Published online 10 September 2008

Nucleic Acids Research, 2008, Vol. 36, No. 19 ***e126***
doi:10.1093/nar/gkn556

Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms

Sharon J. Diskin[1,2,3], Mingyao Li[4], Cuiping Hou[1], Shuzhang Yang[5], Joseph Glessner[6], Hakon Hakonarson[6], Maja Bucan[5], John M. Maris[1,3,*] and Kai Wang[5,6]

[1]Center for Childhood Cancer Research, Children's Hospital of Philadelphia, [2]Genomics and Computational Biology, University of Pennsylvania, [3]Department of Pediatrics, University of Pennsylvania, [4]Department of Biostatistics and Epidemiology, University of Pennsylvania, [5]Department of Genetics, University of Pennsylvania and [6]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Received April 19, 2008; Revised August 13, 2008; Accepted August 18, 2008

ABSTRACT

Whole-genome microarrays with large-insert clones designed to determine DNA copy number often show variation in hybridization intensity that is related to the genomic position of the clones. We found these 'genomic waves' to be present in Illumina and Affymetrix SNP genotyping arrays, confirming that they are not platform-specific. The causes of genomic waves are not well-understood, and they may prevent accurate inference of copy number variations (CNVs). By measuring DNA concentration for 1444 samples and by genotyping the same sample multiple times with varying DNA quantity, we demonstrated that DNA quantity correlates with the magnitude of waves. We further showed that wavy signal patterns correlate best with GC content, among multiple genomic features considered. To measure the magnitude of waves, we proposed a GC-wave factor (GCWF) measure, which is a reliable predictor of DNA quantity (correlation coefficient = 0.994 based on samples with serial dilution). Finally, we developed a computational approach by fitting regression models with GC content included as a predictor variable, and we show that this approach improves the accuracy of CNV detection. With the wide application of whole-genome SNP genotyping techniques, our wave adjustment method will be important for taking full advantage of genotyped samples for CNV analysis.

INTRODUCTION

Many genomics applications involve examination of signal intensity patterns of probes across the genome, and make inference on the gains and losses of genomic elements from examination of these signal intensities at different chromosome regions. These probes vary greatly in size, ranging from hundreds of kilobases for traditional BAC clone-based array-CGH experiments, to dozens of base pairs for oligonucleotide arrays and high-density single nucleotide polymorphism (SNP) genotyping arrays (1). Typically, a signal intensity measure is calculated for each probe or each probe set, and these intensity values are used to make inference on gains or losses of genomic segments. Various data normalization techniques have been developed to better summarize the intensity values between markers and between experiments, and to accurately capture genomic gains and losses, commonly referred to as copy number variations (CNVs) (2,3).

Recently, with the increasing application of high-resolution CNV detection methods, a genome-wide spatial autocorrelation or 'wave' pattern in signal intensity data was described that interferes with accurate CNVdetection (4). We use the term 'genomic wave' to refer to these patterns of signal intensities across all chromosomes, where different samples may show highly variable magnitude of waviness. This phenomenon has been observed before (5), but the first formal description appeared recently for CNV analysis using an array-CGH platform (4). Marioni *et al.* (4) described the presence of genomic waves in their Whole-Genome Tiling Path arrays used for CNV detection, and demonstrated that the wavy patterns they observed appeared to be a 'general feature of aCGH data sets'. They developed a method based on Lowess regression to 'break' the waves and improve CNV calling. Furthermore, Komura *et al.* (6) also described the wavy patterns in signal intensities of Affymetrix arrays, and they reduce the signal noise by incorporating probe and target sequence characteristics in the Genomic Imbalance Map (GIM) algorithm. Nannya *et al.* (7) has also described similar phenomenon in CNV studies on cancer genome by the Affymetrix SNP arrays, and this effect was adjusted by the length and GC content of the PCR products using

*To whom correspondence should be addressed. Tel: +1 215 590 5244. Email: maris@chop.edu
Correspondence may also be addressed to Kai Wang. Tel: +1 267 426 2378; Email: wangk@chop.edu

quadratic regressions, for the purpose of compensation for different PCR conditions.

Besides array-CGH platforms, other CNV detection platforms of similar nature may also be susceptible to genomic waves. In our genotyping experiments using the Illumina HumanHap550 arrays, we have observed obvious genomic waves in many batches of samples. In our studies, even for DNA samples available from commercial cell line repositories, typically ~10% show strong wavy patterns that are visually discernable in the BeadStudio software (Illumina Inc., San Deigo, CA, USA). The presence of genomic waves may adversely affect the performance of CNV calling algorithms and can result in inflated false positive calls. It is of great interest to perform a comprehensive analysis of signal intensity patterns across several SNP genotyping platforms, investigate the causes of genomic waves and find ways to reduce these waves from both experimental and computational perspectives.

In the current study, we first perform a comparative analysis of genomic wave artifacts in several different high-density SNP genotyping arrays, and confirm that genomic waves are not a platform-specific phenomenon. Next, we perform exploratory analysis of local genomic features (such as GC content, gene content and segmental duplication patterns), to find potential genomic features that correlate with genomic waves. We investigate the technical cause of waves by examining potential DNA degradation or protein contamination, and by performing serial dilutions from the same sample to assess the impact of DNA quantity. These experiments allowed us to identify the property of DNA samples that leads to genomic waves and to find ways to reduce genomic waves in the experimental protocol. Finally, we present a method to computationally reduce the effects of genomic waves and show that this approach reduces the wavy patterns of signal intensities and improves the accuracy of CNV detection.

METHODS

Genotyping procedure

All DNA samples genotyped using the Illumina BeadChip are part of an ongoing genome-wide association study in neuroblastoma and satisfied stringent quality control as described elsewhere (8). Genotyping was performed using the Illumina Infinium™ II HumanHap550 BeadChip (Illumina, San Diego, CA, USA) according to methods detailed elsewhere (9,10). All DNA samples were surveyed for quality both by optical density spectrophotometry and the pico-green assay, and samples judged to be of sufficient quality for genotyping were assayed at the Center for Applied Genomics, the Children's Hospital of Philadelphia. The genotyping signal intensity data for Illumina HumanHap1M arrays and Affymetrix Mapping 500K arrays were generated on HapMap individuals, and were acquired from Illumina Inc. and Affymetrix Inc., respectively. The genotyping signal intensity data for Affymetrix genome-wide 6.0 arrays were generated from an internal control collection recruited at Children's Hospital of Philadelphia. For the serial dilution experiment, varying quantities of DNA (187.5, 375, 750, 1500 and 2250 ng) from an anonymous individual were used for genotyping by the Illumina HumanHap550 array.

Derivation of Log R Ratio as signal intensity measure

The Log R Ratio (LRR) value is originally developed on the Illumina platform as a normalized signal intensity measure (11). For each SNP, let the signal intensities for the A and B alleles be denoted as X and Y, respectively. We can then calculate the R-value as $R_{\text{observed}} = X + Y$. As a normalized measure of total signal intensity, LRR is then calculated as $\log_2(R_{\text{observed}}/R_{\text{expected}})$, where R_{expected} is computed from linear interpolation of the canonical genotype clusters (11). For the Illumina SNP arrays, the LRR values can be directly calculated and exported from the BeadStudio software. For the Affymetrix platform, we first extract allele-specific signal intensity values (X and Y) by the Affymetrix Power Tools (http://www.affymetrix.com/support/developer/powertools/index.affx), then construct the canonical genotype clusters using all genotyped samples and calculate the LRR values. The Affymetrix genome-wide 6.0 and the Illumina HumanHap1M arrays contain nonpolymorphic markers to improve genome coverage. For each nonpolymorphic marker in each sample, we take the median value of all samples as the R_{expected} value for computing the LRR values.

