illumina®

DRAGEN TruSight Oncology 500 Analysis Software v2.1 (Local)

User Guide

ILLUMINA PROPRIETARY
Document # 200019138 v00
August 2022

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

Table of Contents

Overview

The Illumina® DRAGEN™ TruSight™ Oncology 500 Analysis Software v2.1 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)
- Exon-level 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. With TruSight Oncology 500 Homologous Recombination Deficiency (HRD), genomic instability score (GIS) is also determined. 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 v2.1 allows for analysis on a single DRAGEN server or split across multiple servers.

Compatibility

The DRAGEN TruSight Oncology 500 Analysis Software v2.1 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 v2.1. Using bcl2fastq does not produce the same results and is discouraged. See the DRAGEN TruSight Oncology 500 Analysis Software v2.1 support pages on the Illumina support website for settings and compatibility information for using DRAGEN TruSight Oncology 500 Analysis Software v2.1 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 or v4.

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 7.9 operating system, or later, is provided.
- Before installing DRAGEN TruSight Oncology 500 Analysis Software, container engine Docker v20.10 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, use the default settings or 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)	Minimum Disk Space (Gb)
NextSeq 500/550 and 550Dx HO flow cell	32–55	82–85	150
NovaSeq 6000 SP flow cell	85–100	250–374	300
NovaSeq 6000 S1 flow cell	164–200	360–665	800
NovaSeq 6000 S2 flow cell	290–460	890–1600	1500
NovaSeq 6000 S4 flow cell	800–1200	2700–4100	3000

When launching an analysis, the software checks that the minimum disk space required is available If the minimum disk space is not available, the software shows an error message and prevents analysis from starting. If disk space is exhausted during a run, the run shows an error and stops analyzing.

Permissions

Illumina recommends the following permissions for installing and running software:

- A non-root user can only run Docker if they're a member of the docker group. For more information on Docker permission requirements and alternatives to running as root, refer to the Docker documentation available on the Docker website.
- Installing and uninstalling the DRAGEN TruSight Oncology 500 Analysis Software and running the system check requires root privileges.
- Run DRAGEN TruSight Oncology 500 Analysis Software without being logged in as a root user. It is possible to run the DRAGEN TruSight Oncology 500 Analysis Softwareas root, but this method is not recommended or required.

Install DRAGEN TruSight Oncology 500 Analysis Software

In addition to DRAGEN TruSight Oncology 500 Analysis Software, the installation script installs the required DRAGEN server software dependencies (DRAGEN 3.10.9 for CentOS 7.9.) The script uninstalls any existing DRAGEN server software versions. To uninstall DRAGEN TruSight Oncology 500 Analysis Software, refer to *Uninstall DRAGEN TruSight Oncology 500 Analysis Software on page 4*.

To run the product successfully on DRAGEN v3 servers, you'll need to update the Docker configuration to store images and other working files in the `/staging` directory of the DRAGEN server, where there is sufficient storage space. To do this, you will need root privileges to reconfigure Docker.

Installation Instructions

1. Contact Illumina Customer Care to obtain the DRAGEN TruSight Oncology 500 Analysis Software installer package. will
2. Download the DRAGEN TSO 500 installation package provided in the email from Illumina. The link expires after 7 days.
3. Make sure that no other analysis is being performed.
   Installing DRAGEN TruSight Oncology 500 Analysis Software while performing other analysis may cause interference between the operation of the installer and running Docker containers.
4. Install Docker v20.10 or later using the install instructions for CentOS provided in the Docker documentation.
5. To move any existing docker data from `/var/lib/docker` to `/staging/docker`, enter the following command:

   ```
   rsync -aqxP /var/lib/docker /staging/docker
   ```

6. Change the output directory in the Docker configuration file.
   The default location of the configuration file is `/etc/docker/daemon.json`. You can change the default location during Docker installation by using the `–config-flag` parameter.

   If you have not configured Docker to store the configuration file in a non-default location, do as follows.

   a. Check if the configuration file exists by using the following command:

   ```
   /etc/docker/daemon.json
   ```

   b. If the file does not exist,use the following command to create the file:

   ```
   touch /etc/docker/daemon.json
   ```

   c. If the following text pairs exist in the `/etc/docker/daemon.json` file, update the pairs to match the following text . If the pairs don’t exist, add them to the file.

