Mutation Detection and CNV Analysis for Illumina Sequencing data from HaloPlex Target Enrichment Panels using NextGENe Software for Clinical Research

Application Note

Authors

John McGuigan, Megan Manion, Kevin LeVan, CS Jonathan Liu

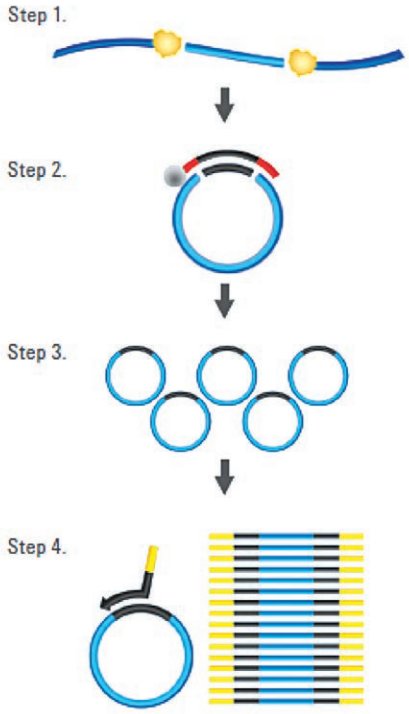


Figure 1. *HaloPlex Workflow*

Introduction

The HaloPlex Target Enrichment System allows for rapid targeted sequencing and library preparation in a single-tube reaction with no need for library preparation and support for up to 96 barcoded samples in one reaction. It works in 4 steps (figure 1).

1. Digest sample DNA
A DNA sample is fragmented using restriction enzymes, and denatured.

2. Hybridize probes
The probe library is added and hybridized to the targeted fragments. Each probe is an oligonucleotide designed to hybridize to both ends of a targeted DNA restriction fragment, thereby guiding the targeted fragments to form circular DNA molecules. The probe also contains a method-specific sequencing motif that is incorporated during circularization. In addition, a sample barcode sequence is incorporated in this step.

3. Purify and ligate targets
The HaloPlex probes are biotinylated and the targeted fragments can therefore be retrieved with magnetic streptavidin beads. The circular molecules are then closed by ligation, a very precise reaction that ensures that only perfectly hybridized fragments are circularized.

4. Amplify targeted fragments with PCR
Only circular DNA targets are amplified, providing an enriched and barcoded amplification product that is ready for sequencing.

This application note covers analysis of this data when sequencing is performed on Illumina HiSeq or MiSeq sequencing systems.

Agilent Technologies

NextGENe allows for simple point-and-click project setup in an easy-to-use Windows environment running on relatively low-cost hardware. It includes several features that make the analysis of this data fast and simple. After alignment the results can be viewed in the NextGENe Viewer with access to dbSNP, dbNSFP, and COSMIC databases.

Figure 2 shows a screenshot from HaloPlex Cardiac Disease Panel targeted sequencing. The called mutation was validated using Sanger sequencing, in this sample.

Procedure

A given dataset can be processed in minutes on a desktop computer running a 64-bit Windows operating system. Several paired-end datasets were processed in this analysis, all sequenced on an Illumina MiSeq Personal Sequencer:

- 2x Cardiomyopathy
- 3x Noonan Spectrum Disorder
- 3x Collagen Tissue Disease

Format Conversion (figure 3)

- The FASTQ files are first converted to FASTA format
- Basecall quality scores are used to filter and trim the raw data using default settings for Illumina data. The "Paired Reads Data" option is selected to keep the paired reads on the same line in each file.
- Adapter sequence is trimmed at the same time according to a text file. Each sequence is on a separate line with 4 columns separated by tabs (NextGENe v2.3.1):
  - Adapter name
  - 5′ Trim Sequence ("-" is used here to skip 5′ trimming)

