NextGENE CNV Detection- Dispersion and HMM

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

Introduction

NextGENe software includes a sophisticated algorithm for copy-number variation (CNV) detection from a wide variety of diploid genome projects, including whole-exome and targeted sequencing panels. Copy number variations are detected by comparing the coverage of specified regions in a "sample" project and one or more "control" projects. One control project can be used directly or multiple can be loaded so that the sample is compared to the best matched control, the average of the controls, or the median of the controls. The coverage ratio (sample divided by sample plus control) is used as the basis for CNV detection. This coverage can be measured as a standard RPKM value (Reads Per Kilobase per Million mapped reads) or as a more intuitive normalized read count value (normalization similar to DEseq [1]).

The software must account for differences in coverage from sample to sample which cause ratios that don't correspond exactly to expected values (1/3, 2/4, or 3/5). A beta-binomial model is fit to the coverage ratio (similar to ExomeDepth software [2]) 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 "Duplication" (increased copy number), "Normal" (little evidence of a CNV), "Deletion", or "Uncalled" (low coverage or short regions). When human genome data is analyzed these calls can be linked to the Database of Genomic Variants (http://dgv.tcag.ca/) so that nearby known structural variations can be visualized. Additionally, each called region receives three Phred-scaled probability scores- one for each potential copy number state. The results are available in a table along with a graphical view, as seen in figure 1. A "block CNV report" makes it possible to quickly exclude short CNV calls that may be due to random noise.

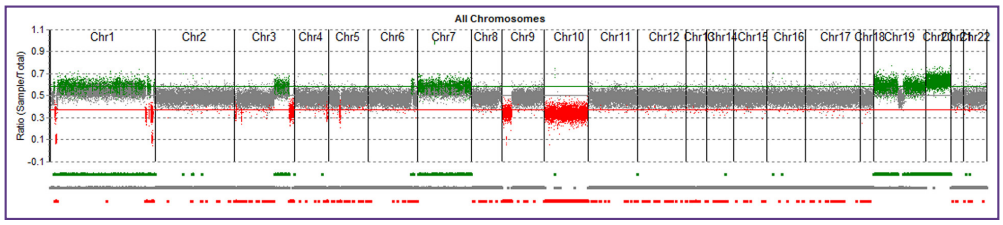

Figure 1: Graphical view of results from a tumor-normal comparison.

Procedure

1. The "CNV" tool is selected from the "Comparisons" menu in the viewer. This algorithm can be selected using Normalized Counts (suggested) or RPKM. Another CNV algorithm is also available (figure 2).
2. One "sample" project and one or more "control" projects are loaded into the tool (figure 2).

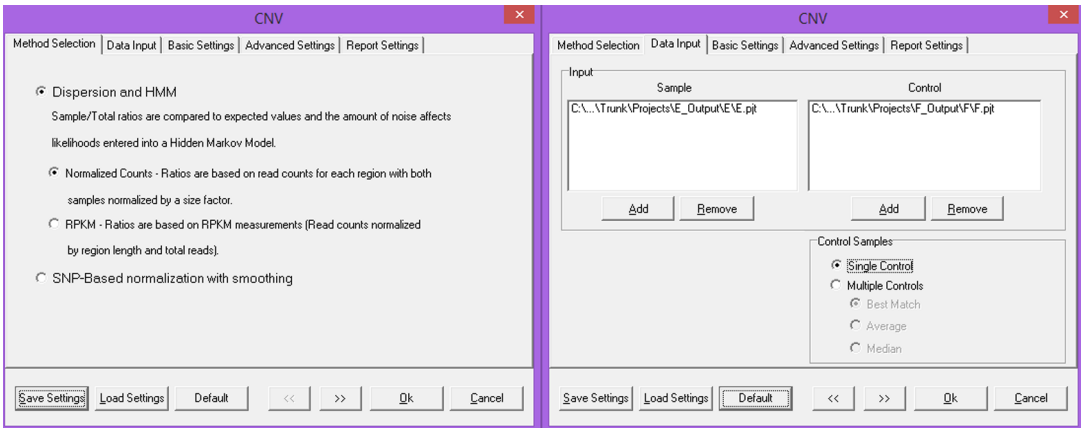

Figure 2: CNV Method Selection (left) and sample/control loading (right)

Procedure (cont.)