Analysis of genomic features

We used a nonoverlapping window approach to test whether the median signal intensity values within each window correlate with particular genomic features, including GC percentage, segmental duplication, gene content, exon content, simple repeat and conserved genomic region. All these features are annotated in the UCSC Genome Browser annotation databases (12,13). The GC percentage data was retrieved from the gc5Base table, the segmental duplication data was retrieved from the genomicSuperDups table, the gene content annotation was retrieved from the refGene table, the exon annotation was also retrieved from the refGene table, the simple repeat annotation was retrieved from the simpleRepeat table and the most conserved genomic region annotation was retrieved from the phastConsElements28way table. We sectioned the genome into 10 kb, 100 kb or 1 Mb nonoverlapping windows; for each window, we calculated the fraction of bases that fall within the annotated regions for each of the genomic features. Windows with less than three SNPs were excluded from analysis. We then calculated the correlation coefficient between the median LRR values and the fraction of annotated bases within each window across the genome.

Derivation of wave factor and GC-wave factor

To quantify the magnitude of signal fluctuation of genotyped samples, it is necessary to develop a summary measure of waviness. This measure should not be susceptible to the presence of CNVs (which generate extreme values in the signal intensity measures), and should be comparable

between different cohorts, or even arrays with different density of markers (for example, the Illumina HumanHap300, HumanHap550 and HumanHap1M arrays). We have developed a score called wave factor (WF) that is based on median absolute deviation of signal intensities. We calculate the median signal intensity values (normalized signal intensity as LRR value) for every 1 Mb nonoverlapping window in the genome and denote them as Y_i (i = 1 to ~3000 for human genome). The windows containing less than 10 SNPs were excluded from the calculation. We then compute the correlation between median signal intensity and local GC content in all windows, and denote this value as r_{GC}. In this study, r_{GC} is calculated using all windows on chromosome 11. The WF score is defined as

$$S_{WF} = (1 - 2 \times I_{\{r_{GC}<0\}}) \times \text{median}\,(|\,Y_i - \text{median}\,(Y_i)|)$$

The first part of the equation is used to assign the sign of the S_{WF}, to help differentiate waves of different directionality. The second part involves a median absolute deviation calculation, which is a similar measure to the commonly used average absolute deviation but is less affected by extreme values in the tail. Therefore, even in the presence of large CNVs in a genotyped sample, the effects on WF score will be reduced or eliminated since these regions are represented in the tail of the distribution.

The variability of signal intensity within each particular sample could be due to multiple reasons, and GC content may only explain partial variability of the WF. To quantify the signal fluctuation that can be attributed to the local GC content, we developed another measure called GC-wave factor (GCWF). This measure is simply the product of WF value and the absolute values of r_{GC}:

$$S_{GCWF} = S_{WF} \times |r_{GC}|$$

Intuitively, the WF and GCWF measures can be understood in this way: the WF function as a proxy for total signal fluctuation, but is more resilient to outliers than the standard deviation measure. The square of r_{GC} can be considered as the fraction of variance explained by local GC content. Therefore, the GCWF score is a summary measure of the signal fluctuation explained by local GC content.

Regression model for signal adjustments

We developed a simple statistical method to adjust signal intensity values at each marker for samples affected by genomic waves. Unlike 'smoothing' based regression methods that try to borrow information from neighboring markers in the adjustment, our method adjusts each marker separately regardless of the signal intensities at neighboring markers, therefore eliminating concerns on smoothing out true CNV boundaries. Suppose there are M (for example, M = ~550K for Illumina HumanHap550 array) markers in a genotyped sample, we collect all the m autosome markers that are at least 1 Mb away from each other (for example, m = ~3K for Illumina HumanHap550 array). This method reduces the number of response variables in regression model, and eliminates the potential dependence between markers due to colocalization in the same genomic region. For each of the m markers, we collect its LRR value as L_j (j = 1, ..., m) and the average GC percentage in the 1 Mb window around the marker, then fit a linear regression model:

$$L_j = \alpha + \beta \times G_j + \varepsilon_j$$

where the model parameters α and β are estimated by the least-squares method. To reduce the effect of markers within CNV regions on the regression coefficients, we restricted the analysis to markers with LRR between −2 and 1. After obtaining these estimated regression parameters, for each of the M marker in the genotyping array, we then calculate the expected signal intensity value based on the GC percentage in the 1 Mb window around the marker. The adjusted signal intensity value is then simply calculated as the observed LRR value minus the expected value (residual in the regression model). The procedures for signal adjustment are implemented in the PennCNV package, available at http://www.openbioinformatics.org/penncnv. The adjustment procedure is available as a stand-alone application that can be used outside of PennCNV, and has also been incorporated directly into the CNV calling procedure within PennCNV.

Quantitative PCR for CNV validation

The copy numbers of six CNV regions were examined by real-time quantitative PCR (Q-PCR) in 48 samples on the ABI Prism 7900HT system (Applied Biosystems, Foster City, CA, USA) using SYBR Green Dye. The primer pairs were designed using PrimerExpress2.0 software (sequences available upon request). The endogenous control was designed to target the *DEC2* gene in chromosome 12, avoiding any known structural variations including CNVs. The ΔΔCt method (User Bulletin #2 for ABI Prism 7700 Sequence Detection System; Applied Biosystems) was employed to quantify the genomic copy numbers by setting a normal number at two copies. For all CNV regions and for all samples, the quantitative copy number estimates by Q-PCR are 1.0E-6 to 3.8E-4 for zero copy genomic regions, 0.79–1.33 for one copy, 1.58–2.52 for two copies, 2.63–3.36 for three copies, 3.62–4.72 for four copies, therefore implicating the high accuracy of Q-PCR for CNV validation.

RESULTS

The presence of genomic waves in SNP genotyping platforms

To examine whether wavy patterns of signal intensities exist in different types of SNP genotyping platforms, we analyzed the intensity data for the Affymetrix Mapping 500K array, Affymetrix genome-wide 6.0 array, and the Illumina HumanHap550 and HumanHap1M arrays (see Methods section). We examined the LRR value, which is a normalized signal intensity measure originally developed for the Illumina platform (11), but we note that it can be adapted to the Affymetrix platform as well (see Methods section). The LRR value for a probe is a measure of the difference between the signal intensity of

the test sample and a pool of reference samples of the same SNP genotype. This measure is a preferred signal intensity summarization measure for CNV detection, since it greatly reduces the signal variability between different markers.

For each SNP array, we selected one example with strong waves and displayed their signal intensities on chromosome 11, where the wavy patterns are especially easy to discern visually (Figure 1). Although, different samples were selected from each platform, when we compared the four arrays, we found the waves had similar periods and phases at similar genomic regions, indicating that regions with inflated or deflated signal intensities in one array are likely to show similar patterns in other arrays. However, the directionality of the magnitude of waves can be identical or opposite between arrays, that is, the waves can be either 'in phase' or 'of opposite phase' with each other: we specifically selected examples where the peaks of one Illumina (or Affymetrix) array correspond to troughs of another Illumina (or Affymetrix) array. In addition, the magnitude of waves can be different between different genotyped samples: for example, the sample in Figure 1D showed more obvious wave patterns than other genotyped samples.

For the four genotyped samples in Figure 1, we calculated their pairwise correlation coefficients of median LRR values in 1 Mb nonoverlapping windows of autosomes (Supplementary Table 1). The absolute values of these correlations range from 0.73–0.91, further demonstrating that the locations of peaks and troughs in signal intensity are consistent across arrays. Therefore, the presence of genomic waves is unlikely to be an artifact caused by probe hybridization or data normalization specific for each particular SNP genotyping platform. It is of interest to investigate what genomic features correlate with the peaks and troughs of the waves.

