The Journal of Molecular Diagnostics, Vol. 14, No. 6, November 2012

http://dx.doi.org/10.1016/j.jmoldx.2012.06.002

Technical Advance

A GC-Wave Correction Algorithm that Improves the Analytical Performance of aCGH

Angela Leo,* Andrew M. Walker,*
Matthew S. Lebo,*† Brant Hendrickson,*
Thomas Scholl,* and Viatcheslav R. Akmaev*

From Research and Development, Esoterix Genetic Laboratories LLC, Westborough; and the Harvard Medical School Genetics Training Program,† Partners Healthcare Center for Personalized Medicine, Boston, Massachusetts*

Array-based comparative genome hybridization (aCGH) is a powerful, data-intensive technique used to identify genomic copy number variation throughout the human genome. The use of aCGH clinically to identify pathogenic copy number aberrations is becoming common, and the statistical and mathematical algorithms used in aCGH data analysis play an important role in determining the performance of these platforms. Interpretation of aCGH data can be complicated by a platform-independent technical artifact described as GC-waves, which are wave patterns in CGH data correlating to regional GC-content of the human genome that can reduce the clinical specificity and sensitivity of aCGH platforms. We describe an automated GC-wave correction algorithm and techniques to understand how the correction affects the analytical performance of aCGH. This GC-correction algorithm was effective at mitigating GC-wave effects. After correction, array data were measurably improved by the algorithm, demonstrating improvements in specificity and sensitivity and in overall data quality. ***(J Mol Diagn 2012, 14:550–559; http://dx.doi.org/10.1016/j.jmoldx.2012.06.002)***

Genomic copy number alterations are the causal lesions for many known genetic diseases and syndromes. The application of array-based comparative genome hybridization (aCGH) as a high-resolution method for detecting genomic copy number changes has significantly advanced the understanding of the human genome, disease, and genetic variation.[1] High-density CGH arrays, which can contain millions of probes selectively spaced throughout the genome, offer increased resolution and sensitivity over cytogenetic methods, such as fluorescence *in situ* hybridization (FISH) and G-banded karyotyping. These proven technologies are known for their reliability in detecting clinically relevant genomic imbalances; however, they have limitations. Current G-banded karyotyping protocols are limited by a detection resolution of 3 to 5 Mb. FISH can assess DNA copy number only in specific targeted loci and also has resolution limited to approximately 50 kb, depending on many factors, including genomic location.[2,3] aCGH combines the genome-wide detection capabilities of G-banded karyotyping with higher resolution than is possible with FISH.

Large studies of patient and control populations using aCGH have led to the discovery of many disease-causing genes and genomic regions.[4,5] In several studies, aCGH demonstrated increased clinical utility over cytogenetic technologies for patients with mental retardation or developmental delays by finding structural aberrations in samples where karyotyping and FISH had previously yielded negative results.[1,6–10] The increased yield of copy number variants detected by aCGH has led to the clinical uptake of the technology; it is commonly used in postnatal diagnostics and increasingly in prenatal settings for certain clinical indications. It has been recommended as a first-line test in individuals with developmental delay, autism spectrum disorder, or multiple congenital anomalies.[11] Expanded use of aCGH has also led to the discovery of polymorphic copy number variable (CNV) regions throughout the human genome.[12] A variety of publicly available databases cataloguing CNVs have been developed (eg, the Database of Genomic Variants).[13] Since aCGH testing identifies CNVs of all sizes throughout the genome, rather than only large aberrations (G-banded karyotyping), or at well-characterized,

Supported by Esoterix Genetic Laboratories LLC.

Accepted for publication June 6, 2012.

Disclosure: All the authors received salary compensation and/or stock from Esoterix Genetic Laboratories LLC.

Supplemental material for this article can be found at *http://jmd.amjpathol.org* or at *http://dx.doi.org/10.1016/j.jmoldx.2012.06.002*.

Address reprint requests to Angela Leo, M.S., Esoterix Genetic Laboratories LLC, 3400 Computer Dr, Westborough, MA 01581. E-mail: leoa@labcorp.com.

pathogenic loci (FISH), databases of previously discovered CNV loci assist in the interpretation of aCGH results.

High-resolution CGH arrays differ from other CGH arrays in that short oligonucleotides rather than longer bacterial artificial chromosomes (BACs) are used for hybridization. The use of shorter but more numerous probes enables the identification of CNVs at greater resolution than can be achieved on BAC arrays. However, each individual oligonucleotide probe yields lower-quality data than a BAC probe owing to the lower specificity and higher potential for noise inherent in probes at this length (25 to 75 bases). Although a single BAC probe is often sufficient to call an aberration, the high-resolution oligo-based arrays need several adjacent probes to confidently identify CNV regions. Thus, oligonucleotide arrays require statistical algorithms to extract results. The aCGH data analysis process uses a sequence of data transformation, data normalization, and data summarization steps. Each of these steps typically involves the definition of algorithm parameters that directly affect the analytical sensitivity and specificity of the aCGH assay. The analysis of array-based copy number data can be complicated by the presence of a common, platform- and method-independent[14–19] artifact that has been described recently as GC-waves.[14] The GC-wave phenomenon is the correlation of the deviation in magnitude and direction of log-ratio values from an expected two-copy baseline with genomic GC-content in aCGH and single-nucleotide polymorphism array-based copy number data. Because there are often several consecutive probes in GC- and AT-rich regions of the genome, the moving average through the probes appears as a wavelike pattern. GC-waves add large-scale variability to the probe signal[14] and interfere with data analysis algorithms as they skew probe data away from expected values. The GC-wave artifact increases the potential for false-positive aberration calls in specific genomic regions and can obscure true aberration calls. Investigation into possible causes of the wave effect have focused on biochemical causes, such as DNA quality,[17,20] DNA isolation protocols, and labeling procedures,[20] as well as on genomic features, such as regional GC and gene content.[17] The causes of GC-waves are not completely understood and are probably multifaceted.

