illumina®

Cancer Analysis Services User Guide

FOR RESEARCH USE ONLY

ILLUMINA PROPRIETARY
Part # 15040893 Rev. C
June 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 **Regulatory**. This Product has not been approved, cleared, or licensed by the United States Food and Drug Administration or any other regulatory entity whether foreign or domestic for any specific intended use, whether research, commercial, diagnostic, or otherwise. This Product is labeled For Research Use Only. Purchaser must ensure it has any regulatory approvals that are necessary for Purchaser's intended uses of this Product.

4 **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.

5 **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).**

6 **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.**

7 **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.**

8 **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.

9 **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

Part #	Revision	Date	Description of Change
15040893	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.
15040893	B	November 2014	Revised documentation to reflect changes in version 3 of the Illumina FastTrack WGS pipeline.
15040893	A	July 2013	Initial release.

Table of Contents

Getting Started

Chapter 1

Cancer Analysis Service

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. High-quality sequence reads are aligned using the Isaac Alignment Software and somatic variant calling is performed using Isaac Somatic Variant Caller (Strelka), a combined Bayesian caller. Two complementary approaches enable detection of large somatic structural variations:

- Read depth analysis by SENECA for somatic structural variant events. See *Copy Number Aberrations (SENECA)* on page 26.
- Discordant paired-end analysis by Isaac Structural Variant Caller (Manta). See *Isaac Structural Variant Caller* on page 24.

Identified small somatic variants are reported with RefSeq annotations, COSMIC annotations, functional consequence predictions, and overlap with gene structure components and regulatory motifs.

This document provides an overview of the source and contents of the main files. Illumina creates these files using the informatics pipeline and information about key algorithms, like Isaac Somatic Variant Caller, Isaac Structural Variant Caller, and SENECA. The main files are to help you understand the Cancer Analysis Service data package that you receive from Illumina.

The following versions of software packages are utilized in the Cancer Analysis Service v4.0.2 pipeline.

Software	Version	Purpose
Isaac Somatic Variant Caller	2.0.14	Somatic SNV and indel caller.
Isaac Structural Variant Caller	0.23.1	Germline and somatic structural variant caller. Candidate variants of less than 50 kb are passed to Isaac Somatic Variant Caller.
SENECA	2.2.2	Somatic copy number aberration (CNA) caller.

NOTE
The BAM files from the whole genome workflow are used as input.

Part # 15040893 Rev. C

Data Delivery

Illumina FTS currently provides data delivery through the following choices.

Illumina Hard Drive Data Delivery

Illumina FastTrack Services 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 FastTrackServices uploads data to a cloud container. Illumina currently supports uploads to the Amazon S3 service. Upload data are organized per upload batch by date under an Illumina_FTS prefix. For example, a sample in a batch uploaded on February 1, 2014 would be found with the prefix Illumina_FTS/20140201/SAMPLE_BARCODE in the container. Contact your FastTrack Services project manager to enable cloud delivery.

Analysis Deliverables

Chapter 2

Analysis Folder Structure 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 Whole Genome Sequencing Service Pipeline are reported for tumor samples, the algorithms used 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 all keyed off the unique sample identifiers for both the cancer [CancerSampleBarcode] and normal sample [NormalSampleBarcode]. Usually, these unique identifiers are the barcodes associated with the cancer and normal samples in the lab (eg, LP600001_DNA-A01) but can be a known sample ID for reference samples (eg, HCC1187).

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 will 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, part # 15040892*.

Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode]/

[CancerSample_Barcode]/

Assembly

[Sample_Barcode].bam—Archival *.bam file for sample.

[Sample_Barcode].bam.bai—Index for *.bam file

[Sample_Barcode].SummaryReport.csv—Summary report in *.csv format

[Sample_Barcode].SummaryReport.pdf—Summary report in *.pdf format

Genotyping

[Sample_Barcode]_idats—Folder containing genotyping intensity data files for the sample (*.idat files) and genotyping sample sheet.

[Sample_Barcode].Genotyping.vcf.gz—Genotyping SNPs mapped to reference in *.vcf format.

[Sample_Barcode].GenotypingReport.txt—Genotyping SNPs tab delimited report.

Variations

[Sample_Barcode].CNV.vcf.gz—Copy number calls (10 kb +) in *.vcf format.

[Sample_Barcode].Indels.vcf.gz—Small Insertion/Deletion calls in *.vcf format.

[Sample_Barcode].SNPs.vcf.gz—Single nucleotide polymorphism (SNVs) calls in *.vcf format.

[Sample_Barcode].SV.vcf.gz—Large Structural Variation calls (51 bp–10 kb) in *.vcf format.

[Sample_Barcode].genome.vcf.gz—Genome *.vcf file containing SNVs, indels, and reference covered regions

[Sample_Barcode].vcf.gz—*.vcf file containing basic annotations and SNV and indel calls.

md5sum.txt—checksum file for confirming file consistency.

[NormalSample_Barcode]/

Assembly

[Sample_Barcode].bam—Archival *.bam file for sample.

[Sample_Barcode].bam.bai—Index for *.bam file

[Sample_Barcode].SummaryReport.csv—Summary report in *.csv format

[Sample_Barcode].SummaryReport.pdf—Summary report in *.pdf format

Genotyping

[Sample_Barcode]_idats—Folder containing genotyping intensity data files for the sample (*.idat files) and genotyping sample sheet.

