

Copy Number Algorithm with Built-in GC Waviness Correction in Genotyping Console™ Software

A technical genome-wide artifact known as GC wave or waviness is a common, systematic issue observed with whole-genome assays. GC waviness occurs independently of array platform^{1,2,3,4} and has been observed by researchers using the Illumina HumanHap550 Genotyping BeadChip, array comparative genomic hybridization (CGH), whole-genome tiling arrays, and the Affymetrix® Genome-Wide Human SNP Array 6.0.

In general, the waviness is highly correlated with the GC content of a genomic region. In Genotyping Console™ (GTC) Software 2.0/2.1/3.0, the copy number (CN) algorithm does not account for the GC content. This wavy artifact (waviness) affects \log_2 ratio calculation; therefore, affected samples may have higher false positive rates for predictions of both duplications and deletions.

In GTC 3.0.1/3.0.2, a correction method based on smoothing has been developed to address this issue. This CN algorithm correction reduces or removes the waviness effect across all chromosomes for most of the samples tested in this white paper, thereby reducing the number of false positive CN segments reported. This adjustment to the algorithm does not introduce a higher false negative rate, and the majority of abnormalities in samples are consistently reported independent of the GC waviness correction.

Correlations between \log_2 ratios, smoothed signals, CN segmentation, and regional GC content

Many observed a spatial “wave” pattern in \log_2 ratios that appears to be genome-wide and/or chromosome-specific in some samples. The waves exhibit larger than expected variations of \log_2 ratios across the observed genomic regions. Data from literature and reported cases indicate that regions within the wave have an increased number of segments of copy number change and that these segments are longer than typical CN segments found in normal populations¹. The GC waviness issue is examined using 76 cytogenetic samples from 2 independent laboratories (59 from laboratory No. 1 and 17 samples from laboratory No. 2). The samples from laboratory No. 1 were processed on 4 different days and the MAPD values ranged from 0.24 to 0.45 with a median of 0.32.

Figure 1A shows a summary of the copy number gains and losses found in chromosome 1p36 for these samples, none of which were thought to have cancer. The colored lines in the graph represent the different copy number gains and losses found in the samples color coded by copy number gains (CN = 3 or 4, blue) and losses (CN = 0 or 1, red).

These samples demonstrate an unusually high number of segments of copy number change and many are longer than usually observed in the normal population (i.e., non-cancerous). Another example of this is evident in the \log_2 ratios, smoothed signals, and the hidden Markov model (HMM) CN states of chromosome 4 region 4p16.3-4p16.1 in a sample with GC waviness from laboratory No. 2 (Figure 2A, 2B, 2C). In an ideal situation, if there were no biologically relevant copy number changes, the distribution of raw \log_2 ratios and smoothed signals would be centered on 0, corresponding to a CN state of 2. In regions where the raw \log_2 ratios and smoothed signals

deviate away from zero, the HMM algorithm is more likely to call markers with CN state other than 2, and therefore more segments are found. The sample from laboratory No. 2 in Figure 2A with GC waviness has elevated \log_2 ratios, and correspondingly has more segments of copy number change identified (Figure 2C).

Further examinations of these regions with waviness have demonstrated that this waviness is correlated with the GC content of the region as reported^{1,3}. Figure 1B shows the correlation between \log_2 ratio (y axis, pink) and regional GC content (y axis, blue) for a given sample with GC waviness on chromosome 1. This correlation between GC content and \log_2 ratios has also been reported in several other platforms, such as the Illumina HumanHap550 and Human1M BeadChip³ and array CGH¹. These regions of increased GC content are again correlated with an increased number of CN gains and losses in this sample (Figure 1B).

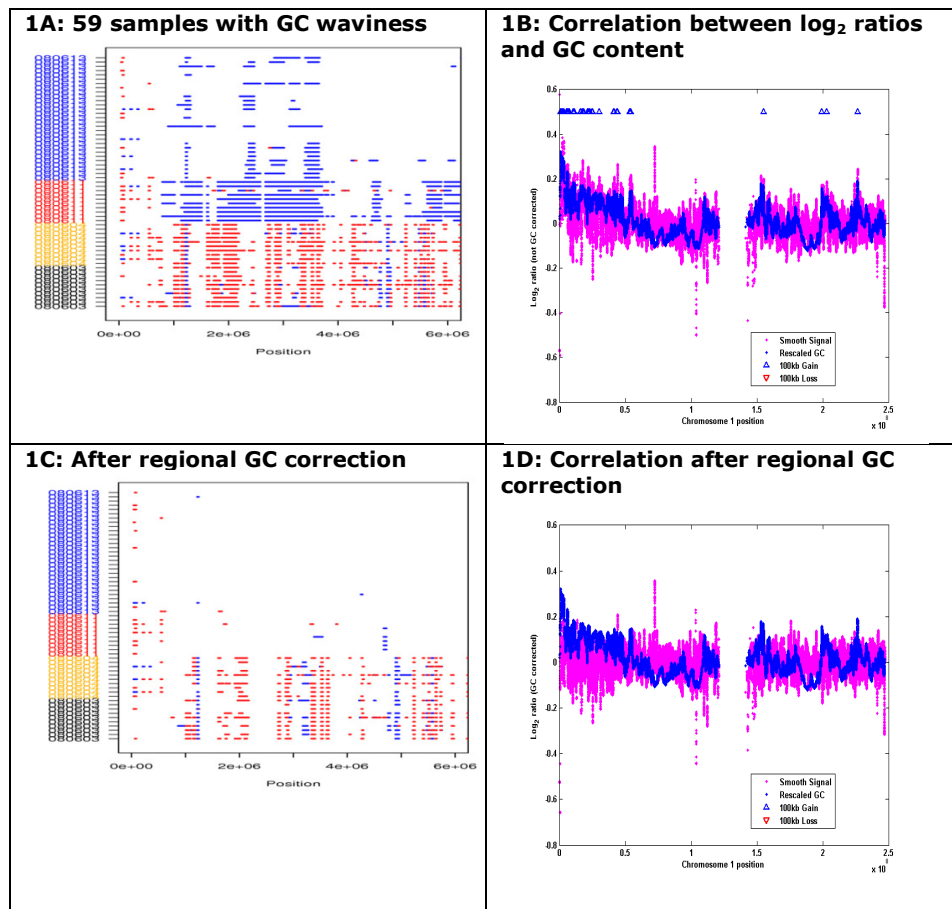


Figure 1: Copy number gains and losses in 59 samples with GC waviness and the same samples after applying GC waviness correction on chromosome 1p36; and the correlations between \log_2 ratios and regional GC content in a sample with MAPD = 0.29 with GC waviness and the correlations after applying GC waviness correction.