Correlation of local genomic features with signal intensity

Since different SNP genotyping arrays show the same phase and period of genomic waves, we assume that certain local genomic features may determine the magnitude of waves in local genomic regions. Previous studies in array-CGH platforms and Affymetrix SNP genotyping platforms suggested correlation between intensity of the clones/probes and the local GC content (4,6). However, GC content is correlated with many genomic features, and these features have not been investigated to find the most likely predictor of waves.

To investigate the relationships between variations in several local genomic features and variation in local signal intensities, we randomly chose two samples with

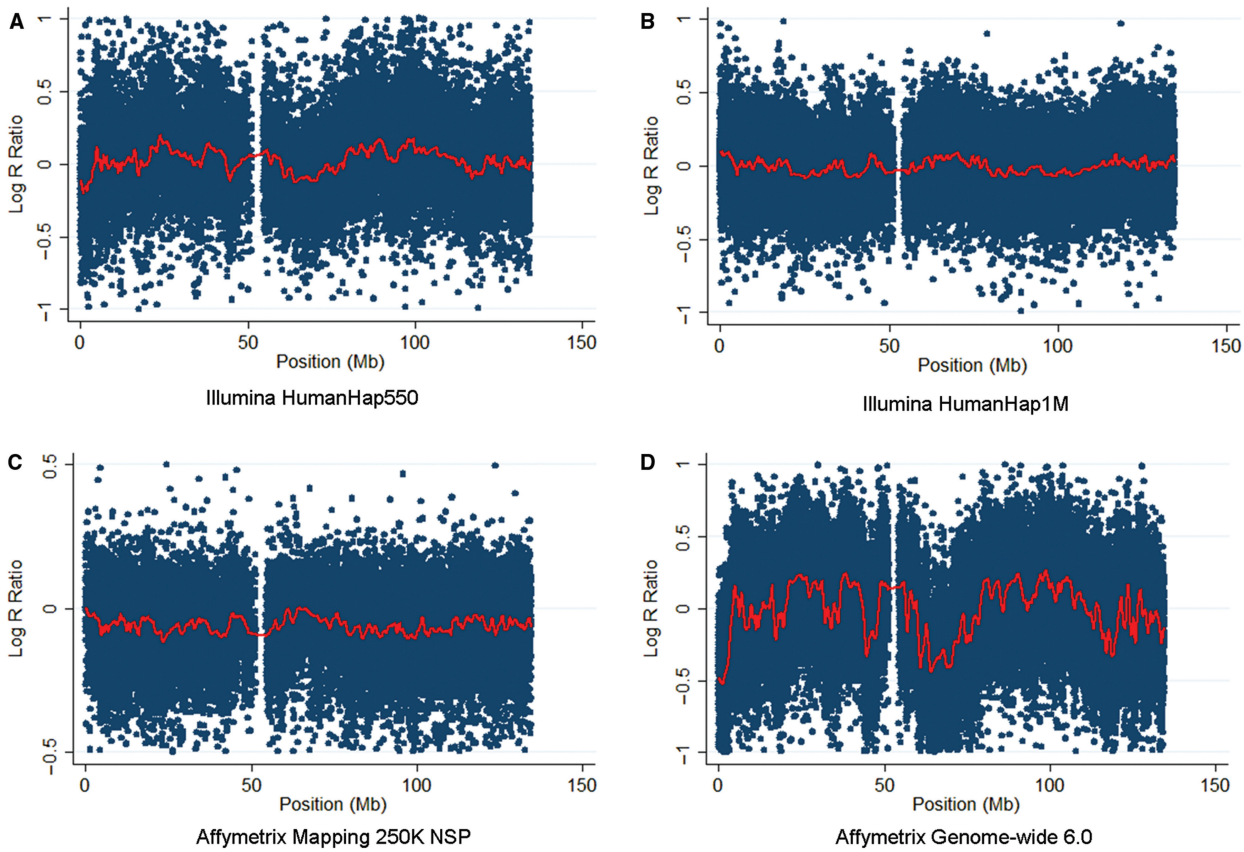

Figure 1. Genomic wave is not platform-specific. An illustration of four representative examples showing the LRR values for chromosome 11 in the Affymetrix 250K NSP (from Affymetrix Mapping 500K array sets) and genome-wide 6.0 arrays, as well as the Illumina HumanHap550 and HumanHap1M arrays. Four different DNA samples were genotyped by these four arrays. In all cases, we observe similar shapes of wavy patterns with identical or opposite peaks and troughs. This indicates that genomic wave is an intrinsic property of the human genome, regardless of the technical platforms.

opposite direction of genomic wave (Figure 2A and E, Table 1). We generated nonoverlapping windows across the entire genome, and calculated the median signal intensity in each of these windows. We then calculated the correlation coefficient of median signal intensities for each window with multiple genomic features in the same window, including the fraction of GC content, the fraction of bases with segmental duplication (14), the fraction of bases within RefSeq transcripts (15), the fraction of bases within exons of RefSeq transcripts (15), the fraction of bases within simple repeat region (16) and the fraction of bases within most conserved genomic regions in 28-way vertebrate multiple alignments (17). The analysis was repeated with a window size of 10, 100 kb and 1 Mb, respectively. Among all genomic features we examined, we found that local GC content is the most correlated factor with median signal intensities. In addition, for both samples, the magnitude of correlation between signal intensity and GC content increases for larger window sizes (Figure 2B–D and F–G). To further illustrate the correlation of the signal intensity values and the six genomic features in 1 Mb windows, we plotted the values along a representative chromosome for the two samples, and demonstrated the clear correspondence of peaks and troughs between signal intensity and GC content (Figure 3). Finally, we plotted the signal intensity values and GC content in 1 Mb windows in the entire genome, and demonstrated a consistent pattern of correspondence across all chromosomes (Supplementary Figure 1). Altogether, our analysis indicates that local GC content is the best predictor of the magnitude of local genomic waves among the features we evaluated. To distinguish the directionality of wavy patterns in different samples, we refer to waves that have positive correlation with GC as positive waves (Figure 2A) and those with negative correlation as negative waves (Figure 2E).

Quantifying the magnitude of genomic waves

Since different DNA samples demonstrate varying magnitude of waviness, we have developed two quantitative measures to summarize the signal fluctuation patterns for genotyped samples. The first measure, called WF, summarizes the total signal fluctuation of a genotyped sample across the genome (see Methods section). The second measure, called GCWF, specifically summarizes the fraction of signal fluctuation that correlates with patterns of GC distribution. Since signal fluctuation can be caused by many factors, including genotyping array defects, presence of large CNVs, cell-line induced rearrangements and local GC content, the GCWF measure effectively measure the fraction of signal fluctuation that can be explained by genome-wide patterns of GC distribution. Both the WF and GCWF use the median absolute deviation of signal intensities, and therefore are less affected by the presence of extreme values (such as those within CNVs) than a standard deviation measure. The WF for samples in Figure 2A and Figure 2E are −0.09 and 0.05, respectively, while the GCWF for them are −0.09 and 0.04, respectively. These measures are applicable to many different types of arrays that measure signal intensities across genome, including array-CGH, SNP genotyping arrays and whole-genome oligonucleotide arrays. Since the magnitude of genomic waves is highly correlated with GC content, we utilize the GCWF in our subsequent analyses.

