requirements apply to Docker.

- Full path to an existing FASTQ folder.
- The FASTQ folder structure conforms to the folder structure in *FASTQ File Organization* on page 11.
- The sample sheet is in the FASTQ folder path, or you can set the path to the sample sheet with the –`-sampleSheet` override command.

Make sure there is sufficient disk space for the analysis to complete. Refer to the `--help` command line argument details for disk space requirements.

FASTQ File Organization

Store FASTQ files in individual subfolders that correspond to a specific Sample_ID. Keep file pairs together in the same folder.

The DRAGEN TruSight Oncology 500 Analysis Software requires separate FASTQ files per sample. Do not merge FASTQ files.

The instrument generates two FASTQ files per flow cell lane, so that there are eight FASTQ files per sample.

```
Sample1_S1_L001_R1_001.fastq.gz
```

- Sample1 represents the Sample ID
- The S in S1 means sample, and the 1 in S1 is based on the order of samples in the sample sheet, so that S1 is the first sample.
- L001 represents the flow cell lane number.
- The R in R1 means Read, so that R1 refers to Read 1.

Running on Multiple DRAGEN Servers

DRAGEN TruSight Oncology 500 Analysis Software can be used to run a subset of samples on different DRAGEN servers to decrease processing time. This is possible using a three stage process called scatter/gather, which consists of demultiplexing, analysis, and result gathering.

The first stage is demultiplexing. Demultiplexing runs once on the entire run folder, generates FASTQ files for each sample in the run, and then separates sample files into respective folders. Once complete, note the output directory containing the sample directories holding the FASTQ files.

The process for scattering the analysis on multiple DRAGEN servers is as follows.

1. Determine how many DRAGEN servers are available to run.
2. Run demultiplexing on a single DRAGEN server.

   > ℹ To sequence runs on multiple DRAGEN server using the XP workflow, modify the sample sheet to include a subset of the lanes. For example, on an S2 flowcell, you can create two modified sample sheets with one containing the samples from Lane 1 and the other from Lane 2. This allows only the sample sheet to be modified instead of copying files between servers. This strategy would use the start from Run Folder commands without the `--demultiplexOnly` option. The entire run folder would need to be copied to each analysis server as demultiplexing would be performed once per server.

3. Transfer the FASTQ folder output from the original DRAGEN server to additional servers. `Logs_Intermediates/FastqGeneration`.
4. Run analysis software using the `--fastqFolder` option on both the original and additional DRAGEN servers.
   - Option 1: Copy the original `SampleSheet.csv` to each server. Then provide a subsetted list to the Bash script on each DRAGEN server with the intended samples/pairs to run.
   - Option 2: Copy and modify the `SampleSheet.csv` to each DRAGEN server to only contain the list of samples/pairs to run.

The software verifies all samples in the sample sheet are contained within the FASTQ folders unless the `--samplePairsIDs` command-line option is present in the analysis launch. Failure to account for these checks results in an error.

Table 1 Commands for Multi Node Analysis

Step	Command
Demultiplexing	`DRAGEN_TSO500.sh --resourcesFolder /staging/illumina/DRAGEN_TSO500/resources --hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable --runFolder /staging/{RunFolderName} --analysisFolder /staging/{DemultiplexAnalysisFolderName} --demultiplexOnly --sampleSheet /staging/illumina/{SampleSheetName}`

Step	Command
Analysis (one server)	`DRAGEN_TSO500.sh --resourcesFolder /staging/illumina/DRAGEN_TSO500/resources --hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable --fastqFolder /staging/{DemultiplexAnalysisFolderName}/Logs_Intermediates/FastqGeneration/ --analysisFolder /staging/{Node1AnalysisFolderName} --sampleSheet /staging/illumina/{SampleSheetName} --samplePairIDs Pair_1,Pair_2`
Analysis (additional servers)	`DRAGEN_TSO500.sh --resourcesFolder /staging/illumina/DRAGEN_TSO500/resources --hashtableFolder /staging/illumina/DRAGEN_TSO500/ref_hashtable --fastqFolder /staging/{DemultiplexAnalysisFolderName}/Logs_Intermediates/FastqGeneration/ --analysisFolder /staging/{Node1AnalysisFolderName} --sampleSheet /staging/illumina/{SampleSheetName} --samplePairIDs Pair_3`
Gather	`DRAGEN_TSO500.sh --analysisFolder /Gathered_Results --resourcesFolder ${RESOURCES} --runFolder ${RUN_FOLDER} --sampleSheet ${SAMPLE_SHEET} --gather /Demultiplex_Output /Node1_Output /Node2_Output`

Analysis Methods

The DRAGEN TruSight Oncology 500 Analysis Software workflow performs the following analysis steps, and then writes analysis output files to the folder specified.

- FASTQ Generation
- DNA analysis using the following methods:
  - DNA Alignment and Realignment
  - Read Collapsing
  - Indel Realignment and Read Stitching
  - Small Variant Calling
  - Small Variant Filtering
  - Copy Number Variant Calling

  - Phased Variant Calling
  - Variant Merging
  - Annotation
  - Tumor Mutational Burden
  - Microsatellite Instability Status
- RNA analysis using the following methods:
  - Downsampling
  - Read Trimming
  - Alignment
  - Duplicate Marking
  - Fusion Calling
  - RNA Fusion Filtering
  - Splice Variant Calling
  - Annotation
  - Fusion Merging
- Quality Control
  - Run QC
  - DNA Sample QC
  - RNA Sample QC
  - Contamination Detection

DNA Analysis Methods

DNA Alignment and Realignment

The alignment step uses the Burrows-Wheeler Aligner (BWA-MEM) with the SAM Tools utility to align DNA sequences in FASTQ files to the hg19 genome. This alignment step generates BAM files (*.bam) and BAM index files (*.bam.bai), which are saved to the DnaAlignment folder. A BAM file is the compressed binary version of a SAM file that is used to represent aligned sequences.

The software performs a second alignment on FASTQ files after the Read Collapsing step completes. The second alignment enables the realignment of sample reads using only unique molecular identifier (UMI) collapsed reads.

