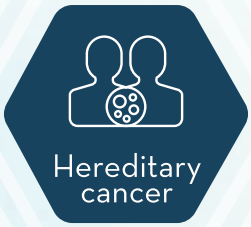
CE IVD

HEREDITARY CANCER SOLUTION BY SOPHiA GENETICS

APPLICATION NOTE



SOPHiA GENETICS®

Introduction

SOPHiA™ artificial intelligence (AI) continuously learns from thousands of patients' genomic profiles and experts' knowledge to accurately detect disease-causing mutations or -predisposing variants. The unmatched analytical power of SOPHiA relies on the genomic information pooled on SOPHiA DDM®, the platform of choice for clinicians performing Next-Generation Sequencing (NGS) testing. Since the launch of SOPHiA DDM, hundreds of hospitals worldwide have benefited from this technology to improve patients' diagnoses. By processing a huge amount of data every day, SOPHiA has learnt how to reduce all the biases introduced behind each library preparation and sequencing.

Thanks to the deep understanding of NGS applications in clinical genomics, we have selected the optimal combination of technologies to develop the Hereditary Cancer Solution by SOPHiA GENETICS. It represents a comprehensive solution with superior analytical performance for the detection of genomic variants associated to hereditary cancers. Moreover, it is the first capture-based kit that obtained the CE-IVD mark for the risk assessment of hereditary breast, ovarian and digestive cancers.

Hereditary Cancer Solution (HCS) by SOPHiA GENETICS

The Hereditary Cancer Solution (HCS) by SOPHiA GENETICS bundles the analytical power of SOPHiA, the collective AI for Data-Driven Medicine, with a capture-based target enrichment kit and full access to SOPHiA DDM. It offers highest on-target rates and lowest noise level, providing optimal input for data analysis. The panel covers the coding regions ± 25 bp of non-coding DNA in exon-flanking regions of 26 most clinically relevant genes (target region of 105 kb) associated to hereditary cancer syndromes. One of the major advantages of the HCS is its highly-optimized probe set that provides a comprehensive coverage of the target regions, resulting in superior data quality starting with as little as 200 ng of genomic DNA.

Knowledge-Driven Kit Design

SaaS Analytical Platform

Integrated library preparation and target capture kit, covering:

ATM, APC, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, EPCAM, FAM175A, MLH1, MRE11A, MSH2, MSH6, MUTYH, NBN, PALB2, PIK3CA, PMS2, PMS2CL, PTEN, RAD50, RAD51C, RAD51D, STK11, TP53, XRCC2*

*Extra probes for the pseudogene *PMS2CL* were added in the design to enable distinction between pseudogene and real disease-causing variants.

SOPHiA analyses complex genomic NGS data by detecting, annotating and pre-classifying genomic variants to help clinicians better diagnose their patients.

- SNVs, Indels and CNVs are accurately detected in all genes of the panel
- ALU insertions are reliably recognized
- Pseudogene variants are efficiently differentiated from real ones

The results are presented in SOPHiA DDM, the platform of choice for clinicians performing routine diagnostic testing.

Thanks to its intuitive user interface and integrated features, variants visualization and interpretation are facilitated, while assuring protection of clinical genomic data.

Workflow

The Hereditary Cancer Solution by SOPHiA GENETICS provides an integrated workflow solution that combines a high-efficiency DNA library preparation kit with an extremely optimized target capture protocol. Within only two working days, users can prepare ready-to-sequence target enriched libraries (Figure 1).

