Published online 7 July 2009 Nucleic Acids Research, 2009, Vol. 37, No. 17 e117
doi:10.1093/nar/gkp559

Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation

Lin Wan[1,2], Kelian Sun[2], Qi Ding[2,3], Yuehua Cui[4], Ming Li[2], Yalu Wen[2], Robert C. Elston[5], Minping Qian[1,2] and Wenjiang J. Fu[2,*]

[1]School of Mathematical Sciences, Peking University, Beijing 100871 China, [2]The Computational Genomics Lab, Department of Epidemiology, Michigan State University, East Lansing, MI 48824, [3]Department of Biochemistry, Michigan State University, East Lansing, MI 48824, [4]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 and [5]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA

Received February 4, 2009; Revised June 15, 2009; Accepted June 16, 2009

ABSTRACT

Affymetrix SNP arrays have been widely used for single-nucleotide polymorphism (SNP) genotype calling and DNA copy number variation inference. Although numerous methods have achieved high accuracy in these fields, most studies have paid little attention to the modeling of hybridization of probes to off-target allele sequences, which can affect the accuracy greatly. In this study, we address this issue and demonstrate that hybridization with mismatch nucleotides (HWMMN) occurs in all SNP probe-sets and has a critical effect on the estimation of allelic concentrations (ACs). We study sequence binding through binding free energy and then binding affinity, and develop a probe intensity composite representation (PICR) model. The PICR model allows the estimation of ACs at a given SNP through statistical regression. Furthermore, we demonstrate with cell-line data of known true copy numbers that the PICR model can achieve reasonable accuracy in copy number estimation at a single SNP locus, by using the ratio of the estimated AC of each sample to that of the reference sample, and can reveal subtle genotype structure of SNPs at abnormal loci. We also demonstrate with HapMap data that the PICR model yields accurate SNP genotype calls consistently across samples, laboratories and even across array platforms.

INTRODUCTION

Human genetic variation studies offer great promise in deciphering the genetics of complex diseases through genome-wide association studies (GWASs) (1), and copy number variation (CNV) studies (2,3), and for both types of study accurate estimation at high resolution of allelic concentration (AC), which refers to the concentration or amount of allelic DNA sequences from the array experiment, is essential for subsequent analysis. Among current platforms, Affymetrix single-nucleotide polymorphism (SNP) arrays have been widely used for SNP genotype calling and CNV inference with low-cost (1–6). Although numerous methods achieve high accuracy (4–18), most studies have paid little attention to the hybridization with mismatch nucleotides (HWMMN) by off-target allele sequences to the probes, which can affect the accuracy, in particular, the accuracy of annotating heterozygous SNPs. In this study, we address this issue and show that HWMMN occurs in all SNP probe-sets and is non-negligible. Ignoring it may lead to biased results and inaccuracy, whereas careful modeling of HWMMN leads to accurate copy number (CN) estimation and SNP genotype calling.

Extensive studies have shown that probe intensities are subject to large variability, and depend on not only the quantities of allelic target sequences, but also the probe-binding affinity. Previously, the perfect match binding of probes to their target sequences has been quantitatively characterized with probe sequence through a positional-dependent-nearest-neighbor (PDNN) or similar model (19–22). Although the PDNN provides a model for probe intensity in perfect match binding and nonspecific binding (20), HWMMN has not been studied thoroughly. We illustrate in this paper that proper modeling of HWMMN can significantly improve the accuracy of AC estimation.

To characterize the HWMMN, we studied the physicochemical properties of sequence binding and generalized the PDNN model to a generalized PDNN (GPDNN)

*To whom correspondence should be addressed. Tel: +1 (517) 353 8623 ext. 113; Fax: +1 (517) 432 1130; Email: fuw@msu.edu model for the binding free energy involving both perfect match hybridization and HWMMN. We then developed a probe intensity composite representation (PICR) model based on a Langmuir-like adsorption principle (19,20) and the GPDNN. In PICR, the intensities of each probe of a given SNP are decomposed into four terms: two terms for specific binding of the two alleles, one term for non-specific binding and an error term. The specific binding depends on the ACs of both alleles and the binding affinity, and the latter can be calculated with the GPDNN model. The parameters of the GPDNN model can be trained with only one array, and the PICR leads to a regression model that yields consistent estimation of ACs by regressing the probe intensities of the probe set for a given SNP on the binding affinity.

