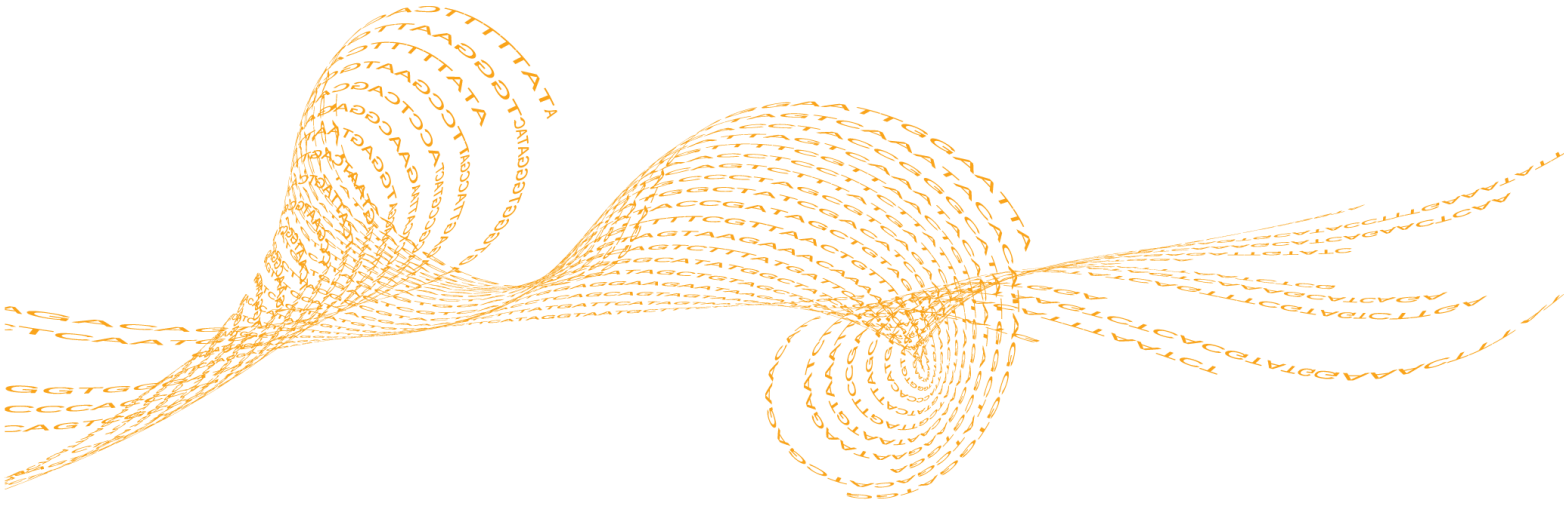
TruSight Tumor 15 v1.0

BaseSpace App Guide

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



ILLUMINA PROPRIETARY
Document # 1000000001132 v02
January 2016

illumina®

Introduction

The BaseSpace® App, TruSight Tumor 15 v1.0, reports on somatic variants of a set of reference panel genes associated with cancer. The app evaluates short regions of amplified DNA, or amplicons, for variants. Focused sequencing of amplicons enables high coverage of particular regions across many samples. This workflow is uniquely suited for detection of variants in formalin-fixed paraffin-embedded (FFPE) samples.

After alignment, the app performs variant calling using the somatic variant caller.

Compatible Libraries

See the BaseSpace support page for a list of library types that are compatible with the TruSight Tumor 15 v1.0 App.

Workflow Requirements

- The minimum read length is 50 bases.
- No minimum number of reads is required. However, use sufficient data for each sample to support an appropriate depth of coverage for variant calling.
- Requires paired-end samples with the same read lengths.
- The maximum number of pairs is 96.
- The maximum size is 25 gigabases per sample.
- Variants are found in the regions covered in the manifest.

Versions

The following components are used in the TruSight Tumor 15 v1.0 App.

