APPLICATION MANUAL

MSK-IMPACT® Flex
powered with SOPHiA DDM™

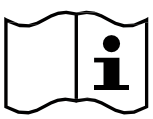
SOPHiA DDM™

SUMMARY INFORMATION

PARAMETER	VALUE
Product name	MSK-IMPACT® Flex powered with SOPHiA DDM™
Sample type	DNA and/or RNA isolated from formalin-fixed, paraffin embedded (FFPE) tumor tissue specimens or from fresh-frozen (FF) samples.
Compatible Library Preparation Kit(s)	DNA: SOPHiA GENETICS™ Universal Library Prep with CUMIN™ Adapters RNA: SOPHiA GENETICS™ Universal Library Prep for fragmented DNA
Gene Panel ID	DNA: SG_MSKIMPACTFLEX_v1 RNA: RCROS_A_v1a
Sequencer	Illumina® NextSeq™ 500/550 Illumina® NextSeq™ 1000/2000 Illumina® NovaSeq™ 6000 Illumina® NovaSeq™ X
Bioinformatics pipeline ID	ILL1XG1S6_FFPE_CNV_NextSeq_18 ILL1XG1S6_FFPE_CNV_NextSeq2000_3 ILL1XG1S6_FFPE_CNV_NovaSeq_13 ILL1XG1S6_FFPE_CNV_NovaSeqX_1
Reference genome	GRCh37 (hg19)
Analytical modules	DNA based analysis: SNV/INDEL, Structural variants, CNV (gene level and exon level), MSI, TMB, HRD genomic integrity, Tumor content and ASCN RNA based analysis: Gene fusions and exon skipping, Gene expression
Product codes	BS0132ILLRSMY13 CS2517ILLRSRY16 CS2517ILBRSRY16 (96 rxn only) DL0121ILLRSM
Document ID and version	SG-06426 v1.2

This Application Manual is applicable to all SOPHiA DDM™ versions.

Please read the Application Manual thoroughly before using this product.



DISCLAIMER

This document and its contents are the property of SOPHiA GENETICS SA and its affiliates ("SOPHiA GENETICS") and are intended solely for the contractual use by its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced, or referenced to in any way whatsoever without the prior written consent of SOPHiA GENETICS.

SOPHiA GENETICS does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document. The instructions in this document must be strictly and explicitly followed by qualified and adequately trained personnel to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood before using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT (S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY.

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

TRADEMARKS

SOPHiA GENETICS™, OncoPortal™ Plus, CUMIN™, PEPPER™, MUSKAT™, GIINGER™, MUSTARD™, MOKA™, and SOPHiA DDM™ are trademarks of SOPHiA GENETICS SA and/or its affiliate(s) in the U.S. and/or other countries. All other names, logos, and other trademarks are the property of their respective owners. UNLESS SPECIFICALLY IDENTIFIED AS SUCH, SOPHiA GENETICS USE OF THIRD-PARTY TRADEMARKS DOES NOT INDICATE ANY RELATIONSHIP, SPONSORSHIP, OR ENDORSEMENT BETWEEN SOPHiA GENETICS AND THE OWNERS OF THESE TRADEMARKS. Any references by SOPHiA GENETICS to third party trademarks are to identify the corresponding third-party goods and/or services and shall be considered nominative fair use under the trademark law.

REVISION HISTORY

DOCUMENT ID / VERSION	DATE	DESCRIPTION OF CHANGE

SG-06426 v1.2	Apr 2026	• *Document Updates*: Minor changes to the text were performed.
		• *8.12 DNA Analysis – Sample tumor content and ASCN analysis*: Added a new section describing the ASCN analysis workflow
		• *9 Precautions and Limitations:*
		  ○ *9.1 Precautions:* Updated to include considerations related to tumor content estimation and ASCN reliability.
		  ○ *9.2 General limitations:* Updated to include limitations specific to tumor content and ASCN analysis.

SG-06426 v1.1	Jan 2026	• *Document Updates*: Minor changes to the text were performed. • *6 Data Upload*: Inserted subsection 6.1 describing manual FASTQ file upload. Existing CLI upload instructions were renumbered to subsection 6.2 .
SG-06426 v1.0	Oct 2025	• Initial release

TABLE OF CONTENTS

1 APPLICATION OVERVIEW

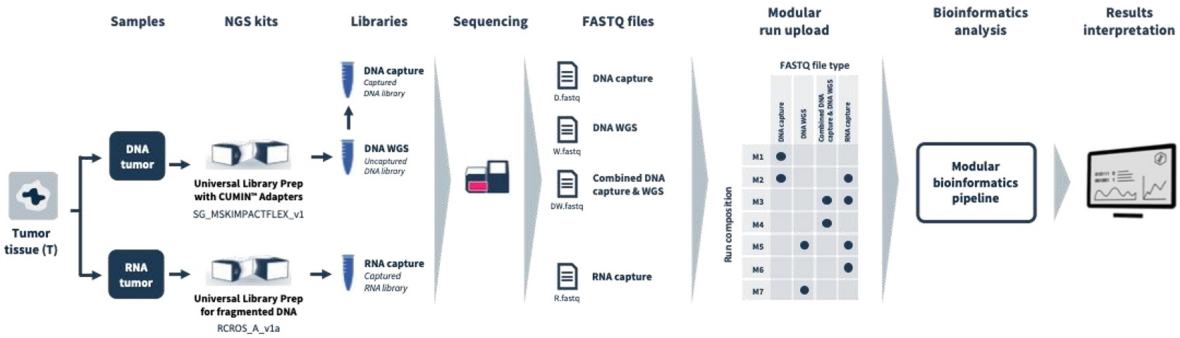
1.1 Intended purpose

MSK-IMPACT® Flex powered with SOPHiA DDM™ is a next-generation sequencing (NGS)–based comprehensive genomic profiling (CGP) solution for research use only (RUO). The assay interrogates 533 DNA genes and 140 RNA genes, enabling modular analysis starting from DNA, and/or RNA. It detects a broad spectrum of genomic alterations including single nucleotide variants (SNVs), insertions/deletions (indels), copy number variations (CNVs, including quantification of allele-specific absolute copy number), loss of heterozygosity (LOH), gene fusions, exon skipping events, and complex signature biomarkers such as homologous recombination deficiency (HRD) genomic instability, tumor mutational burden (TMB), and microsatellite instability (MSI). MSK-IMPACT® Flex powered with SOPHiA DDM™ provides researchers with a versatile solution to study cancer biology, assess or discover biomarkers, and advance translational research. The assay is not intended for use in diagnostic procedures or patient management.

1.2 Technical description of the application

MSK-IMPACT® Flex powered by SOPHiA DDM™ is an end-to-end solution that includes library preparation reagents for processing DNA and/or RNA samples extracted from tumor tissue, together with the SOPHiA DDM™ platform, which hosts the analysis package and provides a user interface for NGS data upload, results visualization, report generation, and download.

The following image provides an overview of the MSK-IMPACT® Flex powered with SOPHiA DDM™ application, from tumor tissue (left) to results visualization (right):



MSK-IMPACT® Flex powered with SOPHiA DDM™ comprises two NGS kits that respectively allow to process DNA and RNA (or total nucleic acid; TNA) samples extracted from tumor tissue, including formalin-fixed, paraffin-embedded (FFPE) samples. The NGS kits can be used either for a DNA-only workflow, for a RNA-only workflow, or as part of a dual DNA-RNA workflow to generate up to three distinct NGS libraries:

1. DNA whole-genome sequencing (WGS) library: a whole-genome DNA library
2. DNA capture library: an enriched DNA library, obtained by performing capture hybridization of the WGS DNA library via the SG_MSKIMPACTFLEX_v1 gene panel (533 genes)
3. RNA capture library: an enriched RNA library, obtained by performing capture hybridization via the RCROS_A_v1a gene panel (140 genes)

MSK-IMPACT® Flex powered with SOPHiA DDM™ is designed with a modular approach, enabling users to sequence in combination DNA capture libraries, RNA capture libraries, and DNA WGS libraries (at ~1x coverage) in various configurations according to their needs. For each tumor tissue analyzed, users can choose from seven possible sequencing modalities:

M1. DNA capture only
M2. DNA capture and matched RNA capture
M3. DNA capture combined with DNA WGS and matched RNA capture
M4. DNA capture combined with DNA WGS
M5. DNA WGS and matched RNA capture
M6. RNA capture only
M7. DNA WGS only

MSK-IMPACT® Flex powered with SOPHiA DDM™ allows to multiplex, within the same run, tumor tissues requiring different sequencing modalities. As shown in the table below, depending on the sequencing modality used for a given sample, one or two FASTQ files will be obtained. Four possible NGS data types are generated:

1. DNA capture FASTQ file (denoted "-D"): a FASTQ file containing high-coverage DNA sequencing data for the genomic regions targeted by the SG_MSKIMPACTFLEX_v1 gene panel
2. DNA WGS FASTQ file (denoted "-W"): a FASTQ file containing low-coverage DNA sequencing data across the whole-genome
3. DNA capture combined with DNA WGS FASTQ file (denoted "-DW"): a FASTQ file containing high-coverage DNA sequencing data for the genomic regions targeted by the SG_MSKIMPACTFLEX_v1 gene panel as well as low-coverage DNA sequencing data across the whole-genome
4. RNA capture FASTQ file (denoted "-R"): a FASTQ file containing high-coverage RNA sequencing data in RCROS_A_v1a target regions

All FASTQ files obtained from the sequencing run can be uploaded to the SOPHiA DDM™ platform via a single upload, regardless of the various sequencing modalities used for each individual tumor tissue. The SOPHiA DDM™ platform automatically recognizes the

sequencing modality for each tissue and performs the appropriate bioinformatics analysis.

The current version of the product does not yet support the combined analysis of FASTQ files obtained using sequencing modality M5 (DNA WGS and matched RNA capture). Users are recommended to upload -W and -R FASTQ files as if they were obtained from independent tumor tissues.

Depending on the sequencing modality, the modular bioinformatics pipeline automatically selects a set of bioinformatics analyses (a detailed description of each individual bioinformatics analysis is provided in TECHNICAL DESCRIPTION OF THE BIOINFORMATIC ANALYSIS) from the following list :

1. DNA-based small variant detection (SNV/INDEL): Identification of short single nucleotide variants and small insertions or deletions.
2. DNA-based gene fusion and exon-skipping detection (Gene fusions and exon skipping): Identification of gene fusions and exon skipping events from DNA data.
3. DNA-based gene-level copy number variation analysis (CNV gene level): Detection of whole-gene amplifications and deletions on targeted genes (based on coverage data from DNA capture).
4. DNA-based exon-level copy number variation analysis (CNV exon level): High-resolution detection of gains and losses at the exon level on a subset of targeted genes (based on coverage data from DNA capture).
5. Microsatellite instability analysis (MSI): Assessment of genomic instability through microsatellite status.
6. Tumor mutational burden analysis (TMB): Calculation of the mutational load across the genome.
7. Tumor content and allele-specific copy number (ASCN) analysis: DNA sample tumor content estimation and large-scale absolute copy number and loss of heterozygosity (LOH) profiling, across the whole genome (based on variant allele fraction information from DNA capture data and coverage information from DNA WGS data).
8. Genomic integrity analysis for HRD (HRD genomic integrity): Evaluation of homologous recombination deficiency through genomic scar signatures.
9. RNA-based fusion and exon-skipping detection (Gene fusions and exon skipping): Identification of gene fusions and exon skipping events from RNA data.
10. Gene expression analysis (Gene expression): Quantification of gene expression levels.

Sample-specific sequencing modality	Libraries sequenced				FASTQ files obtained				Bioinformatic analysis performed									
	DNA capture	DNA WGS	Combined DNA capture & DNA WGS	RNA capture	-D	-W	-DW	-R	SNV/INDEL	DNA gene fusions and exon skipping	CNV gene level	CNV exon level	MSI	TMB	Tumor content and ASCN	HRD genomic instability	RNA gene fusions and exon skipping	Gene expression
M1	●				●				●	●	●	●	●	●	*			
M2	●			●	●			●	●	●	●	●	●	●	*		●	●
M3			●	●			●	●	●	●	●	●	●	●	●	●	●	●
M4			●				●		●	●	●	●	●	●	●	●		
M5		●		●		●		●								●	●	●
M6				●				●									●	●
M7		●				●										●		

● = analysis available, * = analysis performed, but samples rejected.

The table above illustrates the set of bioinformatics analysis performed based on the sequencing modality.

The tumor content and ASCN analysis is initiated for all "-D" and "-DW" FASTQ files analyzed. However, a successful completion of this analysis requires both DNA capture and DNA WGS data (available in modalities M3 and M4). Samples containing exclusively DNA capture data (-D), such as those processed under sequencing modalities M1 and M2, do not include DNA WGS data, resulting in insufficient number of reads available for WGS coverage calculation. As a consequence, samples sequenced following modalities M1 and M2 are rejected from the tumor content and ASCN analysis, as indicated with an asterisk in the table above.

For each tumor tissue analyzed, analytical results obtained from the various NGS data types uploaded are unified and made available for interpretation and reporting via the SOPHiA DDM™ platform.

2 LIBRARY PREPARATION

This application relies on the SOPHiA GENETICS™ Universal Library Prep protocols for library preparation and sequencing.

For instructions on library preparation, refer to the appropriate Instructions for Use (IFU), which accompanies the application. The application is compatible with the following documents, depending on the library preparation setup of the lab (i.e., manual or automated on a liquid handling robot):

For dual DNA-RNA workflow:

- SG-07436 – Instructions for Use for SOPHiA GENETICS™ Universal Library Prep for dual DNA & RNA workflow (*manual*)

For DNA-only workflow:

- SG-06974 – Instructions for Use for SOPHiA GENETICS™ Universal Library Prep with CUMIN™ Adapters (*manual*)
- SG-07738 – Instructions for Use for SOPHiA GENETICS™ Universal Library Prep with CUMIN™ Adapters (*automated on Hamilton® NGS STAR*)
- SG-07806 – Instructions for Use for SOPHiA GENETICS™ Universal Library Prep with CUMIN™ Adapters (*automated on Hamilton® Clinical STARlet*)

For RNA-only workflow:

- SG-07014 – Instructions for Use for SOPHiA GENETICS™ Universal Library Prep for fragmented DNA – Tailored for RNA (*manual*)
- SG-07927 – Instructions for Use for SOPHiA GENETICS™ Hamilton® Clinical STARlet Automation Using the SOPHiA GENETICS™ Universal Library Prep for fragmented DNA (automation workflow does not cover cDNA generation that has to be performed manually according to SG-07014)
- SG-07669 – Instructions for Use for SOPHiA GENETICS™ Hamilton® NGS STAR Automation Using the SOPHiA GENETICS™ Universal Library Prep for fragmented DNA (automation workflow does not cover cDNA generation that has to be performed manually according to SG-07014)

The application-specific parameters to use with the Instructions for Use are as given in the table below

TECHNICAL PARAMETER	DNA (SG_IMPACTFLEX_v1)	RNA (RCROS_A_v1a)
Post-capture amplification PCR cycles	13	15

3 SEQUENCING

3.1 Multiplexing recommendations

MSK-IMPACT® Flex powered with SOPHiA DDM™ is a modular solution that allows users to analyze a set of tumor tissues using up to seven different sequencing modes, within the same sequencing run. The number of tumor tissues that can be multiplexed in a single run depends on the flow-cell capacity and the sequencing modalities selected for each individual tumor tissue.

Regardless of the sequencing modalities being executed, the numbers of reads per sample allocated to DNA capture, RNA capture, and DNA WGS sequencing are fixed, as shown in the following table.

Library type	Recommended total reads per sample (in million)	Recommended total read pairs (fragments) per sample (in million)
DNA capture	51	25.5
RNA capture	11	5.5
DNA WGS	20	10

The following table provides a non-exhaustive list of possible run compositions (multiplexing strategies) that the product allows to execute for different flow-cell capacities.

Flow cell capacity in million reads (million of pairs of reads)	Sequencing Modality	Number of tumor tissues included in the run	Number of libraries included in the sequencing run			Flow cell occupation (%)
			DNA Capture	RNA Capture	DNA WGS	
200M (100M)	DNA capture only	4	4			100%
266M (133M)	DNA capture only	4	4			77%
	DNA capture with matched RNA capture	4	4	4		93%
	DNA capture and DNA WGS with matched RNA capture	4	4	4	2	100%
	DNA capture and DNA WGS	4	4		4	100%

Flow cell capacity in million reads (million of pairs of reads)	Sequencing Modality	Number of tumor tissues included in the run	Number of libraries included in the sequencing run			Flow cell occupation (%)
			DNA Capture	RNA Capture	DNA WGS	
800M (400M)	DNA capture only	16	16			100%
		12	12			77%
		8	8			51%
	DNA capture with matched RNA capture	12	12	12		93%
		12	12	8		88%
		12	12	4		82%
		8	8	8		62%
		8	8	4		57%
	DNA capture and DNA WGS with matched RNA capture	12	12	12	3	100%
		12	12	8	5	100%
		12	12	4	7	100%
		8	8	8	8	82%
		8	8	4	8	77%
	DNA capture and DNA WGS	12	12		9	99%
		8	8		8	71%
2400M (1200M)	DNA capture only	48	48			100%
		32	32			68%
		24	24			51%
		16	16			33%
	DNA capture with matched RNA capture	32	32	32		83%
		24	24	24		62%
		16	16	16		41%
	DNA capture and DNA WGS with matched RNA capture	32	32	32	20	99%
		24	24	24	24	82%
		16	16	16	16	55%
		32	32		32	95%

Flow cell capacity in million reads (million of pairs of reads)	Sequencing Modality	Number of tumor tissues included in the run	Number of libraries included in the sequencing run			Flow cell occupation (%)
			DNA Capture	RNA Capture	DNA WGS	
	DNA capture and DNA WGS	24	24		24	71%
		16	16		16	47%

3.2 Protocol for combining DNA capture, RNA capture, and DNA WGS libraries on a single sequencing run

This section describes the protocol for combining DNA capture, RNA capture and DNA WGS libraries in a single loading pool for sequencing.

3.2.1 Materials

- Individual DNA WGS libraries
- DNA and/or RNA capture library pools
- IDTE
- DNA low-binding 1.5 ml tubes

3.2.2 Instructions

Prior to sequencing, DNA capture, RNA capture, and DNA WGS libraries must be pooled in a single tube. This can be achieved by performing a 3-step procedure.

Step 1: DNA capture pool and RNA capture pool preparation

If DNA capture libraries are to be included in the sequencing run, prepare a single 4 nM pool containing all DNA capture libraries as described in the SOPHiA GENETICS™ Universal Library Prep for dual DNA & RNA workflow Instructions for Use (SG-07436).

If RNA capture libraries are to be included in the sequencing run, prepare a single 4 nM pool containing all RNA capture libraries as described in the SOPHiA GENETICS™ Universal Library Prep for dual DNA & RNA workflow Instructions for Use (SG-07436).

Step 2: DNA WGS pool preparation

If one or more DNA WGS libraries are to be included in the sequencing run, prepare a 4 nM pool of the corresponding DNA WGS libraries, using the following procedure.

1. Determine the molarity of each library using the average size of the library (peak size in base pairs) and concentration (ng/µl) as follows:

$$\text{Library molarity (nM)} = \frac{\text{Library concentration (ng/µl)}}{\text{Average size in base pairs} \times 649.5} \times 10^6$$

2. Transfer 2 µl of each library individually into a new tube and dilute to 10 nM with IDTE.
3. Pool individual libraries at 10 nM by combining 5 µl from each library dilution.
4. Dilute the 10 nM pool of DNA WGS libraries to 4 nM pool to ensure it is at an equimolar ratio with DNA and/or RNA library pool.

Step 3: Mixing of DNA capture pool, RNA capture pool and DNA WGS pool

The mixing strategy outlined below aims at generating sequencing data with the right number of reads for the various sample types (DNA capture, RNA capture and DNA WGS).

The mixing volumes for the DNA capture pool (V_{DNA}), RNA capture pool (V_{RNA}), and DNA WGS pool (V_{WGS}) depend on the number of DNA capture samples (n_{DNA}), the number of RNA capture samples (n_{RNA}) and the number of DNA WGS samples (n_{WGS}). For preparing a 20 ul loading pool for a single sequencing run, the following formula applies:

$$V_{DNA} = 20\mu l \times \frac{n_{DNA} \times 25.5}{n_{DNA} \times 25.5 + n_{RNA} \times 5.5 + n_{WGS} \times 10.0}$$

$$V_{RNA} = 20\mu l \times \frac{n_{RNA} \times 5.5}{n_{DNA} \times 25.5 + n_{RNA} \times 5.5 + n_{WGS} \times 10.0}$$

$$V_{WGS} = 20\mu l \times \frac{n_{WGS} \times 10.0}{n_{DNA} \times 25.5 + n_{RNA} \times 5.5 + n_{WGS} \times 10.0}$$

The following table illustrates the mixing volumes (computed using the formulae above) to be used for a set of specific sequencing run compositions. For other run compositions, mixing volumes must be computed by the user using the formulae above (an Excel file that facilitates volume calculations involved in this step is made available in document SG-08192).

Number of libraries included in sequencing run			Mixing volumes to obtain a 20ul loading pool (ul)		
Number of DNA capture libraries (n_{DNA})	Number of RNA capture libraries (n_{RNA})	Number of DNA WGS libraries (n_{WGS})	DNA capture pool volume (V_{DNA})	RNA capture pool volume (V_{RNA})	DNA WGS pool volume (V_{WGS})
12			20.0		
12	12		16.45	3.55	
12	12	3	15.22	3.28	1.49
8	8	8	12.44	2.68	4.88
8		8	14.37		5.63

3.3 Instructions for sequencing on Illumina® platforms

The following table provides recommendations for paired-end sequencing read length.

SEQUENCER	READ LENGTH (BP)
Illumina® NextSeq® 500/550 **Illumina® NextSeq® 1000/2000** **Illumina® NovaSeq™ 6000** **Illumina® NovaSeq™ X**	2 x 150

The following table provides loading dilution recommendations for different sequencing instruments.

TYPE OF SEQUENCER	LOADING DILUTION
Illumina® NextSeq® 550	1.3 pM (Mid-Output Kit) 1.4 pM (High-Output Kit) Note: Adjust the dilution (1.1 pM to 1.5 pM range) according to the number of clusters obtained in the first run
Illumina® NextSeq® 2000	1000 pM (On-Board Denature/Dilute) 100 pM (Manual Denature/Dilute) *Note: Manual Denature/Dilute was not internally tested by SOPHiA GENETICS*
Illumina® NovaSeq™ 6000	300 pM
Illumina® NovaSeq™ X plus	90-180 pM

4 SAMPLE SHEET INSTRUCTIONS

For sequencing runs performed on Illumina® instruments, a sample sheet is required to enable correct processing of the data. In the context of this product, the sample sheet serves two purposes:

1. It provides the necessary information for demultiplexing raw sequencing reads
2. It is required for Command Line Interface (CLI) upload of FASTQ files to the SOPHiA DDM™ platform

The following sections describe how to prepare the sample sheet for each purpose.

4.1 Instructions relevant for demultiplexing

To perform demultiplexing, the sample sheet must contain the data section completed following the standard Illumina® recommendations for sample sheet preparation.

Sample names must follow a standardized structure to make sure the resulting FASTQ file names clearly indicate the type of NGS data. Each sample name is composed of two strings separated by a "-": <TumorTissueName>-<Postfix>, where:

- <TumorTissueName>: is a string that identifies the tumor tissue
- <Postfix>: is a string that specifies the NGS data type

<TumorTissueName> (e.g. TumorTissue123) should not contain an underscore sign "_", a point "." or any special characters (for example ? ( ) [ ] / \ = + < > : ; , ' " * ^ | &), but may contain dashes "-".

