



TruSight Oncology 500 v2.1 Local App

User Guide

ILLUMINA PROPRIETARY

Document # 1000000113216 v02

September 2021

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).

© 2021 Illumina, Inc. All rights reserved.

All trademarks are the property of Illumina, Inc. or their respective owners. For specific trademark information, see www.illumina.com/company/legal.html.

Table of Contents

Introduction	1
Install TruSight Oncology 500	2
Running TruSight Oncology 500	4
Sample Sheet Requirements	4
Command Line Arguments	6
Starting From BCL Files	6
Starting From FASTQ Files	7
Running TruSight Oncology 500 on Multiple Nodes	7
Running TruSight Oncology 500	8
Running TruSight Oncology 500: NovaSeq 6000	9
Analysis Methods	10
FASTQ Generation	11
DNA Analysis Methods	12
RNA Analysis Methods	18
Quality Control	21
Analysis Output	23
Single Node Analysis Output Folder Structure	23
Multiple Node Analysis Output Folder Structure	26
Combined Variant Output	26
Metrics Output	28
DNA Output	28
RNA Output	31
Uninstall TruSight Oncology 500	36
Troubleshooting	37
Revision History	38
Technical Assistance	39

Introduction

The Illumina® TruSight™ Oncology 500 Local App generates sequencing libraries for DNA and RNA from formalin-fixed, paraffin-embedded (FFPE) tissue samples. The software supports run setup, sequencing, and analysis for the prepared DNA and RNA libraries. DNA library analysis outputs include tumor mutational burden, variant call files for small and complex variants, microsatellite instability, and gene amplifications. RNA library analysis outputs include fusions and splice variant call files. TruSight Oncology 500 Local App supports up to 384 libraries from DNA or RNA or a combination of DNA and RNA libraries. The TruSight Oncology 500 Local App allows for single node or multiple node analysis.

Installation Requirements

The following subsections are the minimum operating requirements for TruSight Oncology 500 Local App analysis.

Hardware

- Minimum:
 - 64 GB RAM
 - 64-bit CPU with 12 physical cores, 48 threads
 - Dual Intel Xeon 64-bit CPU; E5-2620 v4 or later. Processor must support the AVX instruction set.
 - 1 TB HDD free space
- Recommended:
 - 128 GB RAM, 2400 MHz DDR4 RDIMM ECC
 - 64-bit CPU with 24 physical cores, 24 threads
 - 1 TB HDD free space
 - Enterprise-grade solid state drive (SSD)

Software

- md5sum
- Linux CentOS 7.4 operating system, or later
- Container engine Docker CE v17.05 or Singularity 3.1.0s or later
 - If you have Network File System (NFS) mounted volumes for analysis output, use Singularity or contact Illumina Technical Support
- HTTP client such as Curl or Wget

Storage

- Total disk space for app installation is 128 GB.

- The app download file is 50 GB, and is a single *.gzip compressed tar archive. The zipped file can be deleted after decompression.
- The app download file requires 72 GB of space after decompression.
- Network File System (NFS) is recommended for mounted volumes. Use of Common Internet File System (CIFS) can result in pipeline failures during the alignment step.
- A sequencing run using a single node requires a minimum of 400 GB of free space.
 - A single node analysis run can produce 200 GB of output.
- For a sequencing run using multiple nodes, each node used requires 400 GB of free space.
 - A multinode sequencing run can produce 100 GB–4 TB of output.

The free space required to store sequencing and secondary analysis output varies depending on the sequencing system used. Consult the TruSight Oncology 500 Local App support page on the Illumina support website for additional information on storage requirements.

Compatibility

The TruSight Oncology 500 Local App support pages on the Illumina [support website](#) provide information on compatibility with Illumina sequencing systems.

Use BCL Convert to produce FASTQ files for TruSight Oncology 500 Local App. **Using bcl2fastq does not produce the same results and is discouraged.** See the TruSight Oncology 500 Local App support pages on the Illumina [support website](#) for settings and compatibility information for using TruSight Oncology 500 Local App with BCL Convert.

Additional Resources

The TruSight Oncology 500 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
<i>TruSight Oncology 500 Reference Guide (document # 1000000067621)</i>	Information on using the TruSight Oncology 500 kit.

Install TruSight Oncology 500

Installing the TruSight Oncology 500 requires root privileges.

Docker Considerations

Specify a user ID at Docker run time to limit user privileges associated with the account. For

information on Docker permission requirements and alternatives to running as root, see the Post-Installation steps for Linux page of the Docker Docs website. You may require sudo or escalated privileges to load the docker image.

