Agilent HaloPlex Target Enrichment and SureCall Data Analysis: Optimized for the Ion Torrent™ PGM Sequencer

Application Note

Authors

Josh Zhiyong Wang,
Henrik Johansson, Anniek De Witte,
Elin Agne and Karin Zetterman
Agilent Technologies

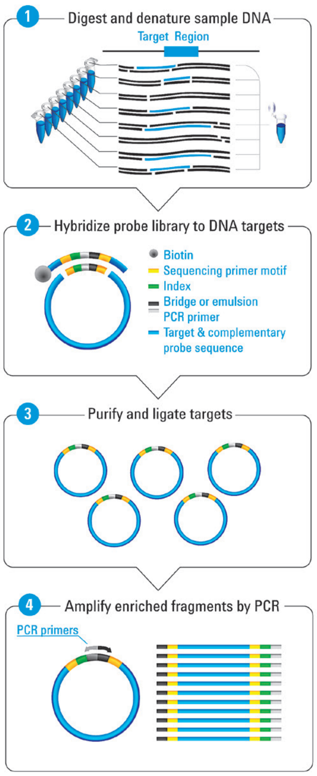

1
Digest and denature sample DNA
Target Region
2
Hybridize probe library to DNA targets
Biotin
Sequencing primer motif
Index
Bridge or emulsion PCR primer
Target & complementary probe sequence
3
Purify and ligate targets
4
Amplify enriched fragments by PCR
PCR primers

Figure 1. HaloPlex workflow
1) Digest and denature sample DNA;
2) Hybridize probe library;
3) Purify and ligate targets;
4) PCR amplify enriched fragments

Introduction

The Agilent HaloPlex target enrichment system is a fast and highly efficient next generation sequencing target enrichment tool. This unique technology requires only 200 ng of input DNA, and the entire protocol can be completed in 6–24 hours (Figure 1). Multiple amplicons with different start and stop coordinates are typically designed to target each individual target region, resulting in sequence confirmation at individual bases from several amplicons, thus achieving high accuracy within the same assay. The HaloPlex target enrichment system has previously been optimized using Illumina sequencing platforms, such as the HiSeq and MiSeq systems, and has now been optimized for use with the Ion Torrent™ Sequencer.

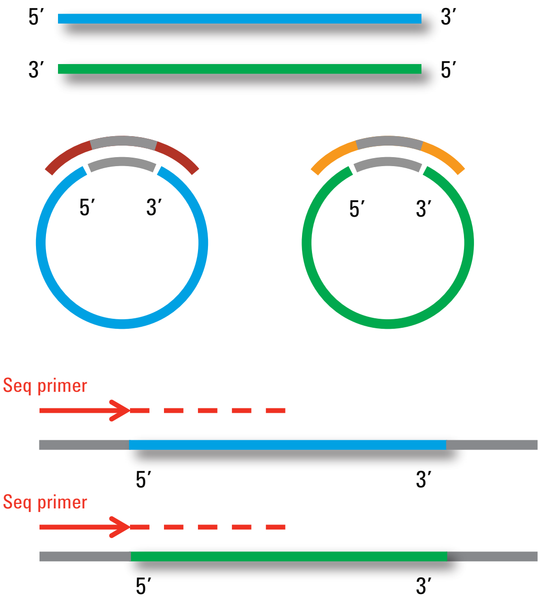
This application note focuses on the performance of the Agilent HaloPlex target enrichment system on the PGM sequencer when used in combination with Agilent's SureCall data analysis software.

HaloPlex Design and Performance on the PGM

HaloPlex "paired end by design" technology produces the high coverage and sequencing efficiency of traditional paired end sequencing without the complex and cumbersome 3-step workflow. Traditional paired end sequencing on the PGM first requires users to perform an initial sequencing run. Then the user must remove the chip and complete a series of enzymatic steps. Finally the user returns the chip onto the instrument and sequences in the reverse direction. Adding to the complexity, this protocol also requires use of a modified primer during the library preparation step prior to the initial sequencing run.

In contrast, Agilent's HaloPlex protocol for the PGM utilizes a "paired end by design" strategy with a single end sequencing run resulting in high coverage and sequencing efficiency with no intermediate manual intervention steps (Figure 2). For each target fragment produced through enzymatic digestion two probes are generated, one for the sense target strand and the other for the antisense target strand. This results in post-capture PCR products from both strands of the same target having their 5' end next to the sequencing primer. When standard single end sequencing is performed on the PGM, sequencing reads will be generated from both ends of the target producing excellent coverage. With this method, Agilent HaloPlex target enrichment attains the high quality results expected from a true paired-end strategy.