For more information on BWA-MEM, see the Burrows-Wheeler Aligner website. For more information on SAM and BAM files, see the Sequence Alignment/Map specification page on GitHub.

Read Collapsing

The read collapsing analysis step executes an algorithm that collapses sets of reads (known as families) with very similar genomic locations into representative sequences using UMI tags. This process allows for the accurate removal of duplicate reads without losing the signal of very low frequency sequence variations. Additionally, UMI collapsing further reduces FFPE deamination artifacts by utilizing duplex collapsing where information from complimentary strands are combined. The read collapsing step produces FASTQ files and associated metrics files in the CollapsedReads output folder. Read collapsing adds the following BAM tags:

- **RX/XU**—UMI.
- **XV**—Number of reads in the family.
- **XW—**Number of reads in the duplex-family, or 0 if not a duplex family.

Indel Realignment and Read Stitching

The Gemini software performs local indel realignment, paired-read stitching, and read filtering to improve small variant calling results. A stitched read is a single read that has been combined from a pair of reads. Reads near detected indels are realigned to remove alignment artifacts. The software takes in a single BAM file and the genome FASTA used to align it and outputs a corresponding single BAM file with stitched, pair-realigned reads. Read pairs with poor map quality or supplementary and secondary alignments from the input BAM are ignored.

For successfully stitched reads, Gemini adds the following BAM tags:

- **XD**—Directional support string indicating forward, reverse, and stitched positions.
- **XR**—Pair orientation (FR or RF).

Small Variant Calling

Pisces software performs somatic variant calling to identify variants at low frequency in DNA samples. Pisces calls small variants in the BAM files that are generated from the StitchedRealigned analysis step.

For each variant candidate, Pisces adds a US field under the Format column in the genome.vcf for the mutant support of the following read type counts:

- Duplex stitched
- Duplex nonstitched
- Simplex forward stitched
- Simplex forward nonstitched
- Simplex reverse stitched
- Simplex reverse nonstitched

This is followed by total support of the same read type counts.

The small variant calling genome.vcf at this step only collects candidate and outputs corresponding read support information. The final variant call is determined in following postprocessing step.

The software component Psara is used to trim the gVCF based on the panel manifest. Variants are included if they overlap with the manifest or are contained within an overlapping indel. Small Variant Filtering determines the final variant call.

For more information, consult the Pisces design document in the Pisces project on GitHub.

Small Variant Filtering

The software component, Pepe, performs post-processing on the small variant calling genome VCFs to polish backgrounds and adjust quality scores. The software filters out variants when error rates do not meet quality thresholds. This analysis step produces genome VCF files and associated error rate files. The minimum read depth for reference calls is 100. The limit of detection for VAF is 5% at the minimum read depth.

Pepe computes two quality scores for each candidate that dynamically adjust for the following conditions:

- Background noise
- Trinucleotide change
- Read support type

For each variant candidate, background noise at the same site is estimated from normal baseline samples of varying qualities. A p-value is calculated using the observed mutant depth, total depth, and background noise using binomial distribution. The p-value is then converted to a variant quality score (AQ). The sample-specific error rate of each trinucleotide change is estimated from different support categories in each sample by using all the positions with an allele frequency less than 1%. For each variant candidate, a likelihood ratio score (LQ) is computed by the corresponding error rate of the observed total and mutant read. A bias score (BFQ) is computed on each variant candidate to evaluate the imbalance of mutant vs total read support between different support groups.

For variants with a Catalogue of Somatic Mutations in Cancer (COSMIC) count > 50, the LQ and AQ thresholds are 20 and the remaining sites are 60. For indel, at least one stitched mutant support is required. For non-COSMIC variant, threshold for BFQ is < 20. In addition, positional information of mutant and WT allele in fragment will be extracted for each variant candidate. A Kolmogorov-Smirnov test will be applied to compute p-value between mutant and WT position. Variants with p-value < 0.05 and median difference > = 0.5 will be filtered and labeled VarBias. The net effect of the read collapsing and variant filtering significantly reduces false positives. For example, false positives in a typical cell-free DNA sample were reduced to < 5 per Mb from ~1500 per Mb.

In addition to the evaluation of the quality scores, certain regions covered in the product manifest are excluded from analysis due to high background noise. All excluded variants are identified in the VCF using a flag. The block list of excluded sites can be obtained on request from your local Illumina representative.

Some regions are known to be difficult to sequence. One example region is the TERT promoter region. Although sequencing can occur at the TERT promoter region, this location might result in low coverage due to the GC rich content of the sequenced region.

Copy Number Variant Calling

The CRAFT copy number variant caller performs amplification, reference, and deletion calling for target CNV genes within the assay. The CRAFT software component counts coverage of each target interval on the panel, performs normalization, calculates fold change values for each gene, and determines the CNV status for each CNV target gene. During normalization steps, coverage biases are corrected using potential variables such as sequencing depth, target size, PCR duplicates, probe efficiency, GC bias, and DNA type. A collection of normal FFPE and genomic DNA samples is used to correct some of these variables. For each target CNV gene, *in silico* data is trained to determine a gene-specific threshold for amplification and deletion. The inputs are collapsed read in BAM format and the outputs are VCF files. Amplifications are annotated as DUP in the VCF file. Deletions status (DEL) are provided for information only and always are marked as LowValidation in the VCF file.

Phased Variant Calling

Scylla rapidly detects multiple nucleotide variants (MNVs) in a given sample. The software uses Scylla to detect specific, clinically relevant mutations in EGFR exon 19 that would otherwise be out of scope for the variant caller. Psara filters the small variant gVCF to a small region in exon 19 of EGFR. Candidate SNPs, MNVs, and indels from this subset of the gVCF are given to Scylla along with the BAM output from Gemini. Scylla uses the original BAM to determine which of these small variants should be phased together into longer MNVs.

At a high level, Scylla identifies variants that are candidates for phasing in the input gVCF and arranges the variants into local neighborhoods. Scylla then mines the sample BAM file for any evidence that these small variants occur in the same clonal sub-populations with each other. This is done by clustering overlapping reads in the neighborhood into a minimal set of clusters, which contain the same variants.