The PICR has the following key features. It (i) utilizes probe sequence information, which is invariant and independent of particular samples, and thus only requires a single array for model parameter training; (ii) applies to small sample studies and cross-laboratory studies as a consequence of accurate modeling of DNA sequence binding through the physico-chemical properties in general, consistently achieves high accuracy in genotype calling across samples, laboratories and array platforms, and is robust with data from different laboratories; (iii) determines genotypes based solely on individual data within each array, and hence requires no across-array normalization; and (iv) yields accurate CN estimation and reveals subtle structure of SNP genotype through the ratio of estimated AC to the concentration of the reference sample.

MATERIALS AND METHODS

Data

Data set I.

HapMap trio data set. The 90 samples of the SNP array of the HapMap trio data set were downloaded from the Mapping 100K data set (http://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx) and the 15 samples were downloaded from the Mapping 500K data set (http://www.affymetrix.com/support/technical/sample_data/500k_data.affx).

All annotation files of the corresponding Affymetrix SNP arrays were downloaded from the Affymetrix website. Genotype annotation of the HapMap Project version 2005-03_16a_phaseI was used. The numbers of annotated SNPs on the Mapping 50K Xba 240 Array and Mapping 250K Nsp array by HapMap 2005-03_16a_phaseI were 41 099 and 67 394, respectively.

Data set II.

Multiple copy X-chromosome data. Data from the Affymetrix 100K SNP arrays hybridized with samples of one to five copies of the X-chromosome (1X to 5X) were downloaded from the Affymetrix Sample Data Sets for Copy Number Analysis (http://www.affymetrix.com/support/technical/sample_data/copy_number_data.affx).

Data set III.

Cancer study data. Seven Xba arrays of good quality from one cancer study were included to examine genotyping accuracy in multiarray genotype calling by combining with the HapMap samples on the Xba arrays.

Description of the Affymetrix SNP array design and notation

The Affymetrix SNP 100K Mapping Array consists of a 50K Xba 240 array and a 50K Hind 240 array. Each array uses 10 quartets to interrogate a single dimorphic SNP site with alleles usually denoted A and B. Each quartet consists of two pairs of probes (one pair of perfect match and mismatch probes for each allele) of a 25-mer oligonucleotide sequence designed to either perfectly match the target sequence or mismatch at a particular SNP site: perfect match A, mismatch A, perfect match B and mismatch B, denoted for short by PA, MA, PB and MB, respectively. Among these 10 quartets, some hybridize with sense strands and the rest with antisense strands. The quartets have different shifts (k) of the nucleotide on the probe sequence (k may take the values −4, −3, −2, −1, 0, 1, 2, 3, 4) from the center nucleotide of the probe sequence ($k = 0$ at position 13 of the 25-mer). It is important to note that mismatch probes have one sure mismatch nucleotide at the center position ($k = 0$), and may also have another mismatch at a shift $k \neq 0$ position (23). Similar to the 100K array design, the Affymetrix Mapping 500K SNP Array consists of a pair of 250K Nsp and 250K Sty arrays, but only six quartets are used to interrogate each SNP. For a given SNP, we denote by $(PA^{ks}, MA^{ks}, PB^{ks}, MB^{ks})$ the quartet of four probes with shift k on strand s ($s = +1$ for a sense strand probe and $s = -1$ for an antisense strand probe).

Characterization of HWMMN

One sample (sample name: NA06985_Xba_B5_4000090) of the 50K Xba 240 Array from the Mapping 100K HapMap Trio data set on the Affymetrix website was randomly selected to illustrate the HWMMN. All 29 192 SNPs are homozygous according to the annotation of the HapMap Project (version: 2005-03_16a_phaseI), and all probes of these SNPs were selected to illustrate the critical effect of the HWMMN. Similar results were obtained from samples other than this particular one and are omitted from this report. For convenience, we introduce the following notation for homozygous SNPs: PM1 for perfect match probes with perfect match binding to its target sequences (e.g. PA probes for ‘AA’ SNPs); MM1 for mismatch probes paired with PM1 (e.g. MA probes for ‘AA’ SNPs); PM2 for perfect match probes of the other allele (e.g. PB probes for ‘AA’ SNPs); and MM2 for mismatch probes paired with PM2 (e.g. MB probes for ‘AA’ SNPs). The probe intensities of all SNPs were plotted against the corresponding PM1 probe intensities, the Pearson correlation coefficients were computed and the regression lines of MM1, PM2 and MM2 intensity on the PM1 intensity through the origin were plotted.

