Copy Number Variation Detection through the Utilization of Unique Algorithms in the CNV Tool in NextGENe Software®

McGuigan J.R., Manion M.R., Liu C.S., Fosnacht J., Wu, Y., Wan, N.
SoftGenetics LLC, 100 Oakwood Ave Suite 350, State College, PA 16803, USA

Abstract

One of the most important steps in accurate determination of copy number variants using a case-control comparison is normalizing the coverage level between the different projects. NextGENe's CNV tool uses unique methods for this global normalization and also for log2 ratio calculation. The global normalization method is based on the median allele coverage of selected heterozygous positions. After normalization, log2 ratios are calculated by selecting one position in each specified region (based on annotation or amplicon location). The positions are selected if they are adequately representative of the region, and (ideally) if they are heterozygous, so that allele frequency changes can be noted.

In this analysis, the tool is used to analyze data from multiple targeted sequencing technologies. A known variant (deletion in the KCNH2 gene) was detected using HaloPlex™ Target Enrichment System Panels and large chromosomal abnormalities (such as trisomy 21) were detected using Ion AmpliSeq™ Panels.

Several additional methods are in development for CNV detection in the next version of NextGENe (v2.3.3) including a method similar to the one described in the recently published ExomeDepth paper[1] which may help account for the greater amount of variation seen in exome capture-based methods.

[1] Plagnol, Vincent, et al. "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling."Bioinformatics (2012).

Methods

Two aligned NextGENe projects are loaded into the CNV tool (figure 1). Ideally the projects were processed with experimental conditions as closely matched as possible, since slight variations in sample prep or sequencing can cause the final coverage to vary from region to region. After global normalization, a median coverage value is generated for each region as specified in a BED file. Based on the coverage, a log2 ratio is calculated for each region, and filters allow regions with "normal" ratios to be hidden from the final report (figure 2).

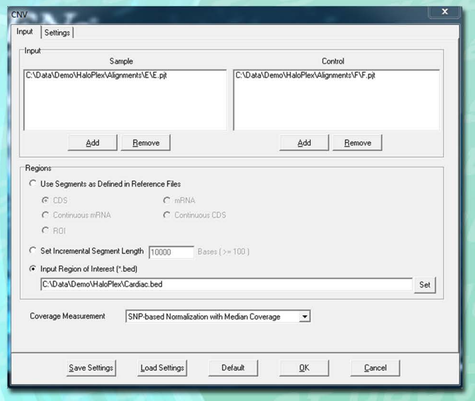
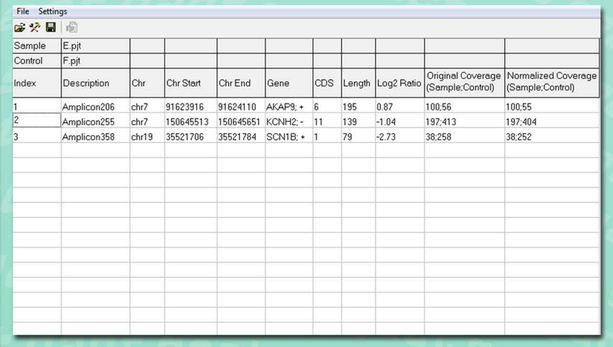


Figure 1: Loading projects in the CNV Tool



Sample	E.pjt									
Control	F.pjt									
Index	Description	Chr	Chr Start	Chr End	Gene	CDS	Length	Log2 Ratio	Original Coverage (Sample;Control)	Normalized Coverage (Sample;Control)
1	Amplicon206	chr7	91623916	91624110	AKAP9; +	6	195	0.87	100;56	100;55
2	Amplicon255	chr7	150645513	150645651	KCNH2; -	11	139	-1.04	197;413	197;404
3	Amplicon358	chr19	35521706	35521784	SCN1B; +	1	79	-2.73	38;258	38;252

Figure 2: Results from comparing two HaloPlex datasets. A known deletion was found in the KCNH2 gene.

Results

As seen in figure 2, a known CNV was found using HaloPlex targeted sequencing data. Additionally, Ion Torrent AmpliSeq Comprehensive Cancer Panel data was analyzed. The report can be saved as a tab-delimited text file and graphs can easily be generated outside of NextGENe using the log2 ratios. Looking at a smoothed average of the log2 ratios reported by NextGENe makes large known copy number variants visible in chr21 and chrX (figure 3) using the Ion Torrent data.