   ```
   {"data-root": "/staging/docker/",
   "experimental": true}
   ```

   If you have explicitly configured docker to have its configuration file at a non-default location do as follows.

   a. Open your Docker configuration file.

   b. If the following text pairs exist in your configuration file, update the pairs to match the the following text. If the pairs don’t exist, add them as follows:

   ```
   {"data-root": "/staging/docker/",
   "experimental": true}
   ```

7. When Docker has been reconfigured, restart Docker using the following command:

```
systemctl enable docker.service
```

8. Copy the install script to the `/staging` directory to store the script in the directory.
9. Use the following command to update the run script permission:

```
chmod +x /staging/install_DRAGEN_TSO500-2.1.0.run
```

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

```
TMPDIR=/staging /staging/install_DRAGEN_TSO500-2.1.0.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.

11. Use the following command to build the DRAGEN server hash table, which runs for approximately 60 minutes:

```
/usr/local/bin/build-hashtable_DRAGEN_TSO500-2.1.0.sh
```

12. Install your DRAGEN server licenses as follows.
    - If your servers is connected to the internet, activate your DRAGEN TSO 500 software licenses as follows.
      a. Test and confirm that the server is connected to the internet.
      Example: `ping www.illumina.com`
      b. To activate the license, enter: `/opt/edico/bin/dragen_lic -i auto`.
    - For servers ***not*** connected to the internet, contact Illumina Customer Service for license information.
13. After installing DRAGEN server licenses, generate a list of installed DRAGEN server licenses by running the following command:

```
/opt/edico/bin/dragen_lic
```

If license installation is successful, the list should include `TSOCombined`. If you have a license for HRD, it should include `TSO500_HRD` and `Genome`.
If the expected licenses are not installed, contact Illumina Customer Care.

Running the System Check

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

```
/usr/local/bin/check_DRAGEN_TSO500-2.1.0.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 Analysis Softwaresuccessfully 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.

If using MacOS, an error can occur if the local settings are not in English. To resolve the error, disable the ability to set environment variables automatically in Terminal settings.

Uninstall DRAGEN TruSight Oncology 500 Analysis Software

The DRAGEN TruSight Oncology 500 Analysis Software installation includes an uninstall script called `uninstall_DRAGEN_TSO500-2.1.0.sh`, which is installed in `/usr/local/bin`.

Executing the uninstall script removes the following assets:

