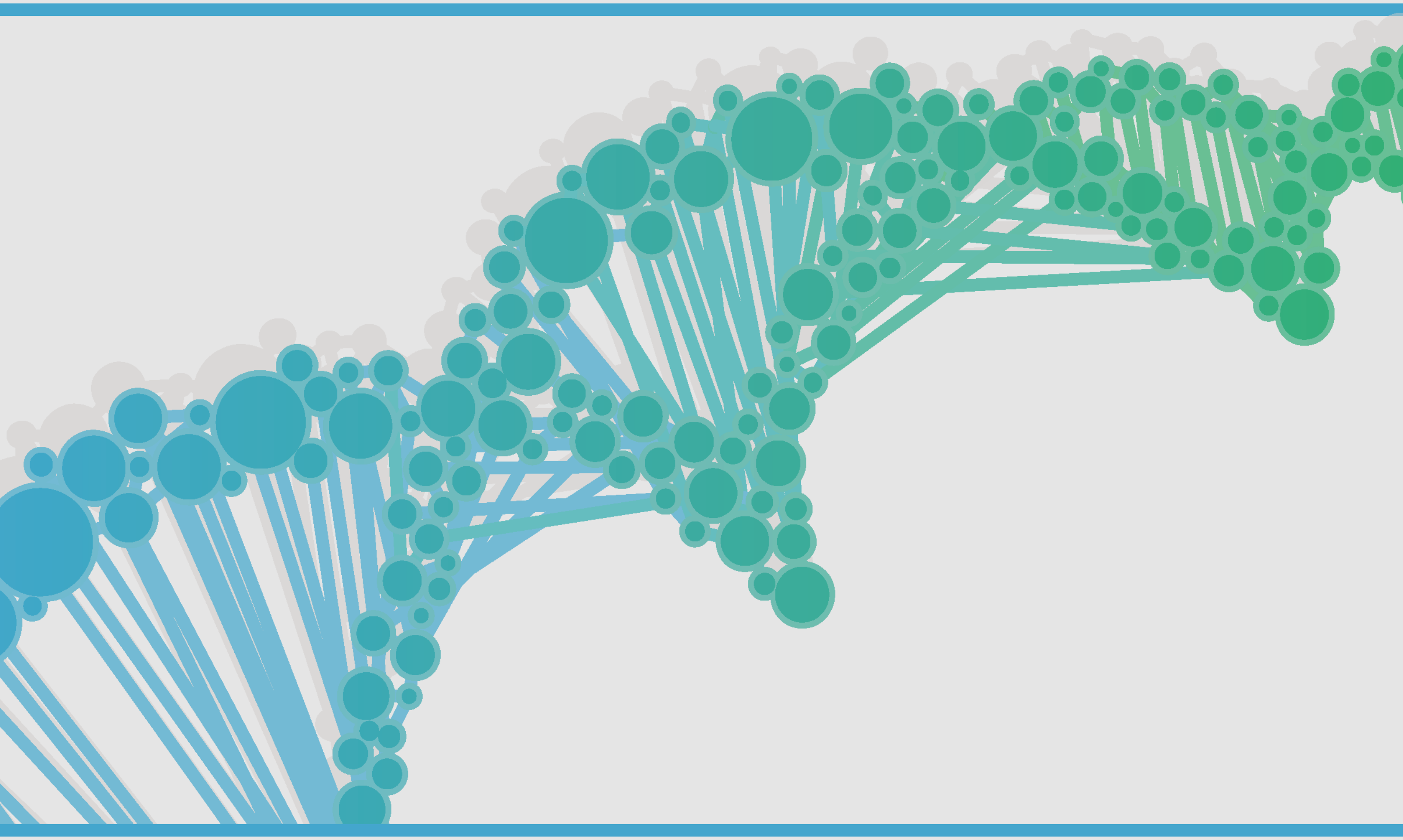

WHITEPAPER

Comparison of the BAM (multiscale reference) algorithm to other methods for CNV detection from NGS

Learn more about the various methods that have been proposed to detect CNVs from NGS data in tumor-normal matched paired colon adenocarcinoma samples.

www.BioDiscovery.com

INTRODUCTION

Copy number variations are a significant source of genetic diversity in humans. The platform of choice to detect genome-wide CNVs has traditionally been microarray, including SNP arrays which can also detect copy neutral LOH regions. Next-generation sequencing (NGS) has advanced rapidly and is the platform of choice for sequence variants. Decreasing costs have made it much more affordable to perform sequencing and the need for deriving copy number from NGS data has been rising. There are several tools available (CoNIFER, xHMM, CNVkit, QDNA) for this but most are not user-friendly; they require strong bioinformatics expertise and use of the command line. Often each tool is adept in only one arena (e.g. cancer or constitutional samples, data from WGS or targeted panels.

Nexus Copy Number is a software solution for copy number estimation from all types of NGS data; in addition, it has an interactive graphical interface which is ideal for researchers. The software has a couple of depth-of-coverage based algorithms to derive copy number from NGS. The BAM ngCGH (matched) method requires matched normals and handles whole-genome sequencing (WGS) or whole exome sequencing (WES) data. The BAM (pooled reference) method (in version 8.0) handled WGS or WES without matched normals.

A new algorithm incorporated into Nexus Copy Number 9.0 is the BAM (multiscale reference) method which improves upon and replaces the BAM (pooled reference) method. BAM (multiscale reference) functions well with both shallow/targeted sequencing data and WGS/WES with normal depth of coverage. It builds a reference file from a set of normal samples and uses adjustable dynamic binning. The method uses a Hidden Markov Model to segment the genome into target areas using the reads in targeted regions and the backbone areas using the off-target reads and additional areas. Coarse binning is used in the backbone areas to provide the copy number baseline as well as large copy number events and fine binning is used in target areas to provide high resolution copy number detection in targeted regions. The adjustable dynamic binning is very flexible allowing adjustment of the minimum bin width based on the depth of coverage. The dynamic binning allows the target regions to get more coverage and the backbone regions, less coverage but the backbone still gets coverage.

Presented here is a comparison between algorithms and platforms of the results from microarray and NGS samples. Three sets of NGS data were used: cancer panels, shallow (low pass) WGS for PGD (preimplantation genetic diagnosis), and WES constitutional samples. NGS data was processed by various methods: CNVkit[1], BAM ngCGH (matched) and BAM (multiscale reference).

Data used:

- Large-cell Lung Carcinoma (LCC) panel data with eight normals including pair matched normals
- Shallow (low pass) PGD (Preimplantation Genetic Diagnosis) WGS data (<1x)
- Constitutional (DiGeorge Syndrome) WES with normal depth of coverage (30x)

METHODS

BAM files from custom NGS panels were obtained for Large-cell Carcinoma of the lung (LCC) samples. The BAM files were processed with two different algorithms within Nexus Copy Number: BAM ngCGH (matched) and BAM (multiscale reference).