Singularity Considerations

Singularity requires an admin user with root privileges for set up, but allows non-admin users to run singularity containers. After configuration, a non-admin user can check Singularity configuration and image information.

Installation Instructions

1. Download the application bundle from the link you receive from Illumina. The link expires after 72 hours.
2. Run the following command. Substitute the link provided to you by Illumina for the placeholder link in the command.

```
curl -o ${DOWNLOAD_PATH}/trusight-oncology-500-ruo-${VERSION}.zip
  "https://${APPLICATION_DOWNLOAD_URL}"
```

3. Extract the application bundle.

After extraction, the bundle contains the following files:

- ./TSO500_RUO_LocalApp/md5sum.txt
- ./TSO500_RUO_LocalApp/trusight-oncology-500-ruo.img
- ./TSO500_RUO_LocalApp/trusight-oncology-500-ruo-dockerimage-ruo-{VERSION}.tar
- ./TSO500_RUO_LocalApp/TruSight_Oncology_500_RUO.sh
- ./TSO500_RUO_LocalApp/resources (folder)

4. To verify the files, calculate the md5 hash value using the following commands.

```
md5sum trusight-oncology-500-ruo.img
md5sum trusight-oncology-500-ruo-dockerimage-ruo-{VERSION}.tar
md5sum TruSight_Oncology_500_RUO.sh
```

5. [Docker] Load the Docker image using the following command.

```
sudo docker load --input trusight-oncology-500-ruo-dockerimage-ruo-
{VERSION}.tar
```

6. [Docker] Make sure that the Docker image contains the following values:

- IMAGE—trusight-oncology-500-ruo
- TAG—\${VERSION}
- IMAGE ID—{Corresponding to VERSION}

7. [Singularity] Make sure that the Singularity image file is contained in the applications bundle.

```
which singularity
```

```
singularity inspect trusight-oncology-500-ruo.img
```

8. Display the help for the Bash script using the following command.

```
./TruSight_Oncology_500_RUO.sh --help
```

Running TruSight Oncology 500

Start the TruSight Oncology 500 Local App software with the bash script called `TruSight_Oncology_500.sh` in the application bundle. The bash script can be executed on the command line and can run the software with Docker or Singularity. For arguments, see [Command Line Arguments on page 6](#). You can start from BCL files or from the FASTQ folder produced by BCL Convert. To avoid permissions issues, Illumina recommends creating a `/scratch` folder with 777 permissions. Position your resources and run input folders in this directory.

The following requirements apply for both methods:

- Full path to the TruSight Oncology 500 resources folder location `<RESOURCE_FOLDER>`. The resources folder contains the manifest file, the hg19 reference genome, sample sheet template, and other analysis resources. Modifying the folder causes analysis errors.
- The TruSight Oncology 500 automatically creates an analysis folder with the path `<ANALYSIS_OUTPUT_PATH>` location. If the folder is not empty, TruSight Oncology 500 produces an error.
- Select one of the following locations:
 - Full path to the sequencing system run folder location.
 - Full path to an existing FASTQ folder location.

Sample Sheet Requirements

A TruSight Oncology 500 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. 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. Different types of sequencing runs may use different index adapters. Use the index IDs included in the TruSight Oncology 500 resource bundle.

Create a Sample Sheet

Follow the steps below to create a TruSight Oncology 500 sample sheet.

Use the appropriate sample sheet for your analysis. Sample sheets can be found on the TruSight Oncology 500 support pages on the Illumina support website.

1. Download the appropriate sample sheet template (SampleSheet.csv) from 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	Details
Sample_ID	The Sample_ID becomes a part of output file names. Sample IDs are not case sensitive. Sample IDs must have the following characteristics: <ul style="list-style-type: none"> • 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	Enter a sample type of either DNA or RNA.
Sample_Name	[Optional] Name of sample.
Sample_Plate	[Optional] Name of the sample plate.
Sample_Well	[Optional] Name of the sample well.
Pair_ID	Use to pair samples and named combined output files.
Index adapter sequences	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	Index 1 sequence.
Index2	Index 2 sequence.
I7_Index_ID	[Optional]
I5_Index_ID	[Optional]
Description	[Optional] Description of the sample.
Lane	[Optional] Used for 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 in one of the following two ways:
 - 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 interface.