HaloPlex "paired end by design" technology



a) Both strands of one digested fragment

b) HaloPlex probes are designed to both strands, different HaloPlex probes hybridizing to different strand of digested fragment

c) Captured amplicons for sequencing with same primer from both strands

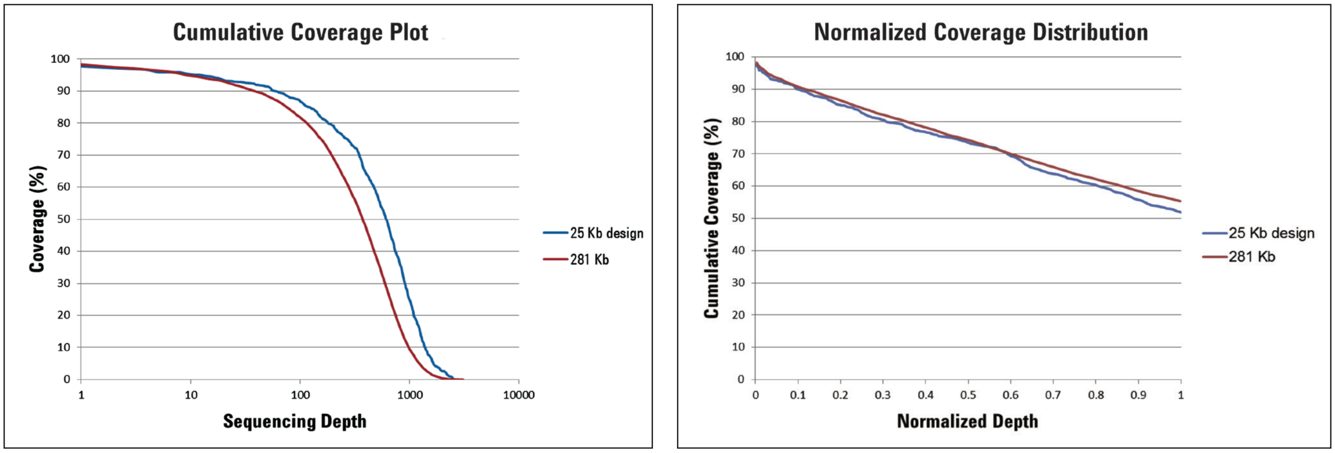
Figure 2. HaloPlex "paired end by design" technology produces the high coverage and sequencing efficiency of traditional paired end sequencing without the complex and cumbersome workflow.

Benefits of HaloPlex technology.

There are several benefits to HaloPlex technology which set it apart from other amplicon based target enrichment methods:

1. Excellent design coverage achieved through a careful selection of restriction enzyme combinations and strong empirical coverage.
2. Multiple amplicons covering the same base position offers intra-assay sequence confirmation.
3. FFPE sample compatibility[1].
4. Up to 2.5 Mb of target region or up to 200 k probes can be processed in a single capture reaction.

As shown in Figure 3, two HaloPlex PGM gene panels were designed to target 25 kb and 281 kb coding exons respectively, and 0.2 million and 1.38 million reads were generated. The end result was 690x and 461x respective average read depths in the regions of interest with over 93% of bases at >20x coverage as well as 81–86% of bases with >100x coverage. Overall, about 90% of bases achieved 10% of average sequencing depth in these two panels, 69x and 46x respectively, a very typical performance for a custom HaloPlex gene panel.



Probe set	Sample	Raw Reads	Pre-processed Reads	Aligned Total	Specificity Region	Specificity ROI	Average Depth Targeted Region	Average Depth ROI	Coverage at 10% of average depth	Coverage at 20% of average depth	Coverage ≥ 1X	Coverage ≥ 10X	Coverage ≥ 20X	Coverage ≥ 100X
25 Kb design	Na12878	2.02E+05	183230	86.76%	91.05%	84.56%	587.2	689.9	89.97%	85.04%	97.68%	95.29%	93.45%	86.82%
281 Kb design	Na12878	1.38E+06	1360087	93.67%	98.97%	77.83%	324.6	460.8	90.64%	86.53%	98.30%	94.83%	93.01%	81.84%

Figure 3. HaloPlex performance metrics on Ion PGM

Special considerations for using HaloPlex with the PGM

Before sequencing on a PGM, a prepared DNA library must be processed using the Ion OneTouch template preparation kit. Life Technologies routinely makes changes to these kits, and there are three special considerations that HaloPlex users must take into account for a successful PGM sequencing experiment:

First, library size selection for emulsion PCR is not required when using HaloPlex. The design selects amplicons in the range of 50–450 bp, and the final HaloPlex library size should be 150–550 bp in length, with ˜ 100 bp from sequencing motifs.