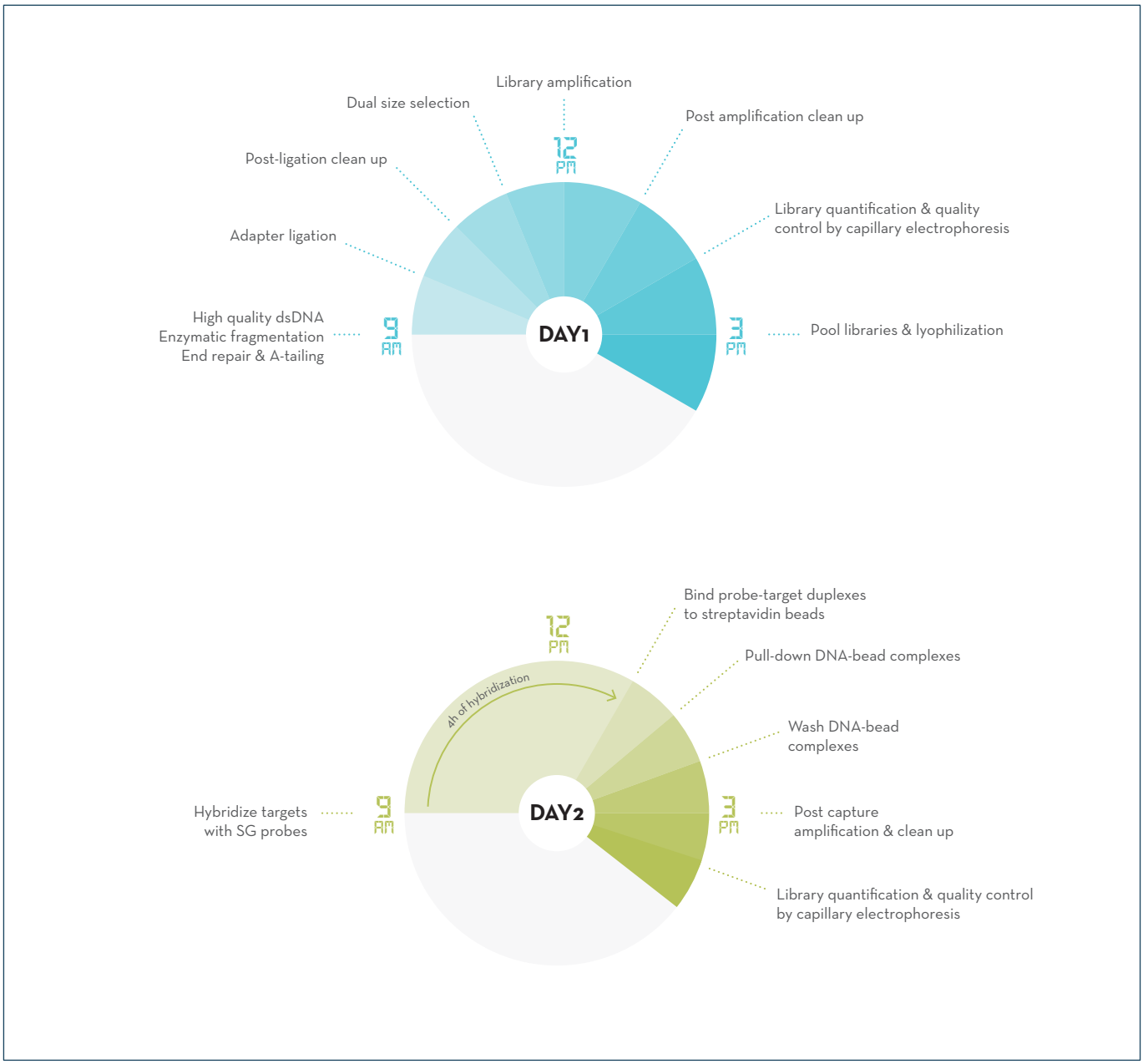


Figure 1: Workflow of the Hereditary Cancer Solution by SOPHiA GENETICS. In only two working days, users can prepare ready-to-sequence target enriched libraries.

Performance Evaluation Study

NGS assays are increasingly being adopted for variant detection in routine genetic testing. For clinical applications, it is crucial to have validated reference sequences to ensure that no artefacts are affecting results and that the solution enables highly reliable detection of variants. Rigorous performance evaluation was therefore conducted using results experimentally verified through alternative methods.

Raw sequencing data and confirmed variants for benchmarking were obtained from 7 sequencing centers using the HCS target-enrichment kit on an Illumina MiSeq® instrument and analyzed by SOPHiA to evaluate the different performance characteristics.

A total of 386 samples were processed either internally or by external partners, 329 of which were from clinical cases. These samples contain a large array of 373 distinct representative mutations and enable estimates of sensitivity for variant calling. Replicates were included to estimate repeatability and reproducibility. For precision, specificity and accuracy, however, larger regions have been characterized by alternative standard methods to also obtain the set of high confidence true negative positions.