Command Line Arguments

Argument	Description
--resourcesFolder	Full path to <code>\${RESOURCE_PATH}/resources</code> created.
--analysisFolder	Full path to analysis folder. This path contains the output from the run analysis. Make sure that the analysis folder is empty prior to launching analysis.
--fastqFolder --runFolder	Full path to an existing FASTQ folder or the sequencing system run folder. Provide either --runFolder, or --fastqFolder, but <i>not</i> both. Cannot be used with --demultiplexonly.
--sampleSheet	[Optional] Full path to SampleSheet.csv file. If you do not specify a path to the SampleSheet.csv, locate the file directly under the --fastqFolder or --runFolder path.
--engine	[Optional] The default is Docker. To specify Singularity, set this input to singularity.
--user	[Optional] For Docker only. Expects a \$UID value. Software leverages this user ID to run subsequent processes.
--remove	[Optional] For Docker only. Removes the subsequent Docker container after a run. Identical to passing --rm to a docker run command.
--version	[Optional] Displays the version of this software.
--help	[Optional] Displays a help screen with available options.
--samplePairIDs	[Optional] Defined in the sample sheet. Provide the comma-delimited Sample Pair IDs that should be processed on this node. Example: "Pair_1, Pair_2"
--isNovaSeq	Used to indicate data is generated on NovaSeq system.

Starting From BCL Files

If starting from BCL (*.bcl) files, TruSight Oncology 500 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.
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 TruSight Oncology 500 App 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 [Starting From FASTQ Files on page 7](#).
- The sample sheet is in the FASTQ folder path, or you can set the path to the sample sheet with the `--sampleSheet` override command.

Running TruSight Oncology 500 on Multiple Nodes

TruSight Oncology 500 allows for analysis to be run in parallel on separate servers. 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 using scatter analysis described here is manual. Once FASTQ files are generated by the demultiplexing step, analysis is started with the `--fastqFolder` argument pointing to the FASTQ sample directories and with one or more sample pair IDs specified indicating which samples to analyze. This configuration runs the TruSight Oncology 500 analysis only on the specified samples. Run this command with different sample pair IDs on different servers. For each analysis command, note the output directory, as these directories are needed to gather the final result.

The third step is to gather the results from demultiplexing and each analysis run, and then generate the final `/Results` directory, which contains contain the aggregated results. The inputs for this command are the `--gather` argument followed by the output directories of the demultiplexing step and each individual analysis run.

Table 1 Commands for Multi Node Analysis

Step	Command
Demultiplexing	<pre>TruSight_Oncology_500.sh --analysisFolder /Demultiplex_Output --resourcesFolder \${RESOURCES} --runFolder \${RUN_FOLDER} --sampleSheet \${SAMPLE_SHEET} --demultiplexOnly</pre>
Analysis (one server)	<pre>TruSight_Oncology_500.sh --analysisFolder /Node1_Output --resourcesFolder \${RESOURCES} --fastqFolder /Demultiplex_Output/Logs_Intermediates/FastqGeneration/ --sampleSheet \${SAMPLE_SHEET} --samplePairIDs Pair_1,Pair_2</pre>
Analysis (additional servers)	<pre>TruSight_Oncology_500.sh --analysisFolder /Node2_Output --resourcesFolder \${RESOURCES} --fastqFolder /Demultiplex_Output/Logs_Intermediates/FastqGeneration/ --sampleSheet \${SAMPLE_SHEET} --samplePairIDs Pair_3</pre>
Gather	<pre>TruSight_Oncology_500.sh --analysisFolder /Gathered_Results --resourcesFolder \${RESOURCES} --runFolder \${RUN_FOLDER} --sampleSheet \${SAMPLE_SHEET} --gather /Demultiplex_Output /Node1_Output /Node2_Output</pre>

Running TruSight Oncology 500

You can run TruSight Oncology 500 Local App using Docker or Singularity. Refer to [Command Line Arguments on page 6](#) for additional commands.

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

To run the workflow analysis using data generated on the Illumina NovaSeq 6000 Sequencing System, refer to [Running TruSight Oncology 500: NovaSeq 6000 on page 9](#)

1. Wait for any running TruSight Oncology 500 container to complete before launching a new analysis. Run the following command to check running containers:

```
sudo docker ps
```

2. Select from one of the following five options:

- Start from BCL files in the run folder.

```

${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--analysisFolder /full/path/to/output/analysisFolder \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--runFolder /full/path/to/runFolder

```

- **Start from BCL files using a separate sample sheet.**

```

${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--runFolder /full/path/to/runFolder \
--analysisFolder /full/path/to/output/analysisFolder \
--sampleSheet /full/path/to/SampleSheet.csv

```

- **Start demultiplex only.**

```

${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--analysisFolder /full/path/to/output/analysisFolder \
--runFolder /full/path/to/runFolder --demultiplexOnly

```

- **Start from FASTQ files.**

```

${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--fastqFolder /full/path/to/fastqFolder \
--analysisFolder /full/path/to/output/analysisFolder