The quantity of DNA causes genomic waves

The presence of positive or negative waves with different magnitudes in different genotyped samples indicates that the directionality of waves reflect certain sample properties. An intuitive explanation is that the LRR is calculated using a fixed set of reference samples, so the wave pattern is a reflection of how similar a given testing sample is to the pool of reference samples. Different directions of signal deviation from population mean result in different directions of waves due to the signal normalization procedure. It is therefore interesting to test what property of DNA samples correlate with the magnitude of wavy patterns.

We initially suspected that DNA quality, such as DNA degradation or protein contamination, could be a direct cause of genomic waves, since strong waves are less frequently observed for DNA samples purchased from standardized cell line repositories than for DNA extracted from patient samples. Using a set of 1444 DNA samples genotyped by the Illumina HumanHap550 arrays, we investigated the relationship between DNA quality and the magnitude of waves. We first calculated the GCWF for each of the 1444 DNA samples. Figure 4A illustrates that DNA purity, as estimated by the 260/280 ratio of DNA samples, does not correlate with GCWF. To investigate the possibility of DNA degradation affecting waviness, we performed an electrophoresis assay (Figure 4B) on a representative set of 18 cases, including six with little genomic wave, six with strong positive waves and six with strong negative waves. We did not observe evidence of increased DNA degradation for samples with positive or negative waves (examples of gel patterns for degraded DNA samples were given in Supplementary Figure 2).

We next investigated whether the quantity of DNA is the predominant factor in generating genomic waves. For the Illumina Infinium assay, the recommended DNA quantity for each sample is 750 ng; however, in many large studies, variable concentrations of DNA are typically available for different samples, and it is generally difficult to control DNA quantity for all genotyped samples accurately. We compared the GCWF for each of the 1444 DNA samples described above with the total quantity of DNA hybridized as estimated from available picogreen DNA concentration measurements (Figure 4C). We observed a positive correlation between DNA quantity and waviness (Pearson correlation coefficient = 0.4371).

To test this effect more directly, we performed a serial dilution experiment and genotyped the same subject on the HumanHap550 array using five different DNA quantities (187.5, 375, 750, 1500 and 2250 ng; Figure 4D). This anonymous subject is chosen since we had sufficiently large amounts of DNA extracted from the same batch of whole blood without whole-genome amplification. We found that when DNA quantity is low (for example, 187.5 ng), the LRR values show patterns of positive

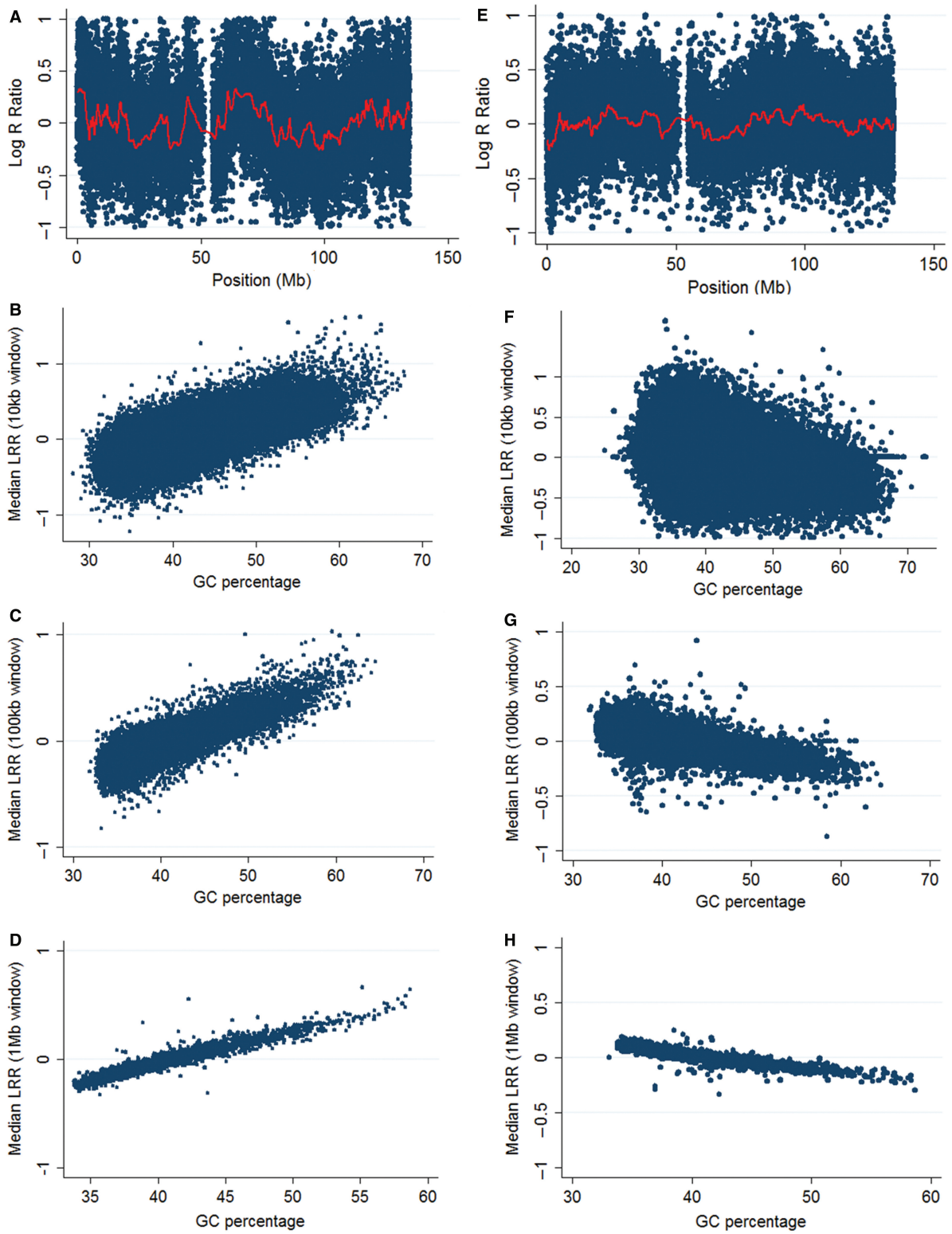

Figure 2. Genomic wave is correlated with GC content. Analysis of the correlation between GC percentage and median values of LRR in nonoverlapping windows across the genome, using 10, 100 kb and 1 Mb window sizes, respectively. (**A**) A sample genotyped by the Illumina HumanHap550 array is chosen and the signal pattern on chromosome 11 is shown. (**B–D**) We observe increasingly higher correlations between these two measures with larger sliding window sizes. The correlation coefficients for panels (B)–(D) are 0.70, 0.85 and 0.96, respectively. (**E**) A sample genotyped by the Illumina HumanHap550 array that has opposite peaks and troughs as (A) is chosen and the signal pattern on chromosome 11 is shown. (**F–G**) We observe increasingly higher correlations between these two measures with larger sliding window sizes.

Table 1. Correlation of wave with genomic features

Genomic features	Sample 1			Sample 2		
	10 kb window	100 kb window	1 Mb window	10 kb window	100 kb window	1 Mb window
GC percentage	0.70	0.85	0.96	−0.47	−0.69	−0.89
Segmental duplication	0.03	0.06	0.17	−0.02	−0.05	−0.17
Gene content	0.07	0.13	0.34	−0.03	−0.09	−0.31
Exon content	0.12	0.36	0.62	−0.09	−0.29	−0.62
Simple repeats	0.18	0.21	0.19	−0.12	−0.23	−0.27
Conserved region	0.04	0.16	0.26	−0.01	−0.12	−0.24

The list of correlation coefficients between median LRR values and several genomic features within nonoverlapping windows. Three different window sizes, including 10, 100 kb and 1 Mb, are examined. Two representative samples with opposite phases are included. The local GC content appears to be best correlated with median signal intensity in sliding windows of varying sizes.

