illumina®

Cancer Analysis Services Guide

FOR RESEARCH USE ONLY

ILLUMINA PROPRIETARY
Document # 15040893 v01
December 2015

Read Before Using this Product

This Product, and its use and disposition, is subject to the following terms and conditions. If Purchaser does not agree to these terms and conditions then Purchaser is not authorized by Illumina to use this Product and Purchaser must not use this Product.

1. **Definitions**. "**Application Specific IP**" means Illumina owned or controlled intellectual property rights that pertain to this Product (and use thereof) only with regard to specific field(s) or specific application(s). Application Specific IP excludes all Illumina owned or controlled intellectual property that cover aspects or features of this Product (or use thereof) that are common to this Product in all possible applications and all possible fields of use (the "**Core IP**"). Application Specific IP and Core IP are separate, non-overlapping, subsets of all Illumina owned or controlled intellectual property. By way of non-limiting example, Illumina intellectual property rights for specific diagnostic methods, for specific forensic methods, or for specific nucleic acid biomarkers, sequences, or combinations of biomarkers or sequences are examples of Application Specific IP. "**Consumable(s)**" means Illumina branded reagents and consumable items that are intended by Illumina for use with, and are to be consumed through the use of, Hardware. "**Documentation**" means Illumina's user manual for this Product, including without limitation, package inserts, and any other documentation that accompany this Product or that are referenced by the Product or in the packaging for the Product in effect on the date of shipment from Illumina. Documentation includes this document. "**Hardware**" means Illumina branded instruments, accessories or peripherals. "**Illumina**" means Illumina, Inc. or an Illumina affiliate, as applicable. "**Product**" means the product that this document accompanies (e.g., Hardware, Consumables, or Software). "**Purchaser**" is the person or entity that rightfully and legally acquires this Product from Illumina or an Illumina authorized dealer. "**Software**" means Illumina branded software (e.g., Hardware operating software, data analysis software). All Software is licensed and not sold and may be subject to additional terms found in the Software's end user license agreement. "**Specifications**" means Illumina's written specifications for this Product in effect on the date that the Product ships from Illumina.
2. **Research Use Only Rights**. Subject to these terms and conditions and unless otherwise agreed upon in writing by an officer of Illumina, Purchaser is granted only a non-exclusive, non-transferable, personal, non-sublicensable right under Illumina's Core IP, in existence on the date that this Product ships from Illumina, solely to use this Product in Purchaser's facility for Purchaser's internal research purposes (which includes research services provided to third parties) and solely in accordance with this Product's Documentation, **but specifically excluding any use that** (a) would require rights or a license from Illumina to Application Specific IP, (b) is a re-use of a previously used Consumable, (c) is the disassembling, reverse-engineering, reverse-compiling, or reverse-assembling of this Product, (d) is the separation, extraction, or isolation of components of this Product or other unauthorized analysis of this Product, (e) gains access to or determines the methods of operation of this Product, (f) is the use of non-Illumina reagent/consumables with Illumina's Hardware (does not apply if the Specifications or Documentation state otherwise), or (g) is the transfer to a third-party of, or sub-licensing of, Software or any third-party software. All Software, whether provided separately, installed on, or embedded in a Product, is licensed to Purchaser and not sold. Except as expressly stated in this Section, no right or license under any of Illumina's intellectual property rights is or are granted expressly, by implication, or by estoppel.

   Purchaser is solely responsible for determining whether Purchaser has all intellectual property rights that are necessary for Purchaser's intended uses of this Product, including without limitation, any rights from third parties or rights to Application Specific IP. Illumina makes no guarantee or warranty that purchaser's

specific intended uses will not infringe the intellectual property rights of a third party or Application Specific IP.

3 **Unauthorized Uses**. Purchaser agrees: (a) to use each Consumable only one time, and (b) to use only Illumina consumables/reagents with Illumina Hardware. The limitations in (a)-(b) do not apply if the Documentation or Specifications for this Product state otherwise. Purchaser agrees not to, nor authorize any third party to, engage in any of the following activities: (i) disassemble, reverse-engineer, reverse-compile, or reverse-assemble the Product, (ii) separate, extract, or isolate components of this Product or subject this Product or components thereof to any analysis not expressly authorized in this Product's Documentation, (iii) gain access to or attempt to determine the methods of operation of this Product, or (iv) transfer to a third-party, or grant a sublicense, to any Software or any third-party software. Purchaser further agrees that the contents of and methods of operation of this Product are proprietary to Illumina and this Product contains or embodies trade secrets of Illumina. The conditions and restrictions found in these terms and conditions are bargained for conditions of sale and therefore control the sale of and use of this Product by Purchaser.

4 **Limited Liability. TO THE EXTENT PERMITTED BY LAW, IN NO EVENT SHALL ILLUMINA OR ITS SUPPLIERS BE LIABLE TO PURCHASER OR ANY THIRD PARTY FOR COSTS OF PROCUREMENT OF SUBSTITUTE PRODUCTS OR SERVICES, LOST PROFITS, DATA OR BUSINESS, OR FOR ANY INDIRECT, SPECIAL, INCIDENTAL, EXEMPLARY, CONSEQUENTIAL, OR PUNITIVE DAMAGES OF ANY KIND ARISING OUT OF OR IN CONNECTION WITH, WITHOUT LIMITATION, THE SALE OF THIS PRODUCT, ITS USE, ILLUMINA'S PERFORMANCE HEREUNDER OR ANY OF THESE TERMS AND CONDITIONS, HOWEVER ARISING OR CAUSED AND ON ANY THEORY OF LIABILITY (WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), STRICT LIABILITY OR OTHERWISE).**

