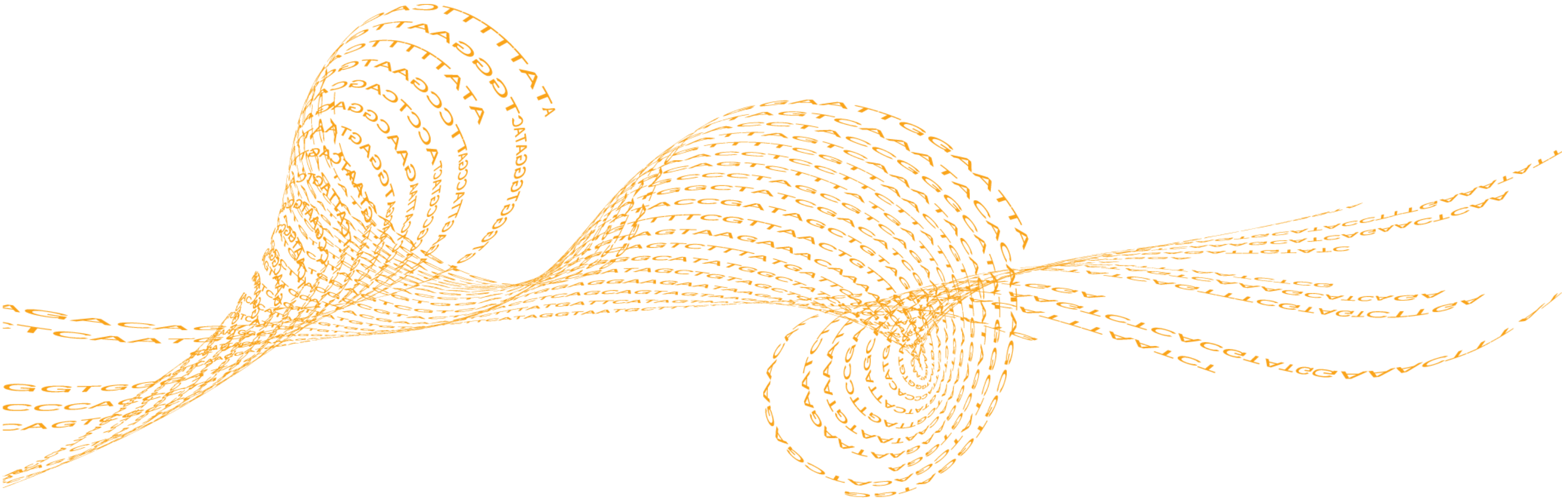
TruSeq Amplicon v2.0

BaseSpace App Guide

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



ILLUMINA PROPRIETARY
Document # 15055857 v02
January 2016

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.

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

Introduction

The BaseSpace® App, TruSeq Amplicon v2.0, analyzes DNA enriched for particular target sequences. The app aligns amplicon reads against the reference specified in the manifest file, and then performs variant analysis. Variants are called for the targeted regions. Statistics reporting accumulates coverage statistics for each target and overall metrics.

Compatible Libraries

See the BaseSpace support page for a list of library types that are compatible with the TruSeq Amplicon v2.0 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.
- Only 1 manifest is used per analysis.
- Paired-end sequencing data is required.
- Sequenced samples have the same read lengths.
- A maximum of 96 samples is possible per analysis.

Versions

The following components are used in the TruSeq Amplicon v2.0 App.

Software	Version
TruSeq Amplicon (BaseSpace Workflow)	2.0.0
Isis (Analysis Software)	2.6.21.7
SAMtools	1.2
Somatic Variant Caller	4.0.13.1
Starling (Isaac Variant Caller)	2.1.4.1
GATK (Variant Caller)	v1.6-23-gf0210b3
IONA (Illumina On-Node Annotation)	1.0.10.37
MarkDuplicates	1.0.1
mono	3.12.1

This app was validated with data from MiSeq and NextSeq systems.