- All scripts (`build-hashtable_DRAGEN_TSO500-2.1.0.sh, check_DRAGEN_TSO500-2.1.0.sh, DRAGEN_TSO500.sh, uninstall_DRAGEN_TSO500-2.1.0.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_TSO500-2.1.0.sh
```

You are not required to uninstall Docker or the DRAGEN server software (3.10.9). To remove Docker, review the install instructions for CentOS provided in the Docker documentation.

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 10. 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.
- Run folder needs to be intact, refer to *Starting From BCL Files* on page 12 for input requirements.
- If the analysis output folder path is different from the default, provide the analysis output folder path. Refer to *Command-Line Options* on page 10.

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 DRAGEN TruSight Oncology 500 Analysis Software supports sample sheets v2.

See the Illumina support site for the appropriate sample sheet template for your run. The compatible sample sheet templates are also provided in the resource file folder that is installed with the software.

The sample sheet is made up of a list of samples and their index sequences, along with optional sample information. DNA samples enriched using TruSight Oncology 500 HRD must be indicated in the Sample Feature column of the sample sheet. 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 sample sheet v2 template, which is available in two locations:
   - TruSight Oncology 500 pages on the Illumina support site
   - In the resource bundle installed with the software (`/staging/illumina/DRAGEN_TSO500/resources/sampleSheet/`)

2. In the BCL Convert Settings section, enter the following required parameters.

Sample Parameter	Required	Details
SoftwareVersion	Yes	Enter `2.1`
AdapterRead1	Yes	If using 8 bp indices: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA If using 10 bp indices: CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
AdapterRead2	Yes	If using 8 bp indices: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT If using 10 bp indices: CTGTCTCTTATACACATCTGACGCTGCCGACGA
AdapterBehavior	Yes	Enter `trim`. This indicates that the BCL Convert software trims the specified adapter sequences from each read.
MinimumTrimmedReadLength	Yes	Enter `35`. Reads with a length trimmed below this point are masked.
MaskShortReads	Yes	Enter `35`. Reads with a length trimmed below this point are masked.

3. In the BCL Convert Data section, enter the following parameters for each sample.

Sample Parameter	Required	Details
Sample_ID	Yes	Must match a Sample_ID listed in [TSO500S_Data] section.
Index	Yes	Index 1 sequence valid for Index_ID assigned to matching Sample_ID in [TSO500S_Data] section.
Index2	Yes	Index 2 sequence valid for Index_ID assigned to matching Sample_ID in [TSO500S_Data] section.
Lane	Only for NovaSeq 6000 XP workflows	Indicates which lane corresponds to a given sample. Enter a single numeric value per row. Cannot be empty if a header is present. Where lane value is specified, Index IDs provided must be unique per lane.

4. In the TSO 500 Data section, enter the following parameters for each sample.

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. • 1–40 characters. • 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`.
Index_ID	Yes	The Index adapter ID used for the sample. Must be unique per sample unless samples are assigned different lanes. Indexes UP, CP, and UDP are valid for DNA samples, but RNA samples can only use indexes UP and UDP.
Sample_Type	Yes	Enter DNA or RNA.
Pair_ID	Yes	Use to pair DNA and RNA samples from the same individual. Use a shared pair ID to link two samples.
Sample_Feature	No	Only required for HRD enriched samples. For DNA samples that have undergone HRD enrichment, enter `HRD` in this column of the sample sheet. If the sample has not undergone HRD enrichment, leave the field empty.
Sample_ Description	No	Sample description must meet the following requirements: • 1–50 characters • Alphanumeric characters with underscores, dashes and spaces. If you enter a underscore, dash, or space, enter an alphanumeric character before and after. Example: Solide-FFPE_213

5. Make sure the sample sheet does not contain any blank lines at the bottom. The lines will cause the analysis to exit unsuccessfully.
6. 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	No	Path to the local analysis folder. The default location is `/staging/DRAGEN_TSO500_Analysis_{timestamp}`. If not using the default location, provide the full path to the local analysis folder. Folder must have sufficient space and must be on an NVMe SSD drive. For example, the `/staging` directory on the DRAGEN server. Refer to table in*Installation Requirements* on page 2 for minimum disk space requirements.
--resourcesFolder	No	Path to the resource folder location. The default location is `/staging/illumina/DRAGEN_TSO500/resources`. If not using the default location, enter the full path to the resource 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.
--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.

Option	Required	Description
--sampleOrPairIDs	No	Provide the comma-delimited sample or pair IDs that should be processed on this node with no spaces. For example, `Pair_1,Pair_2,Sample_1`.
--demultiplexOnly	No	Demultiplex to generate FASTQ only without additional analysis.
--gather	No	Follow this option with any directories whose results should be gathered into a single Results folder.
--hashtableFolder	No	Defaults to the DRAGEN hash table location created upon install. Only needs to be specified if the user changes this location.

Refer to *Command-Line Options* on page 10 for additional commands.

Use full paths when specifying the file paths in the command line. Avoid special characters such as &, *, #, and spaces. When starting from BCL files, only the run folder needs to be specified. The immediate parent directory containing the BCL files does not need to be specified.

When running the analysis software using SSH, Illumina recommends using additional software to prevent unexpected termination of analysis. Illumina recommends `screen` and `tmux`.

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:

   ```
   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 \
   --runFolder /staging/{RunFolderName} \
   --analysisFolder /staging/{AnalysisFolderName}
   ```

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

   ```
   DRAGEN_TSO500.sh \
   --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 \
   --runFolder /staging/{RunFolderName} \
   ```

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

- Start from FASTQ with sample sheet included in the FASTQ folder, and specifying different resources and hash table folders.

```
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 \
--fastqFolder /staging/{FastqFolderName} \
--analysisFolder /staging/{AnalysisFolderName} \
--sampleOrPairIDs “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 starting from BCL files, 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 13.
- 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.

BCL Convert has been set to write UMI sequences to the read headers in the FASTQ files.

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.

   To sequence runs on multiple DRAGEN servers 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 `--sampleOrPairIDs` command-line option is present in the analysis launch. Failure to account for these checks results in an error.
5. Copy the results from demultiplexing and each analysis run onto a single server, and then generate the final `/Results` directory, which contains the aggregated results. To do this, use the `--gather` command followed by the output directories of the demultiplexing step and each individual analysis run.

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} --sampleOrPairIDs 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} --sampleOrPairIDs 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

After sequencing data are collected, they are processed by the DRAGEN TruSight Oncology 500 Analysis Software module to perform quality control, detect variants, determine Tumor Mutational Burden (TMB), Microsatellite Instability (MSI) status, and Genomic Instability Score (GIS), and report results. The following sections describe the analysis methods.