Variant Merging

The software merges the phased variants with the other small variants generated from small variant filtering step and produces a gVCF file. In this process, exact duplicates that match chromosome, position, reference allele, and alternative allele are removed. The following Epidermal Growth Favor Receptor (EGFR) variants are added if found from Phased Variant Calling. All other EGFR variants are filtered out in variant merging.

Table 2 EGFR Variants

Chromosome	Position	Reference Allele	Alternative Allele
chr7	55242463	AAGGAATTAAGAGAAG	A
chr7	55242464	AGGAATTAAGAGA	A
chr7	55242464	AGGAATTAAGAGAAGC	A
chr7	55242465	GGAATTAAGAGAAGCA	G
chr7	55242467	AATTAAGAGAAGCAAC	A
chr7	55242469	TTAAGAGAAGCAACATCTC	T
chr7	55242468	ATTAAGAGAAGCAACATCT	A
chr7	55242466	GAATTAAGAGAAGCAA	G
chr7	55242465	GGAATTAAGA	G
chr7	55242469	TTAAGAGAAGCAA	T
chr7	55242462	CAAGGAATTAAGAGAA	C
chr7	55242466	GAATTAAGAGAAGCAACAT	G
chr7	55242482	CATCTCCGAAAGCCAACAAGGAAAT	C
chr7	55242465	GGAATTAAGAGAAGCAACA	G
chr7	55242467	AATTAAGAGAAGCAACATC	A
chr7	55242469	TTAAGAGAAG	C
chr7	55242467	AATTAAGAGAAGCAACATC	T
chr7	55242469	TTAAGAGAAGCAA	C
chr7	55242467	TTAAGAGAAGCAA	TTGCT
chr7	55242468	ATTAAGAGAAG	GC
chr7	55242469	TTAAGAGAAGCAACATCTCC	CA
chr7	55242465	GGAATTAAGAGAAG	AATTC
chr7	55242465	GGAATTAAGAGAAGCAAC	AAT
chr7	55242467	AATTAAGAGAAGCAAC	T
chr7	55242467	AATTAAGAGAAGCAACATCTC	TCT
chr7	55242469	TTAAGAGAAGCAACATCT	CAA
chr7	55242465	GGAATTAAGAGAAGCAA	AATTC
chr7	55242465	GGAATTAAGAGAAGCAACATC	AAT
chr7	55242468	ATTAAGAGAAGCAAC	GCA

Annotation

The Illumina Annotation Engine Nirvana performs annotation of small variants. The inputs are gVCF files and the outputs are annotated JSON files.

Each variant entry that is processed by Nirvana is annotated with available information from databases such as dbSNP, gnomAD genome and exome, 1000 genomes, ClinVar, COSMIC, RefSeq, and Ensembl. Version information and general details can be retrieved from the header. Each annotated variant is included as a nested dictionary structure in separate lines following the header. Version information for each annotation database is shown in the following table.

Database	Version
gnomAD	2.1
COSMIC	v84
ClinVar	2019-02-04
dbSNP	v151
1000 Genomes Project	Phase 3 v5a
RefSeq	NCBI Homo sapiens Annotation Release 105.20201022
Ensembl	VEP build 91

Tumor Mutational Burden

The tumor mutational burden (TMB) analysis step generates TMB metrics from the annotated small variant JSON file and the gVCF file generated from the small variant filtering analysis step. The annotated JSON file is used to retrieve information regarding individual variants, such as allele counts in public databases and resulting consequences at a transcript level. The gVCF file is used to evaluate the effective panel size denominator.

To remove germline variants from the TMB calculation, the software uses a combination of public database filtering and post-database filtering strategy that uses allele frequency information and variants in close proximity.

First, the component excludes any variant with an observed allele count ≥10 in any of the GnomAD exome, genome, and 1000 genomes database. To filter germline variants that are not observed in the database, the software identifies variants on the same chromosome with an allele frequency within a certain range. If a given variant is not filtered out based on occurrence in the databases, variants on the same chromosome with similar allele frequencies are grouped. If 5 or more similar variants are filtered, the variant of interest is removed from the TMB Calculation. Additionally, variants with an allele frequency ≥ 90% are removed from the TMB calculation as well. The TMB is calculated as follows.

TMB = Eligible Variants / Effective panel size

Eligible Variants	• Variants not removed by the filtering strategy • Variants in the coding region (RefSeq Cds) • Variant Frequency >= 5% • Coverage >= 50X • SNVs and Indels (MNVs excluded) • Nonsynonymous and synonymous variants • Variants with COSMIC count >= 50 excluded
Effective Panel Size	• Total coding region with coverage > 50X • Excluding low confidence regions in which variants are not called

Outputs are captured in a *_TMB_Trace.tsv file that contains information on variants used in the TMB calculation and a *.tmb.json file. The TMB score calculation and configuration details.

Microsatellite Instability Status

The Microsatellite Instability (MSI) status step determines microsatellite instability from the BAM file created in the read stitching analysis step and generates an MSI metric file. The software assesses microsatellite sites for evidence of instability, relative to a set of baseline normal samples that are based on information entropy metrics. The percentage of unstable MSI sites to total assessed MSI sites is reported as a sample-level microsatellite score.

Contamination Detection

The contamination analysis step detects contamination by foreign DNA in the VCF files that the small variant filter step generates. The software determines whether a sample has foreign DNA from the combination of contamination p-value (p-score) and contamination scores.

The contamination score is the sum of all the log likelihood scores across all positions. The p-score represents the significance that SNPs are distributed nonuniformly across the chromosomes. This could indicate a highly rearranged genome and cause false positives for contamination.

In contaminated samples, there are SNPs that have variant allele frequency shifts from 0%, 50%, or 100%. The algorithm collects all the positions that overlap with common SNPs with variant allele frequencies of < 25% or > 75%. Then, the algorithm computes the likelihood that the positions are an error or a real mutation using the following qualifications:

- Estimates the error rate per sample.
- Mutation support.
- Total depth of each position selected.

RNA Analysis Methods

Downsampling

Each sample is downsampled to 30 million RNA reads. This number represents the total number of single reads (ie, R1 + R2, from all lanes). When using the recommended sequencing configurations or plexing, the samples can have fewer reads than the downsampling limit. In these cases, the FASTQ files are left as-is.