A few computational methods for addressing GC-waves have been published. Marioni et al[14] concluded that the wave effect strongly correlates with the GC-content of the probe, and they developed a correction method based on lowess regression to improve CNV calling accuracy for small regions. However, the method is not applicable in the presence of larger aberrations. Large aberrations are often seen in cases of developmental delay, autism spectrum disorder, and multiple congenital anomalies. Nannya et al[19] considered the GC-content of the DNA fragments hybridizing to the array and their size when developing a quadratic regression method for single-nucleotide polymorphism arrays (Affymetrix Inc., Santa Clara, CA). Alternatively, van de Wiel et al[20] created a set of calibration profiles from a subset of previous aCGH results to reduce the GC-waves in data from tumor samples based on ridge regression. Each of these methods is effective in reducing GC-wave patterns in some capacity, but these approaches generally require some previous understanding of expected array results and can lead to a loss of aberration detection sensitivity.[14,17,20] In the clinical setting, it is preferable for a GC-correction algorithm to perform across a broad range of aberration sizes without previous information about the location of aberrations.

We developed a two-part GC-wave correction algorithm, CGH slope and anchored median (cghSAM), which is designed to work with all current array-based copy number platforms. The algorithm normalizes signal log-ratios according to individual probe GC-content and then adjusts a subset of chromosomes based on the observed median signal deviations. The two-step correction accounts for the effect of GC-content on each probe by using a chromosome-specific regression and then targets the chromosomes that are most susceptible to GC bias for further correction.

To evaluate the performance of cghSAM, microarray data generated from 215 human blood samples were analyzed using the correction. We quantified improvements in the overall array performance by analyzing probe distribution before and after correction with cghSAM. The reduction in wave amplitude increased the sensitivity of the assay in the detection of small deletions, as measured using common copy number polymorphic loci. Analytical specificity was simultaneously increased, improving overall CNV calling accuracy.

Materials and Methods

Study Participants

All the data were derived from patient specimens referred to Esoterix Genetic Laboratories LLC (Westborough, MA) for clinical aCGH analysis and were made fully anonymous before the study. This research study was limited to the use of existing data that were made fully anonymous and is exempt under 45 CFR 46.101(b)(4) as defined by the Office of Human Research Protections, US Department of Health and Human Services.

Custom Oligonucleotide Array

A custom CGH array (four arrays per 1 × 3-inch slide, each array containing 44,000 oligonucleotide probes) (Agilent Technologies Inc., Santa Clara, CA) was designed to target 140 known disease-causing regions of the human genome in addition to the subtelomeric and pericentromeric regions of each chromosome. Approximately 500,000 probes were selected from Agilent Technologies Inc.'s high-definition CGH probe database or, in genomic regions where Agilent Technologies Inc. had not designed probes owing to base composition, were designed by Esoterix Genetic Laboratories LLC and were tested for performance in euploid and aneuploid specimens. Empirical data were generated on a sufficient probe pool to enable 10X down sampling of the

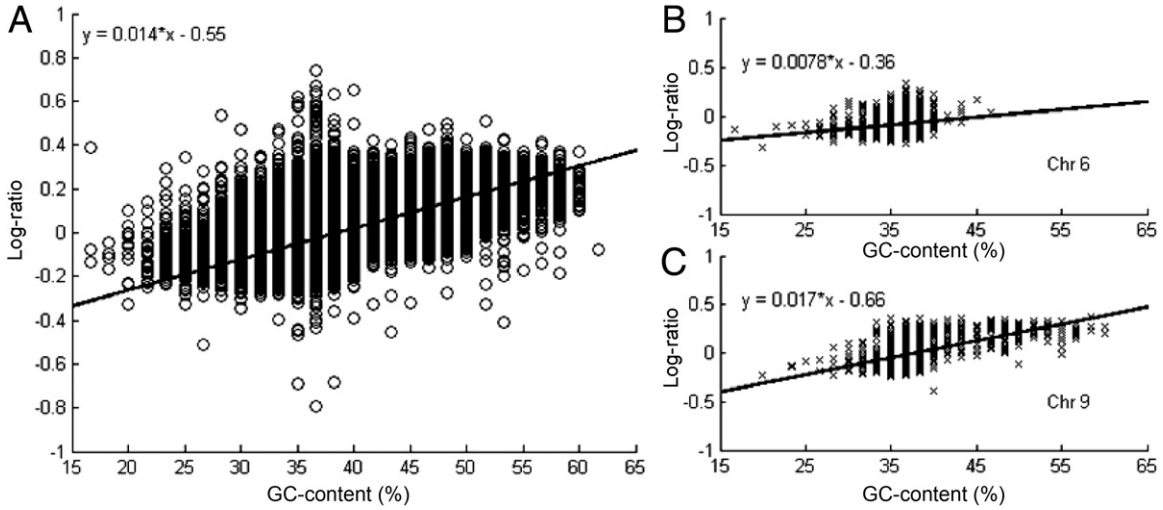


Figure 1. A: Scatterplot of log-ratio versus GC-content for each probe on an array. The linear regression (line and equation) demonstrates the trend between log-ratio signal and GC-content. Scatterplots of the same data for probes mapping to chromosome (Chr) 6 (**B**) and Chr 9 (**C**) that illustrate the variability of the slope and intercept of the regression between individual chromosomes in the same sample.

best-performing probes to compose the final array. The oligonucleotide probes used were 60 bases in length and were spotted once on the array. The feature layout on the array was randomized to prevent positional effects, and the Agilent Technologies Inc.–recommended replicate probe group and normalization controls were included on the array. Probes were distributed to afford maximum aberration detection resolution in regions of known clinical significance (mean probe spacing, 7500 bases). The remainder of the genome was apportioned a lower probe density (mean probe spacing, 125,000 bases).

Array CGH Analysis

Genomic DNA was extracted from whole blood using the QIAamp 96 DNA blood kit (Qiagen Inc., Valencia, CA). All aCGH analyses were performed against reference DNA pooled from six phenotypically normal males or females (Promega Corp., Madison, WI). Samples were sex-matched with the reference pool, except in cases where a sex mismatch was experimentally appropriate. The procedures for digestion, labeling, purification, and hybridization were performed in accordance with the manufacturer's protocols. After hybridization, the array slides were washed using a Little Dipper wash station (SciGene Corp., Sunnyvale, CA) and were scanned in a microarray scanner (G2505B; Agilent Technologies Inc.). The aCGH results were analyzed using the Feature Extraction (version 9.5.3) and DNA Analytics (version 4.0.76) software packages (Agilent Technologies Inc.) with the following settings: centralization threshold 6 and bin size 10, 7 probes and 0.40 log-ratio minimums, fuzzy zero not applied.[21] Aberrations were called in DNA Analytics using the aberration detection method 2 (ADM2) aberration detection algorithm and a score threshold of either 12.9 (before correction) or 10.4 (after correction by cghSAM).