1A: CN gains and losses reported in samples with GC waviness. **1C:** CN gains and losses reported in samples after GC waviness correction. Along the y axis, samples are colored by different batches: blue is batch 1, red is batch 2, orange is batch 3, and black is batch 4. For batch 3, the median MAPD values were around 0.35; for batch 4, the median MAPD values were larger than 0.40. The x axis represents genomic position. Blue inside the box represents copy number gains (3 or 4 from the HMM) and red denotes losses (0 or 1 from the HMM). **1B:** The correlation between \log_2 ratios versus regional GC content in one sample with GC waviness. Y axis: \log_2 ratios; x axis: genome positions. Blue line: regional GC content; pink line: smoothed \log_2 ratios. **1D:** The correlation between \log_2 ratios versus regional GC content after the GC waviness correction in one sample. Y axis: \log_2 ratios; x axis: genome positions. Blue line: regional GC content; pink line: smoothed \log_2 ratios. In figure 1B, there are 23 CN gains (blue triangles) and no losses (red triangles) passing 100 kb and 25 marker-filtering criteria in a sample with MAPD = 0.29.

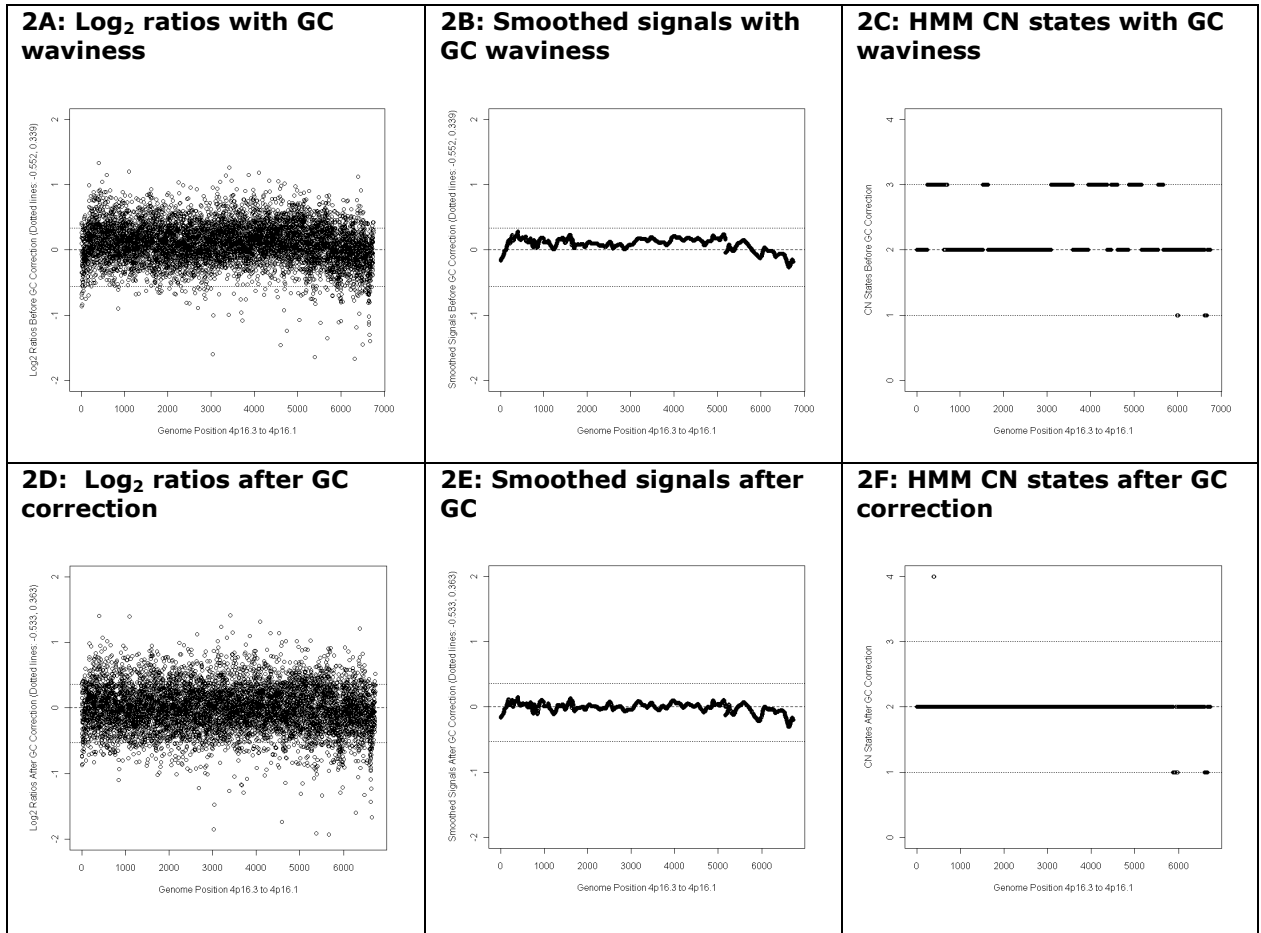


Figure 2: Log₂ ratios, smoothed signals, and HMM CN states from a sample with GC waviness, and the log₂ ratios, smoothed signals, and HMM CN states after applying regional GC correction. 2A and 2D show the log₂ ratios; 2B and 2E show the smoothed signals; 2C and 2F show the HMM CN states for probes on chromosome 4p16.3 – 4p16.1. **2A:** Log₂ ratios with GC waviness. **2B:** Smoothed signals with GC waviness. **2C:** HMM CN states with GC waviness. **2D:** Log₂ ratios after regional GC correction; **2E:** Smoothed signals after regional GC correction. **2F:** CN states after regional GC correction.

