The normal and tumor spectrum of copy number variation: Copy number alterations correlate with changes in gene expression in tumor transcriptome.

Fiona C.L. Hyland, Rajesh Gottimukkala, Ryan Koehler, Xing Xu, Brian B Tuch, Ali Bashir, Vineet Bafna, Rebecca Laborde, Eric Moore, Jan Kasperbauer, David I Smith, Francisco De La Vega, and Asim Siddiqui, Applied Biosystems, Foster City, CA, University of California at San Diego, and Mayo Clinic, Rochester, MN..

Introduction – Copy Number Variation

Copy number variations (CNVs) have been widely observed in mammalian germline DNA and in tumor genomes, and CNVs are increasingly implicated in common disease (for example, mental retardation and schizophrenia) and in cancer progression. In humans, more total nucleotides exhibit variation due to alterations in copy number than due to single nucleotide diversity. Conrad et. al. recently report the discovery of 11,700 CNVs in humans of which 2000 to 4000 are polymorphic in human populations, and estimate that they can genotype up to 40% of polymorphic CNVs in humans with array technologies.

Massively parallel sequencing allows powerful, hypothesis-free genome-wide interrogation of CNVs. In contrast to array methods, with sequencing, genomic coverage data is available at single base resolution. We use the SOLiD^{TM} System to sequence human samples, including NA18507, HuRef, and a matched tumor/normal sample. The SOLiD System offers massively parallel ligation-based next-generation sequencing, with a throughput of 60Mbases per run. The data was analyzed with v3plus CNV algorithms, some within BioScope.

CNV Detection with the SOLiD System: Overview

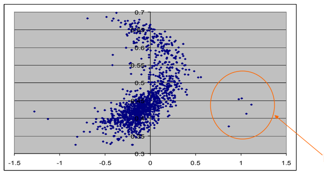
To detect copy number variation using massively parallel sequencing data, the following steps are performed.

1. Mapping – each read is mapped to the genome, and in the case of mate-pair experiments, the reads are paired, allowing interrogation of repetitive regions of the genome. Reads are mapped using the mapreads program in the BioScope framework.
2. Unique Mapping – those reads that map uniquely, and so can be confidently placed, are identified.
3. Normalization – non-overlapping windows of variable size are identified. Specification of the extent of the window differs for single sample and paired sample CNV detection; details below. Normalization is typically easier when comparing two samples to each other than when detecting CNVs in a single sample relative to the expected copy number based on the reference sequence.
4. Segmentation – a Hidden Markov Model (HMM) is used for segmentation, that is, identification of a contiguous set of windows having the same number of copies.
5. Copy number calling and p-value prediction – the HMM is used to predict the integer value of the copy number of each segment, and to predict the statistical significance.
6. Filtering – CNVs failing filtering criteria, such as an insufficient number of contiguous windows, are removed.

CNV Algorithm: Single Sample

We developed a single-sample CNV algorithm as follows. For normalization, we calculate coverage in variable-sized genomic windows that are selected to contain a constant number of mappable positions. (Using windows smoothes stochastic sampling noise but limits resolution.) For the human genome, we predict mappability for various run types (fragment or mate pair) and read length, predicting for each genome position whether it is likely to be capable of having reads uniquely map there or not, based on the degree of homology or repetitiveness elsewhere in the genome. Within these windows, we normalize coverage based on predicted mappability and GC content of this region of the genome; this is analogous to the array-CGH approach of normalizing based on intensity ratio using a matched sample. We then use a hidden markov model (HMM) for segmentation, and we apply empirically derived filters to the contiguous segments to call copy number variants.

