Agilent HaloPlex Target Enrichment System

Design and Analysis of Clinical Research Panels

Application Note

Authors

Anniek De Witte, Ashutosh,
Christian Le Cocq, Jayati Ghosh

Introduction

As Next Generation Sequencing (NGS) becomes more affordable and transitions to use in clinical research for mutation detection, it is essential for researchers to have an efficient end-to-end workflow, including target enrichment panel design, sample preparation, sequencing, and data analysis. Agilent's HaloPlex technology for target enrichment, SureDesign application for panel design, and SureCall application for data analysis combine with today's desktop sequencers to make this workflow a reality.

HaloPlex is a target enrichment system ideally suited to deep sequencing of relatively small panels of genes (up to 5 Mb), such as those required for studies of inherited disorders, cancer, or infectious disease. The Agilent HaloPlex Target Enrichment System enables fast, simple, and efficient analysis of genomic regions of interest for a large number of samples, covering thousands of exons per sample. HaloPlex is compatible with different desktop sequencing and high-throughput platforms.

SureDesign is a web-based design application for designing custom HaloPlex panels. With this tool, researchers can quickly generate high-coverage, high-efficiency designs for targeted re-sequencing.

SureCall is an easy-to-use desktop application combining best in class open source algorithms for end-to-end NGS data analysis from alignment to categorization of mutations.

In this application note more details are provided on how to create panels for clinical research with HaloPlex and how to analyze the resulting sequencing data.

SureDesign

The web-based SureDesign application was used to create HaloPlex custom panels to enrich the coding sequences of genes of interest. Three custom research panels were designed: one for cardiac disease, one for Noonan spectrum disorder, and one for collagen tissue disease.

Figure 1 describes the SureDesign workflow.

The first step is to enter target genes, select the options for regions of interest (ROI) e.g. exons plus flanking bases or UTRs, and specify which genomic databases to use. In this case, the ROI for the designs included all exons of the target genes plus 10 bp of flanking intronic sequence. The exact genomic regions can be reviewed in a design summary and viewed in a genome browser before starting probe design.

In the next step, HaloPlex probes are selected by SureDesign to achieve 3 – 4 fold redundant coverage of the ROI. The probes and a PDF report (Figure 2) are generated within a few minutes. The report contains summary information on the design and the coverage for each target region. For all panels the coverage was over 98% (Table 1).