Reference Genomes

- Human, UCSC hg19

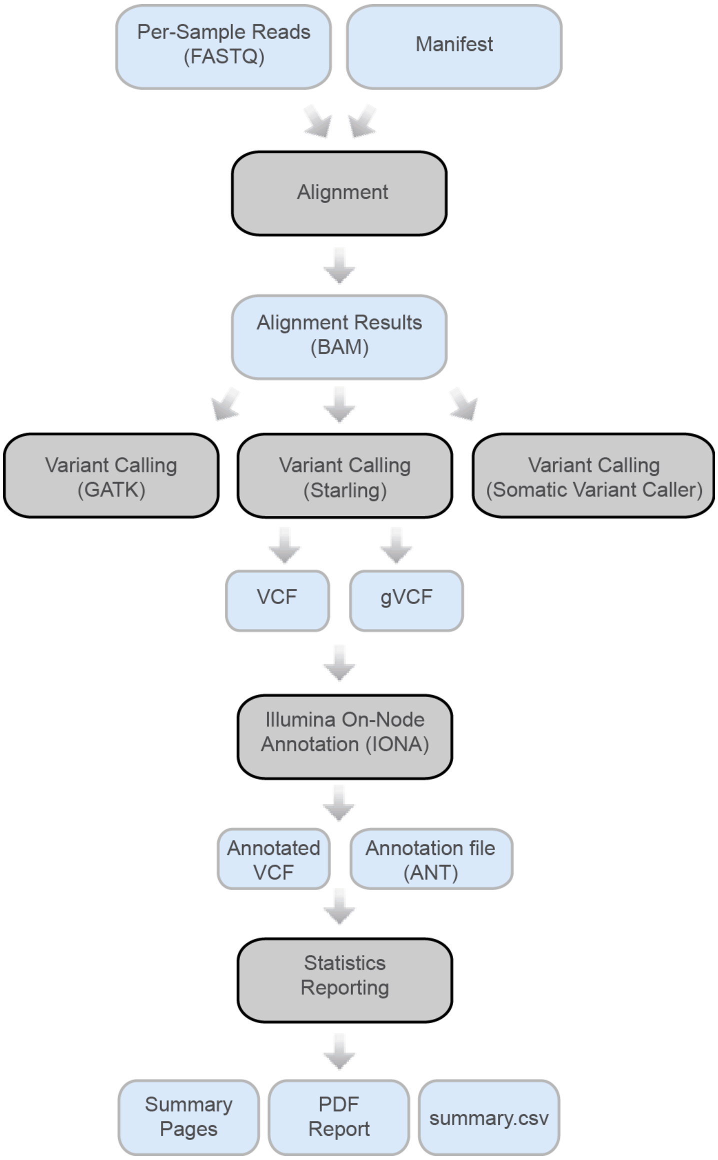
The human reference genome is PAR-Masked, which means that the Y chromosome sequence has the Pseudo Autosomal Regions (PAR) masked (set to N) to avoid mismapping of reads in the duplicate regions of sex chromosomes.

- Cow, UMD3.1
- Mouse, mm9
- Rat, rn4

NOTE
When using a custom manifest file, specify a reference genome in the Build ID column. Reference genome names are case sensitive.

Workflow Diagram

Figure 1 TruSeq Amplicon v2.0 App Workflow



Set Analysis Parameters

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

2 Click **TruSeq Amplicon**.

3 From the drop-down list, select **version 2.0**, 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 that the analysis session starts.

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

6 From the **Sample(s)** field, browse to the sample you want to analyze and select the checkbox. You can select multiple samples.

7 From the **Targeted Amplicons** field, select a panel of targeted amplicons representative of the selected samples.

8 If you selected **Custom Panel** in the **Targeted Amplicons** drop-down list, upload a custom manifest and select the manifest file from the **Custom Manifest File** field. Upload a custom manifest as follows.

a Navigate to your project in BaseSpace.
b Click **Import**.
c Follow the instructions to add the Custom Amplicon manifest file (*.txt) to the project.