NextGENe®
2nd Generation Sequence Analysis Software

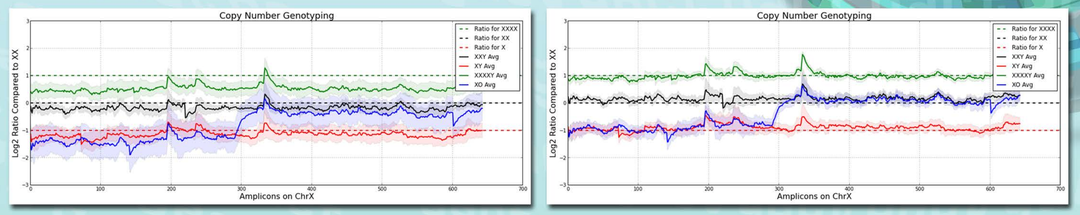


Figure 3: Graphed Log2 ratios calculated by the CNV tool using Ion Torrent AmpliSeq data.
The left side shows the ratios without normalization and the right side is with normalization. The smoothed average is plotted, along with a shaded range representing the minimum and maximum values for replicate samples.

Discussion

As shown in these two examples, the current CNV tool works well with targeted sequencing data. Capture-based methods tend to vary more from sample to sample. In order to detect CNVs in this type of sample, the next version of NextGENe will have an additional CNV detection algorithm available. This method generates complete CNV calls using several steps:
1.) Calculate RPKM values for each exon in a sample-control comparison.
2.) Fit an equation to the data to estimate a dispersion value depending on the coverage (figure 4).
3.) Calculate likelihoods under 3 different scenarios based on a beta distribution of the ratio.
4.) Merge CNV calls from consecutive exons and apply a prior likelihood using a Hidden Markov Model (HMM) (figure 5).

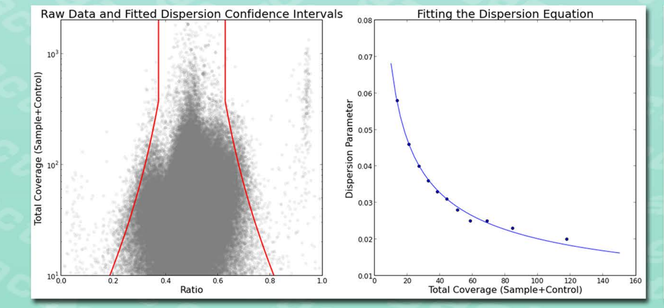


Figure 4: Calling CNVs by Fitting a Dispersion Parameter
Left: Scatterplot of the ratio vs total coverage of each exon in a cancer tumor-normal comparison. The red lines indicate the 97% confidence interval after fitting a dispersion parameter.
Right: Fitting points were generated for every 15,000 exon regions. An exponential equation was fit to model the dispersion across different levels of coverage.

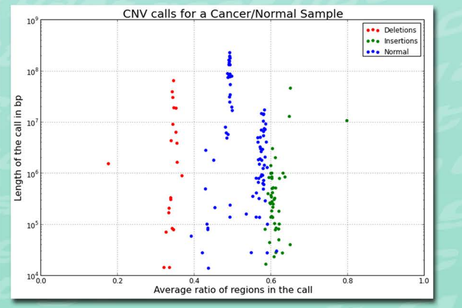


Figure 5: CNV Results
A Hidden Markov Model (HMM) was used to merge multiple-exon calls and apply a prior probability. This tumor-normal comparison resulted in several large CNV calls, including a large insertion on chromosome 7 that had a ratio near 0.8 (4x the normal copy number).

Conclusion

Next-Generation Sequencing (NGS) allows for detection of copy-number variants based on sequencing of targeted panels or even whole-exome sequencing. The amount of variation will vary from sample to sample, but careful analysis and normalization can successfully resolve real CNV changes from high background noise. This type of analysis allows for mutation detection and CNV detection with the same data.

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 HaloPlex data used in this analysis, and Life Technologies for supplying the Ion Torrent data.

SOFTGENETICS®
Software PowerTools for Genetic Analysis
www.softgenetics.com