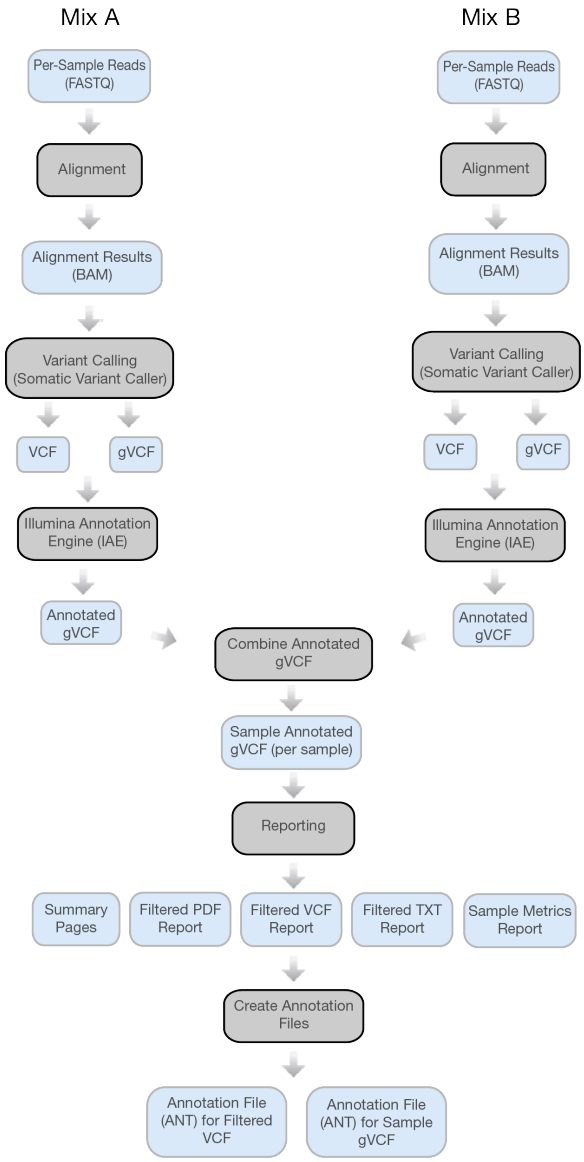
Software	Version
TruSight Tumor 15 (BaseSpace Workflow)	1.0.0
Isis (Analysis Software)	2.6.2.3
Somatic Variant Caller	4.0.7.6
Illumina Annotation Engine	1.1.8.0

Reference Genomes

- Human, UCSC hg19
  The human reference genome is PAR-Masked, which means that the Y chromosome sequence has the Pseudo Autosomal Regions (PAR) masked (set to N) to avoid mismapping of reads in the duplicate regions of sex chromosomes.

Workflow Diagram

Figure 1 TruSight Tumor 15 v1.0 App Workflow



Set Analysis Parameters

1 In BaseSpace, click the **Apps** tab.

2 Click **TruSight Tumor 15**.

3 From the drop-down list, select **version 1.0**, and then click **Launch** to open the app.

4 In the **Analysis Name** field on the app input form, enter the analysis name. By default, the analysis name includes the app name, followed by the date and time the analysis session starts.

5 From the **Save Results To** field, select the project that stores the app results.

6 From the **Sample Pairs** field, select the samples for a pooled sample pair. Select the samples as follows.

   a Click **Select Pairs** to open the Select Pairs window.
   b [Optional] Search for the samples you want to analyze.
   c Select a sample from the search results, or click **Select All** to select all samples, and then drag and drop the samples to the appropriate location (Mix A Samples or Mix B Samples).
   d Click **Confirm**.

Note
Select matching sample pairs with matching sample names. TruSight Tumor 15 v1.0 checks to ensure matching sample names.

7 Click **Continue**.

When analysis is complete, the status of the app session is updated automatically and an email is sent to notify you.

Analysis Methods

The TruSight Tumor 15 v1.0 workflow evaluates short regions of amplified DNA, or amplicons, for variants. Focused sequencing of amplicons enables high coverage of particular regions across many samples.

Alignment

During the alignment step, the banded Smith-Waterman algorithm aligns clusters from each sample against amplicon sequences specified in the manifest file.

The banded Smith-Waterman algorithm performs local sequence alignments to determine similar regions between 2 sequences. Instead of comparing the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths. Local alignments are useful for dissimilar sequences that are suspected to contain regions of similarity within the larger sequence. This process allows alignment across small amplicon targets, often less than 10 bp.

Each paired-end read is evaluated in terms of its alignment to the relevant probe sequences for that read.

- Read 1 is evaluated against the reverse complement of the Downstream Locus-Specific Oligos (DLSO).
- Read 2 is evaluated against the Upstream Locus-Specific Oligos (ULSO).
- If the start of a read matches a probe sequence with no more than 1 mismatch, the full length of the read is aligned against the amplicon target for that sequence.