[Sample_Barcode].Genotyping.vcf.gz—Genotyping SNPs mapped to reference in *.vcf format.

[Sample_Barcode].GenotypingReport.txt—Genotyping SNPs tab delimited report.

Variations

[Sample_Barcode].CNV.vcf.gz—Copy number calls (10 kb +) in *.vcf format.

[Sample_Barcode].Indels.vcf.gz—Small Insertion/Deletion calls in *.vcf format.

[Sample_Barcode].SNPs.vcf.gz—Single nucleotide polymorphism (SNVs) calls in *.vcf format.

[Sample_Barcode].SV.vcf.gz—Large Structural Variation calls (51 bp–10 kb) in *.vcf format.

[Sample_Barcode].genome.vcf.gz—Genome *.vcf file containing SNVs, indels, and reference covered regions

[Sample_Barcode].vcf.gz—*.vcf file containing basic annotations and SNV and indel calls.

md5sum.txt—checksum file for confirming file consistency.

SomaticVariations/

Cancer[CancerSample_Barcode]_Normal[NormalSample_Barcode].SummaryReport.pdf—Summary report in *.pdf format.

Metrics

Cancer[CancerSample_Barcode]_Normal[NormalSample_Barcode].Metrics.json—Metrics in *.json format.

Cancer[CancerSample_Barcode]_Normal[NormalSample_Barcode].somaticCNVs.vcf.gz—Somatic calls for regions with copy number aberrations (CNAs) (10 kb +) and loss of heterozygosity (LOH) in *.vcf format.

Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somaticIndel.vcf.gz—Small Insertion/Deletion somatic calls (1 bp—50 bp) in *.vcf format.

Cancer[CancerSampleBarcode]_Normal

Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somaticSNVs.vcf.gz—Single nucleotide variant somatic calls in *.vcf format.

Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].somaticSVs.vcf.gz—Somatic Structural Variation somatic calls (51 bp—10 kb) in *.vcf format.

md5sum.txt—checksum file for confirming file consistency.

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 [Sample_Barcode].TYPE.vcf.gz.tbi file. It can be used for fast retrieval of targeted regions in the associated vcf.gz file

The tabix index shows up as an additional Cancer[CancerSampleBarcode]_Normal [NormalSampleBarcode].TYPE.vcf.gz.tbi file and 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.

SomaticVariations

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].somaticSNVs.vcf.gz

SNV files contain single nucleotide variations, called through Isaac Somatic Variant Caller, for somatic analysis in VCF 4.1 format.

Table 1 INFO Fields

ID	Description
QSS	Quality score for any somatic SNV (ie, the ALT allele to be present at a significantly different frequency in the tumor and normal).
TQSS	Data tier used to compute QSS.
NT	Genotype of the normal in all data tiers, as used to classify somatic variants. One of {ref, het, hom, conflict}.
QSS_NT	Quality score reflecting the joint probability of a somatic variant and NT.
TQSS_NT	Data tier used to compute QSS_NT.
SGT	Most likely somatic genotype excluding normal noise states.
SOMATIC	Somatic mutation flag.

Table 2 FORMAT Fields

ID	Description
DP	Read depth for tier 1 (used + filtered).
FDP	Number of base calls filtered from original read depth for tier 1.
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.
AU	Number of A alleles used in tiers 1 and 2.
CU	Number of C alleles used in tiers 1 and 2.
GU	Number of G alleles used in tiers 1 and 2.
TU	Number of T alleles used in tiers 1 and 2.

Table 3 FILTER Fields

ID	Description
DP	Greater than 3x chromosomal mean depth in the normal sample.
BCNoise	Fraction of base calls filtered at this site in either sample is ≥ 0.4.
SpanDel	Fraction of reads crossing this site with spanning deletions in either sample is > 0.75.
QSS_ref	Normal sample is not homozygous with the reference or the SNV quality score (ssnv) is < 15 (ie, calls with NT!=ref or QSS_NT < 15).

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

Indel files contain indels, called through Isaac Somatic Variant Caller, for somatic analysis in VCF 4.1 format. Small indels are limited to 50 bp.

Table 4 INFO Fields

ID	Description
QSI	Quality score for any somatic variant (ie, the ALT haplotype to be present at a significantly different frequency in the tumor and normal sample).
TQSI	Data tier used to compute QSI.
NT	Genotype of the normal sample in all data tiers, as used to classify somatic variants. One of ref, het, hom, or conflict.
OVERLAP	Somatic indel possibly overlaps a second indel.
QSI_NT	Quality score reflecting the joint probability of a somatic variant and NT.
TQSI_NT	Data tier used to compute QSI_NT.
SGT	Most likely somatic genotype excluding normal noise states.
SOMATIC	Somatic Mutation flag.
SVTYPE	The type of structural variant.
RU	Smallest repeating sequence unit in inserted or deleted sequence.
RC	Number of times RU repeats in the reference allele.
IC	Number of times RU repeats in the indel allele.
IHP	Largest reference interrupted homopolymer length intersecting with the indel.