Results

Here we demonstrate that the Hereditary Cancer Solution by SOPHiA GENETICS is extremely accurate for routine diagnostics, thanks to SOPHiA, which assures performance metrics comparable to or exceeding those of other NGS tests.

ON-TARGET RATES AND COVERAGE UNIFORMITY

The first performance indicator examined is the rate of fragments that overlap with the target regions covered by the probes of the HCS panel. These so-called on-target rates are often a bottleneck of target enrichment protocols since lower on-target rates significantly reduce the maximum number of samples per run, thus increasing the total cost per sample. For each of the different runs (in all of the 7 centers), we observed very high on-target rates (mean 79,39%), which demonstrates a high specificity of the capture of the probes.

A second important indicator to examine is coverage uniformity. This measure determines which percentage of target regions (CDS ± 25 bp) is within 0.2 and 5 times the median coverage value. A low coverage uniformity would indicate that the assay covers target regions irregularly and therefore that results may be unreliable at some positions. On the contrary, a high coverage uniformity guarantees reliability of the assay across all target regions. For each run of the different sites, we observed very high coverage uniformity across samples (Figure 2).

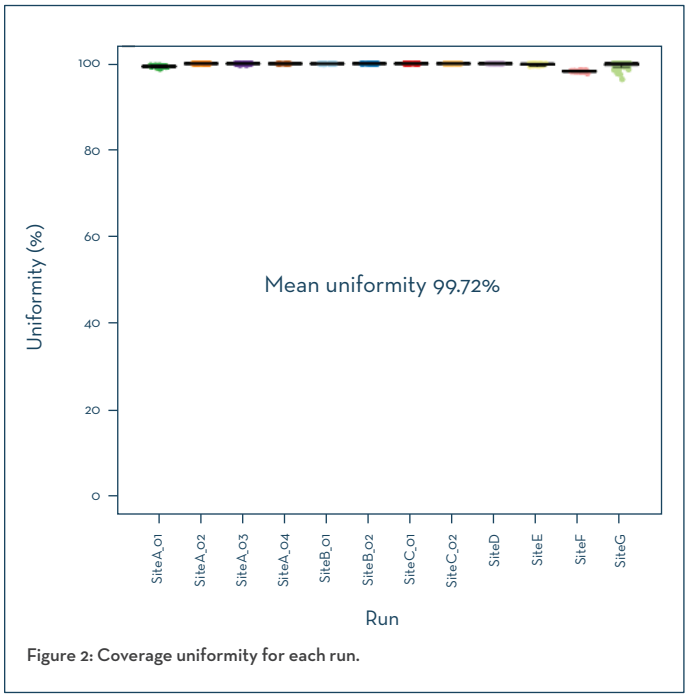


Figure 2: Coverage uniformity for each run.

Finally, we evaluated the total number of positions within each sample where insufficient coverage was reached for high confidence variant calling (positions with less than 50 mapped reads). In most cases, no bases had to be excluded because of low coverage, and only an average of 42 bp and a maximum of 730 bp (respectively 0.04% and 0.78%) had to be removed due to low coverage. Such results are highly satisfactory and demonstrate remarkable quality of the HCS panel.

REPRODUCIBILITY (INTER-RUN SEQUENCING)

24 samples were sequenced across different runs. The graph in Figure 3 displays the correlation of the variant fractions (for each variant) between these replicate samples.
The variant fractions are typically around 0.5 (heterozygous) or close to 1.0 (homozygous), depicted by the coloured dots.

The coefficient of determination (R^2) for all the data points compared to the diagonal shows a very good inter-run reproducibility (R^2 = 0.989). Similar analysis has been performed with intra-run replicates to establish repeatability with comparable results (data not shown).