Figure 1. Normalization for GC-content



Normalization

Sample variable size windows (1kb – 5kb mappable positions)

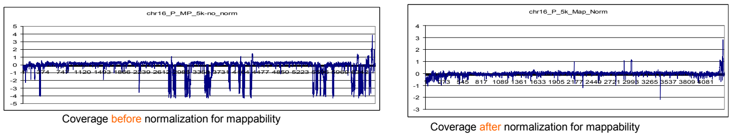
For each window *W*i, calculate local score *S*i

*S*i = log2 (observedi / expectedi)

Observed = Actual coverage (e.g. average 9X on window bases)

Expected = Global coverage

Figure 2. Mappability Normalization corrects for under-coverage of repetitive and homolgous genomic regions

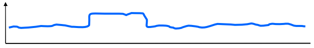


Coverage before normalization for mappability

Coverage after normalization for mappability

Segmentation: Hidden Markov Model

Figure 3. Segmentation



Observations = log ratio (observed coverage/expected)
State = Copy Number = [0 : k-1]
Initial State probabilities = [0.1/k   0.1/k   0.9   0.1/k   0.1/k   0.1/k   0.1/k]
Initial Transition probabilities t = $(1 – e^{- d*0.1})$ d ≈ distance between windows
Emission probabilities : state, variance of coverage

The Expectation Step: Compute the forward and backward probabilities
The Maximization Step: Re-estimate the model parameters using the Baum-Welch algorithm
Find the most probable sequence of states using the Forward-Backward algorithm

Post-Processing

Merge adjacent windows with similar Copy Number. Require:

- Minimum number of adjacent windows
- Minumum Mappability of window
- Maximum P-value
- Exclude 1 MB region around centromere and telomere

Table 1. CNV Concordance with orthogonal data

NA18507 1*35 at 4x	# CNVs	SOLiD with McCarroll	McCarroll with SOLiD	SOLiD with Toronto DB
Min CNV size 10kb Window size 5kb Min Mappability 10%	142	**0.43** 61/142	**0.352** 63/179	**0.887** 126/142
Min CNV size 5kb Window size 5kb Min Mappability 0%	283	**0.332** 94/283	**0.553** 99/179	**0.827** 234/283
Min CNV size 4kb Window size 2kb Min Mappability 10%	326	**0.331** 108/326	**0.620** 111/179	**0.804** 262/326
Min CNV size 2kb Window size 2kb Min Mappability 0%	635	**0.201** 128/635	**0.721** 129/179	**0.718** 456/635

HuRef 2*50 at 30x coverage	# CNVs	SOLiD with Toronto DB
Min CNV size 10kb Window size 5kb Min Mappability 10%	367	**0.807** 296/367
HuRef 2*50 at 8x coverage		
Min CNV size 10kb Window size 5kb Min Mappability 10%	404	**0.767** 310/404

McCarrol et at, 2008

CNV Algorithm: Paired Sample

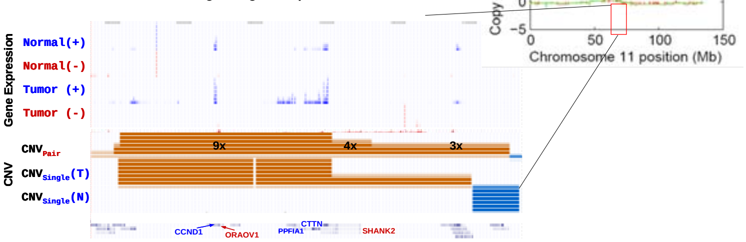
Normalization: In the case of paired-sample normalization, rather than comparing to the predicted mappabilty of the genome, the coverage of the test sample is normalized by comparing to the coverage of the control sample. Systematic issues of mappability, GC content, etc. are expected to be similar between both samples, simplifying normalization; this method is applicable to any species. The window size is variable, determined by fixing the number of positions of the control sample with coverage. To adjust for coverage differences in the samples, coverage of each window is first normalized by mean coverage of that sample. Both samples must be sequenced under the same conditions (e.g. both mate pair, both the same tag length).

RESULTS

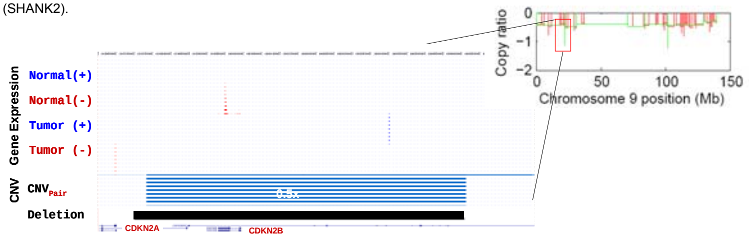
Control: Using the same sample to normalize itself, no CNVs were observed (no false positives) when 3 consecutive windows were required to call a CNV, and 4 CNVs were observed when 2 consecutive windows were required, suggesting a very low false positives rate for paired-sample CNV detection.