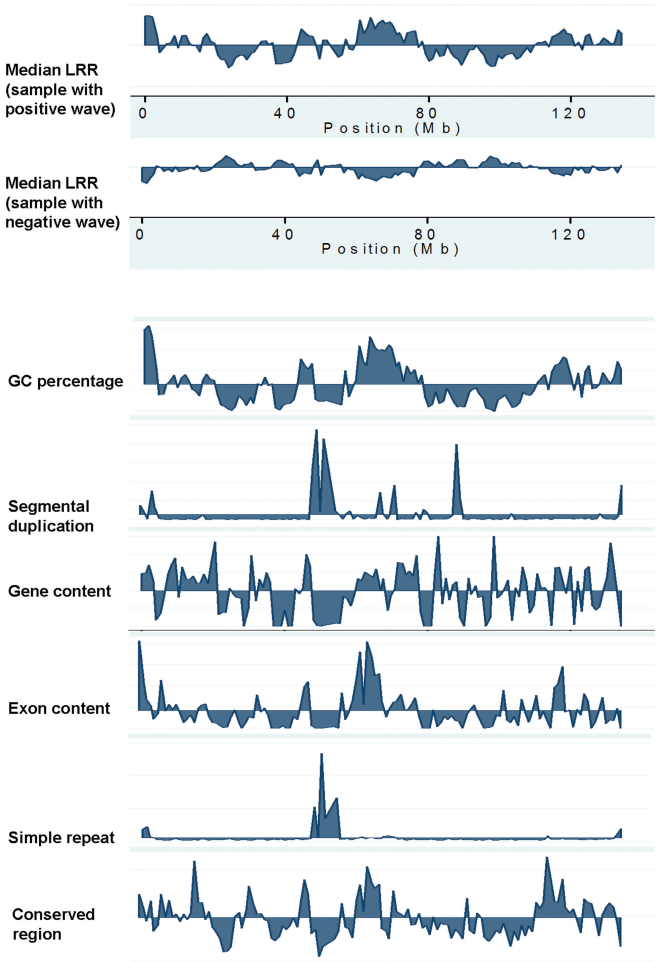

Figure 3. The signal intensity is highly correlated with GC content in sliding windows in chromosome 11. The median LRR values for 1 Mb sliding windows in two samples (one with negative wave and one with positive wave) are plotted against several genomic features, including GC content, segmental duplication, gene content, exon content, simple repeats and conserved region.

waves; conversely, when DNA quantity is high (for example, 2250 ng), the LRR values show patterns of negative waves (Figure 5A). The GCWF values for the five samples with increasing DNA quantity are −0.044, −0.018, 0, 0.012 and 0.023, respectively, reflecting the direction and magnitude of signal fluctuation for each sample. The correlation between GCWF and DNA quantity (logarithm scale) is 0.994, indicating that GCWF is a reliable measure of the DNA quantity, which is usually not well controlled in genotyping experiments.

Adjustment of signal intensity reduces genomic waves

The presence of strong genomic waves creates artificial gains and losses in signal intensities for SNP genotyping arrays, and may lead to spurious CNV calls. To reduce the effect of genomic waves on accurate CNV inference, we developed a regression model to correct and adjust for genomic waves (see Methods section). The essence of our adjustment approach lies in building a regression model for each genotyped sample that correlates the signal intensity at a given marker with GC content within a 1 Mb window centered around the marker, and then calculating the residual for each marker in the array as the adjusted signal intensity. This signal adjustment method does not require a training data set for model construction, and can be applied to many different technical platforms.

To demonstrate the effectiveness of our signal adjustment procedure in reducing genomic waves, we examined the GCWF values for the 1444 genotyped samples before and after signal adjustment. The distribution of GCWF values after adjustment was much tighter around zero, indicating overall lower magnitude of positive and negative waves (Supplementary Figure 3). Before adjustment, the 5th and 95th percentile of GCWF values were −0.0869 and 0.0770, respectively; after adjustment these measures narrowed to −0.0062 and 0.0157. To give some concrete examples, we plotted the signal intensity patterns after wave adjustment for the two samples used in Figure 2, indicating that genomic waves are reduced for both samples with positive and negative waves (Supplementary Figure 4).

To investigate whether our wave adjustment procedure works on other types of arrays, we examined the GCWF values for 270 HapMap samples genotyped by the Affymetrix genome-wide 6.0 arrays. Before adjustment, the 5th and 95th percentile of GCWF values were −0.011 and 0.010, respectively. These HapMap samples do not show strong wavy patterns, since their signal intensities were normalized against a reference clustering file

