- Pipeline Parameters -

Pipeline	ILL1XG1G6_CNV
Gene Panel	SOPHiA DDM Dx Hereditary Cancer Solution (HCS_v1_1)
Sequencer	Illumina MiSeq
Reference genome	hg19
Sample Type	Germline

- Disclaimer -

This file describes the parameters of the algorithm that are set for the product described above. Steps not covered by the CE-ICD claim, and therefore considered Clinical Decision Support features, are marked below as "CDS component" (also highlighted in blue).

- Pipeline Steps: Overview -

SNVs / Indels / CNVs

1. Preprocessing
   1.1. **CDS component:** Collect quality metrics based on the raw fastq files for the quality report
2. Alignment
   2.1. Cut adapters and trim low quality ends from reads (base quality below 20)
   2.2. Align reads to the genome
   2.3. Identify regions with soft clipped reads
      2.3.1. Identify noisy regions with heterogeneous long soft-clipped reads that should be omitted during realignment step
      2.3.2. Identify ALU insertion by comparison with database of known ALU sequences (for details please see limitation section)
      2.3.3. Detect long insertion or deletions (with maximal size 10000bp) and realign long soft-clipped reads present in these regions
      2.3.4. Reassembly reads in regions realigned in previous steps to avoid multiple presentations of the same indel
   2.4. **CDS component:** Calculate read statistics for quality report
   2.5. **CDS component:** Calculate coverage for CNV analysis
3. Variant calling
   3.1. Calculate a sample specific error profile based on Phred score and smooth the profile with linear fitting
   3.2. Collect statistics about all variants found in the data, excluding bases with Phred score below 20
   3.3. Apply statistical test to identify real variants.
      3.3.1. For each variant compute probability of being present at the observed level compared to the expected level of sequencing errors based on the sequencing quality scores and probability of strand bias.
      3.3.2. Remove variants that do not meet minimal requirements: total read depth is at least 30x, while minimum required number of reads supporting alternative variant is above 10 (after excluding bases with Phred quality score below 20), variant fraction is above 1%.
   3.4. Merge variants together if they are on the same allele and with a maximum distance of 2bp (realign indels to see whether there is a representation that fulfills the distance criteria)
   3.5. Re-quantify the variant fraction of variants
4. **CDS component:** CNV detection
   4.1. **CDS component:** CNV analysis (performed if a minimum of 8 samples is present)
   4.2. **CDS component:** CNV annotation
   4.3. **CDS component:** Create CNV report and coverage plot
5. Annotation

5.1. Remove variants:
- 5.1.1. Duplications longer than 500 bp (applies only to duplication found during SNP/indel calling, does not concern duplication reported by CNV analysis)
- 5.1.2. Frequently observed sequencing artefacts (precomputed set of SNPs present with low variant fraction in at least 50% of samples of representative dataset)
- 5.1.3. That are located at a distance larger than 500bp from the target region
- 5.1.4. If the probability of being present at the observed level when compared to the expected level of sequencing errors based on the sequencing quality scores (score P) is identical or higher than the cutoff equal to 10^{-10} (score P expressed in phred scale is identical or smaller than the cutoff equal to 100)
- 5.1.5. If the probability that the strand bias not occurred (score S) is identical or lower than the cutoff equal to 10^{-7} (score S expressed in phred scale is identical or larger than the cutoff equal to 70)
- 5.1.6. If both scores described above only marginally cross the corresponding cutoffs (score P expressed in phred scale after subtraction of 0.25 of score S expressed in phred scale is identical or smaller than the cutoff equal to 100)

5.2. **CDS component:** Calculate transcript dependent variant annotation, as c.DNA, protein notation, exon rank etc.