FASTQ Generation

Sequencing data stored in BCL format is demultiplexed through a process that uses the index sequences unique to each sample to assign clusters to the library from which they originated. Each cluster contains two indexes (i7 and i5 sequences, one at each end of the library fragment), and the combination of those index sequences are used to demultiplex the pooled libraries.

After demultiplexing, this process generates FASTQ files, which contain the sequencing reads for each individual sample library and the associated quality scores for each base call, excluding reads from any clusters that did not pass filter.

DNA Analysis Methods

DNA Alignment and Error Correction

DNA alignment and error correction involves aligning sequencing reads derived from DNA libraries to a reference genome and correcting errors in the sequencing reads prior to variant calling.

DRAGENUMI error correction comprises three main steps:

1. DRAGEN UMI uses its HW accelerated mapper (based on a hash table implementation) to align DNA sequences in FASTQ files to the hg19 reference genome. These alignments are not written to a BAM.
2. The raw alignments are processed to remove errors,including errors introduced during FFPE preservation, PCR amplification, and sequencing. Reads from the same original DNA molecule are tagged with the same unique molecular identifier (UMI) during library preparation. The UMI allows DRAGEN to compare related reads, remove outlier signals, and collapse multiple reads into a single high-quality sequence. 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.
3. DRAGENperforms a final alignment step on the UMI-collapsed reads. These final alignments are then written to a BAM file and a corresponding BAM index file is created.

DRAGEN continues to use these final alignments as input for gene amplification (copy number) calling, small variant calling (SNVs, INDELs, MNVs, DELINs), microsatellite instability (MSI) status determination, and DNA library quality control.

Small Variant Calling and Filtering

DRAGENsupports calling SNVs, insertions, deletions, MNVs, and delins in tumor-only samples by using mapped and aligned DNA reads from a tumor sample as input.t. Variants are detected via both column wise pileup analysis and local *de novo* assembly of haplotypes. The *de novo* haplotypes allow the detection of much larger insertions and deletions than possible through column wise pileup analysis only. DRAGEN insertions and deletions are validated with lengths of at least 0–25 bp. In addition, DRAGEN also uses the *de novo* assembly to detect SNVs, insertions, and deletions that are co-phased and part of the same haplotypes. Any such co-phased variants that are within a window of 15 bp can then be reassembled into complex variants (MNVs and DELINS). The tumor-only pipeline produces a VCF file containing both germline and somatic variants that can be further analyzed to identify tumor mutations. The pipeline makes no ploidy assumptions, enabling detection of low-frequency alleles.

DRAGEN small variant calling includes the following steps:

1. Detects regions with sufficient read coverage (callable regions).
2. Detects regions where the reads deviate from the reference and there is a possibility of a germline or somatic call (active regions).
3. Assembles *de novo* graph haplotypes are assembled from reads (haplotype assembly).
4. Extracts possible somatic or germline calls (events)from column wise pileup analysis.
5. Calibrates read base qualitiesto account for FFPE noise.
6. Computes read likelihoods for each read/haplotype pair.
7. Performs mutation calling by summing the genotype probabilities across all reads/haplotype pairs.
8. Performs additional filtering to improve variant calling accuracy, including using a systematic noise file. The systematic noise file indicates the statistical probability of noise at specific positions in the genome. This noise file is constructed using clean (normal) samples. Regions where noise is common (eg, difficult to map regions) have higher noise values. The small variant caller penalizes those regions to reduce the probability of making false positive calls. Filters used by TSO 500are listed in the DNA Outputs section.

Copy Number Variant Calling

DRAGEN can identify gene amplifications in tumor-only samples.

DRAGEN gene amplification calling include the following steps:

1. Counts target per region covered by the panel.
2. Utilizes preprocessed panel of normal samples to normalize target counts.
3. Corrects for differences in coverage due to the GC bias.
4. Applies a statistical model to calculate scores of a CNV event from observed coverage and makes copy number calls.

Exon Level Copy Number Variant Calling

The BRCA large rearrangement step generates segmentation of the BRCA1 and BRCA2 genes for exon-level CNV detection from the BAM file. Using the same method as CNV calling, the large rearrangement component counts coverage of each target interval of the panel, performs normalization, and calculates the fold change values for each probe across the BRCA genes. Normalization includes GC bias correction, sequencing depth, and probe efficiency using a collection of normal FFPE and genomic DNA samples. Initial segmentation is performed for each gene with circular binary segmentation. The merging of segments is then determined by amplitude, noise, and variance at adjacent segments using thresholds established with *in silico* data. A large rearrangement is reported for genes with more than one segment. Coordinates of the exon-level CNV and the log2 mean fold change for each of the BRCA gene segments are found in the `*_DragenExonCNV.json` file.