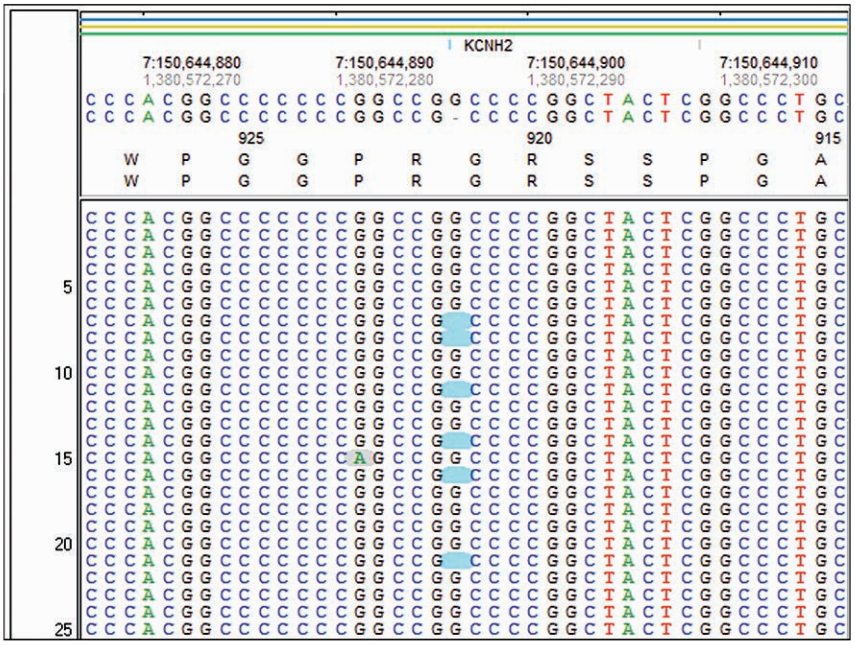


Figure 2. *c.2763delC in KCNH2 detected at approximately 23% frequency*

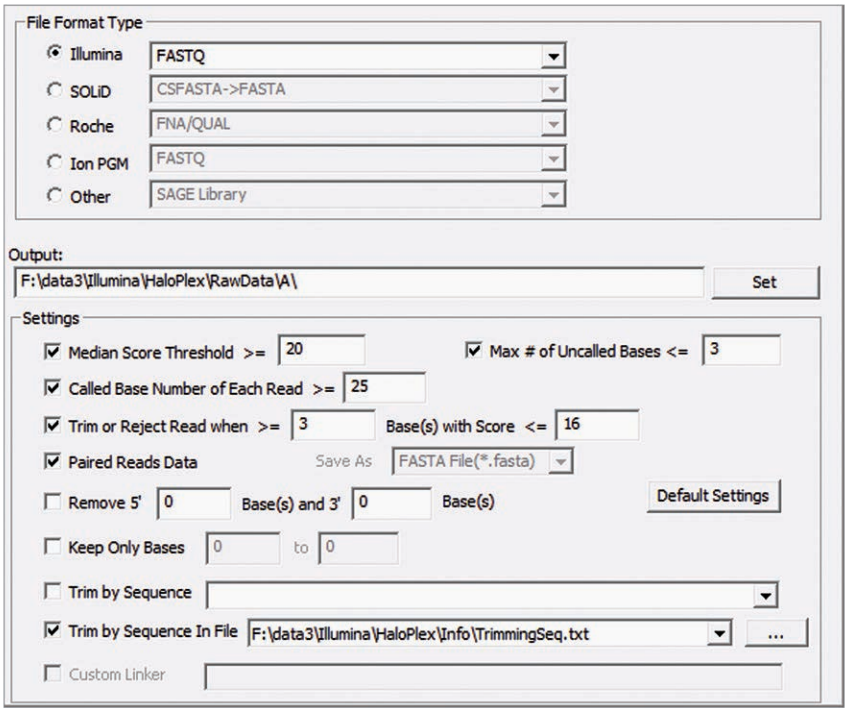


Figure 3. *Format Conversion Settings*

  - 3′ Trim Sequence – two Illumina adapters are listed on separate lines
    - AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
    - AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
  - Trim mode – "P" (Partial Match) is used here because we expect to see the beginning portion of the adapter sequence at the end of the read

Alignment to the human genome

- Alignment is performed against a pre-index human genome reference in order to avoid false positives caused by nonspecific amplification of untargeted regions
- Mismatches are called as variants in the mutation report if they pass the mutation filter. Settings used for the MiSeq projects in this analysis required more than 30x coverage, 5% frequency, and 15 total reads with the mutation (Figure 4).
- The "Load Paired Reads" option is selected in order to keep track of paired read associations.