Despite diminished emulsion PCR efficiency for fragments >330 bp in the Ion 200 bp template preparation kit, size selection to enrich fragments <330 bp is not required with HaloPlex. In fact, trimmed read lengths after PCR as long as 310 bp have been observed, implying respective amplicon lengths of 360–410 bp. Other sequencing metrics such as polyclonal fraction and low quality read fraction from HaloPlex library run reports are well within acceptable ranges and comparable to those observed in Ion Torrent AmpliSeq libraries, in which all library fragments are <330 bp.

To support this claim, empirical data has been generated from multiple panels using HaloPlex target enrichment and standard Ion Torrent template preparation kit. Using the HaloPlex "paired end by design" strategy employing multiple (as many as 16) probes to cover each base position with a combination of long and short probes on both strands, this data shows base coverage within the designed input target region as high as >98% with at least 1X coverage.

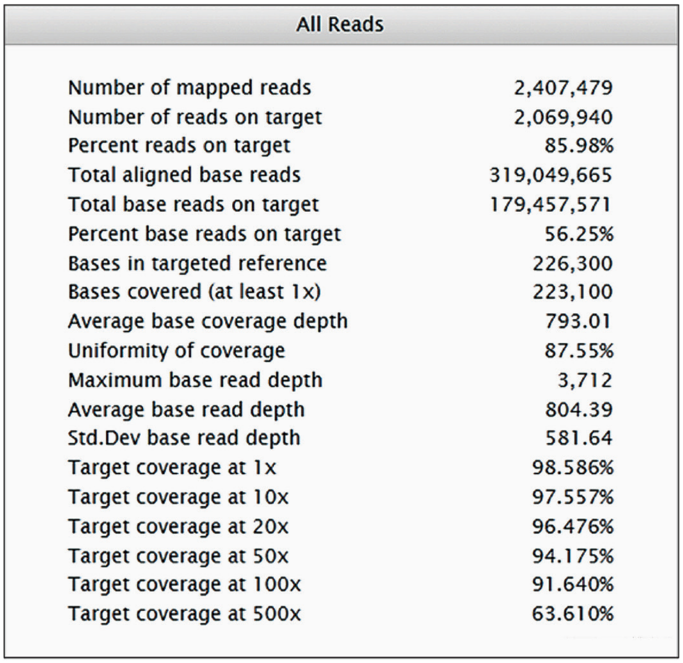
Second, PGM users should choose Ion template preparation kits and sequencing kits which have been validated by Agilent for use with HaloPlex technology to ensure success. Table 1 lists some of these template preparation systems and sequencing kits as well as their Agilent validation status.

Table 1. Related systems and reagent kits for PGM and their compatibility with HaloPlex

Template preparation system	OneTouch	OneTouch DL	OneTouch 2
Template preparation kit	200 template kit v2	200 template kit v2 DL (4480285)	OT2 200 kit (4480974)
Sequencing system	PGM	PGM	PGM
Sequencing kit	200 kit (4474004)	200 kit (4474004)	200 kit v2 (4482006)
Validation with HaloPlex	**No**	**Yes**	**Yes**

Third, it is critical for HaloPlex users to set up their PGM instrument with Torrent Suite v3.2 or newer and to remain up to date with the latest software updates from Life Technologies. As a companion primary and secondary analysis pipeline to the PGM, Torrent Suite software v3.2 offers significant improvements in output file type and generation, base calling, alignment and variant calling, providing higher accuracy for SNP identification and improved indel detection.

Users should also note that when setting up a run with HaloPlex adaptor and library information, guidelines provided in the Agilent HaloPlex protocol should be followed. During a HaloPlex run, one can specify targeted regions using Torrent Suite analysis pipeline resulting in streamlined mapping and on target analysis. Figure 4 shows a HaloPlex run on a PGM with input exonic target regions of 226 kb. Although the total length of designed amplicons in final library cover >500 kb, sequencing reads on target coverage of 86% is achieved as well as on target base coverage of about 56%.