The possible postfixes are:

- "D": DNA capture FASTQ file
- "DW": Combined DNA capture and DNA WGS FASTQ file
- "W": DNA WGS FASTQ file
- "R": RNA capture FASTQ file

Paired DNA and RNA samples must share the same TissueTypeName and differ only by the relevant postfix (e.g., -D and -R). For example, if TumorTissue123 has been processed using the sequencing modality "DNA capture with matched RNA capture" the Data Section table of the sample sheet should contain two rows, with the following Sample Identifiers: TumorTissue123-D and TumorTissue123-R.

In the current version of the product, FASTQ files labeled with the "-DW" postfix are by default processed with GIInger™ Genomic Integrity analysis. To perform the tumor content and ASCN analysis using combined DNA capture and DNA WGS data without triggering the GIInger™ Genomic Integrity analysis, the corresponding combined FASTQ must instead be labeled with the "-D" postfix.

4.2 Instructions relevant for CLI-based FASTQ file upload

To upload data to the SOPHiA DDM™ platform via CLI, the sample sheet must include an additional section, starting with a dedicated header (e.g., [SOPHIA_DDM_Data_v1]). This additional table contains the information required by the SOPHiA DDM™ platform to handle the data upload and automatically trigger the bioinformatics pipeline.

The CLI documentation (https://platform-nl.sophiagenetics.com/uploader/cli/docs/, see "Sample Sheet upload workflow" section) provides instructions on how to include this additional section to the sample sheet. Additional instructions are provided in the DATA UPLOAD section.

The exact same sample names (e.g. TumorTissue123-D) must be used consistently in the sample sheet tables used for demultiplexing and CLI-based FASTQ file upload.

5 DEMULTIPLEXING INSTRUCTIONS

5.1 Procedure

To generate FASTQ files (demultiplexing) from a run on Illumina® NextSeq™ 1000/2000, Illumina® NovaSeq™ 6000, and Illumina® NovaSeq™ X, the following options exist:

1. Customers not using LINUX and the CLI have two options:
    a. Local Run Manager without Analysis Module.
    b. BCL convert (early access BaseSpace app) runs in the BaseSpace Hub and can be accessed at: https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/bcl-convert-early-access.html
2. Customers using LINUX have the following options:
    a. Illumina® BCL convert (standalone software), which runs on Linux and can be accessed at: https://support.illumina.com/sequencing/sequencing_software/bcl-convert/documentation.html?langsel=/us/
    b. Illumina® bcl2fastq v2.0 or higher (bcl2fastq2 Conversion Software v2.20 Software Guide, Document #15051736 v03)

Do not process unique molecular identifiers (UMIs) from the read sequences during demultiplexing, even if the application includes CUMIN™ adapters. The CUMIN™ adapter sequences are automatically processed during the bioinformatics pipeline execution.

The SOPHiA DDM™ Platform uses the Illumina® naming convention and folder configuration to find the FASTQ files. Therefore, it is strongly advised NOT to reorganize or rename the FASTQ files after they are copied from the sequencer, otherwise the run might fail.

Do not upload any 'Undetermined' FASTQ file that might have been generated.

5.2 FASTQ file naming Conventions

The SOPHiA DDM™ Platform follows the Illumina® FASTQ file naming conventions (e.g., **TumorTissue123-D** _S1_L001_R1_001.fastq.gz). The instructions provided in the SAMPLE SHEET INSTRUCTIONS section must be followed.

In addition, the following file specifications are required:

- Sample identifier name "**TumorTissue123**" should not contain an underscore sign "_", a point "." or any special characters.
- A minimum of two files (R1 and R2) per analysis are required to be present in the same folder
- More than two files per sample are acceptable (when reads from multiple lanes are kept separate), provided the files are always in pairs
- All FASTQ files should be in the same folder.

Do not rename FASTQ files originating from demultiplexing. Make sure that sample names are defined as explained in the SAMPLE SHEET INSTRUCTIONS section.

6 DATA UPLOAD

6.1 Instructions for manual upload

This section provides instructions on how to manually upload FASTQ files to the SOPHiA DDM™ platform. For detailed instructions, please visit the Help Center at https://platform.sophiagenetics.com/web/landing/help. Locate the search bar in the top-left corner and enter the keyword *upload settings* to find the correct guide.

Ensure the FASTQ files follow the naming conventions detailed in Sections 4.1 and 5.2 prior to upload.

CNV analysis requires a minimum of eight DNA samples (either DNA capture or combined DNA capture and DNA WGS) to be included in the same analysis request batch. If this requirement is not fulfilled, CNV analysis will not be performed.

6.2 Instructions for upload via Command Line Interface (CLI)

This section provides instructions on how to use the CLI tools to upload FASTQ files to the SOPHiA DDM™ platform. **For installation and usage details, refer to the documentation**: https://platform-nl.sophiagenetics.com/uploader/cli/docs/

The CLI tool allows upload of FASTQ files in one or multiple analysis request batches (i.e. runs), depending on the total volume (in Gigabytes) of data to be analysed:

- If the total size of the FASTQ files to be uploaded and analysed does not exceed the maximal data volume allowed for a single-batch upload, we recommend uploading all data in a single analysis request batch.
- If the total size of the FASTQ files to be uploaded and analysed does exceed the maximal data volume allowed for a single-batch upload, FASTQ files must be grouped and uploaded using two or more analysis request batches.

Please refer to the CLI documentation https://platform-nl.sophiagenetics.com/uploader/cli/docs/#uploadlimit for more information on the maximum allowed data volume for a single upload batch.

6.2.1 Procedure

Prerequisites

Before proceeding with FASTQ files upload:

- Prepare the sample sheet file following the instructions provided in the SAMPLE SHEET INSTRUCTIONS section.
- Make sure that all FASTQ files to be uploaded respect the FASTQ file naming conventions (see the DEMULTIPLEXING INSTRUCTIONS section) and are stored in a dedicated folder.

Step 1: Determine whether the upload can be performed using one or multiple analysis request batches

Measure the total size of the FASTQ files to be uploaded and determine if it does exceed the maximal data volume allowed for a single analysis request batch.

Step 2: Edit the sample sheet to group samples in one or multiple analysis request batch(es)

In the [SOPHIA_DDM_Data_v1] section of the Sample Sheet file, fill in the "Upload_Batch" column to assign each sample to an analysis request batch:

- Single batch: If all samples belong to the same analysis request batch, enter the same numeric value (e.g., "1") for all samples.
- Multiple batches: If samples must be uploaded in two or more batches, assign a distinct numeric value to each batch (e.g., use "1" for all samples in the first batch, "2" for all samples in the second batch, and so on).

More details are provided in the Sample Sheet section of the CLI tool documentation (https://platform-nl.sophiagenetics.com/uploader/cli/docs/#samplesheet).

Individual pairs of DNA and matched RNA FASTQ files must be included in the same analysis request batch. For pairs not respecting this condition, DNA and RNA FASTQ files will be processed and visualized in the SOPHiA DDM™ platform as independent analyses.

CNV analysis requires a minimum of eight DNA samples (either DNA capture or combined DNA capture and DNA WGS) to be included in the same analysis request batch. If this requirement is not fulfilled, CNV analysis will not be performed.

The CLI tool will raise an error if the total size of FASTQ files included in one analysis request batch exceeds the maximal volume allowed.

The following table provides an indicative size for individual FASTQ file pairs (R1+R2) for the four NGS data types possibly obtained using this application (following the sequencing recommendations provided in this application manual).

NGS data type contained in FASTQ file	Typical FASTQ file size (R1+R2)[1]
DNA capture FASTQ file (-D)	3.5 Gigabytes
DNA WGS FASTQ file (-W)	1.4 Gigabytes
Combined DNA capture and DNA WGS FASTQ file (-DW)	4.7 Gigabytes
RNA capture FASTQ file (-R)	0.7 Gigabytes

[1] These values are only indicative and will vary in practice among samples

Step 3: Run the CLI tool command to trigger the upload

Start the upload using the following CLI tool command:

```
python3 sg-upload-v2-wrapper.py new \
     --folder <FASTQ_location> \
     --sampleSheet <sampleSheet_location> \
     --pipeline=<Pipeline_ID> \
     --upload
```

The <Pipeline_ID> parameter determines which bioinformatics pipeline will analyze the NGS data being uploaded. The following table indicates the <Pipeline_ID> parameter to use for data generated with different sequencing platforms.

Sequencing platform used for NGS data generation	<Pipeline_ID> value to use for data upload
Illumina® NextSeq™ 500/550	8689
Illumina® NextSeq™ 1000/2000	8688
Illumina® NovaSeq™ 6000	8656
Illumina® NovaSeq™ X	8376

Note: if your sequencing run includes samples from different assays, it is possible to set a different pipeline ID for each sample directly in the sample sheet in the column named "Pipeline_ID". In this case the --pipeline argument for the CLI upload command must not be used.

More details are provided in the Sample Sheet section of the CLI tool documentation (https://platform-nl.sophiagenetics.com/uploader/cli/docs/#samplesheet).

After completion of the analysis, users will receive a notification by email. If a notification is not received within 24 h from the initiation of the data upload process, please contact the SOPHiA GENETICS support team (see SUPPORT section).

7 RESULT VISUALIZATION AND REPORT GENERATION

The results of the bioinformatic analyses are accessed and visualized via the SOPHiA DDM™ Platform.

Instructions for the use of the platform are provided in dedicated user manuals:

- SOPHiA DDM™ Operation Manual for SOPHiA DDM™ desktop.
- SOPHiA DDM™ Web User Manual for SOPHiA DDM™ web (available in the help menu).
- OncoPortal™ Plus User Manual (available in the help menu in OncoPortal™ Plus).

Output files can be accessed and downloaded from the SOPHiA DDM™ Platform. Output files can be accessed per Request (Run) or per Sample/Group analysis, as follows:

- Request/Run level: all FASTQ files, the QA report, the region_map.tsv, and other files depending on the application.
- Sample/Analysis level: sample-specific files such as alignment bam files, target region coverage statistics, flagged regions, the full variant and fusion tables, sample-specific QA reports, etc.

The SOPHiA DDM™ Platform offers the option of creating a report document (variant report) for each interpretation project, with relevant information about the case and the application, reported variants, interpretive text, and the analyst's conclusions. The report can be downloaded from the platform, and the final version of the report is stored.

Refer to the SOPHiA DDM™ Platform User Manual for instructions on how to generate reports using the interpretation project workflow.

8 TECHNICAL DESCRIPTION OF THE BIOINFORMATIC ANALYSIS

8.1 DNA analysis: data pre-processing and QA

8.1.1 Analysis purpose

This module preprocesses FASTQ files by detecting and trimming CUMIN™ adapters, then align the sequencing reads to the reference genome and generates NGS data quality metrics.

8.1.2 Technical overview of the analysis

The DNA data pre-processing workflow consists of the following steps:

Step 1: Global down-sampling. FASTQ files are down-sampled to a maximum file size of approximately 5.0 GB per file, if applicable.
Step 2: CUMIN™ assignment and trimming. The CUMIN™ adapters used for each sample are identified based on a subset of reads. For each read, the CUMIN™ sequence is decoded, labelled, and trimmed from the read.
Step 3: Alignment. Reads are aligned to the human reference genome (hg19/GRCh37) in paired-end mode.
Step 4: Adaptor trimming. 3' overhanging adapter sequences extending beyond the 5' start position of the mate are trimmed to remove adapter-derived bases.
Step 5: Local down-sampling. Local down-sampling is performed with read coverage cutoffs of 30,000x on extended targets (targets ± 5,000 bp).
Step 6: Read group assignment. Reads are grouped into CUMIN™ groups based on their start–end mapping coordinates and decoded CUMIN™ sequences.
Step 7: Realignment of soft clips. Soft-clipped reads are re-aligned to more accurately represent insertions and deletions, allowing a maximum realignment distance of 2,000 bp.
Step 8: Low coverage regions. Targeted low-coverage regions are extracted based on a threshold of 200 CUMIN™ groups (~molecular coverage).

During these preprocessing steps, different QA metrics are collected.

8.1.3 Description of the results

The key outputs of this step include:

- An aligned and indexed BAM file.
- A PDF QA report summarizing NGS data quality metrics, as well as a set of text files containing the corresponding data.

The pre-processing module also computes summary metrics which allow the user to identify targeted regions with insufficient molecular coverage:

- Low-coverage warnings: a list of genomic intervals within the DNA target region not reaching at least 200x molecular coverage.
- Gene-level molecular coverage statistics: a table providing, for each individual gene, the fraction of targeted regions reaching a given molecular coverage.

BAM files and PDF QA reports are available as downloadable files. Low coverage warnings and gene-level molecular coverage statistics are available via the SOPHiA DDM™ platform (Quality tab) as well as downloadable files.

The following table provides an overview of the downloadable output files produced by this module.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
bwa.bam **bwa.bam.bai**	Sample	Aligned and indexed BAM file.
QA-Report.pdf	Run	NGS data QA report for all samples included in the run.
QA-patient.pdf	Sample	NGS data QA report for individual samples.
input_file_size_table.csv	Run	Text file containing raw data displayed in NGS QA report: Size of FASTQ files uploaded and global down-sampling applied.
read-counts-overview-table.csv	Sample and Run	Text file containing raw data displayed in NGS QA report: Total read count in FASTQ and mapping statistics.
ontarget-mapping-statistics-table-<GENE PANEL NAME >.csv	Sample and Run	Text file containing raw data displayed in NGS QA report: Percentage of reads mapped to the DNA target region.
target-region-coverage-table-< GENE PANEL NAME >.csv	Sample and Run	Text file containing raw data displayed in NGS QA report: molecular coverage statistics in DNA target region.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
pcr-duplicates-table-<GENE PANEL NAME>.csv	Sample and Run	Text file containing raw data displayed in NGS QA report: PCR duplicates statistics measured in the DNA target region.
soft-clip-percentage-table.csv	Sample and Run	Text file containing raw data displayed in NGS QA report: Percentage of reads carrying soft-clipped sequences.
flagged_regions.txt	Sample	Text file containing raw data associated to low-coverage warnings.
exon_coverage_stats_v3.txt **Exon_coverage_stats.txt**	Sample and Run	Text file containing raw data used compute gene-level molecular coverage statistics

8.2 RNA analysis: data pre-processing and QA

8.2.1 Analysis purpose

This module preprocesses FASTQ files by detecting and trimming CUMIN™ adapters, aligning reads to the reference genome, and generating quality metrics for the processed data, providing the foundation for downstream analyses.

8.2.2 Technical overview of the analysis

The RNA data pre-processing workflow consists of the following steps:

- **Step 1: Global down-sampling.** FASTQ files are down-sampled to a maximum file size of approximately 1.0 GB per file, if applicable.
- **Step 2: CUMIN™ assignment and trimming.** The CUMIN™ adapters used for each sample are identified based on a subset of reads. For each read, the CUMIN™ sequence is decoded, labelled, and trimmed from the read.
- **Step 3: Alignment.** Reads are aligned to the human reference genome (hg19/GRCh37) using STAR (version 2.70f_0328) in both paired-end and single-end modes. This step produces three alignment files, which serve different purposes:
    - **star.bam:** STAR alignment output containing paired-end reads that align to single genomic locations or across exon–exon junctions in a linearly spliced manner. Exon-skipping reads and gene fusion reads may be included.
    - **star_SE.bam**: STAR alignment output containing single-end reads that align to single genomic locations or across exon–exon junctions, including both canonical splicing and exon-skipping events.
    - **STAR_SE.Chimeric.out.bam**: STAR alignment output containing reads that map to two or more distinct genomic locations not consistent with normal splicing, indicative of potential gene fusion events.
- **Step 4: Read group assignment.** Reads are grouped into CUMIN™ groups based on their start–end mapping coordinates and decoded CUMIN™ sequences.

During these pre-processing steps, different QA metrics are collected.

8.2.3 Description of the results

The key outputs of this step include:

- Indexed BAM files produced via STAR alignment.
- PDF QA report summarizing NGS data quality metrics, as well as a set of text files containing the corresponding data.

BAM files and PDF QA reports are available as downloadable files.

The following table provides an overview of the downloadable output files produced by this module.

FILE NAME	RUN-vs.SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
QA-Report.pdf	Run	QA report : DNA and RNA merged QA report
QA-patient.pdf	Sample	QA report : RNA QA report
star.bam **star.bam.bai**	Sample	STAR alignment output containing paired-end reads that align to single genomic locations or across exon–exon junctions in a linearly spliced manner. Exon-skipping reads and gene fusion reads may be included.
star_SE.bam **star_SE.bam.bai**	Sample	STAR alignment output containing single-end reads that align to single genomic locations or across exon–exon junctions, including both canonical splicing and exon-skipping events.
STAR_SE.Chimeric.out.bam **STAR_SE.Chimeric.out.bam.bai**	Sample	STAR alignment output containing reads that map to two or more distinct genomic locations not consistent with normal splicing. These reads are indicative of potential gene fusion events.

8.3 DNA analysis: SNV and INDEL detection and annotation via PEPPER™ and MOKA™

8.3.1 Analysis purpose

The SNV & INDEL analysis module is designed to detect with high sensitivity short variants present in a sample with an underlying variant allele fraction (VAF) of 5% (the cutoff for reporting variants is VAF > 1%). Detected variants are reported relative to the hg19/GRCh37 reference genome (used for variant calling), as well as with respect to hg38/GRCh38.

Each variant is annotated to provide transcript-specific descriptions. Gene and transcript-level identifiers are also included.

The genomic regions in scope for SNV & INDEL analysis span 1.4 Mb and include 533 genes. For 531 genes, the entire coding sequences (CDS) are covered, while for 2 genes (ARID1B, ASXL2), CDS regions are not fully covered. In addition to coding sequences, the SNV & INDEL analysis also covers selected clinically relevant non-coding regions. A detailed description of the targeted regions is provided in the section Description of the genomic regions in scope of SNV and INDEL analysis.

8.3.2 Technical overview of the analysis

SNV & INDEL analysis is performed after DNA data preprocessing, and uses the BAM file as a primary input. All the information included in the BAM file is leveraged for SNV & INDEL calling, except for:

- Sequencing base calls from reads with mapping quality <20.
- Non-redundant sequencing base calls: base calls from a CUMIN™ group when no other reads in the group cover the same genomic position.

SNV & INDEL calling and annotation is performed in four main steps.

Step 1: Identification of putative variants

A pileup of the BAM file is performed. The information of alternative base calls is used to obtain a preliminary list of putative variants. Neighboring pileup locations (within a 50bp window) carrying alternative base calls (separated by at most 3bp) are treated as single putative complex variants (i.e. DELINs).

For each putative variant, a set of metrics are collected and used to establish the statistical significance (i.e., a confidence score) of the signal. The preliminary list of putative variants is restricted by retaining only:

- Variants whose presence is supported by at least 1% of the reads.

- Variants whose presence is supported by reads coming from at least 2 different CUMIN™ groups.
- Variants whose signal is deemed of sufficient statistical significance (confidence score > -100).

Step 2: VAF quantification

For each putative variant, VAF is computed using a probabilistic method that considers the number of reads supporting the variant, as well as the CUMIN™ group read-assignment information.

Step 3: Variant calling and filtering

The list of putative variants is further restricted by excluding:

- Variants beyond target regions with a padding of 500 bp.
- Duplications longer than 500 bp.

The remaining variants are classified as high-confidence and low-confidence variants based on the following filters:

- **off_target:** Variants located in genomic regions that do not overlap the SNV & INDEL target region.
- **problematic_region:** Variants located in genomic regions where SNV and INDEL calling may be unreliable, for example due to the presence of homologous sequences, low-complexity regions, repetitive elements, or extreme GC content.
- **homopolymer_region:** Variants located within a homopolymer region of length >7 bp.
- **low_coverage:** Variants in genomic regions with total read coverage <50.
- **low_variant_fraction:** Variants with VAF < 2%.
- **low_alt_molecule:** Variants supported by <5 CUMIN™ groups.
- **high_background_noise:** Variants with VAF lower than a predefined genomic-position specific background noise level (precomputed based on NGS data obtained from a representative set of DNA samples).
- **strand_bias:** Variants with strand Odds Ratio (SOR) exceeding 4 (this filter only applies to SNVs with VAF < 50%).
- **low_molecular_support:** This filter may enforce a more stringent requirement on the minimum number of supporting CUMIN™ groups depending on the deamination level found in the sample.
- **low_quality:** Variants associated to signal not reaching a sufficient level of statistical significance (confidence score < -25).

Each variant is evaluated against all defined filters. The variant is classified as **high confidence** only when no filter is triggered.

Step 4: Tertiary annotation

For each variant identified, this step computes transcript-specific annotations according to HGVS coordinate normalization and notation guidelines (cDNA and protein notation). It provides functional information on the variant's coding consequence, as well as positional and contextual details, including exon rank, distance to the nearest exon, reference and alternate codon sequences, and reference and alternate amino acid sequences. Transcript-level information (RefSeq identifiers) and gene-level information (HGNC symbols and OMIM gene numbers) are also included.

Next, up to 27 external databases are queried via genomic coordinates matches to retrieve variant-level information. This includes dbSNP identifiers, allele frequencies from gnomAD (genomes, exomes, and structural variants), the 1000 Genomes Project, ExAC, and ESP5400; prediction scores from dbNSFP (SIFT, PolyPhen2, MutationTaster), REVEL, and dbscSNV; and clinical significance assertions from ClinVar and ClinGen.

Finally, variants are annotated with licensed databases, ensuring continuous access to high-quality, curated data. Licensed resources include COSMIC, Genomenon, Genomenon-CKB, OMIM, BRCA Exchange, PolyPhen2, CADD, DECIPHER, and dbNSFP.

8.3.3 Description of the results

The SNV & INDEL analysis outputs a list of tertiary-annotated short variants detected in the DNA sample. The results are made available via the SOPHiA DDM™ platform as well as downloadable files.

The SOPHiA DDM™ platform Help Center includes a description of the various fields displayed in the variant table. The Description of fields included in the SNV and INDEL – full variant table section provides a description of the variant attributes present in the downloadable variant table file.

The following table provides an overview of the downloadable output files produced by this module.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
full_variant_table.txt	Sample and Run	List of SNV/INDEL calls with extended annotation in text format
full_variant_table.vcf	Sample	List of SNV/INDEL calls with annotation in VCF format

8.3.4 Precautions and limitations

- High sensitivity in SNV and INDEL analysis requires a molecular coverage of 200x. Low coverage regions significantly increase the risk of false negatives and are reported with warnings.
- Some genomic regions within the scope of SNV and INDEL analysis are designated as *problematic regions*. Variants detected in these regions are systematically classified as **low confidence**. The following genes are considered as problematic:
    - **HLA-A**, **HLA-B**, **HLA-C** (due their high polymorphism)
    - **H3C14 and H3C13** (due to high sequence homology)

  In addition, the following genes contain exons that are considered problematic due to high sequence homology:
    - **EIF1AX (NM_001412, exon 1)**
    - **FANCD2 (NM_001018115, exons 14–17, 21–23, 25, 28)**
    - **KMT2C (NM_170606, exons 7–8, 14–21, 24, 34)**
    - **MST1 (NM_020998, exons 1–2, 4, 6–10, 12–18)**
    - **NOTCH2 (NM_024408, exons 1–4)**
    - **PDPK1 (NM_002613, exons 3–6, 8–10)**
    - **PMS2 (NM_000535, exons 9, 11–15)**
    - **SLFN11 (NM_001104589, exon 5)**
    - **STK19 (NM_004197, exon 7)**
    - **ZRSR2 (NM_005089, exon 8)**
- All genomic regions located within homopolymers of length **>7** are systematically annotated as *homopolymer regions*. Variants detected in these regions are systematically classified as **low confidence**.
- Local realignment allows detection of deletions up to **1,000 bp**. Deletions exceeding this length are outside the SNV & INDEL analysis scope. The sensitivity for deletions as a function of their size has not been systematically evaluated.
- Insertions or deletions that extend beyond the targeted regions may not be detected if one or both breakpoints lie outside the assay target design. As a result, events partially overlapping the target boundaries may not be reliably identified.
- Variants that create alleles highly divergent from the reference genome (e.g., multiple SNVs, multinucleotide variants, or large INDELs) may reduce the hybridization efficiency of the capture probes for the affected regions. Such alleles may be inefficiently captured, potentially resulting in **false negative** results or **underestimation of variant allele frequency**.
- Genomic regions with low-complexity nucleotide sequences, nucleotide bias, repeats of any length (e.g. mono-, di- or trinucleotide repeats, transposable elements, Alu repeats, etc.) and high sequence homology are subject to an increased risk of false positive and false negative results.