GPDNN model of binding free energies of perfect match and mismatch probe sequences

Assume a target sequence S^T is hybridized to a probe sequence S^P with no mismatch nucleotide. Denote the free energy of the sequence binding with nucleotides s_l^P and s_{l+1}^P along the probe sequence, according to the PDNN model for perfect match hybridization, by $E^{\{S^P,S^T\}}$ (19–22):

$$E^{\{S^P,S^T\}} = \sum_{l=1}^{24} \omega_l \lambda(s_l^P, s_{l+1}^P), \tag{1}$$

where ω_l is a weight factor that depends on the position of consecutive bases along the oligonucleotide, s_l^P is the l-th nucleotide of probe sequence S^P and λ is the stacking energy of the pair of nearest-neighbors s_l^P and s_{l+1}^P.

A probe sequence can also bind with sequences of non-complementary alleles through the HWMMN with up to two mismatch nucleotides (see 'Results' section). We thus generalized the PDNN model in Equation (1) to the GPDNN model for bindings with up to two mismatch nucleotides by replacing pair-wise stacking energy with nucleotide triplet stacking energy at mismatch positions, as follows:

(i) One mismatch. Assuming a target sequence S^T is hybridized to a probe sequence S^P with one mismatch nucleotide at position $(13 + j)$ of probe S^P only (note that 13 is the central position of probe S^P), the free energy with one mismatch is denoted by $E_1^{\{S^P,S^T\}}$:

$$E_1^{\{S^P,S^T\}} = \left[\sum_{\substack{l=1\\ l\neq 12+j,13+j}}^{24} \theta_l^j \lambda(s_l^P, s_{l+1}^P)\right] + \kappa^j \delta\left(\{s^P_{12+j} s^P_{13+j} s^P_{14+j}\},\{s^T_{12+j} s^T_{13+j} s^T_{14+j}\}\right), \tag{2}$$

where s_l^P is the l-th nucleotide of the probe sequence S^P, s_l^T is the nucleotide of the target sequence S^T corresponding to the l-th nucleotide of S^P in the hybridization; similar to ω_l in Equation (1), θ_l^j is a position weight factor at position l of the probe sequence with a mismatch nucleotide at position $13 + j$, δ is the free energy of the nucleotide-triplet at the mismatch position $13 + j$ in the hybridization and κ^j is the positional weight factor for the mismatch nucleotide depending on the shift j (κ^0 was set equal to 1 for simplicity).

(ii) Two mismatches. Assume a target sequence S^T is hybridized to a probe sequence S^P with two mismatch nucleotides at positions 13 and $(13 + j)$ $(j \neq 0)$ of probe sequence S^P. Denote the free energy with two mismatch nucleotides by $E_2^{\{S^P,S^T\}}$:

$$E_2^{\{S^P,S^T\}} = E_1^{\{S^P,S^T\}} + \xi^j\left(\{s^P_{12+j} s^P_{13+j} s^P_{14+j}\},\{s^T_{12+j} s^T_{13+j} s^T_{14+j}\}\right), \tag{3}$$

where $E_1^{\{S^P,S^T\}}$ is the free energy that the hybridization would have if there were only one mismatch at position 13, as calculated in Equation (2). The second term ξ^j is an adjustment of the free energy of the nucleotide-triplet due to the additional mismatch at position $(13 + j)$ of the probe sequence S^P.

There is a total of 1974 parameters for the above GPDNN free energy model: 24 positional parameters ω for nucleotide pairs, 16 parameters λ for nucleotide pairs, 22×9 positional parameters θ (22 positional parameters for nucleotide pairs for each mismatch nucleotide × 9 different mismatch positions due to shift), 8 positional parameters κ^j for mismatch nucleotide triplets, 64×3 parameters δ for nucleotide triplets (64 parameters for each nucleotide triplet and each central nucleotide of the triplet has three different mismatch nucleotides) and 192×8 parameters ξ^j ($64 \times 3 = 192$ parameters for each position of the second mismatch nucleotide triplet × 8 different positions of the second mismatch nucleotide at $j \neq 0$). All these parameters of the GPDNN model can be estimated from a single Mapping 50K Xba array.