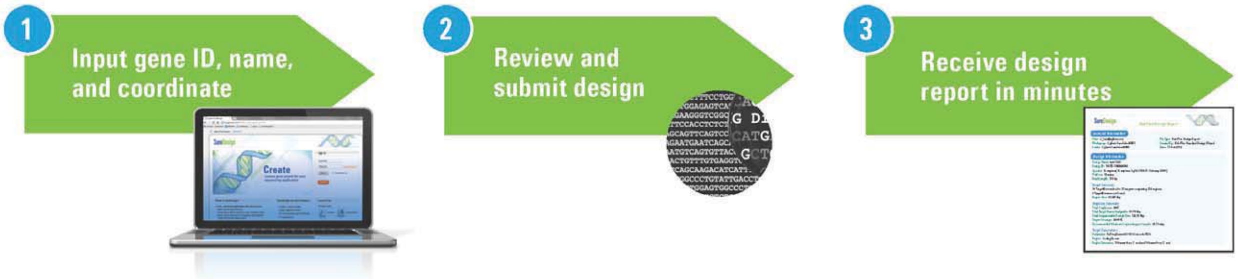
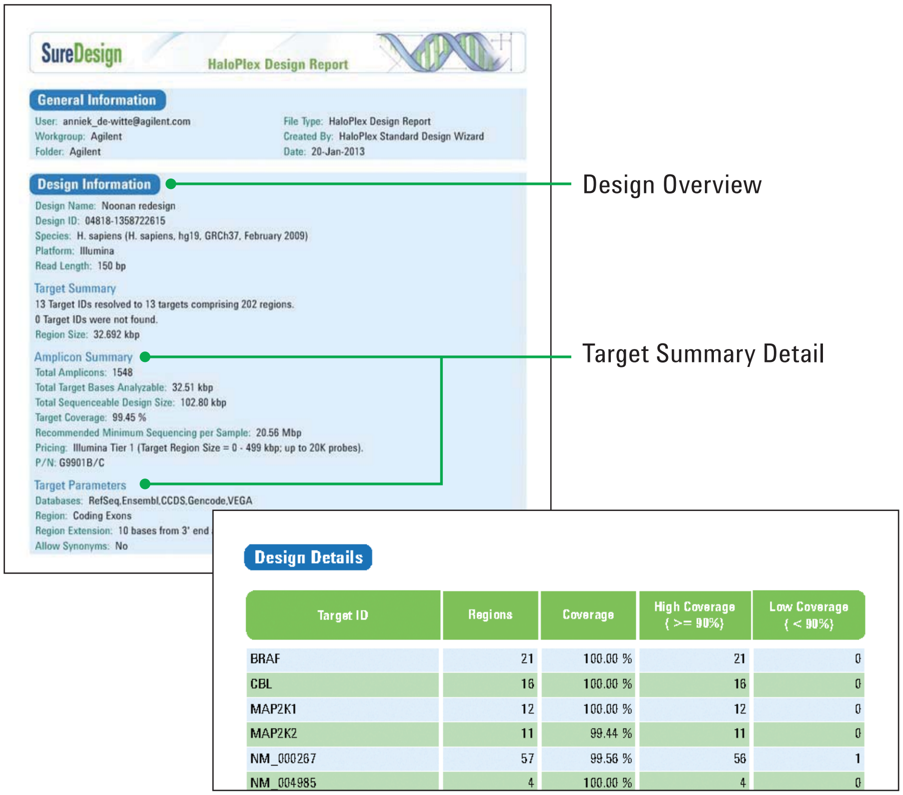

1
Input gene ID, name, and coordinate
2
Review and submit design
3
Receive design report in minutes

Figure 1. SureDesign workflow. A HaloPlex design can be created in a couple of minutes.



SureDesign HaloPlex Design Report

General Information

User: anniek_de-witte@agilent.com
Workgroup: Agilent
Folder: Agilent
File Type: HaloPlex Design Report
Created By: HaloPlex Standard Design Wizard
Date: 20-Jan-2013

Design Information

Design Name: Noonan redesign
Design ID: 04818-1358722615
Species: H. sapiens (H. sapiens, hg19, GRCh37, February 2009)
Platform: Illumina
Read Length: 150 bp

Target Summary

13 Target IDs resolved to 13 targets comprising 202 regions.
0 Target IDs were not found.
Region Size: 32.692 kbp

Amplicon Summary

Total Amplicons: 1548
Total Target Bases Analyzable: 32.51 kbp
Total Sequenceable Design Size: 102.80 kbp
Target Coverage: 99.45 %
Recommended Minimum Sequencing per Sample: 20.56 Mbp
Pricing: Illumina Tier 1 (Target Region Size = 0 - 499 kbp; up to 20K probes).
P/N: G9901B/C

Target Parameters

Databases: RefSeq,Ensembl,CCDS,Gencode,VEGA
Region: Coding Exons
Region Extension: 10 bases from 3' end
Allow Synonyms: No

Design Details

Target ID	Regions	Coverage	High Coverage (>= 90%)	Low Coverage (< 90%)
BRAF	21	100.00 %	21	0
CBL	16	100.00 %	16	0
MAP2K1	12	100.00 %	12	0
MAP2K2	11	99.44 %	11	0
NM_000267	57	99.56 %	56	1
NM_004985	4	100.00 %	4	0

Figure 2. Example PDF report from SureDesign for a Noonan spectrum disorder HaloPlex custom design.