Correction methodology

To correct for this waviness caused by the increased GC content of some regions of the genome, a smoothing method was developed. For each sample, the markers were separated into 25 bins based on the equally spaced percentiles of the average GC count in the surrounding 500 kb for that marker. Each bin was further subdivided by their marker type: copy number or SNP, and enzyme fragment (Nsp, Sty, Nsp + Sty). Because there are no CN probes on Sty-only fragments, this results in 5 sub-bins per major bin for a total of 125 bins (5 sub-bins for each of the 25 major bins). For the autosomal markers in each bin, the \log_2 ratios are adjusted so the median \log_2 ratio is zero, and the interquartile ranges (IQR) are equalized across the bins. The \log_2 ratios are then scaled so that the final interquartile ranges (IQRs) match the original IQRs. Next, for each bin on each autosome, the median of the adjusted \log_2 ratios is calculated and then the median for each bin is calculated across the autosomes. This final median of median for each autosome is used as the final adjustment for the \log_2 ratios for all markers, including those on chromosome X and Y (see Appendix A for details on this calculation).

After applying the GC waviness correction to the 59 samples from laboratory No.1, the copy number gains and losses reported by HMM found in the samples are significantly reduced for batch 1 and batch 2 (Figure 1C). Many long copy number gains (in blue) disappeared after GC waviness correction. For batches 3 and 4, quite a few copy number losses (in red) remain after GC waviness correction. For batch 3, the median MAPD values were around 0.35, and for batch 4, the median MAPD values were larger than 0.40.

Among these significantly noisy samples, the GC waviness correction does not seem to completely eliminate false positives. These 59 samples were also processed in Genotyping Console™ Software 2.1 (without regional GC correction) and in GTC 3.0.1 with regional GC correction. The data shows that GC waviness correction significantly reduces number of long CN segments (500 kb + 25 markers) reported in most of the samples (Figure 3). One of the samples with unknown gender does not show significant reduction in the number of CN segments reported, mainly on sex chromosomes X and Y. In figure 1D, all of the gain segments are removed in one of the 59 samples after the GC waviness correction. The GC waviness correction adjusts smoothed \log_2 ratios and makes smoothed \log_2 ratios narrower and centered on zero (Figure 1D).

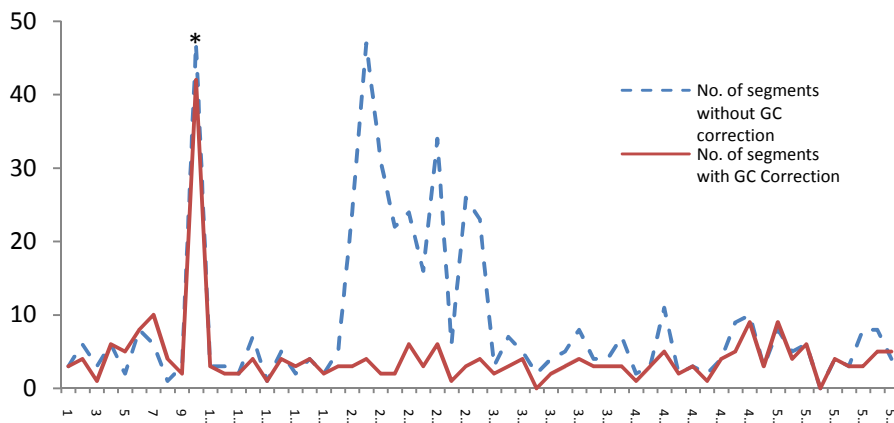


Figure 3: Number of CN segments for 59 samples before and after regional GC correction. X axis: sample 1-59; y axis: number of CN segments ≥ 500 kb and consisted of at least 25 markers. Samples 1, 2, 3, 4, 6, 7, 10, and 16 have high MAPD values (>0.4).

*This sample has unknown gender before and after regional GC correction. MAPD value > 0.40 . Most CN segments reported are located on sex chromosomes X and Y.

However, MAPD value is not an indicator of the existence of GC waviness. Samples with relatively high MAPD values may not show GC waviness, while samples with relatively low MAPD values may show GC waviness. In the sample from laboratory No. 2, the skewed \log_2 ratio distribution is corrected after the GC waviness correction (Figure 2D). The distributions of \log_2 ratios and smoothed signals (Figure 2E) after GC waviness correction are now centered on zero along with a reduction in the number of segments (Figure 2F). Additional examination of other regions and chromosomes in this sample indicates that this correction has reduced the waviness across almost all chromosomes (Figure 4). In Figure 4, chromosome 9 shows a copy number loss. The signal showing this loss is blended with high waviness before GC correction. After GC correction, the waviness is reduced and the loss signal stands out.