Read Trimming

Reads are trimmed to 76 base pairs for further processing.

DRAGEN server RNA Alignment and Fusion Detection

DRAGEN server aligns RNA reads in a transcript-aware mode using the human hg19 genome containing unplaced contigs (i.e. chrUn_gl regions) and uses GENCODEv19 transcript annotations to identify splice sites. DRAGEN server identifies and marks duplicate read alignments using start and end coordinates of alignments (adjusted for soft clipped reads).

Fusion and splice variant calling only use deduped fragments to score variants. DRAGEN server identifies fusion candidates using chimeric split read alignments (pairs of primary and supplementary alignments) against multiple genes. DRAGEN server scores and filters reads based on the various features of each candidate such as the number of supporting reads, mapping quality of supporting reads, and sequence homology between parent genes.

The inputs to DRAGEN server are trimmed reads in FASTQ format. The outputs include a BAM file which contains duplicate-marked read alignments, a SJ.out.tab file which contains unannotated splice junctions, and a CSV file which contains fusion candidates.

Splice Variant Calling

Splice variant calling is performed using internally developed software. The inputs are BAM files and SJ.out.tab files from DRAGEN server. Junctions from SJ.out.tab are filtered first using splice annotations from GENCODEv19, and then further filtered using a baseline from a cohort of non-tumor FFPE samples of varying tissue types. Splice junctions appearing on an allow list (referred to in some files as a whitelist) are not filtered. The allow list contains ARv7, MET exon 14 skipping, and EGFRvIII. The outputs are VCF files, which are the final output, and TSV file containing intergenic variants, which are used in merging. Splice variants are scored from 0–10 as shown in the following table.

Table 3 Scored Features in Splice Variant Caller

Score Component	Splice Feature	Scored Range	Coefficient
Split reads	split_unique_reads_alt	0–10	1

Annotation

The IlluminaAnnotation Engine performs annotation of splice variants. The inputs and outputs are VCF files.

RNA Merge

Most fusion events are called by the DRAGEN fusion caller, which leverages supplemental and soft-clipped alignments to detect fusion breakpoints. Occasionally, fusion breakpoints are close enough together that reads supporting the fusion have gapped alignments (N in the CIGAR string). These fusion events are detected during splice variant calling (a separate step in the TSO 500 workflow). The RNA fusion merge step combines fusions detected during DRAGEN fusion calling and splice variant calling into a single output containing all detected fusions.

Quality Control

The DRAGEN TruSight Oncology 500 Analysis Software v1.1 includes several quality control analyses.

Run QC

The Run Metrics report provides suggested values to determine if run quality results are within an acceptable range using InterOp files from the sequencing run folder. Quality thresholds vary between systems and are automatically detected based on the sequencing run folder. The tables below provide run metric and quality threshold information for different systems.

Table 4 High Throughput Systems (NovaSeq 6000 System)

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
PCT_PF_READS (%)	Total percentage of reads passing filter	≥55.0	All
PCT_Q30_R1 (%)	Percentage of Read 1 reads with quality score equal to or above 30	≥80.0	All
PCT_Q30_R2 (%)	Percentage of Read 2 reads with quality score equal to or above 30	≥80.0	All

Table 5 Low Throughput Systems (NextSeq 500/550 Systems)

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
PCT_PF_READS (%)	Total percentage of reads passing filter	≥80.0	All
PCT_Q30_R1 (%)	Percentage of Read 1 reads with quality score equal to or above 30	≥80.0	All
PCT_Q30_R2 (%)	Percentage of Read 2 reads with quality score equal to or above 30	≥80.0	All

DNA Library QC Metrics for Sample

The inputs for DNA Library QC Metrics for Sample are DNA alignment, read collapsed BAM, indel realignment, read stitching BAM, and CRAFT normalized BinCount.tsv files. The metrics and guideline thresholds can be found in the MetricsOutput.tsv file.

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
CONTAMINATION_ SCORE and CONTAMINATION_ P_VALUE	The contamination score from based on VAF distribution of SNPs. The contamination p-value is used to assess highly rearranged genomes and only needed when contamination score is above USL. A p-score less than 0.05 suggest that the sample has likely large-scale rearrangements that could lead to high contamination scores without actual sample contamination.	Contamination Score ≤ 3106 OR Contamination Score > 3106 and Contamination p-value ≤ 0.049	All
MEDIAN_EXON_ COVERAGE	Median exon fragment coverage across all exon bases.	≥ 150	Small variant TMB

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
PCT_EXON_50X	Percent exon bases with 50X fragment coverage.	≥ 90.0	Small variant TMB
MEDIAN_INSERT_SIZE	The median fragment length in the sample.	≥ 70	Small variant TMB
USABLE_MSI_SITES	The number of MSI sites usable for MSI calling.	≥ 40	MSI
COVERAGE_MAD	Median Absolute Deviation. Represents the median normalized deviation across all regions used for CNV calling.	≤ 0.210	CNV
MEDIAN_BIN_COUNT_CNV_TARGET	The median raw bin count per CNV target.	≥ 1.0	CNV

RNA Library QC

The inputs for RNA Library QC are RNA alignment. Metrics and guideline thresholds can be found in the MetricsOutput.tsv file.

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
MEDIAN_CV_GENE_500X	The median CV for all genes with median coverage > 500x. Genes with median coverage > 500x are likely to be highly expressed. Higher CV median > 500x indicates an issue with library preparation (poor sample input and/or probes pulldown issue).	≤ 93	Fusion Splice
MEDIAN_INSERT_SIZE	The median fragment length in the sample.	≥ 80	Fusion Splice
TOTAL_ON_TARGET_READS	The total number of reads that map to the target regions.	≥ 9000000	Fusion Splice

Analysis Output

When the analysis run completes, the DRAGEN TruSight Oncology 500 Analysis Software generates an analysis output folder in a specified location.

To view analysis output, navigate to the analysis output folder and select the files that you want to view.