Table 5 FORMAT fields

ID	Description
DP	Read depth for tier 1.

ID	Description
DP2	Read depth for tier 2.
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 for tiers 1 and 2 (weak support or insufficient indel breakpoint overlap).
DP50	Average tier 1 read depth within 50 bases.
FDP50	Average tier 1 number of base calls filtered from original read depth within 50 bases.
SUBDP50	Average number of reads below tier 1 mapping quality threshold aligned across sites within 50 bases.

Table 6 FILTER Fields

ID	Description
DP	Greater than 3x chromosomal mean depth in the normal sample.
Repeat	Sequence repeats more than 8x in the reference sequence.
iHpol	Indel overlaps an interrupted homopolymer longer than 14x in the reference sequence.
BCNoise	Average fraction of filtered base calls within 50 bases of the indel is > 0.3.
QSI_ref	Normal sample is not homozygous with the reference or the indel quality score (somatic indel) is < 30 (ie, calls with NT!=ref or QSI_NT < 30).

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

The somatic SV file contains structural variants from 50 bp to 10 kb from the large indel and Isaac Structural Variant Caller called within the sample in VCF 4.1 format. The VCF file contains the following fields.

Table 7 INFO Fields

ID	Description
BND_DEPTH	Read depth at local translocation break-end.
BND_PAIR_COUNT	Confidently mapped reads supporting this variant at this break-end (it is possible that mapping is not confident at remote break-end).
CIEND	CIGAR alignment for each alternate indel allele.
CIGAR	Number of samples with data.
CIPOS	Confidence interval around POS.

ID	Description
DOWNSTREAM_AIR_COUNT	Confidently mapped reads supporting this variant at this downstream break-end (it is possible that mapping is not confident at upstream break-end).
END	End position of the variant described in this record.
HOMLEN	Length of base pair identical microhomology at event breakpoints.
HOMSEQ	Sequence of base pair identical microhomology at event breakpoints.
IMPRECISE	Imprecise structural variation.
MATE_BND_EPTH	Read depth at remote translocation mate break-end.
MATEID	String, ID of mate break-end.
PAIR_COUNT	Read pairs supporting this variant where both reads are confidently mapped.
SOMATIC	Somatic mutation.
SOMATICSCORE	Somatic variant quality score.
SVINSLEN	Integer, length of microinsertion at event breakpoints.
SVINSSEQ	Sequence of microinsertion at event breakpoints.
SVLEN	Difference in length between REF and ALT alleles.
SVTYPE	Type of structural variant.
UPSTREAM_PAIR_COUNT	Confidently mapped reads supporting this variant at the upstream break-end (it is possible that mapping is not confident at downstream break-end).

Table 8 FORMAT Fields

ID	Description
PR	Spanning paired read support for the ref and alt alleles in the order listed.
SR	Split reads for the ref and alt alleles in the order listed, for reads where P(allele|read) > 0.999.

Table 9 ALT Fields

ID	Description
BND	Translocation break-end.
COMPLEX	Unknown Candidate Type.
DEL	Deletion.
DUP:TANDEM	Tandem Duplication.
INS	Insertion.
INV	Inversion.

Table 10 FILTER Fields

ID	Description
MaxDepth	**Indicates that the normal sample site depth is greater than 3.0x of the mean chromosome depth.**
MaxMQ0Frac	For a small variant (< 1000 bases) in the normal sample, the fraction of reads with MAPQ0 around either break-end exceeds 0.4.
MinSomaticScore	Somatic score is less than 30.

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

The somatic CNA file contains copy number aberrations and loss of heterozygosity calls from the CNA module. This file is in VCF 4.1 format and contains the following fields.

Table 11 INFO Fields

ID	Description
SVTYPE	Type of structural variant (see ALT fields).
END	End position of the variant described in this record.
CN	Copy number genotype for imprecise events.
LOH	Loss of heterozygosity indicator.

Table 12 ALT Field

ID	Description
CNV	Copy number variable region.

Summary Report

The **Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode].SummaryReport.pdf** report contains an overview of the somatic analysis results for the samples, including the following sections:

- Sample Information—This section contains information associated with the samples from the provided sample manifest.
- Purity/Ploidy Estimates—This section details the estimated purity and ploidy for the cancer sample output from the Copy Number Aberration module. For more details, see *Ploidy and Purity Calculation* on page 26.
- Somatic Small Variants Summary—These 2 tables provide the total number of SNVs and Somatic Indels overlapping known variants and genes, exons, and coding regions. All counts are based on annotation and use only PASS filter variants.
- Somatic Structural Variants Summary—This table breaks CNA and Somatic SV output into the classes of variants called and their overlap with annotated genes. All counts are based on PASS filter variants.
- Circos Plot of Somatic Variations
- Depth/B allele Plot

For more information about this summary report, see the technical support note Molecular Characterization of Tumors Using next-generation sequencing.

The **[Sample_Barcode].SummaryReport.pdf** report contains an overview of the germline analysis results from the normal sample. For detailed information on this report, see the *Whole-Genome Sequencing Services User Guide, part # 15040892*.

Circos Plot of Somatic Variations

The Circos plot provides visualization of somatic small variation, ploidy, and structural variations reported in the somatic variation files (VCF). The Circos plot displays somatic variation data in tracks with chromosomes circularly arranged. Following is an example legend. Labels are described from inside the circle to the outside.