Alignments that include more than 3 indels are filtered from alignment results. Filtered alignments are written in alignment files as unaligned and are not used in variant calling.

Variant Calling

Variant calling is performed with the somatic variant caller.

Variant Calling

Developed by Illumina, the somatic variant caller identifies variants present at low frequency in the DNA sample.

The somatic variant caller identifies SNPs in 3 steps:

- Considers each position in the reference genome separately
- Counts bases at the given position for aligned reads that overlap the position
- Computes a variant score that measures the quality of the call using Poisson model.

Variants are first called for each pool separately. Then, variants from each pool are compared and combined into a single output file. If a variant meets the following criteria, the variant is marked as PASS in the variant call (VCF) file:

- The variant is present in both pools
- Has a cumulative depth of 1000 or an average depth of 500x per pool
- Has a variant frequency of ≥ 3% as reported in the merged VCF file

A locus for a mutation or reference is classified as a no call under the following conditions:

- The variant frequency is near the signal noise level between 1% and 2.6%
- The variant quality is < Q30
- The depth is < 500

- Significant strand bias is detected
- The indel occurs in a homopolymer region

Read Stitching

For each paired read, a minimum of 10 bases must overlap between Read 1 and Read 2 to be a candidate for read stitching. The minimum threshold of 10 bases minimizes the number of reads that are stitched incorrectly due to a chance match. Candidates for read stitching are scored as follows:

- For each possible overlap of 10 base pairs or more, a score of 1 – MismatchRate is calculated.
- Perfectly matched overlaps have a MismatchRate of 0, resulting in a score of 1.
- Random sequences have an expected score of 0.25.
- If the best overlap has a score of ≥ 0.9 ***and*** the score is ≥ 0.1 higher than any other candidate, then the reads are stitched together at this overlap.

Although the stitched reads are aligned as 1, in the BAM file the stitched alignment is split into individual alignments.

During variant calling, stitched reads are processed together. A consensus read is generated by taking the base call and quality score of the read with the higher Q-score in the overlap region. When the Q-score is the same, but the base call differs, a "no call" is used at that position. Sometimes read stitching can improve the accuracy of variant calling.

Paired-end reads that cannot be stitched are converted to 2 single reads in the FASTQ file.

Illumina Annotation Engine (IAE)

Annotation with IAE populates the *.vcf file with the following values in the INFO column CSQ key:

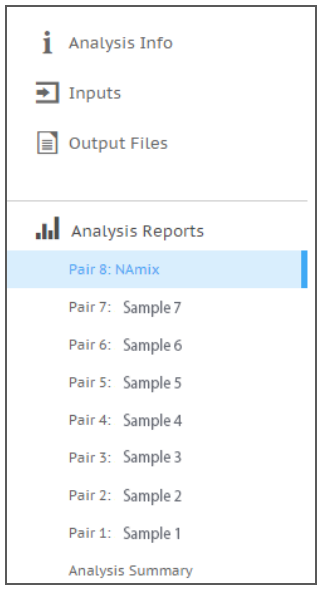
- Genotype Index
- Transcript ID
- Functional consequence
- Canonical transcript flag
- HGVS coding nomenclature
- HGVS protein nomenclature
- HGNC gene symbol

Annotation through IAE is available for alignments against the human reference genome: UCSC build hg19.

Analysis Output

To view the results, click the **Projects** tab, then the project name, and then the analysis.

Figure 2 Output Navigation Bar



After analysis is complete, access the output through the left navigation bar.

- **Analysis Info**—Information about the analysis session, including log files.
- **Inputs**—Lists the samples and settings specified for the analysis session.
- **Output Files**—Output files for the sample.
- **Summary Analysis Report**—Analysis metrics for the aggregate results, displayed when multiple samples are analyzed.
- **Sample Analysis Reports**—Analysis reports for each sample.

Analysis Info

The Analysis Info page displays the analysis settings and execution details.

Row Heading	Definition
Name	Name of the analysis session.
Application	App that generated this analysis.
Date Started	Date and time the analysis session started.
Date Completed	Date and time the analysis session completed.
Duration	Duration of the analysis.
Session Type	Number of nodes used.
Status	Status of the analysis session. The status shows either Running or Complete.

