Evaluation of the NextGENe CNV Caller for Use in the Clinical Laboratory

ARUP LABORATORIES
An enterprise of the University of Utah and its Department of Pathology

The University of Utah
Department of Pathology

T Lewis[1], J Durtschi[2], B Brulotte[2], S Dames[1], G Pont-Kingdon[1], R Mao[2,3], and P Bayrak-Toydemir[2,3]
[1]ARUP Institute for Clinical and Experimental Pathology, [2]ARUP Laboratories,
[3]ARUP Department of Pathology, University of Utah Health Sciences Center, Salt Lake City, Utah

INTRODUCTION

Using Next Generation Sequencing (NGS) data to detect exonic level deletions/duplications is very cost effective and timesaving for diagnostic laboratories, which will be replacing the traditional MLPA or array CGH. Analysis of allele frequency and read depth are the two most common methods. We used read depth data to detect copy number variation and examined the robustness of CNV Caller in the NextGENe software, version 2.4.11.

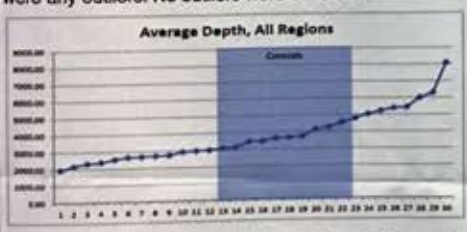
METHODS

NGS read depth data was generated using a 0.5 Mb Agilent SureSelect custom target enrichment panel that captured 1193 exons from 83 genes followed by sequencing on the Illumina NextSeq platform. Samples were sequenced on Illumina NextSeq instruments. Fastq data was processed by BWA aln alignment, and Picard MarkDuplicates for duplicate removal. The average read depth of coverage was 3342.

The NextGENe software utilizes the ratio of read depth of the sample to a control set then uses a Hidden Markov Model to account for dispersion (noise) and calculate copy number variations at each region. Ten samples with normal aCGH results were established as the control set. 138 samples were compared against the gender matched controls as were seven samples known by aCGH to harbor a copy number variation.

Selection of Controls

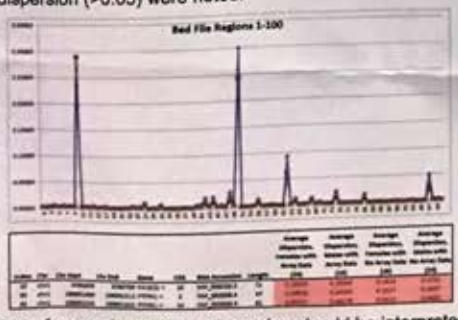
Controls were selected from 30 males with normal aCGH results. Ten control samples were selected from the middle range of the average depth of coverage. The samples were examined pairwise to determine if there were any outliers. No outliers were observed.



To control for gender, the 10 samples were computationally altered to represent males (single chrX) or females (2 chrX).

Dispersion

When dispersion at a region is high, the likelihood of any one call is low and the confidence interval is wide. 93 unique samples were divided into four groups and the average dispersion for each bed file region for each set was calculated. Thirty-one regions of consistent high dispersion (>0.05) were noted.



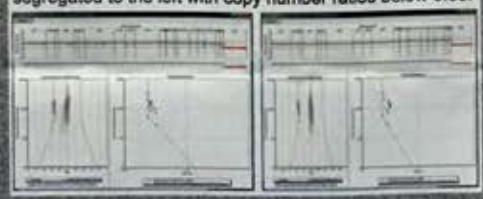
Data from regions of high dispersion should be interpreted with caution as true positives may not be called. Special attention should be given to the ratio. In instances with ratio values consistent with a deletion or duplication, but no HMM call due to high dispersion, the finding should be verified with a second method.

Gender Mismatches

In the SureSelect capture examined, there are 6 genes targeted on the X chromosome: DMD, IL2RG, BTK, XIAP, SH2DIA, and CD40LG. These six genes encompass 122 regions in the .bed file.