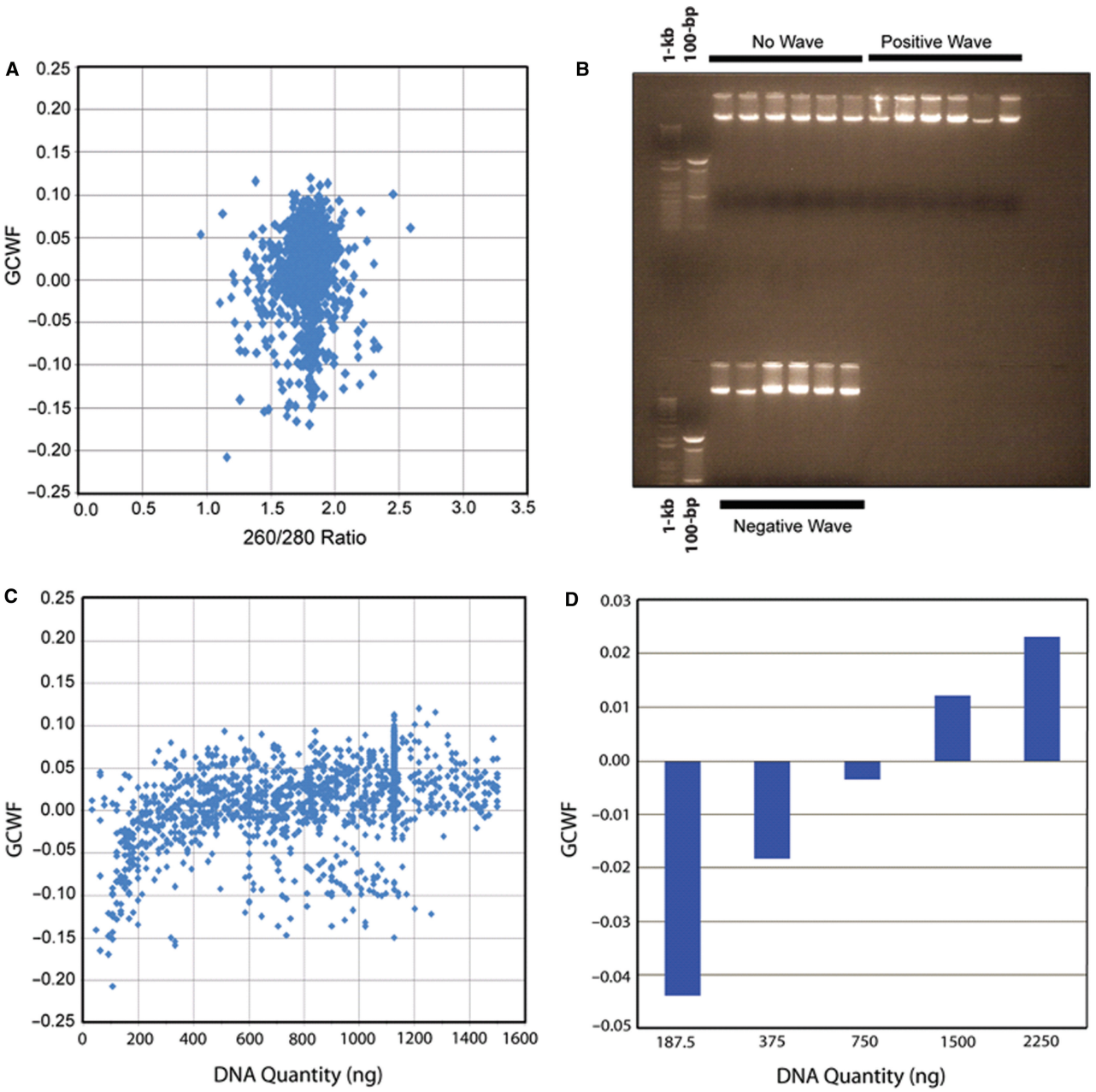

Figure 4. Genomic wave is correlated with DNA quantity, but not quality. (**A**) Plot of GCWF measure against 260/280 ratio for DNA of 1444 neuroblastoma patients (Pearson correlation coefficient = 0.0918). (**B**) Gel electrophoresis assay for 750 ng DNA from 18 genotyped samples (six without waves, six with positive waves and six with negative waves) shows no evidence of DNA degradation, which would appear as smears rather than clear bands (see Supplementary Figure 2 for examples). The largest size marker for the 100-bp ladder lane is 2072 bp, while the largest size marker for the 1-kb ladder lane is 12 kb. (**C**) Plot of GCWF against total quantity of DNA for 1444 samples (Pearson correlation coefficient = 0.4371). Samples with an initial estimated concentration > 100 ng/μl were diluted to 75 ng/μl, explaining the clustering at 1125 ng. (**D**) Plot of DNA quantity versus GCWF for serial dilutions of DNA from a single sample.

constructed from themselves. Nevertheless, after adjustment, these measures narrowed to −0.0046 and 0.005, respectively. Therefore, the wave adjustment procedure can be applied to different technical platforms.

Since Lowess regression has been previously used to reduce genomic waves in array-CGH platforms (4), we investigated its application to SNP genotyping arrays. The method relies on correlating signals from neighboring markers and smoothes the signal intensities continuously along the chromosome. Unlike array-CGH studies, where each CNV may be represented by one or very few clones (probes), the SNP genotyping arrays may reveal CNVs covered by a few SNPs or many hundreds of SNPs. Therefore, the smoothing procedure may not work well for CNVs that are of vastly different sizes. To demonstrate this, we applied Lowess regression with several different window sizes on two samples affected by strong genomic waves (Supplementary Figures 5 and 6). When smaller

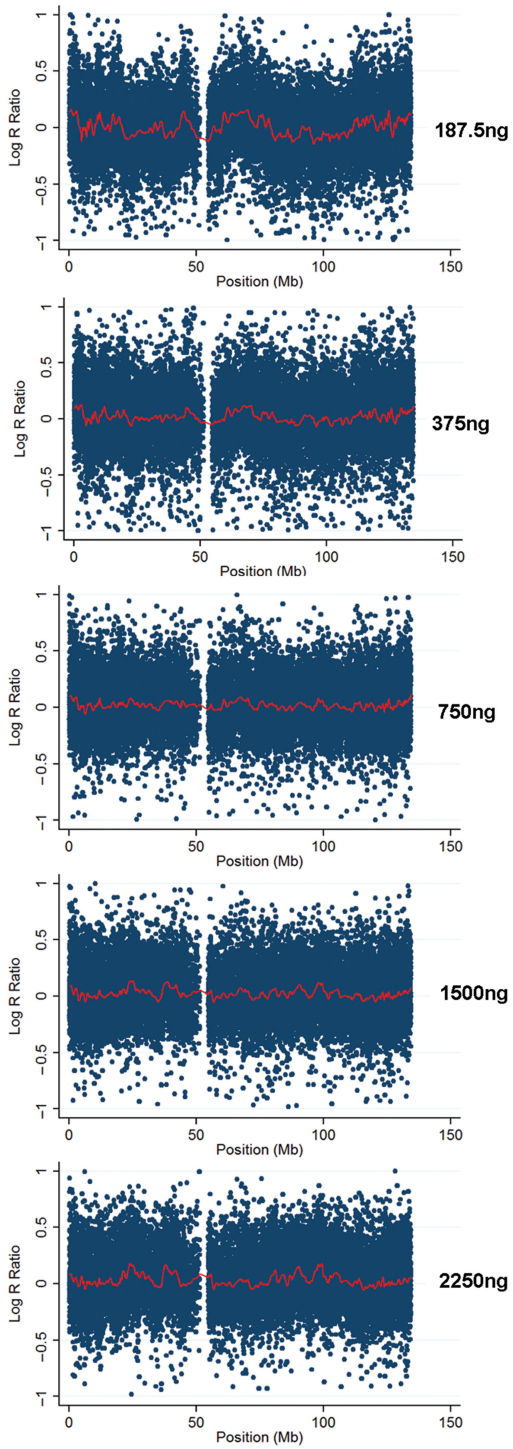

Figure 5. A comparison of signal intensities along chromosome 11 for five duplicated samples with different DNA quantities genotyped by the Illumina HumanHap550 arrays. The LRR signal intensity shows wavy patterns when the DNA quantity deviates from the recommended quantity (750 ng). The directionality of the waves is reversed when increasing amounts of DNA is used for genotyping.

window sizes are used in the Lowess regression, the wavy patterns in signals are successfully eliminated, but the true CNV also disappears from the adjusted signal intensities. When a larger window size is used, the decreased signal intensity in the CNV region is preserved, however, the wave adjustment no longer works well. Therefore our signal adjustment procedure, which operates on each SNP independent of signals in neighboring SNPs, is better suited for SNP genotyping platforms.

Signal intensity adjustment improves CNV detection

To demonstrate that wave adjustment leads to improved CNV detection, we analyzed a recently genotyped three-generation pedigree with 51 individuals (Supplementary Figure 7). The DNA were extracted from fibroblast cell lines, and were then subject to Illumina HumanHap550 arrays without appropriate dilution to ensure optimal DNA quantity used in genotyping. As a result, a large fraction of the samples show visually discernable genomic waves and this hinders accurate CNV inference. Excluding the three samples that failed genotyping, for the remaining 48 samples, the GCWF values range from −0.09 to 0.01, with interquantile range of −0.05 to −0.02. After signal adjustment, all samples have GCWF values between −0.004 and 0.004, with interquantile range of −0.002 to 0.001, suggesting effective elimination of wavy patterns in these samples by our wave adjustment procedure.

One major concern on the wave adjustment is that genuine CNVs may be abolished by the procedure that attempts to flatten wavy signal patterns across the genome. To investigate this issue, we selected six CNV regions and measured their copy numbers in 48 individuals by Q-PCR. The six regions include three deletions (with copy number 0 and 1) and three duplications (with copy number 3 and 4), and each of them is present in 6–29 individuals in the pedigree. We then applied the PennCNV algorithm (18) on the adjusted signal intensity on the 48 individuals, and found that there are only three false negative calls without false positive calls for these six CNV regions (Table 2). We note that the same false negative calls were also made in the absence of signal adjustment. Although wave adjustment did not reduce the false negative rate here (3.5%, 3/85 CNV calls), the adjustment procedure preserves signals from genuine CNVs in these CNV regions and does not introduce false positive calls. To further validate this visually, we plotted examples of signal intensities before and after adjustment for samples with experimentally validated CNVs (Supplementary Figure 8), and show that the signal adjustment procedure reduces the overall magnitude of genomic waves but has little effects on the signals of the true CNV regions.