NOTE
Specify the reference genome in the header of the manifest file. See *Reference Genomes* on page 3.

9 From the **Variant Caller** field, select a variant caller.
For tumor samples, use the Somatic Variant Caller.

10 If using the somatic variant caller, specify the **Somatic Variant Caller Threshold (percentage)**.
Set to 5 by default. Variants with a frequency below the specified threshold are not reported in VCF files. Lower threshold values can result in false positive variants.

11 Set the **Read Stitching** option.
When enabled, reads that overlap > 10 bases between Read 1 and Read 2 are combined to create a single (longer) read for alignment.

12 From the **Annotation** field, select a preferred gene and transcript annotation reference database.

13 Click **Continue**.
The TruSeq Amplicon v2.0 App begins analysis of the sample.
When analysis is complete, the status of the app session is updated automatically and an email is sent to notify you.

Analysis Methods

The TruSeq Amplicon v2.0 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 any of the following variant calling options.

- GATK
- Isaac Variant Caller
- Somatic Variant Caller

GATK

Developed by the Broad Institute, the Genome Analysis Toolkit (GATK) first calls raw variants for each sample read. Then GATK analyzes the variants against known variants, and applies a calibration procedure to compute a false discovery rate for each variant. Variants are flagged as homozygous (1/1) or heterozygous (0/1) in the VCF file sample column.

The GATK best practices were guidelines for the app; they are described here: www.broadinstitute.org/gatk/guide/topic?name=best-practices.

For more information about GATK, see www.broadinstitute.org/gatk.

> DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5): 491-8.

Starling (Isaac Variant Caller)

Starling identifies single nucleotide variants (SNVs) and small indels using the following steps:

- **Read filtering**—Filters out reads failing quality checks.
- **Indel calling**—Identifies a set of possible indel candidates and realigns all reads overlapping the candidates using a multiple sequence aligner.
- **SNV calling**—Computes the probability of each possible genotype given the aligned read data and a prior distribution of variation in the genome.
- **Indel genotypes**—Calls indel genotypes and assigns probabilities.
- **Variant call output**—Generates output in a VCF file and a compressed genome variant call (gVCF) file. See *VCF File Format* on page 12 and *Genome VCF Files* on page 14.

Indel Candidates

Input reads are filtered by removing any of the following reads:

- Reads that failed base calling quality checks
- Paired-end reads not marked as a proper pair
- Reads with a mapping quality < 20

Indel Calling

The variant caller proceeds with candidate indel discovery and generates alternate read alignments based on the candidate indels. As part of the realignment process, the variant caller selects a representative alignment to be used for site genotype calling and depth summarization by the SNV caller.

SNV Calling

The variant caller runs a series of filters on the set of filtered and realigned reads for SNV calling without affecting indel calls. First, any contiguous trailing sequence of N base calls is trimmed from the ends of reads. Using a mismatch density filter, reads having an unexpectedly high number of disagreements with the reference are masked, as follows:

- The variant caller treats each insertion or deletion as a single mismatch.
- Base calls with more than 2 mismatches to the reference sequence within 20 bases of the call are ignored.
- If the call occurs within the first or last 20 bases of a read, the mismatch limit is applied to a 41-base window at the corresponding end of the read.
- The mismatch limit is applied to the entire read when the read length is 41 or shorter.

Indel Genotypes

The variant caller filters out all bases marked by the mismatch density filter and any N base calls that remain after the end-trimming step. These filtered base calls are not used for site-genotyping but appear in the filtered base call counts in the variant caller output for each site.

All remaining base calls are used for site-genotyping. The genotyping method heuristically adjusts the joint error probability that is calculated from multiple observations of the same allele on each strand of the genome. This correction accounts for the possibility of error dependencies.