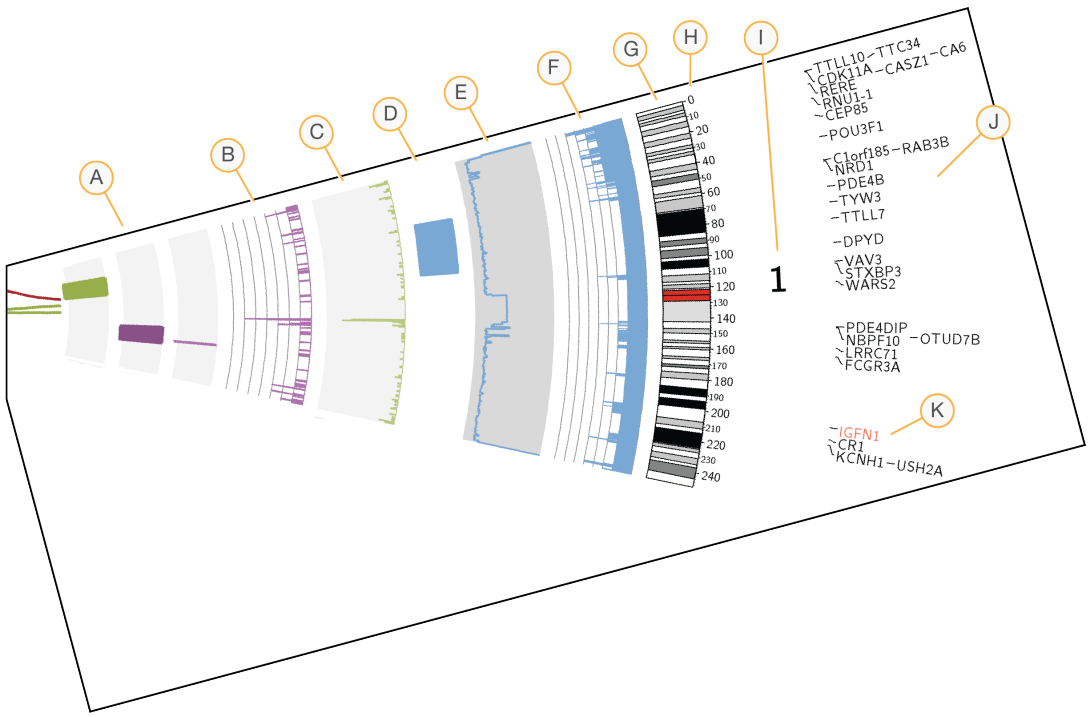


Table 13 Circos Plot Legend

Legend	Label (From Inner Circle to Outer Circle)	Description
A	Somatic structural variants	The somatic structural variants detailed in Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode].somaticSVs.vcf.gz are plotted in the center of the plot. • Green links—Segmental duplications (at the center of the circle). • Green boxes—Inversions (the first inner track). • Purple boxes—Deletions (the second track). The width of the boxes indicates the length of SVs. • Purple bars—Insertion breakpoints (the third track). • Red links—Translocations. The end of the links indicates the 2 breakpoints of SVs.
B	Number of somatic indels per Mb	The density of PASS somatic indels reported in Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode].somaticIndels.vcf.gz in 1 Mb windows. The scale of Y-axis in the histogram indicates the counts.
C	Number of somatic SNVs per Mb	The density of PASS somatic SNVs reported in Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode].somaticIndels.vcf.gz in 1 Mb windows, arbitrarily scaled in a histogram with Y-axis pointing inward.
D	Copy-neutral loss of heterozygosity (LOH)	The LOH regions with SNP calls in the normal genome but a homozygous reference call in the tumor genome, in Cancer[CancerSampleBarcode]_Normal[NormalSampleBarcode].CNAs.vcf.gz.

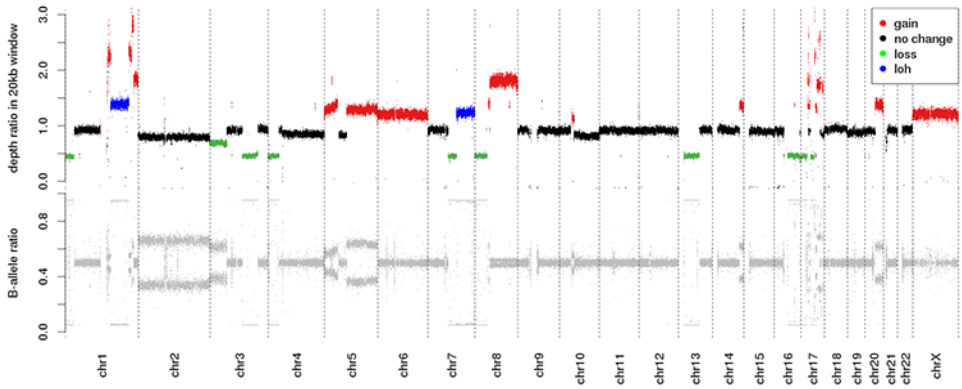
Legend	Label (From Inner Circle to Outer Circle)	Description
E	B allele frequency	The B allele ratios calculated by SENECA are used in the ploidy and purity estimation.
F	Called level	The copy number aberrations from Cancer [CancerSampleBarcode]_Normal [NormalSampleBarcode].CNAs.vcf.gz file. The scale of Y-axis in the histogram indicates the called level.
G	Karyotype	The standard Circos ideogram defining the chromosome position, identity, and color of cytogenetic bands.
H	Chromosome position	The reference coordinates along the chromosome (in megabases)
I	Chromosome number	Chromosome number: 1, 2,…,22, X, Y.
J	HGNC symbols for genes harboring variants	HGNC genes impacted by somatic SNVs. Genes containing SNVs in the coding region with an HGNC symbol are labeled.
K	Genes of nonsynonymous variants	Genes identified in (J) resulting in nonsynonymous changes in the coding region are highlighted in red.