3. 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. There are options to easily exclude analysis of sex chromosomes and the mitochondrial chromosome when analyzing human samples if BED files are not used (Figure 3).
4. Analysis parameters are adjusted (Figure 3).
   a. 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. The number of fitting points is automatically set based on the number of regions but it may be set manually instead. 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. Manual fitting (either with an equation or a single value for all regions) can be used instead of automatic fitting. This is useful for small targeted panels.
   b. Minimum coverage and region length can be used for excluding some regions prior to fitting and further analysis. When RPKM is used, the values will be very small when large regions are being used and so the minimum will have to be lowered significantly.
   c. 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.
   d. Estimated Sample Purity can be useful for tumor samples that are not pure (contaminated with some normal cells). The expected CNV ratios are adjusted based on the purity, allowing for CNV calls in regions with ratios that are more similar to normal.
5. Processing is performed. After the report is finished generating, a graphical view of the results can be accessed using the  button.

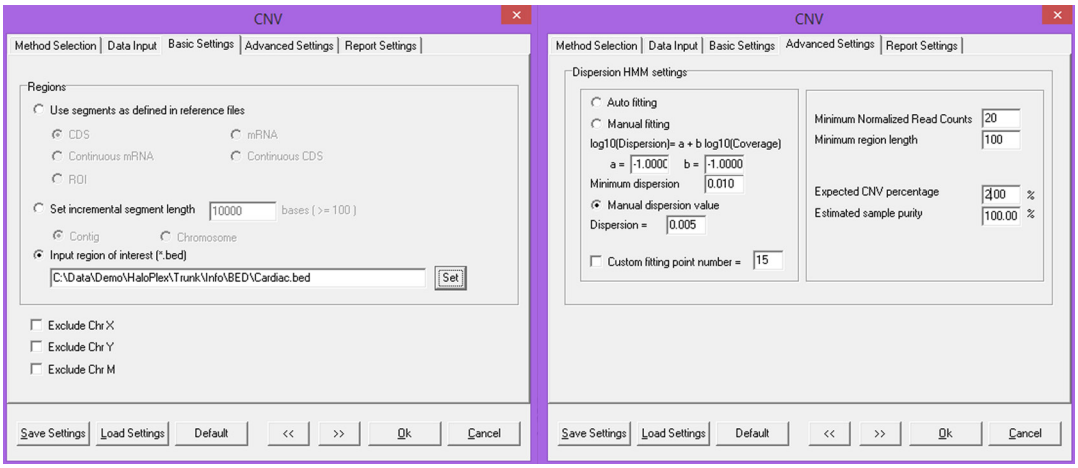

Figure 3: Analysis settings including region specification (left) and modeling settings (right)

Results

Figure 4 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 3, a specific dispersion value (0.005) was used for each region. The only reported CNV using these filters (Normal and Uncalled regions were hidden) is a known heterozygous deletion in the KCNH2 gene.

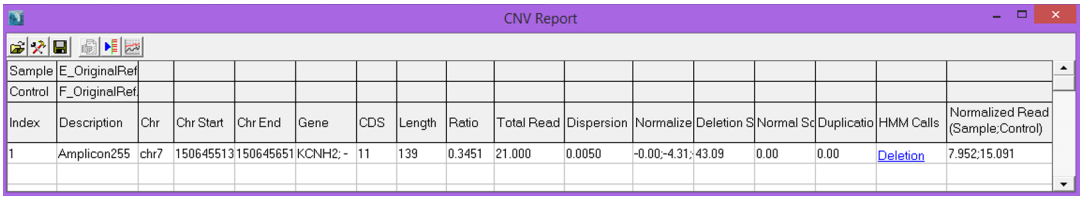

Figure 4: Portion of the CNV Report from a HaloPlex Cardiac Panel Comparison

The graphical report initially shows every region in the genome, but chromosomes can be selected for review one-at-a-time. Figure 5 shows the full graphical view for an Ion Torrent Comprehensive Cancer Panel project with chromosome 11 selected. The top panel shows the ratio for each region (expected ratios are 0.6 for heterozygous duplication, 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 11 (purple) is compared to the data for all chromosomes (gray) in the lower-left chart, it is possible to see the regions that were called as part of a known heterozygous deletion. The lower-right graph shows dispersion fitting results. Automatic fitting was used with 1% expected CNV, 100 minimum reads, and 80 bp minimum region length.

