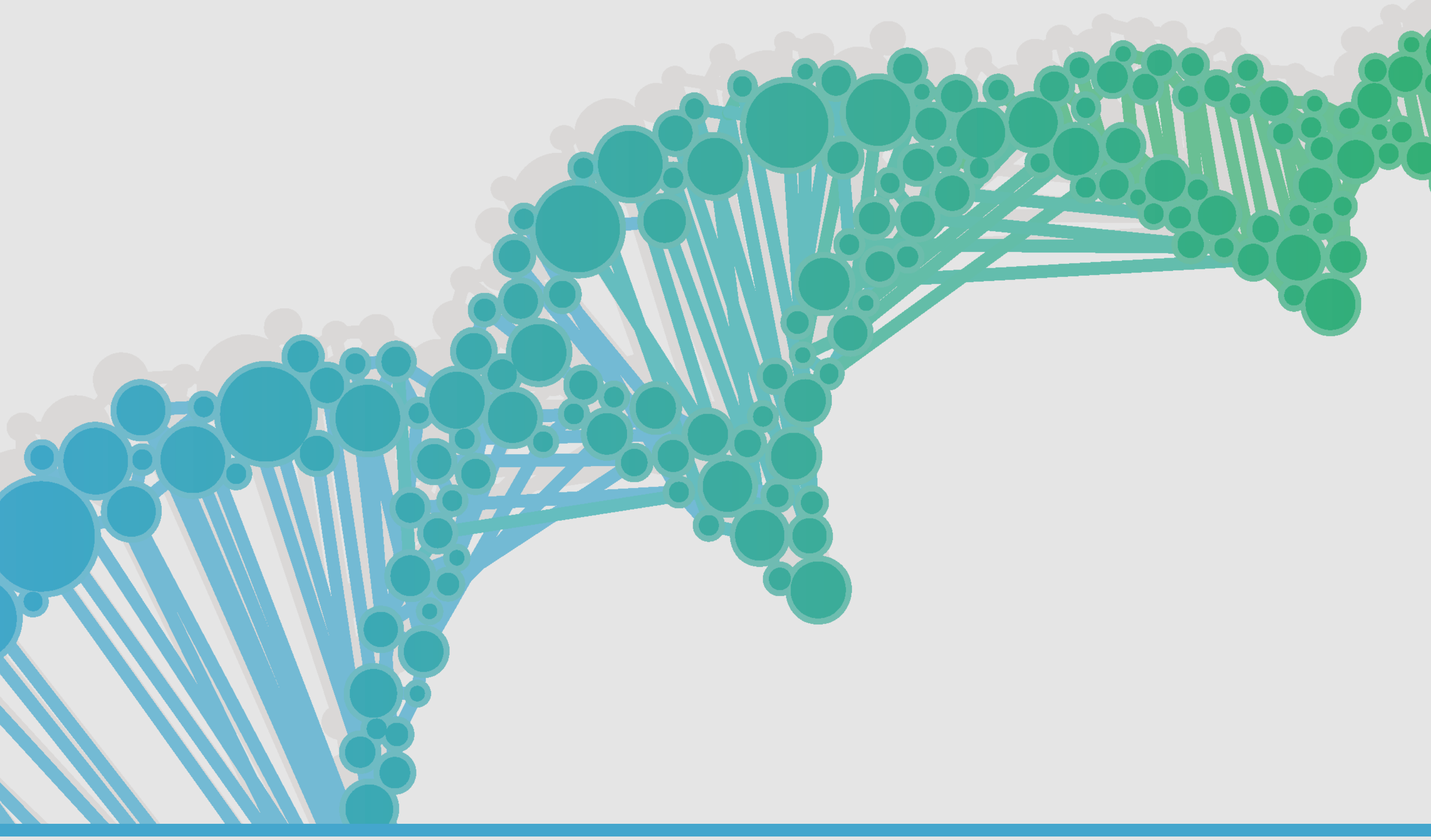
BioDiscovery



WHITEPAPER

Analyzing NGS Data For Copy Number Events

Next-generation sequencing (NGS) is mainly used to obtain sequence variants (SNVs). However, getting copy number variants (CNVs) from NGS has gained momentum in both research and clinical applications.