```

- **Start from FASTQ files with only paired samples.**

```

${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--analysisFolder /full/path/to/output/analysisFolder \
--fastqFolder /full/path/to/fastqFolder \
--samplePairIDs "Pair_1,Pair_2" \
--sampleSheet /full/path/to/SampleSheet.csv

```

Running TruSight Oncology 500: NovaSeq 6000

To run TruSight Oncology 500 using the Illumina NovaSeq 6000 Sequencing System, follow the below steps:

1. Wait for any running TruSight Oncology 500 container to complete before launching a new analysis. Run the following command to check running containers:

```
sudo docker ps
```

2. Select from one of the following five options:

- **Start from BCL files in the run folder.**

```

${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \

```

```
--analysisFolder /full/path/to/output/analysisFolder \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--runFolder /full/path/to/runFolder \
--isNovaSeq
```

- **Start from BCL files using a separate sample sheet.**

```
${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--runFolder /full/path/to/runFolder \
--analysisFolder /full/path/to/output/analysisFolder \
--sampleSheet /full/path/to/SampleSheet.csv \
--isNovaSeq
```

- **Start demultiplex only.**

```
${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--analysisFolder /full/path/to/output/analysisFolder \
--runFolder /full/path/to/runFolder --demultiplexOnly \
--isNovaSeq
```

- **Start from FASTQ files.**

```
  ${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
  \
  --resourcesFolder ${RESOURCE_FOLDER}/resources \
  --fastqFolder /full/path/to/fastqFolder \
  --analysisFolder /full/path/to/output/analysisFolder \
  --isNovaSeq
```

- **Start from FASTQ files with only paired samples.**

```
${APP_PATH}/{Version}/TruSight_Oncology_500_RUO.sh --user=$UID --remove \
--resourcesFolder ${RESOURCE_FOLDER}/resources \
--analysisFolder /full/path/to/output/analysisFolder \
--fastqFolder /full/path/to/fastqFolder \
--samplePairIDs "Pair_1,Pair_2" \
--sampleSheet /full/path/to/SampleSheet.csv \
--isNovaSeq
```

Analysis Methods

The TruSight Oncology 500 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
 - Contamination Detection
- 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

FASTQ Generation

BCL files are demultiplexed and the software generates intermediate analysis files in the FASTQ format. FASTQ files use a text format to represent sequences. Each file contains reads for each sample and the associated quality scores. Any controls used for the run and clusters that did not pass the filter are excluded. Each FASTQ file contains reads for only one sample, and the FASTQ file name includes

the name of that sample. This step requires the SampleSheet_dna.csv and SampleSheet_rna.csv files generated by the analysis module for demultiplexing. The sample sheets can be found in the FastqGeneration step folder.

TruSight Oncology 500 uses BCL-Convert for FASTQ generation. FASTQ results from TruSight Oncology 500 will not be viewable in Illumina Sequencing Analysis Viewer (SAV).

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.

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. Amplification 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 workflow 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 Factor 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	55242482	CATCTCCGAAAGCCAACAAGGAAAT	C
chr7	55242466	GAATTAAGAGAAGCAACAT	G
chr7	55242465	GGAATTAAGAGAAG	AATTC
chr7	55242465	GGAATTAAGAGAAGCAAC	AAT
chr7	55242469	TTAAGAGAAGCAACATCTC	T
chr7	55242467	AATTAAGAGAAGCAACATC	A
chr7	55242469	TTAAGAGAAG	C
chr7	55242467	AATTAAGAGAAGCAACATC	T
chr7	55242465	GGAATTAAGA	G
chr7	55242467	AATTAAGAGAAGCAACATCTC	TCT
chr7	55242467	AATTAAGAGAAGCAAC	T
chr7	55242464	AGGAATTAAGAGAAGC	A
chr7	55242466	GAATTAAGAGAAGCAA	G
chr7	55242464	AGGAATTAAGAGA	A
chr7	55242469	TTAAGAGAAGCAA	T
chr7	55242465	GGAATTAAGAGAAGCAACATC	AAT
chr7	55242469	TTAAGAGAAGCAACATCT	CAA
chr7	55242463	AAGGAATTAAGAGAAG	A

chr7	55242468	ATTAAGAGAAGCAACATCT	A
chr7	55242467	AATTAAGAGAAGCAACA	TTGCT
chr7	55242462	CAAGGAATTAAGAGAA	C
chr7	55242465	GGAATTAAGAGAAGCAA	AATTC
chr7	55242469	TTAAGAGAAGCAA	C
chr7	55242467	AATTAAGAGAAGCAAC	A
chr7	55242469	TTAAGAGAAGCAACATCTCC	CA
chr7	55242468	ATTAAGAGAAG	GC
chr7	55242465	GGAATTAAGAGAAGCA	G
chr7	55242468	ATTAAGAGAAGCAAC	GCA
chr7	55242465	GGAATTAAGAGAAGCAACA	G
chr7	55249011	AC	CCAGCGTGGAT

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
100 Genomes	Phase 3 v5a
RefSeq	VEP build 91
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 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 $\geq 90\%$ are removed from the TMB calculation as well. The TMB is calculated as follows.