5 **ILLUMINA'S TOTAL AND CUMULATIVE LIABILITY TO PURCHASER OR ANY THIRD PARTY ARISING OUT OF OR IN CONNECTION WITH THESE TERMS AND CONDITIONS, INCLUDING WITHOUT LIMITATION, THIS PRODUCT (INCLUDING USE THEREOF) AND ILLUMINA'S PERFORMANCE HEREUNDER, WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), STRICT LIABILITY OR OTHERWISE, SHALL IN NO EVENT EXCEED THE AMOUNT PAID TO ILLUMINA FOR THIS PRODUCT.**

6 **Limitations on Illumina Provided Warranties. TO THE EXTENT PERMITTED BY LAW AND SUBJECT TO THE EXPRESS PRODUCT WARRANTY MADE HEREIN ILLUMINA MAKES NO (AND EXPRESSLY DISCLAIMS ALL) WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THIS PRODUCT, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR ARISING FROM COURSE OF PERFORMANCE, DEALING, USAGE OR TRADE. WITHOUT LIMITING THE GENERALITY OF THE FOREGOING, ILLUMINA MAKES NO CLAIM, REPRESENTATION, OR WARRANTY OF ANY KIND AS TO THE UTILITY OF THIS PRODUCT FOR PURCHASER'S INTENDED USES.**

7 **Product Warranty**. All warranties are personal to the Purchaser and may not be transferred or assigned to a third-party, including an affiliate of Purchaser. All warranties are facility specific and do not transfer if the Product is moved to another facility of Purchaser, unless Illumina conducts such move.

- a **Warranty for Consumables**. Illumina warrants that Consumables, other than custom Consumables, will conform to their Specifications until the later of (i) 3 months from the date of shipment from Illumina, and (ii) any expiration date or the end of the shelf-life pre-printed on such Consumable by Illumina, but in no event later than 12 months from the date of shipment. With respect to custom Consumables (i.e., Consumables made to specifications or designs made by Purchaser or provided to Illumina by, or on behalf of, Purchaser), Illumina only warrants that the custom Consumables will be made and tested in accordance with Illumina's standard manufacturing and quality control processes. Illumina makes no warranty that custom Consumables will work as intended by Purchaser or for Purchaser's intended uses.
- b **Warranty for Hardware**. Illumina warrants that Hardware, other than Upgraded Components, will conform to its Specifications for a period of 12 months after its shipment date from Illumina unless the Hardware includes Illumina provided installation in which case the warranty period begins on the date of installation or 30 days after the date it was delivered, whichever occurs first ("Base Hardware Warranty"). "Upgraded Components" means Illumina provided components, modifications, or enhancements to Hardware that was previously acquired by Purchaser. Illumina warrants that Upgraded Components will conform to their Specifications for a period of 90 days from the date the Upgraded Components are installed. Upgraded Components do not extend the warranty for the Hardware unless the upgrade was conducted by Illumina at Illumina's facilities in which case the upgraded Hardware shipped to Purchaser comes with a Base Hardware Warranty.
- c **Exclusions from Warranty Coverage**. The foregoing warranties do not apply to the extent a non-conformance is due to (i) abuse, misuse, neglect, negligence, accident, improper storage, or use contrary to the Documentation or Specifications, (ii) improper handling, installation, maintenance, or repair (other than if performed by Illumina's personnel), (iii) unauthorized alterations, (iv) Force Majeure events, or (v) use with a third party's good not provided by Illumina (unless the Product's Documentation or Specifications expressly state such third party's good is for use with the Product).
- d **Procedure for Warranty Coverage**. In order to be eligible for repair or replacement under this warranty Purchaser must (i) promptly contact Illumina's support department to report the non-conformance, (ii) cooperate with Illumina in confirming or diagnosing the non-conformance, and (iii) return this Product,

transportation charges prepaid to Illumina following Illumina's instructions or, if agreed by Illumina and Purchaser, grant Illumina's authorized repair personnel access to this Product in order to confirm the non-conformance and make repairs.

e **Sole Remedy under Warranty**. Illumina will, at its option, repair or replace non-conforming Product that it confirms is covered by this warranty. Repaired or replaced Consumables come with a 30-day warranty. Hardware may be repaired or replaced with functionally equivalent, reconditioned, or new Hardware or components (if only a component of Hardware is non-conforming). If the Hardware is replaced in its entirety, the warranty period for the replacement is 90 days from the date of shipment or the remaining period on the original Hardware warranty, whichever is shorter. If only a component is being repaired or replaced, the warranty period for such component is 90 days from the date of shipment or the remaining period on the original Hardware warranty, whichever ends later. The preceding states Purchaser's sole remedy and Illumina's sole obligations under the warranty provided hereunder.

f **Third-Party Goods and Warranty**. Illumina has no warranty obligations with respect to any goods originating from a third party and supplied to Purchaser hereunder. Third-party goods are those that are labeled or branded with a third-party's name. The warranty for third-party goods, if any, is provided by the original manufacturer. Upon written request Illumina will attempt to pass through any such warranty to Purchaser.