Metrics Output

The `MetricsOutput.tsv` file contains the following quality control metrics for all samples:

- QC metrics for the following metrics:
  - Small variant calling (SVC)
  - TMB
  - MSI
  - CNV
  - Fusion
  - Splice variant calling
- Run QC metrics, analysis status, and contamination

This TSV file also includes expanded DNA library QC metrics per sample, based on total reads, collapsed reads, chimeric reads, and on-target reads. Analysis using RNA samples also produces RNA library QC metrics and expanded RNA library QC metrics per sample based on total reads and coverage.

The `MetricsOutput.tsv` file is a final combined metrics report with sample status, key analysis metrics, and metadata in a *.tsv file. Sample metrics within the report indicate guideline-suggested lower specification limits (LSL) and upper specification limits (USL) for each sample in the run.

For troubleshooting information, refer to *Troubleshooting* on page 40.

Run Metrics

Run metrics from the analysis module indicate the quality of the sequencing run.

Review the following metrics to assess run data quality:

Metric	Description	Recommended Threshold
PCT_PF_READS	Percentage of reads on the sequencing flow cell that pass the filter.	≥ 55.0
PCT_Q30_R1	Percentage of bases with a quality score ≥ 30 from Read 1.	≥ 80.0
PCT_Q30_R2	Percentage of bases with a quality score ≥ 30 from Read 2.	≥ 80.0

The values in the Run Metrics section are listed as NA in the following situations:

- If the analysis was started from FASTQ files.
- If the analysis was started from BCL files and the InterOp files are missing or corrupt.

Single Node Analysis Output Folder Structure

This section describes the content of output folders generated from analysis run on a single node.

Single output folder structure is as follows.

- 📁Logs_Intermediates
  - 📁AlignmentCollapser
    - 📁Subfolders per DNA sample ID containing UMI collapsed BAM files and raw DRAGEN metrics.
  - 📁Annotation—Contains annotation output logs.
  - 📁Cleanup—Contains Output cleanup logs.
  - 📁CnvCaller
  - 📁CombinedVariantOutput—CombinedVariant output logs.
  - 📁Contamination
    - 📁Subfolders per sample ID containing the contamination metrics JSON.
    - 📁Contamination output logs
  - 📁DnaFastqValidation

- DnaQCMetrics
  - Subfolders per sample ID containing the aligned, collapsed, and stitched metrics JSON files.
  - DnaQCMetrics output logs
- FastqDownsample
  - Subfolders per sample ID containing FASTQ files.
  - FastqGeneration output logs
- FastqGeneration
  - Subfolders per sample ID containing FASTQ files.
  - FastqGeneration Output Logs
- MergedAnnotation
- MetricsOutput
- Msi
  - Subfolders per sample ID containing the MSI metrics JSON.
  - Msi output logs
- PhasedVariants—Several folders with this name might exist in the output folder structure.
  - Subfolders per sample ID containing the phased variant metrics JSON.
  - Psara and Scylla output logs
- ResourceVerification
- RnaAlignmentFusionCaller
  - Subfolders per sample ID containing the duplicate marked aligned BAM and index files.
  - RnaAlignmentFusionCaller output logs
- RnaAnnotation
  - Subfolders per sample ID containing the annotated VCF file.
  - dsdm json]
- RnaFastqValidation
- RnaFusionMerge
- RnaQCMetrics
  - Subfolders per sample ID containing the aligned, collapsed, and stitched metrics JSON files.
  - RnaQCMetrics output logs
- RnaSpliceVariantCalling
  - Subfolders per sample ID containing the splice variants VCF.
  - dsdm JSON

  - 📁RunQc
    - 📄RunQC Metrics JSON file
    - 📁RunQC Output logs
  - 📁SampleAnalysisResults
  - 📁SamplesheetValidation—Contains sample sheet validation output logs.
  - 📁SmallVariantFilter
    - 📁Subfolders per sample ID containing the error rate tables.
    - 📁SmallVariantFilter output logs
  - 📁StitchedRealigned
    - 📁Subfolders per sample ID containing the stitched, realigned BAM and index files. The StitchedRealigned BAM is the final output BAM for DNA.
    - 📁StitchedRealigned output logs
  - 📁Tmb
    - 📁Subfolders per sample ID containing the TMB metrics JSON.
    - 📁TMB output logs
  - 📁TrimFastq
    - 📁Subfolders per sample ID containing FASTQ files.
    - 📄dsdm JSON
  - 📁VariantCaller
    - 📁Subfolders per sample ID containing the unfiltered genome VCF file.
    - 📁VariantCaller output logsg
  - 📁VariantMatching
- 📁Results
  - Metrics Output TSV
  - 📁Sample ID—The following outputs are produced for each sample:
    - 📄Combined Variant Output TSV
    - 📄TMB Trace TSV
    - 📄Small Variant Genome VCF
    - 📄Small Variant Genome Annotated JSON
    - 📄Copy Number Variant VCF
    - 📄All Fusion CSV
    - 📄Splice Variant VCF

        - Splice Variant Annotated JSON

Multiple Node Analysis Output Folder Structure

This section describes the content of output folders generated from analysis. Analysis output folder structure for analysis using multiple nodes is as follows.

- Demultiplex_Output
    - Logs_Intermediates—Contains FASTQ files per sample.
- Node1_Output—The following outputs are produced for each node used.
    - Logs_Intermediates
    - Results—Contains results only for the samples run on the node.
- Gathered_Results
    - Logs_Intermediates
- Results—Contains results for all samples from all nodes used.

Combined Variant Output

File name: `{SampleID}_CombinedVariantOutput.tsv`

The combined variant output file contains the variants and biomarkers in a single file that is based on a single sample. If using pair ID, the file is based on paired DNA and RNA samples (ie from the same individual). The output contains the following variant types and biomarkers:

- Small variants (including EGFR complex variants)
- Gene amplifications
- TMB
- MSI
- Fusions
- Splice variants

The combined variant output file also contains Analysis Details and Sequencing Run Details sections. The details of each is listed in the following table.