$TMB = \text{Eligible Variants} / \text{Effective panel size}$

Eligible Variants

- Variants not removed by the filtering strategy.
- Variants in the coding region (RefSeq Cds).
- Variant Frequency $\geq 5\%$.
- Coverage $\geq 50X$.
- 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 that contains 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.

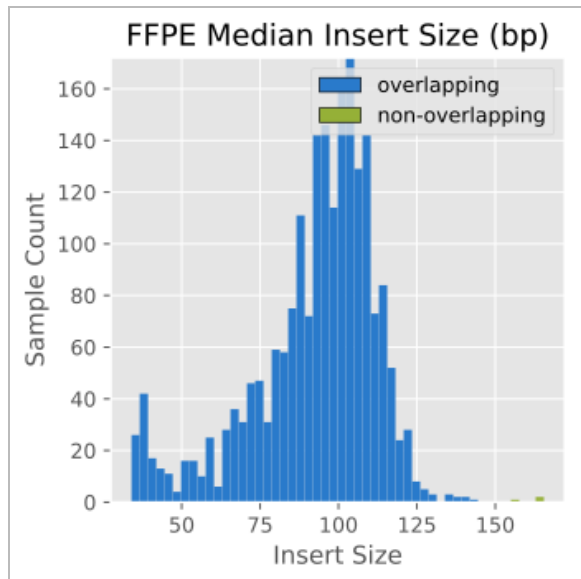
RNA Analysis Methods

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. From internal testing, fragment sizes in RNA FFPE samples hover around 100 bp, so the majority of reads at 76 bp are overlapping (see). While STAR alignment performs stitching to handle overlapping reads, internal testing using simulated data indicates that performance is improved with fewer overlapping reads.



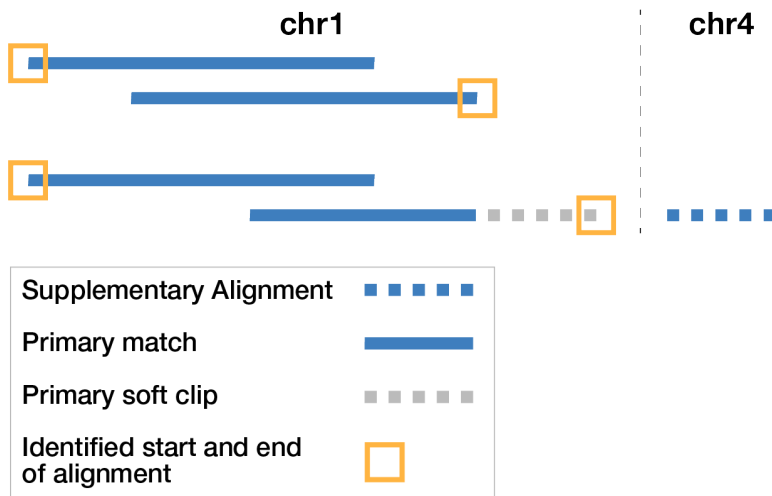
Alignment

The STAR Aligner aligns RNA reads to the human hg19 genome containing unplaced contigs (ie, chrUn_gl regions) and uses GENCODEv19 to identify splice sites. STAR also includes stitching logic to handle overlapping reads.

Duplicate Marking

Duplicate marking is performed using an internally developed tool based on the Picard duplicate marking algorithm. The start and end coordinates of alignments (adjusted for soft clipping) are used to determine whether fragments are overlapping or not. Fusion and splice variant calling only use deduped fragments to score variants. Only primary alignments are considered, supplementary and secondary alignments are not. The alignment with the highest read score is chosen as the unique fragment.

Figure 1 Picard Alignment Based Duplicate Marking.



Fusion Calling

The fusion calling step uses the Manta fusion caller. Manta discovers, assembles, and scores large-scale SVs. Manta only considers candidate fusions with at least 3 unique supporting reads, one of which must be a split read (a single read crossing the fusion breakpoint). The inputs are BAM files and the outputs are VCF files.

RNA Fusion Filtering

The RNAFusionFilter scores fusions and applies filters based on contig alignment to the genome and other features. It also determines which gene is on the 3' end and which gene is on the 5' end based on gene annotations and aligning the contig to the genome. The inputs are BAM files and VCF files and the outputs are .csv format.