5.3. Classify variants as **low confidence variants:**
- 5.3.1. if their variant fraction is too low. Cutoffs used: 15% for Indels, 20% for SNVs (filter name: low_variant_fraction)
- 5.3.2. In case of indels in homopolymers of length 10 or longer (filter name: homopolymer_region)
- 5.3.3. In case of variants outside of the target region of the panel (filter name: off_target)
- 5.3.4. In case of variants with coverage below 30x (filter name: low_coverage).
- 5.3.5. In case of the variants which variant fraction does not differ statistically from the variant fraction of frequent sequence context artefacts identified based on representative sequencing data (filter name: likely_sequence_context_artefact)
- 5.3.6. In case of variants inside of problematic regions (e.g., low complexity regions) (filter name: problematic_region).

chr	start	end
2	48018282	48018315
3	37067075	37067120
5	112111309	112111311
11	94152620	94152662
17	59757835	59757859

5.4. **CDS component:** Extract information about this variant from databases:
- 5.4.1. **CDS component:** OMIM, dbSNP identifier
- 5.4.2. **CDS component:** allele frequency from 1000 genome project, ExAC, ESP5400, gnomAD
- 5.4.3. **CDS component:** prediction scores from dbNSFP: SIFT, PolyPhen2, MutationTaster, GERP, LRT, PhyloP
- 5.4.4. **CDS component:** clinical significance from ClinVar

5.5. **CDS component:** Merge all the annotations into one full variant table and vcf file

5.6. Create full variant table (full_variant_table_secondary.txt) with the results of the secondary results only

5.7. Calculate sample specific low coverage regions, default cutoff = 50x

5.8. Add warning flag to variants
- 5.8.1. that overlap sample specific low coverage regions (class: low_coverage)
- 5.8.2. that overlap predefined panel and sequencer specific flagged regions. Reasons for additional flagging can be noisy regions, pseudogenes, low complexity regions. See table for details.

chr	start	end	class	comment
2	47635515	47635560	noisy region	MSH2_ex02
2	48032732	48032820	noisy region	MSH6_ex07
3	37067103	37067240	noisy region	MLH1_ex13
7	6036932	6037079	noisy region	PMS2_ex07
3	178937334	178937548	pseudogene homology (gene)	PIK3CA
3	178937712	178937865	pseudogene homology (gene)	PIK3CA
7	6012845	6013198	pseudogene homology (gene)	PMS2
7	6017194	6017413	pseudogene homology (gene)	PMS2
7	6018202	6018352	pseudogene homology (gene)	PMS2
7	6022430	6022647	pseudogene homology (gene)	PMS2
7	6026365	6027276	pseudogene homology (gene)	PMS2
7	6776682	6777592	pseudogene homology (pseudogene)	PMS2CL
7	6781291	6781508	pseudogene homology (pseudogene)	PMS2CL
7	6785694	6785844	pseudogene homology (pseudogene)	PMS2CL
7	6786642	6786861	pseudogene homology (pseudogene)	PMS2CL
7	6790855	6791257	pseudogene homology (pseudogene)	PMS2CL
22	29085098	29085228	pseudogene homology (gene)	CHEK2

6. Postprocessing
   6.1. **CDS component:** Create quality plots for the report
   6.2. **CDS component:** Generate quality report

- Limitations and Cautions -
SNVs / Indels / CNVs

See instructions For Use for a detailed list of limitations.

Revision History

VERSION / DATE	DESCRIPTION OF CHANGE
V 2.0	Updated from ILL1XG1G3_CNV to ILL1XG1G6_CNV. This update includes the following changes: • Integration into the SOPHiA DDM web application • Creation of full variant table and flagged region files, focusing on secondary analysis results Further changes to the document: • Identified Clinical Decision Support components • Added cross-reference to the limitations in the IFUs • Completed information about reference genome
V 1.0	Initial document

Document Approvals

Approved Date: 10 May 2022

Approval Task Verdict: Approve	Tamara Maas, (tmaas@sophiagenetics.com) Technical Review 10-May-2022 12:48:56 GMT+0000

Task: QA Approval Verdict: Approve	Maria Nacca, (mnacca@sophiagenetics.com) Quality Assurance Approval 10-May-2022 12:59:44 GMT+0000