8 **Indemnification**.

a **Infringement Indemnification by Illumina**. Subject to these terms and conditions, including without limitation, the Exclusions to Illumina's Indemnification Obligations (Section 9(b) below), the Conditions to Indemnification Obligations (Section 9(d) below), Illumina shall (i) defend, indemnify and hold harmless Purchaser against any third-party claim or action alleging that this Product when used for research use purposes, in accordance with these terms and conditions, and in accordance with this Product's Documentation and Specifications infringes the valid and enforceable intellectual property rights of a third party, and (ii) pay all settlements entered into, and all final judgments and costs (including reasonable attorneys' fees) awarded against Purchaser in connection with such infringement claim. If this Product or any part thereof, becomes, or in Illumina's opinion may become, the subject of an infringement claim, Illumina shall have the right, at its option, to (A) procure for Purchaser the right to continue using this Product, (B) modify or replace this Product with a substantially equivalent non-infringing substitute, or (C) require the return of this Product and terminate the rights, license, and any other permissions provided to Purchaser with respect this Product and refund to Purchaser the depreciated value (as shown in Purchaser's official records) of the returned Product at the time of such return; provided that, no refund will be given for used-up or expired Consumables. This Section states the entire liability of Illumina for any infringement of third party intellectual property rights.

b **Exclusions to Illumina Indemnification Obligations**. Illumina has no obligation to defend, indemnify or hold harmless Purchaser for any Illumina Infringement Claim to the extent such infringement arises from: (i) the use of this Product in any manner or for any purpose outside the scope of research use purposes, (ii) the use of this Product in any manner not in accordance with its Specifications, its Documentation, the rights expressly granted to Purchaser hereunder, or any breach by Purchaser of these terms and conditions, (iii) the use of this Product in combination with any other products, materials, or services not supplied by Illumina, (iv) the use of this Product to perform any assay or other process not supplied by Illumina, or (v) Illumina's compliance with specifications or instructions for this Product furnished by, or on behalf of, Purchaser (each of (i) – (v), is referred to as an "Excluded Claim").

c **Indemnification by Purchaser**. Purchaser shall defend, indemnify and hold harmless Illumina, its affiliates, their non-affiliate collaborators and development partners that contributed to the development of this Product, and their respective officers, directors, representatives and employees against any claims, liabilities, damages, fines, penalties, causes of action, and losses of any and every kind, including without limitation, personal injury or death claims, and infringement of a third party's intellectual property rights, resulting from, relating to, or arising out of (i) Purchaser's breach of any of these terms and conditions, (ii) Purchaser's use of this Product outside of the scope of research use purposes, (iii) any use of this Product not in accordance with this Product's Specifications or Documentation, or (iv) any Excluded Claim.

d **Conditions to Indemnification Obligations**. The parties' indemnification obligations are conditioned upon the party seeking indemnification (i) promptly notifying the other party in writing of such claim or action, (ii) giving the other party exclusive control and authority over the defense and settlement of such claim or action, (iii) not admitting infringement of any intellectual property right without prior written consent of the other party, (iv) not entering into any settlement or compromise of any such claim or action without the other party's prior written consent, and (v) providing reasonable assistance to the other party in the defense of the claim or action; provided that, the party reimburses the indemnified party for its reasonable out-of-pocket expenses incurred in providing such assistance.

e **Third-Party Goods and Indemnification**. Illumina has no indemnification obligations with respect to any goods originating from a third party and supplied to Purchaser. Third-party goods are those that are labeled or branded with a third-party's name. Purchaser's indemnification rights, if any, with respect to third party goods shall be pursuant to the original manufacturer's or licensor's indemnity. Upon written request Illumina will attempt to pass through such indemnity, if any, to Purchaser.

Revision History

Document	Date	Description of Change
Document # 15040893 v01	December 2015	• Revised documentation to reflect changes in version 6 of the Illumina FastTrack Cancer Analysis Service pipeline. • Renamed Isaac Structural Variant Caller, Isaac Somatic Variant Caller, and Isaac Variant Caller to Manta, Strelka, and Starling, respectively.
Part # 15040893 Rev. C	June 2015	• Revised documentation to reflect changes in version 4 of the Illumina FastTrack Cancer Analysis Service pipeline. • Renamed Strelka and Manta to Isaac Somatic Variant Caller and Isaac Structural Variant Caller, respectively.
Part # 15040893 Rev. B	November 2014	Revised documentation to reflect changes in version 3 of the Illumina FastTrack WGS pipeline.
Part # 15040893 Rev. A	July 2013	Initial release.

Table of Contents

Getting Started

Chapter 1

Introduction

The Cancer Analysis Service Informatics Pipeline leverages a suite of proven algorithms that are optimized for the complexities of tumor samples to deliver a set of accurate somatic variants. Sequence reads are aligned and run through the Whole Genome Sequencing workflow. The BAM files are then used as inputs into the Cancer Analysis Service pipeline.

Software Packages

The Cancer Analysis Service pipeline uses the following software packages. For the software versions used, see the Software Versions table in the summary PDF report included with each deliverable.

Software	Description	Links
Strelka	Joint tumor/normal small-variant caller.	• Reference • Availability
Manta	Germline and Somatic SV (structural variant) caller. Calls SVs between 50 bp and 10 kb. Candidate variants < 50 bp are passed to Canvas.	• Reference • Availability
Canvas	Somatic CNV (copy number variant) caller > 10 kb. CNV calls under 10 kb are produced, but set to a filtered status.	• Reference: In Prep • Availability
Illumina Annotation Engine	Internal Annotation pipeline. Modeled from Ensembl's Variant Effect Predictor (VEP).	• Availability: Internal Only

Data Delivery

Illumina FastTrack Services currently provides data delivery through the following choices.

Illumina Hard Drive Data Delivery

Illumina FTS ships data on 1 or more hard drives. The hard drives are formatted with the NTFS file system and can optionally be encrypted.

The data on the hard drive are organized in a folder structure with 1 top-level folder per sample or analysis.

Illumina Cloud Data Delivery

Illumina FTS uploads data to a cloud container. Illumina currently supports uploads to the Amazon S3 service. Upload data are organized per upload batch by date with an Illumina_FTS prefix. For example, a sample in a batch uploaded on February 1, 2015 would be found in the container with the prefix Illumina_FTS/20140201/SAMPLE_BARCODE. Contact your FastTrack Services project manager to enable cloud delivery.