Depth/B Allele Plot

The B allele plot provides the B allele frequency detected by SENECA (sensitive detection copy numbers in cancer package).

The top graph provides the ratio of the tumor read depth to the normal read depth after normalizing for sequencing coverage. Each point represents a 20 kb genomic region. Points are classified as either copy number gain (red), copy number loss (green), or copy number unchanged (black).

Figure 1 Example Graph



The bottom graph provides the variant allele frequencies in the tumor sample at dbSNP positions where the normal sample is heterozygous.

Data Integrity

The md5sum.txt file is provided as a means of checking 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 one 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 and require approximately 30–45 minutes to complete. 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

Analysis Overview Introduction

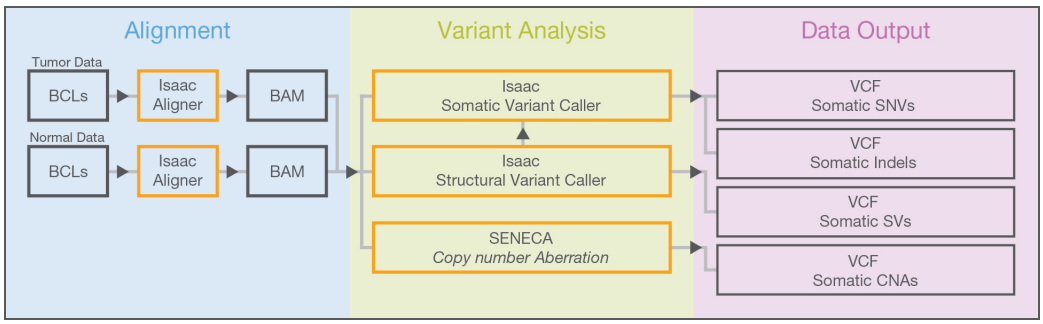
The somatic variant calling pipeline uses 2 aligned sequence files (*.bam files) as inputs–a normal *.bam and a tumor *.bam. In the tumor analysis pipeline, these *.bam files are the result of the whole-genome sequencing pipeline described in the *Whole-Genome Sequencing Services User Guide, part # 15040892*. These *.bam files are then processed through 3 interconnected callers:

- Isaac Somatic Variant Caller
- Isaac Structural Variant Caller
- Copy Number Aberration Caller (SENECA).

Isaac Somatic Variant Caller and SENECA are described in the following sections. For information on the Isaac Structural Variant Caller, see *Isaac Structural Variant Caller* on page 24.

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



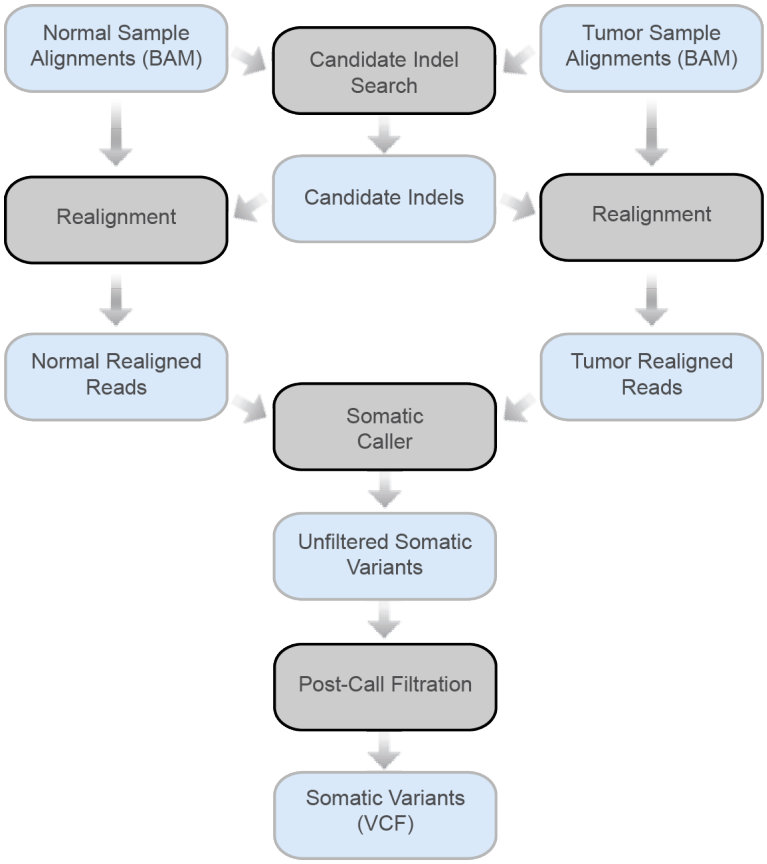
Isaac Somatic Variant Caller

The Isaac Somatic Variant Caller detects somatic SNVs and indels in sequencing data from a tumor and matched normal sample, 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 Isaac Somatic Variant Caller Method



NOTE
For a detailed overview of Isaac Somatic Variant Caller methods, go to www.ncbi.nlm.nih.gov/pubmed/22581179.