Analysis Details	Sequencing Run Details

- Pair ID
- DNA Sample ID (if DNA is run)
- RNA Sample ID (if RNA is run)
- Output Date
- Output Time
- Module Version
- Pipeline Version (Docker Image Version #)
- Run Name
- Run Date
- DNA Sample Index ID (if DNA is run)
- RNA Sample Index ID (if RNA is run)
- Instrument ID
- Instrument Control Software Version
- Instrument Type
- RTA Version
- Reagent Cartridge Lot Number

Combined variant output produces small variants with blank fields in the following situations:

- The variant has been matched to a canonical RefSeq transcript on an overlapping gene not targeted by TruSight Oncology 500.
- The variant is located in a region designated iSNP, iIndel, or Flanking in the `TST500_Manifest.bed` file located in the Resources folder.

Variant Filtering Rules

- **Small Variants**—All variants with the FILTER field marked as PASS in the merged genome VCF are present in the combined variant output.
  - Gene information is only present for variants belonging to canonical transcripts that are within the Gene Allow List–Small Variants.
  - Transcript information is only present for variants belonging to canonical transcripts that are within the Gene Allow List–Small Variants.
- **Copy Number Variants**—Copy number variants must meet the following conditions:
  - FILTER field marked as PASS.
  - ALT field is <DUP>.
- **Fusion Variants**—Fusion variants must meet the following conditions:
  - Passing variant call (KeepFusion field is true).
  - Contains at least one gene on the fusion allow list.
  - Genes separated by a dash (-) indicate that the fusion directionality could be determined. Genes separated by a slash (/) indicate that the fusion directionality could not be determined.
- **Biomarkers TMB/MSI**—Always present when DNA sample is processed.
- **Splice Variants**—Passing splice variants that are contained on genes EGFR, MET, and AR.

DNA Output

Merged Small Variant gVCF

File name: `{SAMPLE_ID}_MergedSmallVariants.genome.vcf`

The merged variant genome variant call file combines the small variant genome VCF (output of variant filtering) and clinically relevant variants in EGFR exon 19 from phased variant calling. This contains information on all candidate small variants evaluated. The variant status is determined by the FILTER column in the genome VCF as follows.

ALT	Filter	Note
.	PASS	WT
., A, C, G, etc.	LowDP	No call (DP < 100X, insufficient depth to confidently detect variants with VAF ≥ 5%).
A, C, G, etc.	PASS	PASS variants
A, C, G, etc.	LowSupport	Filtered variant candidate: • Fail AQ or LQ • 0 stitched support for indel or variant in homopolymer context.
A, C, G, etc.	Blocklist	An excluded list of sites (referred to as blacklist in some files). Refer to the *Small Variant Filtering* on page 16 section for more information.
A, C, G, etc.	LowVarSupport	Filtered variant candidate with mutant support < 1.

Merged Small Variant Annotated JSON

File name: `{SAMPLE_ID}_MergedSmallVariantsAnnotated.json.gz`

The merged small variants annotated file provides variant annotation information for all nonreference positions from the merged genome VCF including pass and nonpass variants.

TMB Trace

The TMB trace file provides comprehensive information on how the TMB value is calculated for a given sample. All passing small variants from the small variant filtering step are included in this file. To calculate the numerator of the `TmbPerMb` value in the TMB JSON, set the TSV file filter to use the `IncludedInTMBNumerator` with a value of `True`.

The TMB trace file is not intended to be used for variant inspections. The filtering statuses are exclusively set for TMB calculation purposes. Setting a filter does not translate into the classification of a variant as somatic or germline.

Column	Description
Chromosome	Chromosome
Position	Position of variant
RefCall	Reference base
AltCall	Alternate base
VAF	Variant allele frequency
Depth	Coverage of position
CytoBand	Cytoband of variant
GeneName	Name of gene if applicable. A semicolon delimited list is used for multiple genes.
VariantType	Type of the variant: SNV, insertion, deletion, MNV
CosmicIDs	Cosmic IDs, if multiple concatenated by “;”
MaxCosmicCount	Maximum Cosmic study count
AlleleCountsGnomadExome	Variant allele count in gnomAD exome database
AlleleCountsGnomadGenome	Variant allele count in gnomAD genome database
AlleleCounts1000Genomes	Variant allele count in 1000 genomes database
MaxDatabaseAlleleCounts	Maximum variant allele count over the three databases.
GermlineFilterDatabase	TRUE if variant was filtered by the database filter
GermlineFilterPRoxi	TRUE if variant was filtered by the proxi filter
CodingVariant	TRUE if variant is in the coding region
Nonsynonymous	TRUE if variant has any transcript annotations with nonsynonymous consequences
IncludedinTMBNumberator	TRUE if variant is used in the TMB calculation

Copy Number VCF

The copy number VCF file contains CNV calls for DNA libraries of the amplification genes targeted by DRAGEN TruSight Oncology 500 Analysis Software v1.1. The CNV call indicates fold change results for each gene classified as reference, deletion, or amplification.

The value in the QUAL column of the VCF is a Phred transformation of the p-value where Q=-10xlog10 (p-value). The p-value is derived from the t-test between the fold change of the gene against rest of the genome. Higher Q-scores indicate higher confidence in the CNV call.

In the VCF notation, <DUP> indicates the detected fold change (FC) is greater than a predefined amplification cutoff. <DEL> indicates the detected FC is less than a predefined deletion cutoff for that gene. This cutoff can vary from gene to gene.

<DEL> calls have only been validated with *in silico* data sets. As a result, all <DEL> calls have LowValidation filter in the VCF.

Each copy number variant is reported as a fold change on normalized read depth in a testing sample relative to the normalized read depth in diploid genomes. Given tumor purity, you can infer the ploidy of a gene in the sample from the reported fold change.

Given tumor purity X%, for a reported fold change Y, you can calculate the copy number n using the following equation:

$$n = [(200 * Y) - 2 * (100 - X)]/X$$

For example, a tumor purity at 30% and a MET with fold change of 2.2x indicates that 10 copies of MET DNA are observed.

RNA Output

Splice Variant VCF

The splice variant VCF contains all candidate splice variants targeted by the analysis panel identified by the RNA analysis pipeline. The following filters can be applied for each variant call:

Filter Name	Description
LowQ	Splice Variant Score is < a Passing Quality Score threshold value of 1.
PASS	Splice Variant Score is ≥ a Passing Quality Score threshold value of 1.
LowUniqueAlignments	All splice junction supporting reads map to a unique genomic interval near at least one of the two splice sites.