- In complex genomic regions, alignment ambiguities may lead to *split calls*, in which equivalent representations of the same variant are reported as distinct events. This may result in **underestimated variant allele frequencies** or **false negative** calls.

8.3.5 Description of fields included in the SNV and INDEL – full variant table

FIELD CATEGORY	FIELD CATEGORY	FIELD DESCRIPTION
Genomic description (hg19)	type	Variant type (SNP/INDEL)
	refGenome	Reference genome (hg19) used for variant calling
	chromosome	Chromosome
	genome_position	Genomic position
	first1	Normalized genomic position (3' alignment, in direction of strand)
	last1	Normalized genomic position (5' alignment, in direction of strand)
	ref	Reference base (in reference genome)
	alt	Alternative base (variant)
	ref1	Normalized reference allele (3' alignment, in direction of strand)
	alt1	Normalized alternative allele (3' alignment, in direction of strand)
Gene information	gene	HGNC gene symbol where the variant is located
	OMIM	OMIM gene ID
	gene_strand	DNA strand of the gene where the variant is located
	gene_boundaries	Indicates if the variant falls within a gene's region or outside
	gene_role	Gene role in cancer context (Oncogene or Tumor Suppressor Gene)
Genomic description (hg38)	hg38_refGenome	Alternative reference genome used to provide alternative variant genomic coordinates
	lift_diagnostic	Indicates the tool used for variant lifting between reference genomes

FIELD CATEGORY	FIELD CATEGORY	FIELD DESCRIPTION
	hg38_chrom	Chromosome (in alternative reference genome)
	hg38_pos	Genomic position (in alternative reference genome)
	hg38_ref	Reference base (in alternative reference genome)
	hg38_alt	Alternative base (variant, for alternative reference genome hg38)
Warning and Quality	filter	Indicates whether a variant has passed or failed internal quality control filters (the list of filters potentially applied to a variant is provided in the section "Technical overview of the analysis")
	flagged_region_id	Link the variant to a specific warning region from the warnings tab
	matchStatus	Match status of the variant to external catalog information (Exact match vs Partial match)
Signal Quantification	var_percent	Variant fraction computed based on molecular coverage (CUMIN™ groups)
	depth	Read coverage depth
	depth_uniq	Molecular coverage depth (CUMIN™ groups)
	refNum	Number of reads supporting the reference allele
	refNum_uniq	Number of molecules (CUMIN™ groups) supporting the reference allele
	altNum	Number of reads supporting the alternative allele
	altNum_uniq	Number of molecules (CUMIN™ groups) supporting the alternative allele
Transcript Annotation	tx_id	Transcript ID in annotation database
	tx_name	Transcript symbol in annotation database
	tx_version	Transcript version in annotation database
	refSeqId	Transcript symbol in RefSeq database
	refSeqIdVersion	Transcript version in RefSeq database

FIELD CATEGORY	FIELD CATEGORY	FIELD DESCRIPTION
	refSeq	Reference codon
	altSeq	Alternative codon
	refAA	Reference amino acid
	altAA	Alternative amino acid
	c.DNA	c.DNA notation of the variant
	codingConse-quence	Predicted protein coding consequence
	exon_id	Exon identifier (legacy system)
	exon_rank	Exon identifier
	cds_rank	Rank of the CDS
	pos_in_exon	Variant position in exon (strand specific)
	dist2exon	Distance to closest exon
	dist2cds	Distance to closest CDS
	protein	p. notation of the variant
HGVS nomenclature	HGVS_gnomen	HGVS genomic (g.) description (variant description relative to the genomic reference sequence)
	HGVS_cnomen	HGVS coding DNA (c.) description (variant description relative to the coding DNA sequence of the transcript)
	HGVS_pnomen	HGVS protein (p.) description (variant description relative to the amino acid sequence of the protein)
Population frequency databases	dbSNP	dbSNP rsID
	g1000	Allele frequency (from 1000 Genomes Project database)
	GnomAD	Allele frequency (from Genome Aggregation Database database)
	ExAC	Allele frequency (from Exome Aggregation Consortium database)
	esp5400	Allele frequency (from NHLBI Exome Sequencing Project database)
Variant classification databases	overlapKnown	rsid of pathogenic Clinvar entries
	id_clinvar	Identifier of the variant according to ClinVar

FIELD CATEGORY	FIELD CATEGORY	FIELD DESCRIPTION
	CLNSIG	Variant pathogenicity according to ClinVar
	CLNREVSTAT	Supporting evidence for ClinVar's pathogenicity assessment
	id_cosmic_coding	COSMIC ID from COSMIC coding variants catalog
	id_cosmic_non_coding	COSMIC ID from COSMIC non-coding variants catalog
	BRCA_Pathogenicity	Variant pathogenicity according to BRCA Exchange
	uniprot_acc	UniProt accession number
In-Silico Predictors	LJB_PhyloP	dbNSFP's precomputed PhyloP score
	LJB_LRT	dbNSFP's precomputed LRT score which has been normalised
	LJB_GERP	dbNSFP's precomputed GERP++_RS score
	LJB_SIFT	dbNSFP's precomputed SIFT score which has been normalised
	LJB_PolyPhen2	dbNSFP's precomputed PolyPhen2 score which has been normalised
	LJB_PolyPhen2_HumDiv	dbNSFP's precomputed PolyPhen2_ HumDiv score which has been normalised
	LJB_MutationTaster	dbNSFP's precomputed MutationTaster score which has been normalised
Internal Reference	id	Internal variant ID
	annotation_id	Variant ID in annotation database
	sgid	Variant ID in annotation database (alternative representation of annotation_id)
	hg38_sgid	Variant ID in internal annotation database (hg38)
	multiTranscriptId	TranscriptId of closest transcript within 10Kbp

8.3.6 Description of the genomic regions in scope of SNV and INDEL analysis

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
ABL1	NM_005157	3503	3503	100	3649	146	
ABRAXAS1	NM_139076	1320	1320	100	1320	0	
ACVR1	NM_001111067	1620	1620	100	1620	0	
AGO1	NM_012199	2764	2764	100	2764	0	
AGO2	NM_012154	2770	2770	100	2770	0	
AKT1	NM_001014431	1573	1573	100	1573	0	
AKT2	NM_001626	1576	1576	100	1576	0	
AKT3	NM_005465	1570	1570	100	1624	54	
ALB	NM_000477	1970	1970	100	1970	0	
ALK	NM_004304	5153	5153	100	5153	0	
ALOX12B	NM_001139	2256	2256	100	2256	0	
AMER1	NM_152424	3418	3418	100	3418	0	
ANKRD11	NM_013275	8102	8102	100	8102	0	
APC	NM_000038	8682	8682	100	9457	775	
APLNR	NM_005161	1153	1153	100	1153	0	
AR	NM_000044	2843	2842	99.96	3004	162	AR exon3 covered with a -4bp/+5bp padding
ARAF	NM_001654	1971	1971	100	1984	13	
ARHGAP35	NM_004491	4560	4560	100	4560	0	
ARID1A	NM_006015	7058	7058	100	7058	0	
ARID1B	NM_001374820	7199	6945	96.47	7114	169	85.9% of ARID1B exon1 6:157098809-157100610 is covered
ARID2	NM_152641	5718	5718	100	5722	4	
ARID5B	NM_032199	3667	3667	100	3681	14	
ASS1	NM_054012	1379	1379	100	1379	0	
ASXL1	NM_015338	4756	4756	100	4772	16	
ASXL2	NM_018263	4438	4425	99.71	4425	0	ASXL2 exon3 of size 3bp

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
							(2:26058407-26058409) is not covered
ATM	NM_000051	9791	9791	100	10399	608	
ATR	NM_001184	8405	8405	100	8405	0	
ATRX	NM_000489	7829	7829	100	7829	0	
ATXN7	NM_001177387	2958	2958	100	3032	74	
AURKA	NM_003600	1292	1292	100	1297	5	
AURKB	NM_004217	1115	1115	100	1174	59	
AXIN1	NM_003502	2689	2689	100	2689	0	
AXIN2	NM_004655	2632	2632	100	2632	0	
AXL	NM_021913	2885	2885	100	2885	0	
B2M	NM_004048	390	390	100	390	0	
BABAM1	NM_001033549	1070	1070	100	1070	0	
BAP1	NM_004656	2360	2360	100	2360	0	
BARD1	NM_000465	2444	2444	100	2444	0	
BBC3	NM_001127240	826	826	100	826	0	
BCL10	NM_003921	732	732	100	732	0	
BCL2	NM_000633	740	740	100	773	33	
BCL2L1	NM_138578	722	722	100	722	0	
BCL2L11	NM_138621	627	627	100	1099	472	
BCL6	NM_001706	2201	2201	100	2201	0	
BCOR	NM_001123385	5408	5408	100	5408	0	
BCORL1	NM_001379451	5488	5488	100	5488	0	
BIRC3	NM_182962	1895	1895	100	1895	0	
BLM	NM_000057	4464	4464	100	4464	0	
BMPR1A	NM_004329	1709	1709	100	1709	0	
BRAF	NM_004333	2481	2481	100	2620	139	
BRCA1	NM_007294	5812	5812	100	6038	226	
BRCA2	NM_000059	10517	10517	100	10517	0	
BRD4	NM_058243	4279	4279	100	4537	258	
BRIP1	NM_032043	3940	3940	100	4090	150	
BTK	NM_000061	2160	2160	100	2272	112	
CALR	NM_004343	1344	1344	100	1344	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
CARD11	NM_032415	3705	3705	100	3705	0	
CARM1	NM_199141	1987	1987	100	1987	0	
CASP8	NM_001080125	1707	1707	100	1813	106	
CBFB	NM_022845	624	624	100	655	31	
CBL	NM_005188	2881	2881	100	2881	0	
CCND1	NM_053056	938	938	100	938	0	
CCND2	NM_001759	920	920	100	920	0	
CCND3	NM_001760	929	929	100	987	58	
CCNE1	NM_001238	1343	1343	100	1343	0	
CCNQ	NM_152274	795	795	100	795	0	
CD274	NM_014143	933	933	100	933	0	
CD276	NM_001024736	1695	1695	100	1695	0	
CD58	NM_001779	813	813	100	817	4	
CD79A	NM_001783	731	731	100	731	0	
CD79B	NM_001039933	753	753	100	753	0	
CDC42	NM_001791	626	626	100	726	100	
CDC73	NM_024529	1766	1766	100	1766	0	
CDH1	NM_004360	2809	2809	100	2809	0	
CDK12	NM_016507	4613	4613	100	4613	0	
CDK4	NM_000075	982	982	100	982	0	
CDK6	NM_001145306	1051	1051	100	1051	0	
CDK8	NM_001260	1525	1525	100	1525	0	
CDKN1A	NM_078467	515	515	100	627	112	
CDKN1B	NM_004064	617	617	100	617	0	
CDKN2A	NM_058195 NM_000077	419 501	419 501	100 100	1366 1366	947 865	
CDKN2B	NM_004936	437	437	100	518	81	
CDKN2C	NM_078626	527	527	100	527	0	
CEBPA	NM_004364	1087	1087	100	1192	105	
CENPA	NM_001809	463	463	100	463	0	
CHEK1	NM_001274	1551	1551	100	1551	0	
CHEK2	NM_007194	1772	1772	100	1911	139	
CIC	NM_015125	5027	5027	100	5027	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
CMTR2	NM_001099642	2323	2323	100	2323	0	
COP1	NM_022457	2396	2396	100	2396	0	
CREBBP	NM_004380	7639	7639	100	7639	0	
CRKL	NM_005207	942	942	100	942	0	
CRLF2	NM_022148	852	852	100	852	0	
CSDE1	NM_001242891	2725	2725	100	2725	0	
CSF1R	NM_005211	3129	3129	100	3129	0	
CSF3R	NM_000760	2661	2661	100	2857	196	
CTCF	NM_006565	2284	2284	100	2284	0	
CTLA4	NM_005214	712	712	100	712	0	
CTNNB1	NM_001904	2486	2486	100	2486	0	
CTR9	NM_014633	3772	3772	100	3772	0	
CUL3	NM_003590	2467	2467	100	2561	94	
CXCR4	NM_003467	1079	1079	100	1106	27	
CYLD	NM_001042355	3022	3022	100	3022	0	
CYP19A1	NM_000103	1602	1602	100	1602	0	
CYSLTR2	NM_020377	1051	1051	100	1051	0	
DAXX	NM_001141970	2339	2339	100	2356	17	
DCUN1D1	NM_020640	850	850	100	850	0	
DDR1	NM_001297654	2912	2912	100	2994	82	
DDR2	NM_006182	2728	2728	100	2728	0	
DICER1	NM_177438	6029	6029	100	6271	242	
DIS3	NM_014953	3087	3087	100	3165	78	
DNAJB1	NM_006145	1053	1053	100	1053	0	
DNMT1	NM_001379	5251	5251	100	5318	67	
DNMT3A	NM_022552	2959	2959	100	3104	145	
DNMT3B	NM_006892	2782	2782	100	2828	46	
DOT1L	NM_032482	4894	4894	100	5515	621	
DPYD	NM_000110	3308	3308	100	3357	49	
DROSHA	NM_013235	4455	4455	100	4455	0	
DUSP4	NM_001394	1225	1225	100	1395	170	
E2F3	NM_001949	1468	1468	100	1496	28	
EED	NM_003797	1446	1446	100	1531	85	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
EGFL7	NM_201446	902	902	100	902	0	
EGFR	NM_005228	3913	3913	100	4479	566	
EIF1AX	NM_001412	505	505	100	505	0	EIF1AX exon1 fully encompassed by problematic region due to high homology
EIF4A2	NM_001967	1334	1334	100	1334	0	
EIF4E	NM_001130678	784	784	100	915	131	
ELF3	NM_004433	1196	1196	100	1196	0	
ELOC	NM_005648	369	369	100	369	0	
EP300	NM_001429	7555	7555	100	7555	0	
EPAS1	NM_001430	2773	2773	100	2773	0	
EPCAM	NM_002354	1035	1035	100	1035	0	
EPHA3	NM_005233	3122	3122	100	3148	26	
EPHA5	NM_004439	3294	3294	100	3301	7	
EPHA7	NM_004440	3167	3167	100	3175	8	
EPHB1	NM_004441	3115	3115	100	3115	0	
ERBB2	NM_004448	4038	4038	100	4206	168	
ERBB3	NM_001982	4309	4309	100	4440	131	
ERBB4	NM_005235	4207	4207	100	4207	0	
ERCC2	NM_000400	2513	2513	100	2576	63	
ERCC3	NM_000122	2499	2499	100	2499	0	
ERCC4	NM_005236	2861	2861	100	2861	0	
ERCC5	NM_000123	3711	3711	100	3711	0	
ERF	NM_006494	1687	1687	100	1687	0	
ERG	NM_182918	1540	1540	100	1825	285	
ERRFI1	NM_018948	1419	1419	100	1419	0	
ESR1	NM_001122740	1868	1868	100	1874	6	
ETAA1	NM_019002	2841	2841	100	2841	0	
ETV1	NM_001163147	1475	1475	100	1625	150	
ETV6	NM_001987	1439	1439	100	1597	158	
EZH1	NM_001991	2434	2434	100	2445	11	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
EZH2	NM_004456	2446	2446	100	2446	0	
EZHIP	NM_203407	1522	1522	100	1522	0	
FANCA	NM_000135	4798	4798	100	4963	165	
FANCB	NM_001018113	2660	2660	100	2660	0	
FANCC	NM_000136	1817	1817	100	1977	160	
FANCD2	NM_001018115	4786	4786	100	4921	135	FANCD2 exons 14,15,16,17, 21, 22, 23, 25 and 28 are fully encompassed by problematic region due to high homology
FANCE	NM_021922	1711	1711	100	1711	0	
FANCF	NM_022725	1135	1135	100	1135	0	
FANCG	NM_004629	2009	2009	100	2009	0	
FANCI	NM_001113378	4357	4357	100	4357	0	
FANCL	NM_018062	1268	1268	100	1283	15	
FANCM	NM_020937	6377	6377	100	6385	8	
FAS	NM_000043	1098	1098	100	1098	0	
FAT1	NM_005245	14027	14027	100	14027	0	
FBXW7	NM_033632	2234	2234	100	2692	458	
FGF19	NM_005117	681	681	100	681	0	
FGF23	NM_020638	786	786	100	786	0	
FGF3	NM_005247	750	750	100	750	0	
FGF4	NM_002007	651	651	100	651	0	
FGFR1	NM_001174067	2742	2742	100	2825	83	
FGFR2	NM_000141	2636	2636	100	2910	274	
FGFR3	NM_000142	2591	2591	100	2781	190	
FGFR4	NM_213647	2579	2579	100	2643	64	
FH	NM_000143	1633	1633	100	1633	0	
FLCN	NM_144997	1850	1850	100	2008	158	
FLT1	NM_002019	4317	4317	100	4593	276	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
FLT3	NM_004119	3222	3222	100	3222	0	
FLT4	NM_182925	4392	4392	100	4444	52	
FOXA1	NM_004496	1439	1439	100	1439	0	
FOXF1	NM_001451	1160	1160	100	1160	0	
FOXL2	NM_023067	1141	1141	100	1141	0	
FOXO1	NM_002015	1988	1988	100	1988	0	
FOXP1	NM_001244814	2194	2194	100	2549	355	
FUBP1	NM_003902	2135	2135	100	2208	73	
FYN	NM_153047	1715	1715	100	1890	175	
GAB1	NM_207123	2285	2285	100	2285	0	
GAB2	NM_080491	2131	2131	100	2131	0	
GATA1	NM_002049	1292	1292	100	1292	0	
GATA2	NM_032638	1493	1493	100	1643	150	
GATA3	NM_002051	1382	1382	100	1385	3	
GEN1	NM_001130009	2857	2857	100	2857	0	
GLI1	NM_005269	3431	3431	100	3431	0	
GNA11	NM_002067	1150	1150	100	1150	0	
GNA13	NM_006572	1174	1174	100	1174	0	
GNAQ	NM_002072	1150	1150	100	1150	0	
GNAS	NM_000516	1315	1315	100	4186	2871	
GNB1	NM_002074	1113	1113	100	1113	0	
GPS2	NM_004489	1084	1084	100	1084	0	
GREM1	NM_013372	565	565	100	565	0	
GRIN2A	NM_001134407	4515	4515	100	4515	0	
GSK3B	NM_002093	1422	1422	100	1422	0	
H1-2	NM_005319	652	652	100	652	0	
H2BC5	NM_021063	391	391	100	391	0	
H3-3A	NM_002107	441	441	100	441	0	
H3-3B	NM_005324	441	441	100	441	0	
H3-4	NM_003493	421	421	100	421	0	
H3-5	NM_001013699	418	418	100	418	0	
H3C1	NM_003529	421	421	100	421	0	
H3C10	NM_003536	421	421	100	421	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
H3C11	NM_003533	421	421	100	421	0	
H3C12	NM_003535	421	421	100	421	0	
H3C13	NM_001123375	421	421	100	421	0	Also known as HIST2H3D, fully covered by problematic region due to high homology
H3C14	NM_021059	421	421	100	421	0	Also known as HIST2H3C, fully covered by problematic region due to high homology
H3C2	NM_003537	421	421	100	421	0	
H3C3	NM_003531	421	421	100	421	0	
H3C4	NM_003530	421	421	100	421	0	
H3C6	NM_003532	421	421	100	421	0	
H3C7	NM_021018	421	421	100	421	0	
H3C8	NM_003534	421	421	100	421	0	
HDAC2	NM_001527	1607	1607	100	1607	0	
HGF	NM_000601	2367	2367	100	2393	26	
HLA-A	NM_002116	1178	1178	100	1178	0	Fully covered by problematic regions due to high polymorphism
HLA-B	NM_005514	1159	1159	100	1159	0	Fully covered by problematic regions due to high polymorphism
HLA-C	NM_002117	1181	1181	100	1181	0	Fully covered