Algorithm Development

We developed a two-step data correction algorithm (cghSAM) that uses normalized log-ratio signal intensities from two DNA samples involved in comparative hybridization. Step 1 is a data slope correction that is based on robust linear regression of log-ratio signal intensities to the probe GC-content. Step 2 is a chromosome-based normalization of the residual log-ratio bias based on historical chromosomal signal ratio medians.

Step 1—Slope Correction

The algorithm fits a robust linear regression to the probe log-ratio values plotted against probe GC-content. It uses the slope of the linear regression (GC slope) to correct individual probe log-ratios. Some of the previously published algorithms used the slope of the genome-wide GC-content linear regression as part of GC correction.[14,17,19] We observed a similar trend in the data but also found that the GC slope varied widely depending on which chromosome the probe was mapped (Figure 1; see also Supplemental Figure S1 at *http://jmd.amjpathol.org*). To avoid overcorrection, cghSAM was designed to correct probe log-ratios on a per-chromosome basis. The algorithm sorts the probe data by chromosome and uses robust regression to derive the chromosome-specific linear regression slope and the y-intercept estimate for each of the 24 chromosomes. The algorithm then calculates the median GC-content percentage of all the probes on a chromosome and uses the regression slope of that chromosome and the y-intercept to derive the log-ratio baseline for that chromosome (Figure 2A). The correction factor for probe *i* on chromosome *c* is defined as the difference between the calculated value and the log-ratio baseline (equation 1)

$$CorrectionFactor_i = LRBaseline_c - m_c \times PercentGC_i - b_c \quad (1)$$

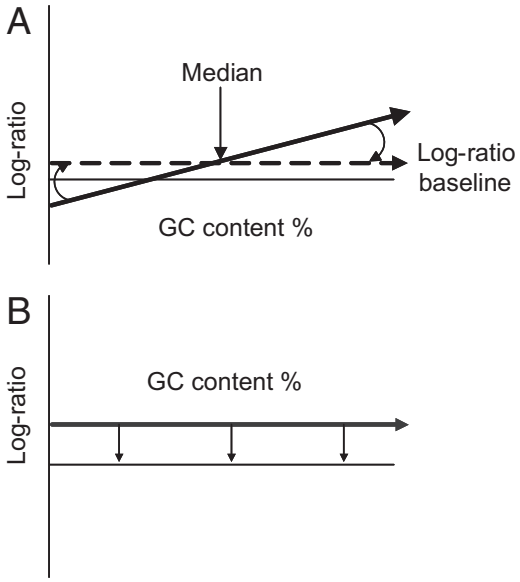


Figure 2. cghSAM correction strategy. **A:** Step 1 correction: the cghSAM algorithm calculates the median GC percentage for all probes on each chromosome. The log-ratio baseline for a particular chromosome is determined using its slope and intercept. Individual probe log-ratios are adjusted by their correction factor (the difference between the solid and dashed lines). **B:** Step 2 correction: some chromosome medians may be skewed above or below the baseline owing to GC-content. Chromosomal adjustment is used to correct the bias.

where m is the regression slope of chromosome c and b is the y-intercept for that chromosome.

Each probe is adjusted by its correction factor, and the median log-ratio of the corrected chromosome is stored for further calculations. After this correction, a subset of nonaberrant chromosomes that consistently skew above or below the expected baseline throughout the data set may require further correction (see step 2) (Figure 2B).

Step 2—Chromosomal Adjustment

Slope correction alone can be insufficient to fully normalize aCGH data. To target the genomic regions most affected by GC-waves, we designed a second step that adjusts the most consistently skewed chromosomes. Using a subset of the microarray data, we analyzed the medians of the 22 autosomes after step 1 correction. A subset of chromosomes had medians that were consistently skewed from the expected diploid baseline when two copies were present and that required further adjustment. This adjustment is designed to minimize the offset from the expected baseline seen in Figure 2B. However, automated adjustments of signal data for individual chromosomes inadvertently fail to detect large aberrations.[14,17,20] In cghSAM, only chromosomes that require a second adjustment step would be susceptible to this problem, as the first step corrects by slope only, without taking intercept into account. To prevent overcorrection, cghSAM uses mathematical safeguards to avoid overnormalization in truly aberrant regions by ensuring that any adjustment made falls within the expected range of adjustment for a nonaberrant sample. For the subset of chromosomes that consistently required adjustment after the step 1 correction, henceforth known as anchor chromosomes, we defined a typical pattern of median log-ratios between these chromosomes.

The selected chromosomes formed anchor set $\boldsymbol{A}$ with anchor values $a_1, \ldots, a_N$. Since any of the anchor chromosomes could potentially contain aberrant regions in a given sample, we chose to eliminate a subset of anchor chromosomes that were most likely to harbor a copy number change based on sample-specific data (Figure 3). For a given array, after step 1 correction, the algorithm defines a set of the step 1–corrected medians of anchor chromosomes $m_1, \ldots, m_N$. To remove outliers or chromosomes potentially harboring large copy number alterations, such as trisomies or large segmental deletions, the algorithm recursively searches set $\boldsymbol{A}$ for the chromosomes that possess signal medians most dissimilar from the anchor pattern. Every chromosome in set $\boldsymbol{A}$ is consecutively skipped in minimization of the sum in equation 2. The smallest sum is recorded, and the chromosome that was omitted in calculation of the minimal sum is designated as an outlier and removed from further anchor calculations, although all the anchor chromosomes are ultimately corrected.

$$\min_{e} \sum_{j=1}^{N} (a_j - e \cdot m_j)^2 \qquad (2)$$

The process is repeated until a predetermined number of chromosomes in set $\boldsymbol{A}$ have been designated as outliers (see *Results*). After all the outliers are removed and M chromosomes remain in the anchor set, cghSAM calculates the coefficient e^* such that the difference between the remaining anchors' medians, $a_1, \ldots, a_M$, and the rescaled medians, $m_1, \ldots, m_M$, is minimized (equation 3). The chromosomal adjustment factors for the array are then defined for all anchor chromosomes as a/e^*, including the previously removed anchor chromosomes. The log-ratio signal intensities for all the chromosomes in the original set $\boldsymbol{A}$ are corrected by subtracting their respective chromosomal adjustment factors.