HaloPlex and Sequencing

Genomic DNA (gDNA) was extracted from eight samples: two cardiac disease, three Noonan spectrum disorder, and three collagen tissue disease. The HaloPlex target enrichment system was used to enrich for the genomic regions of interest. The HaloPlex system uses a single-tube target amplification and removes the need for library preparation to reduce total sample processing time to only 8 hours, without the need for dedicated instrumentation (Figure 3). HaloPlex is compatible with multiple desktop sequencing and high-throughput platforms. In this case, the samples were pooled (5 samples) and run on an Illumina MiSeq Personal Sequencer.

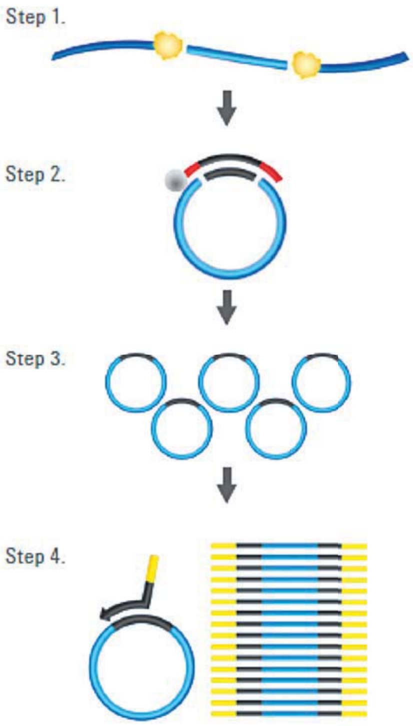


Figure 3. HaloPlex workflow. The HaloPlex protocol is accomplished in 4 basic steps:
Step 1) Digest and denature sample DNA
Step 2) Hybridize oligonucleotide probe library
Step 3) Purify and ligate targets
Step 4) Amplify targeted fragments with PCR

Data Analysis in SureCall

Data was analyzed using Agilent SureCall software. SureCall addresses the critical need for an easy-to-use analysis tool that incorporates the most widely accepted open source libraries and algorithms. Analysis in SureCall begins with raw reads from Illumina HiSeq/MiSeq or Ion Torrent sequencing of genomic DNA enriched with HaloPlex (Figure 4). After removal of the adaptor sequences, the reads are aligned to the genome (BWA or TMAP). SAMTools is used to recalibrate the base call quality scores, perform local realignment, and index the reads for improved performance. SAMTools is also used to identify mutations from the local read pileup at each location and to assess the significance of the mutations. Several tools are then used to provide input for the mutation classification.

Each mutation is evaluated based on its location, amino acid change, effect on protein function (SIFT), and impact on structure and function of the protein (PolyPhen-2). Further information regarding the mutation is then aggregated from various public sources, including NCBI, COSMIC (Catalog of Somatic Mutations in Cancer), PubMed, and Locus Specific Databases. After collecting the various inputs for classification, the proprietary mutation classifier evaluates the significance of the mutation following default or customized guidelines. Each mutation is then categorized, with the user triaging each mutation and reviewing supporting evidence in the built-in viewer, including raw data and confidence measures, as well as links to external databases such as OMIM and dbVar.