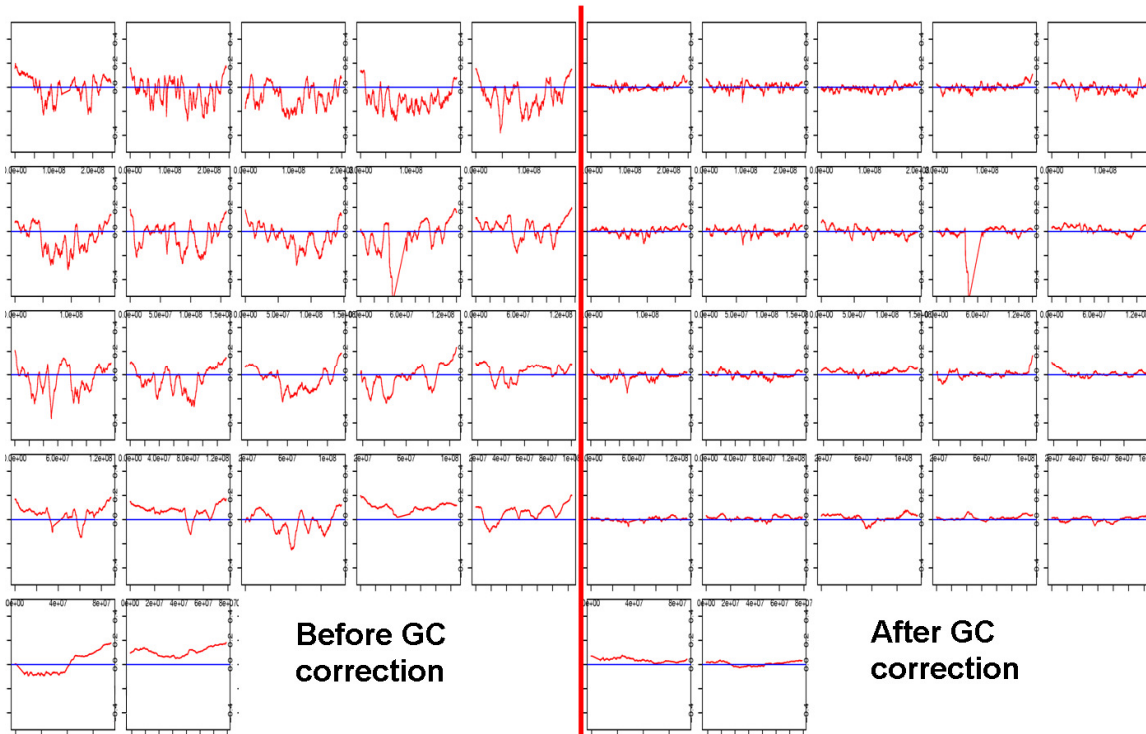


Figure 4: The \log_2 ratios for one sample across all autosomes before and after regional GC correction. Left panel: before regional GC correction; Right panel: after regional GC correction. For each subpanel, the \log_2 ratios are displayed on the y axis and the autosomal position is on the x axis. The autosomes are arranged in order from top left to the bottom right (row 1: chr 1-5, row 2: chr 6-10, row 3: chr 11-15, row 4: chr 16-20, row 5: chr 21-22).

In the 270 HapMap samples processed by Affymetrix, there is minimal GC waviness. In conclusion, the higher than normal \log_2 ratios/CN segments reported in some samples are related to the GC content of the genomic region surrounding the probes. The regional GC correction described above is effective in addressing this issue. After applying the regional GC correction, almost all samples examined in this investigation show improvement with fewer false positive CN segments. Although all samples saw improvement to some degree, not all samples were fixed completely; however, these samples may have other unaddressed issues.

HMM parameter modification with GC waviness correction

The cluster centers for HMM means for each of the copy number states for the SNP Array 6.0 copy number algorithm were based on the median \log_2 values for the chromosome X titration series. The titration series consists of five Coriell DNA samples with chromosome X copy number ranging from one to five, forming a natural titration series. The samples were run in sets of four or five replicates at four different laboratories, totaling 122 microarrays.

Because the GC waviness correction affects the \log_2 ratios, the chromosome X titration series was reexamined after calculating the \log_2 ratios using the GC waviness correction. Figure 5 shows box plots of the \log_2 ratios on chromosome X for the 98 samples that are restricted to X copy number of four or less, before and after regional GC correction. The majority of the data points are very similar, with a slight decrease in dynamic range and slightly narrower box plots after regional GC correction.

The cluster centers for the HMM means, states one to four, for regional GC corrected \log_2 ratios are -0.533, 0, 0.363, and 0.567, which are slightly different than the HMM values used for these four states before GC waviness correction: -0.552, 0, 0.339, 0.543, respectively. The cluster center for CN state 0 is based on the median of the median \log_2 ratios of all chromosome Y markers after GC waviness correction for the females. The \log_2 ratio of each chromosome Y marker is calculated based on an all-male reference, where the reference median \log_2 Y ratio is set to match the HMM mean of -0.533, corresponding to CN state of one after GC waviness correction. In these samples, the median \log_2 ratio was -2 with and without the GC waviness correction, and therefore is used as the HMM mean for CN state 0 on both cases.

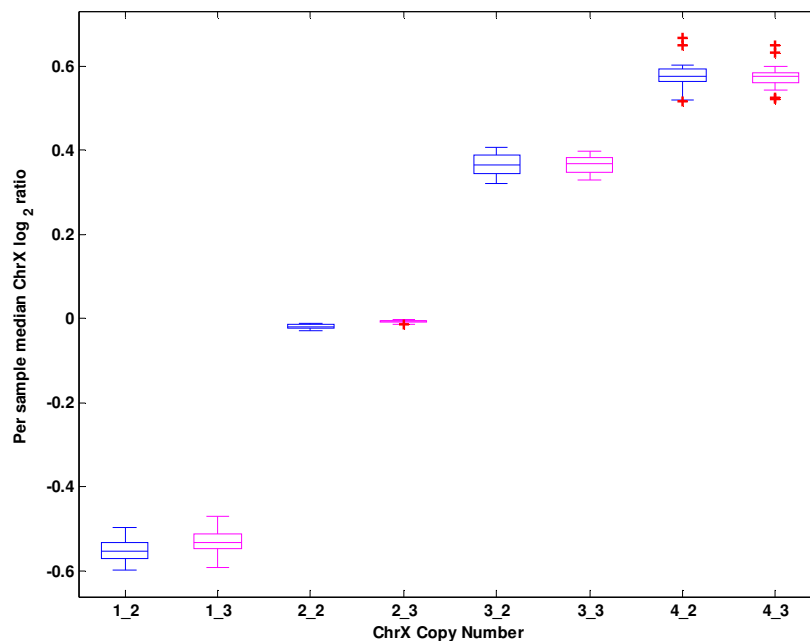


Figure 5: Median \log_2 ratios for chromosome X markers pooled over all samples, excluding the pseudoautosomal regions (PAR) before (blue), and after regional GC correction (pink). Y axis: median chromosome X \log_2 ratios. X axis: CN state (1 through 4); 1(1-4); GC: indicating CN state 1(1-4) with GC waviness correction.

To verify that the regional GC correction does not dramatically affect the signal-to-noise ratio, the relationship between the median \log_2 ratio for each sample and copy number state was examined. As seen in Figure 6, there is a close log-linear relationship between \log_2 ratios and the copy number state. The regional corrected \log_2 ratios have a slightly better ratio, but this is not likely to have a dramatic effect on the data.