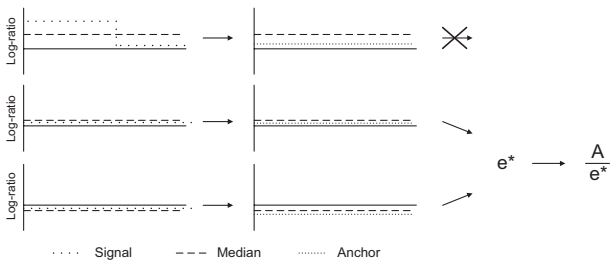


Figure 3. Exclusion of aberrant chromosomes. A schematic representation of the exclusion of potentially aberrant anchor chromosomes from calculation of the cghSAM chromosomal correction factor e^*. Three chromosomes are shown. The chromosome in the **top panel** harbors a segmental duplication, represented by the dotted line. This aberration skews the median probe log-ratio for the chromosome (dashed line) from the expected value of zero (solid line). The chromosomes in the **middle** and **bottom panels** have median probe log-ratios slightly above or below zero owing to GC-bias, but they harbor no aberrations. The sum of all the median log-ratios is minimized to determine an outlier chromosome (in this case the chromosome in the **top panel**). This is done to ensure that overcorrection does not result from use of an aberrant chromosome in the correction factor. After the chromosome in the **top panel** is eliminated from the e^* calculation, e^* is used to correct all chromosomes based on their anchor profiles.

$$e^{*} = \arg\min \sum_{j=1}^{M} (a_j - e \cdot m_j)^2 \quad (3)$$

Algorithm Implementation/Validation

After development, the cghSAM algorithm was implemented in Matlab (The MathWorks Inc., Natick, MA) and was tested on the custom microarray. To evaluate algorithm performance, we compared aberration calling sensitivity and specificity before and after algorithm correction on 215 arrays. We compared probe-by-probe log-ratios before and after correction for a subset of samples to understand overall array performance. We quantified the specificity and sensitivity improvements of the array by comparing aberration calls made before and after correction and by using data generated at a common polymorphic CNV locus. We also used these data to model the expected performance of the ADM2 aberration calling algorithm and to understand the technical capabilities of aberration calling on the array.

Results

Algorithm Parameters

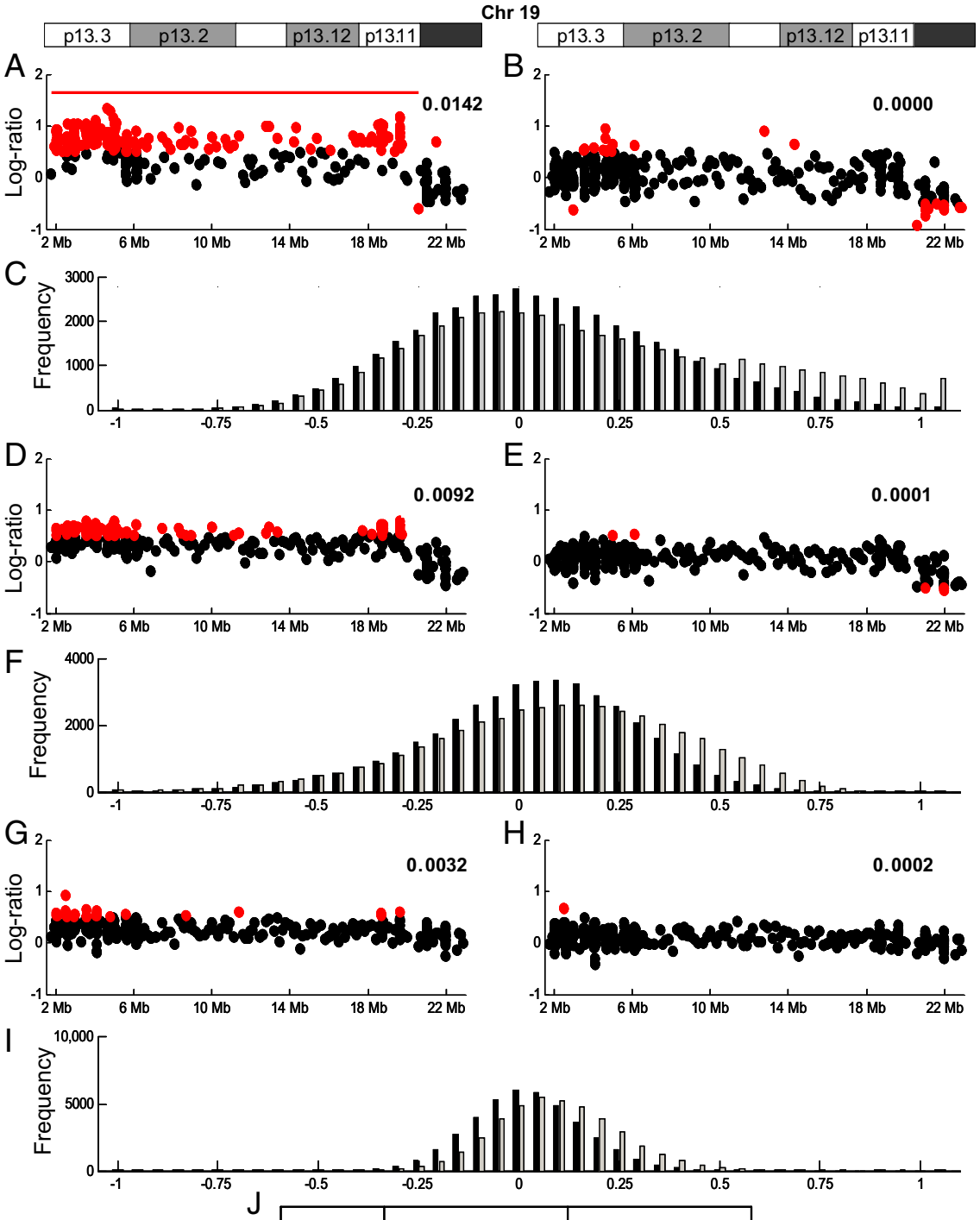
In developing the cghSAM algorithm, we investigated various correction strategies. Initially, we explored the use of a genome-wide linear regression of probe log-ratio plotted against probe GC-content, an approach that has been used by several research groups.[17,19,20] We found that the magnitude of this slope could be used as a quality control metric to identify samples harboring false-positive aberration calls and, in the most extreme cases, samples that could not be corrected because of poor data quality. Further research into the relationship between the GC-slope of the array and the GC-slope of individual chromosomes revealed that for certain chromosomes, the individual GC-slopes were substantially different from the GC-slope of the whole array in the present custom design (Figure 1; see also Supplemental Table S1 at *http://jmd.amjpathol.org*) and on a custom-designed 180K CGH array (Agilent Technologies Inc.), designed within the manufacturer's GC-content specifications (see Supplemental Figure S1 at *http://jmd.amjpathol.org*). To avoid overcorrection of probes on certain chromosomes, the cghSAM algorithm corrects probe log-ratios on a per-chromosome basis based on the individual chromosome regression result. We also examined further refining the correction algorithm by using GC-content in the genomic regions surrounding the individual probes rather than probe sequence but found that this had no effect on the chromosome regressions while being more computationally intensive (data not shown). Figure 1 also demonstrates the value of this array design method in which we used empirical data of probe performance from approximately 500,000 probes to select probes (44,000) that consistently generate the highest data quality. For example, the area of the graph in Figure 1 with log-ratios of 0.1 to −0.1 contains many probes with good performance over a breadth of GC-content percentages. The *in silico* probe selection programs used by array manufacturers restrict the choice of probes to a narrow range of GC percentages, thereby excluding many probes that perform well and unnecessarily limiting probe coverage in regions of skewed GC-content.