Annotation

The Illumina Annotation Engine performs annotation of small variants, CNVs, and exon-level CNVs. 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

Tumor Mutational Burden

DRAGEN is used to compute Tumor Mutational Burden (TMB) in coding regions where there is sufficient coverage.

The following variants are excluded from the TMB calculation:

- Non-PASS variants
- Mitochondrial variants
- MNVs
- Variants that do not meet a minimum depth threshold
- Variants that do not meet the minimum variant allele threshold
- Variants that fall outside the eligible regions
- Tumor driver mutations. Variants with a population allele count ≥ 50 are treated as tumor driver mutations

Germline variants are not counted towards TMB. Variants are determined as germline based on a database and a proxy filter.

Variants with a population allele count ≥ 50 that are observed in either the 1000 Genomes or gnomAD databases are marked as germline. The proxy filter scans the variants surrounding a specific variant and identifies those variants with similar variant allele frequencies (VAF). If the majority of surrounding variants of similar VAF are germline, then the variant is also marked as germline.

The formula for TMB calculation is:

- TMB = *Filtered Variants / Eligible Region size (Mbp)*
- Nonsynonymous TMB = *Filtered Nonsynonymous Variants / Eligible Region size (Mbp)*

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

Microsatellite Instability Status

DRAGENis used to determine the MSI status of a sample. It uses a normal reference file, which was created from a set of normal samples. During sequencing, normal reference files are generated by tabulating read counts for each microsatellite site. The normal file contains the read count distribution for each microsatellite.

MSI calling for a tumor-only sample is performed by first tabulating tumor counts from the read alignments for each microsatellite site. Then, the Jensen-Shannon distance (JSD) is calculated between each pair of tumor and normal baseline samples. DRAGEN determines unstable sites by performing Chi-square testing of tumor JSD and normal JSD distributions. Unstable sites are called if the mean distance difference of the two JSD distributions is greater than or equal to the distance threshold and Chi-square p-value is ≤to the p-value threshold. Lastly, DRAGEN produces an MSI status given assessed site count, unstable site count, the percentage of unstable sites in all assessed sites, and the sum of the Jensen-Shannon distance of all the unstable sites.

Genomic Instability Score

Genomic Instability Score (GIS) is a whole genome signature for homologous recombination deficiency. The GIS is composed of the sum of three components: loss of heterzygosity, telomeric allele imbalance, and large-scale state transition. These components are estimated using the GIS algorithm contracted from Myriad Genetics, which uses an input of the b-allele frequency and coverage across a genome-wide single nucleotide panel. A panel of normal samples is used for both bias reduction and normalization prior to GIS estimation. Final GIS results can be found in the `*.gis.json` file.

Contamination Detection

The contamination analysis step detects foreign human DNA contamination using the SNP error file and pileup file that are generated during the small variant calling and the TMB trace file. The software determines whether a sample has foreign DNA using the contamination score. In contaminated samples, the variant allele frequencies in SNPs shift from the expected values of 0%, 50%, or 100%. The algorithm collects all positions that overlap with common SNPs that have variant allele frequencies of < 25% or > 75%. Then, the algorithm computes the likelihood that the positions are an error or a real mutation. The contamination score is the sum of all the log likelihood scores across the predefined SNP positions with minor allele frequency < 25% in the sample and are not likely due to CNV events.

The larger the contamination score, the more likely there is foreign DNA contamination. A sample is considered to be contaminated if the contamination score is above predefined quality threshold. The contamination score was found to be high in samples with highly rearranged genomes or HRD samples. 1% of HRD samples found to be above the threshold with no evidence for actual contamination.

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.

RNA Alignment and Fusion Detection

RNA alignment and fusion detection uses trimmed reads in FASTQ format as input. The outputs include a BAM file which contains duplicate-marked read alignments, an `SJ.out.tab` file which contains unannotated splice junctions, and a CSV file which contains fusion candidates.