Analysis Deliverables

Chapter 2

Overview

This section details the files and folder structure for the cancer-normal somatic analysis deliverable. Normal and paired tumor samples are batched together at delivery, but each folder follows the same underlying format.

Though results from our Cancer Analysis Service are reported for tumor samples, the algorithms have been designed for and tested on diploid samples, and not heterogenetic tumor samples.

The files and folders generated for the cancer-normal somatic analysis results are based on the unique sample identifiers for both the cancer and normal sample. Usually, these unique identifiers are the barcodes associated with the cancer and normal samples in the lab, but can be a known sample ID for reference samples.

Result Folder Structure

Under each paired tumor-normal sample folder, you can find the following file structure that contains analysis results. Due to the quantity of DNA, samples run using our Nano service do not have genotyping information.

For detailed information on assembly, genotyping, variations files, and descriptions of the algorithms used to generate them, see the *Whole Genome Sequencing Services User Guide, document # 15040892*.

Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode]

- **Metrics**
  - **Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode].Metrics.json**—JSON formatted statistics that mirror statistics from the summary PDF report
- **SomaticVariations**
  - **Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somatic.vcf.gz**—Single nucleotide variant and small Insertion/Deletion somatic calls (1bp – 50bp) in *.vcf format
  - **Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somatic.SV.vcf.gz**—Somatic Structural Variation somatic calls (51bp + to 10kb) and somatic calls for regions with copy number aberrations (CNAs) (10kb +) and loss of heterozygosity (LOH) in *.vcf format
- **md5sum.txt**—Checksum file for confirming file consistency
- **Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].SummaryReport.pdf**—PDF report containing a brief overview of the somatic analysis results for the samples

NOTE

All the VCF files that Illumina provides are compressed and indexed using tabix. For details about tabix, see the tabix manual in SAMtools (at samtools.sourceforge.net/tabix.shtml).

The tabix index shows up as an additional Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].TYPE.vcf.gz.tbi file.TYPE.vcf.gz.tbi file. It can be used for fast retrieval of targeted regions in the associated *.vcf.gz file

NOTE

For some VCF files, a binary format of the annotations and their indexes are contained in corresponding *.vcf.ant and *.vcf.ant.idx files respectively. If the *.vcf.ant file is maintained in the same directory as its VCF file, the annotation information can be visualized alongside the variant call information when imported to VariantStudio.

Somatic Variations Folder

The somatic variations folder contains all the variant calls produced for the somatic analysis. The variant files that Illumina provides conform to the variant call format, VCF 4.1, specifications. For more information on the details of the VCF format, see www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41.

Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somatic.vcf.gz

The somatic small variants VCF file contains small variants ≤ 50 bp that are called using Strelka in VCF 4.1 format and annotated with the Illumina Annotation Engine.

This file contains the following metadata in the FILTER, FORMAT, and INFO fields.

FILTER Fields

ID	Description
BCNoiseIndel	Average fraction of filtered basecalls within 50 bases of the indel > 0.3.
HighDepth	Locus depth is > 3× mean chromosome depth in the normal sample.
LowQscore	The empirically fitted VQSR score is < 2.35.
QSI_ref	Normal sample is not homozygous ref or indel Q-score < 30 (ie, calls with NT! = ref or QSI_NT < 30).

FORMAT Fields

ID	Description
AU	Number of A alleles used in tiers 1 and 2.
CU	Number of C alleles used in tiers 1 and 2.
DP	Read depth for tier 1 (used + filtered).
DP2	Read depth for tier 2.
DP50	Average tier 1 read depth within 50 bases.
FDP	Number of base calls filtered from original read depth for tier 1.
FDP50	Average tier 1 number of base calls filtered from original read depth within 50 bases.
GU	Number of G alleles used in tiers 1 and 2.
SDP	Number of reads with deletions spanning this site in tier 1.
SUBDP	Number of reads below tier 1 mapping quality threshold aligned across this site.

ID	Description
SUBDP50	Average number of reads below tier 1 mapping quality threshold aligned across sites within 50 bases.
TAR	Reads strongly supporting alternate allele for tiers 1 and 2.
TIR	Reads strongly supporting indel allele for tiers 1 and 2.
TOR	Other reads (weak support or insufficient indel breakpoint overlap) for tiers 1 and 2.
TU	Number of T alleles used in tiers 1 and 2.

INFO Fields

ID	Description
ALTMAP	Tumor alternate allele read position MAP.
ALTPOS	Tumor alternate allele read position median.
DP	Combined depth across samples.
IC	Number of times RU repeats in the indel allele.
IHP	Largest reference interrupted homopolymer length intersecting with the indel.
MQ	IMS Mapping Quality.
MQ0	Number of MAPQ == 0 reads covering this record.
NT	Genotype of the normal sample in all data tiers, as used to classify somatic variants. One of the following {ref,het,hom,conflict].
OVERLAP	Somatic indel possibly overlaps a second indel.
PNOISE	Fraction of panel containing nonreference noise at this site.
PNOISE2	Fraction of panel containing more than 1 nonreference noise obs at this site.
QSI	Quality score for any somatic variant (ie, for the ALT haplotype to be present at a significantly different frequency in the tumor and normal sample).
QSI_NT	Quality score reflecting the joint probability of a somatic variant and NT.
QSS	Quality score for any somatic SNV (ie, for the ALT allele to be present a significantly different frequency in the tumor and normal sample).
QSS_NT	Quality score reflecting the joint probability of a somatic variant and NT.
RC	Number of times RU repeats in the reference allele.