www.BioDiscovery.com

ANALYZING NGS DATA FOR COPY NUMBER EVENTS

Next-generation sequencing (NGS) is mainly used to obtain sequence variants (SNVs). However, getting copy number variants (CNVs) from NGS has gained momentum in both research and clinical applications. Nexus Copy Number 8.0 employs a new algorithm, BAM (pooled reference), to get copy number results from whole genome sequencing (WGS), whole exome sequencing (WES), or targeted panel NGS data. In addition to copy number events (gains and losses), this algorithm can directly extract B-allele frequencies (BAFs) from BAM files, resulting in a BAF plot along with the logRatio plot and the calling of Loss of Heterozygosity (LOH) and allelic imbalance (AI) events.

INTRODUCTION

Many different algorithms were designed to extract copy number information out of the NGS data. These algorithms can be grouped into Paired-End, Split-Read, Read-Depth, Assembly, or a combination of these methods. Even though no method alone can detect all types of CNVs, the Read-Depth method is most similar to the microarray methodology, still considered the gold standard for copy number detection. The BAM (pooled reference) algorithm in Nexus Copy Number 8.0 is a Read-Depth method, which builds a reference file from a group of BAM files and uses this reference file as the baseline for other test samples.

BAM (POOLED REFERENCE) ALGORITHM

The BAM (pooled reference) algorithm requires a reference file to be built first. The reference file is generated using the BAM Reference Builder, a separate graphic interface utility installed with Nexus Copy Number 8.0. A minimum of 10 "normal" BAM files are recommended for this purpose.

The first step in generating the reference file is the creation of an intermediary binary file from each BAM file used to create the reference. This file, called a depth file, represents the depth of coverage. If a region file is provided (used for targeted NGS panel data), one bin is used per region specified in the region file. If a region file is not used, a bin size of 100 bases is used. Values in each bin are the normalized read depth, defined as the number of bases within the bin divided by the product of the length of the bin and the total number of bases within all bins ***(Figure 1)***. A minimum depth in each bin in the minimum fraction of the samples for the reference file is used to exclude the regions with very low read depth. The median value for each bin is store into the reference file, which is then saved within the Nexus Copy Number program.

When the test samples' BAM files are loaded into Nexus Copy Number using BAM (pooled reference), each BAM file is compared with the reference file to get logRatios for all the probes (bins). B-allele frequencies (BAFs) are also generated based on the reads in all the SNP locations, after being checked for minimum reads and SNP call quality. Copy Number and Allelic events are then called using the SNP-FASST2 algorithm by segmenting the LogRatios and BAFs.

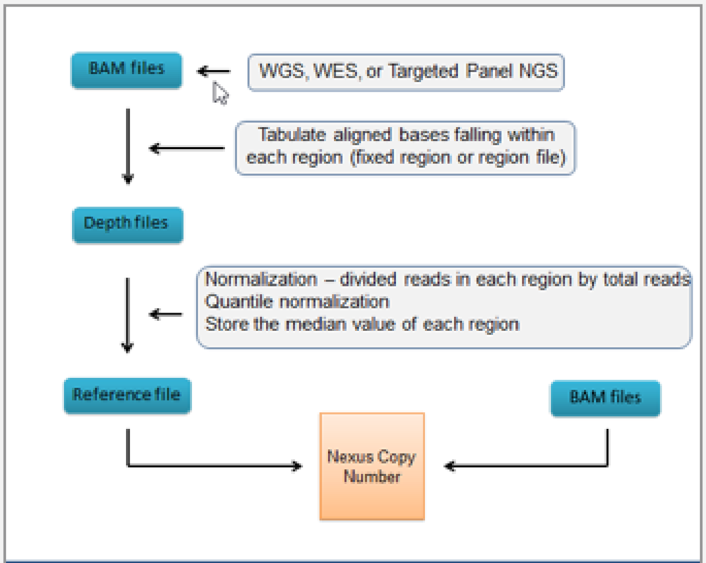

BAM files
WGS, WES, or Targeted Panel NGS
Tabulate aligned bases falling within each region (fixed region or region file)
Depth files
Normalization – divided reads in each region by total reads
Quantile normalization
Store the median value of each region
Reference file
BAM files
Nexus Copy Number