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
							by problematic regions due to high polymorphism
HNF1A	NM_000545	1996	1996	100	2017	21	
HOXB13	NM_006361	875	875	100	875	0	
HRAS	NM_005343	610	610	100	702	92	
ICOSLG	NM_015259	979	979	100	1006	27	
ID3	NM_002167	380	380	100	380	0	
IDH1	NM_005896	1325	1325	100	1325	0	
IDH2	NM_002168	1469	1469	100	1469	0	
IDO2	NM_194294	1324	1324	100	1324	0	
IFNGR1	NM_000416	1540	1540	100	1605	65	
IGF1	NM_001111285	628	628	100	723	95	
IGF1R	NM_000875	4314	4314	100	4314	0	
IGF2	NM_001127598	751	751	100	751	0	
IKBKE	NM_014002	2351	2351	100	2351	0	
IKZF1	NM_006060	1630	1630	100	1630	0	
IL10	NM_000572	587	587	100	587	0	
IL7R	NM_002185	1460	1460	100	1460	0	
INHA	NM_002191	1121	1121	100	1121	0	
INHBA	NM_002192	1301	1301	100	1301	0	
INPP4A	NM_001134224	3174	3174	100	3368	194	
INPP4B	NM_001101669	3005	3005	100	3075	70	
INPPL1	NM_001567	4057	4057	100	4057	0	
INSR	NM_000208	4369	4369	100	4369	0	
IRF4	NM_002460	1436	1436	100	1436	0	
IRS1	NM_005544	3739	3739	100	3739	0	
IRS2	NM_003749	4037	4037	100	4037	0	
JAK1	NM_002227	3705	3705	100	3705	0	
JAK2	NM_004972	3629	3629	100	3629	0	
JAK3	NM_000215	3605	3605	100	3605	0	
JUN	NM_002228	1006	1006	100	1006	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
KBTBD4	NM_018095	1645	1645	100	1735	90	
KDM5A	NM_001042603	5353	5353	100	5353	0	
KDM5C	NM_004187	4943	4943	100	5047	104	
KDM6A	NM_021140	4496	4496	100	4662	166	
KDR	NM_002253	4371	4371	100	4371	0	
KEAP1	NM_203500	1925	1925	100	1925	0	
KIT	NM_000222	3141	3141	100	3144	3	
KLF4	NM_004235	1490	1490	100	1582	92	
KLF5	NM_001730	1414	1414	100	1414	0	
KMT2A	NM_001197104	12279	12279	100	12279	0	
KMT2B	NM_014727	8518	8518	100	8518	0	
KMT2C	NM_170606	15326	15326	100	15326	0	KMT2C exons 7, 8, 14-21, 24 and 34 are fully encompassed by problematic regions due to high homology
KMT2D	NM_003482	17154	17154	100	17154	0	
KMT5A	NM_020382	1139	1139	100	1139	0	
KNSTRN	NM_033286	1041	1041	100	1090	49	
KRAS	NM_004985	607	607	100	737	130	
LATS1	NM_004690	3463	3463	100	3526	63	
LATS2	NM_014572	3337	3337	100	3337	0	
LDB1	NM_001113407	1346	1346	100	1346	0	
LMO1	NM_002315	511	511	100	543	32	
LYN	NM_002350	1659	1659	100	1659	0	
LZTR1	NM_006767	2733	2733	100	2733	0	
MAD2L2	NM_001127325	716	716	100	716	0	
MALT1	NM_006785	2645	2645	100	2645	0	
MAP2K1	NM_002755	1292	1292	100	1316	24	
MAP2K2	NM_030662	1313	1313	100	1313	0	
MAP2K4	NM_003010	1310	1310	100	1353	43	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
MAP3K1	NM_005921	4739	4739	100	4739	0	
MAP3K13	NM_004721	3031	3031	100	3079	48	
MAP3K14	NM_003954	2993	2993	100	2993	0	
MAPK1	NM_002745	1163	1163	100	1163	0	
MAPK3	NM_002746	1220	1220	100	1277	57	
MAPKAP1	NM_001006617	1679	1679	100	1703	24	
MAX	NM_002382	533	533	100	908	375	
MCL1	NM_021960	1083	1083	100	1094	11	
MDC1	NM_014641	6410	6410	100	6410	0	
MDM2	NM_002392	1604	1604	100	1604	0	
MDM4	NM_002393	1573	1573	100	1573	0	
MED12	NM_005120	6984	6984	100	6984	0	
MEF2B	NM_001145785	1187	1186	99.92	1186	0	MEF2B exon9 covered with a -4bp/+5bp padding
MEN1	NM_130799	1923	1923	100	1938	15	
MET	NM_000245	4373	4373	100	4838	465	Includes regions relevant for MET ex14 skipping 7:116411547-116412200, spanning exon 13, intron13 and exon14
MGA	NM_001164273	9428	9428	100	9575	147	
MITF	NM_000248	1350	1350	100	1926	576	
MLH1	NM_000249	2461	2461	100	2461	0	
MLLT1	NM_005934	1800	1800	100	1800	0	
MPL	NM_005373	2028	2028	100	2028	0	
MRE11	NM_005591	2317	2317	100	2317	0	
MSH2	NM_000251	2965	2965	100	3517	552	
MSH3	NM_002439	3654	3654	100	3654	0	
MSH6	NM_000179	4183	4183	100	4189	6	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
MSI1	NM_002442	1229	1229	100	1229	0	
MSI2	NM_138962	1117	1117	100	1272	155	
MST1	NM_020998	2358	2358	100	2403	45	MST1 exons 1,2,4,6-10,12-18 are fully encompassed in problematic regions due to sequence homology
MST1R	NM_002447	4403	4403	100	4403	0	
MTAP	NM_002451	932	932	100	1273	341	
MTOR	NM_004958	8220	8220	100	8220	0	
MUTYH	NM_001128425	1810	1810	100	1810	0	
MYC	NM_002467	1395	1395	100	1395	0	
MYCL	NM_001033082	1215	1215	100	1340	125	
MYCN	NM_005378	1415	1415	100	1532	117	
MYD88	NM_002468	941	941	100	965	24	
MYOD1	NM_002478	993	993	100	993	0	
NADK	NM_001198994	1906	1906	100	2095	189	
NBN	NM_002485	2425	2425	100	2425	0	
NCOA3	NM_181659	4485	4485	100	4515	30	
NCOR1	NM_006311	7773	7773	100	7829	56	
NEGR1	NM_173808	1135	1135	100	1135	0	
NF1	NM_000267	9027	9027	100	11826	2799	
NF2	NM_000268	1948	1948	100	2170	222	
NFE2L2	NM_006164	1868	1868	100	1868	0	
NFKBIA	NM_020529	1014	1014	100	1014	0	
NKX2-1	NM_001079668	1236	1236	100	1236	0	
NKX3-1	NM_006167	725	725	100	725	0	
NOTCH1	NM_017617	8008	8008	100	8008	0	
NOTCH2	NM_024408	7756	7756	100	7809	53	NOTCH2 exons 1-4 are fully encompassed by

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
							problematic regions due to high sequence homology
NOTCH3	NM_000435	7296	7296	100	7296	0	
NOTCH4	NM_004557	6312	6312	100	6312	0	
NPM1	NM_002520	995	995	100	1014	19	
NRAS	NM_002524	610	610	100	610	0	
NSD1	NM_022455	8311	8311	100	8311	0	
NSD2	NM_001042424	4308	4308	100	4390	82	
NSD3	NM_023034	4544	4544	100	4637	93	
NTHL1	NM_002528	975	975	100	975	0	
NTRK1	NM_002529	2561	2561	100	2703	142	
NTRK2	NM_006180	2697	2697	100	2784	87	
NTRK3	NM_001012338	2700	2700	100	2999	299	
NUF2	NM_031423	1525	1525	100	1525	0	
NUP93	NM_014669	2670	2670	100	2670	0	
PAK1	NM_002576	1778	1778	100	1886	108	
PAK5	NM_177990	2240	2240	100	2240	0	
PALB2	NM_024675	3691	3691	100	3691	0	
PARP1	NM_001618	3275	3275	100	3275	0	
PAX5	NM_016734	1276	1276	100	1307	31	
PBRM1	NM_018313	5039	5039	100	5562	523	
PDCD1	NM_005018	917	917	100	917	0	
PDCD1LG2	NM_025239	882	882	100	882	0	
PDGFRA	NM_006206	3490	3490	100	3502	12	
PDGFRB	NM_002609	3541	3541	100	3575	34	
PDPK1	NM_002613	1811	1811	100	1811	0	PDPK1 exons 3-6 and 8-10 are fully encompassed by problematic regions due to high sequence

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
							homology
PGBD5	NM_001258311	1645	1645	100	1645	0	
PGR	NM_000926	2882	2882	100	2882	0	
PHF6	NM_032458	1188	1188	100	1293	105	
PHOX2B	NM_003924	975	975	100	975	0	
PIK3C2G	NM_004570	4648	4648	100	4781	133	
PIK3C3	NM_002647	2914	2914	100	2914	0	
PIK3CA	NM_006218	3407	3407	100	3407	0	
PIK3CB	NM_006219	3433	3433	100	3433	0	
PIK3CD	NM_005026	3355	3355	100	3355	0	
PIK3CG	NM_002649	3409	3409	100	3409	0	
PIK3R1	NM_181523	2325	2325	100	2467	142	
PIK3R2	NM_005027	2337	2337	100	2337	0	
PIK3R3	NM_003629	1486	1486	100	1547	61	
PIM1	NM_002648	1002	1001	99.90	1001	0	PIM1 exon1 covered with a -4bp/+5bp padding
PLCG2	NM_002661	4118	4118	100	4118	0	
PLK2	NM_006622	2198	2198	100	2198	0	
PMAIP1	NM_021127	185	185	100	185	0	
PML	NM_033238	2739	2739	100	3768	1029	
PMS1	NM_000534	2919	2919	100	2919	0	
PMS2	NM_000535	2739	2739	100	2840	101	PMS2 exons 9 and 11-15 are fully encompassed by problematic regions due to high sequence homology
PNRC1	NM_006813	1004	1004	100	1004	0	
POLD1	NM_002691	3584	3584	100	3662	78	
POLE	NM_006231	7351	7351	100	7351	0	
POT1	NM_015450	2055	2055	100	2055	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
PPARG	NM_015869	1588	1588	100	1654	66	
PPM1D	NM_003620	1878	1878	100	1878	0	
PPP2R1A	NM_014225	1920	1920	100	1920	0	
PPP2R2A	NM_002717	1444	1444	100	1491	47	
PPP4R2	NM_174907	1344	1344	100	1423	79	
PPP6C	NM_002721	988	988	100	1109	121	
PRDM1	NM_001198	2548	2548	100	2567	19	
PRDM14	NM_024504	1786	1786	100	1786	0	
PREX2	NM_024870	5221	5221	100	5456	235	
PRKAR1A	NM_212471	1246	1246	100	1297	51	
PRKCI	NM_002740	1971	1971	100	1971	0	
PRKD1	NM_002742	2919	2919	100	2919	0	
PRKN	NM_004562	1518	1518	100	1518	0	
PRPF8	NM_006445	7428	7428	100	7428	0	
PTCH1	NM_000264	4574	4574	100	4949	375	
PTEN	NM_000314	1302	1301	99.92	1451	150	PTEN exon1 covered with a -4bp/+5bp padding
PTP4A1	NM_003463	572	572	100	652	80	
PTPN11	NM_002834	1932	1932	100	1948	16	
PTPRD	NM_002839	6089	6089	100	6169	80	
PTPRS	NM_002850	6217	6217	100	6217	0	
PTPRT	NM_133170	4703	4703	100	4730	27	
RAB35	NM_006861	666	666	100	666	0	
RAC1	NM_018890	706	706	100	706	0	
RAC2	NM_002872	639	639	100	639	0	
RAD21	NM_006265	2026	2026	100	2026	0	
RAD50	NM_005732	4189	4189	100	4189	0	
RAD51	NM_002875	1110	1110	100	1110	0	
RAD51B	NM_133509	1255	1255	100	1309	54	
RAD51C	NM_058216	1221	1221	100	1225	4	
RAD51D	NM_002878	1087	1087	100	1276	189	
RAD52	NM_134424	1367	1367	100	1419	52	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
RAD54L	NM_001142548	2424	2424	100	2424	0	
RAF1	NM_002880	2107	2107	100	2107	0	
RARA	NM_000964	1469	1469	100	1642	173	
RASA1	NM_002890	3394	3394	100	3412	18	
RB1	NM_000321	3057	3057	100	3388	331	
RBM10	NM_001204468	3228	3228	100	3228	0	
RECQL	NM_002907	2090	2090	100	2090	0	
RECQL4	NM_004260	3838	3838	100	3912	74	
REL	NM_002908	1970	1970	100	1970	0	
REST	NM_001193508	3324	3324	100	3324	0	
RET	NM_020975	3545	3545	100	3577	32	
RHEB	NM_005614	635	635	100	635	0	
RHOA	NM_001664	622	622	100	622	0	
RICTOR	NM_152756	5507	5507	100	5589	82	
RIT1	NM_006912	710	710	100	771	61	
RNF43	NM_017763	2442	2442	100	2442	0	
ROS1	NM_002944	7474	7474	100	7477	3	
RPS6KA4	NM_003942	2489	2489	100	2489	0	
RPS6KB2	NM_003952	1599	1599	100	1599	0	
RPTOR	NM_020761	4348	4348	100	4348	0	
RRAGC	NM_022157	1270	1270	100	1270	0	
RRAS	NM_006270	717	717	100	717	0	
RRAS2	NM_012250	675	675	100	688	13	
RTEL1	NM_001283009	4243	4243	100	4315	72	
RUNX1	NM_001754	1523	1523	100	1728	205	
RXRA	NM_002957	1489	1489	100	1489	0	
RYBP	NM_012234	697	697	100	697	0	
SCG5	NM_001144757	689	689	100	689	0	
SDHA	NM_004168	2145	2145	100	2145	0	
SDHAF2	NM_017841	541	541	100	541	0	
SDHB	NM_003000	923	923	100	923	0	
SDHC	NM_003001	570	570	100	677	107	
SDHD	NM_003002	520	520	100	520	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
SERPINB3	NM_006919	1243	1243	100	1243	0	
SERPINB4	NM_002974	1243	1243	100	1243	0	
SESN1	NM_014454	1756	1756	100	1868	112	
SESN2	NM_031459	1543	1543	100	1543	0	
SESN3	NM_144665	1579	1579	100	1579	0	
SETD2	NM_014159	7905	7905	100	7905	0	
SETDB1	NM_001145415	4086	4086	100	4143	57	
SF3B1	NM_012433	4165	4165	100	4245	80	
SH2B3	NM_005475	1798	1798	100	1934	136	
SH2D1A	NM_002351	427	427	100	427	0	
SHOC2	NM_007373	1829	1829	100	1829	0	
SHQ1	NM_018130	1844	1844	100	1844	0	
SLFN11	NM_001104589	2746	2746	100	2746	0	SLFN11 exon 5 is fully encompassed by problematic regions due to high sequence homology
SLX4	NM_032444	5645	5645	100	5645	0	
SMAD2	NM_001003652	1504	1504	100	1504	0	
SMAD3	NM_005902	1368	1368	100	1452	84	
SMAD4	NM_005359	1769	1769	100	1776	7	
SMARCA2	NM_003070	5103	5103	100	5103	0	
SMARCA4	NM_001128849	5390	5390	100	5399	9	
SMARCB1	NM_003073	1248	1248	100	1452	204	
SMARCD1	NM_003076	1678	1678	100	1678	0	
SMARCE1	NM_003079	1336	1336	100	1336	0	
SMO	NM_005631	2484	2484	100	2484	0	
SMYD3	NM_001167740	1407	1407	100	1407	0	
SOCS1	NM_003745	646	646	100	646	0	
SOS1	NM_005633	4232	4232	100	4232	0	
SOX17	NM_022454	1265	1265	100	1265	0	
SOX2	NM_003106	964	964	100	964	0	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
SOX9	NM_000346	1560	1560	100	1560	0	
SPEN	NM_015001	11145	11145	100	11145	0	
SPOP	NM_001007228	1215	1215	100	1215	0	
SPRED1	NM_152594	1405	1405	100	1405	0	
SPRTN	NM_032018	1520	1520	100	1555	35	
SRC	NM_198291	1721	1721	100	1721	0	
SRSF2	NM_003016	686	686	100	686	0	
STAG2	NM_001042749	4137	4137	100	4137	0	
STAT3	NM_139276	2543	2543	100	2558	15	
STAT5A	NM_003152	2565	2565	100	2626	61	
STAT5B	NM_012448	2544	2544	100	2544	0	
STAT6	NM_003153	2754	2754	100	2754	0	
STK11	NM_000455	1392	1392	100	1392	0	
STK19	NM_004197	835	835	100	1187	352	STK19 exon 7 is fully encompassed by problematic regions due to high sequence homology
STK40	NM_032017	1408	1408	100	1423	15	
SUFU	NM_016169	1575	1575	100	1591	16	
SUZ12	NM_015355	2380	2380	100	2380	0	
SYK	NM_003177	2038	2038	100	2038	0	
TAP1	NM_000593	2357	2357	100	2357	0	
TAP2	NM_018833	2072	2072	100	2211	139	
TBX3	NM_016569	2312	2312	100	2312	0	
TCF3	NM_001136139	2136	2136	100	2382	246	
TCF7L2	NM_001146274	1949	1949	100	2422	473	
TEK	NM_000459	3605	3605	100	3605	0	
TENT5C	NM_017709	1186	1186	100	1186	0	
TERT	NM_198253	3559	3559	100	3611	52	TERT Promoter regions included: 5:1295156-

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
							1295166; 5:1295223-1295233; 5:1295237-1295255; 5:1295344-1295354
TET1	NM_030625	6521	6521	100	6521	0	
TET2	NM_001127208	6099	6099	100	6188	89	
TFE3	NM_006521	1828	1828	100	1828	0	
TGFBR1	NM_004612	1602	1602	100	1614	12	
TGFBR2	NM_003242	1774	1774	100	1859	85	
TMEM127	NM_017849	747	747	100	835	88	
TMPRSS2	NM_001135099	1730	1730	100	1752	22	
TNFAIP3	NM_006290	2453	2453	100	2453	0	
TNFRSF14	NM_003820	932	932	100	932	0	
TOP1	NM_003286	2508	2508	100	2508	0	
TP53	NM_000546	1282	1282	100	1383	101	
TP53BP1	NM_001141980	6214	6214	100	6214	0	
TP63	NM_003722	2183	2183	100	2360	177	
TRAF2	NM_021138	1606	1606	100	1606	0	
TRAF7	NM_032271	2213	2213	100	2213	0	
TRIP13	NM_004237	1429	1429	100	1433	4	
TSC1	NM_000368	3705	3705	100	3705	0	
TSC2	NM_000548	5834	5834	100	6217	383	
TSHR	NM_000369	2395	2395	100	2538	143	
U2AF1	NM_006758	803	803	100	880	77	
UGT1A1	NM_000463	1652	1652	100	1653	1	
UPF1	NM_002911	3587	3587	100	3620	33	
USH2A	NM_206933	16319	16319	100	16333	14	
USP8	NM_005154	3547	3547	100	3547	0	
VEGFA	NM_001171623	779	779	100	1385	606	
VHL	NM_000551	672	672	100	672	0	
VTCN1	NM_024626	899	899	100	899	0	
WT1	NM_024426	1669	1669	100	1689	20	

Gene Name	Transcript reference seq id	CDS +/- 5bp size [bp]	CDS +/- 5bp overlap with SNV/INDEL target [bp]	CDS +/- 5bp percentage overlap with target [bp]	Target region footprint [bp]	Number of additional bases covered [bp]	Comments
WWTR1	NM_001168280	1263	1263	100	1263	0	
XIAP	NM_001167	1554	1554	100	1554	0	
XPO1	NM_003400	3456	3456	100	3456	0	
XRCC2	NM_005431	873	873	100	873	0	
YAP1	NM_001130145	1605	1605	100	1617	12	
YES1	NM_005433	1742	1742	100	1742	0	
ZFHX3	NM_006885	11202	11202	100	11202	0	
ZFTA	NM_001144936	2087	2087	100	2087	0	
ZNRF3	NM_001206998	2901	2901	100	2901	0	
ZRSR2	NM_005089	1559	1559	100	1559	0	ZRSR2 exon 8 is fully encompassed by problematic regions due to high sequence homology

8.4 DNA analysis: gene-fusion and exon-skipping analysis

8.4.1 Analysis purpose

This analysis detects genomic junctions arising from structural variants in targeted DNA sequencing data and assesses their potential functional impact at the transcript level. Junctions supported by sequencing evidence are identified and evaluated individually. Only junctions consistent with gene fusions or exon-skipping events are retained and reported.

This analysis focuses on a subset of genomic regions known to harbor clinically relevant gene fusion events. In these regions, the DNA panel targets selected intronic sequences to enable the detection of DNA junctions that may give rise to aberrant transcripts. The complete list of exonic and intronic regions included in the analysis is provided in the Description of genomic regions targeted by the DNA fusion analysis section.

All detected events are annotated with respect to transcripts. The annotation process integrates external knowledge bases to add clinical significance where available (e.g. ChimerDb, COSMIC).

8.4.2 Technical overview of the analysis

This analysis involves five main steps after DNA data preprocessing.

Step 1: Candidate junction identification

The evidence for junctions obtained from DNA sequence data consists of:

- **Split reads**: cases where different parts of a single read map to two different locations in the genome.
- **Discordant read-pairs**: cases where the two reads in a read-pair map to genomic locations that are further apart than expected from the insert-size distribution (i.e. the size of the sequenced molecules).

To effectively identify split reads, the algorithm individually realigns all soft-clip sequences of size >=20bp.

Split reads and discordant read pairs are grouped based on their mapping coordinates to identify candidate junctions and their corresponding breakpoints.

Step 2: Statistical evaluation of candidate junctions and score assignment

Each candidate junction is evaluated by a multivariate statistical model. The features considered by the statistical model include:

- Number of CUMIN™ groups supporting the junction (including split reads and discordant pairs).
- Lengths of soft-clips supporting the junction.
- Mismatches (relative to reference genome) in reads supporting the junction.
- Sequencing quality scores.
- Sequence complexity (repetitiveness) of sequences with unique supporting reads.
- Mapping quality of reads supporting the junction.
- Diversity in start-end mapping positions of unique supporting reads.
- Distance between the junction breakpoints (if both breakpoints are on the same chromosome).

The statistical model (informed by the expected behavior of the features in the presence/absence of a true variant) evaluates the supporting features of each putative junction and produces a score reflecting the likelihood that the junction represents a true biological event rather than a technical artifact.

Step 3: Calling

Calling is performed by **retaining all candidate junctions** that meet the following criteria:

- The algorithm score computed in step 2 is ≥ 0.1, indicating sufficient confidence that the junction represents a true biological event.
- Support is observed from ≥ 2 CUMIN™ groups, providing adequate molecular evidence.
- Both breakpoints map within intragenic regions, consistent with potential gene fusions or exon-skipping events.
- For events consistent with potential gene fusions, the transcripts of both genes involved must maintain their native 5'-3' transcription direction.
- For events with breakpoints on the same chromosome, the distance between them must be ≥ 1000 bp, as the algorithm is optimized for larger structural variants (structural variants between 50 and 1000 bp are outside the scope of this analysis).

Candidate junctions that satisfy all these criteria are retained and reported as potential gene fusion or exon-skipping events.

Step 4: Filtering

Variants are classified as high-confidence and low-confidence variants based on the following filters:

- **off_target:** Both breakpoints fall outside the genomic regions in scope of the analysis.

- **low_probability_score**: The algorithm score is < 0.5.
- **low_molecular_support**: Fewer than 3 CUMIN™ groups support the variant, reflecting weak molecular support.
- **problematic_region**: At least one breakpoint overlaps a genomic region known to have an increased risk of false positives in this panel.

Each variant is evaluated against all defined filters. The variant is classified as high confidence only when no filter is triggered.

Step 5: Annotation

Variants are annotated with their type, gene, gene partners, genomic breakpoints in genomic coordinates and reading frame status. In cases where multiple transcripts from different genes share the same breakpoint, the junction may be reported as multiple gene fusions.

Variants are finally cross-referenced against curated databases, including COSMIC_SV and ChimerDB.

8.4.3 Description of the results

This analysis outputs a list of tertiary-annotated gene-fusion and exon-skipping events detected in the DNA sample. The results are made available via the SOPHiA DDM™ platform as well as downloadable files.

The SOPHiA DDM™ platform Help Center includes a description of the various fields displayed in the variant table. The Definition of variant attributes reported in the full_DNA_fusion_table.txt section provides a description of the variant attributes present in the downloadable variant table file.

The following table provides an overview of the downloadable output files produced by this module.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
full_DNAfusion_table.txt	Sample	Table with high-and-low confidence gene fusion and exon skipping calls

8.4.4 Precautions and limitations

- Structural variants occurring in or near repetitive or transposable elements may lead to false positive calls.

- Structural variants involving pseudogene regions may result in false positive calls due to sequence similarity with functional genes.
- In rare cases, distinct junctions with very close breakpoints may be interpreted as a single event, potentially leading to false negatives.
- Variant fractions are often underestimated due to reduced capture efficiency of DNA fragments spanning junctions.
- In rare cases, structural variants outside the defined target regions may be considered on target because of padding applied to the target intervals. This padding ensures that variants with minor deviations from the target region remain in scope.
- In samples with an unusually high proportion of soft-clipped reads, the likelihood of false positive results may be increased.