The BAM ngCGH (matched) processing is based on the ngCGH algorithm developed by Sean Davis at the NCI[2]; it computes pseudo-log ratios using simple coverage counting on the tumor relative to the normal. Each tumor sample was processed with its pair-matched normal.

BioDiscovery's BAM (multiscale reference) algorithm is a read-depth method that uses a pooled reference file to generate pseudo-log ratios based on the reads. It also generates B-allele frequencies based on the reads at SNP locations. Eight normal samples were used to create the pooled reference using the MultiScale BAM Reference Builder Utility packaged with Nexus Copy Number software.

After pseudo-probes were generated from the NGS methods, segmentation was performed using BioDiscovery's SNP-FASST2 segmentation algorithm. Affymetrix SNP 6.0 CEL files were loaded directly into Nexus Copy Number 9.0 and processed using BioDiscovery's SNP-FASST2 segmentation algorithm.

RESULTS

LCC tumor samples

For the tumor data set, an LCC sample was processed with CNVkit, BAM ngCGH (matched), and BAM (multiscale reference). As BAM ngCGH (matched) requires use of a pair-matched normal, several different normal samples were used as the reference to see if and what difference it would make. The sample LCC2 was processed via BAM ngCGH (matched) using its pair-matched normal (LCC2 normal) as well as the normal samples of seven other LCC samples. Most of the samples with matched normals gave good results except #5 which was quite noisy ***(Figure 1)***. The quality score shows noisiness of data; the lower the quality score, the better the results.

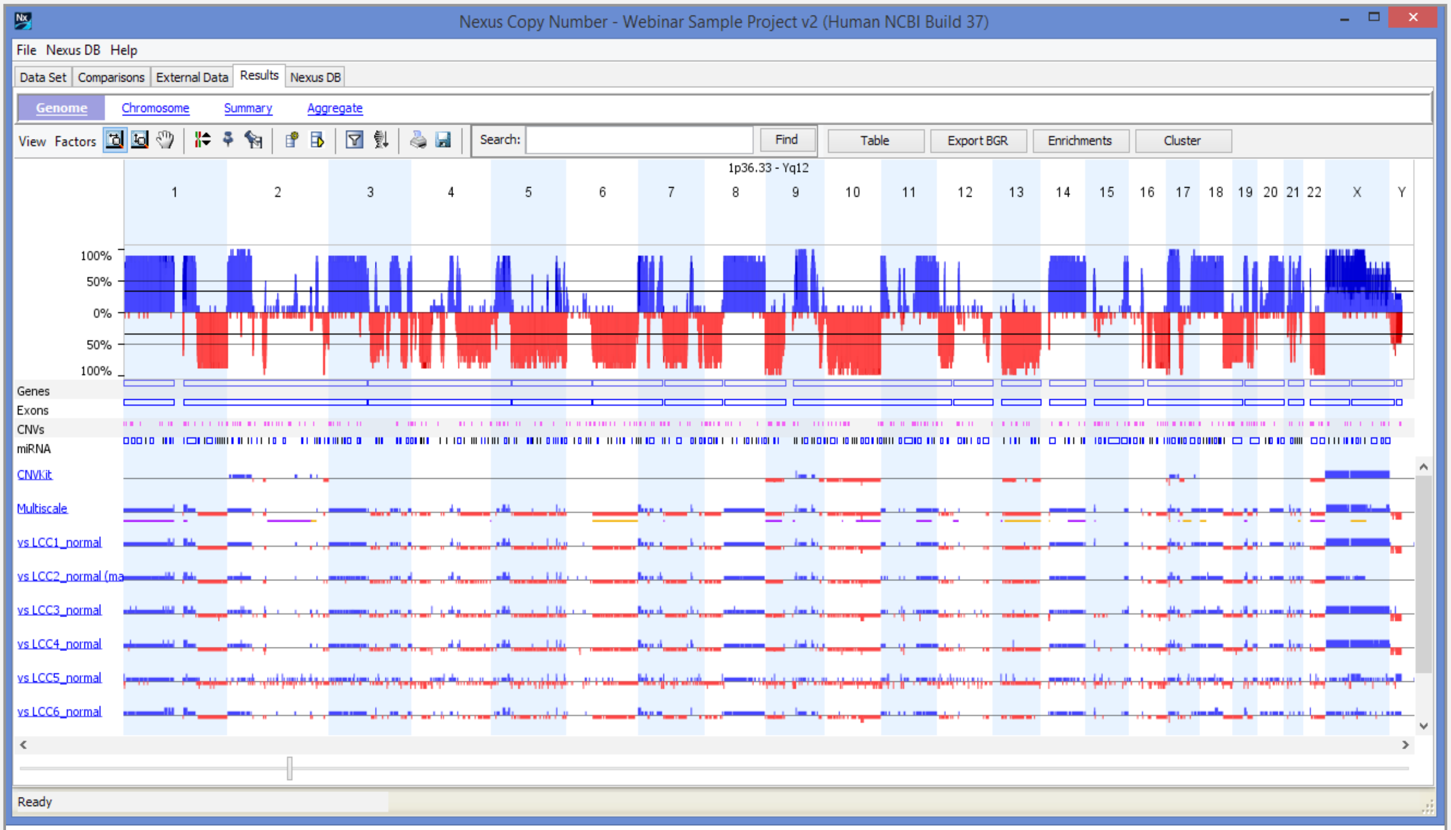

Figure 1. Overlap of regions of change from copy number estimation with CNVkit, BAM (multiscale reference), and BAM ngCGH (matched). Blue indicates gain, red indicates loss. The sample name "CNVkit" indicates algorithm used; BAM (multiscale reference) algorithm was used for the sample "MultiScale". The rest of the samples were processed with BAM ngCGH (matched) using different normal samples for the pair-matched reference. The sample name (e.g. "vs LCC2_normal") indicates the reference sample. Concordance of copy number estimation is highly dependent on the reference sample.

Figure 2 is a plot of the quality scores of all samples. Of all the tumor/normal pairs, the pair-matched sample (LCC2 tumor – LCC2 normal) gave the lowest quality score. But using the multiscale algorithm with a pooled reference of all eight normal samples, the quality score was the lowest among all including the pair-matched reference. If using pair-matched normal with the BAM ngCGH (matched) algorithm, a good quality score is returned. If the normal is from a different sample (e.g. LCC2 tumor – LCC3 normal), the quality of the results is not consistent; some pairs have a decent quality score and some can be quite bad. If combining several normal samples together for a pooled reference as with the BAM (multiscale reference) algorithm, a good quality score is obtained. If an outlier sample is in there (e.g. LCC5), it doesn't affect the overall quality.