Table 3 Scored Features

Score Component	Fusion Feature	Scored Range	Coefficient
Split reads	SplitAlt	0–10	.15
Paired reads	PairedAlt	0–5	.15
Alt allele fraction	$(\text{splitAlt} + \text{PairAlt}) / (\text{SplitRef} + \text{PairRef})$	0–.01	.1
Fusion contig align length (bp)	ContigAlign1 ContigAlign2	12–76	.4
Breakend homology (bp)	BreakpointHomology	2–20	-.2
Fusion contig align length (bp) (For fusions on different chromosome: dist-DistMax)		100000– 2000000	.1
Coverage after breakend (bp)	min (CoverageGene1, CoverageGene2)	0–500	.1

Splice Variant Calling

Splice variant calling is performed using internally developed software. The inputs are BAM files and SJ.out.tab files from STAR. Junctions from SJ.out.tab are filtered first using splice annotations from GENCODEv19, and then further filtered using a baseline from a cohort of nontumor FFPE samples of varying tissue types. Splice junctions appearing on a whitelist are not filtered. The whitelist 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 fusion merging. Splice variants are scored from 0–10 as shown in the table below.

Table 4 Scored Features in Splice Variant Caller

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

Annotation

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

Fusion Merging

Fusions identified through the fusion calling and RNA fusion filtering are combined with the intergenic calls made during splice variant calling. Additionally, each precise fusion event from the RNA fusion filter is recalculated for read count support. The inputs are intergenic splice variant TSV files, fusion

filter CSV files, and BAM files. The outputs are CSV files.

Quality Control

The TruSight Oncology 500 Local App 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 5 High Throughput Systems

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 6 Low Throughput 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 Sample QC

The inputs for DNA Sample QC 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 \leq 3106 OR Contamination Score $>$ 3106 and Contamination p-value \leq 0.049	All
MEDIAN_EXON_COVERAGE	Median exon fragment coverage across all exon bases.	\geq 150	Small variant TMB
PCT_EXON_50X	Percent exon bases with 50X fragment coverage.	\geq 90.0	Small variant TMB
MEDIAN_INSERT_SIZE	The median fragment length in the sample.	\geq 70	Small variant TMB
COVERAGE_MAD	Median Absolute Deviation. Represents the median normalized deviation across all regions used for CNV calling.	\leq 0.210	CNV
MEDIAN_BIN_COUNT_CNV_TARGET	The median raw bin count per CNV target.	\geq 1.0	CNV
USABLE_MSI_SITES	The number of MSI sites usable for MSI calling.	\geq 40	MSI

RNA Sample QC

The inputs for RNA Sample 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 TruSight Oncology 500 generates an analysis output folder in a user-specified location.

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

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
 - SamplesheetValidation—Contains sample sheet validation output logs.
 - RunQc
 - RunQC Metrics JSON file
 - RunQC Output logs
 - FastqGeneration
 - Subfolders per sample ID containing FASTQ files.
 - FastqGeneration Output Logs
 - FastqValidation—Contains FASTQ validation output logs.
 - DnaAlignment

- Subfolders per sample ID containing the aligned BAM and index files.
- DnaAlignment output logs
- CollapsedReads
 - Subfolders per sample ID containing the aligned BAM and index files.
 - CollapsedReads output logs
- DnaRealignment
 - Subfolders per sample ID containing the realigned BAM and index files.
 - DnaRealignment 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
- PhasedVariants—Several folders with this name may exist in the output folder structure.
 - Subfolders per sample ID containing the phased variant metrics JSON.
 - Psara and Scylla output logs
- Msi
 - Subfolders per sample ID containing the MSI metrics JSON.
 - Msi output logs
- VariantCaller
 - Subfolders per sample ID containing the unfiltered genome VCF file.
 - VariantCaller output logs
- SmallVariantFilter
 - Subfolders per sample ID containing the error rate tables.
 - SmallVariantFilter output logs
- Contamination
 - Subfolders per sample ID containing the contamination metrics JSON.
 - Contamination output logs
- Annotation—Contains annotation output logs.
- CvDnaAlignment
 - Subfolders per Sample ID containing the complex-aligned BAM and index files.
 - CvDnaAlignment output logs
- MarkDuplicates
 - Subfolders per sample ID containing the duplicate marked aligned BAM and index files. The MarkDuplicates BAM file is the final output BAM for RNA.
 - MarkDuplicates output logs