Candidate Indel Search

The Isaac Somatic Variant Caller caller 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 the Isaac Structural Variant (SV) Caller. Isaac SV Caller provides external candidate indels to Isaac Somatic Variant Caller 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 Isaac Variant Caller is run on the individual samples. Specifically, where a minimum number of nominating reads is required for candidacy in Isaac Variant Caller, Isaac Somatic Variant Caller requires the same minimum number of nominating reads from the combined input. Isaac Somatic Variant Caller requires that at least one sample contains a minimum fraction of supporting reads among the sample reads for candidacy.

For more information on the Isaac Variant Caller, see the *Whole-Genome Sequencing Services User Guide, part # 15040892*.

Realignment

For every read that intersects a candidate alignment, the Isaac Somatic Variant Caller 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

The Isaac Somatic Variant Caller uses a Bayesian probability model similar to the one used for germline variant calling in the Isaac 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 Isaac Somatic Variant Caller 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:

$\theta=(f_t, f_n)$

In the equation above, f_t and f_n represent the allele frequencies of the tumor and normal samples, respectively.

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.

Post-Call Filtration

Heuristic filters remove several types of improbable calls resulting from data artifacts that cannot be easily represented in the somatic probability model. These filters act as a final step to separate out the final set of somatic calls reported by Isaac Somatic Variant Caller.

Input Data Filtration

Isaac Somatic Variant Caller uses 2 tiers of input data filtration during somatic small variant calling:

- **Tier 1**—A more stringent filtering to ensure high quality calls
- **Tier 2**—A lower filtration stringency

Initially, candidates are called using a subset of the data with more stringent tier 1 filtering. If the method produces a nonzero quality score for any SNV or indel, the potential somatic variant is called again using data with a lower tier 2 stringency. The lower quality from the 2 tiers is selected for output. However, if the tier 2 quality is 0, the call is eliminated.

For somatic SNVs and indels, Isaac Somatic Variant Caller produces a general somatic quality score, Q(ssnv), or Q(somatic indel). This score indicates the probability of the somatic variant and a joint probability of the somatic variant and a specific normal genotype, Q(ssnv+ntype), or Q(somatic indel+ntype). The 2 tier evaluation is applied to each of these qualities separately, as follows:

Q(ssnv) = min(Q(ssnv | tier1), Q(ssnv | tier2))

Q(ssnv+ntype) = min(Q(ssnv+ntype | tier1), Q(ssnv+ntype | tier2))

The tier used for each quality value is provided in the Isaac Somatic Variant Caller output record for each somatic variant. If the most likely normal genotype is not the same at tier 1 and tier 2, then the normal genotype is reported as a conflict in the output.

Using 2 data tiers enables an initial somatic call based on high-quality data. Given a potential call, using 2 data tiers removes support for the putative somatic allele in the normal sample from lower quality data. The following table lists the primary data filtration levels that are changed between tier 1 and tier 2.