All Reads	
Number of mapped reads	2,407,479
Number of reads on target	2,069,940
Percent reads on target	85.98%
Total aligned base reads	319,049,665
Total base reads on target	179,457,571
Percent base reads on target	56.25%
Bases in targeted reference	226,300
Bases covered (at least 1x)	223,100
Average base coverage depth	793.01
Uniformity of coverage	87.55%
Maximum base read depth	3,712
Average base read depth	804.39
Std.Dev base read depth	581.64
Target coverage at 1x	98.586%
Target coverage at 10x	97.557%
Target coverage at 20x	96.476%
Target coverage at 50x	94.175%
Target coverage at 100x	91.640%
Target coverage at 500x	63.610%

Figure 4. Coverage analysis in targeted region from Torrent Suite software

Agilent SureCall Data Analysis Software

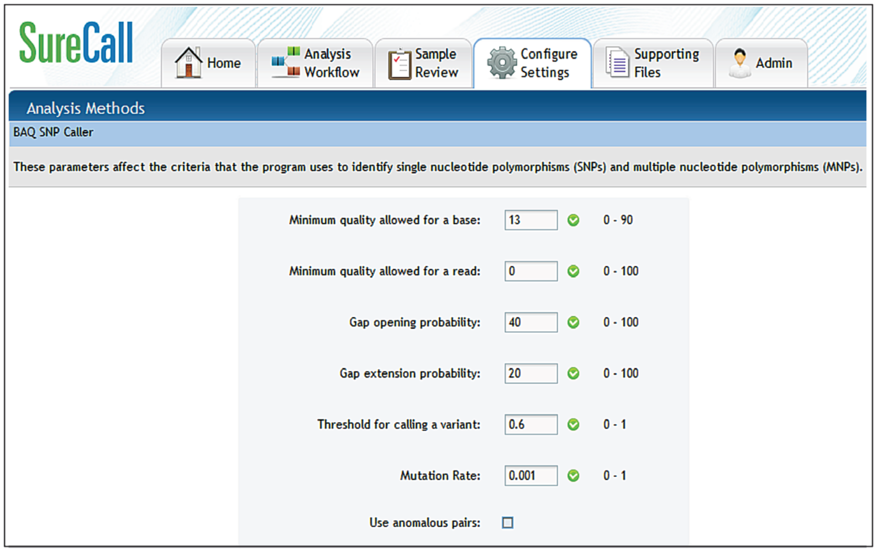
The final step of a targeted sequencing experiment is data analysis, especially tertiary analysis to determine the biological implications of the underlying data. To manage this analysis, Agilent has developed SureCall data analysis software designed specifically to detect and report mutations in the target regions of HaloPlex sequencing data[2]. It is provided to HaloPlex users free of charge.

SureCall can effectively analyze either aligned BAM files or unmapped files originating from Torrent Suite software using SAMtools algorithms for SNP calling. With powerful built-in tertiary analysis tools, SureCall classifies each identified mutation into a distinct category based on mutation attributes and on known annotation in prominent genome databases including NCBI clinSNP, COSMIC and GWAS disease catalog. With identified SNP/indel variants displayed in the mutation table and easy scrolling and filtering functions, the embedded genome viewer allows visualization of individual sequencing reads, read depth, number of quality reads (reads that pass the specified quality filter), and coverage at selected variant locations.

PGM users may experience difficulty in obtaining accurate sequence information through stretches of homopolymers as well as difficulty with the Ion PGM's tendency to generate false positive indel calls. It has been noted that using the PGM sequencing kit introduced in May 2012, the false positive indels occur at a frequency of about 0.32 per 100 bp raw sequence, or 320,000 per 100 Mb raw sequence[3]. More recently introduced PGM sequencing kits would likely have improved performance metrics; however, indel and homopolymer miscalls caused by a combination of sequencing chemistry and software issues are still commonly observed in PGM sequencing data. During an Ion PGM sequencing run, the chemistry generates signals of increasing similarity for homopolymers of increasing length, and the software could have difficulty distinguishing these increasingly similar signals. Therefore, it is important for an Ion PGM user to be prepared to adjust the parameter settings in the Torrent Suite variant caller and to manually examine all identified indel mutations.

Within Torrent Suite variant caller software, parameters in the dibayes section affect only SNPs, while parameters in other sections (torrent-variant-caller, torrent-variant-caller-highcov, long-indel-assembler and filter-indels) affect only indels. For example, increasing thresholds for "min_mapping_quality_score" and/or "min_allele_frequency" could increase the specificity of the indel mutation identified.