Although it is expected that a regression fitted to the corrected data would coincide with a baseline for diploid chromosomes, for certain chromosomes, we found that the corrected regression was consistently skewed above the baseline for GC-rich chromosomes and below the baseline for AT-rich chromosomes (Figure 2B). The nine chromosomes with medians consistently skewed from the baseline were the five most AT-rich chromosomes (4, 13, 5, 6, and 3) (negative offset) and the four most GC-rich chromosomes (19, 22, 17, and 16)[22] (positive offset) (see Supplemental Table S2 at *http://jmd.amjpathol.org*). This result suggested that further correction was necessary for these GC- and AT-rich chromosomes to reduce the log-ratio skew.

The second step of the algorithm involves adjusting the most GC- and AT-rich chromosomes to the log-ratio baseline for diploid samples. We identified nine chromosomes, a_1—a_9 (the most GC- and AT-rich, listed in the previous paragraph and in Supplemental Table S2 at *http://jmd.amjpathol.org*), that consistently required additional correction. In each sample, we chose to eliminate 40% of these chromosomes from the adjustment coefficient calculation in equation 3 based on the minimization of equation 2 (Figure 3). We chose 40% in this implementation because large aberrations on more than three of the autosomes are rarely observed in human specimens when aCGH is used in a reproductive testing setting, the application for this array. An exception to this assumption would be triploidy, but triploidy cannot be detected by any aCGH assay using solely copy number probes because of the normalization protocols used. For samples in which a larger percentage of aberrations could be expected, such as cancer samples, the percentage of outliers in the implementation could be increased. Several studies have discussed the sensitivity of automated correction algorithms for detecting CNVs over a broad range of sizes.[14,17,20] One way to avoid a potential loss of sensitivity is to train the normalization algorithm to recognize a typical wave pattern in nonaberrant (control) cases. An advantage of this method is that aberrant samples are not needed to calibrate the algorithm; all data necessary to calibrate the algorithm come from normal samples. Supplemental Figure S2 (available at *http://jmd.amjpathol.org*) details the workflow required to configure and implement cghSAM correction.

Algorithm Performance

The effect of the correction algorithm on reducing the GC-wave is shown for specific GC-rich regions with different GC-slopes (Figure 4, A–I). As shown in the three examples, over the whole genome, the correction moves the distribution closer to zero (the expected diploid log-ratio value) and tightens the distribution around zero, thus removing much of the GC-wave effect (Figure 4J). The effect of the cghSAM correction is more profound in samples with large GC-slopes (larger-magnitude GC-waves;



Chr 19 P arm	Before (Avg LR ± SD)	After (Avg LR ± SD)
Top	0.078 ± 0.398	0.013 ± 0.315
Middle	0.079 ± 0.313	0.024 ± 0.263
Bottom	0.058 ± 0.171	-0.006 ± 0.163

Figure 4. Precorrection and postcorrection data from three arrays with GC-waves. Three examples of diploid aCGH data before (**A, D,** and **G**) and after (**B, E,** and **H**) cghSAM correction show the results for individual probes on the p arm of chromosome (Chr) 19 for varying GC-wave amplitudes (**A**>**D**>**G**; also see the GC correction slope in the **top right** corner of the panels). Data from probes with log-ratios >0.5 or <−0.5 are displayed in red. The **top** sample contained a false-positive duplication call that was removed after correction (**red line** in **A** versus **B**). **C, F,** and **I:** The effect of cghSAM across the whole array for the same specimens is shown as histograms of probe number (*x* axis) versus log-ratios (*y* axis) before (gray) and after (black) correction. **J:** The results from the histograms are quantified to show the average (Avg) and SD of log-ratios (LRs) before and after correction for all probes on the arrays.

Figure 4, C and E) and affects fewer probes in samples with smaller GC-slopes (Figure 4I), as samples with lower GC-slopes have better data quality and need less data correction.