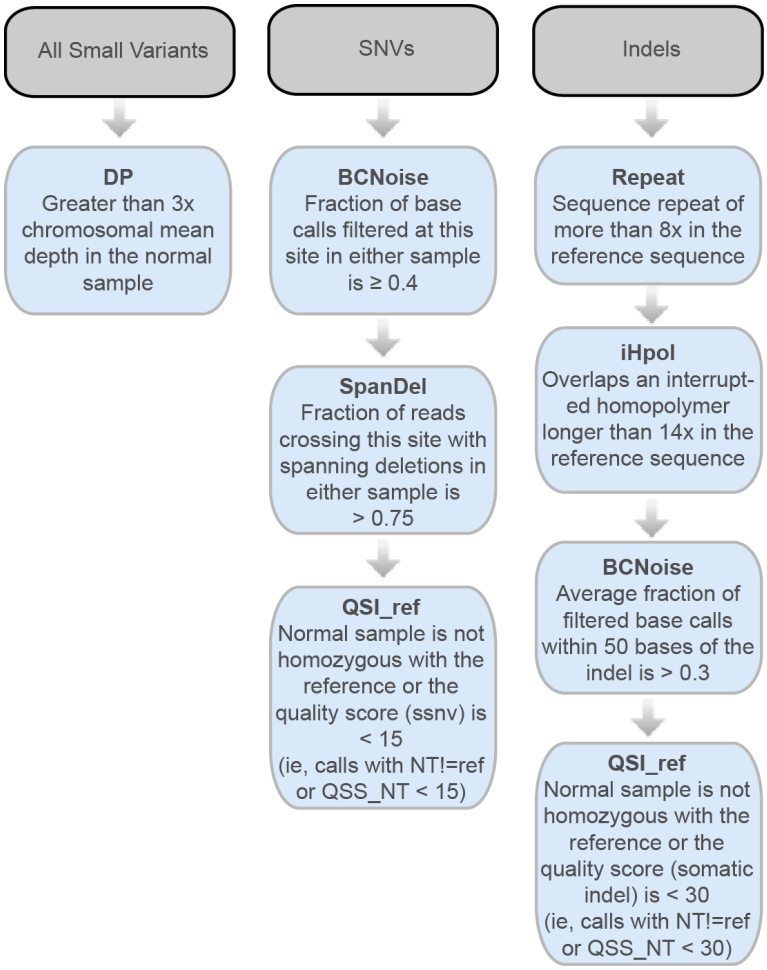
Table 14 Tiered Filtration Parameters

Parameter	Tier 1 Value	Tier 2 Value
Min paired-end alignment score	20	0
Min single-end alignment score	10	0
Single-end score rescue?	No	Yes
Include unanchored pairs?	No	Yes
Include anomalous pairs?	No	Yes
Include singleton pairs?	No	Yes
Mismatch density filter - max mismatches in window	3	10

Additional Filtration

Additional filters are applied after the somatic caller completes. A single candidate somatic call can be annotated with several filters, as described in the FILTER fields tables *SomaticVariations* on page 9.

Figure 4 Additional Filtration



Quality Filtration Levels

Only somatic calls originating from homozygous reference alleles in the normal sample are reviewed for validation and included in the output.

- Somatic SNVs are reported if the normal genotype is equal to the reference *and* Q(ssnv+ntype) ≥ 15.
- Somatic indels are reported if the normal genotype is equal to the reference *and* Q(somatic indel+ntype) ≥ 30.

NOTE
The value Q(ssnv+ntype) is associated with the VCF key **QSS_NT**.
The value Q(somatic indel+ntype) is associated with the VCF key **QSI_NT**.

Isaac Structural Variant Caller

Isaac Structural Variant (SV) Caller 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.

Method Overview

Isaac SV Caller 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**
   In this step, the entire genome is scanned 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**
   The second step is to analyze individual graph edges or groups of highly connected edges to discover and score SVs associated with the edges. These 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

Isaac SV Caller 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, Isaac SV Caller 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). Isaac SV Caller 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 Isaac SV Caller 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

Isaac SV Caller 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

Isaac SV Caller 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 Isaac SV Caller 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 Isaac SV Caller 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.

Copy Number Aberrations (SENECA)

The copy number aberrations module is also referred to as SENECA (SEnsitive detection of copy NumbErs in CAncer). It identifies copy number aberrations (CNAs) in heterogeneous tumor samples that exhibit contamination with normal tissues, aneuploidy, and loss of heterozygosity (LOH) that can confound correct copy assignment and lead to erroneous CNA calls.

The algorithm workflow comprises 2 distinct steps:

- Segmentation of data into regions with putatively distinct copy numbers.
- Calculation of ploidy and purity with a final copy number assignment.

As input, SENECA uses aligned sequences from tumor and matched normal samples (in *.bam format) and annotation information about the location of known variants in dbSNP, regional alignability, and the location of gaps in dbSNP.

Segmentation

SENECA is a count-based method to assign copy number state. It compares coverage between tumor and normal samples. Specifically, it bins read coverage using nonoverlapping 1 kb windows to derive counts in tumor and normal samples, and it then takes the ratio of the 2 counts. Bins are skipped during segmentation when they overlap low alignability regions in more than 20% of their size.

Independently, SENECA calculates B allele ratios at dbSNP positions from a tumor BAM file, and it keeps only SNVs that are heterozygous in the corresponding normal sample. Segmentation is carried out independently for copy number and B allele ratios.

Ploidy and Purity Calculation

Following segmentation, SENECA performs ploidy and purity calculations. These calculations are based on the principle that for each value of ploidy and purity and a selected copy number, the values of *B allele* and read count ratios are inferred. For example, for copy number state 1 (1 deleted allele of a diploid genome), the *B allele* ratio is always near 0 because only 1 allele is present. However, if a tumor sample has only 70% percent purity because of the presence of the normal genome as background, the *B allele* ratio increases due to the presence of a heterozygous normal allele. The low percentage of purity results in a final B allele ratio of 0.15.

SENECA fits a multivariate Gaussian distribution to copy data and *B allele* ratio data on a two-dimensional grid of varying ploidy and purity. On the grid, each state encodes ploidy and purity values. In addition, SENECA uses a separate state encoding copy neutral LOH and copy gain LOH to identify loss-of-heterozygosity events.

Ploidy and purity associated with the model having highest log-likelihood are then used to assign a copy number state to each segment. When both segments and copy numbers are estimated, a quality score for copy number assignment is computed using a likelihood ratio test. This test compares the likelihood of a current copy number assignment to a likelihood of assigning 1 more or 1 less copy. Results of the likelihood ratio test are then reported as a Q-score field in the VCF file using the following transformation: 2*log (s1/s2), where s1 is a sum of squares for selected model and s2 is a sum of squares for the next nearest model. Q-score threshold of 1.5 provides a good trade-off between sensitivity and specificity.