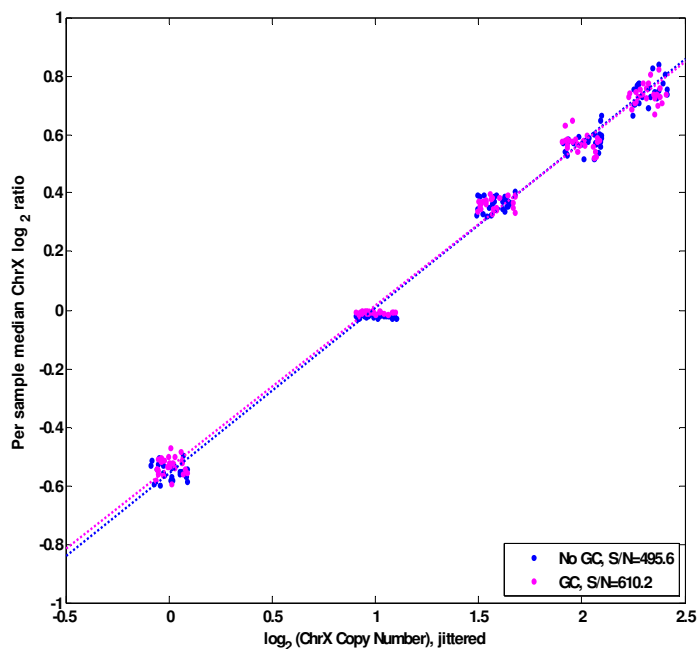


Figure 6: Median \log_2 ratios for chromosome X are plotted against the \log_2 of the CN state for all 98 samples, both before (blue) and after (red) regional GC correction. The CN values for these samples are jittered to see the variation in medians without overlapping the data points. The jittered points are fitted with two lines representing with and without regional GC correction. The estimate of signal range to sample variation ratio (S/N) is estimated by dividing the slope of the fitted line by the estimated standard deviation (SD) of the error (higher is better). Y axis: median chromosome X \log_2 ratios. X axis: jittered \log_2 ratios.

Impact of GC waviness correction on false negative CN segmentation

False negatives, (i.e., regions where $CN \neq 2$, but are called $CN = 2$) are a concern in copy number studies and are not easily identified. Using the chromosome X titration data, it is possible to use copy number calls and segmentation on the X chromosome as an indication of effect of the regional GC correction on the false negative rate. In the absence of natural CN variation on the X chromosome in these samples, an ideal algorithm would see exactly one region (or segment) spanning the entire X chromosome excluding the pseudoautosomal regions (PAR). Even though there is likely some natural CN variation on the X chromosome, an algorithm that predicts fewer segments with CN states different than the expected CN state for the titration is probably more accurate.

Based on this assumption, the false negative rate of the CN algorithm with and without the regional GC correction was compared at the marker and segmentation levels. The proportion of markers in each sample that have the same CN state calls as expected based on the titration information increased when analyzed using the regional GC correction (Figure 7). There were six samples of expected CN state = 3 that were dramatically improved by the regional GC correction. Further examination of the samples that had very high "correct" CN call rates (in excess of 98.5 percent)

indicated that samples with CN = 1 appeared to distribute on both sides of the diagonal, showing no net benefit from the GC waviness correction. However, the higher CN samples mostly clustered above the diagonal, showing a small improvement from the GC waviness correction (Figure 7, right panel). This improvement at the marker level is paralleled when examining the effect at the level of CN segmentation. After applying the regional GC correction, fewer segments that disagree with the expected chromosomal CN for all samples were identified. The most significant improvements were observed in CN = 3 state cases (Figure 7, right panel). This suggests that the GC waviness correction more accurately reports chromosomal CN deviations and hence has a lower false negative rate.

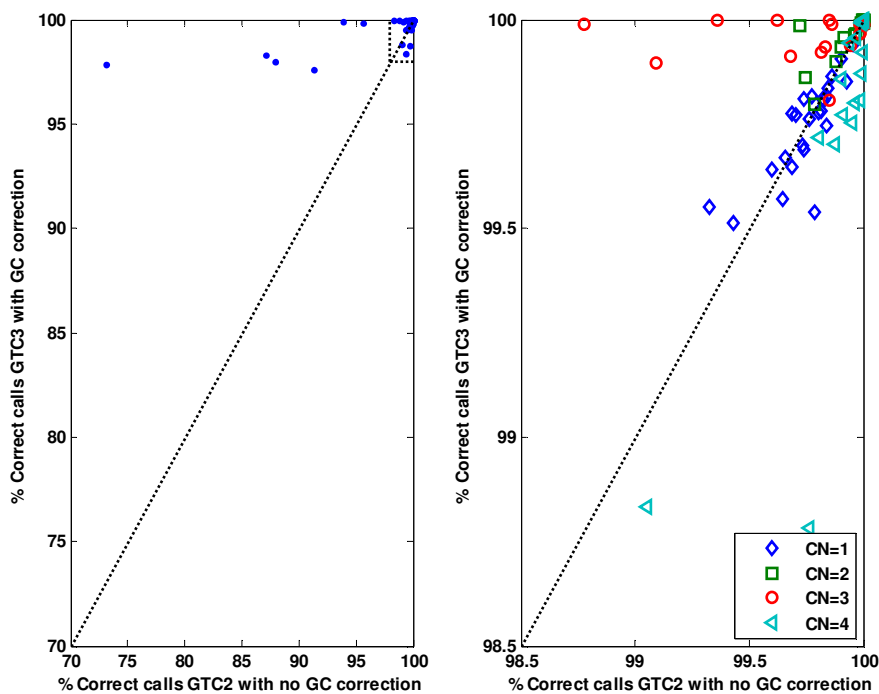


Figure 7: The regional GC correction increases the number of CN state calls that match the expected value based on the chromosome X titration. The percentage of calls matching the titration value without regional GC correction is shown on the x axis, and the percentage using regional GC correction is shown on the y axis. Samples above the diagonal have a higher proportion of "correct" CN values after the regional GC correction; samples below the diagonal have a higher proportion of "correct" CN values with no GC correction. The right panel shows a higher magnification view of the outlined area on the right. The color and shape of the points corresponds to the copy number state as identified in the legend.

The following figures show box plots of the decrease in the number of segments per microarray given the X chromosomal CN after the GC waviness correction is applied (Figure 8). Positive values indicate a decrease, i.e., likely higher accuracy. The box plots show that for all CN possibilities, there is similar median accuracy. In almost all cases a substantial minority of cases improve in accuracy. CN = 3 replicates seem to improve across the board.