DRAGEN aligns RNA reads in a transcript-aware mode using the human hg19 genome containing unplaced contigs (ie,, chrUn_gl regions) and uses GENCODEv19 transcript annotations to identify splice sites. DRAGEN 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 identifies fusion candidates using chimeric split read alignments (pairs of primary and supplementary alignments) against multiple genes. DRAGEN 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 DRAGEN RNA Fusion caller identifies gene fusions by searching for chimeric reads spanning two distinct parent genes. Based on the chimeric reads, DRAGEN first creates a list of fusion candidates, then scores the candidates to report the list of high confidence fusion calls from the candidate pool.

DRAGEN RNA Fusion caller performs the following steps:

1. Generates fusion candidate generation based on split read alignment.
2. Recruits additional evidence from fusion supporting discordant read pairs and soft-clipped reads.
3. Computes fusion candidate features such as gene coverage, read mapping quality, alternate allele frequency, gene homology, alignment anchor length, and breakpoint distance from exon boundary.

4. Scores and ranks the fusion candidates using a logistic regression model.
5. Selects a final list of fusion calls based on score and other filters including number of supporting reads, unique read alignment count, read through transcripts, and fusions matching the enriched regions.

Splice Variant Calling

RNA splice variant calling is performed for RNA sample libraries. Candidate splice variants (junctions) from RNA Alignment are compared against a database of known transcripts and a splice variant baseline of non-tumor junctions generated from a set of normal FFPE samples from different tissue types. Any splice variants that match the database or baseline are filtered out unless they are in a set of junctions with known oncological function. If there is sufficient read support, the candidate splice variant is kept. This process also identifies candidate RNA fusions.

RNA Fusion Merging

Fusions identified during RNA fusion calling are merged with fusions from proximal genes identified during RNA splice variant calling. These are then annotated with gene symbols or names with respect to a static database of transcripts (GENCODE Release 19). The result of this process is a set of fusion calls that are eligible for reporting.

RNA Splice Variant Annotation

The Illumina Annotation Engine annotates detected RNA splice variant calls with transcript-level changes (ie, affected exons in a gene's transcript) with respect to RefSeq. This RefSeq database is the same RefSeq database used by the small variant annotation process.

Analytical Performance Testing

Illumina tested the analytical performance of variant calling using several approaches. The first approach covers the entire workflow including library preparation, sequencing, and secondary analysis. Illumina tests a diverse selection of variants with this approach. When the variant-calling pipeline is expanded to call a new variant class, this approach is always utilized.

To further expand the scope of variants tested and ensure that a wide range of clinically relevant variants, including rare variants, are examined, we also leverage *in silico* testing. For this testing approach, variants of interest are extracted from public databases like Cosmic and ClinVar. Each variant is simulated at different VAF levels by spiking in mutant reads into a normal FFPE background. The simulated reads match the expected quality of typical FFPE samples: fragment length, error rate, and family size. After the simulation, the samples with spiked-in variants are processed via the pipeline and the sensitivity and corresponding c95 level are determined. This approach was recently used to complement the laboratory testing to evaluate the performance of complex variants and larger insertions and deletions.

Quality Control

The DRAGEN TruSight Oncology 500 Analysis Software v2.1 includes several quality control analyses.

Run QC

The Run Metrics section of the Metrics Output report provides sequencing run quality metrics along with suggested values to determine if they are within an acceptable range. The overall percentage of reads passing filter is compared to a minimum threshold. For Read 1 and Read 2, the average percentage of bases ≥ Q30, which gives a prediction of the probability of an incorrect base call (Q-score), are also compared to a minimum threshold. The following tables below run metric and quality threshold information for different systems.

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.

Table 2 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 3 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

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

DNA Sample QC

DRAGEN TruSight Oncology 500 Analysis Software v2.1 uses QC metrics to assess the validity of small variant calling, TMB, MSI, and gene amplifications for DNA libraries that pass contamination quality control. If the library fails one or more quality metrics, then the corresponding variant type or biomarker is not reported, and the associated QC category in the report header displays FAIL. Additionally, a companion diagnostic result may not be available if it relies on QC passing for one or more of the following QC categories.

DNA library QC results are available in the `MetricsOutput.tsv` file, refer to *Metrics Output on page 25* for details.