Log Files

File Name	Description
AmpliconCoverage_M#.tsv	Contains details about the resulting coverage per amplicon per sample. M# represents the manifest number.
AmpliconRunStatistics.xml	Provides amplicon-related information.
AnalysisError.txt	Workflow standard error output (contains error messages created when running the workflow).
AnalysisLog.txt	Workflow standard output (contains details about workflow steps, command line calls with parameters, timing, and progress).
CompletedJobInfo.xml	Contains information about the completed analysis session.
Logging.zip	Contains all detailed log files for each step of the workflow.
SampleSheet.csv	Sample sheet.
TST_15-A-Manifest.txt, TST_15-B-Manifest.txt	Manifest files used in the analysis.

Inputs

The Inputs page lists the samples and settings specified for the analysis session.

Output Files

The Output Files page provides access to the output files for each sample pair analysis.

- **BAM Files**—Aligned sequences and quality scores in the BAM (*.bam) file format.
- **VCF Files**—Variant calls in the VCF (*.vcf) file format.
- **Genome VCF Files**—Variants, references, and no calls for all sites in the genome in the genome VCF (gVCF) file format.
- **Annotation File**—Detailed annotations in a binary file format.
- **Filtered Reports**—A filtered summary of the analysis.
- **Summary File**—Statistics for each sample.
- **Sample Metrics Report**—Calculations for the number of bases in the library that have ≥ 500 coverage, and percentage of on-target reads.

BAM File Format

A BAM file (*.bam) is the compressed binary version of a SAM file that is used to represent aligned sequences up to 128 Mb. SAM and BAM formats are described in detail at https://samtools.github.io/hts-specs/SAMv1.pdf.

BAM files use the file naming format of SampleName_S#.bam, where # is the sample number determined by the order that samples are listed for the run.

BAM files contain a header section and an alignments section:

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

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

- **RG:** Read group, which indicates the number of reads for a specific sample.
- **BC:** Barcode tag, which indicates the demultiplexed sample ID associated with the read.
- **SM:** Single-end alignment quality.
- **AS:** Paired-end alignment quality.
- **NM:** Edit distance tag, which records the Levenshtein distance between the read and the reference.
- **XN:** Amplicon name tag, which records the amplicon tile ID associated with the read.

BAM index files (*.bam.bai) provide an index of the corresponding BAM file.

VCF File Format

VCF is a widely used file format developed by the genomics scientific community that contains information about variants found at specific positions in a reference genome.

VCF files use the file naming format SampleName_S#.vcf, where # is the sample number determined by the order that samples are listed for the run.

VCF File Header—Includes the VCF file format version and the variant caller version. The header lists the annotations used in the remainder of the file. If MARS is listed, the Illumina internal annotation algorithm annotated the VCF file. The VCF header includes the reference genome file and BAM file. The last line in the header contains the column headings for the data lines.

VCF File Data Lines—Each data line contains information about a single variant.

VCF File Headings

Heading	Description
CHROM	The chromosome of the reference genome. Chromosomes appear in the same order as the reference FASTA file.
POS	The single-base position of the variant in the reference chromosome. For SNPs, this position is the reference base with the variant; for indels or deletions, this position is the reference base immediately before the variant.
REF	The reference genotype. For example, a deletion of a single T is represented as reference TT and alternate T. An A to T single nucleotide variant is represented as reference A and alternate T.
ALT	The alleles that differ from the reference read. For example, an insertion of a single T is represented as reference A and alternate AT. An A to T single nucleotide variant is represented as reference A and alternate T.

Heading	Description
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant and lower probability of errors. For a quality score of Q, the estimated probability of an error is $10^{-(Q/10)}$. For example, the set of Q30 calls has a 0.1% error rate. Many variant callers assign quality scores based on their statistical models, which are high in relation to the error rate observed.