ID	Description
ReadPosRankSum	Z=score from Wilcoxon rank sum test of Alt Vs. Ref read-position in the tumor.
RU	Smallest repeating sequence unit in inserted or deleted sequence.
SGT	Most likely somatic genotype excluding normal noise states.
SNVSB	Somatic SNV site strand bias.
SOMATIC	Somatic mutation.
SVTYPE	Type of structural variant.
TQSI	Data tier used to compute QSI.
TQSI_NT	Data tier used to compute QSI_NT.
TQSS	Data tier used to compute QSS.
TQSS_NT	Data tier used to compute QSS_NT.
VQSR	Recalibrated quality score expressing the phred-scaled probability of the somatic call being a FP observation.

Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somatic.SV.vcf.gz

The somatic structural variants (SV) VCF file contains large variants > 50 bp that are called using Manta and Canvas in VCF 4.1 format and annotated with the Illumina Annotation Engine.

This file contains the following metadata in the ALT, FILTER, FORMAT, and INFO fields.

ALT Fields

ID	Description
BND	Translocation break-end.
CNV	Copy number variable region.
DEL	Deletion.
DUP:TANDEM	Tandem Duplication.
INS	Insertion.
INV	Inversion.

FILTER Fields

ID	Description
CLT10kb	Canvas call with length < 10 kb.

ID	Description
MaxDepth	Normal sample site depth is > 3× of the mean chromosome depth near 1 or both variant break-ends.
MaxMQ0Frac	For a small variant (< 1000 bases) in the normal sample, the fraction of reads with MAPQ0 around either break-end is > 0.4.
MinSomaticScore	Somatic score is < 30.
MGE10kb	Manta DEL or DUP call with length ≥ 10 kb.
q10	Quality < 10.

FORMAT Fields

ID	Description
BC	Number of bins in the region.
CN	Copy number genotype for imprecise events.
MCC	Major chromosome count (equal to copy number for LOH regions).
PR	Spanning paired-read support for the ref and alt alleles in the order listed.
RC	Mean counts per bin in the region.
SR	Split reads for the ref and alt alleles in the order listed, for reads where P(allele|read) > 0.999.

INFO Fields

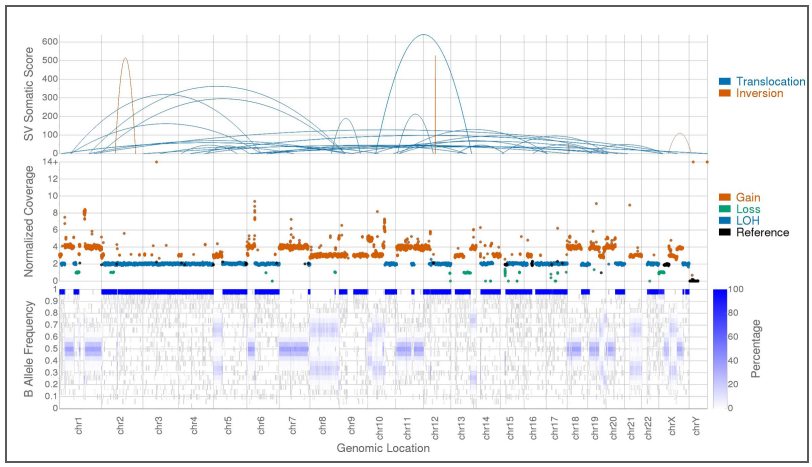
ID	Description
AA	The inferred allele ancestral (if determined) to the chimpanzee or human lineage.
AF1000G	The allele frequency from all populations of 1000 genomes data.
BND_DEPTH	Read depth at local translocation break-end.
CIEND	Confidence interval around END.
CIGAR	CIGAR alignment for each alternate indel allele.
CIPOS	Confidence interval around POS.
clinvar	Clinical significance. Format: GenotypeIndex|Significance
ColocalizedCanvas	Overlapped with a 10 kb+ Canvas call.

ID	Description
cosmic	The numeric identifier for the variant in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Format: GenotypeIndex|Significance
QSQR	Predicted regulatory consequence type. Format: GenotypeIndex|RegulatoryID|Consequence
CSQT	Consequence type as predicted by IAE. Format: GenotypeIndex|HGNC|Transcript ID|Consequence
END	End position of the variant described in this record.
EVENT	ID of event associated to break-end.
EVS	Allele frequency, coverage, and sample count taken from the Exome Variant Server (EVS). Format: AlleleFreqEVS|EVSCoverage|EVSSamples
GMAF	Global minor allele frequency (GMAF). Technically, the frequency of the second most frequent allele. Format: GlobalMinorAllele|AlleleFreqGlobalMinor
HOMLEN	Length of base pair identical microhomology at event breakpoints.
HOMSEQ	Sequence of base pair identical microhomology at event breakpoints.
IMPRECISE	Imprecise structural variation.
INV3	Inversion break-ends open 3' of reported location.
INV5	Inversion break-ends open 5' of reported location.
JUNCTION_ SOMATICSCORE	If the SV junction is part of an EVENT (ie, a multiadjacency variant), this field provides the SOMATICSCORE value only for the adjacency in question.
LEFT_SVINSSEQ	Known left side of insertion of an insertion of unknown length.
MATE_BND_DEPTH	Read depth at remote translocation mate break-end.
MATEID	ID of mate break-end.
phyloP	PhyloP conservation score. Denotes how conserved the reference sequence is between species throughout evolution.
RefMinor	Denotes positions where the reference base is a minor allele and is annotated as though it were a variant.
RIGHT_SVINSSEQ	Known right side of insertion of an insertion of unknown length.
SOMATIC	Somatic mutation.
SOMATICSCORE	Somatic variant quality score.
SVINSLEN	Length of insertion.
SVINSSEQ	Sequence of insertion.

ID	Description
SVLEN	Difference in length between REF and ALT alleles.
SVTYPE	Type of structural variant.