- Tmb
 - Subfolders per sample ID containing the TMB metrics JSON.
 - TMB output logs
- 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
- FusionCalling
 - Subfolders per sample ID containing the genome VCF file
 - FusionCaller output logs
- RnaAlignment
 - Subfolders per sample ID containing the aligned BAM and index files.
 - RnaAlignment output logs
- RnaAnnotation
 - Subfolders per sample ID containing the annotated JSON
 - dsdm json
- RnaFusionFiltering
 - Subfolders per sample ID containing the filtered CSV file
 - dsdm JSON
- RnaFusionMerge
 - Subfolders per sample ID containing the CSV file listing all the fusions.
 - RnaFusionMerge output logs
- RnaMarkDuplicates
 - Subfolders per sample ID containing the marked, aligned BAM and index files.
 - RnaMarkDuplicate output logs
- 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 and fusion TSV files.
 - dsdm JSON
- TrimFastq
 - Subfolders per sample ID containing FASTQ files

- dsdm JSON
 - SampleAnalysisResults
 - Results
- 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. The folder structure differs based on whether multinode analysis was used.

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 paired sample (if using PairID). 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
<ul style="list-style-type: none"> • 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 #) 	<ul style="list-style-type: none"> • 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 will produce small variants with blank fields in either of these 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 Whitelist–Small Variants.
 - Transcript information is only present for variants belonging to canonical transcripts that are within the Gene Whitelist–Small Variants.
- **Copy Number Variants**—Copy number variants must meet the following conditions:
 - FILTER field marked as PASS.
 - ALT field is <DUP>.
- **Combined Variants**—Fusion variants must meet the following conditions:
 - Passing Variant Call (KeepFusion field is true).
 - Contains at least one gene on the fusion whitelist.

- **Biomarkers TMB/MSI**—Always present when DNA sample is processed.
- **Splice Variants**—Passing splice variants that are contained on genes EGFR, MET, and AR.

Metrics Output

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

- QC metrics for small variant calling (SVC)
- TMB
- MSI
- CNV
- Fusion
- RunQc 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.

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: <ul style="list-style-type: none"> • Fail AQ or LQ • 0 stitched support for indel or variant in homopolymer context .
A, C, G,etc.	Blacklist	Position with high background noise. Not available for variant detection.
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

Column	Description
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 mentioned above
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 TruSight Oncology 500 Local App. 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 $Q = -10 \times \log_{10}(p\text{-value})$. 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. indicates the detected fold change (FC) is less than a predefined deletion cutoff for that gene. This cutoff can vary from gene to gene.

 calls have only been validated with *in silico* data sets. As a result, all 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, the copy number n can be calculated 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.

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

Splice Variant Annotated JSON

Each splice variant is annotated using the Illumina Annotation Engine. The following information is captured in the JSON if available:

- 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. Candidate fusions from the splice variant caller are listed in this output with relevant supporting information but are not considered high confidence. Two key output columns in the file describe the candidate fusions: Filter and KeepFusion.

The following table describes the output found in the Filter columns. The output is either a confidence filter or information only as indicated. If none of these filters are triggered, the Filter column displays PASS.

Table 7 Filter Column Output

Filter	Description
Imprecise	(Confidence filter) A low-resolution candidate, not an assembled fusion call.
RepeatOverlap	(Confidence filter) The fusion is tagged as overlapping with a repeat region. Only used as a confidence filter for nonuniquely mapping fusion candidates, otherwise information only.
WeakBreakend	(Confidence filter) The read/alignment evidence on one side of the fusion is weak. Usually this filter indicates that the reads only overlap the fusion by a few base pairs. Alternatively, it can indicate too much homology (no unique sequence).
Homology	(Information only) The fusion contig is a substring of another fusion contig.
DuplicateContig	(Information Only) The two contigs of the fusion are the same contig.
ContigIntragenic	(Confidence filter) The realignment of half-contigs produces alignments that map to the same gene on both sides (or within 1 kB if unannotated).
LowQ	(Confidence filter) Fusion supporting reads (unique) < 5 (+ 1 for every 10 million reads over 16 million reads).
LowDupReads	(Confidence Filter) Fusion supporting reads (duplicate) < 5.
NonExonic	(Information only) Fusion breakpoint does not fall within an exon.
LocalContigAlign	(Information only) Contig realignment found a nonfusion alignment for this contig.
LowFusionRatio	(Information only) Few strong evidence reads compared to wild type reads.
NoReferenceReads	(Information only) No reads on either side of the presumed breakpoint are marked as reference (structurally normal) reads.

The KeepFusion column of the output has a value of True when the RNAFusionFilter score is ≥ 0.45 , none of the confidence filters are triggered, and fusions called by the splice variant caller have a score of 1.