Project review

- The mutation report, expression report, and coverage curve reports are filtered by loading a BED file specifying amplicon or targeted loci regions. The latter two reports provide information about coverage in the regions of interest. The mutation report has many additional filtering options.

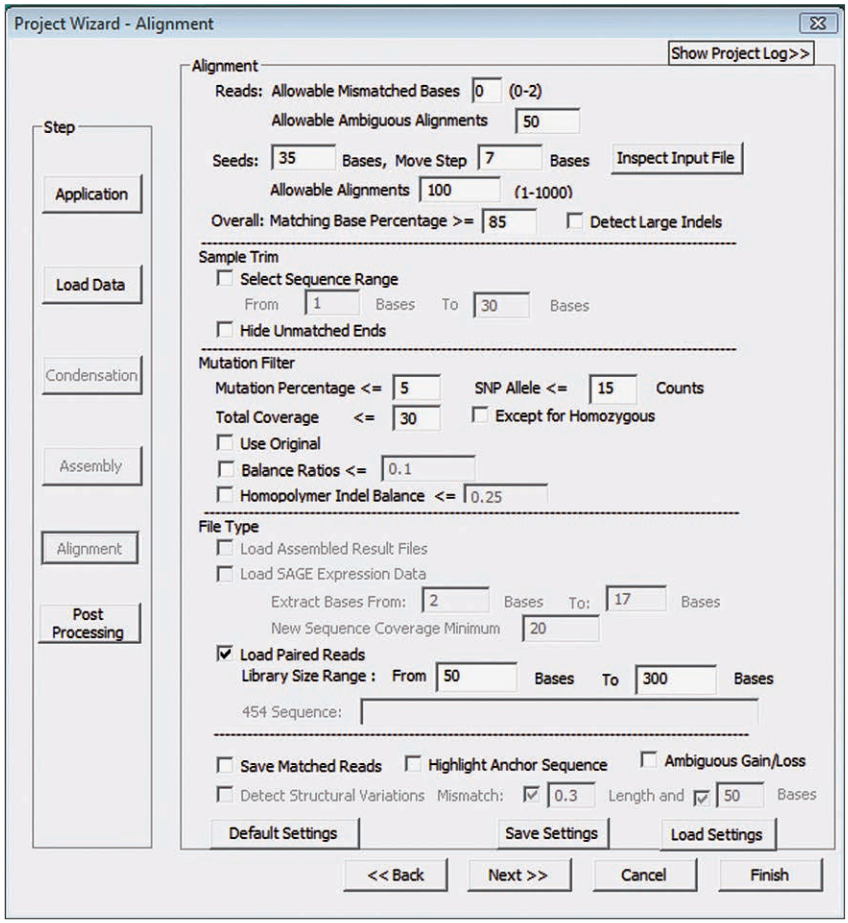


Figure 4. *Alignment and Mutation filter settings*

Results

Figure 5 shows the fraction of converted reads that were aligned for each sample, while figure 6 lists the fraction of converted reads aligned in the targeted regions. Results were very consistent for the samples within each panel. Cardiac Disease Panel results are in blue, Noonan Spectrum Disorder Panel results are in green, and Collagen Tissue Disease Panel results are in red.

Figure 7 shows the average coverage in the targeted regions for each sample and figure 8 shows the fraction of targeted bases that had at least 30x coverage. The information for figures 5, 6, 7, and 8 can easily be obtained from NextGENe's Coverage Curve report. The report (when filtered with a BED file describing targeted regions) will list any regions below a set threshold of coverage and also provide a summary view (figure 9).