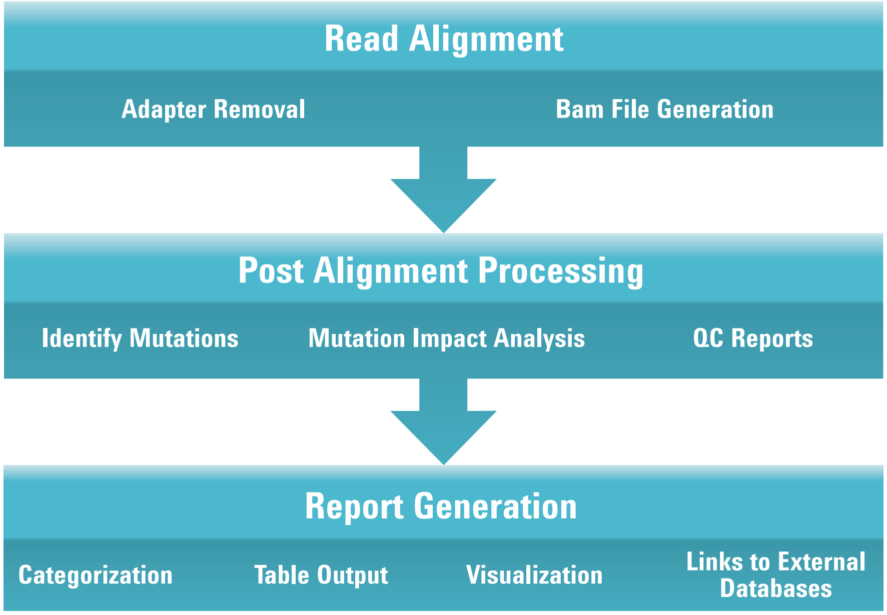


Figure 4. Analysis workflow in SureCall.

QC Metrics and Triage View

SureCall is a simple 3-step workflow based software. After selecting and describing the samples, the analysis is run. SureCall comes preloaded with a default analysis method, but custom analysis methods can also be added.

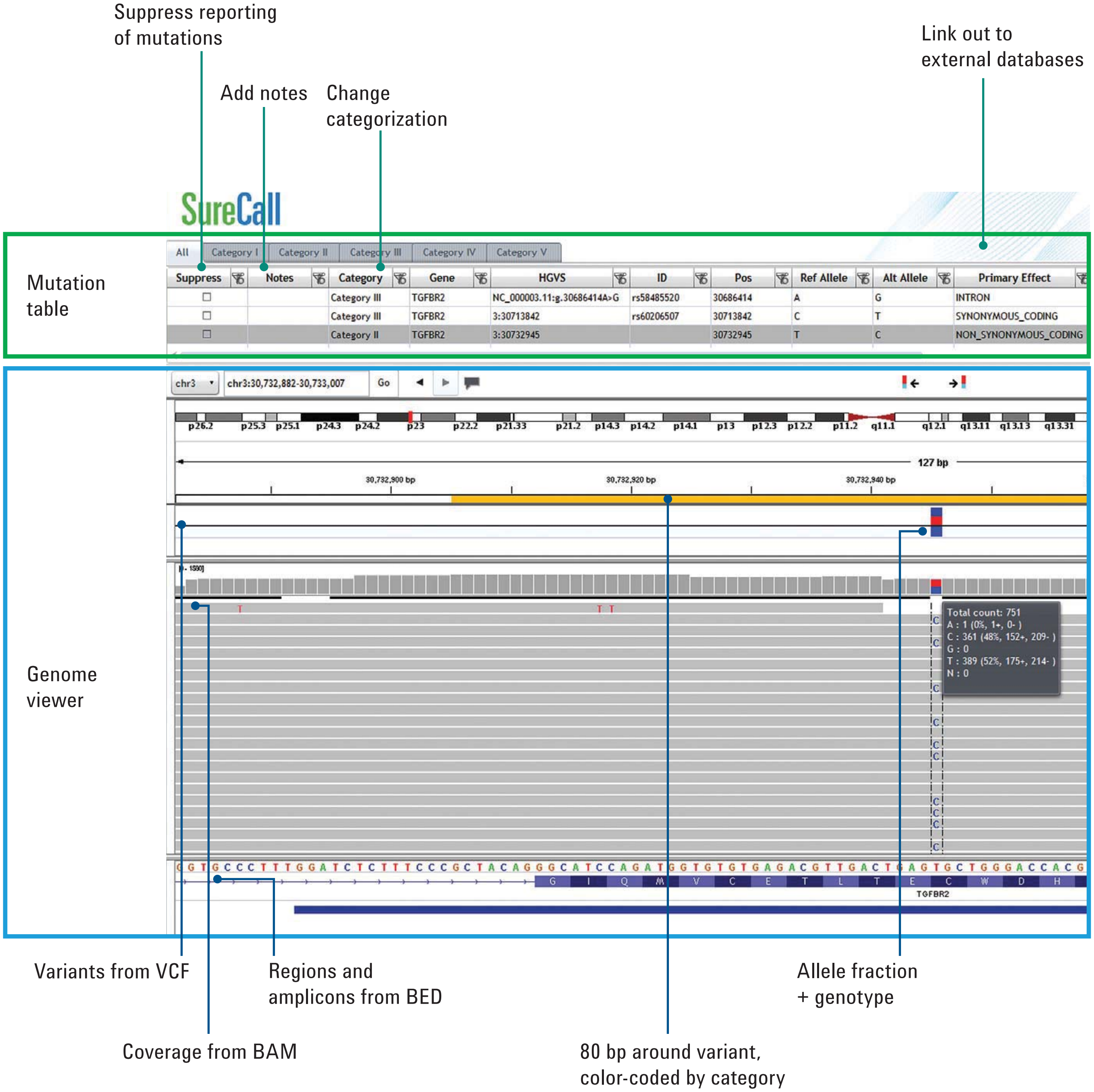
Once the program has analyzed the samples, a QC report can be generated for each sample. Table 1 contains a summary of the key QC metrics from SureCall, calculated for the eight gDNA samples enriched with HaloPlex. The raw data and sample analysis can be manually inspected in the Triage View (Figure 5). From the Triage View, researchers can suppress the reporting of mutations, change mutation categorization assignments, look up annotation information for a mutation in external databases, and add notes to mutations. Table 2 shows the links to external databases that are available in SureCall. Researchers can also compare mutations in a sample with mutations present in other samples. An audit trail is kept of all changes that are made to the analysis of a sample.