Similarly, SureCall software features its own BAQ SNP caller in which parameters can be adjusted to increase the specificity for indel mutation identification (Figure 5). Examples include "Minimum quality allowed for a base", "Gap opening probability", and "Gap extension probability." "Gap opening probability" refers to the probability for an error to occur in the the gap opening sequence. SureCall's initial estimate for the likelihood of a missing base in the sequence in the start of the indel. The permitted value range for this parameter is 0–100 with a default value of 40, indicating 40% likelihood. Increasing this value allows for increasing stringency in identifying indels.



Analysis Methods

BAQ SNP Caller

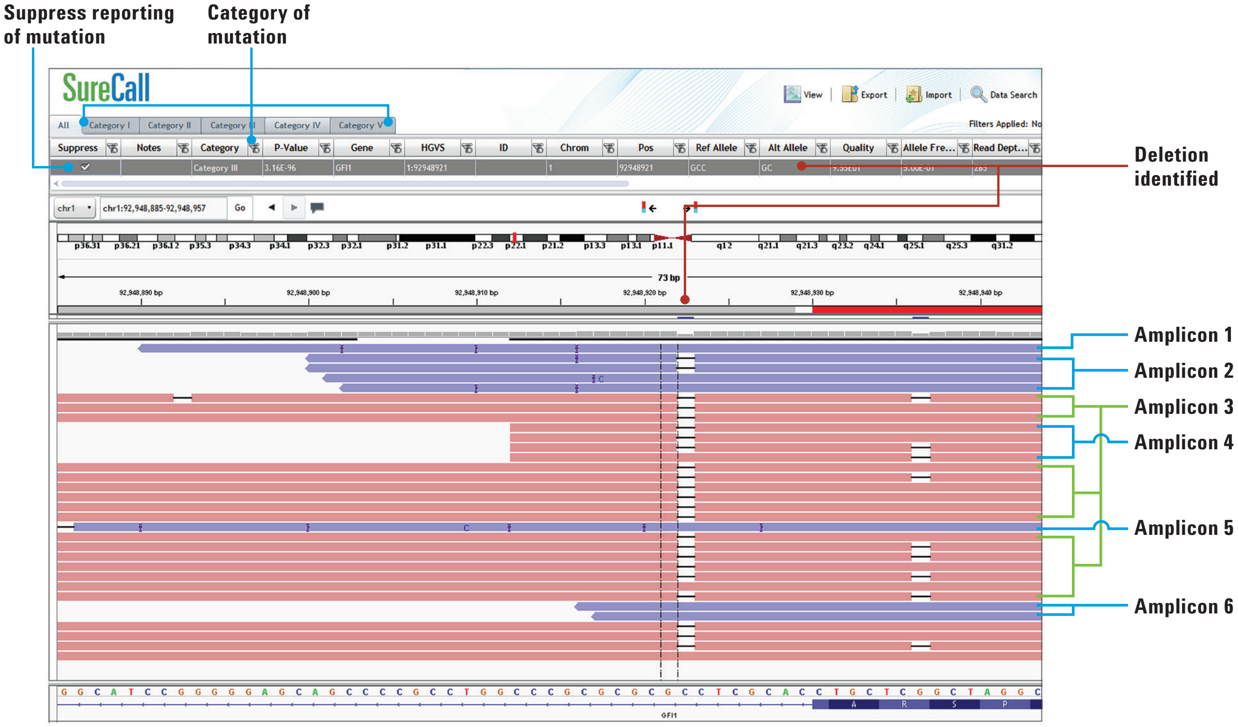
These parameters affect the criteria that the program uses to identify single nucleotide polymorphisms (SNPs) and multiple nucleotide polymorphisms (MNPs).

Minimum quality allowed for a base:	13	✓	0 - 90
Minimum quality allowed for a read:	0	✓	0 - 100
Gap opening probability:	40	✓	0 - 100
Gap extension probability:	20	✓	0 - 100
Threshold for calling a variant:	0.6	✓	0 - 1
Mutation Rate:	0.001	✓	0 - 1
Use anomalous pairs:	□		

Figure 5. Parameters in SureCall BAQ SNP Caller

"Gap extension probability" refers to the probability of an error in the gap extension sequence, which determines SureCall's initial estimate of the likelihood that a missing base in the sequence is a continuation of an indel. The permitted value range for this parameter is 0–100, and the default value is 20, or 20% likelihood for an error. Increasing this value allows for increasing stringency in identifying indels while reducing this value facilitates the identification of longer indels. Users must exercise caution when reducing the "Gap extension probability" as this may lead to the identification of false positives. Finally, indel calling can be disabled altogether.