See the headers in the output for more information about each column.

Splice Variant Annotated JSON

If available, each splice variant is annotated using the Illumina Annotation Engine. The following information is captured in the JSON:

- HGNC Gene

- Transcript
- Exons
- Introns
- Canonical
- Consequence

All Fusions CSV

The all fusions CSV file contains all candidate fusions identified by the RNA analysis pipeline. Two key output columns in the file describe the candidate fusions: Filter and KeepFusion.

The following table describes the semicolon-separated output found in the Filter columns. The output is either a confidence filter or information only as indicated. If none of the confidence filters are triggered, the Filter column contains the output PASS, else it contains the output FAIL.

Table 6 Filter Column Output

Filter	Filter Type	Description
DOUBLE_BROKEN_EXON	Confidence filter	If both breakpoints are distant from annotated exon boundaries, the number of supporting reads do not satisfy a high threshold requirement (≥ 10 supporting reads).
LOW_MAPQ	Confidence filter	All fusion supporting read alignments at either of the breakpoints have MAPQ < 20.
LOW_UNIQUE_ALIGNMENTS	Confidence filter	All fusion supporting read alignments map to a unique genomic interval at either of the breakpoints.
LOW_SCORE	Confidence filter	The fusion candidate has probabilistic score as determined by the features of the candidate.
MIN_SUPPORT	Confidence filter	The fusion candidate has very few fusion supporting reads (< 5 supporting read pairs).
READ_THROUGH	Confidence filter	The breakpoints are cis neighbors (< 200 kbp) on the reference genome.
ANCHOR_SUPPORT	Information only	Read alignments of fusion supporting reads are not long enough (12 bp) at either of the two breakpoints.
HOMOLOGOUS	Information only	The candidate is likely a false candidate generated because the two genes involved have high gene homology.
LOW_ALT_TO_REF	Information only	The number of fusion supporting reads is < 1% of the number of reads supporting the reference transcript at either of the two breakpoints.
LOW_GENE_COVERAGE	Information only	Each breakpoint in an enriched gene has fewer than 125 bp with nonzero read coverage.
NO_COMPLETE_SPLIT_READS	Confidence filter	For every fusion-supporting split read, the total number of aligned bases across two breakpoints is less 60% of the read length.
UNENRICHED_GENE	Confidence filter	Neither of the two parent genes is in the enrichment panel.

The KeepFusion column of the output has a value of TRUE when none of the confidence filters are triggered.

Refer to the headers in the output for more information about each column.

Table 7 Fusion Columns

Fusion Object Field	Source
Gene A	The gene associated with the A side of the fusion. A semicolon delimited list is used for multiple genes.
Gene B	The gene associated with the B side of the fusion. A semicolon delimited list is used for multiple genes.
Gene A Breakpoint	**[Information only]** The chromosome and offset of the Gene A side of the fusion.
Gene A Location	Location of the breakpoint within Gene A: • **IntactExon**—Matches exon boundary • **BrokenExon—**Inside an exon • **Intronic—**Within an intron • **Intergenic—**No gene overlap (currently excluded) If multiple genes are in Gene A, then semicolon separated list of locations. This column is used internally to identify genes to report when a breakpoint occurs in a region overlapping multiple genes. Occasionally, additional values are listed for genes that were excluded from the GeneA list
Gene A Sense	Boolean indicating whether left/right breakpoint order suggests fusion transcript is in the same sense of Gene A. If multiple genes are in Gene A, then semicolon separated list of bools.
Gene A Strand	Strand of Gene A, + for forward, – for reverse.
Gene B Breakpoint	**[Information only]** The chromosome and offset of the Gene B side of the fusion.
Gene B Location	Location of the breakpoint within Gene B: • **IntactExon—**Matches exon boundary • **BrokenExon—**Inside an exon • **Intronic—**Within an intron • **Intergenic—**No gene overlap (currently excluded) If multiple genes in Gene B, then semicolon separated list of locations. This column is used internally to identify genes to report when a breakpoint occurs in a region overlapping multiple genes. Occasionally, additional values are listed for genes that were excluded from the GeneB list.
Gene B Sense	Boolean indicating whether left/right breakpoint order suggests fusion transcript is in the same sense of Gene B. If multiple genes are in Gene B, then semicolon separated list of bools.

Fusion Object Field	Source
Gene B Strand	Strand of Gene B, + for forward, – for reverse.
Score	The quality of fusion as determined by DRAGEN server.
Filter	The filter associated with the fusion as determined by the respective caller. Results from different callers are not equivalent.
Ref A Dedup	Gene A uniquely mapping reads paired across or split by the junction. Does not support fusion. Duplicate reads are not included.
Ref B Dedup	Gene B uniquely mapping reads paired across or split by the junction. Does not support fusion. Duplicate reads are not included.
Alt Split Dedup	Uniquely mapping reads split by the junction. Supports fusion. Duplicate reads are not included.
Alt Pair Dedup	Uniquely mapping reads paired across junction. Supports fusion. Duplicate reads are not included.
KeepFusion	The determination whether the fusion should be kept or dropped from the list of fusions.
Fusion Directionality Known	Whether fusion directionality is known and indicated by gene order.

When using Microsoft Excel to view this report, genes that are convertible to dates (such as MARCH1) automatically convert to dd-mm format (1-Mar) by Excel. The following are fusion allow list genes:

- ABL1
- AKT3
- ALK
- AR
- AXL
- BCL2
- BRAF
- BRCA1
- BRCA2
- CDK4
- CSF1R
- EGFR

- EML4
- ERBB2
- ERG
- ESR1
- ETS1
- ETV1
- ETV4
- ETV5
- EWSR1
- FGFR1
- FGFR2
- FGFR3
- FGFR4
- FLI1
- FLT1
- FLT3
- JAK2
- KDR
- KIF5B
- KIT
- KMT2A
- MET
- MLLT3
- MSH2
- MYC
- NOTCH1
- NOTCH2
- NOTCH3
- NRG1
- NTRK1
- NTRK2
- NTRK3

- PAX3
- PAX7
- PDGFRA
- PDGFRB
- PIK3CA
- PPARG
- RAF1
- RET
- ROS1
- RPS6KB1
- TMPRSS2

Block List

The block list represents high noise regions in the panel where false positive variant calls are likely produced. As a result, all positions in the gVCF are marked as `Filter=blacklist` to indicate variant call results are not reliable in such regions.