This method treats the highest-quality base call from each allele and strand as an independent observation and leaves the associated base call quality scores unmodified. Quality scores for subsequent base calls for each allele and strand are then adjusted. This adjustment is done to increase the joint error probability of the given allele above the error expected from independent base call observations.

Variant Call Output

After the SNV and indel genotyping methods are complete, the variant caller applies a final set of heuristic filters to produce the final set of calls in the output.

The output in the genome variant call (gVCF) file captures the genotype at each position and the probability that the consensus call differs from reference. This score is expressed as a Phred-scaled quality score.

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

Illumina On-Node Annotation (IONA)

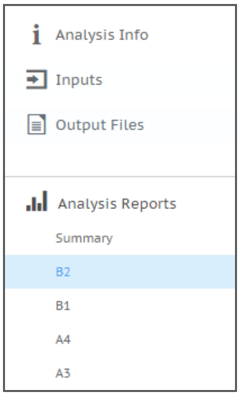
Annotation with IONA populates several values in the VCF file, including dbSNP ID (in the ID column), and some values in the INFO column. More detailed and extensive annotations are stored in a binary ANT file. This binary file can be imported into VariantStudio. For more information, see www.illumina.com/informatics/research/biological-data-interpretation/variantstudio.html.

Ensembl or RefSeq annotation through IONA is available for alignments against the human reference genome: UCSC build hg19 / Ensembl build GRCh37 / NCBI build 37.2.

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
AmpliconRunStatistics.xml	Provides amplicon-related information.

File Name	Description
CompletedJobInfo.xml	Contains information about the completed analysis session.
Logging.zip	Contains all detailed log files for each step of the workflow.
RunInfo.xml	Identifies the boundaries of the reads (including index reads).
SampleSheet.csv	Sample sheet.
SampleSheetUsed.csv	A copy of the sample sheet, generated at the beginning of the workflow.
[Manifest name].txt	Manifest files used in the analysis.

Output Files

The Output Files page provides access to the output files for each sample 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.
- **Summary File**—Statistics for each sample.

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.
ID	The rs number for the SNP obtained from dbSNP.txt, if applicable. If there are multiple rs numbers at this location, the list is semicolon delimited. If no dbSNP entry exists at this position, a missing value marker ('.') is used.
REF	The reference genotype. For example, a deletion of a single T is represented as reference TT and alternate T. An A to T single nucleotide variant is represented as reference A and alternate T.
ALT	The alleles that differ from the reference read. For example, an insertion of a single T is represented as reference A and alternate AT. An A to T single nucleotide variant is represented as reference A and alternate T.
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant and lower probability of errors. For a quality score of Q, the estimated probability of an error is $10^{-(Q/10)}$. For example, the set of Q30 calls has a 0.1% error rate. Many variant callers assign quality scores based on their statistical models, which are high in relation to the error rate observed.

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. • **LowQual**—The variant quality (QUAL) 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. Used with the Somatic Variant Caller and GATK.
INFO	Possible entries in the INFO column include: • **AC**—Allele count in genotypes for each ALT allele, in the same order as listed. • **AF**—Allele Frequency for each ALT allele, in the same order as listed. • **AN**—The total number of alleles in called genotypes. • **CD**—A flag indicating that the SNP occurs within the coding region of at least 1 RefGene entry. • **DP**—The depth (number of base calls aligned to a position and used in variant calling). • **Exon**—A comma-separated list of exon regions read from RefGene. • **FC**—Functional Consequence. • **GI**—A comma-separated list of gene IDs read from RefGene. • **QD**—Variant Confidence/Quality by Depth. • **TI**—A comma-separated list of transcript IDs read from RefGene.
FORMAT	The format column lists fields separated by colons. For example, GT:GQ. The list of fields provided depends on the variant caller used. 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. • **DP**—Approximate read depth; reads with MQ=255 or with bad mates are filtered. • **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. • **PL**—Normalized, Phred-scaled likelihoods for genotypes. • **SB**—Strand bias at this position. Larger negative values indicate less bias; values near 0 indicate more bias. Used with the Somatic Variant Caller and GATK. • **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 (< 3%) 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.