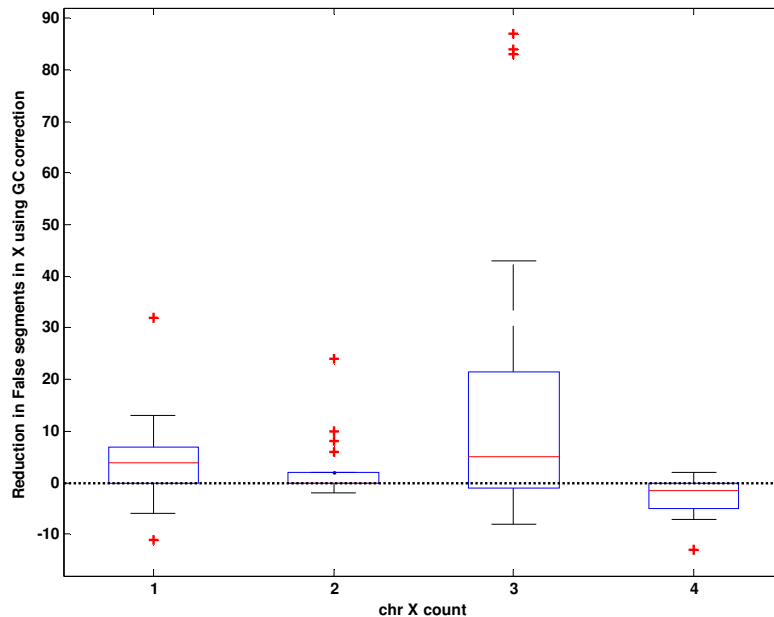


Figure 8: Box plots of the difference in number of segments detected before and after regional GC correction for the chromosome X titration data. The number of segments identified after regional GC correction was subtracted from the number identified without regional GC correction; thus, positive values indicate a decrease and suggest higher accuracy.

In conclusion, the regional GC correction reduces the number of false negatives, especially in samples with CN > 2, while also reducing the number false positives, mostly in samples with CN = 2.

Impact of GC waviness correction on CN segmentation in samples with abnormalities

To ensure that the regional GC correction did not diminish the copy number algorithm's ability to detect real CN changes, a data set consisting of samples from four cell lines exhibiting different known cytogenetic abnormalities was analyzed. The samples were processed at 4 separate laboratories for a total of 16 microarrays. The abnormalities associated with the samples were the sister syndromes Angelman/Prader-Willi (trisomy 13), Duchenne muscular dystrophy (DMD, del Xp21.1), and Smith-Magenis (del 17p11.2). The smallest of these abnormalities is the DMD-del Xp21.1, with a deletion to CN = 0 estimated at around 662 kb in this sample, or about 560 markers. As shown in Figures 1 through 12, the known cytogenetic abnormality is clearly visible with and without the regional GC correction.

The CN calls for the region associated with the Prader-Willi/Angelman abnormality can be seen in Figure 9. Examination reveals that the region of CN change in both cases is almost exactly the same with and without the correction. There might be a false positive region in laboratory 1 near the 2.625×10^7 genome position because it is not found by other laboratories. In laboratory 4,

there is a region broken into one large region and one small region near 2.12×10^7 genome position before GC correction. After GC correction, this breakpoint disappears and only one longer region is reported, the same as the data reported by the other three laboratories. The two small regions near 2.1×10^7 genome position reported in laboratories 1 and 2 might be due to the change of CN state of markers located in the breakpoints.

In the sample with the Smith-Magenis abnormality, the affected region is again identified with and without the regional GC correction. A region only found laboratory 1 but not in other laboratories is probably due to the change of CN state of markers located in the breakpoint near the 1.82×10^7 genome position. This might be true for other regions within 1.83 to 1.84×10^7 genome positions (Figure 10). For the sample with the DMD abnormality, again, all four labs were able to identify the region with and without the regional GC correction (Figure 11).

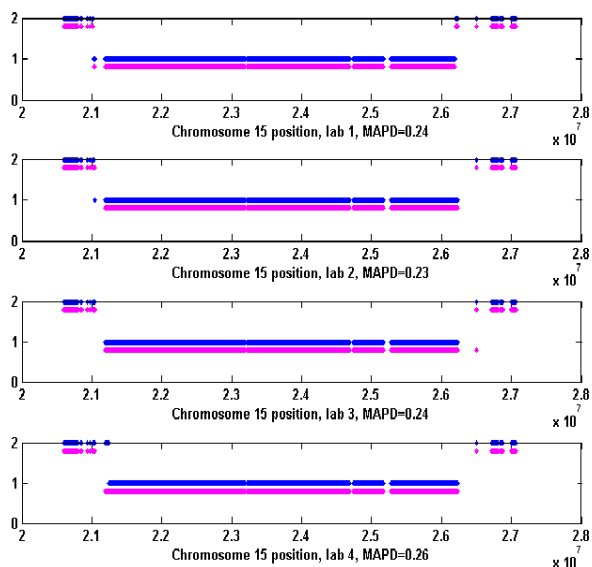


Figure 9: Copy number state calls for chromosome 15 for the samples run at different labs with the Prader-Willi/Angelman abnormality before and after regional GC correction. The chromosome position is displayed along the x axis and the copy number state is on the y axis. The CN state calls are shown in blue for the analysis without the regional GC correction, and in pink for those with regional GC correction.

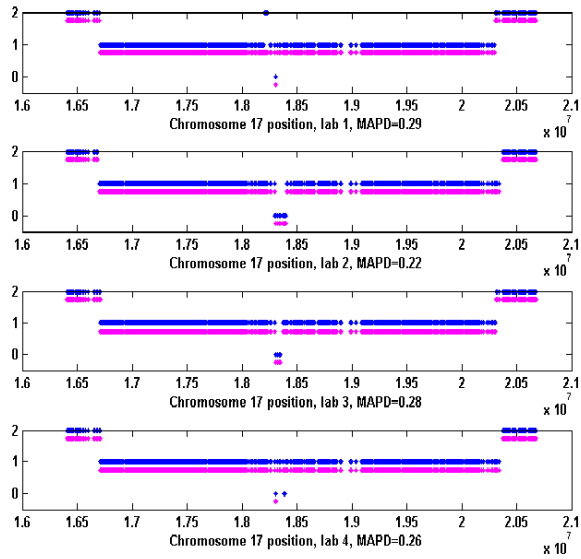


Figure 10: Copy number state calls for chromosome 17 for the samples run at different labs with the Smith-Magenis abnormality before and after regional GC correction. The chromosome position is displayed along the x axis and the copy number state is on the y axis. The CN state calls are shown in blue for the analysis without the regional GC correction, and in pink for those with regional GC correction.