Figure 1. *BAM (pooled reference) workflow highlights the steps to build a reference file and to extract copy number results from BAM files in Nexus Copy Number 8.0.*

GC CORRECTION FOR DATA WAVINESS

Correction of data waviness related to GC contents in sequencing regions has a big impact on the data quality. In Nexus Copy Number, data quality is a measurement of the variance of logRatio probe distribution, after excluding the probes for the true biological events (e.g. copy number gains or losses). For NGS data, a probe (or pseudo-probe) is the bin size used to get the depth of the sequence reads.

To test the correction method suited for NGS data, different GC correction schemes (GC contents in regions of different sizes with or without the GC contents of the probes themselves) were applied to five TCGA COAD whole exome sequencing BAM files, after they are loaded and processed in Nexus Copy Number using the BAM (pooled reference) algorithm. The best quality after GC correction was found to come from the 50kb region size with or without the probes ***(Figure 2)***.

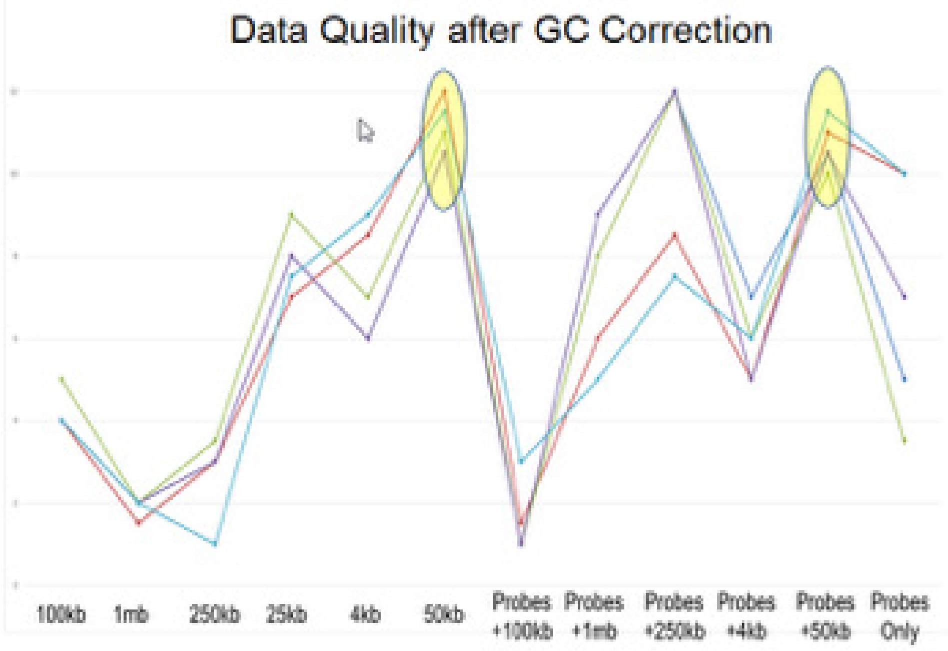


Figure 2. *Best data quality comes from correction method employing the 50kb region size with or without the probes (yellow circled areas).*

COPY NUMBER AND ALLELIC EVENTS FROM NGS

Just like results from SNP array data, logRatios and BAFs are obtained after BAM files are processed by the BAM (pooled reference) algorithm. With both logRatio and BAF plots, copy number events corroborates with allelic events ***(Figure 3)***. Together they are very helpful for assessing genomic events, and for adjusting logRatio baseline due to tumor ploidy, if necessary. Loss of Heterozygosity (LOH), especially Copy Neutral - Loss of Heterozygosity (CN-LOH), can also be identified ***(Figure 4)***. Practically, BAM (pooled reference) algorithm turns NGS data into SNP array data.

