

CNV Detection from Ion Ampliseq™ Panels in NextGENe® software

September 2013

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

Introduction

NextGENe software 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**. A “block CNV report” makes it possible to quickly exclude short CNV calls that may be due to random noise.

In this analysis, two datasets were downloaded from the Ion Community, aligned by NextGENe software, with Copy Number Variation analysis performed by the new CNV tool. The sample (GM22624) and the control (DNA CEPH individual 1347-02) had been prepared using the AmpliSeq comprehensive cancer panel.

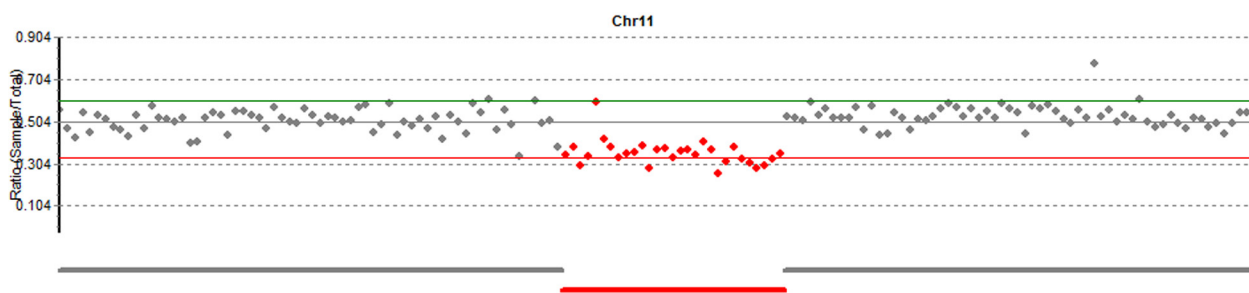




Figure 1: A portion of chromosome 11 showing a known deletion (red)

Procedure

1. Both projects were downloaded as unaligned BAM files. Alignment to the pre-indexed whole-human genome reference was performed in NextGENe software.
2. The projects are loaded into the CNV tool (figure 2), available in the NextGENe software Viewer “Tools” menu. In the future, multiple controls and replicate samples will be supported.
3. A BED file specifying amplicon locations is used to define the regions.
4. The new CNV method is selected for use from a drop-down menu: **“Dispersion and HMM with RPKM”**
5. 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. Here it is set to 1%.
 - 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. The default of 15 is used here.
 - c. “Normal” regions (little evidence of a CNV) and “Uncalled” regions (low coverage) were hidden in the initial report.
6. Processing is performed. After the report is finished generating, a graphical view of the results can be accessed using the  button. The block CNV report can be accessed using the  button.