8.4.5 Definition of variant attributes reported in the full_DNA_fusion_table.txt

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
Genomic description (hg19)	type	Variant type (e.g. gene.fusion)
	id	Variant Identifier combining gene names and genomic breakpoints
	refGenome	Reference genome to describe the variant
	5.prime.pos	Genomic position of the 5' breakpoint (chr:pos)
	5.prime.chromosome	Chromosome of the 5' breakpoint (with “chr” prefix)
	5.prime.seqName	Chromosome of the 5' breakpoint (without “chr” prefix)
	5.prime.pos1	Genomic position of the 5' breakpoint (numeric position only)
	3.prime.pos	Genomic position of the 3' breakpoint (chr:pos)
	3.prime.chromosome	Chromosome of the 3' breakpoint (with “chr” prefix)
	3.prime.seqName	Chromosome of the 3' breakpoint (without “chr” prefix)
	3.prime.pos1	Genomic position of the 3' breakpoint (numeric position only)
	DNAseq	DNA sequence at the breakpoint junction (if available)
Transcript	5.prime.tx	Transcript ID at the 5' breakpoint

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
annotation	5.prime.gene	Gene name at the 5' breakpoint
	5.prime.exon	Exon number at the 5' breakpoint
	5.prime.type	Genomic feature type at 5' breakpoint (e.g., exon, intron)
	3.prime.tx	Transcript ID at the 3' breakpoint
	3.prime.gene	Gene name at the 3' breakpoint
	3.prime.exon	Exon number at the 3' breakpoint
	3.prime.type	Genomic feature type at 3' breakpoint (e.g., exon, intron)
	coding.consequence	Predicted effect on coding sequence (if determinable)
	In-Frame	Whether the variant preserves the reading frame (Yes/No)
	description	Free-text description of the variant (if available)
Signal quantification	read.count	Number of sequencing reads supporting the event
	mol.count	Number of unique molecules (deduplicated reads) supporting the event
	read.percentage	Fraction of supporting reads relative to total reads
	mol.percentage	Fraction of supporting molecules relative to total molecules
Warning and quality	filter	Indicates whether a variant has passed or failed internal quality control filters
Variant classification database	COSMIC	Match to known variants in COSMIC database (if any)
	ChimerDB	Match to known fusion/chimeric events in ChimerDB (if any)
Internal reference	sgid	Internal variant ID

8.4.6 Description of genomic regions targeted by the DNA fusion analysis

Notes:

- Genes denoted with an asterisk ("*") are also included in the RNA gene-fusion and exon-skipping analysis.
- Target regions marked with an asterisk ("*") are only partially covered by the DNA panel. The corresponding genomic intervals targeted by the DNA panel are listed in the column **"Targeted Genomic Coordinates."**

GENE	TRANSCRIPT	TARGETED REGIONS	TARGETED GENOMIC COORDINATES
ALK*	NM_004304	ex20,in19,ex19,in18*,ex18,in17,ex17	2:29446204-29449079; 2:29449320-29450543
BCL2L11 (BIM)	NM_138621	in2	2:111883002-111886500
BRAF*	NM_004333	ex11,in10*,ex10,in9*,ex9,in8*,ex8,in7*,ex7	7:140481372-140481707; 7:140481948-140482367; 7:140482428-140484493; 7:140484734-140486833; 7:140487134-140488436; 7:140488977-140489576; 7:140489817-140490656;7:140490897-140493476; 7:140493537-140495865; 7:140496106-140498205; 7:140498326-140498685; 7:140498926-140500286
CD74	NM_001025159	in6*,ex6,in5,ex5,in4*	5:149782877-149783169; 5:149783410-149785362; 5:149785603-149785821
DNAJB1*	NM_006145	ex3,in2,ex2,in1,ex1	19:14626748-14629166
EGFR*	NM_005228	ex7,in7*,ex8	7:55221700-55222924; 7:55223105-55223644
		ex17,in17,ex18	7:55240672-55241741
		ex24,in24,ex25,in25,ex26,in26,ex27	7:55268005-55270463
ETV6*	NM_001987	ex4,in4*,ex5,in5*,ex6	12:12006357-12007016; 12:12007077-

GENE	TRANSCRIPT	TARGETED REGIONS	TARGETED GENOMIC COORDINATES
			12007976;12:12008217-12012056;12:12012357-12017936;12:12018177-12021356;12:12021477-12021896; 12:12022137-12024201; 12:12024382-12025461;12:12025702-12027081;12:12027322-12027801;12:12027982-12033981;12:12034042-12036201; 12:12036502-12037526
EWSR1*	NM_005243	in7*,ex8,in8,ex9,in9,ex10,in10,ex11, in11*,ex12,in12*,ex13, in13*,ex14	22:29683125-29683286;22:29683467-29683886; 22:29683947-29688806 22:29688987-29689586;22:29689827-29690006;22:29690127-29690306;22:29690727-29690846;22:29690907-29691086; 22:29691147-29692886 22:29693007-29694206; 22:29694507-29694788
FGFR2*	NM_000141	ex18,in17*,ex17,in16,ex16	10:123239367-123241373; 10:123241614-123242333; 10:123242574-123245051
FGFR3*	NM_000142	ex14,in14,ex15,in15,ex16,in16,ex17,in17,ex18	4:1807774-1809020
MET*	NM_000245	ex13,in13,ex14,in14,ex15	7:116411482-116415230
NAB2	NM_005967	in2,ex3,in3,ex4,in4,ex5,in5,ex6,in6,ex7	12:57485783-57489250
NTRK1*	NM_002529	ex3,in3,ex4	1:156834516-156836775

GENE	TRANSCRIPT	TARGETED REGIONS	TARGETED GENOMIC COORDINATES
		ex7,in7*,ex8,in8,ex9,in9,ex10,in10,ex11,in11,ex12,in12,ex13	1:156841411-156842486; 1:156842727-156842846; 1:156842967-156846007
NTRK2*	NM_006180	ex13,in13*,ex14	9:87475951-87479650; 9:87479951-87482351
NUTM1*	NM_175741	in2*	15:34638238-34638813; 15:34639054-34639353; 15:34639414-34639593;15:34639714-34639833;15:34640074-34640169
PAX8*	NM_003466	in10*,ex10,in9*,ex9,in8	2:113977757-113981873; 2:113982114-113982413; 2:113982474-113989322; 2:113989563-113991002; 2:113991243-113992142; 2:113992383-113994177
RELA*	NM_021975	in2-ex2-in1	11:65429561-65430296
RET*	NM_020975	ex7,in7,ex8,in8,ex9,in9,ex10,in10,ex11,in11,ex12	10:43606651-43612184
ROS1*	NM_002944	ex36,in35,ex35,in34,ex34,in33,ex33,in32,ex32,in31*,ex31,in30,ex30	6:117641027-117651681; 6:117651742-117652221; 6:117652342-117652581; 6:117653062-117653181;6:117654142-117654321; 6:117654622-117654741; 6:117655522-117655641;6:117657082-117657441; 6:117657682-117662479

GENE	TRANSCRIPT	TARGETED REGIONS	TARGETED GENOMIC COORDINATES
TFE3*	NM_006521	ex6,in5*,ex5,in4,ex4,in3,ex3	X:48891645-48892546; X:48892847-48892966; X:48893207-48893446;X:48894347-48894586;X:48894827-48894946; X:48895247-48896940
TMPRSS2*	NM_001135099	ex3,in2*,ex2,in1*,ex1	21:42866279-42866745; 21:42866986-42867285; 21:42867526-42868785; 21:42869026-42873286; 21:42873527-42879936
TP53	NM_000546	ex2,in1*,ex1	17:7579835-7580254; 17:7580495-7581334;17:7581575-7581814;17:7582055-7582294;17:7582535-7582774;17:7582895-7583854;17:7584395-7584514;17:7585415-7585534;17:7585835-7586254;17:7586855-7587034;17:7587275-7587454;17:7587515-7588774;17:7589015-7590868

8.5 DNA analysis: gene amplifications and deletions detection and annotation via MUSKAT™ and MOKA™

8.5.1 Analysis purpose

Copy number variants (CNVs) are structural changes in the DNA associated with variations in the number of copies of the affected DNA segments. CNV analysis is based on the MUSKAT™ proprietary algorithm which is designed to detect whole-gene amplification and deletion events by considering the NGS data aligning to genomic regions targeted by the panel.

Each variant is then annotated to provide transcript-specific descriptions and integrate external knowledge base catalogs (e.g., dbVar).

A total of 520 genes targeted by the DNA panel are in scope of whole-gene amplification and deletion CNV analysis. The full list is provided below:

ABL1, ABRAXAS1, ACVR1, AGO1, AGO2, AKT1, AKT2, AKT3, ALB, ALK, ALOX12B, AMER1, ANKRD11, APC, APLNR, AR, ARAF, ARHGAP35, ARID1A, ARID1B, ARID2, ARID5B, ASS1, ASXL1, ASXL2, ATM, ATR, ATRX, ATXN7, AURKA, AURKB, AXIN1, AXIN2, AXL, B2M, BABAM1, BAP1, BARD1, BBC3, BCL10, BCL2, BCL2L1, BCL2L11, BCL6, BCOR, BCORL1, BIRC3, BLM, BMPR1A, BRAF, BRCA1, BRCA2, BRD4, BRIP1, BTK, C11orf95, CALR, CARD11, CARM1, CASP8, CBFB, CBL, CCND1, CCND2, CCND3, CCNE1, CCNQ, CD274, CD276, CD58, CD79A, CD79B, CDC42, CDC73, CDH1, CDK12, CDK4, CDK6, CDK8, CDKN1A, CDKN1B, CDKN2A, CDKN2B, CDKN2C, CEBPA, CENPA, CHEK1, CHEK2, CIC, CMTR2, COP1, CREBBP, CRKL, CRLF2, CSDE1, CSF1R, CSF3R, CTCF, CTLA4, CTNNB1, CTR9, CUL3, CXCR4, CXorf67, CYLD, CYP19A1, CYSLTR2, DAXX, DCUN1D1, DDR1, DDR2, DICER1, DIS3, DNAJB1, DNMT1, DNMT3A, DNMT3B, DOT1L, DPYD, DROSHA, DUSP4, E2F3, EED, EGFL7, EGFR, EIF1AX, EIF4A2, EIF4E, ELF3, ELOC, EP300, EPAS1, EPCAM, EPHA3, EPHA5, EPHA7, EPHB1, ERBB2, ERBB3, ERBB4, ERCC2, ERCC3, ERCC4, ERCC5, ERF, ERG, ERRFI1, ESR1, ETAA1, ETV1, ETV6, EZH1, EZH2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FAS, FAT1, FBXW7, FGF19, FGF23, FGF3, FGF4, FGFR1, FGFR2, FGFR3, FGFR4, FH, FLCN, FLT1, FLT3, FLT4, FOXA1, FOXF1, FOXL2, FOXO1, FOXP1, FUBP1, FYN, GAB1, GAB2, GATA1, GATA2, GATA3, GEN1, GLI1, GNA11, GNA13, GNAQ, GNAS, GNB1, GPS2, GREM1, GRIN2A, GSK3B, H3F3A, H3F3C, HDAC2, HGF, HIST1H1C, HIST1H2BD, HIST1H3B, HIST1H3C, HIST3H3, HLA-B, HNF1A, HOXB13, HRAS, ICOSLG, ID3, IDH1, IDH2, IDO2, IFNGR1, IGF1, IGF1R, IGF2, IKBKE, IKZF1, IL10, IL7R, INHA, INHBA, INPP4A, INPP4B, INPPL1, INSR, IRF4, IRS1, IRS2, JAK1, JAK2, JAK3, JUN, KBTBD4, KDM5A, KDM5C, KDM6A, KDR, KEAP1, KIT, KLF4, KLF5, KMT2A, KMT2B, KMT2C, KMT2D, KMT5A, KNSTRN, KRAS, LATS1, LATS2, LDB1, LMO1, LYN, LZTR1, MAD2L2, MALT1, MAP2K1, MAP2K2, MAP2K4, MAP3K1, MAP3K13, MAP3K14, MAPK1, MAPK3, MAPKAP1, MAX, MCL1, MDC1, MDM2, MDM4, MED12, MEF2B, MEN1, MET, MGA, MITF, MLH1, MLLT1, MPL, MRE11, MSH2, MSH3, MSH6, MSI1, MSI2, MST1, MST1R, MTAP, MTOR, MUTYH, MYC, MYCL, MYCN, MYD88, MYOD1, NADK, NBN, NCOA3, NCOR1, NEGR1, NF1, NF2, NFE2L2, NFKBIA, NKX2-1, NKX3-1, NOTCH1, NOTCH2, NOTCH3, NOTCH4, NPM1, NRAS, NSD1, NSD2, NSD3, NTHL1, NTRK1, NTRK2, NTRK3, NUF2, NUP93, PAK1, PAK5, PALB2, PARP1, PAX5, PBRM1, PDCD1, PDCD1LG2, PDGFRA, PDGFRB, PDPK1, PGBD5, PGR, PHF6, PHOX2B, PIK3C2G, PIK3C3, PIK3CA, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3, PIM1, PLCG2, PLK2, PMAIP1, PML, PMS1, PMS2, PNRC1, POLD1, POLE, POT1, PPARG, PPM1D, PPP2R1A, PPP2R2A, PPP4R2, PPP6C, PRDM1, PRDM14, PREX2, PRKAR1A, PRKCI, PRKD1, PRKN, PRPF8, PTCH1, PTEN, PTP4A1, PTPN11, PTPRD, PTPRS, PTPRT, RAB35, RAC1, RAC2, RAD21, RAD50, RAD51, RAD51B, RAD51C, RAD51D, RAD52, RAD54L, RAF1, RARA, RASA1, RB1, RBM10, RECQL, RECQL4, REL, REST, RET, RHEB, RHOA, RICTOR, RIT1, RNF43, ROS1, RPS6KA4, RPS6KB2, RPTOR, RRAGC, RRAS, RRAS2, RTEL1, RUNX1, RXRA, RYBP, SCG5, SDHA, SDHAF2, SDHB, SDHC, SDHD, SERPINB3, SERPINB4, SESN1, SESN2, SESN3, SETD2, SETDB1, SF3B1, SH2B3, SH2D1A, SHOC2, SHQ1, SLFN11, SLX4, SMAD2, SMAD3, SMAD4, SMARCA2, SMARCA4, SMARCB1, SMARCD1, SMARCE1, SMO, SMYD3, SOCS1, SOS1, SOX17, SOX2, SOX9, SPEN, SPOP, SPRED1, SPRTN, SRC, SRSF2, STAG2, STAT3, STAT5A, STAT5B, STAT6, STK11, STK19, STK40, SUFU, SUZ12, SYK, TAP1, TAP2, TBX3, TCF3, TCF7L2, TEK, TENT5C, TERT, TET1, TET2, TFE3, TGFBR1, TGFBR2, TMEM127, TMPRSS2, TNFAIP3, TNFRSF14, TOP1, TP53, TP53BP1, TP63, TRAF2, TRAF7, TRIP13, TSC1, TSC2, TSHR, U2AF1, UGT1A1, UPF1, USH2A, USP8, VEGFA, VHL, VTCN1, WT1, WWTR1, XIAP, XPO1, XRCC2, YAP1, YES1, ZFHX3, ZNRF3, ZRSR2

The following genes are covered by the DNA panel but are excluded from CNV analysis:

HIST2H3D, HIST2H3C, HIST1H3A, HIST1H3D, HIST1H3E, HIST1H3F, HIST1H3G, HIST1H3H, HIST1H3I, HIST1H3J, HLA-A, HLA-C, H3F3B

A detailed description of the genomic regions considered by the CNV analysis is provided in the SOPHiA DDM™ platform as a downloadable file.

8.5.2 Technical overview of the analysis

The whole-gene amplification and deletion detection algorithm involves five main steps after read alignment.

Step 1: Raw coverage calculation

The raw coverage signal is obtained by performing the count of DNA fragments overlapping with genomic regions targeted by the panel (in the context of CNV detection, target regions are defined as a continuous region covered by probes).

Step 2: Coverage normalization and gene-level copy number quantification

The raw coverage signal of each sample is processed to compensate for GC biases. The resulting target region coverage levels are normalized per sample, and across samples within the same batch. For each gene, a gene-level copy number is obtained by averaging the normalized coverage across all regions belonging to the gene.

Step 3: Sex determination

The sex of the patient is determined automatically from the ratio of coverage over the Y chromosome to that over the autosomes. This information is used to adjust the gene-level copy number signal in genes of the X chromosome.

Step 4: QC and sample rejection

The following QC metrics are computed and used to reject samples from the analysis:

- Average coverage per region: average number of DNA fragments mapped to the target regions analyzed by the CNV module. The acceptance criterion is >= 30.
- Residual noise: measure of noise in the normalized coverage signal. The acceptance criterion is <= 0.25.

Step 5: CNV calling

Whole-gene amplification or deletion calls are obtained by applying the following thresholds to the gene-level copy numbers estimated in the previous step:

- Amplification: if gene-level copy number is larger than 3.25.
- Deletion: if gene-level copy number is lower than 1.25.

For samples identified as male, the copy-number thresholds for reporting amplifications and deletions in chromosome X (in non-pseudoautosomal regions) are divided by a factor 2 (i.e. 1.625 and 0.625 for amplification and deletion respectively). The gene copy-number levels of rejected samples cannot be confidently estimated, and these samples are reported as having undetermined CNV status.

Additional information about the methods is provided in the CNV PDF report.

8.5.3 Description of the results

The CNV analysis outputs a list of tertiary-annotated whole-gene amplification and deletion calls detected in the DNA sample. For each gene analyzed the key results of the analysis are:

- Copy number: numerical value reporting the average copy number of the gene relative to the average copy number of the panel
- CNV status (also referred to as “Type”):
    - Amplification: genes with gene-level copy number larger than 3.25
    - Deletion: genes with gene-level copy number smaller than 1.25
    - Normal: genes with gene-level copy number in the range [1.25 – 3.25]
    - Undetermined: for genes in samples rejected from the analysis

Gene amplifications reported with copy number between 3.25 and 6 could reflect copy number gains in high tumor content samples. Only calls associated to a copy number larger than 6 should be considered as high confidence gene amplification calls.

CNV analysis also computes a set of sample-level quality assessment (QA) metrics which are used to reject samples from the analysis:

- Average coverage per region: Average number of DNA fragments mapped to the target regions analyzed by the CNV module. The acceptance criterion is >= 30.
- Residual noise: Measure of noise in the normalized coverage signal. The acceptance criterion is <= 0.25.
- Noise status: Overall QA status that indicates if the data quality meets the criteria for CNV analysis. The possible values are:
    - Passed: the sample is eligible for CNV analysis
    - Rejected: the sample is rejected from CNV analysis

The results are made available via the SOPHiA DDM™ platform as well as downloadable files.

The following table provides an overview of the downloadable output files produced by this module. The Description of the CNV call attributes section provides a description of the variant attributes present in the downloadable CNV table file.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
<GENE PANEL NAME>-CNV-Report-summary.pdf	Run	Run level CNV report including a description of the methods and limitations, the QA metrics and CNV calls for all samples included in the run, the list of regions in scope of the analysis.
<GENE PANEL NAME>-CNV-Report-Sample-<SAMPLE ID >.pdf	Sample	Sample specific CNV report including the CNV calls and a plot of the normalized coverage across the panel.
full_CNV_table.txt	Sample	List of CNV calls with tertiary annotation in text format.
gene_cnv_table.txt	Sample	List of CNV calls with tertiary annotation in text format complemented with gene related information.
<GENE PANEL NAME>-region-map .txt	Resource File	List of genomic regions in scope of the analysis, including region name, genomic coordinates and associated transcript annotation.

8.5.4 Precautions and limitations

- The copy number levels reported by the CNV module quantify the copy number in the DNA sample and thus depend on the sample tumor content. At low tumor content, the signal from somatic CNVs may be diluted by the germline DNA possibly resulting in False Negatives.
- The copy number levels reported by the CNV module quantify the copy number relative to the average copy number across all genes included in the panel.
- The threshold used for calling gene-amplifications is optimized to reduce the occurrence of false negatives. In high tumor content samples, copy number gains can be reported as amplifications.
- The CNV module is designed to detect copy number changes affecting entire genes. The CNV module does not detect copy number variations that affect only parts of genes (i.e. exon-level CNV events). Exon-level copy number events affecting a large portion of the gene can result in whole-gene CNV calls.
- The CNV module uses cross-sample coverage normalization. If many samples share the same CNV in a given gene, this can bias the analysis for that gene.
- The CNV module is only applied to batches including at least 8 DNA samples. If the total number of non-rejected DNA samples in the batch is below 4, the whole batch is rejected. The performance of the module is expected to improve with an increased number of samples.

- For optimal performance of the Gene Amplification module, all the samples included in the batch must be processed under the same laboratory conditions.
- Processing of highly degraded FFPE samples may produce poor quality coverage signals, which can increase the risk of false positives, false negatives, or sample rejection.
- Genes with extreme GC content may be associated with poor quality coverage signals, making these regions more prone to false positives and false negatives.

8.5.5 Description of the CNV call attributes

The following table provides a definition of the CNV call attributes reported in the full_CNV_table.txt and gene_cnv_table.txt files.

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
Genomic description (hg19)	SVTYPE	The type of CNV (DUP, DEL or CNV, where CNV represents an undetermined CN status).
	refGenome	Reference genome used for variant calling.
	chromosome	Chromosome.
	seqName	Chromosome number.
	inner_start	5' genomic position indicating where the CNV event starts (first genomic position in which NGS data directly supports the presence of the CNV).
	inner_end	3' genomic position indicating where the CNV event ends (last genomic position in which NGS data directly supports the presence of the CNV).
	inner_SVLEN	Size of the CNV event defined using "inner_" genomic coordinates.
	outer_start	5' genomic position indicating where the CNV may start (inferred based on CNV results from nighbouring genes targeted by the panel).
	outer_end	3' genomic position indicating where the CNV may end (inferred based on CNV results from nighbouring genes targeted by the panel).
	outer_SVLEN	Size of the CNV event defined using

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
		"Outer_" genomic coordinates.
	chrom_overlap	Overlap between the CNV event and the chromosome to which it belongs.
	in-ner_genes_count	Number of genes from the panel which are nested in the CNV event.
Gene Information	gene	HGNC gene symbol.
	OMIM	OMIM ID.
	gene_strand	strand the gene is found on either + / -.
	exon_span	List of all genomic regions (see CNV PDF report) in scope of CNV analysis (associated to the gene) named based on a predefined gene transcript.
	kit_exons	Number of genomic regions in scope of CNV analysis (associated to the gene) labled as exons in exon_span
	covered_exons	Total number of exons targeted by the gene panel / Total number of exons in scope of CNV analysis
	gene_complete	Indicates if CNV analysis covers all the exons targeted by the gene panel
Signal Quantification	CN	Copy number value
Transcript Annotation	refSeqId	Transcript symbol in RefSeq database
	refSeqIdVersion	Transcript version in RefSeq database
Variant classification database	annotationdb_id	Variant ID in annotation database (multiple IDs are shown in case of multiple matches)
	match_status	Match status (Exact vs Partial) for each annotation database entry (annotationdb_id) associated to the CNV event
	overlap_metric	Fraction of CNV event overlap with each annotation database entry (annotationdb_id) associated to the CNV event
	public_id	dbVAR ID associated to each annotation database entry (annotationdb_id) associated to the CNV event

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
	dbvar_CLN-SIG_summary	Counts of pathogenic assertions in dbVAR IDs associated to the CNV event
	gno-mAD_AF_glob-alpop	GnomAD global population allele fre-quency of the CNV event
	gnomAD_cod-ingConsequence	Coding consequence of the CNV event
Internal reference	id	internal ID for a given variant greater than or equal to 1. found in gene table maybe a duplicate. Same as sgid. Interna ld. Fini
	id_cnv	Internal variant ID
	parent_id_cnv	Internal ID of larger structural variant in which the variant is found
	id_gene	internal ID for a given gene greater than or equal to 1. Inernal id full stop
	sgid	Variant ID in annotation database (alter-native representation of annotation_id)
	sgid_cnv	The same as SOPHiA ID but specific for CNVs
	sgid_gene	internal SOHPiA ID - alphanumeric hash unique for each gene that can be used to match variants across samples and runs
	tx_id	Transcript ID in annotation database
	region_span	Internal ID of genomic regions in scope of CNV analysis
	inner_start_norm	Rounded version of inner_start
	outer_start_norm	Rounded version of outer_start
	inner_end_norm	Rounded version of inner_end_norm
	outer_end_norm	Rounded version of outer_end_norm

8.6 DNA analysis: exon-level CNV detection via MUSKAT™

8.6.1 Analysis purpose

Partial-gene copy number alterations, such as exon-level deletions or duplications, can have significant functional consequences in cancer. Tumor suppressor genes are particularly sensitive to loss-of-function events, and even deletions affecting a subset of exons can disrupt protein function and drive tumorigenesis. Detecting these subgenic events is therefore important for accurately assessing gene inactivation, which is critical for making clinically relevant interpretations from cancer genomic profiles.