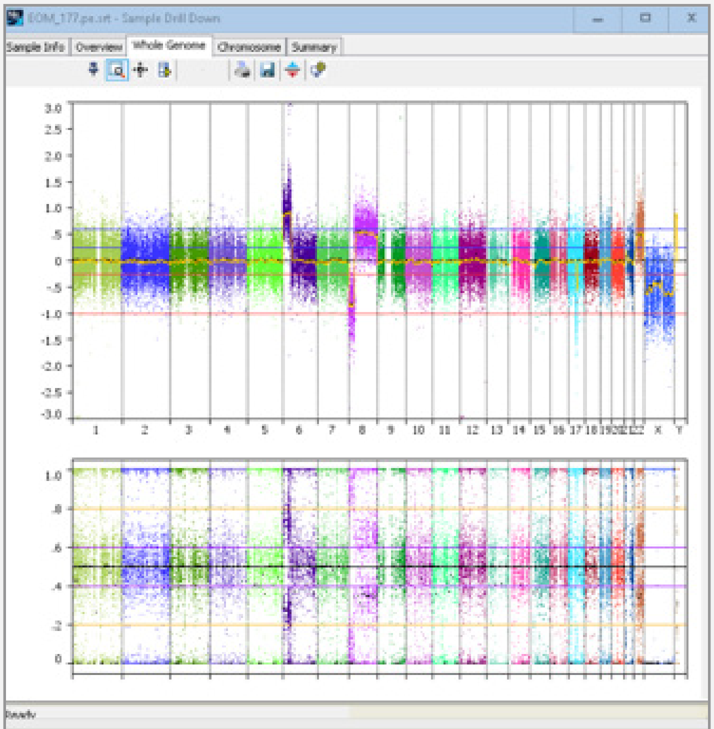


Figure 3. *Whole genome view of a WES sample shows logRatio plot (above) and BAF plot (below).*

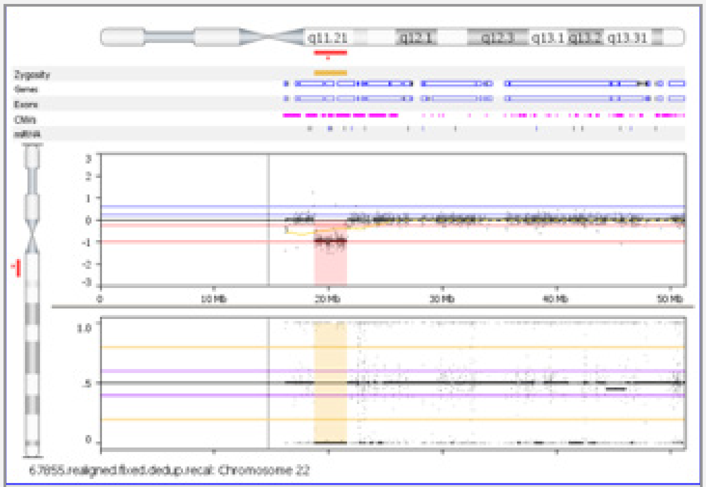


Figure 4. *A DiGeorge syndrome sample displays a copy number loss at 22q11.21, with an accompanied LOH call in the same region (above: logRatio plot, bottom: BAF plot).*

COMPARISON WITH CONIFER AND XHMM

CoNIFER1 and xHMM2 are two popular algorithms to analyze NGS data for copy number events. Like BAM (pooled reference) in Nexus Copy Number, both of these methods calculate the depth of coverage across exonic regions. But unlike it, they require strong bioinformatic expertise from the users, due to the methods' reliance on line commands and scripting. Also, CoNIFER and xHMM are suited for WES data from constitutional samples, while BAM (pooled reference) can analyze WGS, WES, and targeted panel NGS data from constitutional and cancer samples.

There are limitations in using CoNIFER and xHMM. CoNIFER requires at least 50 million mapped reads and a minimum of 8 exome samples to run at a time, while xHMM recommends ~50 exome samples with at least 60 - 100x coverage. In contrast, Nexus Copy Number's BAM (pooled reference) does not have such limitations, even though it suggests a minimum of 10 control samples for building the reference file to improve data quality.