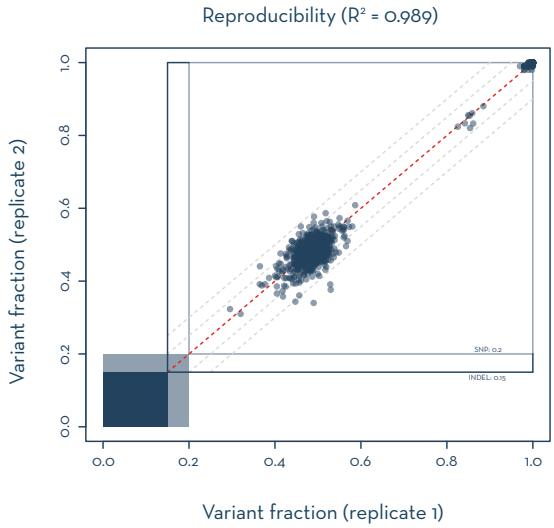


Figure 3: Scatter plot of variant fractions for inter-run replicates. The grey, dotted lines represent the 5% and 10% deviation from identity (diagonal = red dashed line). The blue squares represent the low variant fraction cut-off (SNV = 20%, Indel = 15%).

Top Analytical Performance

Thanks to SOPHiA, the Hereditary Cancer Solution by SOPHiA GENETICS leads to unmatched performance values as shown in Table 1. The performance metrics were calculated on the SNVs and Indels detection inside the target regions. The systematic analysis of the raw data was performed by PEPPER™, SOPHiA GENETICS' patented algorithm, integrated in SOPHiA that is able to call variants with the sensitivity and specificity demanded by clinical grade standards.

Specificity, accuracy and precision were only calculated using fully characterized samples due to the lack of true negative positions in the partially characterized samples. Repeatability and reproducibility were calculated using all positions for all samples and all runs considered.

	PERFORMANCE MEASUREMENT	OBSERVED [LOWER 95% CI]
Sensitivity	100%	99.20%
Specificity	99.99%	99.99%
Accuracy	99.99%	99.99%
Precision	99.86%	96.42%
Repeatability	99.98%	99.98%
Reproducibility	99.93%	99.93%

Table 1: Performance measurements and lower 95% confidence intervals. The 95% CI for sensitivity, specificity, accuracy and precision were calculated on the unique variants in the performance evaluation study, in order to reflect the real diversity of the variants. The 95% CI for repeatability and reproducibility were calculated on all positions for all samples. Repeatability and reproducibility are based on replicates. To determine the confidence intervals in case of 100% sensitivity or other measures, the methods described by Mattocks et al[1] were used. In cases where the measured criterion was less than 100%, the exact method[2] was used to obtain the confidence interval on the binomial probability for sensitivity, specificity, accuracy, and precision. To reflect the true diversity of variants only unique True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) were used in confidence interval calculations.

ALU INSERTIONS

An ALU element is a short stretch of DNA originally characterized by the action of the ALU (*Arthrobacter Luteus*) restriction endonuclease. ALU elements of different kinds occur in large numbers in primate genomes. In fact, ALU elements are the most abundant transposable elements in the human genome and are implicated in several genetic diseases and cancers. Especially in *BRCA1* and *BRCA2 genes*, rearrangements caused by ALU insertions occur quite frequently.[3] Due to their size (~300 bp) they cannot be included in a single read with current sequencing technologies and are therefore not easy to detect. Our algorithms use "soft-clipping" to be able to identify them. Reads that partially map the 5' and 3' ends of the ALU insertion respectively are aligned and mapped against our own database that is based on a list of about 800 ALU sequences. A warning is given in case a transposable element is detected by SOPHiA (Figure 4).

"I am impressed that SOPHiA GENETICS got the analysis right from the first time and could find all the variants and CNVs in a selection of difficult and diverse patients' samples. Thanks to the very intuitive and continuously evolving algorithm, PEPPER, we are now able to detect the ALU insertion in the exon 3 of BRCA2 gene"

Dr Pascale Hilbert, Director of The Molecular Biology Department, Institute of Pathology and Genetics (IPG), Belgium