Panel	Number of genes	Coverage in SureDesign	% reads in covered regions	% regions with zero coverage	% analyzable target regions covered by > 20 reads
Cardiac disease	20	99.3 %	88.8 % ± 0.2 %	0.2 % ± 0.1 %	98.9 % ± 0.4 %
Noonan spectrum disorder	13	98.6 %	85.9 % ± 0.4 %	0.4 % ± 0.1 %	98.8 % ± 0.3 %
Collagen tissue disease	34	98.2 %	88.6 % ± 0.2 %	0.7 % ± 0.2 %	97.1 % ± 0.4 %

Table 1. Coverage of target regions from SureDesign and summary of QC metrics generated in SureCall for two cardiac disease, three Noonan spectrum disorder, and three collagen tissue disease samples.

Right click on column in SureCall	Database
ID	dbSNP
Gene Name	OMIM GeneCard
Position	dbVar
Transcript	NCBI
Transcript ID	Ensembl
Uniprot	Uniprot

Table 2. Links to external databases available from SureCall by right-clicking on specific columns in the triage view table.



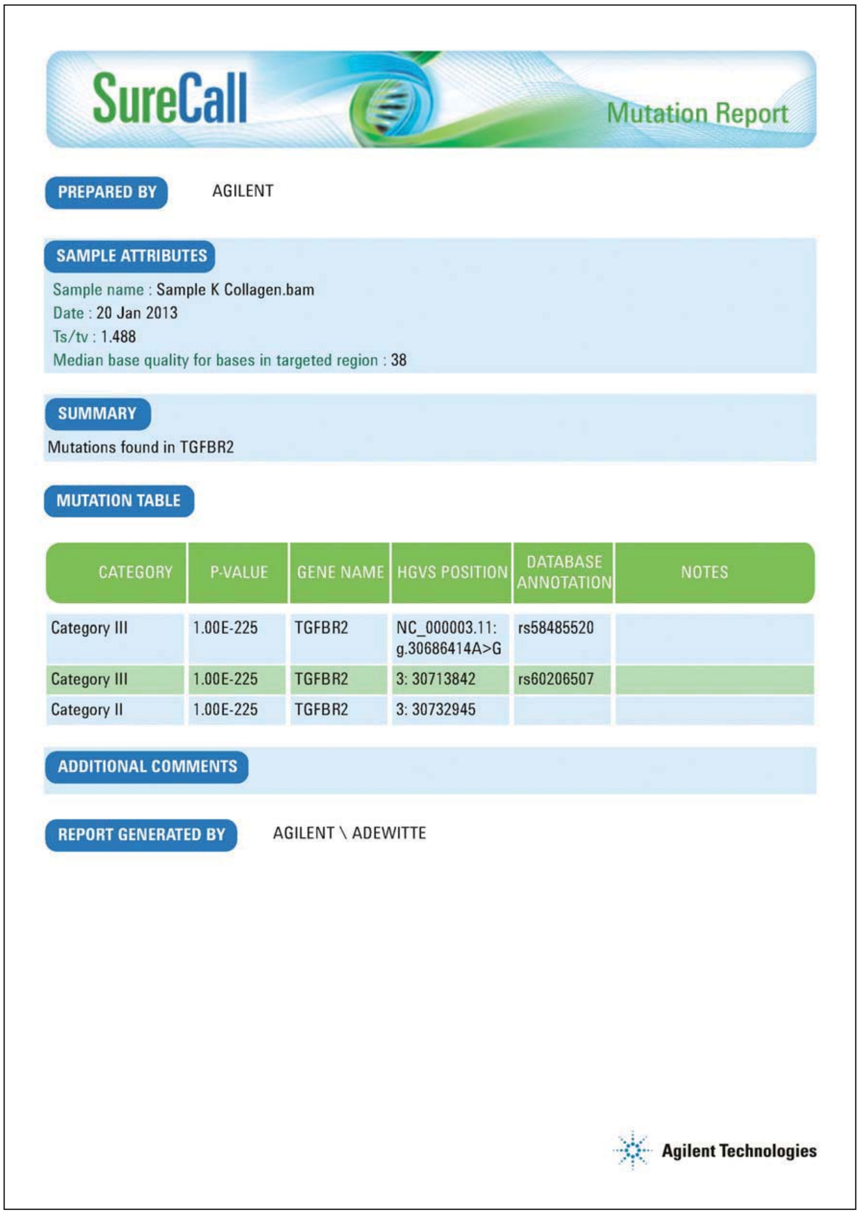
Suppress	Notes	Category	Gene	HGVS	ID	Pos	Ref Allele	Alt Allele	Primary Effect
□		Category III	TGFBR2	NC_000003.11:g.30686414A>G	rs58485520	30686414	A	G	INTRON
□		Category III	TGFBR2	3:30713842	rs60206507	30713842	C	T	SYNONYMOUS_CODING
□		Category II	TGFBR2	3:30732945		30732945	T	C	NON_SYNONYMOUS_CODING

```
Total count: 751
A : 1 (0%, 1+, 0- )
C : 361 (48%, 152+, 209- )
G : 0
T : 389 (52%, 175+, 214- )
N : 0