We analyzed 215 arrays with and without normalization by cghSAM (Table 1). The algorithm removed 12 of 16 false-positive calls (defined as aberrations called in the software that were subsequently shown to be negative by a FISH assay) ranging in size from 2 to 20 Mb. This increased specificity was achieved despite using an ADM2 score threshold of 10.4 (versus 12.9; see *Materials and Methods*) when cghSAM was used, which reduces the stringency for aberration calling, thereby increasing aberration detection sensitivity. After correction, all clinically reportable aberrations that were called before correction were also called, so the correction did not lead to any decrease in clinical sensitivity in the 215 samples.

To further investigate the sensitivity of the array for the calling of small aberrant regions, we assessed assay performance at a common polymorphic locus of the glutathione S-transferase theta-1 gene (*GSTT1*) (chr22: 22,706,139–22,714,284, HG18).[23] The *GSTT1* gene is absent in 15%[24] to 38%[25] of the population owing to a prevalent deletion.[26,27] The present custom array design has sufficient probe density to call aberrations in the 11 kb of chromosome 22 that include the *GSTT1* gene. Deletions and duplications of this gene occur commonly, making it ideal for studying the performance of cghSAM close to the limits of the aberration detection algorithm and over a range of data quality. Such an analysis is often not possible using clinically significant aberrations, which are, by definition, rare. We found that in the present patient population, the *GSTT1* locus was deleted in 35 of 215 patients according to the aCGH results and that these patients were clearly discernable from other copy number states. We calculated a trimodal distribution of $\log_2$ ratio average for *GSTT1* copy number and found that the means of each group corresponded to expected log-ratios for zero, one, and two copy samples against a one-copy reference and that the means of the groups differed by 2.5 SD, a boundary often used in copy number calling algorithms as a benchmark for significance (Figure 5). We found that 106 patients had one copy of *GSTT1* and 74 patients had two copies. From the frequency of zero-copy patients, we expected to see approximately 104 one-copy and 76 two-copy patients, corresponding very closely with the observed allele frequencies. By comparing these copy number calls before and after correction with the observed copy number state at the *GSTT1* locus (determined by the trimodal distribution), we could determine aberration calling sensitivity for two copy samples (Figure 5). The application of cghSAM correction resulted in a 33.8% increase in the rate of detection of a copy number change from one to two copies in the *GSST1* region (Figure 5 and Table 1).

We also modeled assay sensitivity in cghSAM-corrected data by examining the performance of X chromosome probes in female patient samples hybridized against male references. These probes displayed a copy

Table 1. Statistical Characteristics of the aCGH Assay Before and After cghSAM Correction

Characteristic	Before correction	After correction	Result
False-positive calls (No.)	16	4	−12
False-positive call rate (%)	7.4	1.9	−5.5
GSTT1 duplication calls made (copy number change from one to two copies) (No.)	26	51	25
GSTT1 duplication sensitivity (%)	35.1	68.9	33.8

Aberrant regions called before and after cghSAM using ADM2. Software settings are described in *Material and Methods*.

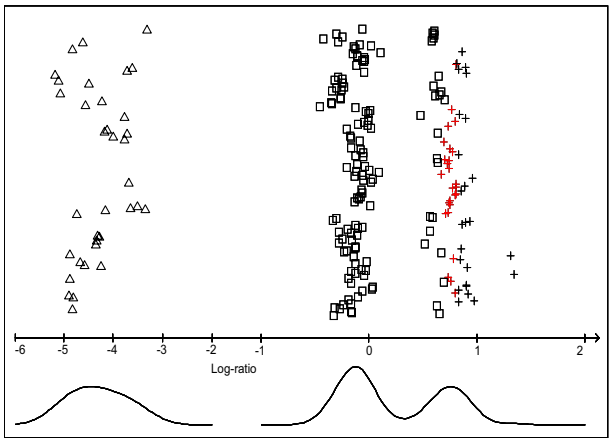


Figure 5. Plot of log-ratios at the *GSTT1* locus. Plot of the average log-ratios for a 10-probe region covering the *GSTT1* locus in 215 samples (triangles, called deletions; plus signs, called amplifications; and squares, no aberration call). Red plus signs denote samples that were not called before cghSAM correction but that were called after correction (n = 25 total). The trimodal distribution density is displayed under the plot.

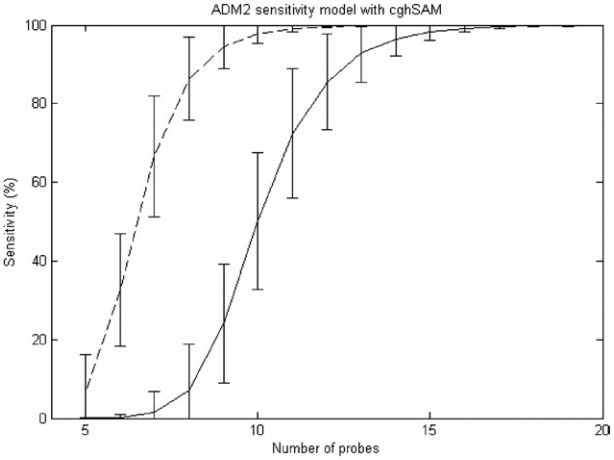


Figure 6. ADM2 sensitivity for deletion calling was modeled as a function of probe number. Probes located on the X chromosome were binned in sets of 5 to 50 probes in samples hybridized with a female sample and a male reference (one X chromosome in the reference to two X chromosomes in the sample) to emulate a heterozygous deletion (a copy number change from two copies in the reference to one copy in the sample) and to understand variability in probe performance. The model is applicable because the absolute value of $\log_2(2/1)$ is equal to the absolute value of $\log_2(1/2)$. The mean deletion detection sensitivity with ADM2 set to 10.4 (dashed line) (the post-correction score threshold) and 12.9 (solid line) (the precorrection score threshold) is pictured, with the error bars denoting ±1 SD.