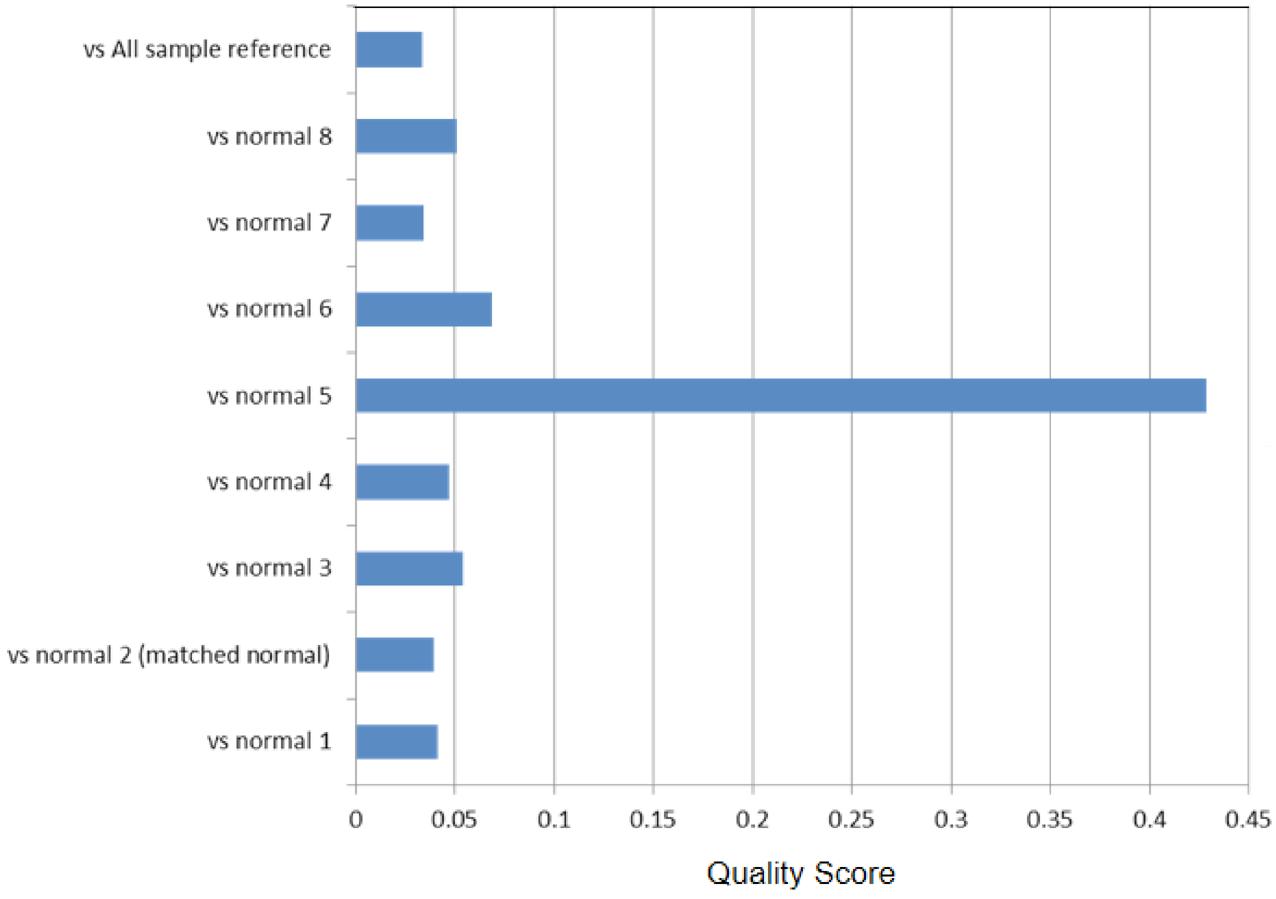

Figure 2. Quality scores of the various processed samples. LCC2 tumor with LCC5 normal as the reference (vs normal 5) gave the highest (worst) quality score. The score was almost 0.45 whereas the rest were all under 0.1. The best quality score was obtained using a pooled reference of eight normal samples (vs All sample reference).

Looking at the pseudo-probes plot of each sample ***(Figure 3)***, one can see that they are very similar. Note that CNVkit only produces pseudo probes for the targeted regions (chr 2, 9. 10, 13, 22). The BAM (multiscale reference) algorithm also estimates B-allele frequency (BAF) from BAM files ***(Figure 4)***. A zoomed in view of all three samples shows ample number of pseudo-probes and good concordance for the TP53 region ***(Figure 5)***.

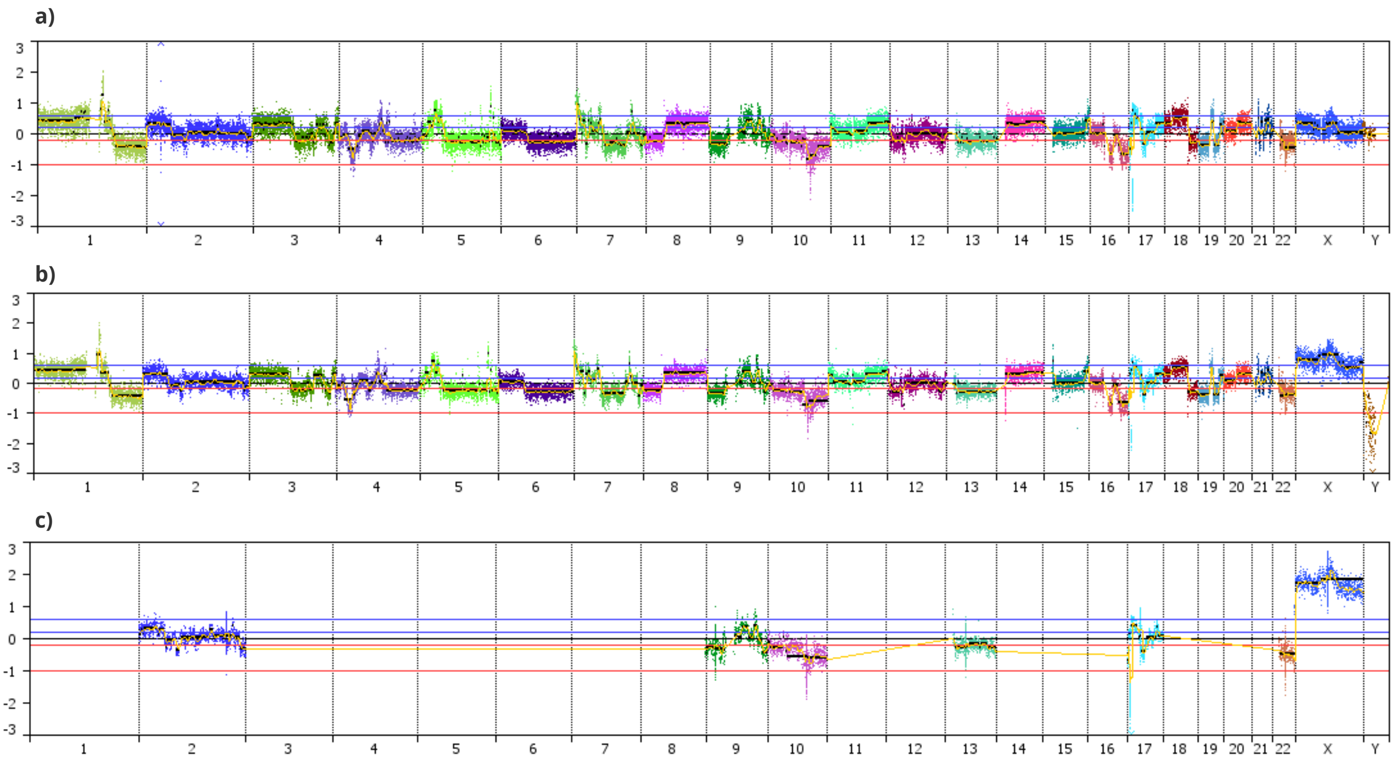

Figure 3. Whole genome probes plot of the three samples. a) BAM ngCGH (matched) vs. pair-matched normal. b) BAM (multiscale reference) c) CNVKit. All three algorithms produced pseudo-probes with good consensus for the targeted regions.