```

Figure 5. Example of review of a mutation found in a collagen tissue disease sample. A category II mutation was found in the TGFBR2 gene at position 30,732,945.

Mutation Report

Following manual inspection of the sample in the Triage View a Mutation Report can be generated by the researcher. The Mutation Report contains information about the mutations found in the sample, including the chromosomal location of the mutation with HGVS nomenclature, as well as the categorization of each mutation. In addition, researchers can include notes associated with the mutation or the sample. Figure 6 shows the Mutation Report for one of the collagen tissue disease samples.



SureCall

Mutation Report

PREPARED BY AGILENT

SAMPLE ATTRIBUTES

Sample name : Sample K Collagen.bam
Date : 20 Jan 2013
Ts/tv : 1.488
Median base quality for bases in targeted region : 38

SUMMARY

Mutations found in TGFBR2

MUTATION TABLE

CATEGORY	P-VALUE	GENE NAME	HGVS POSITION	DATABASE ANNOTATION	NOTES
Category III	1.00E-225	TGFBR2	NC_000003.11: g.30686414A>G	rs58485520	
Category III	1.00E-225	TGFBR2	3: 30713842	rs60206507	
Category II	1.00E-225	TGFBR2	3: 30732945		

ADDITIONAL COMMENTS

REPORT GENERATED BY AGILENT \ ADEWITTE

Agilent Technologies

Figure 6. Mutation Report for a collagen tissue disease sample.