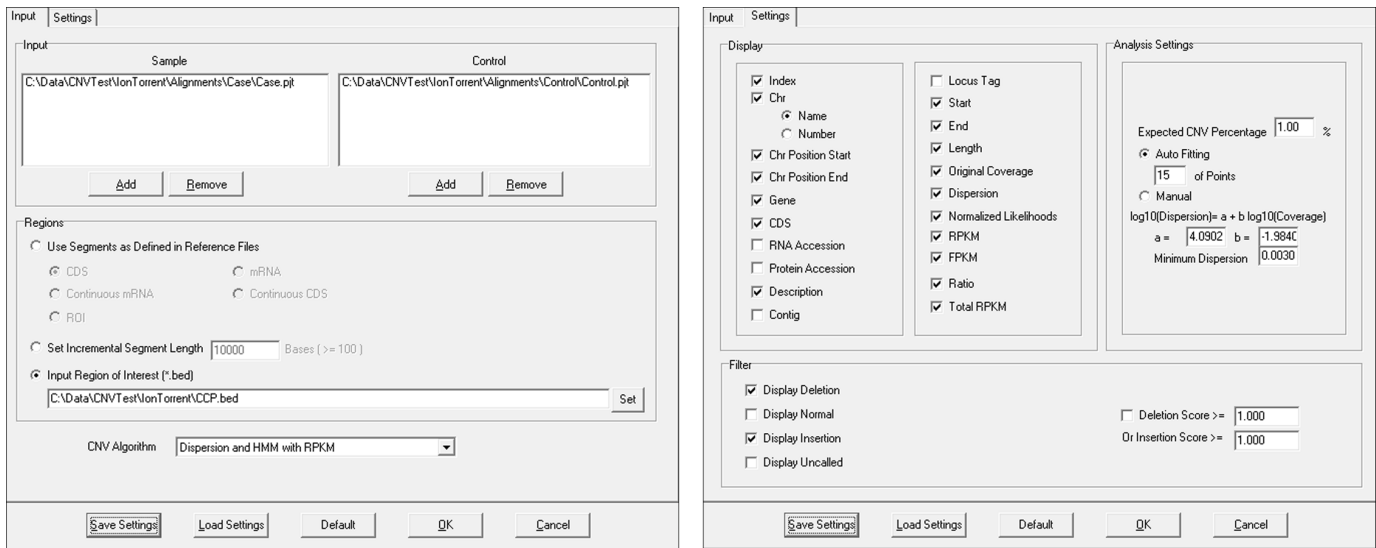


Figure 2: Running the CNV Tool

Results

Alignment of each dataset (approximately 2.3 and 3.4 million reads for the case and control respectively) took approximately 11 minutes on an 8-core laptop computer with 8 GB of RAM.

Figure 3 shows the block CNV report results with single-amplicon CNV calls ignored. The expected heterozygous deletion was found in EXT2, spanning 29 amplicons. Every other call consisted of 4 or fewer consecutive amplicons. The report can be adjusted to filter for calls of any number of consecutive amplicons. The coverage ratio for this known deletion (0.35) is extremely close to the expected ratio for a heterozygous deletion (1/3).

Sample	Case.pjt	Control.pjt	Chr	Chr Start	Chr End	Gene	Number of Regions	Length	Median Ratio	Median Disp	Median Del S	Median Ins Score	Max Del S	Max Ins S	HMM Call	RPKM(Sample,C
AMPL24107756 ; AMPL241080421			chr1	144912032	144912288	PDE4DIP	3	257	0.37	0.0048	3.27	0.00	7.85	0.00	Deletion	638.60;1005.98
AMPL234945360 ; AMPL234953554			chr1	144994881	145016062	PDE4DIP	4	21182	0.56	0.0030	-0.00	0.18	0.00	6.88	Insertion	1899.09;1458.68
AMPL228147507 ; AMPL228147507			chr2	42513441	42515392	EML4	2	1952	0.26	0.0088	31.44	0.00	62.01	0.00	Deletion	431.79;1009.85
AMPL241845298 ; AMPL235884794			chr2	48033361	48033652	MSH6	3	292	0.25	0.0126	19.14	0.00	68.69	0.00	Deletion	264.13;781.42
AMPL326849834 ; AMPL326849834			chr4	55960961	55961136	KDR	2	176	0.74	0.0121	0.00	28.80	0.00	57.54	Insertion	804.43;264.11
AMPL222794331 ; AMPL222794331			chr5	112176626	112176888	APC	2	263	0.99	0.0969	0.00	2.70	0.00	5.01	Insertion	392.36;5.48
AMPL238691315 ; AMPL238691315			chr7	128846219	128846451	SMO	2	233	0.79	0.0414	0.00	2.23	0.00	4.41	Insertion	456.33;117.05
AMPL223129514 ; AMPL223129514			chr10	88649782	88649988	BMPR1A	2	207	0.63	0.0084	0.00	13.97	0.00	27.87	Insertion	901.62;509.47
AMPL293993809 ; AMPL293993809			chr10	97956798	97959856	BLNK	2	3059	0.33	0.0064	39.65	0.00	76.38	0.00	Deletion	543.33;1118.53
AMPL230546735 ; AMPL230851855			chr11	44129201	44265837	EXT2	29	136637	0.35	0.0242	0.12	0.00	53.83	0.02	Deletion	266.44;485.28
AMPL328491631 ; AMPL328491631			chr12	43833775	43833914	ADAMTS20	2	140	0.37	0.0053	18.79	0.00	37.20	0.00	Deletion	909.40;1590.08
AMPL222443847 ; AMPL222443847			chr15	91341452	91341577	BLM	2	126	0.58	0.0030	-0.00	0.80	-0.00	0.91	Insertion	1715.91;1255.55

Figure 3: CNV Block CNV Report- consecutive amplicons with the same call are combined.