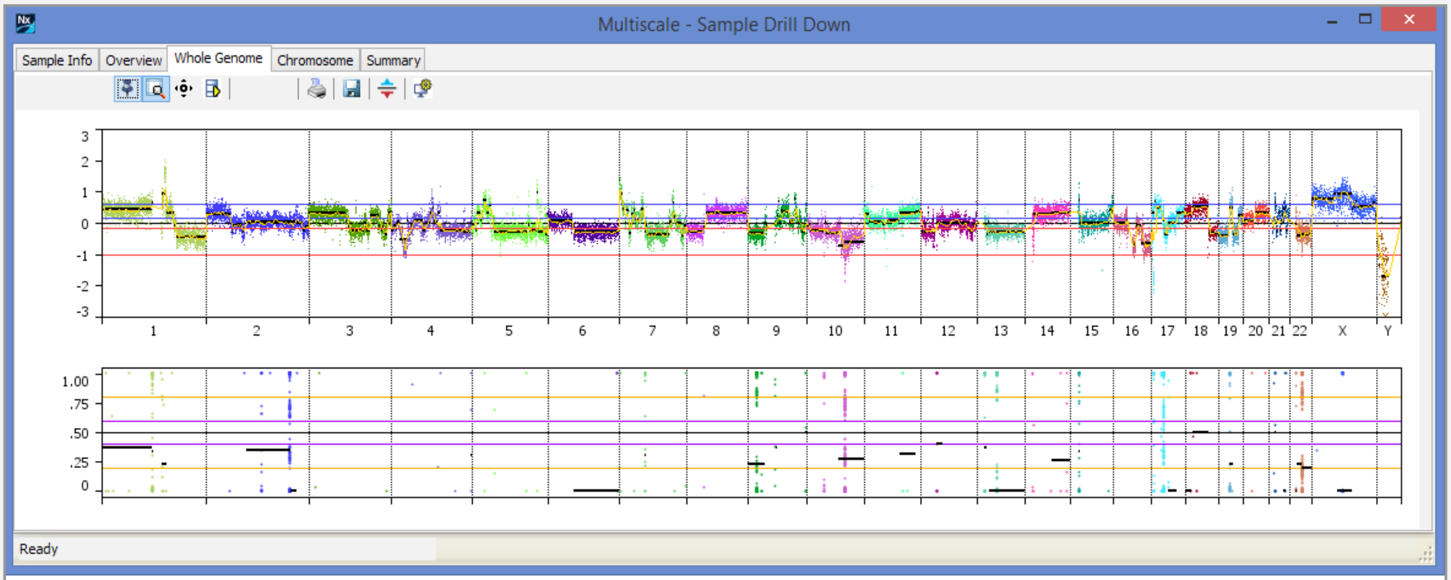

Figure 4. Whole genome probes plot and BAF plot of the BAM (multiscale reference) sample. The BAM (multiscale reference) algorithm also estimates BAF (lower part of figure).

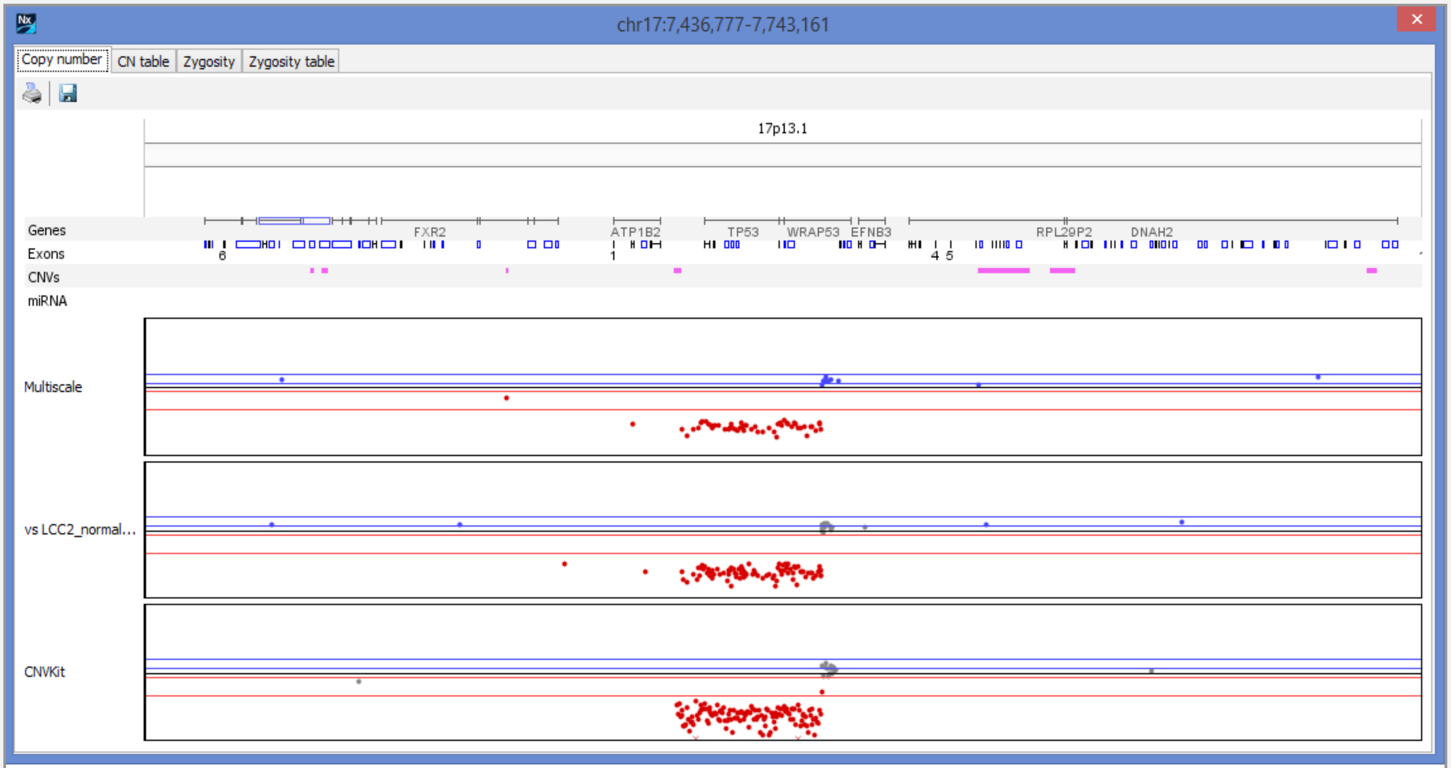

Figure 5. Probes plot of all three samples. All samples displayed good concordance for the TP53 region showing ample number of pseudo-probes and high copy loss.

