

New CNV Algorithm in NextGENe v2.3.4

June 2013

John McGuigan, Jacie Wu, Ni Shouyong, CS Jonathan Liu

Introduction

NextGENe version 2.3.4 includes a sophisticated new algorithm for copy-number variation (CNV) detection from a wide variety of projects, including whole-exome and targeted sequencing panels. Copy number variations are detected by comparing the coverage (RPKM) of specified regions in a “sample” project and a “control” project. The coverage ratio (sample divided by sample plus control) is used as the basis for CNV detection. A beta-binomial model is fit to the coverage ratio (similar to the recently published ExomeDepth software [1]) in order to model the amount of dispersion (noise). Likelihood values are calculated based on the dispersion measurements and coverage ratios. These probabilities are then entered into a Hidden Markov Model (HMM) to make CNV classifications for each region.

The resulting report gives a simple classification for each region- either “Insertion” (increased copy number), “Normal” (little evidence of a CNV), “Deletion”, or “Uncalled” (due to low coverage). Additionally, each region receives two phred-scaled probability scores- one for insertions and one for deletions. The results are available in a table along with a graphical view, as seen in figure 1.

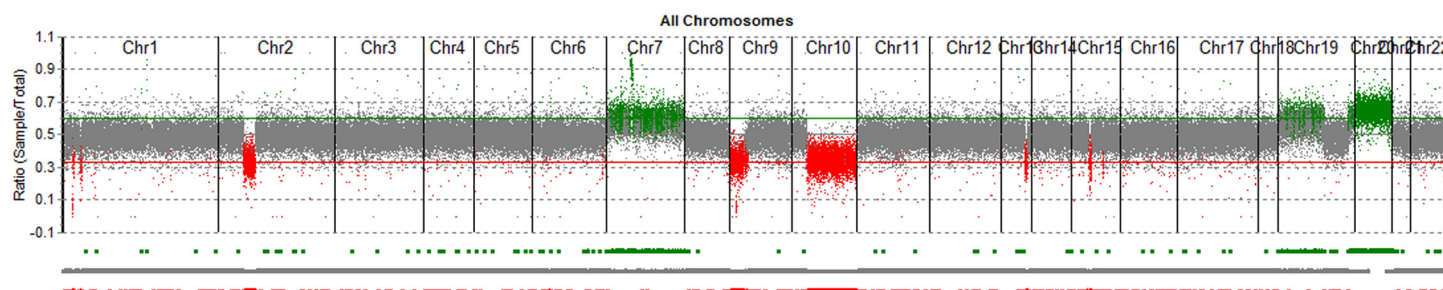



Figure 1: New Graphical CNV View: results from a tumor-normal comparison.

Procedure

1. One “sample” project and one “control” project are loaded into the CNV tool (figure 2), available in the NextGENe Viewer “Tools” menu. In the future, multiple controls and replicate samples will be supported.
2. The regions are identified- either by annotation, incremental length, or a BED file. A BED file specifying amplicon locations is suggested for targeted sequencing projects, and exon locations are useful for whole-exome sequencing.
3. The new CNV method is selected for use from a drop-down menu: “Dispersion and HMM with RPKM”
4. Analysis parameters are adjusted.
 - a. Expected CNV frequency is the prior estimate for the fraction of regions that should be classified as being a CNV. The setting is used during fitting and as a parameter in the HMM.
 - b. For automatic fitting, the raw data is grouped to generate “fitting points” describing the dispersion at a given level of coverage. A line is fit to these points and used to calculate the dispersion value for each region. As a rule of thumb, there should be at least 4 to 5 fitting points and at least 100 raw data points per fitting point.
 - c. A custom equation can be specified instead of using automatic fitting. This is useful for small targeted panels- a single, low dispersion value can be used for each region.
 - d. Filters can be adjusted before or after processing. Regions can be removed from the report based on classifications or based on the scores for insertions or deletions.

5. Processing is performed. After the report is finished generating, a graphical view of the results can be accessed using the  button.