The graphical report initially displays every region in the genome, but chromosomes can be selected for review one-at-a-time. Figure 4 illustrates the full graphical view with chromosome 11 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 11 (purple) is compared to the data for all chromosomes (gray) in the lower-left chart, it is easy to see that a few amplicons have a lower-than-normal ratio (outside the fitted interval). The lower-right graph shows dispersion fitting results.

The fitting process worked well for this data. The linear fit of dispersion to coverage was good (correlation was 0.829) and 97.61% of regions were inside the 99% confidence interval (lower left).

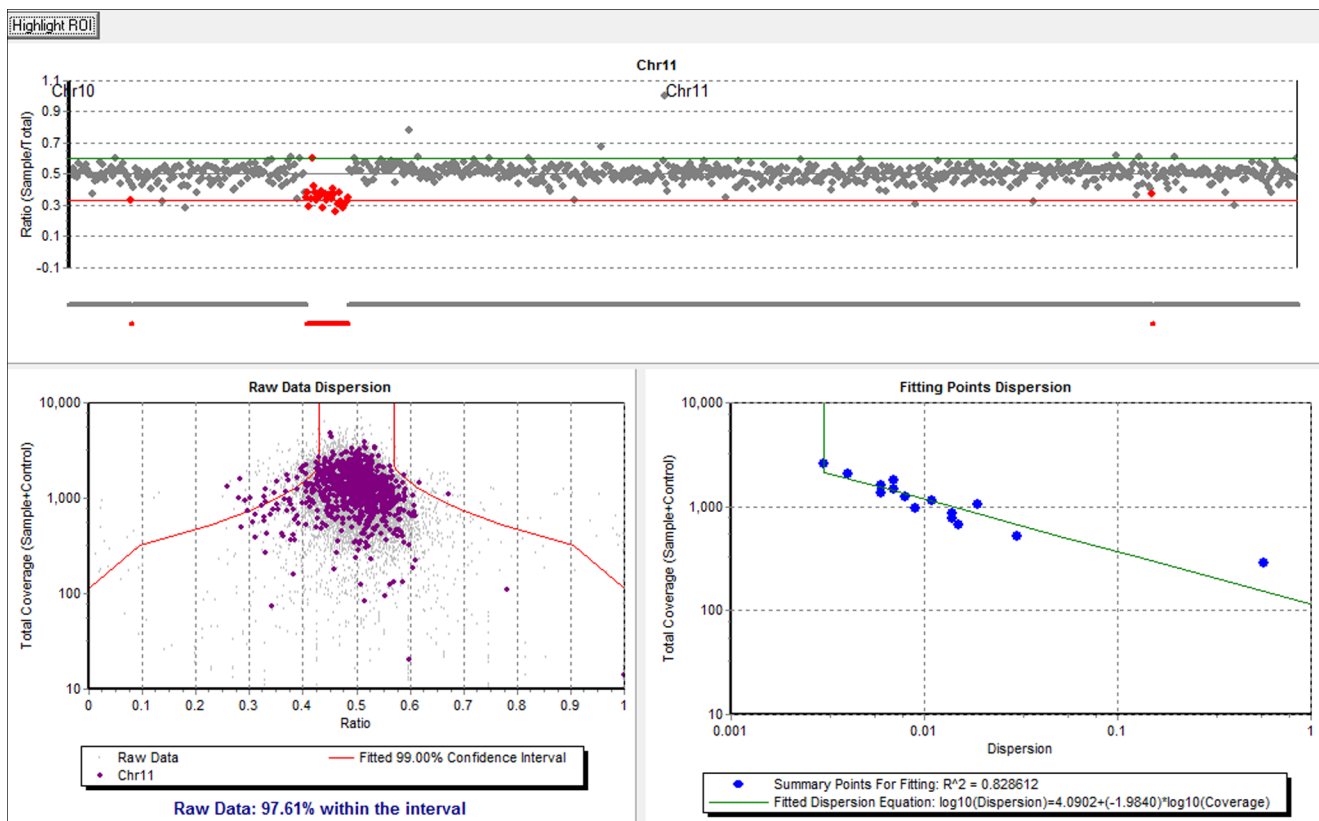


Figure 4: Graphical results with chromosome 11 selected

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 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).