Tumor/Normal: We sequenced an oral squamous cell carcinoma (OSCC) and a matched normal sample to 0.8x coverage with the SOLiD System. We also sequenced the whole transcriptome of the tumor and normal samples using a new total RNA based protocol, and examined the correlation between copy number variation and changes in gene expression. We observed a significantly positive correlation between CNV and gene expression. These results suggest that gene duplication and deletion are key mechanisms driving the transcriptional profile changes of these tumor samples. The identified CNV segments offer insight into genes associated with the initiation or progression of cancer.

Fig 4: Examples of genomic regions that show strong correlations between CNVs and changes of gene expression.

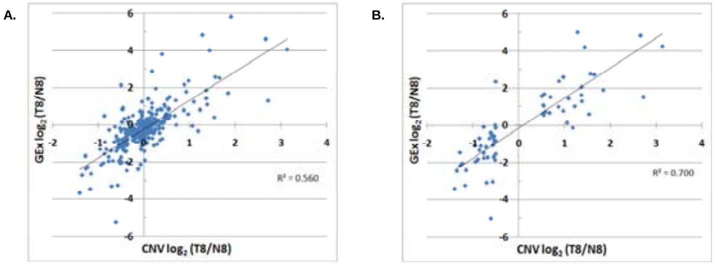


The most strongly amplified region, with 9-fold more copies in the tumor than normal is chr11:68,503,204-69,987,273. This region contains several differentially expressed (red and blue tracks) genes of interest: cyclin D1 (CCND1), oral cancer overexpressed 1 (ORAOV1), protein tyrosine phosphatase receptor (PPFIA1), cortactin (CTTN), and SH3 and multiple ankyrin repeat domains 2 (SHANK2).



This region (chr9:21,973,361-22,061,522), which shows evidence of having a single copy deletion in the tumor, contains two genes of interest: cyclin-dependent kinase inhibitor 2B (CDKN2B) and cyclin-dependent kinase inhibitor 2A (CDKN2A).

Figure 5. Large structural mutations are strongly correlated with tumor-specific changes in gene expression.



A) A strong correlation (*R = 0.73*) is observed between changes in copy number and changes in gene expression for patient 8. **B**) The correlation is stronger (*R = 0.84*) if only meaningful copy number changes (i.e., those greater than 1.4-fold) are considered.

CONCLUSIONS

CNVs can be accurately detected using the SOLID System of next-generation sequencing. We demonstrated 89% concordance with the Toronto database for single-sample CNVs larger than10kb, but we have also shown high concordance with CNVs larger than 2kb, suggesting a very high true positive rate. In addition we show good correlation with orthogonal data sets.

In contrast to detection of CNVs with array technology, CNV detection with sequencing can detect copy number increases at least as accurately as copy number decreases.

Paired-sample CNV detection has an extremely low false discovery rate. Detection of CNVs in tumor genomes is strongly positively correlated with changes in gene expression, suggesting a causative mechanism for transcriptional alterations in tumorgenesis and/or growth.

REFERENCES

McCarrol SA *et al.*, Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genetics* 40, 1166-1174 (2008)
Chiang DY *et al.*, High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* 6, 99 - 103 (2009)
Conrad D.F. *et al*. Origins and functional impact of copy number variation in the human genome. *Nature* doi:10:1038 (2009)

ACKNOWLEDGEMENTS

Thanks to Yutao Fu for running mappability calculations. Thanks to Matthew Muller, Catalin Barbacioru, Christina Bormann-Chung, Cinna Monighetti, Joel Brockman, Jeff Schageman, Jian Gu, Scott Kuersten, Robert Setterquist, Yongming Sun, Chunlin Xiao, Heather Peckham of Applied Biosystems for generating data and for data analysis.

TRADEMARKS/LICENSING

For Research Use Only. Not for use in diagnostic procedures.

Applied Biosystems is part of Life Technologies • 5791 Van Allen Way • Carlsbad, CA 92008 • www.appliedbiosystems.com