PGD samples (low pass/shallow WGS <1x):

Two PGD samples (#1 and #2) were run via BAM ngCGH (matched) and BAM (multiscale reference). Since the algorithm requires a pair-matched reference, the sample was run using two different reference samples (one male and one female). The ngCGH2 sample is female and the ngCGH1 sample is male; both were run using a male and a female reference sample.

Figure 6 shows good consensus among the copy number changes across samples via both algorithms but BAM ngCGH (matched) results are dependent on the choice of reference file. The first ngCGH2 sample (female) was run against a male reference and therefore shows a gain on the X chromosome (***Figure 7a***). The second ngCGH2 sample was run against a female reference and therefore the sex chromosome calls were made correctly (***Figure 7b***). Since the BAM ngCGH (matched) algorithm depends on a single reference sample, the results could have a greater variance due to the reference sample used. As the BAM (multiscale reference) method uses a pooled reference (***Figure 7c***), the results are cleaner and more consistent.

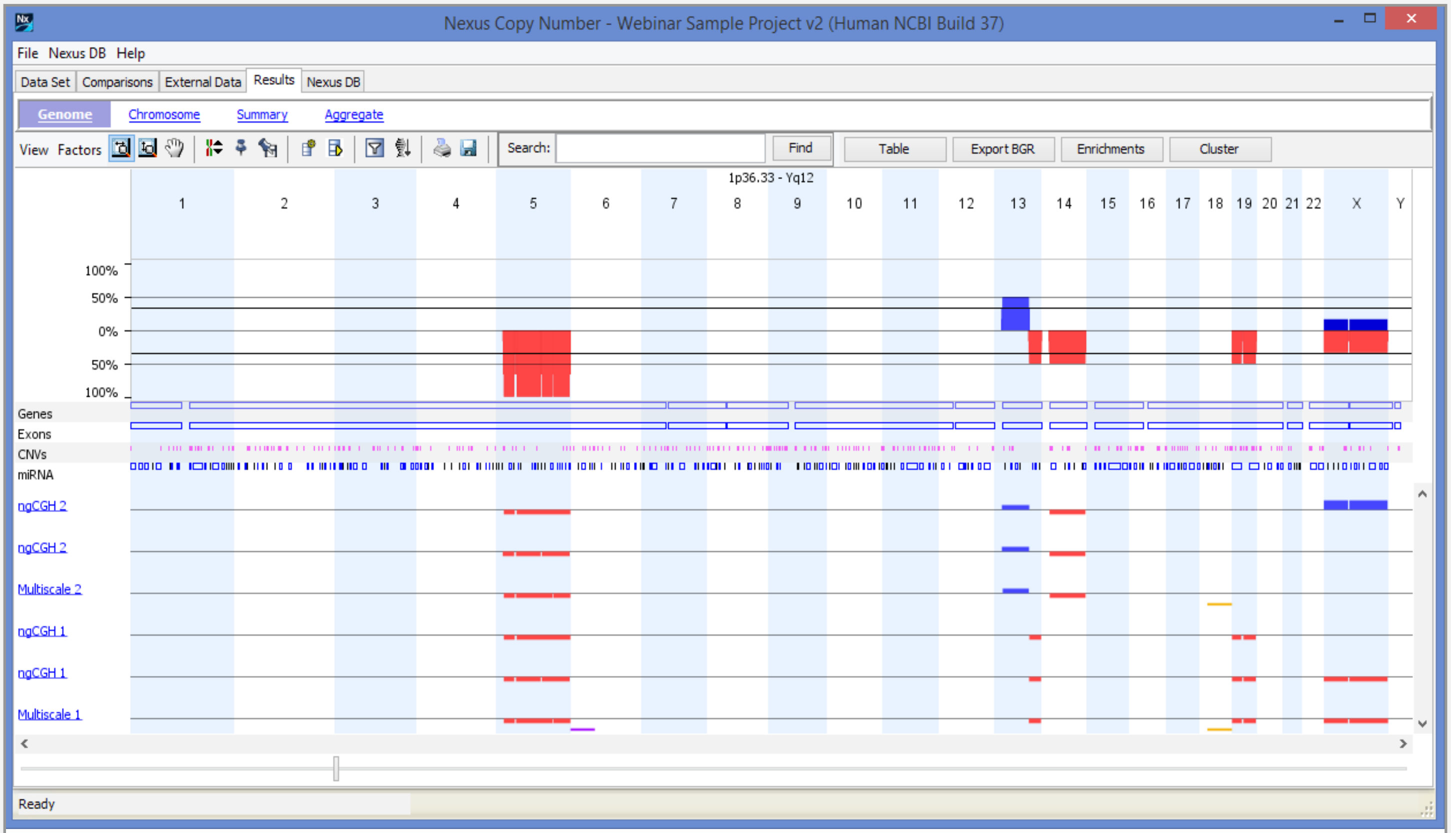

Figure 6. Copy number changes across the genome of two PGD samples processed via BAM ngCGH (matched) and BAM (multiscale reference) algorithms. Overall samples show high concordance among gains and losses.