Annotation File

Illumina On-Node Annotation (IONA) generates a binary annotation file (*.ant), which contains consequences for all affected transcripts. The annotations are more detailed than the annotations in the VCF file. You can view this binary file in VariantStudio. For more information, see www.illumina.com/informatics/research/biological-data-interpretation/variantstudio.html.

Summary File

The TruSeq Amplicon v2.0 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 results folder for each sample and the aggregate results.

Statistic	Definition
Sample ID	IDs of samples reported in the file.
Sample Name	Names of samples reported in the file.
Run Folder	Run folders for samples reported in the file.
Reference genome	Reference genome selected.
Manifest	The manifest file used for analysis. This file specifies the targeted regions for the aligner and variant caller.
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 targeted 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.

Statistic	Definition
Total probe bases	Total number of bases that aligned to the probe sequences (ULSO and DLSO) and are soft-clipped in the BAM files. Numbers are calculated per read, and over both reads.
Total aligned non-probe bases	Total number of bases that aligned to the reference, excluding bases aligning to the probe sequences. This number is the same as the number of bases aligned in the BAM file (probe sequence bases are soft-clipped). Numbers are calculated per read, and over both reads.
Total PF bases	The number of bases passing filter for the sample. Numbers are calculated per read, and over both reads.
Percent Q30 bases	The percentage of bases with a quality score of 30 or higher. Numbers are calculated per read, and over both reads.
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.
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 reads to the targeted region divided by the number of targeted regions.
Uniformity of Coverage (Pct > 0.2*mean)	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.
SNVs, Insertions, Deletions (All)	Total number of variants present in the data set.
SNVs, Insertions, Deletions	Total number of variants present in the data set that pass the quality filters.
SNV Ts/Tv ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of 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.
SNVs, Insertions, Deletions (Percent found in dbSNP)	100*(Number of variants in dbSNP/Number of variants).
SNVs, Insertions, Deletions in genes	Number of variants that fall into a gene.

Statistic	Definition
SNVs, Insertions, Deletions in exons	Number of variants that fall into an exon.
SNVs, Insertions, Deletions in coding regions	Number of variants that fall into a coding region.
SNVs, Insertions, Deletions in UTR regions	Number of variants that fall into an untranslated region (UTR).
SNVs, Insertions, Deletions in splice site regions	Number of variants that fall into a splice site region.
Stop gained SNVs, Insertions, Deletions	Number of variants that cause an additional stop codon.
Stop lost SNVs, Insertions, Deletions	Number of variants that cause the loss of a stop codon.
Non-synonymous SNVs, Insertions, Deletions	Number of variants that cause an amino acid change in a coding region.
Synonymous SNVs	Number of variants that are within a coding region, but do not cause an amino acid change.
Frameshift Insertions, Deletions	Number of variants that cause a frameshift.

Sample Analysis Reports

The TruSeq Amplicon v2.0 App provides an overview of statistics per sample on the Analysis Reports sample pages. To download statistics in the TruSeq Amplicon Sequencing Report, click **PDF Summary Report**.

Amplicon Summary

Table 1 Amplicon Summary Table

Statistic	Definition
Manifest (PDF only)	The name of the manifest used in the analysis.
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 targeted region.

Table 2 Read Level Statistics Table

Statistic	Definition
Total Aligned Reads	The total number of reads passing filter present in the data set that aligned to the reference genome.
Percent Aligned Reads	The percentage of reads passing filter that aligned to the reference genome.

Table 3 Base Level Statistics Table

Statistic	Definition
Percent Q30	The percentage of bases with a quality score of 30 or higher.
Total Aligned Bases	The total number of bases present in the data set that aligned to the reference genome.
Percent Aligned Bases	The percentage of bases that aligned to the reference genome.
Mismatch Rate	The average percentage of mismatches across both reads 1 and 2 over all cycles.