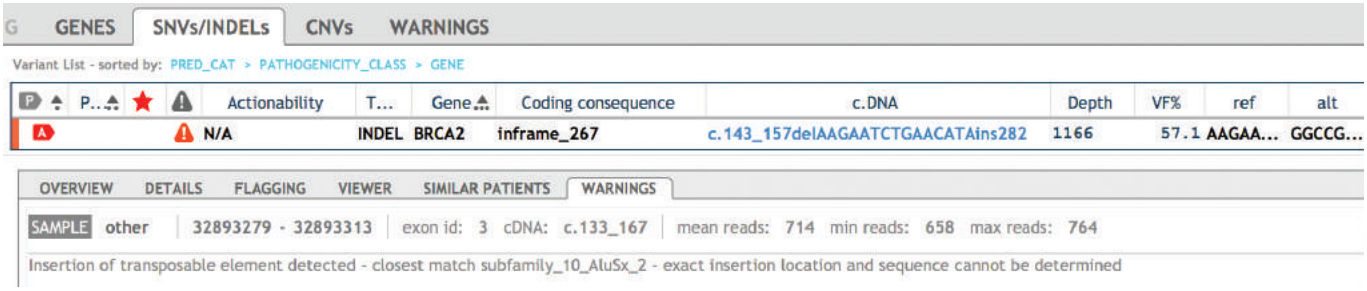


Figure 4: Display of ALU insertion warning in SOPHiA DDM.

CNV DETECTION

The term "Copy Number Variation" (CNV) refers to a difference in the dosage of genes or genomic fragments when compared with a normal reference human genome. The term was originally used to describe genomic fragments, which ranged in size from at least 1kb to several million bases that had a variable copy number. However, higher resolution CNV analysis has revealed the existence of increasingly smaller CNVs (100 to 1,000 bp) in the human genome. SOPHiA through MUSKAT™, our advanced machine learning based algorithm, is able to detect CNVs without having any control samples.

MUSKAT™

MUSKAT enables the detection of any CNV up to a resolution of 200 bp. Common CNV detection techniques can only detect long rearrangements. The CNV analysis is performed by analyzing the coverage levels of the target regions across samples within the same run. For each sample, the algorithm automatically selects a set of reference samples from the same run, based on the similarity of coverage patterns.

Moreover, using the reference samples, the coverage is normalized by sample and by the target region, and CNV calling is performed using a Hidden-Markov-model algorithm.

As a result, the most probable copy number for each target region is determined. Additionally, a confidence level is calculated both at the sample level and at the target-region level. To assess the sensitivity of our CNV detection algorithm, we performed 5 sequencing runs using the HCS on an Illumina MiSeq instrument with in total 59 distinct samples. 23 of these samples contained MLPA-confirmed CNVs.

In this study (data not shown) we have demonstrated that:

- All MLPA-confirmed CNVs were correctly identified with HCS
- Observed sensitivity, specificity, robustness and accuracy of CNV detection always reached 100%
- Performing CNV detection on inter-run replicates and reference samples analyzed with HCS yielded the same list of CNVs
- The results obtained by analyzing twice the raw data were strictly identical to each other

PSEUDOGENE HANDLING

SOPHiA can easily differentiate between gene and pseudogene variants. Nevertheless, those variants still need to be interpreted with care: even if a variant can clearly be assigned *PMS2* or *PMS2*-like by sequence, a definite *PMS2* or *PMS2CL* gene location cannot be assigned, due to the high rate of gene conversion in the homologous regions. In particular, exon 11 through 15 of the *PMS2* gene have highly homologous counterparts in the pseudogene *PMS2CL*. The differences in the reference genome between the corresponding regions of the gene and the pseudogene are named Paralogous Sequence Variants (PSV). In gene conversion occurrences, PSVs can be exchanged between the gene and the pseudogene locations. Therefore, a special consideration is needed when assigning reads, variants and CNVs to the gene or the pseudogene.

In SOPHiA DDM, it is possible to distinguish the following three types of PSVs (Figure 5):

- Stable PSVs that only rarely undergo gene conversion
- Group-polymorphic PSVs (group of neighboring PSVs on the scale of an exon or larger) that often undergo gene conversion and are frequently linked together
- Individual-polymorphic PSVs that undergo gene conversion events frequently and independently of the neighboring PSVs

When, a gene conversion event is detected, the user is given a warning in SOPHiA DDM.