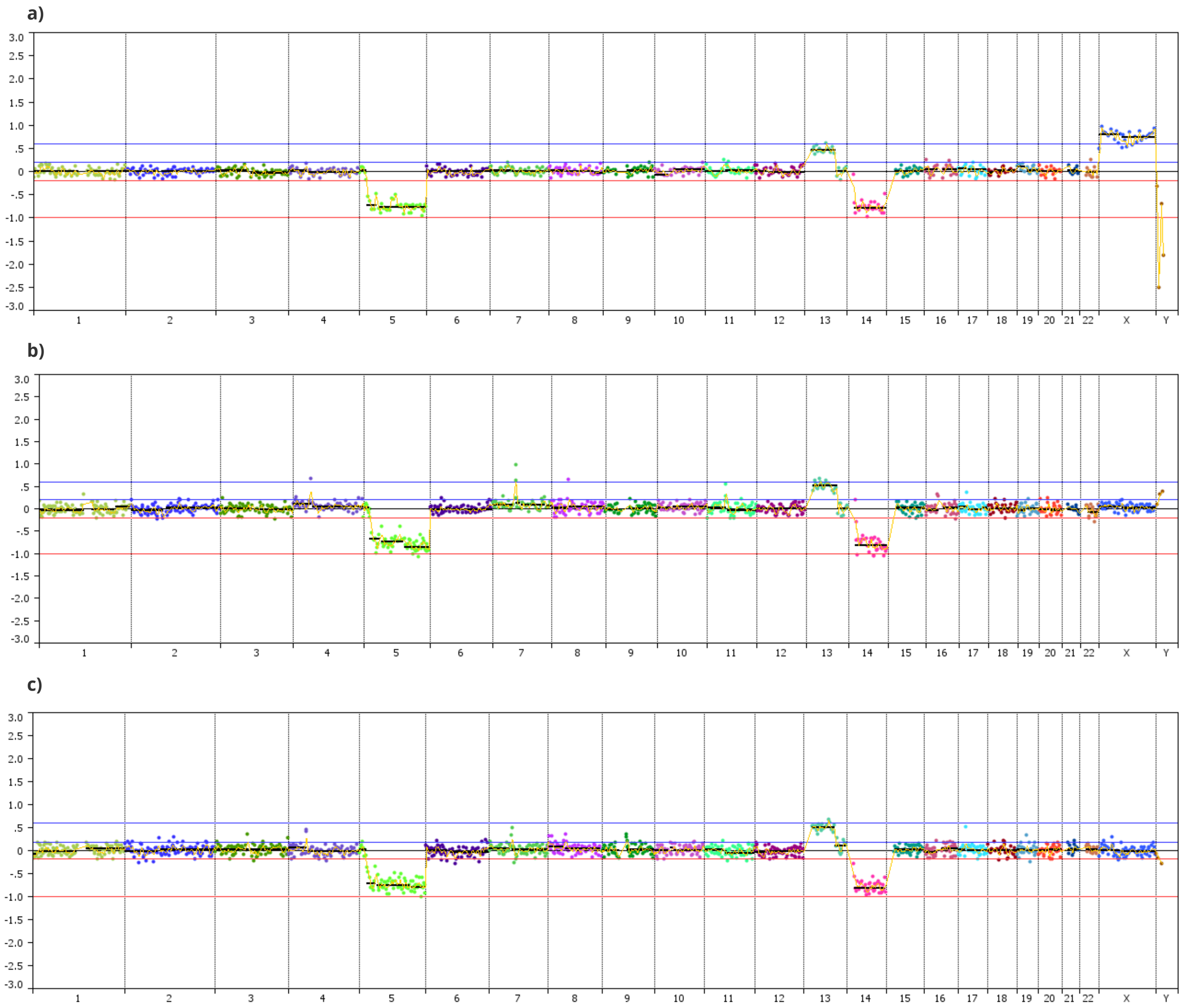

Figure 7. Whole genome probes plot of three samples. a) BAM ngCGH (matched) vs. sex mismatched normal. b) BAM ngCGH (matched) vs. pair-matched normal. c) BAM (multiscale reference). The probes profiles of BAM ngCGH (matched) with pair matched normal and BAM (multiscale reference) show good consensus.

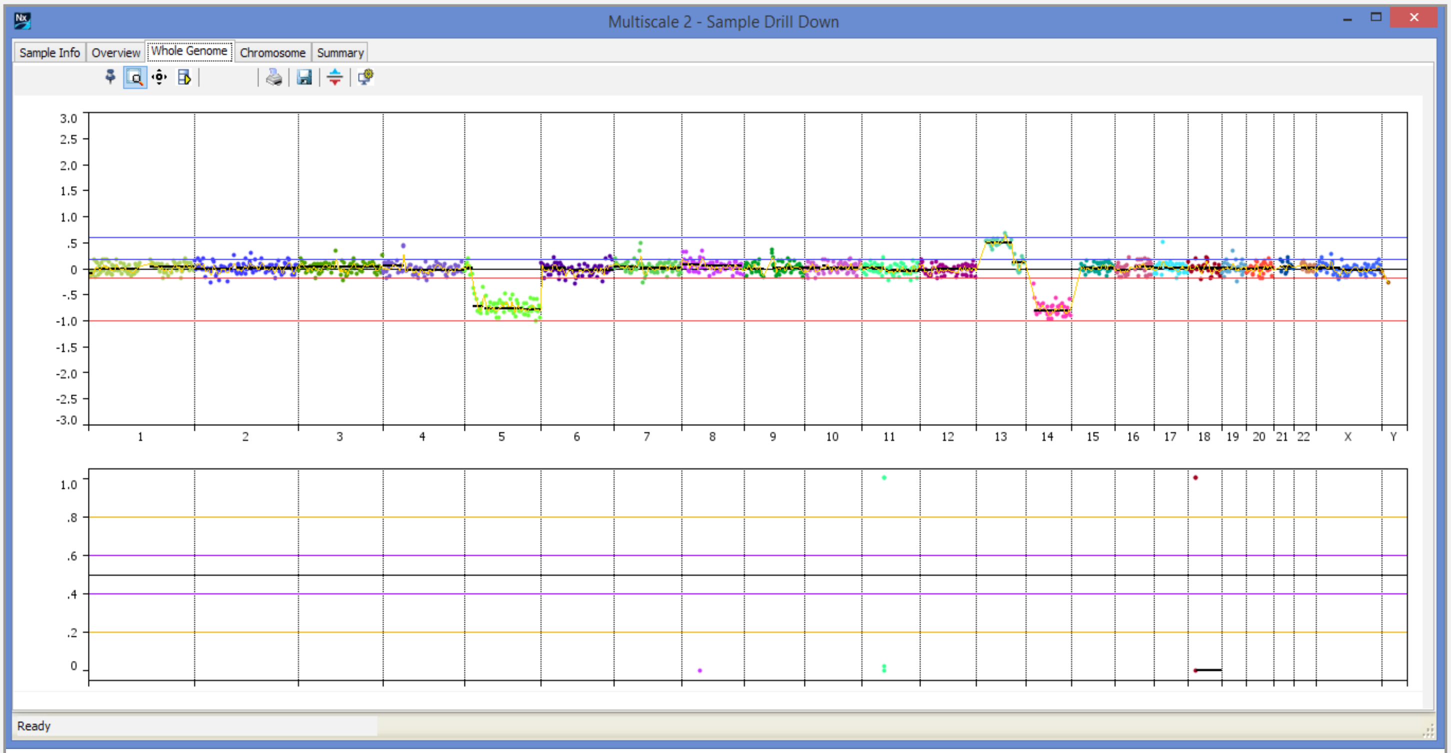

Figure 8. Whole genome probes plot and BAF plot of the BAM (multiscale reference) sample. Top part is the log ratio plot and the bottom is the BAF plot. The BAF plot is virtually empty because there are very few SNP probes to get LOH information for shallow sequencing samples due to the low coverage.

DiGeorge Syndrome samples (WES with normal depth of coverage 30x):

For the WES constitutional samples, there were both microarray and WES samples. The array sample was an Affymetrix CytoScan HD array. The WES sample was run using BAM ngCGH (matched) and BAM (multiscale reference) algorithms. The DiGeorge Syndrome is characterized by a loss on chromosome 22q11.2.