Another concern on wave adjustment is that it may create false positive CNVs, by making signals in some genomic regions deviate from zero. Since false positive calls are more difficult to assess than false negatives, to investigate this issue, we took advantage of the three large sibships (each with 7–12 siblings) in this pedigree (Supplementary Figure 7). We define 'unique' CNV calls in the sibships as those detected in only one individual (but not in the parents or other family members). A 'unique' CNV can occur if (i) the CNV is a false positive call; (ii) both parents have normal copies, and the offspring has copy number change due to de novo event; or (iii) at least one parent has copy number change but

Table 2. Comparison of computationally generated CNV calls with Q-PCR results on six CNV regions on 48 subjects in a large pedigree

CNV region	Aberrant copy number	Number of subjects with CNVs by Q-PCR	Number of subjects with CNV calls (Number of overlapping with Q-PCR results)	
			With signal adjustment	Without signal adjustment
4p12	3	19	16 (16)	16 (16)
4q13.2	1	6	6 (6)	6 (6)
6q14.1	0, 1	29	29 (29)	29 (29)
6q27	4	10	10 (10)	10 (10)
12p13.31	3	15	15 (15)	15 (15)
15q14	1	6	6 (6)	6 (6)

The CNV calling results are the same before and after signal adjustment on these six regions, and only three false negative CNV calls were made, indicating a false negative rate of 3.5% without or with wave adjustment. In all cases, the exact copy numbers (0, 1, 3 or 4) for CNVs were identified correctly.

due to false negative calls, the copy number change is not detected in the parents and are also not detected in any other siblings. In the context of this large pedigree, since the probability of (i) far exceeds both (ii) and (iii), it is reasonable to treat those 'unique' CNVs as false positive calls. Before signal adjustment, we identified 42 unique CNV calls in the three sibships; after signal adjustment, only 15 unique CNV calls were detected, suggesting a reduction in false positive calls with wave adjustment.

To further validate the performance of the wave adjustment procedure, we analyzed the five samples used in the serial dilution experiment (Figure 5). Before adjustment, the GCWF values for the five samples ranged from −0.044 to 0.023, which correlated with the DNA quantity (Figure 5B). However, after adjustment, the GCWF values for the five samples were 0, −0.001, 0, 0 and −0.003, respectively, supporting the effectiveness of wave adjustment procedure. We next generated CNV calls in the BeadStudio software using the cnvPartition algorithm (the default CNV-calling algorithm developed by Illumina), as well as the PennCNV algorithm (18) without and with wave adjustment (Figure 6). The cnvPartition algorithm is used here as an alternative approach to independently support CNV calls in the absence of experimental data. The CNV calls are represented as colored bars in the corresponding chromosome positions in the graph for each sample and for each algorithm. Using the cnvPartition algorithm, we detected 16, 10, 7, 13 and 11 CNV calls for the five samples with increasing DNA quantity. With the PennCNV algorithm without wave adjustment, we detected 24, 14, 14, 11 and 13 CNV calls for the five samples. Clearly, for both algorithms, the sample with the lowest DNA quantity (and the strongest wavy pattern) had more CNV calls, reflecting possible false positive calls. The concordance between the five samples was generally poor (Supplementary Tables 2 and 3).

We next performed signal intensity adjustment and generated CNV calls by PennCNV (cnvPartition is a built-in algorithm that cannot be applied on adjusted signal intensities). With wave adjustment, the PennCNV algorithm generated 17, 11, 11, 8 and 15 calls, for the five samples with increasing DNA quantity. Despite having fewer CNV calls after adjustment, a higher concordance rate is achieved between five samples (Supplementary Table 4): six CNV regions were consistently detected in all samples, including three CNVs that have identical boundaries in all samples. In comparison, without adjustment, only two and five CNV regions were consistently detected by cnvPartition and PennCNV, respectively. This data set further demonstrates that the wave adjustment procedure reduces spurious CNV calls and increases the concordance rate on duplicated samples.

DISCUSSION

In this article we described the impact of genomic waves on the quality of signal intensity data generated using several high-density SNP genotyping platforms, and investigate the underlying mechanism for their presence. Our experimental analysis illustrated the importance of DNA preparation for the Illumina Infinium platforms in achieving the best genotyping data quality, and this is likely the case for other microarray platforms, such as the Affymetrix platform. In addition, we showed that our computational approach can reduce the effects of genomic waves and salvage data for CNV analysis when wavy signal patterns are present.

We proposed two measures for wavy patterns of any genotyped sample. The main advantages of these measures are that they are not dependent on an external data model or reference sample, and can be applied to many different technical platforms and different array designs. We further showed that the GCWF measure correlates strongly with DNA quantity, and can be used to evaluate the effectiveness of reducing genomic waves. Besides GCWF, we have also tried other measures of autocorrelation, including the correlation of signal intensities of neighboring markers with lag 1, lag 10, lag 100 and lag 500 distance (Supplementary Table 5). Although, the autocorrelation measure does reflect the general trend of correlation between nearby markers (with positive values), they are poor predictors of DNA quantity and were not useful in evaluating the magnitude or directionality of waves.

Although we have applied our method to correct genomic waves on the Illumina platforms, this method may be readily extended to other whole-genome arrays, such as array-CGH with BAC clones or whole-genome oligo–nucleotide arrays. Unlike SNP genotyping arrays, these arrays utilize hybridization of nonpolymorphic probes, however the data normalization techniques (especially the derivation of LRR) could still be applied to such data for reducing variation across markers, and then for building regression models for signal intensity adjustments. Similarly, for SNP genotyping arrays with nonpolymorphic markers, the LRR values can also be derived using the same multisample normalization approach.

Unlike previously described methods that use 'smoothing' techniques by borrowing information from signals of adjacent markers within each sample (4), our method borrows information on the GC content of the