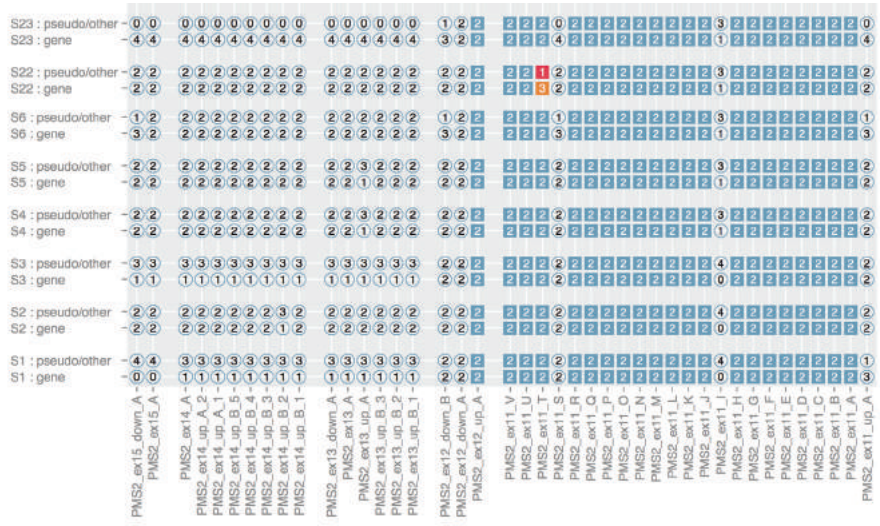


Figure 5: Display of stable, group-polymorphic and individual-polymorphic PSVs in SOPHiA DDM. Solid squares correspond to stable PSVs. Blue squares denote normal copy number of the allele while red or orange squares show reduced or increased copy numbers, respectively. A gene-conversion event in any given PSV would appear as a combination of an increased copy number in one allele and a reduced copy number in the other. Empty circles correspond to group-polymorphic and individual-polymorphic PSVs. For these categories the color corresponds to the variation of the total copy number. A red or orange color indicates a copy number variation (CNV) either in the gene or in the pseudogene region.

Summary

The Hereditary Cancer Solution by SOPHiA GENETICS (HCS) bundles a superior capture-based target enrichment kit with SOPHiA AI and full access to SOPHiA DDM platform. The solution covers the coding regions of the 26 most clinically relevant hereditary cancer genes.

We demonstrate that SOPHiA allows to reach unmatched performance. Hence, the Hereditary Cancer Solution by SOPHiA GENETICS is the most accurate application for clinical routine diagnostics.

Our solution offers:

- High on-target rates and coverage uniformity with very few low coverage regions
- Close to perfection analytical performance values
- CNV detection in all genes of the panel
- Detection of difficult variants like long Indels and ALU insertions
- Differentiation of *PMS2* and *PM2CL* variants

PRODUCT	PRODUCT CODE
Hereditary Cancer Solution by SOPHiA GENETICS (48 rxn) CE-IVD (for use on Illumina MiSeq®)	BSC.K0012. L001.048.C.I001.GL
Hereditary Cancer Solution by SOPHiA GENETICS (48 rxn) CE-IVD (for use on Illumina NextSeq®)	BSC.K0012. L001.048.C.I004.GL
Hereditary Cancer Solution by SOPHiA GENETICS (32 rxn) CE-IVD (for use on Illumina MiSeq®)	BSC.K0011. L001.032.C.I001.GL
Hereditary Cancer Solution by SOPHiA GENETICS (48 rxn) RUO (for use on Thermo Fisher Scientific Ion PGM™)	BSC.K0015. L001.048.R.I007.GL
Hereditary Cancer Solution by SOPHiA GENETICS (48 rxn) RUO (for use on Thermo Fisher Scientific Ion Proton™)	BSC.K0015. L001.048.R.I008.GL
Hereditary Cancer Solution by SOPHiA GENETICS (48 rxn) RUO (for use on Thermo Fisher Scientific Ion S5™)	BSC.K0015. L001.048.R.I009.GL

References

1. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V, Wallace A, EuroGentest Validation Group (2010)

2. Clopper C, Pearson ES (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404-413

3. Smith TM, Lee MK, Szabo CI, Jerome N, McEuen M, Taylor M, Hood L, King MC. Complete Genomic Sequence and Analysis of 117kb of Human DNA Containing the Gene BRCA1. Cold Spring Harbor Laboratory Press (1996). ISSN1054-9803/96