VCF File Annotations

Heading	Description
FILTER	If all filters are passed, **PASS** is written in the filter column. • **LowDP**—Applied to sites with depth of coverage below a cutoff. • **LowGQ**—The genotyping quality (GQ) is below a cutoff. • **LowVariantFreq**—The variant frequency is less than the given threshold. • **R8**—For an indel, the number of adjacent repeats (1-base or 2-base) in the reference is greater than 8. • **SB**—The strand bias is more than the given threshold.
INFO	Possible entries in the INFO column include: • **CSQ**—Consequence as predicted by Illumina Annotation Engine (IAE). • **DP**—The depth (number of base calls aligned to a position and used in variant calling).
FORMAT	The format column lists fields separated by colons. For example, GT:GQ. Available fields include: • **AD**—Entry of the form X,Y, where X is the number of reference calls, and Y is the number of alternate calls. • **GQ**—Genotype quality. • **GQX**—Genotype quality. GQX is the minimum of the GQ value and the QUAL column. In general, these values are similar; taking the minimum makes GQX the more conservative measure of genotype quality. • **GT**—Genotype. 0 corresponds to the reference base, 1 corresponds to the first entry in the ALT column, and so on. The forward slash (/) indicates that no phasing information is available. • **NL**—Noise level; an estimate of base calling noise at this position. • **SB**—Strand bias at this position. Larger negative values indicate less bias; values near 0 indicate more bias. • **VF**—Variant frequency; the percentage of reads supporting the alternate allele.
SAMPLE	The sample column gives the values specified in the FORMAT column.

Genome VCF Files

Genome VCF (gVCF) files are VCF v4.1 files that follow a set of conventions for representing all sites within the genome in a reasonably compact format. The gVCF files

include all sites within the region of interest in a single file for each sample.

The gVCF file shows no-calls at positions with low coverage, or where a low-frequency variant (< 2.6%) occurs often enough (> 1%) that the position cannot be called to the reference. A genotype (GT) tag of **./.** indicates a no-call.

For more information, see sites.google.com/site/gvcftools/home/about-gvcf.

Per-Library and Merged VCF Files

The TruSight Tumor 15 v1.0 App generates 2 sets of variant call files:

- Per-library VCF and gVCF files
- Merged gVCF files

Output Files

Pair Folder—Contains merged gVCF (*.genome.vcf) files.

Libraries—Contains per-library VCF (*.vcf) files, and per-library gVCF (*.genome.vcf) files.

Per-Library VCF Files

Using the somatic variant caller, variants are called in the Mix A library and the Mix B library to produce an independent set of VCF files for each library. The set of per-library VCF files include both VCF and gVCF files.

Per-library VCF files use the following naming convention, where S# represents the order the sample is listed in the sample sheet, which the app defines:

- SampleName_S#.genome.vcf—Reports all sites within the region of interest for a single library.
- SampleName_S#.vcf—Reports variants only for a single library.

Merged gVCF Files

The software selects specific coordinates from the gVCF files generated for Mix A and Mix B to create a final merged VCF file for the sample.

Merged gVCF files are written to the Analysis folder.

Annotation File

The Illumina Annotation Engine (IAE) generates a binary an annotation file (*.ant). The merged *.vcf and *.Report.vcf file information is also provided in the annotation file. You can view this binary file in VariantStudio. For more information, see www.illumina.com/informatics/research/biological-data-interpretation/variantstudio.html.

The annotation files are located in each pair folder in the Output Files.

Filtered Reports

Filtered report files are written to Pair folders in the Output Files.

TruSight Tumor 15 v1.0 generates a TruSight Tumor 15 Report that provides a filtered summary of the analysis, consisting of sample information, identified variants, and details about no calls.

The Filtered Reports follow the naming convention of Sample_Report.pdf. Reports are provided in PDF, TXT, and VCF file formats.

Variants identified as specified in the report definition

- Detected SNVs, Insertions, and Deletions

Statistic	Definition
Gene	The gene where the SNV, insertion, or deletion is detected.
Amino Acid Change	Human Genome Variation Society (HGVS) protein notation.
Variant Type	Consequence on protein function.
Nucleotide Change	HGVS nucleotide notation.
Variant Frequency	Fraction of reads in which the variant was detected.
Transcript	Ensembl canonical transcript.

No Calls

Statistic	Definition
Gene	The gene where the no call is located.
Chromosome	The chromosome where the no call is located.
Coordinate	The coordinate where the no call is located.
Failed Filter	The reason for the no call. • **Low Variant Frequency**—The variant frequency is less than the given threshold. • **Low Coverage**—The depth of coverage is below a cutoff. • **Low Genotype Quality**—The genotyping quality (GQ) is below a cutoff. • **Indel Reference Repeat**—For an indel, the number of adjacent repeats (1-base or 2-base) in the reference is greater than 8. Identical to R8 in the VCF File Annotations Filter entry. • **Strand Bias**—The strand bias is more than the given threshold. Identical to SB in the VCF File Annotations Filter entry.