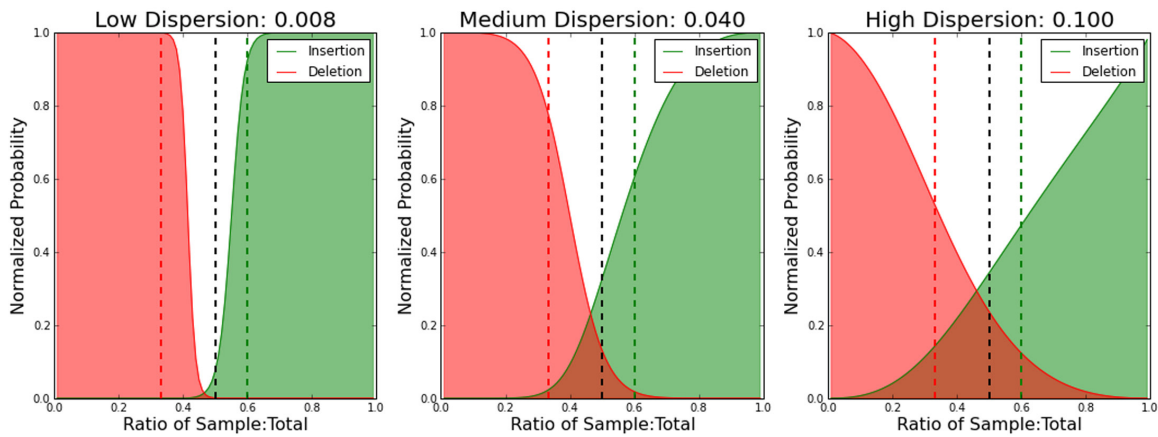


Figure 5: Normalized likelihoods at different dispersion values

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).

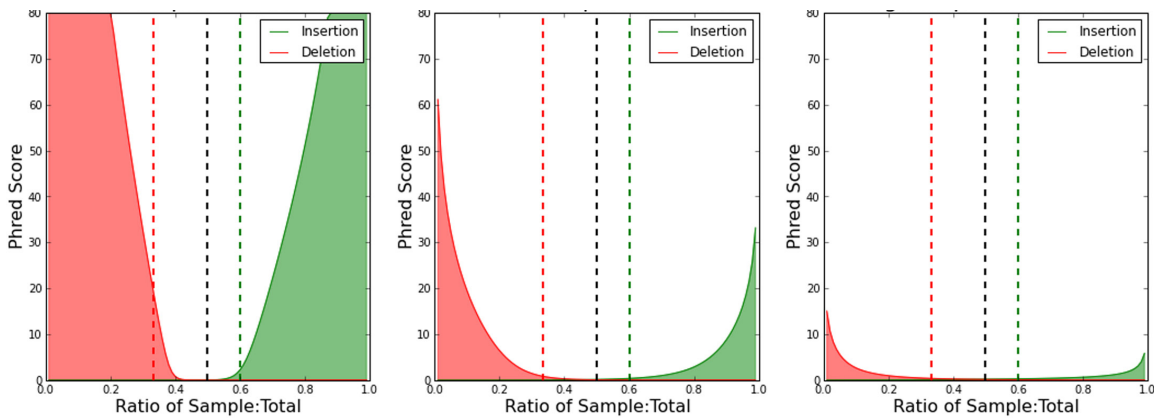


Figure 6: Distribution of Phred Scores across all possible ratios for three different levels of dispersion.

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 Life Technologies for making the data available on the Ion Torrent Community.

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.

Trademarks are property of respective owners