All three showed consistent results across chromosome 22 with the q11.2 loss ***(Figure 9)***. Looking at the probes view shows that the sequencing platform presents fewer pseudo-probes but the calls are consistent ***(Figure 10)***. The Affymetrix CytoScan HD is a very dense array and therefore shows more probes than WES. The BAM (multiscale reference) sample shows a very clean BAF plot ***(Figure 11)***. The 22q11.2 deletion region shows no heterozygous probes, confirming a single copy loss.

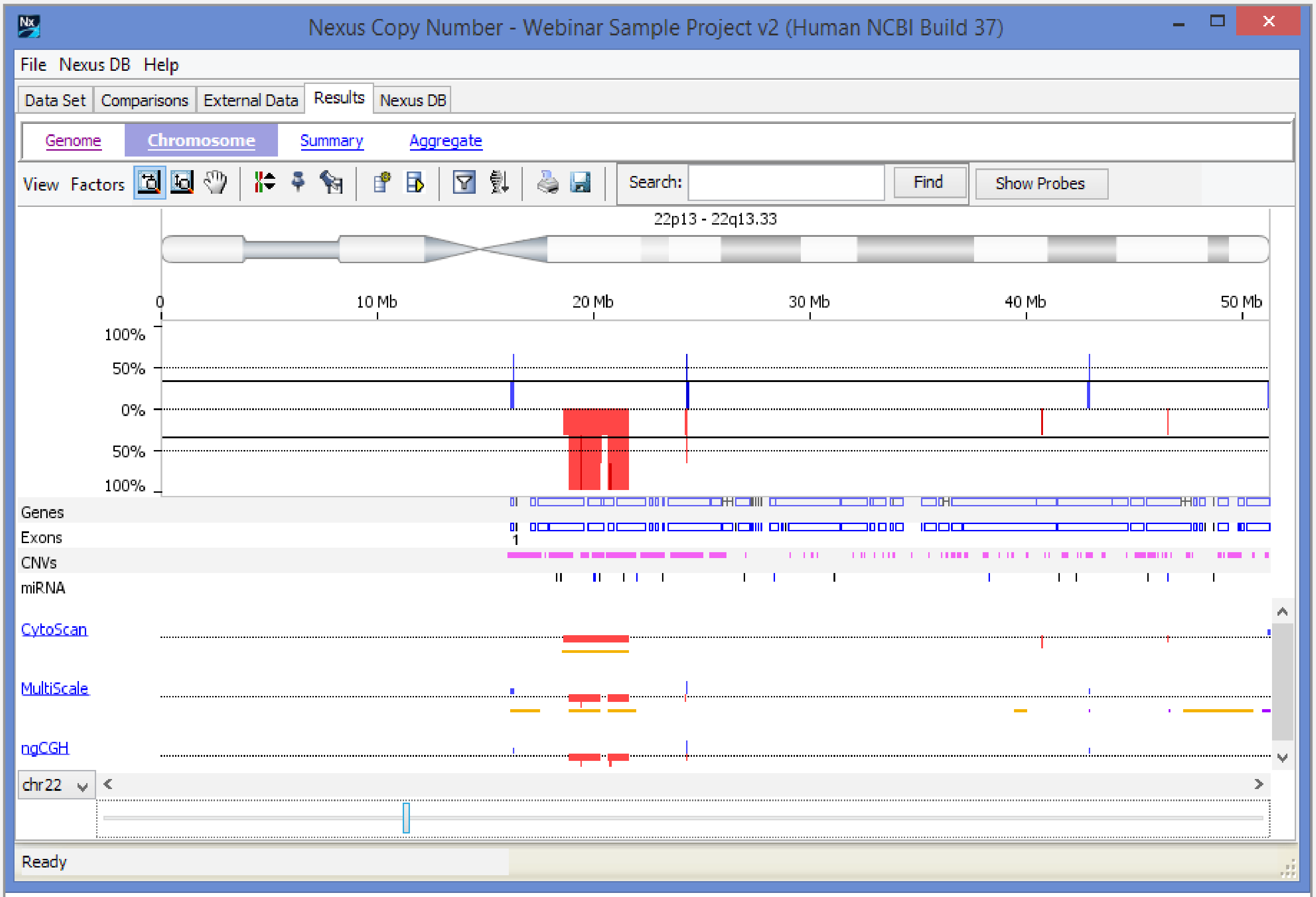

Figure 9. chr 22q11.2 loss across all samples. All samples (CytoScan, BAM multiscale reference, and BAM ngCGH matched) exhibited the loss characteristic of DiGeorge Syndrome.