Small Variants Summary

This table provides metrics about the number of SNVs, deletions, and insertions.

Table 4 Small Variants Summary Table

Statistic	Definition
Total Passing	The total number of variants present in the data set that passed the variant quality filters.
Percent Found in dbSNP	100*(Number of variants in dbSNP/Number of variants).
Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.
Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges between purine and pyrimidine bases (for example, A to T).

Table 5 Variants by Sequence Context Table

Statistic	Definition
Number in Genes	The number of variants that fall into a gene.
Number in Exons	The number of variants that fall into an exon.
Number in Coding Regions	The number of variants that fall into a coding region.

Statistic	Definition
Number in UTR Regions	The number of variants that fall into an untranslated region (UTR).
Number in Splice Site Regions	The number of variants that fall into a splice site region.

To view the guidelines for calculating variation consequences, visit the Ensembl website: uswest.ensembl.org/info/genome/variation/predicted_data.html#consequences.

Table 6 Variants by Consequence Table

Statistic	Definition
Frameshifts	The number of variants that cause a frameshift.
Non-synonymous	The number of variants that cause an amino acid change in a coding region.
Synonymous	The number of variants that are within a coding region, but do not cause an amino acid change.
Stop Gained	The number of variants that cause an additional stop codon.
Stop Lost	The number of variants that cause the loss of a stop codon.

Variants Table

Use the Variants Table to view, sort, filter, and export a subset of the data provided in the VCF files.

NOTE
The Variants table is provided on the Sample Analysis Reports page, and not in the TruSeq Amplicon Sequencing Report.

Table 7 Variants Table

Statistic	Definition
Chromosome (Chr)	Name of reference chromosome.
Position (Pos)	Position within reference chromosome.
Reference Allele (Ref)	The reference allele.
Variant Allele (Alt)	The alt allele.
Variant Type (Type)	Type of variant, including single nucleotide variant (SNV), insertion, and deletion.
Sequence Context (Context)	Location of the variant based on annotations of the reference genome.
Consequence	Predicted transcript consequence as described at uswest.ensembl.org/info/genome/variation/predicted_data.html#consequences.

Statistic	Definition
dbSNP ID (dbSNP)	Identifier in the Single Nucleotide Polymorphism Database (dbSNP), a free public archive for genetic variation within and across different species developed and hosted by the National Center for Biotechnology Information (NCBI).
COSMIC ID (COSMIC)	The numeric identifier for the variant in the Catalogue of Somatic Mutations in Cancer (COSMIC) database.
ClinVar	Clinical significance based on the freely accessible, public archive of reports of the relationships among human variations and phenotypes
Variant Quality (Qual)	Phred-scaled quality score indicating how confident we are in this asserted haplotype.
Variant Frequency (Alt Freq)	Proportion of the variant allele among all alleles being considered.
Total Depth	Number of reads aligned at this position.
Reference Allele Depth (Ref Depth)	Number of reads containing the reference allele.
Variant Allele Depth (Alt Depth)	Number of reads containing the variant allele.
Strand Bias	Strand bias is a type of sequencing bias in which 1 DNA strand is favored over the other, which can result in incorrect evaluation of the amount of evidence observed for one allele versus the other.