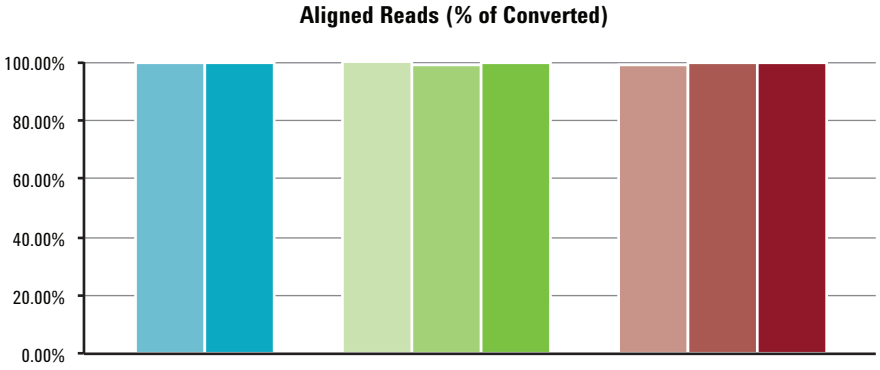


Figure 5. *Alignment results for the entire reference genome*

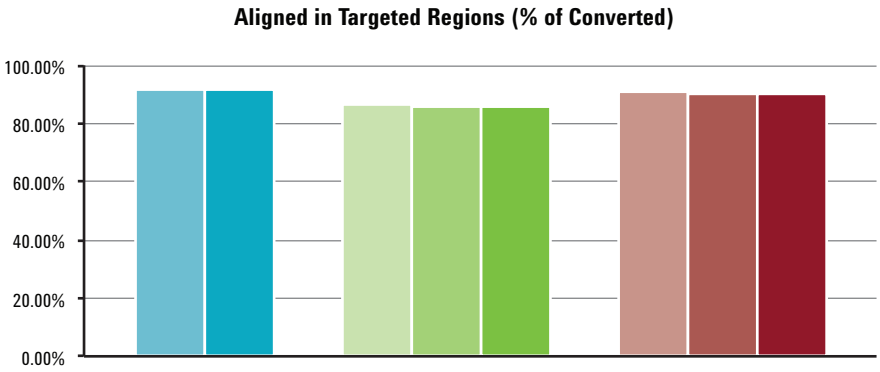


Figure 6. *Alignment results for the targeted regions*

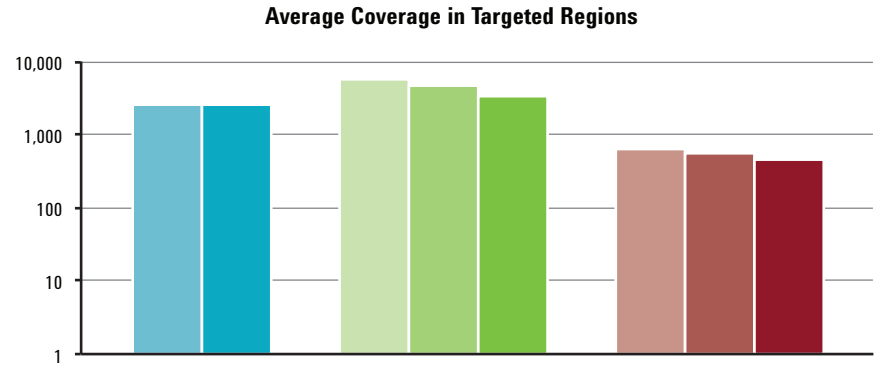


Figure 7. *Average coverage in the targeted regions*

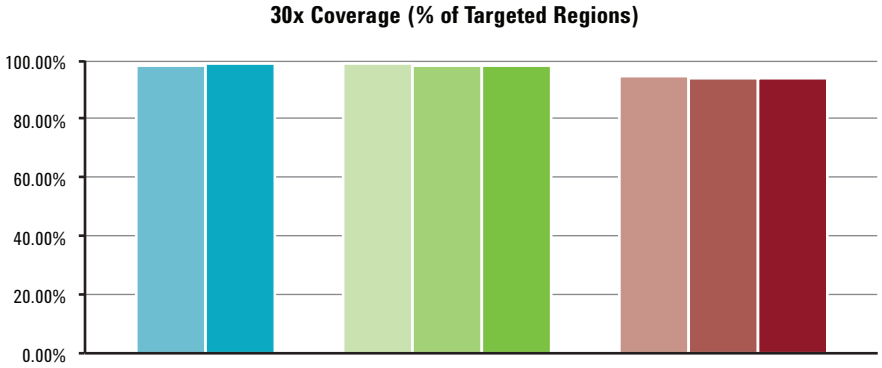


Figure 8. *Fraction of the targeted regions with 30 or more coverage*

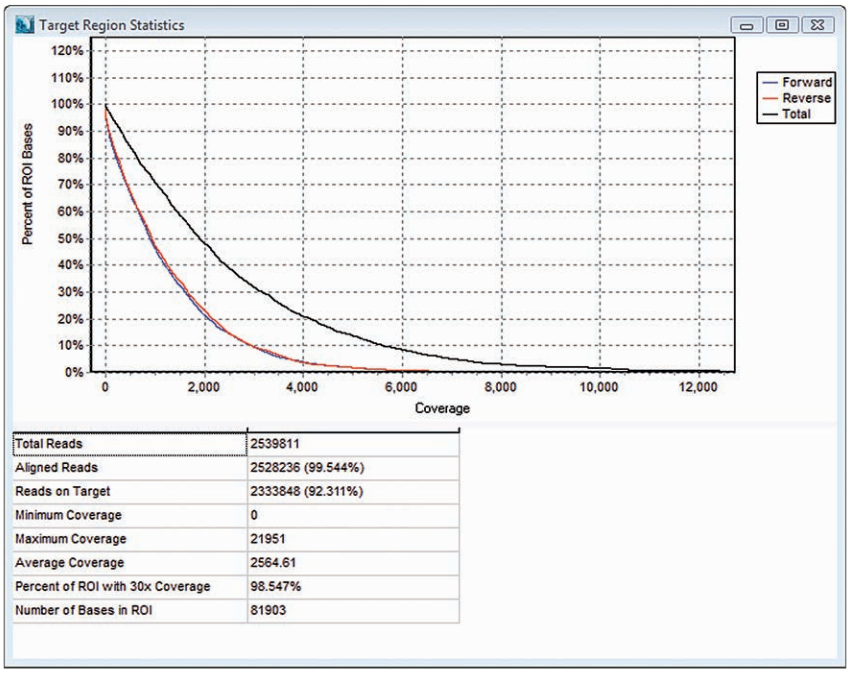