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
CONTAMINATION_SCORE	The contamination score is based on VAF distribution of SNPs.	Contamination Score ≤ 1457	All
MEDIAN_EXON_COVERAGE	Median exon fragment coverage across all exon bases.	≥ 150	Small variant TMB
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

Metric	Description	Recommended Guideline Quality Threshold	Variant Class
GENE_SCALED_MAD	Median of absolute deviations (MAD) from the median of the normalized count following a correction for the median normalized count for each gene of each CNV target region after excluding genes with a potential deletion.	≤ 0.134	CNV
MEDIAN_BIN_COUNT_CNV_TARGET	The median raw bin count per CNV target.	≥ 1.0	CNV
PCT_TARGET_HRD_50X	**[HRD]** Percent HRD probes with 50X fragment coverage. Only reports for HRD samples.	≥ 50.0	GIS

RNA Sample QC

The input for RNA Library QC is RNA alignment. Metrics and guideline thresholds can be found in the `MetricsOutput.tsv` file, refer to *Metrics Output on page 25* for details.

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
GENE_MEDIAN_COVERAGE	The median deduped coverage across all genes in the RNA panel (55 genes).	N/A*	Fusion Splice

* To avoid failing RNA samples unnecessarily, Illumina does not recommend a universal threshold to determine RNA sample quality. RNA expression varies significantly across tissue types and a small panel size (55 genes), which makes normalization challenging. Tissue-specific thresholds could be considered for normalization.

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:

- DNA library QC metrics for:
  - Small variant calling
  - TMB

  - MSI
  - CNV
  - GIS (if TruSight Oncology 500 HRD is run)
- RNA library QC metrics
- 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 include 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.

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

AdditionalSarjMetrics— Contains per pair ID calculations to support the PCT_TARGET_250X metric.

Annotation—Contains outputs for small variant annotation.

Subfolders per sample ID—Contains the aligned small variants JSON.

CombinedVariantOutput

Subfolders per pair ID—Contains the combined variant output TSV files.

A combined output log file.

Contamination

Subfolders per DNA sample ID—Contains the contamination metrics JSON and output logs.

DnaDragenCaller

Subfolders per sample ID—Contains the aligned BAM and index files, small variant VCF and gVCF, copy number variant VCF, MSI JSON, and QC outputs in CSV format.

DnaDragenExonCNVCaller

Subfolders per DNA sample ID—Contains the exon-level CNV JSON along with supporting calculation and QC files.

DnaFastqValidation—Contains the FASTQ validation output log for DNA samples.

FastqDownsample

- Subfolders per RNA sample ID—Contains FASTQ files and output logs.
- FastqDownsample output

Gis—Contains GIS-related files for HRD samples.

- Subfolders per HRD sample ID—Contains the GIS JSON, along with supporting calculation and QC files.

LrAnnotation

- Subfolders per DNA sample ID—Contains the annotated exon-level CNV JSON.

LrCalculator

- Subfolders per DNA sample ID—Contains the exon-level CNV VCF.

MetricsOutput

- Subfolders per pair ID—Contains the metrics output TSV files.
- A combined output log file.

ResourceVerification—Contains the resource file checksum verification logs.

RnaAnnotation

- Subfolders per RNA sample ID—Contains the annotated splice variant JSON.

RnaDragenCaller

- Subfolders per sample ID—Contains the aligned BAM, fusion candidates CSV, and QC outputs in CSV format.

RnaFastqValidation—Contains the FASTQ validation output log for RNA samples.

RnaFusion

- Subfolders per RNA sample ID—Contains the All Fusions CSV and Fusion Processor logs.

RnaQcMetrics

- Subfolders per RNA sample ID—Contains the RNA QC metrics JSON.

RnaSpliceVariantCalling

- Subfolders per RNA sample ID—Contains the splice variants VCF.

Run QC—Contains the Run QC metrics JSON, Intermediate Run QC metrics JSON, and log file.

SampleAnalysisResults

- Subfolders per pair ID—Contains the Sample Analysis Results JSON and detailed log file.

SampleSheetValidation—Contains the Intermediate sample sheet and validation log.

Tmb

- Subfolders per DNA sample ID—Contains the TMB metrics CSV, TMB trace TSV, and related files and logs.
  - `passing_sample_steps.json`—Contains the steps passed for each sample ID.

`pipeline_trace.txt`—Contains a summary and troubleshooting file that lists each Nextflow task executed and the statusstatus (for example, COMPLETED or FAILED).

`run.log`—Contains a complete trace-level log file describing the Nextflow pipeline execution.

`run_report.html`—Contains high-level run statistics (performance, usage, etc.)

`run_timeline.html`—Contains timeline-related information about the analysis run.