Summary Report

This PDF report contains a brief overview of the somatic analysis results for the samples and contains the following sections.

Section	Description
Tumor Sample Information	Contains information associated with reads and alignment quality from the tumor sequence input to the workflow.
Normal Sample Information	Contains information associated with reads and alignment quality from the normal sequence input to the workflow.
Somatic Small Variants Summary	Contains counts for the various types of reported small variants in the small variants VCF file, split up by totals, sequence context, and consequence (as calculated by the Illumina Annotation Engine).
Somatic Structural Variants Summary	Contains counts for the various types of reported large variants in the structural variants VCF file and counts of variants located in genes (as calculated by the Illumina Annotation Engine).
Structural Variants, Tumor Sample Coverage, and Allele Frequency	Details the estimated purity and ploidy for the cancer sample output from Canvas. For more information, see the Canvas Software Design Document. This section also includes a graph showing structural inversions, windowed copy number aberration and loss of heterozygosity plots, and B-allele frequency plots from Canvas. Figure 1 Example of Structural Variants, Tumor Sample Coverage, and Allele Frequency Graph 

Data Integrity

The md5sum.txt file is provided to check the integrity of the sample files and folders. Immediately after sample quality check, the md5sums, or compact digital fingerprint, for every file in the directory tree are generated. If media failures compromise data integrity, you can use the md5sum tool to find the inconsistencies. Use the tool to compare the hash from the provided md5sum file to the hash generated from the downloaded file.

On a Unix system, you can use the following commands to perform an md5sum check, assuming the utility is installed:

```
% cd [Sample_Barcode]
% md5sum -c md5sum.txt
```

The check verifies every file in ~30–45 minutes. Any errors are listed in the output.

In Windows, there are various command line and GUI tools available to perform an md5sum check. The Cygwin tools provide a utility identical to Linux.

Analysis Overview

Chapter 3

Overview

The somatic variant calling pipeline uses a normal BAM file and a tumor BAM file as input. In the tumor analysis pipeline, these BAM files are the result of the whole-genome sequencing pipeline described in the *Whole-Genome Sequencing Services User Guide (document # 15040892)*.

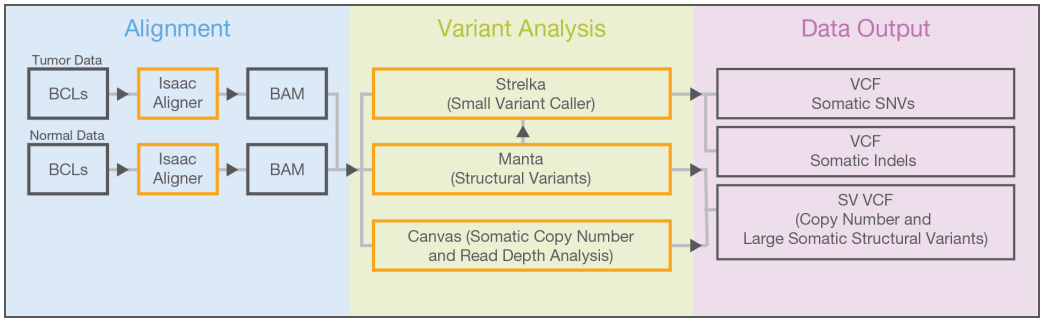
These BAM files are then processed through 3 interconnected callers:

- Somatic Small Variant Calls (Strelka)
- Large Indel and Structural Variant Caller (Manta)
- Copy Number Variations Caller (Canvas)

During the first stage of the pipeline, the tumor and normal BAM files run through a combined indel realignment operation. This realignment operation is used as the input for further processing. During calling, putative calls and *de novo* reassembled sections of sequence are passed between the callers to produce internally consistent variant calls.

All 3 callers use statistical models that operate on the combined tumor and normal reads as input instead of the variants. The statistical models use combined calling instead of subtraction of variant calls. Using combined calling produces superior results. However, subtraction of the calls from the normal and tumor whole genome results often do not match the somatic calls from a combined caller. For example, you can find a somatic variant that was not called in the tumor WGS sample because the combined caller is operating on the reads.

Figure 2 Cancer Analysis Pipeline



Strelka (Somatic Small Variant Caller)

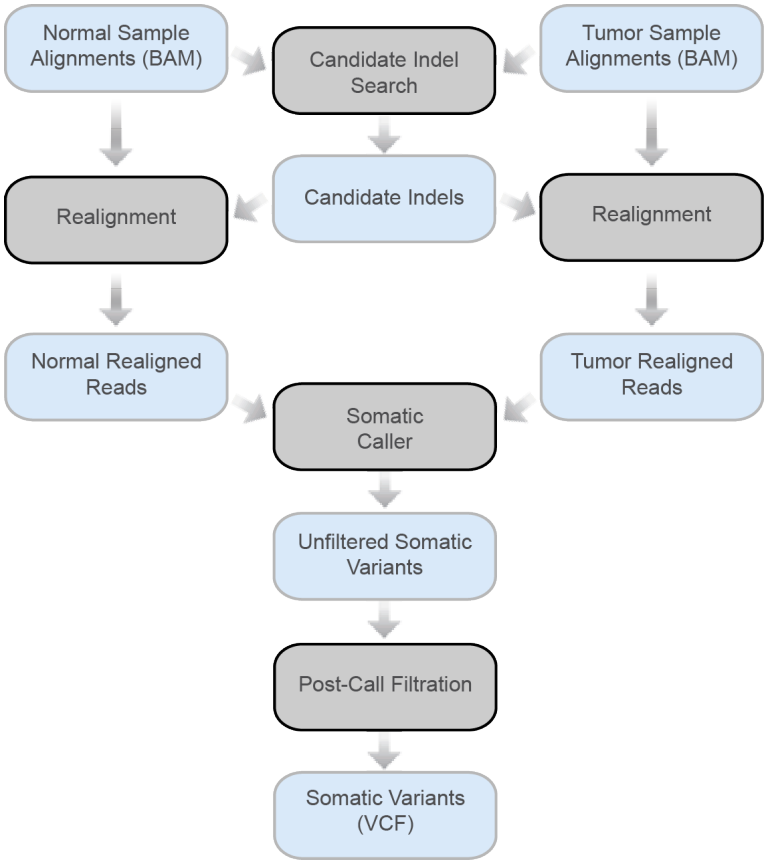
The somatic small variant calling method (Strelka) detects somatic SNVs and indels in sequencing data from a tumor and matched normal sample. For more information, see the publication Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs or post to the Strelka mailing list.