PICR model for probe intensity

Models similar to the Langmuir adsorption equation have been used to model microarray hybridization by Zhang *et al.* (19,20): the binding affinity ϕ between the probe sequence and the target sequence in hybridization is modeled as a function of the free energy E, by an adsorption function $\varphi(x) = 1/(1 + e^x)$. Based on this adsorption model, the probe intensity I is given by $I = N\varphi(E) + b + \varepsilon$, where E is the binding free energy of the hybridization, N is the concentration, or CN, of the sequences in binding, b is the baseline intensity and ε is the measurement error of intensity (19,20).

Although this probe intensity model combined with the PDNN model has been applied to perfect match hybridization (19), such an approach requires *a priori* genotype information and may only be applied to homozygous SNPs. The difficulty with heterozygous SNPs arises in the mismatch hybridization of allele A and allele B to probe sequences (18), where each perfect match probe also has one mismatch to one of the alleles of a heterozygous SNP, i.e. perfect match probes bind through a perfect match to one allele and through a mismatch to the other allele of the target sequence (see 'Results' section for more details).

To address this, we developed a PICR model that provides a decomposition of the probe intensity of all probes in the probe set of an SNP as a function of ACs and binding affinity based on Zhang's adsorption equation and the GPDNN model. This PICR model requires no *a priori* SNP genotype information; instead, it provides estimates of ACs with a statistical regression model that later on will be used to determine the SNP genotype.

A probe sequence may be hybridized to a target sequence, either of allele A (denoted by S^{TA}) or of allele B (denoted by S^{TB}) for homozygous SNPs, or of both alleles for heterozygous SNPs. Thus, we model probe intensities in each probe quartet of shift k on strand s for a given SNP with two terms for specific binding to alleles A and B of the target sequence, one term for background nonspecific binding and an error term.

The model for the probe intensities of quartet (PA^{ks},MA^{ks},PB^{ks},MB^{ks}) is thus given as:

$$\begin{cases} \vdots \\ I_{PA,ks} = N_A\varphi(E^{\{S^{PA,ks},S^{TA}\}}) + N_B\varphi(E_1^{\{S^{PA,ks},S^{TB}\}}) \\ \qquad +b_{PA,ks} + \varepsilon_{PA,ks} \\ I_{PB,ks} = N_A\varphi(E_1^{\{S^{PB,ks},S^{TA}\}}) + N_B\varphi(E^{\{S^{PB,ks},S^{TB}\}}) \\ \qquad +b_{PB,ks} + \varepsilon_{PB,ks} \\ I_{MA,ks} = N_A\varphi(E_1^{\{S^{MA,ks},S^{TA}\}}) + N_B\varphi(E_{t_k}^{\{S^{MA,ks},S^{TB}\}}) \\ \qquad +b_{MA,ks} + \varepsilon_{MA,ks} \\ I_{MB,ks} = N_A\varphi(E_{t_k}^{\{S^{MB,ks},S^{TA}\}}) + N_B\varphi(E_1^{\{S^{MB,ks},S^{TB}\}}) \\ \qquad +b_{MB,ks} + \varepsilon_{MB,ks} \\ \vdots \end{cases}, \qquad \mathbf{4}$$

where $t_k = 2$ for $k \neq 0$ or $t_k = 1$ for $k = 0$; N_A and N_B are ACs for alleles A and B, respectively; the free energy terms inside $\varphi(x)$ are estimated by the GPDNN model, and the baselines $b_{PA,ks}, b_{MA,ks}, b_{PB,ks}$ and $b_{MB,ks}$ are model intercepts for the probes. Here, we assume the same baseline for probes of a given SNP on the same strand(s). The measurement errors $\varepsilon_{PA,ks}$, $\varepsilon_{MA,ks}$, $\varepsilon_{PB,ks}$ and $\varepsilon_{MB,ks}$ are assumed independent and normally distributed with mean 0 and a common variance.