- Results
  - Metrics Output TSV (all pair IDs)
  - Pair ID—The following outputs are produced for each sample:
    - Combined Variant Output TSV
    - Metrics Output TSV
    - TMB Trace TSV
    - Small Variant Genome VCF
    - Small Variant Genome Annotated JSON
    - Copy Number Variant VCF
    - GIS JSON
    - Exon-level CNV VCF
    - Exon-level CNV Annotated JSON
    - 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.
- NodeX_Output—The following outputs are produced for each node used.
  - Logs_Intermediates
  - Results—Contains results only for the samples run on the node.
- Results—Contains results for all samples from all nodes.

Combined Variant Output

File name: `{Pair_ID}_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 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
- **[HRD]** GIS
- Exon-level CNVs

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) • Sample Feature (**HRD**) • 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 hard-filtered 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.
- **Biomarker GIS**—Present only when TruSight Oncology 500 HRD is run.
- **Exon-Level CNV**—Exon-level CNVs must meet the following conditions:
  - BRCA1 or BRCA2 contains at least one affected exon
  - ALT field is <DUP> or <LOSS>

DNA Output

Small Variant gVCF

File name: `{SAMPLE_ID}_hard-filtered.gvcf.gz`

The small variant genome variant call file contains information on all candidate small variants evaluated, including complex variants up to 15 bp from phased variant calling across the entire TSO 500 panel.

The variant status is determined by the FILTER column in the genome VCF as follows.

Filter	Note
PASS	PASS variants.
base_quality	Site filtered because median base quality of alt reads at this locus does not meet threshold.
filtered_reads	Site filtered because the fraction of reads is too large.

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_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.
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.
no_reliable_supporting_read	Site filtered because no reliable supporting somatic read exists.
read_position	Site filtered because median of distances between start/end of read and this locus is below threshold.
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 (SQ) does not meet threshold.
systematic_noise	Site filtered based on evidence of systematic noise in normal sample.
excluded_regions	Site overlaps with VC excluded regions bed.

Small Variant Annotated JSON

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

The small variants annotated file provides variant annotation information for all nonreference positions from the 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 v2.1. 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 the 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 DRAGEN RNA 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 4 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 5 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=excluded_regions` 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% occurrence in six or more of the 60 baseline samples

Troubleshooting

Failure Type	Actions
Software	Open the log file `./{AnalysisFolder}/Logs_Intermediates/pipeline_trace.txt`. This log file displays each pipeline step run by the Nextflow workflow manager software. If a step fails, it is marked as FAILED. Each step generates logs files that are stored in step-specific subfolders in the `Logs_Intermediates` folder. Review the logs files in the relevant `Logs_Intermediates` folder for the step to identify potential sources of error.
Samples	Open the combined metrics output results file `./{AnalysisFolder}/Results/{PairId}/MetricsOutput.tsv`. If a sample fails an analysis step, the Pair ID containing that sample shows the failure under FAILED_STEPS in the Analysis Status section, and COMPLETED_ALL_STEPS are as False. If available, review the individual log files for the failed steps under `./{AnalysisFolder}/Logs_Intermediates` to identify potential sources of error.
Multinode Gather	If the following error appears, check whether the sample or pair ID was included multiple times during separate node analysis runs, before being gathered together. If the error exists, rerun one of the analyses without the duplicate and reattempt gathering. `ERROR:Gather:Destination file ... already exists - check if the same sample ID is in multiple input folders`

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.	%

Metric	Description	Units
PCT_EXON_100X	Percent of exon bases with greater than 100X coverage.	%
PCT_READ_ ENRICHMENT	Percentage of reads that cross any part of the target region vs total reads. This metric considers only non-HRD probe panel probes; therefore, if HRD probes are applied for the sample, lower values could be observed.	%
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.	%
PCT_TARGET_50X	Percent target bases with greater than 50X coverage.	%
[HRD] ALLELE_ DOSAGE_RATIO	Allele dosage ratio measure of noise calculated during the GIS step.	Count
[HRD] MEDIAN_TARGET_ HRD_COVERAGE	The median target coverage of HRD probes.	Count

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
GENE_MEDIAN_COVERAGE	The median coverage depth of all genes in the panel.	Count
GENE_ABOVE_MEDIAN_CUTOFF	Number of genes above the median coverage cutoff.	Count

Resources & References

The DRAGEN TruSight Oncology 500 Analysis Software v2.1 support pages on the Illumina support site provide additional resources. These resources include training, compatible products, and other considerations. Always check support pages for the latest versions.

Revision History

Document	Date	Description of Change
Document # 200019138 v00	August 2022	Initial release

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	
Spain	+34 800 300 143	+34 911 899 417

Region	Toll Free	International
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.

illumina®