Figure 9 – Coverage Curve Summary report for a Cardiac Disease Panel sample.

All eight samples were successful in finding previously validated variants. One sample had a CNV variant that was detected using a new analysis that will be available in NextGENe v2.3.1. Mutation calling results are summarized in table 1. Figure 10 shows the deletion found in the first Noonan Spectrum Disorder sample.

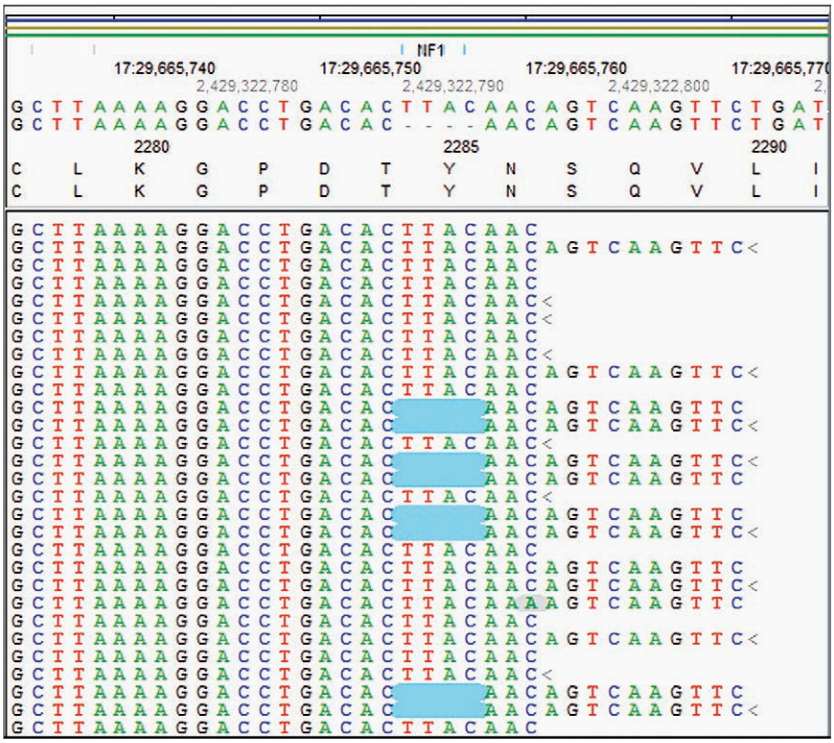
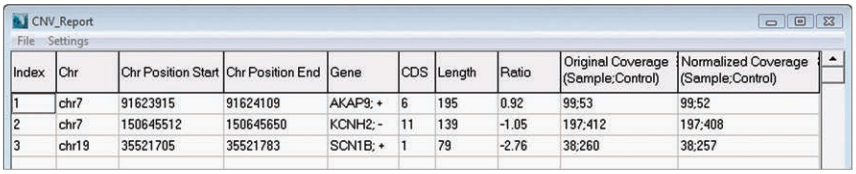