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

Table 8 Fusion Columns

Fusion Object Field	Source
Caller	(Either RNAFusionFilter or SpliceGirl) The algorithm used to identify the fusion.
Gena A	The gene associated with the A side of the fusion. A semicolon-delimited list is used for multiple genes.

Fusion Object Field	Source
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 B Breakpoint	(Information only) The chromosome and offset of the Gene B side of the fusion.
Score	The quality of fusion as determined by the respective caller. Results from different callers are not equivalent.
Filter	The filter associated with the fusion as determined by the respective caller. Results from different callers are not equivalent.
Precise/Imprecise	(RNAFusionFilter Only) Whether the algorithm could identify the precise fusion coordinates. Coordinates listed for imprecise fusions are based on strongest statistical evidence.
Intragenic Call	(SpliceGirl Only) List any genes associated with a splice overlapping the fusion. A semicolon-delimited list is used for multiple genes.
Ref A Split	Gene A uniquely mapping reads spanning the junction. Does not support fusion. Duplicate reads included.
Ref A Pair	Gene A uniquely mapping reads paired across junction. Does not support fusion. Duplicate reads included.
Ref B Split	Gene B uniquely mapping reads spanning the junction. Does not support fusion. Duplicate reads included.
Ref B Pair	Gene B uniquely mapping reads paired across junction. Does not support fusion. Duplicate reads included.
Alt Split	Uniquely mapping reads split by the junction. Supports fusion. Duplicate reads included.
Alt Pair	Uniquely mapping reads paired across junction. Supports fusion. Duplicate reads included.
CandidateAlt	(RFF Imprecise Reads Only) The number of reads and pairs that potentially support this candidate before refinement and scoring.
Contig	Sequence of fusion. Can be used to determine fusion directionality. (RFF Only)

Fusion Object Field	Source
ContigAlign1	(RFF Only) Length of Gene A in contig.
CntigAlign2	(RFF Only) Length of Gene B in contig.
KeepFusion	The determination whether the fusion should be kept or dropped from the list of fusions.
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.
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 whitelist 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

Uninstall TruSight Oncology 500

The procedure to uninstall the TruSight Oncology 500 Local App removes the following assets:

- TruSight Oncology 500 Local App
- Containers (if you performed an analysis)
- Singularity image
- Docker image
- Bash script

1. Wait for any running TruSight Oncology 500 Local App container to complete before uninstalling. Run the following command to check running containers.

When a run is active, at least one Docker container includes the image `trusight-oncology-500`.

```
$ sudo docker ps
$ CONTAINER ID IMAGE COMMAND CREATED STATUS PORTS NAMES
```

2. If there are active runs, stop run execution.

```
# using the container id, found in the above command:
$ sudo docker stop <container_id>
```

3. Remove all Docker containers that include the `trusight-oncology-500` image.

```
# similar to the above, identify the container id:
$ sudo docker ps -a|grep trusight-oncology-500
# then remove each container:
$ sudo docker rm $(sudo docker ps -a|grep trusight-oncology-500| \ awk '
{print $1}')
```

4. Remove Docker from the image and previous runs.

```
$ sudo docker rmi -f trusight-oncology-500:${VERSION}
```

5. Remove downloaded assets.

```

$ rm ${APP_PATH}md5sum.txt
$ rm -rf ${APP_PATH}/resources
$ rm ${APP_PATH}/{trusight-oncology-500-ruo-dockerimage-ruo-2.1.0.60.tar}
$ rm ${APP_PATH}/{trusight-oncology-500-ruo.img}
$ rm ${APP_PATH}/{Trusight_Oncology_500_RUO.sh}

```

Troubleshooting

Failure Type	Actions
Software	Open the log file <code>./{analysisFolder}/Logs_Intermediates/TruSight-Oncology-500-pipeline-<timestamp>.log</code> file. The log file displays all commands run by the workflow and the exit code for each analysis step. If a step fails, review standard output and standard error log files in the folder <code>./{analysisFolder}/Logs_Intermediates/.</code>
Samples	Open the final sample biomarker report log file <code>./{analysisFolder}/Results/SampleID/CombinedVariantOutput.tsvt</code> . 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 <code>./{analysisFolder}/Logs_Intermediates/{FailedStep}/.</code>

Revision History

Document	Date	Description of Change
Document # 1000000113216 v02	September 2021	Corrected minimum read depth for reference calls and limit of detection for VAF. Updated EGFR variants table.
Document # 1000000113216 v01	November 2020	Added recommendation to create /scratch folder Included separate V1 and V2 sample sheet information Added reference to Pisces GitHub page
Document # 1000000113216 v00	April 2020	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	

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.

© 2021 Illumina, Inc. All rights reserved.

illumina[®]