The gene-level and exon-level (GLEL) CNV analysis complements the whole-gene amplification & deletion CNV analysis by performing CNV detection at the exon-level resolution. This module is designed to: i) detect copy number gains and losses affecting only part of a gene, and ii) refine whole-gene CNV calls by determining whether a reported whole-gene event involves only specific exons.

The GLEL CNV analysis is applied to a subset of 49 genes covered by the whole-gene amplification & deletion CNV analysis:

APC, ARID1A, ATM, BAP1, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CDK12, CHEK1, CHEK2, DICER1, EGFR, EPCAM, FANCA, FANCD2, FANCL, FH, FLCN, MET, MLH1, MRE11, MSH2, MSH6, NBN, NF1, NF2, PALB2, PMS2, PPP2R2A, PTCH1, PTEN, RAD51B, RAD51C, RAD51D, RAD54L, RB1, SDHA, SDHB, SDHC, SMARCA4, SMARCB1, STK11, SUFU, TP53, TSC1, TSC2, WT1

8.6.2 Technical overview of the analysis

The technical details of the GLEL module are documented in the GLEL Manual (SG-08061), which describes the module independently of the application. The table below provides the parameter values used for this specific application.

PARAMETER ID	NAME	DESCRIPTION	VALUE
QA1	GLEL coverage threshold	Samples with an “Average nb. fragments in genomic regions” lower than the value will be rejected.	30
QA2	GL residual noise threshold	Gene-level CNV analysis residual noise threshold. Samples with a value exceeding this threshold will be rejected.	0.25
QA3	GLEL gene residual noise threshold	Gene-level and exon-level CNV analysis gene residual noise threshold. Genes with a value exceeding this threshold will be rejected.	0.15

PARAMETER ID	NAME	DESCRIPTION	VALUE
QA4	GLEL sample residual noise threshold	Gene-level and exon-level CNV analysis sample residual noise threshold. Samples with a value exceeding this threshold will be rejected.	0.11
TH1	GL gain threshold	Gene-level coverage signal threshold for gain. Genes with a coverage signal greater than or equal to this value will be reported as “Gain”	3.25
TH2	GL suspected loss threshold	Gene-level coverage signal threshold for suspected loss. Genes with a coverage signal lower than or equal to this value will be reported as “Suspected loss”	1.25
TH3	GL loss threshold	Gene-level coverage signal threshold for loss. Genes with a coverage signal lower than or equal to this value will be reported as “Loss”	0.9

8.6.3 Description of the results

The GLEL module generates a CNV variant table containing the following fields for each gene analyzed:

- **Gene-level status:** Indicates whether the gene is affected by a whole-gene event or an exon-level event. The possible statuses are:
  - **Gain**: Tumor cells carry a gene-level gain.
  - **Loss**: Tumor cells carry a homozygous gene-level deletion (i.e., zero copies of the gene are present in tumor DNA).
  - **Suspected loss**: Tumor cells are suspected to carry a homozygous gene-level deletion. In samples with high tumor content and/or high tumor ploidy, a heterozygous gene-level deletion may be sufficient to trigger a Suspected loss call.
  - **Exon-level:** The gene is not affected by a whole-gene CNV but carries an intragenic CNV.
  - **Normal**: No gene-level CNVs nor intragenic CNVs are detected.
  - **Rejected**: The gene-level status cannot be computed due to insufficient data quality.

- **Exon-level status:** Indicates any exon-level events affecting the gene and specifies their type**.** The possible statuses are**:**
  - **Normal**: No exon-level CNVs are detected.

    - **Exon-gain**: The gene is affected by an exon-level gain. The regions of the gene impacted by the event are identified and reported.
    - **Exon-loss**: The gene is affected by an exon-level loss. The regions of the gene impacted by the event are identified and reported.
    - **Suspected exon-loss**: The gene is affected by an exon-level CNV. The regions of the gene suspected to be impacted by a loss are identified. The biological interpretation of the exon-level CNV has low confidence. The gene could be affected by an exon-gain, and not by an exon-loss. For details, see SG-08061.
    - **Uncharacterized:** The gene is affected by an exon-level CNV. The regions of the gene impacted by the event could not be identified. The gene is affected either by an exon-gain or by an exon-loss.
    - **Complex rearrangement**: The exon-level coverage signal reflects the presence of multiple exon-level CNV events.
    - **Rejected**: The exon-level status cannot be computed due to insufficient data quality.
- **Segment regions, coverage levels and interpretations:** Reports and interprets the coverage levels values measured within a gene.
- **Gene QA status:** Reports if the exon-level coverage signal of the gene of interest is of insufficient quality. If Gene QA status is Rejected, the gene is rejected.

The current version of the GLEL-CNV module does not support tertiary annotation. In all outputs generated by the module, exon nomenclature follows an internal convention. The GLEL-CNV_regions.txt file, which can be downloaded from the module, provides a table that enables interpretation of exon nomenclature used.

Additionally, the GLEL module computes the following quality metrics:

- **Nb. fragments:** Total number of DNA fragments (in millions) available for CNV analysis. This QA metric is provided for information purposes only.
- **Average nb. fragments in genomic regions:** Average number of DNA fragments mapped to the target regions analyzed by the GLEL CNV module. The acceptance criterion is >= 30.
- **Coverage normalization status:** This QA metric reports whether the GLEL CNV module could not reliably compute the gene-level coverage signal, due to an abnormal coverage profile. If this QA metric is Fail, the QA status of the sample is Rejected.
- **Gene-level analysis residual noise:** Measure of noise in the data used to compute gene-level coverage signal. The acceptance criterion is <= 0.25.
- **Exon-level analysis residual noise:** Measure of noise in the data used to compute exon-level coverage signal. The acceptance criterion is <= 0.11.

- **QA status:** Indicates the outcome of the GLEL CNV analysis, as either successful completion or rejection due to insufficient data quality. The possible sample QA statuses are:
    - **Pass**: All QA metrics meet the acceptance criteria.
    - **Rejected**: NGS data are deemed of insufficient quality.

The results are made available in the SOPHiA DDM™ platform via downloadable files. The following table provides an overview of the output files produced.

FILE NAME	RUN-vs-SAMPLE SPECIFIC OUTPUT	DESCRIPTION
GLEL -CNV_results.tsv	Run	Text file containing GLEL CNV results for all samples in the run.
GLEL -CNV_QC.tsv	Run	Text file containing GLEL QC metrics for all samples in the run.
GLEL -CNV_Report.pdf	Run	PDF report describing GLEL CNV results for all samples in the run.
GLEL -CNV_Report_ sample_<SAMPLE ID>.pdf	Sample	PDF report describing GLEL CNV results for individual samples.
GLEL-CNV_regions.txt	Resource File	List of regions in scope for GLEL CNV analysis, including region name, genomic coordinates and associated transcript annotation.

8.6.4 Precautions and limitations

- The copy number levels reported by the module quantify the copy number in the DNA sample and thus depend on the sample tumor content. At low tumor content, the signal from somatic CNVs may be diluted by the germline DNA possibly resulting in False Negatives.
- The module reports copy number levels, which are quantified relative to the average copy number across all genes in the panel.
- The ability to detect an intragenic CNV within a gene of interest depends on:

    - The size of the CNV (i.e., the number of genomic regions affected by the CNV).
    - The size of the gene (i.e., the total number of genomic regions considered by the GLEL CNV module for the gene of interest).
      Small intragenic CNVs occurring within small genes may be missed or may cause the gene of interest to be excluded from the analysis.
- The ability to detect CNVs depends on the magnitude of the event. For example, the ability to detect a copy number change from 2 to 3 (gain of an extra copy) is lower compared to detecting a change from 2 to 1 (heterozygous deletion).
- The GLEL CNV module uses cross-sample coverage normalization. If many samples share the same CNV in a given gene, the analysis for that gene can be biased.
- The GLEL CNV module is only applied to batches including at least 8 DNA samples. If the total number of non-rejected DNA samples in the batch is below 4, the whole batch is rejected. The performance of the module is expected to improve with an increased number of samples.
- For optimal performance of the module, all the samples included in the batch must be processed under the same laboratory condition.
- Processing of highly degraded FFPE samples may produce poor quality coverage signals, which can increase the risk of false positives, false negatives, or sample rejection.
- Genomic regions with extreme GC content may be associated with poor quality coverage signals, making these regions more prone to false positives and false negatives.
- The GLEL CNV module only considers a subset of genomic regions targeted by the DNA panel (refer to “GLEL-CNV _regions.txt” downloadable file).

8.7 DNA analysis: microsatellite instability (MSI) detection with MUSTARD

8.7.1 Analysis purpose

Microsatellite instability (MSI) is a molecular phenotype exhibiting the accumulation of insertion-deletion mutations at short tandem repeats or microsatellites in tumor cells due to DNA mismatch repair (MMR) deficiency. The MSI analysis aims to assess microsatellite instability by analyzing NGS data from a predefined set of homopolymer regions targeted by the DNA panel.

8.7.2 Technical overview of the analysis

The detection of MSI relies on the observation of instability in the length of a set of 117 homopolymers targeted by the DNA panel SG_MSKIMPACTFLEX_v1. The MSI detection algorithm involves four main steps after read alignment.

Step 1: Computation of locus-specific homopolymer length distributions and coverage metrics

The coverage or each homopolymer locus is assessed and the following quantities are measured:

- **Effective Coverage**: Count of reads retained for MSI calculation. Only reads covering the entire homopolymer length plus a 3bp anchor before and after the homopolymer without any mismatch are considered. This quantity is measured separately for forward and reverse strands.
- **True Coverage**: Total read count covering the homopolymer loci including reads discarded from MSI analysis.

For each read retained for MSI calculation, the homopolymer length is measured and used to create a locus-specific homopolymer length distribution. The locus-specific homopolymer length distribution is determined for the forward and reverse strand separately as the two may differ depending on the sequencing direction.

Step 2: Computation of locus-specific MSI scores using a reference distribution

For each homopolymer locus, a locus-specific MSI score is calculated independently for forward and backward homopolymer length distributions. This metric assesses the deviation of the locus-specific homopolymer distribution measured in the sample compared to a locus-specific reference homopolymer length distribution. This deviation considers the proportion of reads that do not match the reference distribution (directly related to the tumor content of the sample) and the difference in homopolymer length measured at the locus compared to the reference. The locus-specific MSI score ranges from 0

(perfectly stable) to 1 (maximally unstable). The reference homopolymer length distributions were established using a set of microsatellite stable samples from various cancer types.

The locus-specific MSI score is the average of the locus-specific MSI score measured for forward and reverse strands. Locus-specific MSI scores are computed only for homopolymer loci with effective coverage > 50x in at least one of the strands.

Step 3: Sample-specific MSI score and sample-specific QA metrics

Locus-specific coverage metrics and locus-specific MSI scores are aggregated into the following sample-specific metrics.

- **Num_loci_used**: Number of homopolymer loci for which a locus-specific MSI score can be computed (i.e. number of homopolymer loci with effective coverage >50x).
- **Pct_loci_used**: The percentage of homopolymer loci for which a locus-specific MSI score can be computed (i.e. number of homopolymer loci with effective coverage >50x divided by the total number of homopolymer loci).
- **Mean_effective_coverage**: Homopolymer effective coverage averaged across all loci
- **Mean_coverage**: Mean effective coverage formatted for reporting purposes (rounded to the nearest integer).
- **Mean_true_coverage**: Homopolymer true coverage averaged across all loci.
- **Median_score**: Sample specific MSI score computed by taking the median of locus-specific MSI scores.
- **Mean_score**: Average of locus-specific MSI scores.

Step 4: MSI status determination

The MSI status of a particular sample is established based on the median MSI score and on the percent of loci used for MSI analysis. The possible statuses are defined below:

1. **Reject**: When the percentage of the homopolymer loci retained (Pct_loci_used) for MSI calculation is below 70%, the sample is rejected due to insufficient data quality.
2. **MSS** (Microsatellite Stable): When the median MSI score is below 0.01 and the percentage of loci used for calculation is above 70%, the sample is considered “stable “.
3. **MSI-High** (High confidence Microsatellite Instability, also referred to as **MSI-H**): When the median MSI score is greater than 0.02 and the percentage of loci used for calculation is above 70%, the sample is considered “unstable”.
4. **Indeterminate** (also referred to as Low confidence Microsatellite Instability **MSI-LC**): When the median MSI score is in the range [0.01, 0.02] and the percentage of loci used for calculation is above 70%, the sample is considered “Indeterminate”.

This status reflects samples with intermediate MSI scores associated with an increased risk of misclassification. We recommend confirmation of the MSI status via an orthogonal method for samples with an indeterminate call.

Notes:

- In case of rejected sample(s) due to low coverage, we recommended re-processing the sample(s).
- The thresholds for defining MSS, MSI-LC and MSI-H were set based on 115 clinical samples with known and concordant status for MSI (MSS/MSI-H) and immunochemistry status (proficient mismatch repair / deficient mismatch repair) and coming from various cancer types.

8.7.3 Description of the results

For each sample analyzed, the main results computed by MSI analysis are:

- **Pct loci used:** QA metric measuring the percentage of homopolymer loci fulfilling the coverage criteria to be considered for MSI score determination. When lower than 70%, a sample is rejected from MSI analysis.
- **MSI score:** The microsatellite instability score of the sample, which ranges between 0 and 1.
- **MSI status**: The microsatellite instability status of the sample. Four outcomes are possible:
    - **Reject**: When the sample is rejected from MSI analysis due to insufficient data quality
    - **MSS** (Microsatellite Stable): Non-rejected samples with MSI score below 0.01.
    - **MSI-High** (High confidence Microsatellite Instability): Non-rejected samples with MSI score greater than 0.02.
    - **Indeterminate** (also referred to as Low confidence Micro Satellite Instability **MSI-LC**): Non-rejected samples with intermediate MSI score comprised in the range [0.01, 0.02].

The following additional sample-specific metrics, defined in the previous section, are reported for information purposes: Number of loci used, Mean true coverage, Mean effective coverage, Mean coverage, Mean score.

The results from the MSI analysis are available for download through the SOPHiA DDM™ platform. The following table provides an overview of each output file.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
<GENE PANEL NAME>-MSI-Report.pdf	Run	PDF report including the main MSI analysis results, MSI QA metrics and visualization of MSI scores at individual homopolymer loci
MSI_summary.csv	Run	Table with MSI analysis results and MSI QA metrics

8.7.4 Precautions and limitations

- The MSI analysis algorithm was developed using a cohort composed primarily of colorectal, endometrial, and stomach cancers. The user shall interpret with care the score for other cancer types.
- The presence of germline polymorphisms at the considered homopolymer sites increases the risk of false positive MSI calls.
- For runs with noise levels significantly higher than those from the run used to define the reference homopolymer length distributions, the MSI score may be overestimated, increasing the risk of false positive calls.
- Low tumor content or high tumor heterogeneity may increase the risk of false negative results. The limit of detection (LOD) was established at approximately 20% tumor content, based on an *in-silico* dilution study performed using six MSI positive FFPE tumor samples.

8.8 DNA analysis: tumor mutational burden (TMB) measurement

8.8.1 Analysis purpose

The Tumor Mutational Burden (TMB) is a measure of the total number of somatic mutations present in a tumor genome, expressed as a number of mutations per megabase (mut/Mb). A high TMB value reflects an increased number of tumor-specific mutations, which can lead to the generation of neoantigens and potentially enhance tumor immunogenicity, making TMB an important biomarker for predicting response to immune checkpoint inhibitors.

8.8.2 Technical overview of the analysis

TMB is measured only in coding regions (CDS) within the target areas used for SNV and INDEL analysis, excluding problematic regions, for a total of 1.28 Mb. Specifically, TMB is determined by counting the number of somatic variants detected in CDS regions that reach a minimum molecular coverage of 200x. Variants considered for TMB are derived from the variant calling results, and the calculation is performed in three main steps.

Step 1: Data-driven determination of genomic regions used for TMB analysis

The molecular coverage, assessed based on CUMIN™ groups, of the genomic regions considered for TMB analysis is evaluated. Only regions with molecular coverage ≥200x are classified as *callable TMB regions* and retained for TMB calculation; regions not meeting this threshold are excluded.

Two quality assessment metrics are generated during this step:

- **Callable_length_Mb**: The total length of *callable TMB regions*, reported in megabases.
- **Callable_length_fraction**: The fraction of the predefined TMB genomic regions that meet the callable criteria based on molecular coverage.

Step 2: Identification of somatic variants to be included in the TMB calculation

The subset of variants to be included in the TMB calculation is selected from the full list of high-confidence variants identified during the SNV and INDEL analysis. The following exclusion criteria are applied:

- **Callable region filter**: Variants exclusively located within the callable TMB regions (as defined in Step 1) are retained; variants outside these regions are excluded.

- **Germline variant exclusion**: Putative germline variants are identified and excluded using the following approaches:
    - Reference to public databases to exclude common polymorphisms.
    - Application of a Hidden Markov Model (HMM) that leverages variant fraction information to identify additional germline-origin variants, which are not documented in public databases as being polymorphisms.
- **Driver mutation exclusion**: Known driver mutations defined as those with more than 50 entries in the COSMIC database are excluded.
- **Multi-nucleotide variants (MNV) exclusion:** MNVs (or phased variants) are excluded.
- **Variant fraction filter**: Variants with an observed variant fraction outside the range of 2%–90% are excluded.

Step 3: TMB calculation

The TMB score is calculated by dividing the number of variants retained after Step 2 by the total size of the callable TMB regions determined in Step 1. The result is expressed as the number of mutations per megabase (Mb).

8.8.3 Description of the results

For each sample, the main results produced by TMB analysis are:

- **TMB (all):** TMB score calculated by considering the total number of high-confidence variants inferred to be of somatic origin (excluding driver mutations) with an observed variant fraction between 2% and 90%. The value is reported as the number of mutations per megabase (Mb).

TMB (non-synonymous): Same as TMB (all) but restricted to non-synonymous variants.

The TMB analysis also computes and reports the following quality metrics:

- **Variants**: Total number of variants retained for TMB calculation.
- **Variant_non_synonymous**: Total number of non-synonymous variants retained for TMB calculation.
- **callabe_length_Mb**: Total size of genomic regions considered for TMB calculation (in megabases).
- **callabe_length_fraction**: Proportion of genomic regions in scope of TMB analysis that meet the molecular coverage criteria required to be effectively used for TMB calculation.

The key TMB results are available in the SOPHiA DDM™ platform. Additional metrics are available via downloadable files. The following table provides an overview of the output files produced by TMB analysis.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
TMB.txt	Sample	Table with main TMB analysis results
full_variant_table_TMB.txt	Sample	SNV INDEL full variant table annotated to indicate which variants are retained for TMB calculation

8.8.4 Precautions and limitations

1. The TMB analysis does not compute a TMB status.
2. The TMB analysis reports quality metrics but does not include a formal sample rejection mechanism.
3. The TMB may be overestimated in samples from patients with ethnicities underrepresented in public databases (DBs)
4. The TMB may be underestimated in samples with tumor content higher than 90%.
5. The TMB may be underestimated in samples where a low tumor content leads to false negative SNV/INDEL results.
6. In samples featuring specific combinations of large-scale copy number aberrations and tumor contents, TMB may be underestimated due to misclassification of somatic variants as germline variants.
7. Low NGS data quality causing false positive calls can lead to TMB overestimation.

8.9 DNA analysis: HRD genomic integrity (GI) detection with GIINGER™

8.9.1 Analysis purpose

The HRD genomic instability (GI) analysis aims to detect HRD by assessing the degree of genomic instability. Functional homologous recombination repair is necessary for the error-free repair of double strand breaks and for maintaining genome integrity. A consequence of HRD is the loss of genomic integrity via the accumulation of genomic aberrations due to the cell inability to repair double strand breaks. The GI analysis assesses these genomic aberrations using WGS. More specifically, our method uses a deep learning algorithm that has been trained to recognize patterns of genomic instability in the WGS coverage profile. The algorithm analyzes low-pass WGS data to generate a general GI index, which reflects the level of genomic integrity. In addition, a GI status applicable to ovarian cancer samples is computed.

8.9.2 Technical overview of the analysis

During the GI analysis, the following algorithmic steps are performed sequentially for each sample of interest: first the low-pass WGS data are pre-processed and undergo quality assessment, next the GI index is computed, and lastly a proposed GI status is assigned to the sample. These steps are described in the sections below.

Step 1: Data preprocessing and sample quality assessment

1. WGS paired-end reads are mapped to the human reference genome and processed to trim adaptors and low-quality base calls. The WGS coverage profile is computed and normalized. When NGS data are generated by combining DNA WGS with DNA capture, this preprocessing step automatically excludes from the analysis the genomic regions that are enriched via capture hybridization.

2. The normalized WGS coverage profile undergoes QA based on the metrics defined in the next section.

Step 2: GI index calculation

The GI index is obtained by processing the normalized WGS coverage profile using a proprietary deep learning algorithm that has been trained to recognize patterns of genomic instability.

Step 3: GI status determination

A GI status that applies to ovarian cancer samples only is determined by combining the sample QA status and the GI index. The possible GI statuses are described in the next section.

Step 4: GI warning determination

A GI warning signaling the presence of sample properties known to bias GI analysis (e.g., insufficient tumor content) is determined based on GI QA metrics. The possible GI warnings are described in the next section.

8.9.3 Description of the results

For each sample, the main results produced by the GI analysis are:

- **QA Status**: The quality assessment status of the NGS data used for GI analysis. One of the following GI QA statuses is assigned to the sample:
    - **Pass**: The quality of the data is sufficient to compute a GI index and status.
    - **Fail**: The quality of the data does not meet the criteria required to reliably compute a GI index. A QA status Fail can result from processing of highly degraded DNA samples and/or inappropriate execution of the NGS workflow.
- **GI Index**: The GI index is a scalar value, ranging between -20 and 20, which quantifies the level of genomic integrity. High GI indices reflect low levels of genomic integrity. Low GI indices reflect high levels of genomic integrity.
- **GI Status**: The genomic integrity status of a sample. The relevance of the GI status has only been assessed in the context of ovarian cancer. Three outcomes are possible:
    - **GI positive**: Ovarian cancer samples with a GI index greater than or equal to 0.
    - **GI negative**: Ovarian cancer samples with a GI index lower than 0.
    - **GI rejected**: Ovarian cancer samples with QA status Fail that are discarded from the GI analysis. The GI index associated with samples with GI rejected status may be inaccurate and is reported for information purposes only.
- **GI Warning**: The warning type triggered during the GI analysis. If no warnings are triggered, this value is empty. Two warnings are possible:
    - **Low purity**: Indicates that the sample tumor content is low, potentially causing an underestimation of the GI index.
    - **Flat profile**: Indicates that the sample tumor content is extremely low or that the tumor sample does not feature large scale copy number changes.