Figure 10. *A 4 bp deletion in the NF1 gene found in a Noonan Spectrum Disorder sample*

Table 1. *Mutation Results*

Panel	Gene	Mutation	Frequency	Chr	Position	Note
Cardiac Disease	KCNH2	Del Exon 11	N/A	7	~150,645,530–150,645,631	CNV
	KCNH2	delC	22.58	7	150,644,896	
Noonan Spectrum Disorder	NF1	delTTAC	8.87	17	29,665,754	
	SOS1	G>A	45.83	2	39,250,272	
	RAF1	C>T	45.43	3	12,645,699	rs80338796
Collagen Tissue Disease	COL5A2	G>A	44.78	2	189,906,320	
	TGFBR2	T>C	48.05	3	30,732,945	
	FBN1	delTG	40.68	15	48,752,496	

The new CNV analysis module is based on the median coverage in each targeted region after global normalization. A log-2 ratio is calculated in a case-control comparison. A negative ratio around -1 indicates a deletion in the sample relative to the control, while a positive ratio indicates duplication. In this analysis the two cardiac disease panel samples were compared, with the first treated as the sample and the second treated as the control. Three possible CNV changes with a ratio value greater than 0.6 were found, including the deletion of exon 11 in the KCNH2 gene. Results are shown in figure 11.


CNV_Report

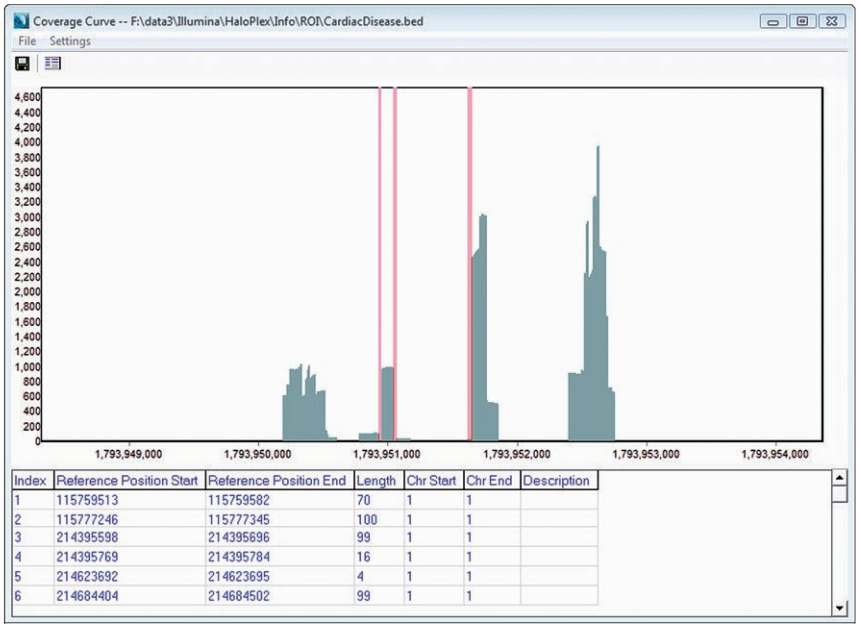
File Settings

Index	Chr	Chr Position Start	Chr Position End	Gene	CDS	Length	Ratio	Original Coverage (Sample;Control)	Normalized Coverage (Sample;Control)
1	chr7	91623915	91624109	AKAP9; +	6	195	0.92	99;53	99;52
2	chr7	150645512	150645650	KCNH2; -	11	139	-1.05	197;412	197;408
3	chr19	35521705	35521783	SCN1B; +	1	79	-2.76	38;260	38;257