The analysis is based on the following assumptions:

- The normal sample is a mixture of diploid germline variation and noise.
- The tumor sample is a combination of the normal sample and somatic variation. It is assumed that the somatic variation and the normal noise can occur at any allele frequency ratio.

For SNVs, but not for indels, the normal noise component is further modeled as a combination of single-strand and double-strand noise.

Figure 3 Strelka Method



NOTE
For a detailed overview of Strelka methods, go to www.ncbi.nlm.nih.gov/pubmed/22581179.

Saunders,C.T., Wong,W.S., Swamy,S. *et al.* (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28, 1811–1817. doi: 10.1093/bioinformatics/bts271

Candidate Indel Search

Strelka scans through the genome using sequence alignments from the normal sample and tumor sample together to find a joint set of candidate indels. The information in sequence alignments is supplemented with externally generated candidate indels discovered by Manta. Manta provides external candidate indels to Strelka for indels of size 50 and below.

Candidate indels are used for realignment of reads, during which each candidate indel is evaluated as a potential somatic indel. Any other types of indels are considered noise indels. If a better alignment is not found, these indels are allowed to remain in the read alignments; otherwise, they are not used.

The candidate indel thresholds are designed so that the joint candidate indel set is at least the combined set found if the Small Variant Caller (Starling) is run on the individual samples. Specifically, where a minimum number of nominating reads is required for candidacy in Starling, Strelka requires the same minimum number of nominating reads from the combined input. Strelka requires that at least 1 sample contains a minimum fraction of supporting reads among the sample reads for candidacy.

For more information on Starling, see the *Whole-Genome Sequencing Services User Guide, document # 15040892*.

Realignment

For every read that intersects a candidate alignment, the Strelka attempts to find the most probable alignments including the candidate indel and excluding the candidate indel. Typically, the alignment excluding the candidate indel aligns to the reference, but occasionally an alternate indel that overlaps or interferes with the candidate is found to be more likely. The indel caller uses the probabilities of both alignments as part of the indel quality score calculation, whereas only a single alignment (usually the most probable) is preserved for SNV calling.

Somatic Caller

Strelka uses a Bayesian probability model similar to the one used for germline variant calling in the Starling Small Variant Caller or in external tools such as GATK. Using this model, our objective is to compute the posterior probability $P(\theta \mid D)$, which is the probability of the model state θ conditioned on the observed sequencing data.

In a germline variant caller, the state space of the model is conventionally a discrete set of diploid genotypes. For SNVs, the set of possible states is G= {"AA,CC,GG,TT,AC,AG,AT,CG,CT,GT"}.

The Strelka model instead approximates continuous allele frequencies for each allele:

$f=\{f_A, f_C, f_G, f_T\}$

The allele frequencies are restricted to allow a maximum of 2 nonzero frequencies. Any additional alleles observed in the data are treated as noise.

Another departure from typical germline calling methods is that the state space of the model is the allele frequency of both the tumor and the normal sample. In the following equation, f_t and f_n represent the allele frequencies of the tumor and normal samples, respectively.

$\theta=(f_t, f_n)$

The final somatic variant quality value reported by the model is computed from the probability that the allele frequencies are unequal (ie, $f_t \neq f_n$) given the observed sequence data.

Manta (Large Indel and Structural Variant Caller)

The large indel and structural variant calling method (Manta) is a structural variant caller for short sequencing reads. It can discover structural variants of any size and score these variants using both a diploid genotype model and a somatic model (when separate tumor and normal samples are specified). Structural variant discovery and scoring incorporate both paired read fragment spanning and split read evidence.

For more information, see the publication Manta: Rapid detection of structural variants and indels for clinical sequencing applications or the Manta GitHub.

Chen,X., Schulz-Trieglaff,O., Shaw,R. *et al.* (2015) Manta: Rapid detection of structural variants and indels for clinical sequencing applications. *Bioinformatics*. Advance online publication. doi: 10.1101/024232

Method Overview

Manta works by dividing the structural variant discovery process into 2 primary steps– scanning the genome to find SV associated regions and analysis, scoring, and output of SVs found in such regions.

1. **Build SV association graph**
   Scan the entire genome to discover evidence of possible SVs and large indels. This evidence is enumerated into a graph with edges connecting all regions of the genome that have a possible SV association. Edges can connect 2 different regions of the genome to represent evidence of a long-range association, or an edge can connect a region to itself to capture a local indel/small SV association. These associations are more general than a specific SV hypothesis, in that many SV candidates can be found on 1 edge, although typically only 1 or 2 candidates are found per edge.
2. **Analyze graph edges to find SVs**
   Analyze individual graph edges or groups of highly connected edges to discover and score SVs associated with the edges. The substeps of this process include:
   - Inference of SV candidates associated with the edge.
   - Attempted assembly of the SVs break-ends.
   - Scoring and filtration of the SV under various biological models (currently diploid germline and somatic).
   - Output to VCF.