The GI analysis also computes a set of quality assessment (QA) metrics that are used to establish the GI QA status and GI warnings:

- **Total nb. of fragments**: The total number of DNA fragments (paired-end reads) that are properly mapped.
- **Nb. WGS fragments**: The total number of DNA fragments available for the raw WGS coverage profile calculation. DNA fragments mapping to genomic regions that are enriched via capture hybridization, and thus excluded in the preprocessing step, do not contribute to the count. If the number of WGS fragments is lower than 4 million, the QA status is Fail.
- **Percentage WGS fragments**: Fraction of the total number of WGS fragments over the total number of fragments.
- **Residual noise**: The residual noise is computed by measuring the standard deviation of the normalized WGS coverage profile with respect to the smoothed WGS coverage profile. If the residual noise is larger than 0.15, the QA status is Fail.
- **Signal**: The signal is computed by measuring the standard deviation of the smoothed WGS coverage profile.
- **SNR**: The signal-to-noise ratio is defined as the ratio of signal with respect to residual noise.
- **Purity/ploidy ratio**: The ratio between sample tumor content and sample ploidy is estimated by measuring the strength of the signal induced in the normalized WGS coverage profile by a copy number change. If the purity/ploidy ratio is lower than 0.1 and the SNR is lower than 0.5, a flat profile warning is emitted. If the purity/ploidy ratio is lower than 0.15 and the signal is lower than 0.1, a low purity warning is emitted.

The GI results are available in the SOPHiA DDM™ platform. Results are also provided via downloadable files. The following table provides an overview of the output files produced by GI analysis.

FILE NAME	RUN vs. SAMPLE SPECIFIC OUTPUT	DESCRIPTION
<GENE PANEL NAME>-GI-Report.pdf	Run	PDF report including the main GI analysis results, GI QA metrics and visualization of normalized WGS coverage profile

FILE NAME	RUN vs. SAMPLE SPECIFIC OUTPUT	DESCRIPTION
<GENE PANEL NAME>-GI-Report_sample_<SAMPLE ID>.pdf	Sample	PDF report including the main GI analysis results, GI QA metrics and visualization of normalized WGS coverage profile
GI_status_table.txt	Run	Table with the main GI analysis results
GI_qc_table.txt	Run	Table with GI QA metrics and GI QA status

8.9.4 Precautions and limitations

- The biological relevance of the GI status has only been assessed for ovarian cancer samples.
- The limit of detection for genomic instability in ovarian cancer samples is a tumor content of 30%.
- The GIINGER™ algorithm has been developed to measure genomic integrity in ovarian cancer samples. Users shall interpret with care the significance of the score in other cancer types.
- The reported GI QA metrics and index are obtained by rounding up the full precision values used for QA and GI status calculations to the first or second decimal (depending on the metric).

8.10 RNA analysis: Gene fusion and exon-skipping detection with CARDAMOM

8.10.1 Analysis purpose

This analysis is designed to detect gene fusion and exon-skipping events, and to identify RNA alterations suggestive of kinase domain duplication.

Gene fusions, exon skipping, and kinase domain duplication events are RNA-level alterations that can result in aberrant or constitutively active proteins with oncogenic potential. Gene fusions occur when exons from two distinct genes are joined, typically as a result of chromosomal rearrangements. Exon-skipping events arise when one or more exons are omitted during mRNA splicing, while kinase domain duplication events occur when the region encoding a kinase domain is duplicated within the transcript.

To detect these RNA alterations, the assay targets specific regions within selected genes known to be recurrently affected in cancer, with regions defined according to the type of alteration analyzed. The assay applies dedicated detection strategies for gene fusions, exon-skipping events, and RNA alterations indicative of kinase domain duplication.

Gene fusion detection covers a total of 135 genes. The fusion detection strategy is partner-agnostic, allowing the identification of fusion transcripts involving any gene partner, provided that at least one partner gene is targeted by the assay. A complete list of genes covered by the panel is provided in the section Target regions for RNA fusion analysis. For each gene, the assay defines the set of targeted exons, indicating whether they are used for detection when the gene functions as a 5' or 3' fusion partner.

Exon-skipping detection focuses on predefined transcript regions where exon loss results in detectable junctions between non-consecutive exons. The exon-skipping events analyzed by this assay include:

- ALK (NM_004304): ex1–ex18; ex1–ex4
- AR (NM_001348061): ex3-ex4 (ARV7, cryptic exon inclusion)
- BRAF (NM_004333): ex1–ex11; ex1–ex9; ex2–ex9; ex3–ex11; ex3–ex9
- EGFR (NM_201282): ex1–ex8
- ERBB2 (NM_004448): ex15–ex17
- MET (NM_000245): ex13–ex15; ex14–ex16
- NFE2L2 (NM_006164): ex1–ex3; ex1–ex4
- NOTCH1 (NM_017617): ex1–ex28; ex20–ex28; ex2–ex28; ex2–ex29
- PDGFRA (NM_006206): ex7–ex10

RNA alterations indicative of kinase domain duplication are identified based on exon junctions supporting the presence of duplicated kinase-coding regions within the transcript. The RNA alterations analyzed by this assay include:

- BRAF (NM_004333): ex18–ex10
- EGFR (NM_005228): ex25–ex18
- FGFR1 (NM_015850): ex18–ex10

The detection method reports the observed exon junctions (e.g., *BRAF ex18→ex10*) that may indicate the presence of a kinase domain duplication. However, it does not determine whether the duplication spans the entire region of kinase domain.

8.10.2 Technical overview of the analysis

This analysis involves four main steps after RNA data preprocessing.

Step 1: Candidate event identification

Chimerically-aligned reads are analyzed to identify potential chimeric transcripts (i.e., gene fusions, exon-skipping events, and RNA alterations indicative of kinase domain duplication). Candidate gene fusions are detected by examining chimeric reads mapped to two distinct genes, while for a predefined list of exon-skipping events and kinase domain duplications, candidate variants are identified by evaluating chimeric reads mapped within a single gene.

Step 2: Calling

The validity of candidate events is evaluated using a Bayesian statistical model that estimates the posterior probability of each event representing a true biological signal rather than a sequencing artifact. The model integrates multiple parameters, including:

- **Coding consequence of the gene product**, to evaluate the predicted impact on transcript functionality.
- **Estimated break-point position** for the gene fusion, derived from the mapping and alignment of supporting reads.
- **Unique molecule count**, representing the number of distinct RNA molecules (as estimated based on CUMIN™ groups) supporting the detected junction.
- **Genomic context**, encompassing the proximity, orientation, and sequence characteristics of the involved regions.

Statistically significant events are called when supported by at least three CUMIN™ groups.

Multiple distinct breakpoints may be reported for the same gene pair, reflecting the potential presence of alternative fusion transcript variants. The resulting calls are then subjected to filtering to determine event confidence.

Step 3: Filtering

Variants are classified as high-confidence and low-confidence variants based on the following filters:

- **off-target:** Variants falling outside the targeted regions.
- **low-posterior-probability:** Variants with a posterior probability below 0.5, indicating that NGS data do not strongly support the presence of the event.
- **low-molecular-count:** Variants supported by fewer than 10 unique CUMIN™ groups.
- **same-gene-family:** Variants in which both affected regions occur within genes from the same gene family, which may result in false positives due to sequence similarity.
- **homologous-genes:** Variants involving highly homologous genes that are not annotated in the same family, which may result in false positives due to sequence similarity.
- **readthrough:** Variants where the affected regions correspond to adjacent genes annotated as “readthrough” in HGNC, representing transcriptional artifacts rather than true events.
- **intronic-mapping:** Variants in which at least one affected region ends in an intronic region, potentially reflecting splicing intermediates or misalignment.
- **pseudogene:** Variants in which at least one breakpoint or affected region maps to a pseudogene, which can produce false-positive signals due to sequence similarity with functional genes.

Each variant is evaluated against all defined filters. The variant is classified as high confidence only when no filter is triggered.

Step 4: Annotation

High-confidence and low-confidence events are annotated against curated databases such as COSMIC_SV and ChimerDB to provide biological and clinical context. Transcript prioritization is also applied to report relevant transcripts only.

8.10.3 Description of the results

The analysis for detecting RNA fusion, exon-skipping, and RNA alterations indicative of kinase domain duplication generates a list of tertiary-annotated variants detected in the RNA sample. The section RNA variant table: Field Descriptions in this manual provides a

detailed explanation of the variant attributes included in the downloadable variant table file.

RNA alterations indicative of kinase domain duplication are reported in the variant table with "type = exon.skipping". These events can be distinguished from true exon-skipping events because the reported 3' exon number is smaller than the reported 5' exon number. For example, the event *BRAF ex18–ex10* indicates a junction where Exon 18 is followed by Exon 10.

The results are made available through the SOPHiA DDM™ platform as well as in downloadable files.

The following table provides an overview of the output files produced.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
full_fusion_table.txt	Sample	Table listing detected RNA fusion, exon-skipping and RNA alterations indicative of kinase domain duplication

8.10.4 Precautions and limitations

- The detection of RNA fusions, exon-skipping, and RNA alterations indicative of kinase domain duplication events relies on sufficient transcript expression. Variants occurring in transcripts with low or absent expression may not be detected.
- Gene fusions with breakpoints located in the 3' UTR regions are not detected.
- Fusions involving IGH genes are only partly annotated due to the absence of a canonical reference transcript.
- Readthrough transcripts formed between adjacent genes located within 100 kb are not reported (these events typically result from normal transcriptional readthrough rather than genomic rearrangements).
- RNA alterations indicative of kinase domain duplication are identified based on detected junctions only. It cannot be concluded from the data that the full kinase domain is duplicated in the sample.

8.10.5 Target regions for RNA fusion analysis

Note: Exons denoted with an asterix "*" are also in the scope of DNA-based fusion and exon-skipping analysis. Asterisks next to genes indicate that at least one exon is also in the scope of DNA-based fusion and exon-skipping analysis.

GENE	REFSEQ TRANSCRIPT ID	EXONS TARGETED WHEN THE GENE IS 5' PARTNER	EXONS TARGETED WHEN THE GENE IS 3' PARTNER
ACVR2A	NM_001616		1,2,3
AKT1	NM_005163		2,3,4,5
AKT2	NM_001626	11	2,5
AKT3	NM_005465	6,7,8	2,3,4,9
ALK*	NM_004304		1,2,4,6,8,10,12,14,16,17, 18*,19*,20*,21,22, 23,26
AR	NM_000044	2,3,4,5,6,7,8	1
AR	NM_001011645		1
ARHGAP26	NM_015071		2,10,11,12
ARHGAP6	NM_006125		2
AXL	NM_021913	18,19,20	11
BCOR	NM_001123385	2,4,6,7,10,12,14,15	2,3,4,5,6,7,8,9,11
BCOR	NM_017745		4,8
BRAF*	NM_004333	2,3,7*,8*,10*,13,14,18	2,3,4,5,7,8*,9*,10*,11*,12,15,16
BRD3	NM_007371	9,10,11,12	
BRD4	NM_058243	10,11,12,13,14	2
CAMTA1	NM_015215	3	8,9,10
CCNB3	NM_033031		2,3,4,5,6,7
CCND1	NM_053056	2,3,4,5	1,2,3,4
CHMP2A	NM_014453	2,3,4	2,3,4
CIC	NM_015125	14,15,16,17,18,19,20	12
CRTC1	NM_015321	2,3,4	1
CSF1	NM_000757	5,6,7,8,9	2,3,4,5,6
CSF1	NM_172212	9	
CSF1R	NM_005211		11,12,13
CTNNB1	NM_001904		1
DNAJB1*	NM_006145	2*	1
EGF	NM_001963		16,17,18,19

GENE	REFSEQ TRANSCRIPT ID	EXONS TARGETED WHEN THE GENE IS 5' PARTNER	EXONS TARGETED WHEN THE GENE IS 3' PARTNER
EGFR*	NM_005228	24*,25*,26*	7,8*,9,14,15,16,17,18*,19,20
EPC1	NM_025209	9,10,11	
ERBB2	NM_004448	15,23,24,25,26	4,5,13,15,17
ERBB4	NM_005235		2,3,4,14,15,16,17,18,23
ERG	NM_004449		2,3,4,5,6,7,8,9,10
ESR1	NM_001122742	3,4,5,6,7,8	3,4,5,6,7,8
ESR1	NM_000125	2,3,4,5,6,7,8	1,5,6,7
ESRRA	NM_004451	2,3	
ETV1	NM_004956		3,4,5,6,7,8,9,10,11,12,13
ETV4	NM_001986		2,3,4,5,6,7,8,9,10
ETV5	NM_004454		2,3,7,8,9
ETV6*	NM_001987	1,2,3,4*,5*,6	2,3,4,5*,6*,7
EWSR1*	NM_005243	4,5,6,7*,8*,9*,10*,11*,12*,13*,14	8*
FGF1	NM_000800	4	2
FGFR1	NM_015850	12,17	2,3,4,5,6,7,8,9,10,11,17
FGFR2*	NM_000141	16*,17*,18	2,3,5,6,7,8,9,10
FGFR3*	NM_000142	16*,17*,18	3,5,8,9,10,11,12,13,14
FGR	NM_005248		2,3
FOS	NM_005252	4	
FOSB	NM_006732		1,2
FOXO1	NM_002015	2,3	1,3
FOXO4	NM_005938		2
FUS	NM_004960	3,4,5,6,7,8,9,10,11,13,14	
GLI1	NM_005269	4,5,6,7	4,5,6,7
GRB7	NM_005310		10,11,12
GREB1	NM_014668		2,3
HMGA2	NM_003483	2,3,4,5	
IGF1R	NM_000875		13,14,15
INSR	NM_000208	20,21,22	2,12,13,14,15,16,17,18,1

GENE	REFSEQ TRANSCRIPT ID	EXONS TARGETED WHEN THE GENE IS 5' PARTNER	EXONS TARGETED WHEN THE GENE IS 3' PARTNER
			9
JAK2	NM_004972	9,10,11,12,22	6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
JAK3	NM_000215	23	10,11,12,17,18,19
JAZF1	NM_175061	2,3,4	
KANSL1	NM_001193466	2,4	10,11,12
KIT	NM_000222	1	1,8
KRAS	NM_004985	1,2,3,4,5	1,2,3,4
MAML2	NM_032427	2	2,3
MAP2K1	NM_002755		2
MAP3K8	NM_005204		1,2,3,4,5,6,7,8
MAST1	NM_014975		7,8,9,18,19,20,21
MAST2	NM_015112	15,16,17	2,3,5,6
MBTD1	NM_017643	15,16,17	3
MDM2	NM_002392	2,4,6,8,10	5,9
MEAF6	NM_001270875	4,5	
MET*	NM_000245	2,13*	2,4,5,6,13,14*,15*,16,17
MGEA5	NM_012215		4,5,6,7,8,9,12,13,14,15
MKL2	NM_014048		11,12,13
MN1	NM_002430	2	1
MSMB	NM_002443	2,3,4	
MUSK	NM_005592	15	7,9,10,12,13,14
MYB	NM_001130173	7,8,9,11,12,13,14,15,16	
MYBL1	NM_001080416	8,9,10,11,12,13,14,15	
MYC	NM_002467		1,2
MYOD1	NM_002478	2	1,2
NCOA1	NM_147223	10,11,12,13,14,15	10,11,12,13,14,15
NCOA2	NM_006540	14	11,12,13,14,15,16,22

GENE	REFSEQ TRANSCRIPT ID	EXONS TARGETED WHEN THE GENE IS 5' PARTNER	EXONS TARGETED WHEN THE GENE IS 3' PARTNER
NCOA3	NM_006534	20	2,13,14,15,16
NFATC2	NM_012340	10	2,3,9
NFE2L2	NM_006164	5	1,2,3,4
NFIB	NM_001369458		5,10,11
NFIB	NM_005596	2,9	
NOTCH1	NM_017617	2,4,24,29,30,31	5,24,25,26,27,28,29
NOTCH2	NM_024408	5,6,7	24,25,26,27,28,29
NR4A3	NM_173200	8	2,3,4,5,7
NRG1	NM_013957	4,8	8
NRG1	NM_001159996		1,3,4,5
NRG1	NM_004495		1,2,3,4,5
NRG1	NM_013959	3	1
NRG1	NM_013962		1
NTRK1	NM_001007792		2
NTRK1*	NM_002529		1,2,3,4*,5,6,7,8*,9*,10*,11*,12*,13*,14
NTRK2*	NM_006180	11,14	4,5,6,7,8,9,10,11,12,13,14*,15,16,17,18
NTRK3	NM_001012338		18
NTRK3	NM_001007156		15
NTRK3	NM_002530	13,14,15,17	3,4,5,6,7,8,9,10,12,13,14,15,16
NUMBL	NM_004756		2,3
NUTM1*	NM_175741		2,3*,4,5,6
PAX3	NM_181459	3,5,6,7,8	2,4,8
PAX8*	NM_003466	2,6,7,8*,9*,10*	3
PDGFB	NM_002608		2,3
PDGFD	NM_025208	7	5,6

GENE	REFSEQ TRANSCRIPT ID	EXONS TARGETED WHEN THE GENE IS 5' PARTNER	EXONS TARGETED WHEN THE GENE IS 3' PARTNER
PDGFRA	NM_006206	7	10,11,12,13,14,15
PDGFRB	NM_002609		8,9,10,11,12,13,14
PHF1	NM_024165	10,11,12	1,2
PHKB	NM_000293	4	
PIK3CA	NM_006218		2,15
PKN1	NM_002741		10,11,12,13
PLAG1	NM_002655	1,2,3	1,2,3,4
PPARG	NM_015869		1,2,3,5
PRDM10	NM_020228		13,14
PRKACA	NM_002730	2,4,5,6,9	2,4,5,6,9
PRKACB	NM_182948	3,4,7,8,9	1,3,4,7,8,9
PRKCA	NM_002737		4,5,6,9,15
PRKCB	NM_002738		1,3,7,8,9
PRKCD	NM_006254	18	9,10,11,12,15
PRKD1	NM_002742		2,10,11,12,13
PRKD2	NM_016457		10,11,12,13
PRKD3	NM_005813		10,11,12,13
RAD51B	NM_133509	3,4,5,6,7,8,9	8
RAF1	NM_002880	4,5,6,7,8,9	2,4,5,6,7,9,10,11,12
RELA*	NM_021975	11	1,3*,4
RET*	NM_020975	8*,9*,10*,12,13,14	8*,9*,10*,11*,12*,13,14
RET*	NM_020630		2,4,6,8*,9*,10*,11*,12*,13,14
ROS1*	NM_002944		2,4,7,31*,32*,33*,34*,35*,36*,37
RSPO2	NM_178565		1,2,3
RSPO3	NM_032784		2
SS18	NM_001007559	4,5,6,8,9,10	2,3,4,5,6,10
SS18L1	NM_198935	1,2,3,8,9,10	1
STAT6	NM_001178078		1,2,3,4,5,6,7,15,16,17,18,19,20

GENE	REFSEQ TRANSCRIPT ID	EXONS TARGETED WHEN THE GENE IS 5' PARTNER	EXONS TARGETED WHEN THE GENE IS 3' PARTNER
TAF15	NM_139215	5,6,7,9	6,7
TCF12	NM_207036	4,5,6	
TERT	NM_198253	3,9,15	2,3,5,7,10,11,12
TFE3*	NM_006521	2,3*,4*,5*,6	2,3,4*,5*,6*,7,8
TFEB	NM_007162	9,10	1,2,3,4,5,6
TFG	NM_006070	3,4,5,6,7,8	6
THADA	NM_022065	24,25,26,27,28,29,30,31,36,37	
TMPRSS2*	NM_001135099	2*,3,4,5,6	
TMPRSS2*	NM_005656	2*,3,4,5,6	
USP6	NM_004505		1,2,3
VGLL2	NM_182645	2,3,4	1
WWTR1	NM_015472	3,4	3,4
YAP1	NM_001130145	2,3,4,5,6,7,9	1,2,3,4,8
YWHAE	NM_006761	5	5

8.10.6 RNA variant table: Field Descriptions

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
Genomic description (hg19)	type	Variant type (e.g., gene.fusion, exon.skipping)
	id	Variant identifier combining gene names and genomic breakpoints
	refGenome	Reference genome to describe the variant
	5.prime.pos	Genomic position of the 5' breakpoint (chr:pos)
	5.prime.chromosome	Chromosome of the 5' breakpoint (with "chr" prefix)
	5.prime.seqName	Chromosome of the 5' breakpoint (without "chr" prefix)
	5.prime.pos1	Genomic position of the 5' breakpoint (numeric position only)

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
	3.prime.pos	Genomic position of the 3' breakpoint (chr:pos)
	3.prime.chromosome	Chromosome of the 3' breakpoint (with "chr" prefix)
	3.prime.seqName	Chromosome of the 3' breakpoint (without "chr" prefix)
	3.prime.pos1	Genomic position of the 3' breakpoint (numeric position only)
	DNAseq	DNA sequence at the breakpoint junction (if available)
Transcript annotation	5.prime.tx	Transcript ID at the 5' breakpoint
	5.prime.gene	Gene name at the 5' breakpoint
	5.prime.exon	Exon number at the 5' breakpoint
	5.prime.type	Genomic feature type at 5' breakpoint (e.g., exon, intron)
	3.prime.tx	Transcript ID at the 3' breakpoint
	3.prime.gene	Gene name at the 3' breakpoint
	3.prime.exon	Exon number at the 3' breakpoint
	3.prime.type	Genomic feature type at 3' breakpoint (e.g., exon, intron)
	coding.consequence	Predicted effect on coding sequence (if determinable)
	In-Frame	Whether the variant preserves the reading frame (Yes/No)
	description	Free-text description of the variant (if available)
Signal quantification	read.count	Number of sequencing reads supporting the event
	mol.count	Number of unique molecules (deduplicated reads) supporting the event
	read.percentage	Fraction of supporting reads relative to total reads
	mol.percentage	Fraction of supporting molecules relative to total molecules
Warning and quality	filter	Indicates whether a variant has passed or

FIELD CATEGORY	FIELD NAME	FIELD DESCRIPTION
		failed internal quality control filters. The list of filters potentially applied is provided in the section “Technical overview of the analysis”
Variant classification database	COSMIC	Match to known variants in COSMIC database (if any)
	ChimerDB	Match to known fusion/chimeric events in ChimerDB (if any)
Internal reference	sgid	Internal variant ID

8.11 RNA analysis: Gene expression analysis with PAPRIKA

8.11.1 Analysis purpose

The core function of the gene expression analysis module is to quantify transcript abundance estimates from targeted RNA sequencing data. Accurate measurement of gene expression is essential to characterize the transcriptome in tumor samples and to detect biologically relevant differences across conditions. The method distinguishes between RNA- and DNA-derived reads ensuring that expression measurements reflect true mRNA content even in the presence of contaminating genomic DNA. This analysis computes gene expression levels normalized with respect to a set of control genes to correct for variability in input material and experimental conditions, enabling consistent comparison across samples.

A subset of genes targeted by the RNA panel (N=56) are in scope for gene expression quantification:

AKT1, AKT3, ALK, ALPK1, ARHGA-P26, BCOR, BRAF, CSF1, CTNNB1, DICER1, EGFR, ERBB2, ERG, ESR1, ETV1, ETV4, FGFR1, FGFR2, FGFR3, FUS, HMGA2, HRAS, JAK2, JAK3, KRAS, MAP2K1, MDM2, MET, MYBL1, MYOD1, NCOA1, NFE2L2, NFIB, NOTCH1, NRAS, NRG1, NTRK1, NTRK2, NTRK3, PDGFRA, PIK3CA, PKN1, PRKACA, PRKACB, PRKCD, RAD51B, RAF1, RET, ROS1, SS18L1, SS18, STAT6, TCF12, TFE3, THADA, YAP1

8.11.2 Technical overview of the analysis

Gene expression analysis includes three algorithmic steps after RNA data preprocessing.

Step 1: RNA signal quantification

The gene expression analysis focuses on a set of genomic positions included in the RNA panel to quantify gene expression. These positions are referred to as *gene expression viewpoints*. At viewpoints, the fragment coverage is processed to quantify the number of RNA molecules, estimated based on CUMIN™ groups, while compensating for potential biases originating from DNA contamination. This step allows for robust quantification of expressed transcripts, even in samples with residual DNA.