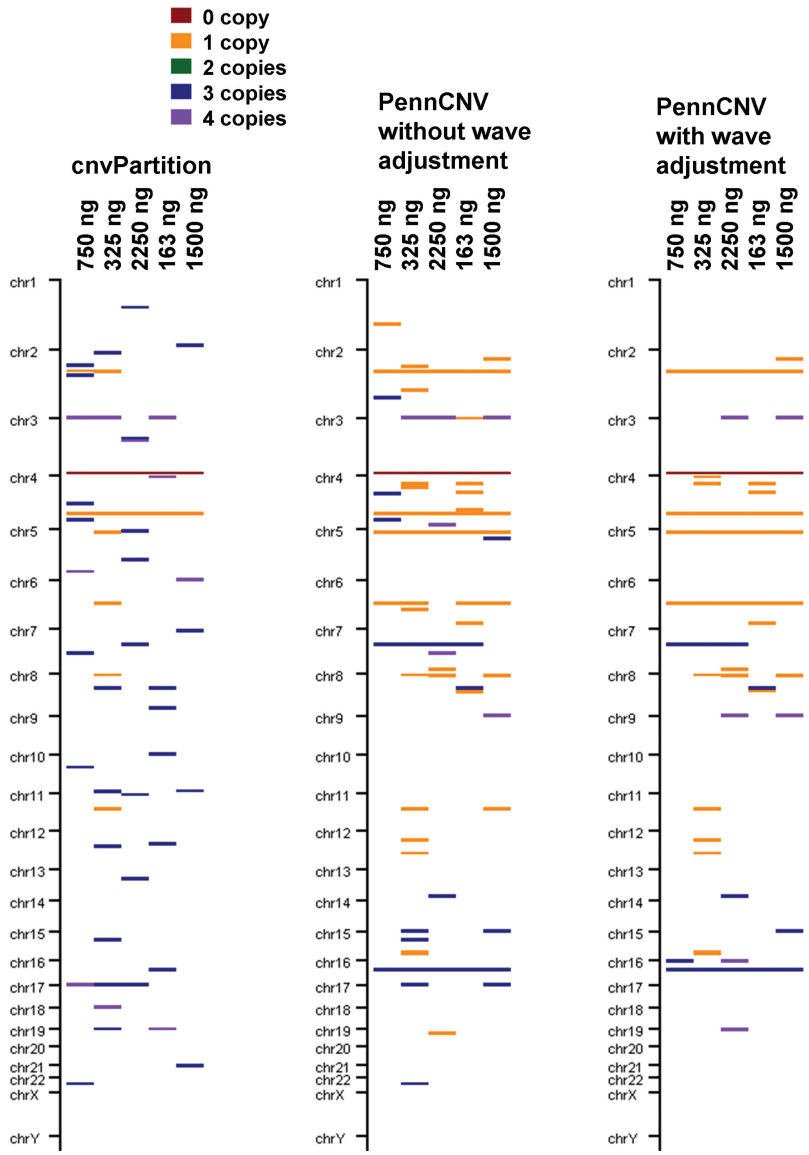

Figure 6. DNA quantity and waves impact CNV calling. CNV calling results from the cnvPartition algorithm, and the PennCNV algorithm without and with signal adjustment, as implemented in the BeadStudio software, on five duplicated DNA samples with different quantities. A 5-SNP threshold was used in PennCNV so that the number of calls was comparable to cnvPartition. The color and the thickness of the bars represent the copy number and the size of the CNV calls, respectively. DNA quantity and signal intensity waves severely affect the accuracy of CNV calling; however, after signal adjustment, higher specificity and higher concordance rate in CNV calls are achieved.

surrounding genomic region. The advantage of our approach is that each marker in the testing sample is independently adjusted regardless of neighboring markers in the same testing sample, so that the boundaries of a true CNV in a testing sample will not be affected by neighboring normal copy markers. In addition, unlike the 'smoothing' techniques that are trained on signal intensity only, our model is trained upon genome-wide GC distributions that are identical for each sample, effectively using more available information for improved model construction. Finally, Lowess regression is computationally intensive and may not scale up well for even a single chromosome with hundreds of thousands of markers. Unlike the smoothing approach, our method only uses a subset of autosome markers that are at least 1 Mb away from each other when building the regression model. While this simple approach circumvents the problem of dependency between neighboring markers, this also raises a concern on the model stability. To investigate this issue, we tested building models with 10 different sets of markers on the two samples used in Figure 2, and found that the regression model parameters are quite stable and are little affected by the use of different markers in the training process. The resulting GCWF values after adjustment are virtually identical.

Our wave adjustment approach shares some similarities with the improved version of the GIM algorithm (6) and with the algorithm proposed by Nannya *et al.* (7), since these approaches all incorporated GC content information into the model building. The improved GIM

algorithm takes into account of the GC percentage of 40 kb of sequence surrounding each SNP, and then uses a robust regression to determine the optimal degree of polynomials in place of least-squares regression. The Nannya *et al*. algorithm applies an empirical formula: the regression model contains both the length and the GC content of the PCR fragment that contains SNP, as well as the squares of these two measures, into a quadratic regression form. Our algorithm has several distinct differences: first, we used linear least-squares regression, due to the simplicity and elegance of the model, and due to the linear relationships observed in Figure 2. Second, we used GC content of 1 Mb window around each marker in the regression model, since this long-range GC content appears to be better correlated with the variation of signal intensities. Third, we used markers that are at least 1 Mb away from each other in the model building to eliminate potential dependence between markers, that is, correlated signal intensities of neighboring markers due to factors unrelated to GC content, such as being covered by the same CNV. Despite these differences, it is clear from all studies that genome-wide GC patterns provide a strong basis for signal adjustment in SNP genotyping platforms.

In summary, we describe the presence of genomic waves in several high-density SNP genotyping platforms and present experiments and analyses to investigate the major causes of these patterns. Moreover, we propose a simple computational procedure that generates adjusted signal intensities to reduce the effects of genomic waves and to improve the accuracy of CNV inference. With the increasing use of SNP genotyping platforms for genome-wide association studies and genome-wide CNV analysis, our method will be useful for taking full advantage of all available samples.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the technical staff at the Center for Applied Genomics, Children's Hospital of Philadelphia for performing the genotyping experiments. We thank the four reviewers in their constructive comments and insightful suggestions on additional experiments.

FUNDING

National Institutes of Health (grant T32 HG000046 to S.J.D.); Institute Development Award to the Center for Applied Genomics from the Children's Hospital of Philadelphia (H.H.); National Institutes of Health (grant R01 MH604687 to M.B.); NARSAD distinguished Investigator Award (M.B.); National Institutes of Health (grant R01 CA124709 to J.M.M.); the Giulio D'Angio Endowed Chair and Alex's Lemonade Stand Foundation (J.M.M.). Funding for open access charge: R01 CA124709.

Conflict of interest statement. None declared.

REFERENCES

1. Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.
2. Scherer,S.W., Lee,C., Birney,E., Altshuler,D., Eichler,E.E., Carter,N., Hurles,M. and Feuk,L. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
3. Feuk,L., Carson,A.R. and Scherer,S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
4. Marioni,J.C., Thorne,N.P., Valsesia,A., Fitzgerald,T., Redon,R., Fiegler,H., Andrews,T.D., Stranger,B.E., Lynch,A.G., Dermitzakis,E.T. *et al*. (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.*, **8**, R228.
5. Frilyand,J., Snijders,A.M., Pinkel,D., Albertson,D.G. and Jain,A.N. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.
6. Komura,D., Shen,F., Ishikawa,S., Fitch,K.R., Chen,W., Zhang,J., Liu,G., Ihara,S., Nakamura,H., Hurles,M.E. *et al*. (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.*, **16**, 1575–1584.
7. Nannya,Y., Sanada,M., Nakazaki,K., Hosoya,N., Wang,L., Hangaishi,A., Kurokawa,M., Chiba,S., Bailey,D.K., Kennedy,G.C. *et al*. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res.*, **65**, 6071–6079.
8. Maris,J.M., Mosse,Y.P., Bradfield,J.P., Hou,C., Monni,S., Scott,R.H., Asgharzadeh,S., Attiyeh,E.F., Diskin,S.J., Laudenslager,M. *et al*. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.*, **358**, 2585–2593.
9. Steemers,F.J., Chang,W., Lee,G., Barker,D.L., Shen,R. and Gunderson,K.L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31–33.
10. Illumina. (2006) Whole-genome genotyping with Sentrix HumanHap550 genotyping BeadChip and the Infinium II assay. *Illumina Technical Bulletin*, Pub. No. 370-2006-017.
11. Peiffer,D.A., Le,J.M., Steemers,F.J., Chang,W., Jenniges,T., Garcia,F., Haden,K., Li,J., Shaw,C.A., Belmont,J. *et al*. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.
12. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E., Siepel,A. *et al*. (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
13. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al*. (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
14. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
16. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
17. Miller,W., Rosenbloom,K., Hardison,R.C., Hou,M., Taylor,J., Raney,B., Burhans,R., King,D.C., Baertsch,R., Blankenberg,D. *et al*. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome browser. *Genome Res.*, **17**, 1797–1808.
18. Wang,K., Li,M., Hadley,D., Liu,R., Glessner,J., Grant,S.F.A., Hakonarson,H. and Bucan,M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.