CNA Output

SENECA produces a genome-wide plot, a per-chromosome plot, and a VCF file.

The genome-wide plot shows distribution of copy number and B allele ratios; copy number ratios are classified as either gains (red) or losses (green). Estimated purity and ploidy values are listed at the top of the plot.

Per-chromosome plots list the distribution of either B allele or copy number ratios. Plots also report identified segments as black lines using a second Y-axis of copy number states, which range from 0 to a maximum of 9 copies. Capped segments are indicated in red. Segments exhibiting LOH are indicated in green. Uncapped values are reported in a VCF file generated in VCF 4.1 format.

CNV VCF

The following metadata is used in INFO fields of the Copy Number Variations (CNV) VCF file.

ID	Description
<CNV>	Indicates that the alternate allele is reported.
SVTYPE	Specifies the structural variant type, which is CNV in a CNV VCF file.
CN	Reports the copy number of each segment.
LOH	A binary indicator of the presence or absence of LOH for a given segment.

The following VCF example record shows a copy number gain of 6.

```
chr2 140982000 chr2_141842999 G <CNV> 1.75 PASS
   SVTYPE=CNV;END=141842999;CN=6;LOH=0
```

The following VCF example record shows a copy number gain of 3 and an LOH event.

```
chr2 141843000 chr2_205542999 A <CNV> 3.71 PASS
   SVTYPE=CNV;END=205542999;CN=3;LOH=1
```

Appendix

Illumina FastTrack Services Annotation Pipeline

The Illumina FastTrack Services Annotation Pipeline provides variant annotation for Single Nucleotide variants (SNVs), insertions, and deletions (indels). All annotations are provided in the INFO field of *[Sample_Barcode].vcf.gz* file and documented in the header.

Larger variants (CNAs, SVs) are not annotated with the full pipeline. The annotation database is queried for each of the small variants input to the pipeline. Both positional and allelic annotations can be returned for a given variant. After querying the annotation database, novel variants (variants for which no annotation exists) are then processed with VEP. If VEP does not return an annotation for the variant, it will remain unannotated.

Annotation Database Sources

The following table includes sources for the annotation databases.

Table 15 List of Annotation Database Sources

Source	Version	Release Date
Variant Effect Predictor	72	06/01/2013
1000 Genomes Allele Frequencies	v3, Release 20110521	04/30/2012
ClinVar	20130905	09/05/2013
COSMIC	65	05/28/2013
dbSNP	137	06/16/2012
HGNC/RefSeq Mapping	Updated daily	07/01/2013
NHLBI Exome Variant Server	v.0.0.20 ESP6500SI-V2	06/07/2013
phastCons	N/A	12/06/2009

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 16 Illumina General Contact Information

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

Table 17 Illumina Customer Support Telephone Numbers

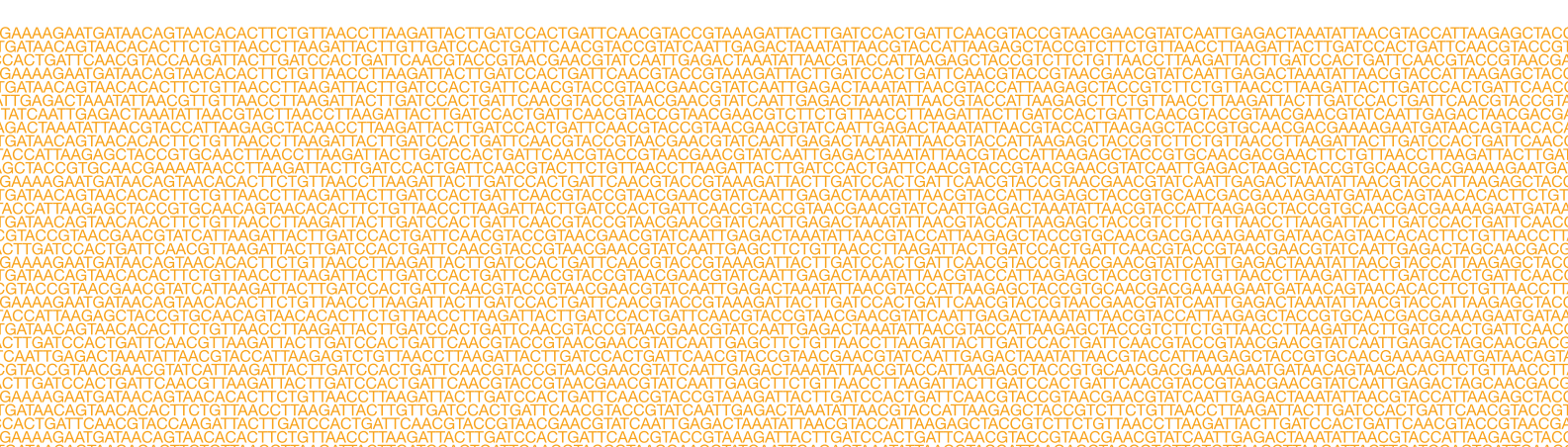
Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Italy	800.874909
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850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

Safety Data Sheets

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

Product Documentation

Product documentation in PDF is available for download from the Illumina website. Go to support.illumina.com, select a product, then click **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