Step 2: Quantification of RNA signal at gene-level resolution

For each gene in scope of the analysis, a single value representing the gene-level RNA signal is obtained by computing the median number of RNA molecules (estimated based on CUMIN™ groups) across all viewpoints within the gene. By averaging out the variability observed at individual viewpoints, this step improves the stability of the gene expression measurement.

Step 3: Gene expression measurement

For each sample, the final gene expression values are calculated by normalizing the gene-level RNA signal to the mean RNA signal of a predefined set of control genes

included in the panel. The control gene set has been established based on stable expression across multiple tissue types. The control genes are exclusively used for normalization purposes. The associated gene-expression values are thus not provided. This normalization step corrects for potential technical variability, such as differences in RNA input quantity or quality, and enables reliable comparison of gene expression profiles across samples.

8.11.3 Description of the results

The analysis produces a table that reports the normalized expression value for each sample, run and gene:

- **Normalized expression values:** Gene counts normalized by control gene expression, allowing cross-sample comparability.

Results are provided through the SOPHiA DDM™ platform as a downloadable file. The following table provides an overview of the output files produced by the gene expression analysis.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
normalized_gene_expression_counts.txt	Run	This table contains normalized gene expression values for each sample-gene pair. The sample names provided correspond to the analysis SGID identifiers in the SOPHiA DDM™ platform.

8.11.4 Precautions and limitations

- The accuracy of expression measurement may differ across genes depending on the transcript structure and presence of multiple isoforms, impacting the number of available gene expression viewpoints.
- Normalized gene expression values are defined with respect to a predefined set of control genes. Distortions may be introduced into the data if these control genes are not stably expressed across the samples.
- The method does not perform comparative analysis needed to identify transcriptomic changes such as the identification of up- and down-regulated genes.
- Gene expression values may be biased in case of low RNA input, low RNA quality, low sample tumor content, or insufficient sequencing depth.

8.12 DNA analysis: Sample tumor content and allele specific copy number (ASCN) analysis

8.12.1 Analysis purpose

Tumor DNA samples generally consist of a mixture of DNA originating from tumor cells and from normal cells. The fraction of tumor-derived DNA present in the sample, commonly referred to as tumor content or tumor purity, is an important property, as it directly affects the interpretation of various genomic signals. In addition, the ability of NGS assays to detect genomic alterations and features are often limited by sample tumor content, making this parameter valuable for interpreting analytical results.

Large-scale copy number alterations which are often observed in tumor cells, such as *gains* and *losses*, affect the coverage signal as well as the allele fractions of germline-origin heterozygous variants (BAF). The magnitude of these signals can be exploited to estimate sample tumor content and tumor ploidy (i.e., the average copy number of the tumor genome).

When tumor content and tumor ploidy are known, the coverage and BAF signals can be used to infer total copy number and B-allele copy number (i.e. the number of copies of the minor allele) across the tumor genome. This analysis is often referred to as allele-specific copy number analysis (ASCN analysis). As a result, regions with loss of heterozygosity (LOH) can be identified, and copy number alterations can be characterized more precisely, distinguishing, for example, homozygous from heterozygous deletions.

The ASCN analysis module uses a proprietary method (patent pending) which combines variant allele fractions measured from targeted sequencing data with the genome-wide coverage signal obtained via low-pass WGS to estimate sample tumor content and tumor ploidy. Using these estimates, the module performs genome-wide ASCN analysis, generating profiles of total copy number and B-allele copy number across the entire genome. Finally, for each gene included in the targeted panel, the module reports gene-level estimates of total copy number, B-allele copy number, and loss of heterozygosity (LOH).

8.12.2 Technical overview of the analysis

ASCN analysis uses the following inputs measured from a tumor sample:

- Coverage signal: A genome-wide coverage profile, derived from the low-pass WGS portion of the data.
- B-allele frequency signal (BAF signal): The set of variant allele frequencies (VAFs) associated to single-nucleotide variants identified from the targeted sequencing portion of the data.

The analysis is performed in five main steps, which are described below.

Step 1a: Data preprocessing (BAF signal preprocessing)

SNVs identified in the targeted sequencing portion of the data are filtered to retain only those suitable for ASCN analysis. Specifically, SNVs are retained if they: 1. Are suspected to originate from heterozygous germline loci (i.e. heterozygous SNPs). 2. Have sufficient local coverage to ensure accurate VAF measurements. The raw VAF values are then adjusted to correct for biases that can arise in capture-based sequencing. The resulting set of VAF values constitute the preprocessed BAF signal.

Step 1b: Data preprocessing (Coverage signal preprocessing)

The low-pass WGS coverage profile is normalized using the same algorithm as implemented in our proprietary GIInger™ module.

Step 2: Data quality assessment (QA)

The preprocessed data undergo a quality assessment (QA) step to determine whether they are of sufficient quality to proceed with ASCN analysis. During this step, quality metrics are computed and used to determine a sample QA status. The sample QA status can take the following values:

- **Pass:** All quality metrics meet the acceptance criteria; the preprocessed data are suitable for ASCN analysis.
- **Fail:** One or more quality metrics do not meet the acceptance criteria; the preprocessed data are not suitable for ASCN analysis.

The QA metrics and the associated acceptance criteria used to determine the QA status are provided in Section 8.12.3.

Step 3: Sample tumor content and tumor ploidy estimation

A Hidden Markov Model (HMM) is used to jointly model the coverage and BAF signals across the genome. In this model, the hidden states represent possible combinations of total copy number and B-allele copy number in the tumor cells. The model assumes that all tumor-derived DNA in the sample originates from a single dominant tumor clone and that the normal cells do not carry large-scale copy number aberrations.

The HMM model is fitted to the coverage and BAF signals, and an optimization procedure is used to identify the tumor content and tumor ploidy values that allow the HMM to best explain the observed signals.

Step 4: Genome-wide allele-specific copy number analysis

Using the estimated tumor content and tumor ploidy obtained in Step 3, the HMM is applied to infer allele-specific copy number states along the genome. This results in genome-wide profiles of total copy number and B-allele copy number. The set of possible total copy numbers computed by the module is {0,1,2,3,4,5,6,7,>7}, where “>7” indicates 8 or more copies.

In addition to these discrete total copy number levels, the HMM model includes a state designed to capture genomic regions where the observed data do not match a discrete total copy number level but fall between two consecutive discrete total copy number levels. This may occur, for example, in the presence of subclonal copy number aberrations. In such cases, the data are assigned to this non-specific HMM state, indicating that the total copy number and B-allele copy number cannot be reliably determined for that region.

Because the BAF signal is derived from the targeted sequencing portion of the data, it may not be uniformly distributed across the genome, as it can only be captured in regions included in the targeted panel. As a result, genome-wide B-allele copy number estimates may be less precise in genomic regions with sparse BAF observations. In addition, in case of noisy data or in the presence of subclonal copy number alterations that are not assigned to the non-specific HMM state described above, the data observed in some genomic regions may not be well explained by the HMM model. To account for these situations, the module assigns a confidence level to each inferred total copy number and B-allele copy number state. This confidence level reflects how reliable each inferred copy number is, with lower confidence typically occurring in regions where: the BAF signal is sparse, signals are noisy or inconsistent with the expected values, or near transitions between different copy number states.

By combining the total and B-allele copy number profiles, and the associated confidence scores, the ASCN module identifies regions of the genome that are affected by loss of heterozygosity (LOH). LOH is defined as regions in which the B-allele copy number is zero, while total copy number remains greater than zero. As a result, a genome-wide LOH profile is produced.

Finally, the genome-wide ASCN profiles computed in this step undergo an automated assessment to evaluate the overall reliability of the result. When appropriate, sample-level warnings are issued to indicate an overall reduced confidence in the results. The list of possible warnings is provided in the section 8.12.3: Description of the results

Step 5: Gene-specific allele-specific copy number analysis

For each gene included in the targeted sequencing panel, the genome-wide ASCN profiles generated in Step 4 are utilized to derive gene-level results.

Using the genome-wide ASCN results, the module determines for each gene the corresponding values of: total copy number, B-allele copy number, and LOH status (with the associated confidence level).

If the genomic region corresponding to a gene does not uniquely overlap a single ASCN segment (for example, if the gene spans a breakpoint between two regions with different copy number states), gene-specific ASCN results are not reported for that gene.

8.12.3 Description of the results

For each sample, the ASCN module provides the following results:

- **QA metrics and status**: A set of QA metrics and a QA status indicating whether the NGS data are suitable for ASCN analysis (see Step 2)
- **Tumor content and tumor ploidy results**: Estimates of sample tumor content and tumor ploidy (see Step 3), and associated warnings (see Step 4)
- **Genome-wide ASCN results:** Genome-wide total copy number, B-allele copy number, and LOH status profiles (see Section 8.12.2, Step 4)
- **Gene-specific ASCN results:** Readouts of genome-wide ASCN results for the set of genes targeted by the assay (see Step 5)

QA metrics and status

For each sample, the following QA metrics are computed to determine whether the input data are suitable for ASCN analysis:

- **Nb. of DNA Fragments:** The total number of DNA fragments available for WGS coverage profile calculation. If the Nb. of DNA Fragments is lower than 4M, the QA status is *Fail*.
- **Coverage Residual Noise:** A measure of the noise in the normalized WGS coverage profile used as input to the ASCN module. If the Coverage Residual Noise is larger than 0.2, the QA status is *Fail*.
- **Number of SNPs:** The number of putative heterozygous SNPs retained as suitable for ASCN analysis. If the Number of SNPs is lower than 500, the QA status is *Fail*.
- **BAF Residual Noise:** A measure of the noise in the BAF signal used as input to the ASCN module. If the BAF Residual Noise is larger than 0.125, the QA status is *Fail*.

Tumor content and ploidy results

For each sample, the following results are computed:

- **Tumor content:** Percentage of tumor-derived DNA present in the sample.
- **Tumor ploidy:** Average total copy number in tumor cells, computed across the entire genome.

- **Warning:** Outcome of the automated assessment performed to evaluate the overall reliability of tumor content, tumor ploidy and ASCN results. Three possible warnings can be issued:

    - **Low SNR:** The ratio between the magnitudes of the signal (expected given the estimated tumor content and tumor ploidy) and the noise present in the data is low. In this situation, the copy number–related signals in the data are weak relative to the noise observed in the data. In the case of a Low SNR warning, tumor content, tumor ploidy, and ASCN results are reported, but the risk of inaccurate estimates is increased.
    - **Silent CN profile:** No clear signatures of large-scale copy number alterations are detectable in the coverage and BAF signals. This warning indicates either that the sample tumor content is extremely low or that the tumor genome does not feature large-scale copy number changes. In the case of a Silent CN profile warning, tumor content, tumor ploidy, and ASCN results are not reported.
    - **Biologically Unexpected CN profile:** The genome-wide ASCN results correspond to a copy number pattern that is considered highly unusual from a biological perspective (for example, a large proportion of the tumor genome is inferred to have a total copy number equal to 0). In the case of a Biologically Unexpected CN profile warning, tumor content, tumor ploidy, and ASCN results are reported, but the risk of inaccurate estimates is increased.

Genome-wide ASCN results

For each sample, the ASCN module generates the following genome-wide profiles:

- **Total copy number profile:** Annotates genomic regions across the entire genome with the estimated total copy number in the tumor sample. Possible values are 0, 1, 2, 3, 4, 5, 6, 7, >7, or NA. “>7” indicates 8 or more copies. NA indicates that the total copy number cannot be reliably determined for that genomic region.
- **B-allele copy number profile:** Annotates genomic regions across the entire genome with the estimated B-allele copy number in the tumor sample. Following standard conventions, the B-allele copy number corresponds to the number of copies of the allele present at the lowest copy number. Possible values range from 0 up to half of the total copy number, or NA. NA indicates that the B-allele copy number cannot be reliably determined for that genomic region.
- **LOH profile:** Annotates genomic regions across the entire genome with the inferred loss of heterozygosity (LOH) status. The following categories are reported:

    - **LOH (high confidence):** LOH detected with high confidence
    - **LOH (low confidence):** LOH detected with low confidence
    - **No LOH:** Data do not support the presence of LOH

- **Undetermined:** LOH status cannot be reliably determined

Gene-specific ASCN results

For each sample, gene-specific ASCN results are derived by utilizing the genome-wide ASCN profiles at the genomic location of each gene included in the targeted sequencing panel. For each gene, the following results are reported:

- **Total copy number:** The total copy number for the specific gene. Possible values are the same as described for the *Genome-wide ASCN results: Total copy number profile*.
- **B-allele copy number:** The B-allele (minor allele) copy number for the specific gene. Possible values are the same as described for the *Genome-wide ASCN results: B-allele copy number profile*.
- **Gene LOH status and confidence:** The LOH status for the gene, along with the associated confidence level. Possible values are the same as described for the *Genome-wide ASCN results: LOH profile*.

To generate gene-specific results, the genomic region corresponding to a gene must be entirely contained within a contiguous region of the genome (referred to as a *segment*) where both the total copy number and B-allele copy number are uniform. When gene-specific ASCN results can be established, the module reports the genomic coordinates of the specific genome-wide ASCN profile segment to which the gene belongs.

Tumor content and ASCN results are available in the SOPHiA DDM™ platform and are also provided as downloadable files. The SOPHiA DDM™ platform displays QA metrics and status, as well as tumor content and tumor ploidy results. It also includes a genomic browser that allows users to visualize the following tracks: B-allele signal ("B-allele frequency"), coverage signal ("Normalized coverage"), genome-wide total copy number profile ("Absolute CN"), genome-wide LOH profile ("LOH"), and gene-specific ASCN results ("Genes"). The following table provides an overview of the results available via downloadable files.

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
purity-ploidy-run-level-report.pdf	Run	PDF report describing Tumor Content and ASCN analysis results for all samples included in the run.
purity-ploidy-report.pdf	Sample	PDF report describing Tumor Content and ASCN analysis

FILE NAME	RUN- vs. SAMPLE-SPECIFIC OUTPUT	DESCRIPTION
		results for individual samples.
purity-ploidy-run-level-results.tsv	Run	Text file containing QA metrics, QA status, Tumor Content and Tumor Ploidy results for all samples included in the run.
purity-ploidy-results.tsv	Sample	Text file containing QA metrics, QA status, Tumor Content and Tumor Ploidy results for individual samples.
purity-ploidy-run-level-gene-results.tsv	Run	Text file containing gene-specific ASCN results for all samples included in the run.
purity-ploidy-gene-results.tsv	Sample	Text file containing gene-specific ASCN results for individual samples.

NOTE: In the current version of the product, genome-wide ASCN results are not provided via downloadable files.

8.12.4 Precautions and limitations

- Processing of highly degraded FFPE samples may result in poor-quality data, increasing the risk of sample rejection or inaccurate results.
- The module is designed to operate on combined DNA capture and DNA WGS data. In the absence of DNA WGS data, the module rejects the sample due to an insufficient number of DNA WGS reads for coverage calculation.
- The accuracy of tumor content, ploidy, and ASCN estimation depends on the strength of the underlying genomic signals, which is influenced by tumor content. In samples with low tumor content, copy number and B-allele frequency signals may be insufficiently pronounced, which may result in sample rejection, warnings, or reduced accuracy of the reported results. Verification studies showed that results are robust for samples with tumor content above approximately 30%, while below this level the reliability of results may vary depending on the specific CNV profile and data quality.
- The analysis is designed to report tumor content, ploidy, and ASCN profiles representative of the dominant tumor clone. While the model includes mechanisms to accommodate signals that do not conform to integer copy number states (e.g., arising from subclonal alterations), it does not explicitly resolve subclonal structure. As a result, subclonal events may be reported as undetermined,

approximated to the closest integer copy number, or, in rare cases, may affect the accuracy of tumor ploidy estimation.

- The estimation of tumor content, ploidy, and ASCN profiles relies on the presence of copy number patterns in the tumor sample that allow the HMM model to fit the coverage and BAF signals. In some cases, multiple combinations of tumor content and ploidy may fit the observed data with a similar degree of accuracy. The module reports the solution considered most likely, but due to this inherent ambiguity, there is a risk that the reported solution may not correspond to the actual tumor copy number profile.
- ASCN profiles are derived using estimated tumor content and tumor ploidy; inaccuracies in these estimates may propagate to the reported ASCN profiles.
- ASCN profiles may be less accurate in genomic regions located near transitions between adjacent ASCN segments (i.e., breakpoints, e.g., regions where a copy number change occurs).
- B-allele copy number profiles and LOH detection relies on the availability of variant allele frequency (VAF) data from the targeted sequencing panel. Because SNPs are not uniformly distributed across the genome, LOH may be undetermined in regions with sparse VAF coverage.
- Tumor content, ploidy, and ASCN analysis includes the X chromosome only for samples determined to be female. In male or sex-undetermined samples, the X chromosome is excluded from the analysis.
- Gene-specific ASCN results are derived from genome-wide ASCN profiles based on DNA WGS data, which do not include coverage from the targeted gene regions. Very focal copy number alterations affecting only a targeted gene and not reflected in the VAF signal may therefore be missed.

9 PRECAUTIONS AND GENERAL LIMITATIONS OF THE APPLICATION

The following are general precautions and limitations that apply to this application. For module-specific precautions and limitations, please refer to the section TECHNICAL DESCRIPTION OF THE BIOINFORMATIC ANALYSIS.

9.1 Precautions

- DNA and RNA sample processing and NGS data generation shall follow the instructions provided in this application manual and use only supported sequencing instruments. Any deviation may compromise data quality and result accuracy.
- Users shall take care to avoid cross-sample contamination prior to the indexing step of library preparation. Contamination occurring before library indexing may impact downstream analyses, including variant calling, TMB, MSI, HRD genomic integrity, and gene expression. The use of unique dual indexes (UDI) mitigates cross-sample contamination occurring after library indexing.
- Users should generate NGS data targeting the recommended number of paired-end reads per sample type, as described in this Application Manual. The bioinformatics pipeline can process sequencing data in excess; however, very large excess may be downsampled. Generating substantially more sequencing data provides no additional benefit.
- The pipeline relies on the NGS data type specified by the user via the FASTQ file name (e.g., -W, -D, -R, -DW) to determine the type of analysis to perform. SOPHiA DDM™ does not verify that the content of the FASTQ file is consistent with the specified NGS data type. Mislabeled files will be processed according to the user-specified type, which may produce invalid results. Users must ensure that FASTQ files are correctly named and labeled to obtain valid results.
- Users shall only upload data generated according to the sequencing modalities described in this Application Manual. The application does not support the analysis of NGS data obtained using a sequencing modality that combines DNA WGS (-W) and RNA capture (-R). Uploading NGS data generated with unsupported modalities may cause errors and prevent the complete analysis from being displayed in SOPHiA DDM™.
- All FASTQ files labeled with the postfix “-DW” will automatically undergo GIInger™ Genomic Integrity analysis. To perform Tumor Content and ASCN analysis (requiring combined DNA capture and DNA WGS data) without triggering GIInger™ Genomic Integrity analysis, NGS data generated from combined DNA capture and DNA WGS must be labeled with the “-D” postfix.

- All FASTQ files labeled with the postfix “-D” will automatically undergo Tumor Content and ASCN analysis. The successful completion of this analysis requires both DNA capture and DNA WGS data. In the absence of DNA WGS data, samples are rejected from the Tumor content and ASCN analysis, due to insufficient number of reads available for WGS coverage calculation.
- A given fusion or exon-skipping event may be detectable only in RNA sequencing or only in DNA sequencing results due to biological or technical factors. Consequently, some variants may be identified by only one of the two analyses.
- If a gene fusion or exon-skipping event is detected in both DNA sequencing and RNA sequencing results, the application reports each call independently. The application does not consolidate these calls into a single event. Users shall assess whether multiple calls correspond to the same underlying mutation.
- The whole-gene amplification and deletion CNV module, and the exon-level CNV module report results separately. The exon-level CNV module may identify intragenic events that refine or modify the call reported by the whole-gene CNV module. Users shall review the results of both modules to ensure accurate interpretation of CNV findings.
- The exon-level CNV module has stricter data quality requirements than the whole-gene amplification and deletion CNV module. A sample may yield a conclusive whole-gene CNV result but be rejected from exon-level analysis; in such cases, the whole-gene CNV results remain valid.
- The product performs two independent gene-level copy number variation (CNV) analyses. The Gene Amplification and Deletions analysis is based on coverage from DNA capture data and measures relative copy number (i.e., reported copy numbers are influenced by tumor content and ploidy). The Tumor Content and ASCN analysis is based on coverage from DNA WGS data (in combination with VAF from DNA capture data) and reports absolute copy number in the tumor. Tumor content and ASCN analysis aims at complementing the Gene Amplifications and Deletions analysis. In the current version of the product, these two analyses are performed and reported independently and are not reconciled into a single, unified result. Discrepancies between results may occur for various reasons, including reduced sensitivity of ASCN analysis for detecting focal CNV events. Results from both analyses should be interpreted with caution and in the context in their respective methodological limitations.

9.2 General limitations

- This product is for research use only (RUO) and not for use in diagnostic procedures.
- This application was developed and tested using FFPE-derived DNA and RNA samples. While the application may be used with other sample types, such as fresh frozen DNA and RNA samples, its performance has not been established

under these conditions. While the application may be used with Total Nucleic Acid (TNA) as an alternative input for RNA analysis, its performance has not been established under this condition. TNA is not recommended as input for DNA analysis, as it may lead to inaccurate or unreliable results.

- In the current version of the application, quality indicators related to CUMIN™ pre-processing are available only for DNA data. While CUMIN™ pre-processing is also applied to RNA data, the corresponding RNA quality indicators are not available in SOPHiA DDM™. Detailed CUMIN™ quality metrics for RNA samples can be found in the pipeline QA report.
- The current version of this application performs tumor-only analysis and does not yet include tumor–normal matched analysis. Without a matched normal, germline variants may be present in the results.
- The results produced by the exon-level CNV analysis, and the gene expression analysis are provided in SOPHiA DDM™ exclusively as downloadable files.
- The accuracy and reliability of the results depend on the quality and integrity of the DNA and RNA samples. Highly degraded or poor-quality FFPE samples may generate low-quality NGS data, increasing the risk of false negative and false positive results.
- The accuracy and sensitivity of the results depend on the tumor content of the analyzed DNA or RNA samples. Samples with low tumor content may lead to false negative results due to insufficient representation of tumor-derived material.
- NGS data may feature insufficient local read depth in certain regions of the DNA panel, which may reduce the sensitivity of the results. This may occur even when sample multiplexing recommendations are followed and the recommended average number of reads per sample is achieved. Contributing factors include low sample quality, suboptimal NGS data quality, uneven read allocation among multiplexed samples, low capture quality resulting in poor coverage uniformity, or consistently reduced read depth in challenging regions (e.g., AT-rich regions).
- Highly degraded RNA samples may result in sequencing data with a high proportion of soft-clipped reads. Processing such NGS data can increase computational load and lead to longer bioinformatics analysis turnaround times.

10 SUPPORT

In case of difficulty using the SOPHiA DDM™ Platform, please consult the troubleshooting section of the "General information about usage of SOPHiA DDM™" document, visit our support portal, e-mail support@sophiagenetics.com, or contact our support line by telephone at:

- EU: +41 21 561 34 75
- US: +1 617 313 7957
- FR: +33 5 47 51 01 29
- AU: +61 2 51 10 7564

Any serious incident occurring in relation to the device should be promptly reported to SOPHiA GENETICS and the competent authorities of the member state where the user is established.

Do not use components that are damaged. Contact the SOPHiA GENETICS support team if there are any concerns with the kits using the support channels mentioned above.

Document Approvals

Approved Date: 21 Apr 2026

Approval Verdict: Approve	Technical Approval 21-Apr-2026 06:28:10 GMT+0000
QA Approval Verdict: Approve	Quality Assurance Approval 21-Apr-2026 07:52:25 GMT+0000