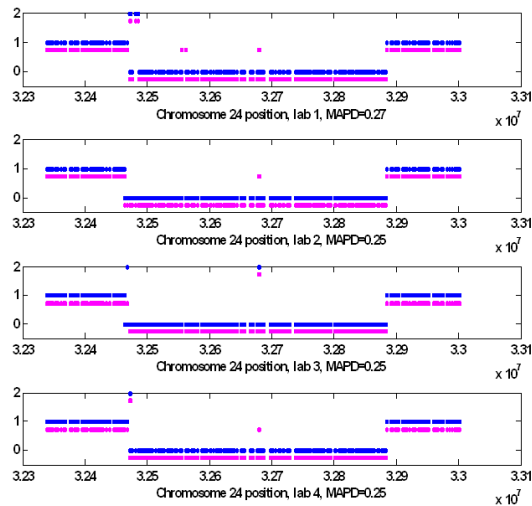


Figure 11: Duchenne muscular dystrophy (DMD) abnormality in samples run at different labs before and after regional GC correction. The chromosome position is displayed along the x axis and the copy number state is on the y axis. For each lab, the blue line indicates the CN state calls without regional GC correction. The pink line for each pair indicates the CN state calls with regional GC correction. Chromosome 24 represents chromosome X.

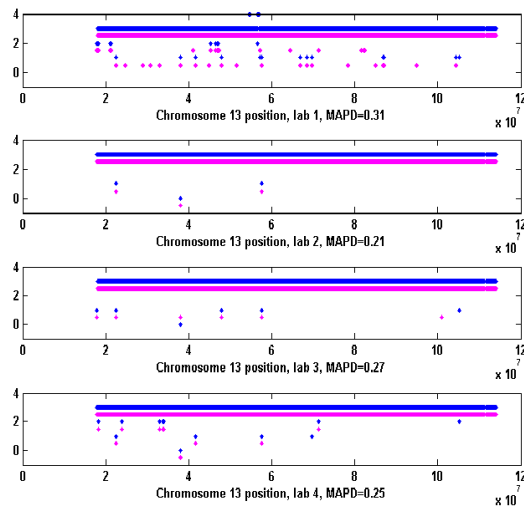


Figure 12: Chromosome 13 trisomy abnormality in samples run at different labs before and after regional GC correction. The chromosome position is displayed along the x axis and the copy number state is on the y axis. For each lab, the blue line indicates the CN state calls without regional GC correction. The pink line for each pair indicates the CN state calls with regional GC correction.

In samples with chromosome 13 trisomy, the affected region was identified with and without the regional GC correction (Figure 12). Laboratories 2, 3, and 4 have fewer regions with CN < 3. The data from laboratory 1 is noisier and shows a slight increase in the number of regions with CN < 3 with regional GC correction, but the abnormality is still easily identified.

In conclusion, the major CN abnormalities in the abnormal samples were consistently reported before and after applying the regional GC correction.

Normalization and signal summarization with GC waviness correction

In Genotyping Console™ Software 3.0.1/3.0.2, the normalization and signal summarization of the CN algorithm is done before the GC waviness correction. Due to the CN algorithm transition from MatLab to Affymetrix Power Tools (APT) in GTC 3.0.1/3.0.2, the normalization and signal summarization is done slightly differently compare to those in GTC 2.0/2.1. In GTC 2.0/2.1, the algorithms for copy number (CN) estimation on the Genome-Wide SNP Array 6.0 were implemented in a compiled MatLab executable. To provide a robust implementation of the algorithms and to ensure that the same algorithm is available to both command-line and GUI users, these algorithms were ported to APT.

The transition from MatLab to APT led to some minor differences in the normalization and signal summarization algorithms with minimal impact on CN analysis results. The proportion of copy number markers that disagree between GTC 2.0/2.1 and GTC 3.0.1/3.0.2 is typically around 1 in 10,000, based on an example data set of 50 cytogenetic samples. Both GTC 2.0/2.1 and GTC 3.0.1/3.02 use adapter-type normalization⁵ followed by quantile normalization. The difference between GTC 2.0/2.1 and GTC 3.0.1/3.0.2 quantile normalization is that GTC 2.0/2.1 excludes the oligo-B1 and oligo-B2 probes used for gridding, whereas GTC 3.0.1/3.0.2 normalizes using all probes on the array.

The oligo-B1 and oligo-B2 probes comprise 3.97 percent of the probes on the array and will have a minimal impact on the overall signal estimates. Following normalization, signal summaries for the probe sets are generated. The CN probe sets consist of only one probe per probe set and no summarization is needed. The SNP probe sets consist of more than one probe per probe set and require the summarization step.

Briefly, in GTC 2.0/2.1, the intensities for all of the probes in the probe sets for each allele are logged and summarized for each allele using median polish. Then the exponentiated result from each allele is summed to produce the final signal for the SNP probe set. In GTC 3.0.1/3.0.2, signal summarization of SNPs using PLIER⁶ with non-standard options is applied to perfect-match (PM) probes (as there are no mismatch [MM] probes on the SNP Array 6.0), which makes it behave very much like a median polish.

In addition, PLIER has been extensively tested in combination with the BRLMM-P+⁷ genotyping algorithm in APT. PLIER does not require the logarithmic transformation of the data, so the log transformation and exponentiation steps are skipped. After the signal summarization, GTC 2.0/2.1 and GTC 3.0.1/3.0.2 generate log₂ ratios for each probe set by dividing the signal for the probe set by the median signal for the probe set across a reference set of samples. Post-processing median-autosome corrections to log₂ ratios (as described in the GTC 2.1 manual) are the same in both GTC 2.0/2.1 and GTC 3.0.1/3.0.2.

Allelic difference calculation with GC waviness correction

In general, allelic difference⁵ is calculated by subtracting the signal of one allele from the other and normalizing this by dividing the difference by the median difference for the same genotype calls in the reference set. Because regional GC content affects the signals, the allelic difference calculation also needs to be adjusted for this effect. The allelic difference for samples analyzed with the regional GC correction is calculated as follows.

For a given SNP probe set i , its allele probe sets are labeled A and B. The allelic difference adjustment is calculated by first taking the difference in probe set A summary signal and B summary signal (both unlogged).

$$D_j = S_{A,j} - S_{B,j}$$

This difference is then adjusted similar to the log₂ ratios described above. Again, the probe sets are segregated into 125 bins based on GC content of the surrounding 500 kb and enzyme fragment. Next, the difference is adjusted by dividing the anti-log₂ of the factor used to adjust the median log₂ ratios to zero and the IQR adjustments followed by dividing by the anti-log₂ of the median of the median for the log₂ ratios for each individual autosome (see Appendix B).