Figure 11. *CNV analysis results*

Discussion

When choosing alignment settings, it is important to consider the expected results. Increasing the minimum depth of coverage will reduce the number of false positives (even at lower mutation frequencies), but it may also decrease sensitivity. Setting a minimum number of mutant allele reads will allow for detection of low frequency variants in high coverage regions without allowing low frequency false positives to be called in low coverage regions. The coverage curve summary report is very useful for measuring potential loss of sensitivity because it reports the regions that did not meet a minimum level of coverage (figure 12).



Index	Reference Position Start	Reference Position End	Length	Chr Start	Chr End	Description
1	115759513	115759582	70	1	1	
2	115777246	115777345	100	1	1	
3	214395598	214395696	99	1	1	
4	214395769	214395784	16	1	1	
5	214623692	214623695	4	1	1	
6	214684404	214684502	99	1	1	

Figure 12. *The Coverage Curve Report for a cardiomyopathy panel. Portions of targeted regions with less than 50x coverage are highlighted in pink and reported in a table.*

NextGENe's whole genome alignment algorithm has three steps- Match perfect reads, match reads with some number of mismatches, and finally a seeded alignment. Shorter amplicons will benefit from the second step because there may not be enough bases on either side of a mutation to align perfectly matching seeds. Longer seed sizes will improve alignment specificity, and fewer seeds will improve speed. After alignment several projects can be compared and filtered against one another. Compound heterozygous, shared/different, low coverage, and Mendelian inheritance filtering are all possible (figure 13).

After calling mutations, some functional prediction information is available from the dbNSFP v1.x database (1). This includes PolyPhen-2, SIFT, MutationTaster, LRT, and PhyloP in addition to 1000 genomes frequencies. The Sanger COSMIC (Catalog of Somatic Mutations in Cancer) database can also be imported (2).

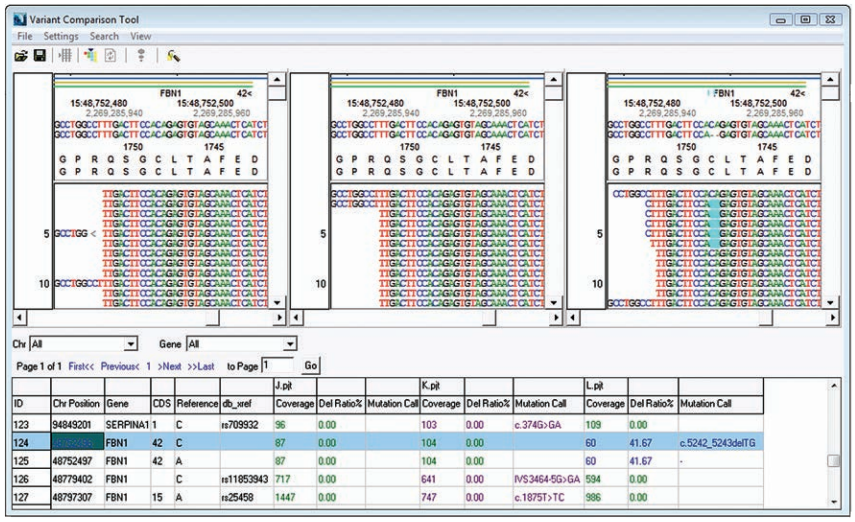


Figure 13. *Three Collagen Tissue Disease projects. A two bp deletion found in one of the projects is highlighted.*

Acknowledgements

We would like to thank Agilent Technologies and Berivan Baskin (Clinical Genetics, Uppsala University Hospital; The Centre for Applied Genetics; The Hospital for Sick Children) for supplying the data used in this analysis.

References

1. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* **32**, 894-899 (2011).

2. Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945-D950 (2010).

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

Learn more about NextGENe:
www.softgenetics.com

NextGENe® Next Generation Sequencing Software for Biologists

Published in USA, September 24, 2012
Publication Number 5991-1203EN

Agilent Technologies