System Setup and User Roles

SureCall is a client/server system. Data analysis is performed on the client while the primary function of the server is to host the data. SureCall client and server can be installed on one machine. Alternatively, the system can be set up with a central server that multiple clients can connect to simultaneously.

SureCall provides secure data access through three user roles: Technician, Scientist, and Administrator. Each of the three user roles has unique permissions to access a selected set of features and data in the software (Table 3).

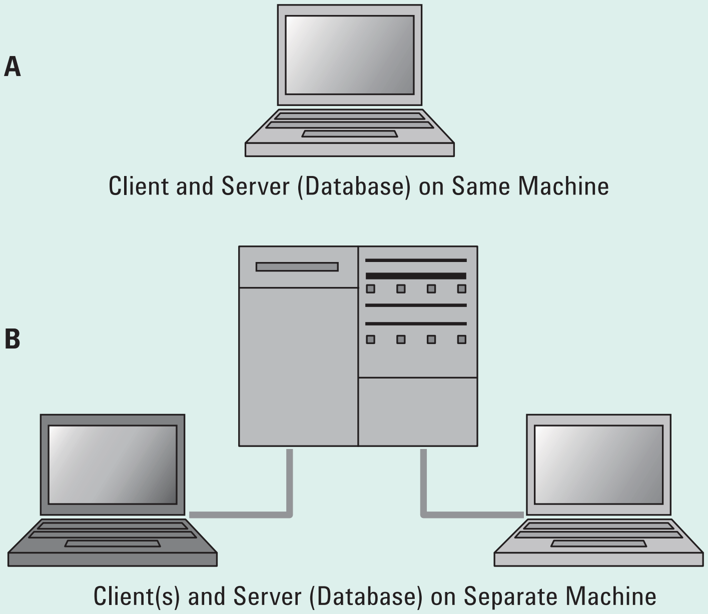

A
Client and Server (Database) on Same Machine
B
Client(s) and Server (Database) on Separate Machine

Figure 7. System setup with client and server installed on one machine (A) or on one server with multiple clients (B).

Role	Capabilities
Technician	• Run analysis workflows • Add sample information • Monitor workflow jobs • Triage samples ◦ Check in/out samples ◦ Add notes ◦ Suppress mutations ◦ Change category assigned to mutations ◦ Compare mutations across samples
Scientist	Technician tasks, plus: • Configure analysis methods • Configure categorizations • Sign-off results and generate reports • Unlock results
Administrator	Complete system access, including all Technician and Scientist tasks, plus: • Add users and roles • Change database connection settings for client systems

Table 3. Specific capabilities allowed for the three different user roles within SureCall.

Conclusion

In this application note HaloPlex, a simple target enrichment protocol that can be used to capture genomic regions of interest for clinical research, is described. Custom panels can be designed quickly and easily with SureDesign. Results described here demonstrate the high performance of the HaloPlex assay to capture regions of interest. Data produced by the HaloPlex assay can be easily analyzed in SureCall software without the need for an extensive bioinformatics infrastructure.

SureCall, which is available at no additional cost to Agilent target enrichment customers, addresses all the needs of HaloPlex users for analysis, visualization, contextualization and summarization of results. This easy-to-use tool performs all of the analysis steps from raw sequence alignment, mutation categorization, and visualization with options for linking to external databases.

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 SureDesign, Haloplex and SureCall:
www.agilent.com/genomics/ngs

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

PR7000-0531

This information is subject to change without notice.

Published in USA, June 24, 2016
Publication Number 5991-1919EN