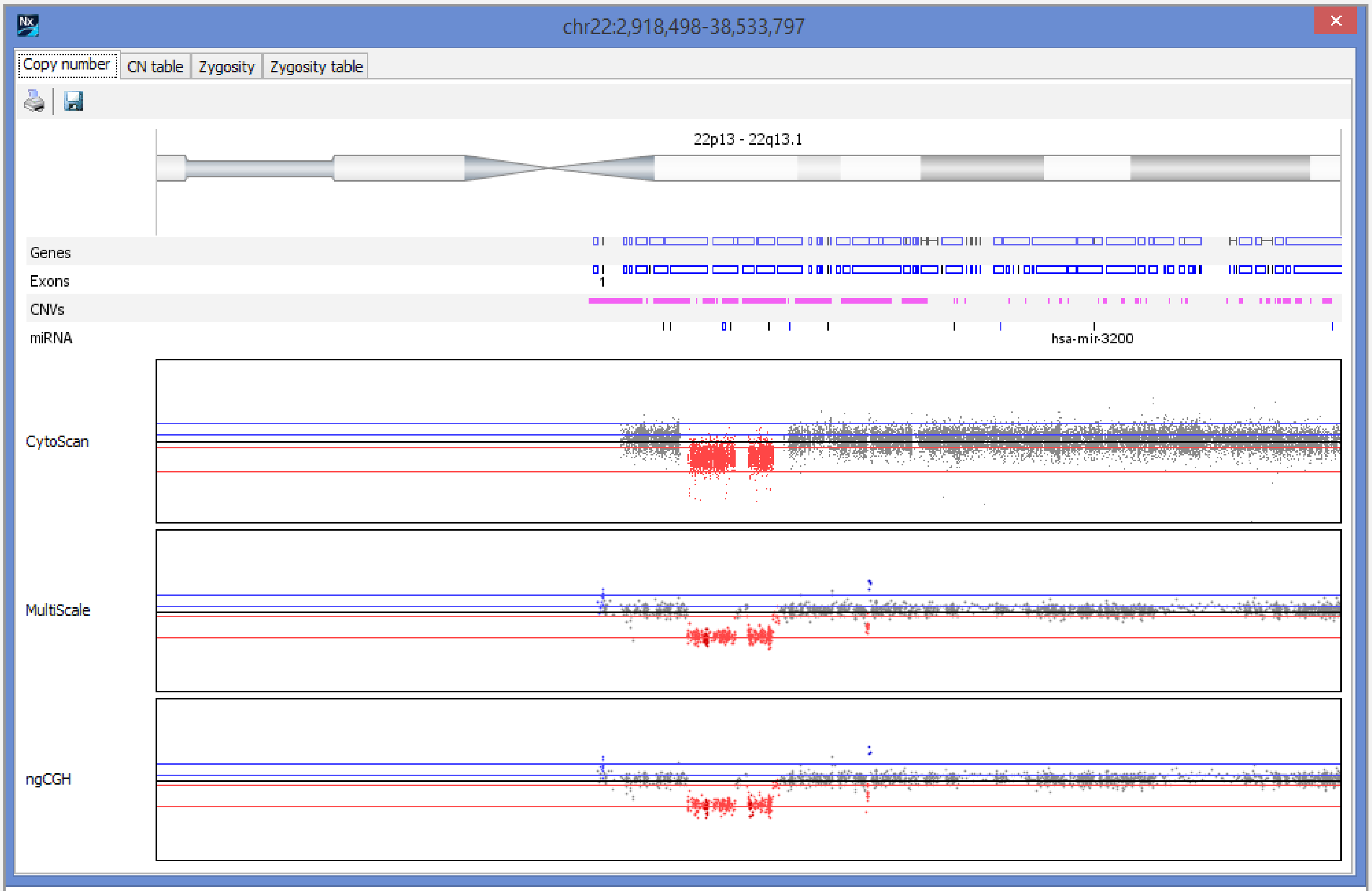

Figure 10. Probes view of chr 22q11.2 loss across all samples. All samples (CytoScan, BAM multiscale reference, and BAM ngCGH matched) exhibited the loss characteristic of DiGeorge Syndrome but the array sample has more probes.

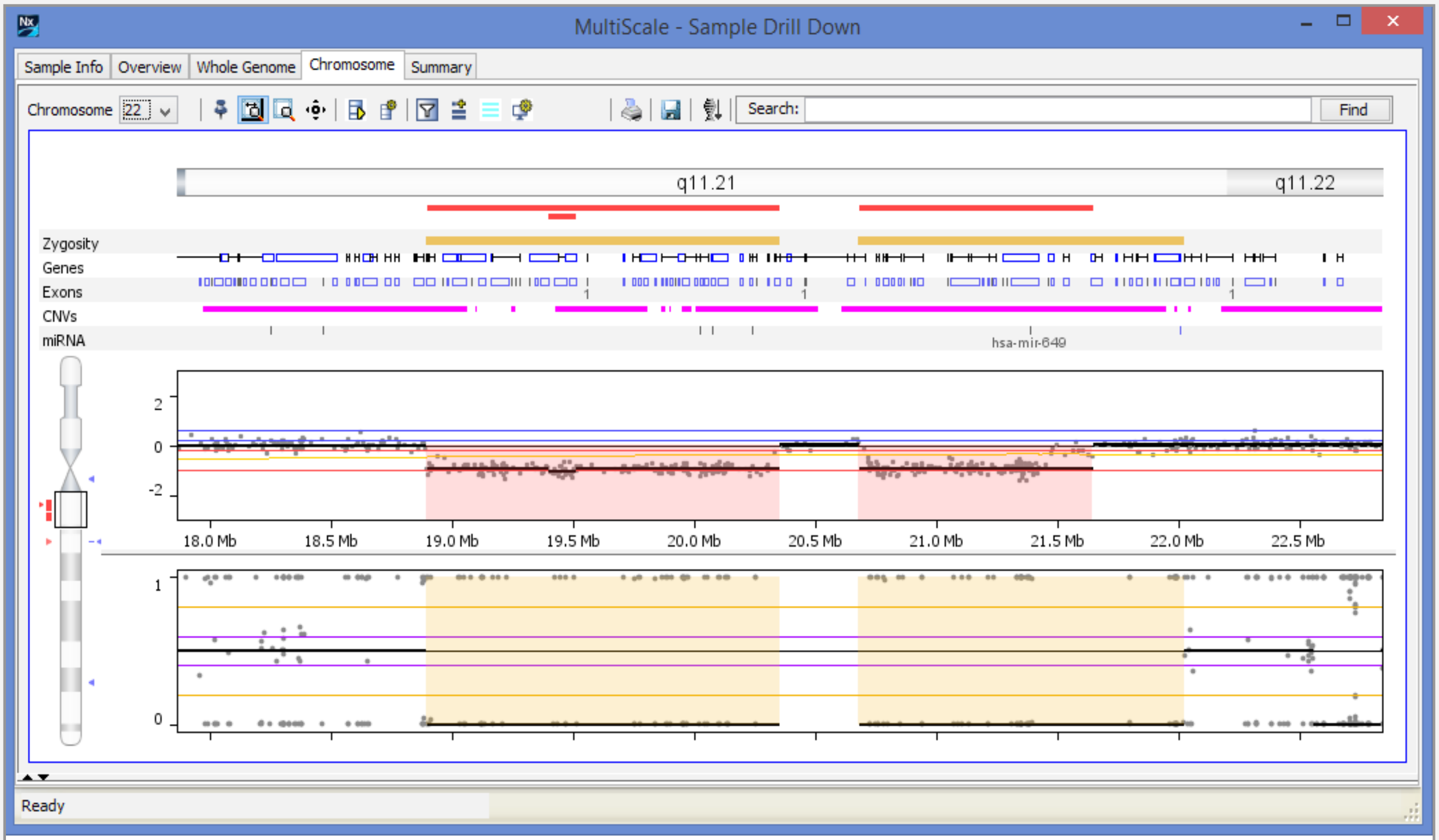

Figure 11. BAM (multiscale reference) DiGeorge Syndrome sample showing genomic changes, log ratio plot and BAF plot. The BAF plot shows only A/B allele probes.

DISCUSSION

SNP arrays have been the gold standard technology for copy number variation estimation but newer sequencing technologies as well as new software developments for CNV estimation from NGS are increasingly providing improved CN results from sequencing platforms. BioDiscovery's new algorithm for obtaining copy number from NGS data, BAM (multiscale reference), is highly reliable for CNV estimation from various types of NGS data. There are several algorithms for CNV estimation from sequencing results but many are not user-friendly and require command line knowledge to run scripts or are designed to work with data from selected NGS platforms. BAM (multiscale reference) algorithm is well suited for all types of NGS data from targeted panels and low pass (shallow) sequencing to WGS/WES with normal depth of coverage.

Allele calling is particularly useful when evaluating somatic tumor results, for confirmation of copy number estimation and baseline ploidy calling, and to determine potential areas of copy neutral loss of heterozygosity. The BAM (multiscale reference) processing is able to combine copy number estimation with allele calling results and therefore is on par with SNP-based arrays. Nexus Copy Number supports visualization and downstream analysis from all of these copy number estimation algorithms, along with the added feature of comparing and simultaneously viewing results from different methods.

REFERENCES

1. CNVkit
   https://github.com/etal/cnvkit

2. ngCGH - Tools for producing pseudo-cgh of next-generation sequencing data.
   https://github.com/seandavi/ngCGH