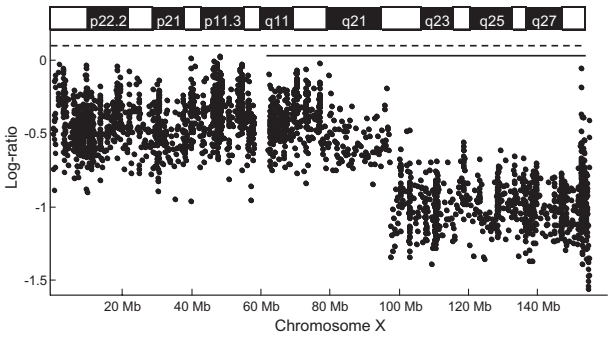


Figure 7. Increased mosaic aberration detection after correction. The solid line represents a mosaic deletion from Xq11 to Xq21 and a single copy loss from Xq22 to Xq28 detected before correction with cghSAM. The dashed line shows the expanded deletion call that encompasses an additional 1251 probe deletion from Xp23 to Xp11.

number change from one to two copies (and, thus, the equivalent absolute log-ratio to a heterozygous deletion), and, by binning the probes in sets of 5 to 50 probes per aberration, we could estimate the aberration calling sensitivity with the ADM2 algorithm at the precorrection and postcorrection score thresholds (Figure 6). The present model supported the conclusion that the cghSAM correction improved the sensitivity of analysis by demonstrating that copy number changes from one to two copies are more likely to be called at the postcorrection score threshold using cghSAM-corrected data. This model illustrates the performance of the ADM2 aberration detection algorithm over a range of probe performances and error measurements.

Many of the aberrations found in the 215 samples in this study were large and well above the configured log-ratio thresholds chosen to balance clinical specificity and sensitivity. The aberrations in these samples were detected before correction and also after correction with cghSAM. However, in one sample, after correction we detected a mosaic deletion from Xp23 to Xp11, in addition to the deletion from Xq11 to Xq28 that was detected before correction (Figure 7). The incremental 1251 probe deletion was detected after correction with an average log-ratio of −0.41 and was confirmed by FISH (11 of 30 cells had deletion of one copy of chromosome X, and the remaining 19 of 30 cells had deletion from Xq22 to Xq28 on one copy of chromosome X; Figure 7). Improvement in the data quality after correction led to the discovery of a clinically significant aberration that could aid in the diagnosis and potentially treatment of this patient. Several smaller deletions and duplications were also newly called as copy number alterations after cghSAM correction for other samples tested, ranging in size from 11 kb to 2 Mb, although these were generally classified as either benign or uncertain variants after review by clinicians (data not shown).

Discussion

As aCGH technologies have moved from BAC to oligonucleotide probes to increase aberration calling resolution, mathematical and statistical analyses have become essential in CNV detection. Algorithmic approaches that are designed to reduce systemic noise in aCGH data are highly desirable. We developed an algorithm (cghSAM) for use in an automated workflow that reduces GC-content–related signal variability in aCGH data. We tested this algorithm on 215 microarray samples, and we used aberration calling sensitivity and specificity and overall probe log-ratio distribution before and after cghSAM correction to evaluate the performance of cghSAM.

Performance of the cghSAM algorithm was evaluated by four criteria. Genome-wide probe performance was analyzed by comparing probe log-ratio distribution before and after algorithmic correction, and the algorithm improved overall data quality as measured by average probe log-ratio and SD in samples over a range of GC-wave magnitudes. It is critically important in a clinical setting that correction by cghSAM does not inadvertently prevent aberration detection. In the entire data set, all cytogenetically confirmed calls were retained after correction. We also examined a small, well-characterized, copy number polymorphic locus, *GSTT1*, as a model for deletion detection. Here, the application of cghSAM resulted in improved accuracy of 33.8%.

The initial data set of 215 arrays contained 16 instances where array findings were not confirmed by cytogenetic assays, for a 7.4% false-positive rate before correction. The cghSAM correction algorithm eliminated the false-positive aberration calls in 12 of the samples; the remaining 4 false-positive samples (1.9%) showed especially strong genome-wide GC-slopes and demonstrated the value of the analysis inherent in cghSAM as a quality control metric. The cghSAM algorithm improved array performance, including overall sensitivity and specificity, without any loss of sensitivity for cytogenetically confirmed aberrations. The improvement in sensitivity in small regions can partially be attributed to the more permissive ADM2 score threshold enabled by the reduction in wave amplitude, which was achieved with a simultaneous increase in specificity.

In general, quantifying the sensitivity of aCGH arrays is difficult because of the large amount of data generated per sample and the difficulty in confirming small aberrations throughout the genome. By using probes present on the array at a common polymorphic locus, we quantified array performance over a large data set using established methods. Prevalent CNVs are especially useful for this purpose since characteristics such as aberration size and genomic context can be tailored to specific experiments.

Understanding the performance of the array over many samples at a specific locus also enabled us to model the performance of the ADM2 aberration detection algorithm over a wide range of probes. When calling small aberrations with scores near thresholds, the performance of individual probes can have a substantial effect on whether a call is made. Therefore, while the *GSTT1* region was a good measure of the sensitivity improvement on this array, it is not a sufficient model for probe performance over the entire array. By using sex-mismatched samples exhibiting a one- to two-copy change for the entire X chromosome, we could more accurately model and understand the range of probe performance over the entire array and the likelihood of calling small aberrations

over the entire array at different aberration calling thresholds. We found that without sufficient probe coverage at a particular locus, even aberrations containing perfectly performing probes will often be missed owing to algorithm score thresholds. Analyzing the performance of aberration detection algorithms in this way is important in understanding the true resolution of an array and can be useful for creating new array designs.

The cghSAM algorithm was developed using a custom 4 × 44K CGH array (Agilent Technologies Inc.) with probes designed outside the manufacturer's suggested GC-content range to provide a more complete interrogation of copy number throughout the genome than was possible using the manufacturer's specifications. By developing this algorithm, we could use probes in GC- and AT-rich regions of the human genome that would otherwise have been inaccessible to accurate aCGH analysis. The algorithm was effective in correcting GC-waves in regions where these custom-designed probes were used, enabling more comprehensive and accurate copy number detection at high resolution throughout the genome, independent of GC-content. Although it was developed for the present custom platform, the algorithm could be used for any oligonucleotide array format containing copy number probes and is effective in correcting aCGH data even in array designs that do not use probes of high and low GC-content. The importance of correcting GC-waves for probes within the manufacturer's suggested GC-content is well established as Agilent Technologies Inc. now includes a GC correction algorithm in their own aCGH analysis software packages. To use the presented algorithm, a sufficient number of samples (either aberrant or nonaberrant) would have to be run to understand the GC-wave patterns and chromosome biases present, but the number of chromosomes or genomic regions present in the anchor set, as well as the number or percentage of regions removed from the correction calculation for potentially harboring aberrations, can be modified to meet the needs of a particular sample pool. Although we optimized the algorithm parameters for use in samples being tested for microdeletion and/or microduplication syndromes, if samples suspected of harboring many large or mosaic aberrations were being analyzed, such as cancer samples, these parameters could be optimized differently. Otherwise, the only input needed by the algorithm would be probe sequence and genomic location.