A group of 19 control samples and 9 DiGeorge syndrome patient samples (8 with 22q11.21 deletion and 1 with CN - LOH in the same region) were prepared with Agilent WES SureSelect V5 protocol and run on Illumina HiSeq2500. The WES BAM files were analyzed by CoNIFER, xHMM, BAM (pooled reference) in Nexus, and the results were compared with each other and with the Affymetrix CytoScan HD array results3. Both ROC curves for sensitivity vs specificity and the call accuracy comparisons indicate that Nexus Copy Number's BAM (pooled reference) has the best behavior ***(Figure 5 & 6)***.

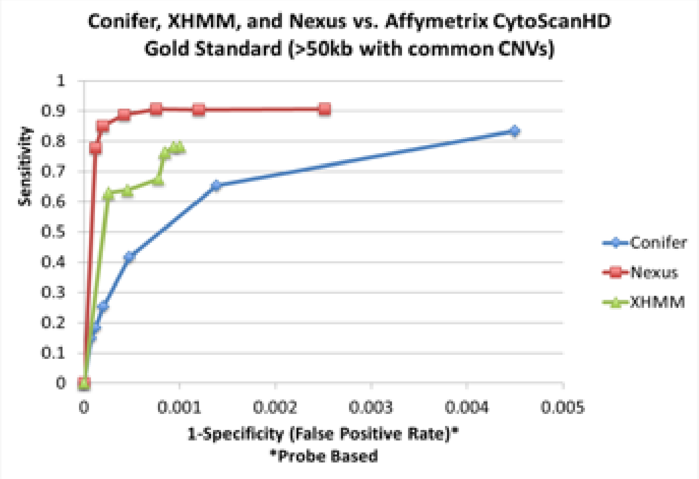


Figure 5. *Receiver operator curves (ROC) of CoNIFER, xHMM, and BAM (pooled reference) in Nexus Copy Number on the same DiGeorge samples.*

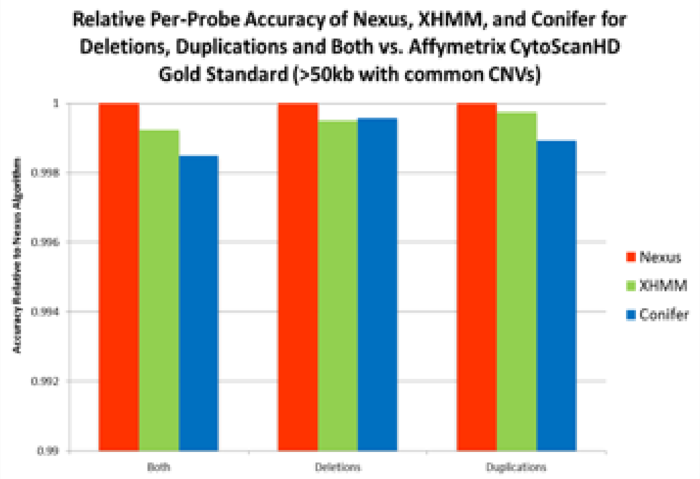


Figure 6. *Relative accuracy for calling CNVs with CoNIFER, xHMM, and BAM (pooled reference) in Nexus Copy Number on the same DiGeorge samples.*

SUMMARY

Nexus Copy Number 8.0 provides a new method, BAM (pooled reference), to streamline the process for analyzing copy number and allelic events from WGS, WES, and targeted panel NGS data. Similar to other functions in Nexus Copy Number, this method presents to the users an easy-to-use graphic interface, unlike other algorithms that would require the users to go through line commands and scripting. This tool adds great value to Nexus Copy Number, which now allows scientists to analyze for CNVs from aCGH arrays, SNP arrays, and all kinds of NGS data.

REFERENCE

1. Krumm, et. al., Copy number variation detection and genotyping from exome sequence data. Genome Res. 2012 Aug; 22(8): 1525-1532.

2. Fromer, et. al., Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet. 2012 Oct 5;91(4):597-607.

3. Darbro, et. al. (University of Iowa), Unpublished data.