This adjustment assumes that the GC effect is a multiplicative effect that will affect the A signal equally as the B signal, and that the per-bin adjustment X to the log₂ ratios R arises as a global log additive effect equally applicable to the original summary signal. Under these assumptions, this is equivalent to dividing the (signal A - signal B) by 2^X . Similarly, if M = median-autosome-median value, used to adjust the log₂ ratios of the summary CN signals, is assumed to arise from a global multiplicative shift in the sample signal, it is appropriate to use this to adjust the allelic difference by dividing by 2^M . Both of these adjustments are done before using the reference to standardize the difference; by construction GC waviness is relative to the reference and hence has no GC waviness.

There is one more difference in the allelic difference calculation compared to those in GTC 2.0/2.1. In reference generation, GTC 2.0/2.1 used Birdseed⁸ whereas GTC 3.0.1/3.0.2 uses the existing APT implementation of BRLMM-P+. This difference means the median differences stored in the reference are slightly different between GTC 2.0/2.1 and GTC 3.0.1/3.0.2. The reason to switch to BRLMM-P+ is to harmonize the genotype calling between reference calling workflow and single sample workflow.

The GC waviness correction does not impact the LOH analysis because the adjusted \log_2 ratios are not used by the LOH analysis algorithm. To see the difference of LOH between GTC 2.0/2.1 and GTC 3.0/3.0.1/3.0.2, please refer to the LOH white paper, available at www.affymetrix.com.

Appendix A: Details for GC correction algorithm

A smoothing method has been developed to decrease variability of \log_2 ratios in regions of high regional GC content. The method is applied to initial \log_2 ratios for SNP markers (also called probe sets) in a particular array. This initial \log_2 ratio for each SNP marker is calculated after quantile normalizing and summarizing probe intensities into SNP marker signal.

For the CN probes, summary signal is the normalized intensity only. For each sample, markers are divided into 25 bins based on the equally spaced percentiles of the average GC count (GC content) in the upstream/downstream 250 kb for a particular marker (500 kb total). Within each of the 25 bins, the markers are subdivided based on their type: CN/SNP marker type, enzyme fragment type (Nsp, Sty, Nsp + Sty), which gives five sub-bins per major bin, as there are no CN probes in Sty-only fragments, for a total of $5 \times 25 = 125$ bins. For the autosomal markers in each bin, the median \log_2 ratio of each bin is adjusted to zero and IQRs are equalized across all the bins. For example, if summary marker j \log_2 ratio in Bin i is represented by $R_{i,j}$ then given Bin i , adjustment is:

$$X_i = \text{median}_{\{j \in \text{Bin}_i\}}(R_{i,j})$$

Next, the interquartile range (IQR) of the \log_2 ratios with bin median is subtracted:

$$IQR_i = \text{IRQ}_{\{j \in \text{Bin}_i\}}(R_{i,j} - X_i)$$

An initial adjustment of given by the following formula is formed:

$$\hat{R}_{i,j} = (R_{i,j} - X_i)/IQR_i$$

The IQRs of all the adjusted \log_2 ratios (including the X and Y chromosomes) are multiplied by a factor that makes the IQRs of the adjusted \log_2 ratios equal to the IQRs of the original \log_2 ratios. If we use the notation to represent possible values of a subscript i or j , after this step, the \log_2 ratios are:

$$\hat{R}_{i,j}^* = \hat{R}_{i,j} * \text{IQR}(R_{.,.})/\text{IQR}(\hat{R}_{.,.})$$

Finally, the \log_2 ratios of all markers (including X and Y markers) in a bin are adjusted using the median of the medians of the \log_2 ratios for each autosome.

$$M = \text{median}_{\{\text{autosomes } i=1,2,\dots,22\}}(\text{median}(\hat{R}_{i,.}^*))$$

Thus, the final adjusted \log_2 ratio is:

$$\hat{R}_{i,j}^{**} = \hat{R}_{i,j}^* - M$$

Appendix B: Allelic difference calculation adjustment

For a given SNP, j , let its allele probe sets be A and B. The allelic difference adjustment is calculated by first taking the difference in probe set A summary signal and B summary signal (both unlogged).

$$D_j = S_{A,j} - S_{B,j}$$

Depending on which of the above 125 bins the SNP is in, this difference is adjusted by first dividing the antilog_2 of the adjustment calculated in the GC waviness algorithm described in Appendix A, and then dividing that by the antilog_2 of M = median-autosome-median value, also described in Appendix A. So, given Bin i , adjustment is applied to the initial allelic difference as:

$$\hat{D}_{j,i} = (S_{A,j} - S_{B,j}) \cdot 2^{-X_i}$$

Followed by the global correction:

$$\hat{D}_{j,i} = (S_{A,j} - S_{B,j}) \cdot 2^{-X_i} \cdot 2^{-M}$$

While calculating the reference, these differences in the allele summaries are grouped by an estimate of whether they are AA, AB, or BB genotypes, and the median difference across all samples in the reference given AA, given AB, and given BB is stored in the reference. Let these three values for a given SNP be denoted D_{AA} , D_{AB} , D_{BB} (they are different for every SNP). Assume $D_{AA} > D_{AB} > D_{BB}$ (if it's the other way round, reverse the calculation below). The final allele difference is calculated by taking the adjusted difference D and finding if it is $> D_{AB}$, in which case we calculate the final value as:

$$\hat{D}^* = (\hat{D} - D_{AB}) / |D_{AA} - D_{AB}|$$

or if it is $< D_{AB}$, in which case we calculate the final value as:

$$\hat{D}^* = (\hat{D} - D_{AB}) / |D_{BB} - D_{AB}|$$

References

- ¹Marioni J. C., *et al.* Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology* **8**(10):R228 (2007).
- ²Fridlyand J., Snijders A. M., Pinkel D., Albertson D. G., Jain A. N. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**:132-153 (2004).
- ³Diskin S. J., *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research* **36**:e126 (November 2008).
- ⁴Komura D., *et al.* Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Research* **16**:1575-1584 (2006).
- ⁵Affymetrix, Inc. User manual: Genotyping Console™ Software 2.1 (2008).
- ⁶Affymetrix, Inc. Technical Note: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation (2005).
- ⁷Affymetrix, Inc. White Paper: BRLMM-P: A Genotype Calling Method for the SNP Array 5.0 (2007).
- ⁸<http://www.broad.mit.edu/mpg/birdsuite/birdseed.html>