For indels encountered using the Ion PGM system, HaloPlex technology combined with SureCall software offers a straightforward solution to examine and suppress false positive indel calls. Within SureCall, one can easily identify indels by filtering on the ID column in the mutation table. Figure 6 illustrates an example of a deletion mutation from a constitutional sample that was identified in sequencing data. The mutation was present for the indicated base position in reads from three different HaloPlex amplicons (amplicons 2, 3, 4) but not in the reads from the other three different HaloPlex amplicons (amplicons 1, 5, 6) that also covered the indicated base position. Therefore, it is a false positive indel mutation and can easily be identified using the "Suppress" function. (Note that amplicons 3 and 4 are from the positive strand orientation, while amplicons 1, 2, 5 and 6 are from the negative strand orientation, as HaloPlex probes for the Ion PGM are designed for both strands, per the "paired-end by design" strategy).


Suppress reporting of mutation
Category of mutation
Deletion identified
Amplicon 1
Amplicon 2
Amplicon 3
Amplicon 4
Amplicon 5
Amplicon 6
SureCall
View
Export
Import
Data Search
All
Category I
Category II
Category III
Category IV
Category V
Filters Applied: No
chr1
chr1:92,948,885-92,948,957
Go
p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 p11.1 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2
73 bp
92,948,890 bp
92,948,900 bp
92,948,910 bp
92,948,920 bp
92,948,930 bp
92,948,940 bp
G G C A T C C G G G G G A G C A G C C C C G C C T G G C C C G C G C G C C T C G C A C C T G C T C G G C T A G G C
A R S P
GFI1

Suppress	Notes	Category	P-Value	Gene	HGVS	ID	Chrom	Pos	Ref Allele	Alt Allele	Quality	Allele Fre...	Read Dept...
✓		Category III	3.16E-96	GFI1	1:92948921		1	92948921	GCC	GC	9.55E01	5.00E-01	285

Figure 6. Suppressing a false positive indel mutation in a constitutional sample using SureCall software to analyze an Ion PGM sequencing run. The red and blue colors represent reads from the + or – strand, respectively.

Conclusions

The Agilent HaloPlex target enrichment protocol for the Ion Torrent PGM sequencer provides high sequence coverage and high accuracy while eliminating the need for library size selection. For a streamlined workflow from design to data analysis, Agilent SureCall software offers a simple and effective data analysis option free of charge.

Special considerations should be taken when performing HaloPlex experiments with an Ion PGM, including the use of Ion Torrent platform specific reagent kits (emulsion PCR kits and sequencing kits) and proper setup of Torrent Suite software.

For commonly occurring indel mutations in Ion PGM data, HaloPlex "paired-end by design" technology produces superior indel mutation confirmation through analysis of both strands and through the use of multiple amplicons covering the same bases within identified indels. Unlike actual paired-end sequencing on the Ion PGM, this all occurs within the same assay. SureCall software provides an intuitive sequencing analysis platform to monitor, edit and suppress identified indel mutations from Ion PGM data as necessary.

SureCall can be downloaded at www.agilent.com/genomics/surecall

References

1. "HaloPlex Target Enrichment from FFPE Tissues". Agilent Application Note Publication Number 5991-0666EN, 2012.

2. "Agilent HaloPlex Target Enrichment– Design and Analysis of Clinical Research Panels". Agilent Application Note Publication Number 5991-1919EN, 2013

3. "Performance Comparison of Benchtop high-throughput Sequencing Platforms" Nature Biotechnology, 30, 434-439, 2012

Find a local Agilent customer center
www.agilent.com/genomics/contactus

USA and Canada
1-800-227-9770
Agilent_inquiries@agilent.com

Europe
Info_agilent@agilent.com

Asia Pacific
Inquiry_isca@agilent.com

Learn more about HaloPlex:
www.agilent.com/genomics/haloplex

Learn more about SureCall:
www.agilent.com/genomics/surecall

Ion Torrent™ PGM™ are trademarks of Life Technologies.

For Research Use Only. Not for Use in Diagnostic Procedures.

This information is subject to change without notice.

Published in USA, April 19, 2013
Publication Number 5991-2229EN