In summary, we presented an automated data correction algorithm designed to reduce systemic data variability caused by the platform-independent GC-wave artifact and discussed methods for evaluating array performance. This algorithm systemically adjusts individual probe signal intensities and, in doing so, reduces GC-wave amplitude, making CNV calling more accurate. The algorithm can be used on any array platform that incorporates copy number probes, and it does not require aberrant samples to optimize performance to a specific platform. Now that aCGH is common clinical practice, this algorithm will enable more accurate detection of CNVs, and these methods will facilitate a better understanding of the performance of aCGH platforms.

References

1. Baldwin EL, Lee JY, Blake DM, Bunke BP, Alexander CR, Kogan AL, Ledbetter DH, Martin CL: Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray. Genet Med 2008, 10:415–429
2. Shevell M, Ashwal S, Donley D, Flint J, Gingold M, Hirtz D, Majnemer A, Noetzel M, Sheth RD: Practice parameter: evaluation of the child with global developmental delay: report of the Quality Standards Subcommittee of the American Academy of Neurology and the Practice Committee of the Child Neurology Society. Neurology 2003, 60:367–380
3. Shaffer L, Ledbetter D, Lupski J: Molecular cytogenetics of contiguous gene syndromes: mechanisms and consequences of gene dosage imbalance. Metabolic and Molecular Basis of Inherited Disease. Edited by CR Scriver, AL Beaudet, WS Sly, D Valle, B Childs, KW Kinzler, B Vogelstein, New York, McGraw Hill, 2001, pp 1291–1324
4. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C: The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 2008, 82:685–695
5. Slavotinek AM: Novel microdeletion syndromes detected by chromosome microarrays. Hum Genet 2008, 124:1–17
6. Shaffer LG, Kashork CD, Saleki R, Rorem E, Sundin K, Ballif BC, Bejjani BA: Targeted genomic microarray analysis for identification of chromosome abnormalities is 1500 consecutive clinical cases. J Pediatr 2006, 149:98–102
7. De Gregori M, Ciccone R, Magini P, Pramparo T, Gimelli S, Messa J, et al: Cryptic deletions are a common finding in "balanced" reciprocal and complex chromosome rearrangements: a study of 59 patients. J Med Genet 2007, 44:750–762
8. Christian SL, Brune CW, Sudi J, Kumar RA, Liu S, Karamohamed S, Badner JA, Matsui S, Conroy J, McQuaid D, Gergel J, Hatchwell E, Gilliam TC, Gershon ES, Nowak NJ, Dobyns WB, Cook EH Jr: Novel submicroscopic abnormalities detected in autism spectrum disorder. Biol Psychiatry 2008, 63:1111–1117
9. Baptista J, Mercer C, Prigmore E, Gribble SM, Carter NP, Maloney V, Thomas NS, Jacobs PA, Crolla JA: Breakpoint mapping and array CGH in translocations: comparison of a phenotypically normal and an abnormal cohort. Am J Hum Genet 2008, 82:927–936
10. Van den Veyver IB, Patel A, Shaw CA, Pursley AN, Kang SH, Simovich MJ, Ward PA, Darilek S, Johnson A, Neill SE, Bi W, White LD, Eng CM, Lupski JR, Cheung SW, Beaudet AL: Clinical use of array comparative genome hybridization (aCGH) for prenatal diagnosis in 300 cases. Prenat Diagn 2009, 29:29–39
11. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH: Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 2010, 86:749–764
12. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al: Global variation in copy number in the human genome. Nature 2006, 444:444–454
13. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nat Genet 2004, 36:949–951
14. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, Tavare S, Hurles ME: Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol 2007, 8:R228
15. Carter N: Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 2007, 39:S16–S21
16. Lepretre F, Villenet C, Quief S, Nibourel O, Jacquemin C, Troussard X, Jardin F, Gibson F, Kerckaert JP, Roumier C, Figeac M: Waved aCGH: to smooth or not to smooth. Nucleic Acids Res 2010, 38:e94
17. Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K: Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res 2008, 36:e126

18. Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai AK, Liu JS, Chen R, Liu XS: Model-based analysis of two-color arrays (MA2C). Genome Biol 2007, 8:R178
19. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S: A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 2005, 14:6071–6079
20. van de Wiel MA, Brosens R, Eilers PH, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B: Smoothing waves in array CGH tumor profiles. Bioinformatics 2009, 9:1099–1104
21. Agilent oligonucleotide array-based CGH for genomic DNA analysis. CGH protocol version 4.0. Agilent Technologies, 2006
22. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR: Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci 2008, 105:16266–16271
23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. Genome Res 2002, 12:996–1006
24. Chen H, Sandler DP, Taylor JA, Shore DL, Liu E, Bloomfield CD, Bell DA: Increased risk for myelodysplastic syndromes in individuals with glutathione transferase theta 1 (GSTT1) gene defect. Lancet 1996, 347:295–297
25. Pemble S, Schroeder KR, Spencer SR, Meyer DJ, Hallier E, Bolt HM, Ketterer B, Taylor JB: Human glutathione S-transferase theta (GSTT1): cDNA cloning and the characterization of a genetic polymorphism. Biochem J 1994, 300:271–276
26. Hallier E, Jager R, Deutschmann S, Bolt H, Peter H: Glutathione conjugation and cytochrome P-450 metabolism of methyl chloride in vitro. Toxicol In Vitro 1990, 4:513–517
27. Wiebel F, Dommermuth A, Thier R: The hereditary transmission of the glutathione transferase hGSTT1-1 conjugator phenotype in a large family. Pharmacogenetics 1999, 9:251–256