Capabilities

Manta can detect all structural variant types that are identifiable in the absence of copy number analysis and large scale *de novo* assembly. Detectable types are enumerated in this section.

For each structural variant and indel, Manta attempts to align the break-ends to base pair resolution and report the left-shifted break-end coordinate (per the VCF 4.1 SV reporting guidelines). Manta also reports any break-end microhomology sequence and inserted sequence between the break-ends. Often the assembly fails to provide a confident explanation of the data. In such cases, the variant is reported as IMPRECISE, and scored according to the paired-end read evidence alone.

The sequencing reads provided as input to Manta are expected to be from a paired-end sequencing assay that results in an inwards orientation between the 2 reads of each DNA fragment. Each read presents a read from the outer edge of the fragment insert inward.

Detected Variant Classes

Manta is able to detect all variation classes that can be explained as novel DNA adjacencies in the genome. Simple insertion/deletion events can be detected down to a configurable minimum size cutoff (defaulting to 51). All DNA adjacencies are classified into the following categories based on the break-end pattern:

- Deletions
- Insertions
- Inversions
- Tandem Duplications
- Interchromosomal Translocations

Known Limitations

Manta cannot detect the following variant types:

- Nontandem repeats/amplifications
- Large insertions—The maximum detectable size corresponds to approximately the read-pair fragment size, but note that detection power falls off to impractical levels well before this size.
- Small inversions—The limiting size is not tested, but in theory detection falls off below ~200 bases. So-called microinversions might be detected indirectly as combined insertion/deletion variants.

More general repeat-based limitations exist for all variant types:

- Power to assemble variants to break-end resolution falls to 0 as break-end repeat length approaches the read size.
- Power to detect any break-end falls to (nearly) 0 as the break-end repeat length approaches the fragment size.
- The method cannot detect nontandem repeats.

While Manta classifies novel DNA-adjacencies, it does not infer the higher level constructs implied by the classification. For instance, a variant marked as a deletion by Manta indicates an intrachromosomal translocation with a deletion-like break-end pattern. However, there is no test of depth, b-allele frequency, or intersecting adjacencies to infer the SV type directly.

Canvas (Copy Number Variations Caller)

Canvas is a tool for calling copy number variants (CNVs) from human DNA sequencing data. Canvas can work with either germline data or paired tumor/normal samples. The primary input is aligned reads in BAM format and the primary output is a report VCF file that gives the copy number status of the genome.

For a description of Canvas and its algorithms, see the Canvas Software Design Document.

Appendix

Illumina FastTrack Services Annotation Pipeline

The FastTrack pipeline is an internal pipeline that provides the following annotations.

NOTE
These versions are specific to the time of publication of this document and can change with later updates. To determine the versions used, see the VCF file headers.

Source	Version	Release Date
dbSNP	144	06/06/2015
COSMIC	v73	06/06/2015
1000 Genomes Project	Phase 3 v5a	05/27/2013
EVS	V2	11/13/2013
ClinVar	Unknown	09/02/2015
phyloP	hg19	11/10/2009

In addition, the following annotations are added:

- Consequence predictions on RefSeq and Ensembl transcripts (modeled from VEP)
- Annotations in regulatory elements (modeled from VEP)
- Gene/transcript identifiers and their relationship between RefSeq, Ensembl, HGNC, and known synonyms (Gene Index)

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1 Illumina General Contact Information

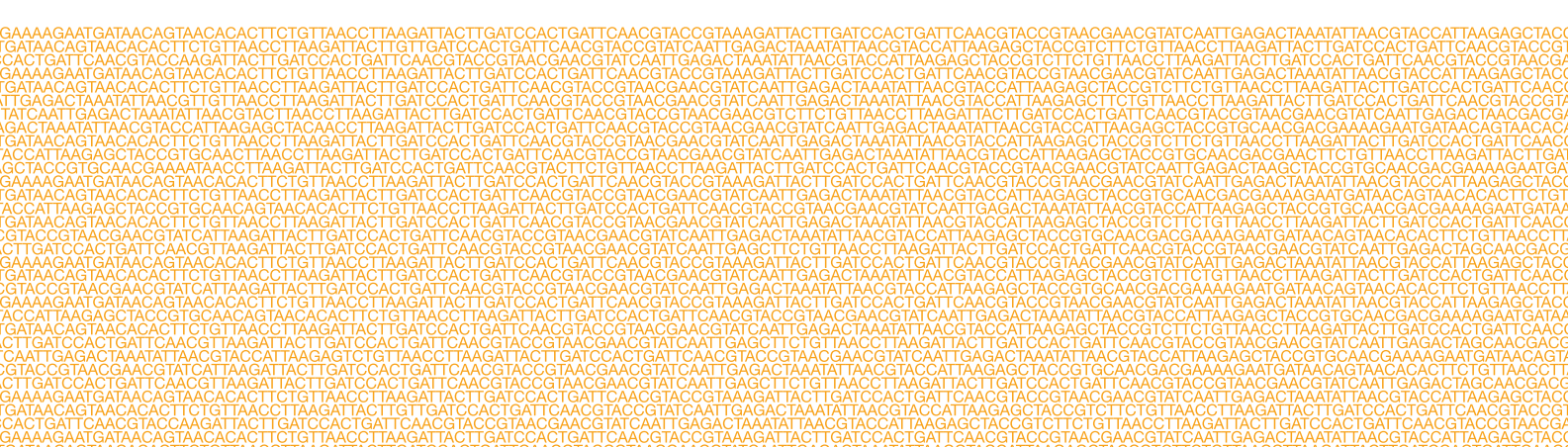
Website	www.illumina.com
Email	techsupport@illumina.com

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