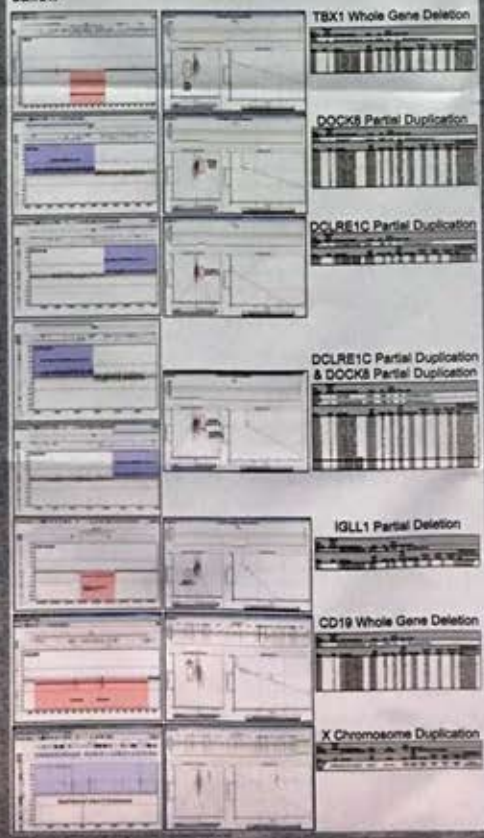
Two female were analyzed against the male control set and all 122 regions on the X chromosome appear as duplicated for each sample. In the raw data dispersion graph, the X chromosome regions are clearly seen as segregated to the right with copy number ratios exceeding 0.6.

Conversely, two male samples were analyzed using the feminized control set in which case the regions on the X chromosome appear as a deletion. In the raw data dispersion graph, the X chromosome regions are segregated to the left with copy number ratios below 0.33.



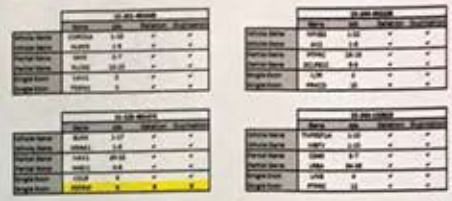
Biological Positives

Seven samples were identified by array comparative genomic hybridization (aCGH) as having copy number variation in regions targeted by the NGS Agilent Sure-Select capture. In each sample, the known copy number variation was detected by the NextGENe CNV caller.



Artificial Positives

Read depth on four samples was bioinformatically altered to create artificial copy number variations. Reads in targeted regions were either reduced or increased by 50%. Copy number variations in three sizes were created: whole gene, three contiguous exons, or single exons. The creation of the artificial positive samples occurred in bioinformatics and the altered bam files were provided to R&D as blinded samples.

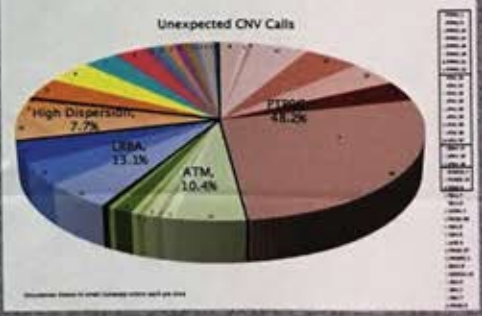


One variation, PSTPIP exon 6, was not detected as either a deletion or duplication. This region is one of the 31 regions marked as having high dispersion.

Unexpected Variation

A total of 138 samples with passing QC were examined; 93 of the samples had accompanying arrays while 45 did not. All results, with the exception of known positives, was compiled. Forty-seven of the samples had no calls. In the remaining 91 samples, there were 222 unexpected copy number variations representing 38 bed file regions (3.2%).

Variation in four genes account for 79.7% (177/222) of the calls. The four genes are: PTPRC, LRBA, ATM, and TXBN. Exons previously identified as having high dispersion accounted for 17 of the calls. 28 calls remained, representing 20 samples and 9 genes.



CONCLUSION

A set of 10 control samples were selected to analyze a data set of 138 samples. Controls were bioinformatically altered for gender.

Thirty-one of the 1193 bed file regions analyzed were found to have consistently high dispersion. Results from these regions should be interpreted with caution.

Seven known positive controls were analyzed using the NextGENe CNV caller. In each instance, the copy number variation was accurately identified by the software.

The 138 samples analyzed for unexpected calls. Forty-seven of the samples had no calls; in the remaining 91 samples, there were 222 calls of copy number variation. 80% of the calls were in four specific genes: PTPRC, LRBA, ATM, and TXBN. Previously identified regions of high dispersion accounted for another 17 calls, leaving 28 calls to be evaluated.

Preliminary results indicate that the CNV Caller in the NextGENe software is robust enough to consider for use in the clinical lab. Positive results need to be confirmed by a second method.