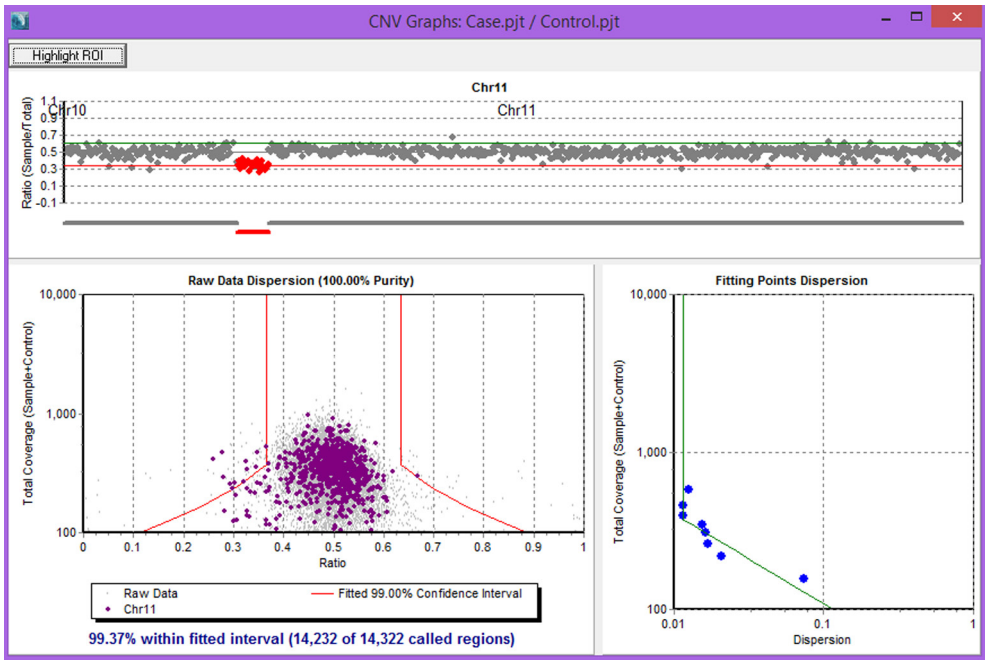

Figure 5: Results for an Ion Torrent Comprehensive Cancer Panel comparison, with chromosome 11 selected

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 usually 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 (CI). 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 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 Duplication = 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 + Duplication + Deletion = 1) beta-binomial distributions (figure 6). 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).

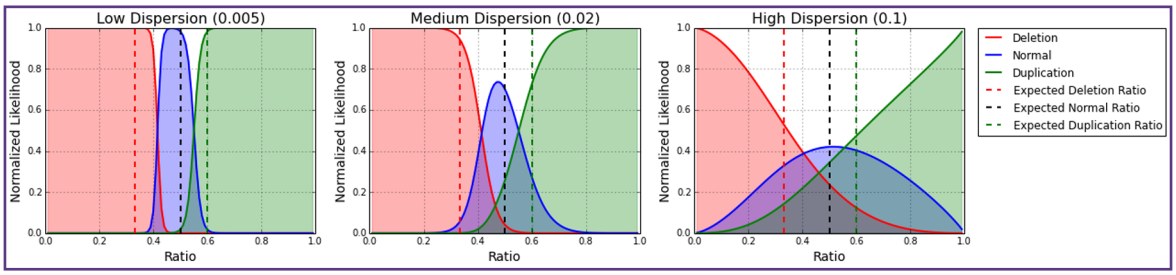

Figure 6: Normalized likelihoods at different dispersion values

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. They 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 7). Generally deletion calls can be more confident than duplication calls because the expected heterozygous ratio (0.333) is farther away from the normal ratio (0.5) than the heterozygous duplication ratio is (0.6).

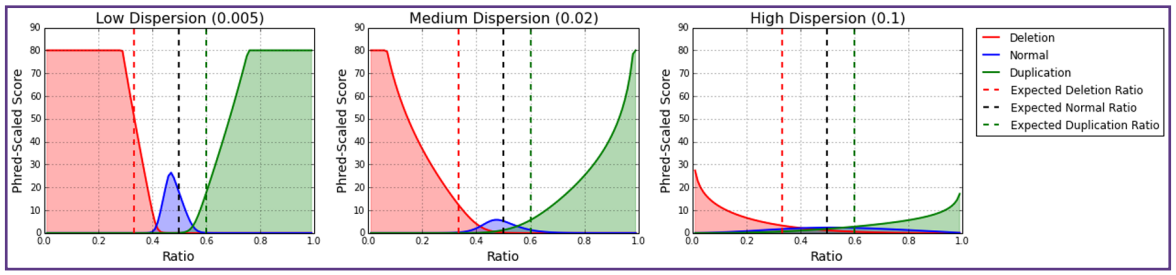

Figure 7: Distribution of Phred Scores across all possible ratios for three different levels of dispersion.

The best CNV results will come from 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 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 and Life Technologies for the publically accessible Ion Torrent data.

References

1. Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." Genome biol 11.10 (2010): R106.
2. 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.