Coverage Summary

Table 8 Coverage Summary Table

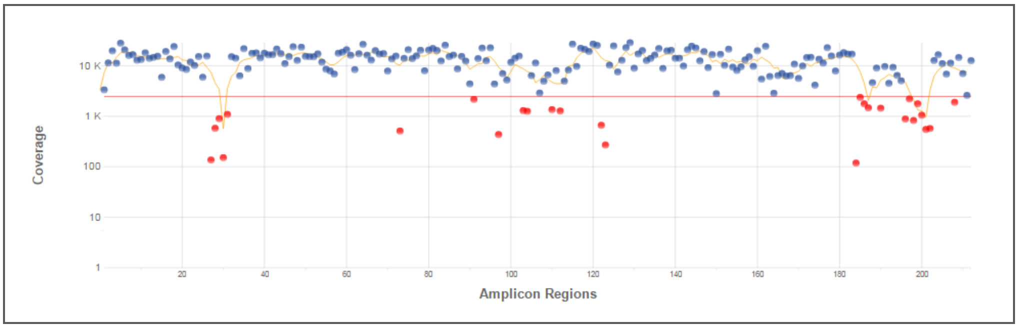
Statistic	Definition
Amplicon Mean Coverage	The total number of aligned reads to the targeted region divided by the number of targeted regions.
Uniformity of Coverage (Pct > 0.2*mean)	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.

Coverage by Amplicon Region Plot

The Coverage by Amplicon Region plot shows the number of bases plotted against the amplicon region. It has the following features:

- Amplicon regions with coverage values less than the low coverage threshold (0.2 * amplicon mean coverage) are highlighted in red.
- The horizontal red line marks the low coverage threshold.
- The orange line marks the moving average of all coverage values.
- Off-targets are excluded in the plot.

Figure 3 Example Coverage by Amplicon Region Plot



Coverage values for each amplicon are detailed in the downloadable Export (CSV) file.

Summary Analysis Report

The TruSeq Amplicon v2.0 App provides an Aggregate Summary for all samples. Statistics are plotted against samples with tables providing additional metrics.

Amplicon Summary

Table 9 Read Level Statistics Table

Statistic	Definition
Total Aligned Reads (R1/R2)	The total number of reads passing filter present in the data set that aligned to the reference genome. Numbers are per read.
Percent Aligned Reads (R1/R2)	The percentage of reads passing filter that aligned to the reference genome. Numbers are calculated per read.
Overall Aligned Reads	The percentage of reads passing filter that aligned for the sample across both reads.

Table 10 Base Level Statistics Table

Statistic	Definition
Total Aligned Bases (R1/R2)	The total number of bases present in the data set that aligned to the reference genome. Numbers are calculated per read.
Overall Total Aligned Bases	The total number of bases present in the data set that aligned to the reference genome across both reads (R1 and R2). The value is the sum of the individual Total Aligned Bases values.
Percent Aligned Bases (R1/R2)	The percentage of bases that aligned to the reference genome. Numbers are calculated per read, and the average over both reads.
Overall Percent Aligned Bases	The percentage of bases that aligned to the reference genome across both reads (R1 and R2). The value is the average of the individual Percent Aligned Bases values.
Percent Q30 (R1/R2)	The percentage of bases with a quality score of 30 or higher. Numbers are calculated per read.
Mismatch Rate (R1/R2)	The average percentage of mismatches across both reads 1 and 2 over all cycles. Numbers are calculated per read.

Small Variants Summary

Table 11 SNVs Table

Statistic	Definition
SNVs	The total number of SNVs present in the data set that passed the variant quality filters.
SNV Ts/Tv Ratio	Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges between purine and pyrimidine bases (for example, A to T).
SNV Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.

Table 12 Insertions Table

Statistic	Definition
Insertions	The total number of insertions present in the data set that passed the variant quality filters.
Insertion Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.

Table 13 Deletions Table

Statistic	Definition
Deletions	The total number of deletions present in the data set that passed the variant quality filters.
Deletions Het/Hom Ratio	Number of heterozygous variants/Number of homozygous variants.

Coverage Summary

Table 14 Coverage Summary Table

Statistic	Definition
Amplicon Mean Coverage Depth	The total number of aligned reads to the targeted region divided by the number of targeted regions.

Revision History

Document	Date	Description of Change
Document # 15055857 v02	January 2016	Reorganized topics, updated writing style.
Document # 15055857 v01	October 2015	Release supporting TruSeq Amplicon v2.0.

Notes

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 15 Illumina General Contact Information

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

Table 16 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