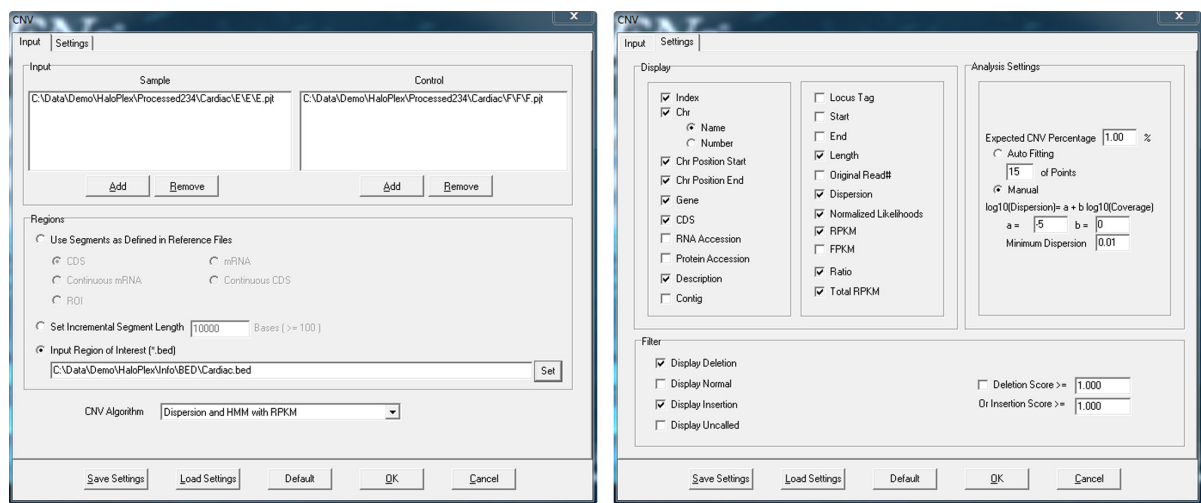


Figure 2: Running the CNV Tool

Results

Figure 3 shows the report from a HaloPlex Cardiac panel project. A manual fitting was used because of the small size of the panel and very low amount of noise. As seen in figure 2, a specific dispersion value (0.01) was used for each region by making this the minimum value and then adjusting the equation to be much lower (slope of 0 and intercept of -5). One of the three reported CNVs (Normal and Uncalled regions were hidden) is a known heterozygous deletion in the KCNH2 gene.


													
Sample E.pjt													
Control F.pjt													
Index	Description	Chr	Chr Start	Chr End	Gene	CDS	Length	Ratio	Total RPKM	Dispersion	Deletion Score	Insertion Score	HMM Calls
1	Amplicon66	chr1	237823266	237823394	RYP2; +	55	129	0.75	18.80	0.0100	0.00	22.02	Insertion
2	Amplicon255	chr7	150645513	150645651	KCNH2; -	11	139	0.31	96.64	0.0100	15.50	0.00	Deletion
3	Amplicon358	chr19	35521706	35521784	SCN1B; +	1	79	0.13	2178.39	0.0100	80.00	0.00	Deletion
													RPKM(Sample;Control)

Figure 3: Portion of the CNV Report from a HaloPlex Cardiac Panel Comparison

The graphical report initially shows every region in the genome, but chromosomes can be selected for review one-at-a-time. Figure 4 shows the full graphical results for the comparison in figure 1 except chromosome 10 is selected. The top panel shows the ratio for each region (expected ratios are 0.6 for heterozygous insertion, 0.5 for normal, and 0.333 for heterozygous deletion) and the location of CNV calls (lines below the graph). The lower-left graph shows the ratio-vs-coverage plot for every region. When data from chromosome 10 (purple) is compared to the data for all chromosomes (gray) in the lower-left chart, it is easy to see that a large portion of regions have a lower-than-normal ratio. In fact, about 3/4 of exons in the chromosome appear to be deleted. The lower-right graph shows dispersion fitting results. Automatic fitting was used, with 15 points and 6% expected CNV.

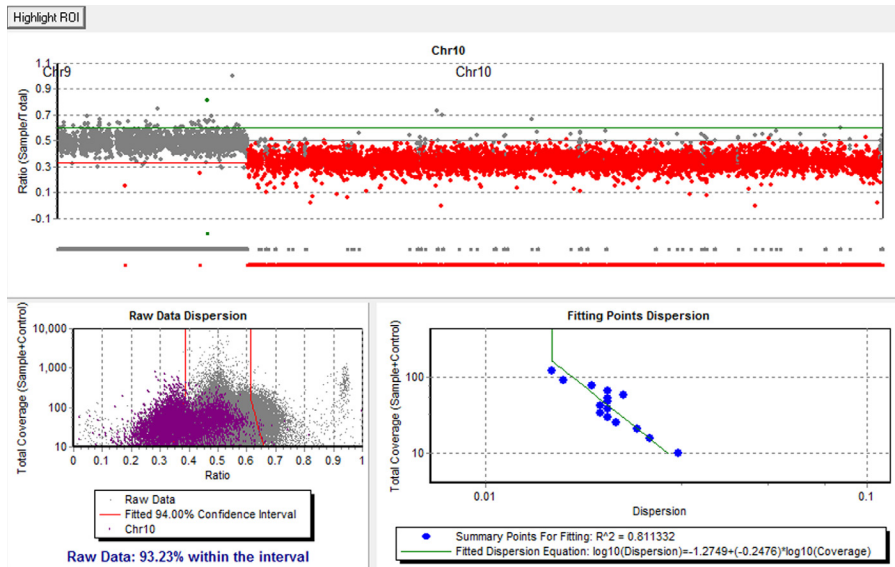


Figure 4: Results for a whole-exome tumor-normal comparison with chromosome 10 selected

Discussion

The goal of fitting the equation is to measure the amount of dispersion (noise) present in “normal” regions. The coverage ratio is expected to be equal to 0.5 for regions in the absence of a CNV. There is some randomness expected for this value, with higher-coverage regions showing a tighter distribution around the expected value than lower-coverage regions. The software first splits the data up into groups based on the total coverage, generating a summary “fitting point” for each group based on measured dispersion and the median coverage. A line is fit to these “fitting points” and the equation for this line is used to calculate dispersion for every individual region.