Summary File

The TruSight Tumor 15 v1.0 App produces an overview of statistics for each sample and the aggregate results in a comma-separated values (CSV) format: *.summary.csv. These files are located in the Libraries folder for each sample.

Statistic	Definition
Sample ID	IDs of samples reported in the file.

Statistic	Definition
Sample Name	Names of samples reported in the file.
Run Folder	BaseSpace specific intermediate run folder for samples reported in the file.
Manifest	The manifest file used for analysis. This file specifies the targeted regions for the aligner and variant caller.
Reference genome	Reference genome selected.
Number of amplicon regions	The number of amplicon regions that were sequenced.
Total length of amplicon regions	The total length of the sequenced bases in the target region.
Total PF reads	The number of reads passing filter for the sample.
Total aligned reads	The total number of reads passing filter present in the data set that aligned to the reference genome. Numbers are calculated per read, and over both reads.
Percent aligned reads	The percentage of reads passing filter that aligned to the reference genome. Numbers are calculated per read, and over both reads.
Total PF bases	The number of bases passing filter for the sample.
Total aligned bases	The total number of bases present in the data set that aligned to the reference genome. Numbers are calculated per read, and over both reads.
Percent aligned bases	The percentage of bases that aligned to the reference genome. Numbers are calculated per read, and over both reads.
Percent Q30	The percentage of bases with a quality score of 30 or higher. Numbers are calculated per read.
Mismatch rate	The average percentage of mismatches across both reads 1 and 2 over all cycles. Numbers are calculated per read.
Amplicon mean coverage	The total number of aligned bases to the targeted region divided by the targeted region size.
Uniformity of coverage	The percentage of amplicon regions with coverage values greater than the low coverage threshold, where the low coverage threshold is defined as (0.2 * Amplicon Mean coverage).
SNVs, Insertions, Deletions	Total number of variants present in the data set that pass the quality filters.
SNVs, Insertions, Deletions (Percent found in dbSNP)	100*(Number of variants in dbSNP/Number of variants).

Statistic	Definition
SNV Ts/Tv ratio	The number of Transition SNVs that pass the quality filters divided by the number of Transversion SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).
SNVs, Insertions, Deletions Het/Hom ratio	Number of heterozygous variants/Number of homozygous variants.

Sample Metrics Report

The Sample Metrics Report provides information for each sample in each library. The report provides calculations from the gVCF file that indicate the number of bases in the library that have ≥ 500 coverage. The report also shows the percentage of reads in each library that aligned to the manifest, which are referred to as on-target reads.

The Sample Metrics report information is also displayed in the Analysis Summary section of the Analysis Reports. See *Analysis Summary* on page 15.

The report includes the following information:

- Sample Name
- For each Sample Name, the percentage of bases that have ≥ 500 coverage for Mix A and Mix B.
- For each Sample Name, out of the total number of reads, the percentage of reads on-target for Mix A and Mix B.

NOTE
For amplicon coverage specifications, see the *TruSight Tumor 15 Data Sheet (document # 1170-2015-003)* on the Illumina website.

Sample Analysis Reports

Click the Analysis Report for any pair to view the per-sample filtered report (*.Report.pdf). This report is identical to the PDF filtered report located in the Pair folders in the Output Files. For more information, see *Filtered Reports* on page 12.

Click **PDF Summary Report** to download the report.

Analysis Summary

The Analysis Summary report provides information for each sample in each library, and is populated based on data from the SampleMetricsReport.txt file. See *Sample Metrics Report* on page 15.

The report includes the following information:

- Report date and time
- Sample Name
- For each Sample Name, the percentage of bases that have ≥ 500 Read 1 and Read 2 coverage for Mix A and Mix B.
- For each Sample Name, out of the total number of reads, the percentage of reads on-target for Mix A and Mix B.

Revision History

Document	Date	Description of Change
Document # 1000000001132 v02	January 2016	Reorganized topics, updated writing style.
Document # 1000000001132 v01	November 2015	Initial release.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1 Illumina General Contact Information

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

Table 2 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Japan	0800.111.5011
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
China	400.635.9898	Singapore	1.800.579.2745
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	Taiwan	00806651752
Hong Kong	800960230	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000
Italy	800.874909		

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

Product documentation—Available for download in PDF from the Illumina website. Go to support.illumina.com, select a product, then select **Documentation & Literature**.

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