The block list includes the following genes:

- HLA-A
- HLA-B
- HLA-C
- KMT2B
- KMT2C
- KMT2D
- chrY
- Any position with VAF > 1% occurred in six or more of the 60 baseline samples

Troubleshooting

Failure Type	Actions
Software	Open the log file `./{analysisFolder}/Logs_Intermediates/TruSight-Oncology-500-pipeline-<timestamp>.log file`. The log file displays all commands run by the software and the exit code for each analysis step. If a step fails, review standard output and standard error log files in the folder `./{analysisFolder}/Logs_Intermediates/`.
Samples	Open the final sample biomarker report log file `./{analysisFolder}/Results/SampleID/CombinedVariantOutput.tsv`. If a sample fails an analysis step, the step name appears in the [SAMPLE STATUS] section of the report in the Failed Steps field. Review the log files for the step in `./{analysisFolder}/Logs_Intermediates/{FailedStep}/`.
Samples	Open the log file `./{analysisFolder}/Logs_Intermediates/TruSight-Oncology-500-pipeline-<timestamp>.log file`. If the sample limit exceeds 50 GB, the following message appears with a time stamp and the file path for the sample: `TSO500_SOLID_LIMIT_BASES_PER_SAMPLE - TSO500 Solid exceeded sample limit`

DNA Expanded Metrics

DNA expanded metrics are provided for information only. They can be informative for troubleshooting but are provided without explicit specification limits and are not directly used for sample quality control. For additional guidance, contact Illumina Technical Support.

Metric	Description	Units
TOTAL_PF_READS	Total reads passing filter.	Count
MEAN_FAMILY_SIZE	The sum of the reads in each family divided by the number of families after correction, collapsing, and filtering on supporting reads.	Count
MEDIAN_TARGET_COVERAGE	The median coverage of bases.	Count
PCT_CHIMERIC_READS	Percent of chimeric reads	%
PCT_EXON_100X	Percent of exon bases with greater than 100X coverage	%

Metric	Description	Units
PCT_READ_ENRICHMENT	Percentage of reads that cross any part of the target region vs total reads	%
PCT_USABLE_UMI_READS	The percentage of reads with usable UMIs.	%
MEAN_TARGET_COVERAGE	The mean coverage of bases.	Count
PCT_ALIGNED_READS	Percent of reads that aligned to the reference genome.	%
PCT_CONTAMINATION_EST	Percent of contamination of the sample.	%
PCT_PF_UQ_READS	Percent unique reads passing filter.	%
PCT_TARGET_0.4X_MEAN	Percent target bases with target coverage greater than 0.4 times the mean.	%
PCT_TARGET_100X	Percent target bases with greater than 100X coverage.	%
PCT_TARGET_250X	Percent target bases with greater than 250X coverage.	%

RNA Expanded Metrics

RNA expanded metrics are provided for information only. They can be informative for troubleshooting but are provided without explicit specification limits and are not directly used for sample quality control. For additional guidance, contact Illumina Technical Support.

Metric	Description	Units
PCT_CHIMERIC_READS	Percentage of reads that are aligned as two segments which map to non-consecutive regions in the genome.	%
PCT_ON_TARGET_READS	Percentage of reads that cross any part of the target region vs total reads. A read that partially maps to a target region is counted as on target.	%
SCALED_MEDIAN_GENE_COVERAGE	Median of median base coverage of genes scaled by length. An indication of median coverage depth of genes in the panel.	Count
TOTAL_PF_READS	Total number of reads passing filter.	Count

Metric	Description	Units
GENE_ MEDIAN_ COVERAGE	The median coverage depth of all genes in the panel.	Count
PER_GENE_ MEDIAN_ COVERAGE	The median deduped coverage for each gene. This metric is found in `Logs_Intermediates/RnaQCMetrics`. In the RnaQCMetrics folder, there are subfolders for each sample that contain a `{SampleName}_GeneCoverage.tsv` file.	Count

Technical Assistance

For technical assistance, contact Illumina Technical Support.

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

Illumina Technical Support Telephone Numbers

Region	Toll Free	International
Australia	+61 1800 775 688	
Austria	+43 800 006249	+43 1 9286540
Belgium	+32 800 77 160	+32 3 400 29 73
Canada	+1 800 809 4566	
China		+86 400 066 5835
Denmark	+45 80 82 01 83	+45 89 87 11 56
Finland	+358 800 918 363	+358 9 7479 0110
France	+33 8 05 10 21 93	+33 1 70 77 04 46
Germany	+49 800 101 4940	+49 89 3803 5677
Hong Kong, China	+852 800 960 230	
India	+91 8006500375	
Indonesia		0078036510048
Ireland	+353 1800 936608	+353 1 695 0506
Italy	+39 800 985513	+39 236003759
Japan	+81 0800 111 5011	
Malaysia	+60 1800 80 6789	
Netherlands	+31 800 022 2493	+31 20 713 2960
New Zealand	+64 800 451 650	
Norway	+47 800 16 836	+47 21 93 96 93
Philippines	+63 180016510798	
Singapore	1 800 5792 745	
South Korea	+82 80 234 5300	

Region	Toll Free	International
Spain	+34 800 300 143	+34 911 899 417
Sweden	+46 2 00883979	+46 8 50619671
Switzerland	+41 800 200 442	+41 56 580 00 00
Taiwan, China	+886 8 06651752	
Thailand	+66 1800 011 304	
United Kingdom	+44 800 012 6019	+44 20 7305 7197
United States	+1 800 809 4566	+1 858 202 4566
Vietnam	+84 1206 5263	

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
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

For Research Use Only. Not for use in diagnostic procedures.