The dispersion value is used to calculate parameters for a beta distribution, which is used to generate a confidence interval. A higher dispersion value gives a broader CI because the ratios are expected to be more widely dispersed. If the expected CNV frequency is 10%, the software will calculate fitting points by incrementing the dispersion value until it produces an appropriate 90% (equal to 100%-10%) confidence interval (CI) of ratios. An appropriate confidence interval is one where the lower half of the CI is lower than the 5th percentile ratio of the real data (because Insertion = 5% and Deletion = 5% in this case), or the upper half of the confidence interval is greater than the 95th percentile. This one-sided fitting allows the software to be tolerant of CNVs that cause the raw data to have an asymmetrical distribution.

Dispersion values calculated for each region are used to generate normalized (probability of Normal + Insertion + Deletion = 1) beta-binomial distributions (figure 5). When dispersion in a given region is high, the likelihood for any one call is low except for extreme ratio values (close to 0.0 or 1.0).

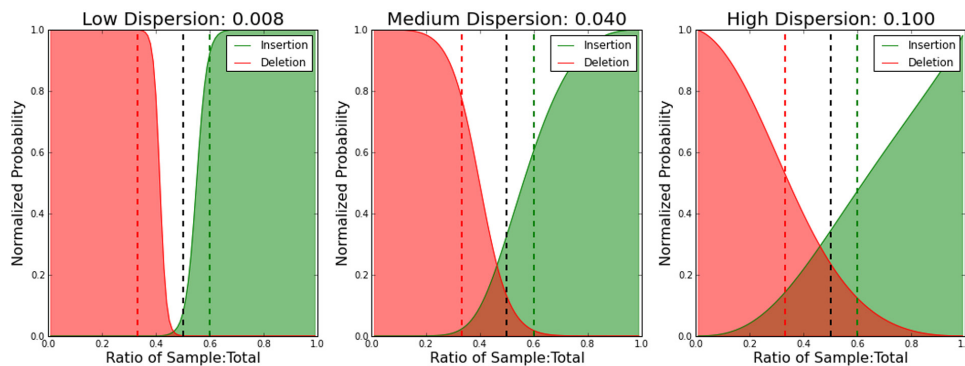


Figure 5: Normalized likelihoods at different dispersion values

The HMM used to make CNV calls makes some assumptions. The initial likelihood of each state is related to the expected CNV frequency, as is the probability of transitioning from a “normal” region to a region with a CNV. Once a region is called as a CNV, the next region is assumed to have a 50% chance of continuing that CNV or going back to normal. This transition probability enables the HMM to both ignore possibly erroneous ratios from single regions and also identify long CNVs where no individual region in the call has a very high probability.

Phred scores are also calculated using these likelihoods, by comparing the probability of obtaining the ratio if the region was an insertion or deletion (at least heterozygous) compared to the probability if it was a normal region. Phred scores are capped at 80, equivalent to a 99.999999% probability. Phred scores are much lower if the dispersion is high, because there is less certainty about the classifications (figure 6) and are higher if more regions are expected to contain CNVs. Generally deletion calls can be more confident than insertion calls because the expected heterozygous ratio (0.333) is farther away from the normal ratio (0.5) than the heterozygous insertion ratio is (0.6).

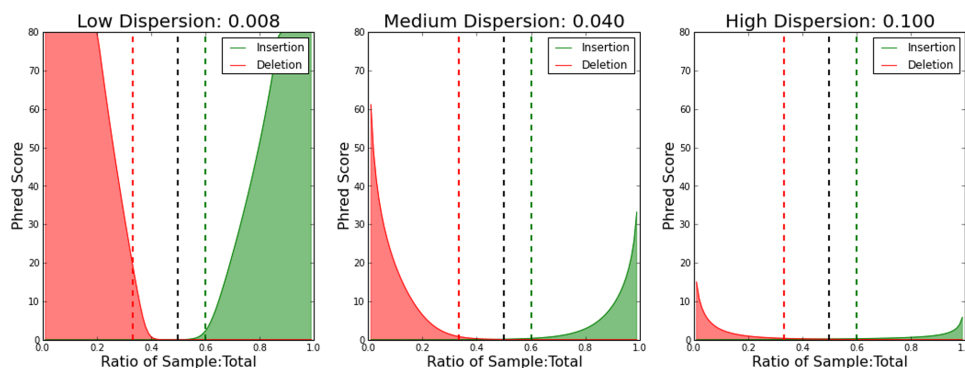


Figure 6: Distribution of Phred Scores across all possible ratios for three different levels of dispersion.

The best CNV results will come from two projects with very little dispersion- this means samples that are prepared as similarly as possible (generally sequenced as part of the same run). However, this automatic data fitting process can allow for any two projects to be compared- poorly matching projects will just have lower quality scores and fewer CNV calls.

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.

References

- [1] Plagnol, Vincent, et al. "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling." *Bioinformatics* 28.21 (2012): 2747-2754.