Parameter estimation for the GPDNN model

The parameters of the GPDNN model, i.e. the effects of the nucleotide pairs for perfect match and the nucleotide triplets for mismatch, and the position effect, were estimated with one randomly selected training sample—a HapMap sample 50K Xba 240 array (NA06985_Xba_B5_4000090). The estimation procedure was implemented by iteratively fitting the probe intensity data to the regression model [Equation (4)] based on the estimated binding affinity through the GPDNN model to minimize the squared loss function $l = \sum (\hat{I}_i - I_i)^2$ with the sum being over all probes on this training array, and estimating the parameters of the GPDNN model through a Monte Carlo method for given free energy of probes calculated from the binding affinity from Equation (4) with estimated CNs and background (see Supplementary Data for the details of parameter estimation). The nonlinear function ϕ of free energy E in Equation (4) may lead to biased estimation of model parameters and thus requires bias adjustment through functional data analysis (see detailed description in Supplementary Data).

Estimation of ACs and total concentration

For a homozygous SNP with genotype 'AA', $N_B = 0$ in theory. The target sequences bind to PA probes through perfect match hybridization, and bind to the PB probes through HWMMN. The HWMMN may make the PB probe intensity well above 0, which would cause probe intensity-based genotype calling methods to yield an incorrect heterozygous genotype 'AB' rather than 'AA'. Similarly, $N_A = 0$ is expected for a homozygous SNP with genotype 'BB'. However, binding to a heterozygous SNP with genotype 'AB' is complex in that (i) PA probes bind with target sequences of allele A and allele B through perfect match hybridization and hybridization with one mismatch nucleotide, respectively, and (ii) PB probes bind with target sequences of allele B and allele A through perfect match hybridization and hybridization with one mismatch nucleotide, respectively. Furthermore, the two mismatch probes (MA and MB) bind with the target sequences through HWMMN with one or two mismatches. In theory, at an SNP locus with no insertion or deletion, both ACs N_A and N_B of an 'AB' SNP are expected to be positive and close to each other ($N_A \approx N_B$). We refer to the sum of the two ACs $N_{\text{total}} = N_A + N_B$ as the total concentration. The total concentration $N_{\text{total}} = 0$ for a homozygous deletion, and $N_{\text{total}} > 0$ otherwise.

Equation (4) provides a general expression for probe intensity by ACs N_A and N_B for SNPs of all genotypes, 'AA', 'BB' and 'AB', where the binding free energy and affinity are fixed function of probe sequences, and are precalculated by the GPDNN model. For a given SNP, the ACs N_A and N_B are estimated by a regression model [Equation (4)] using all the probe intensities of the SNP. In doing so, the regression intercept (nonspecific binding) is SNP dependent.

CN estimation for cell-line data of multiple copies of the X-chromosome

To assess the performance of the PICR in CN estimation, we compared the estimated concentration with the true CN of the multiple copies of the X-chromosome in Data Set II. The total concentration (N_{total}) of each SNP was first computed for each sample based on the PICR model. These N_{total} of the SNPs on each array were then normalized across samples by multiplying by a sample-specific constant such that the normalized median N_{total} of each sample achieved the same level across samples. To examine the CN estimation at each SNP, the Pearson correlation coefficient between the normalized N_{total} and the true CN (1, 2, 3, 4 and 5) of the five samples was calculated as a measure of the relative agreement.

For comparison with the estimated ACs based on mean intensity of PM probes, the probe intensities were first normalized with the quantile normalization method (24) across samples. Second, for each normalized sample and a given SNP, the AC of allele A (B) was assigned a value of the mean intensity of PM probes of allele A (B). The estimated total AC of each SNP was calculated as the sum of the two estimated ACs of alleles A and B.

Comparison of genotyping between PICR, CRLMM and the mean-intensity method via the receiver operating characteristic (ROC) curve

At a given SNP, the PICR genotype calling was based on the estimated ACs. First, the two-dimensional plane (N_A, N_B) was divided into three regions by two lines $N_B = cN_A$ and $N_B = (1/c)\ N_A$ through the origin and a parameter $c > 1$: the region above the line $N_B = cN_A$ for 'BB' SNPs, the region below the line $N_B = (1/c)\ N_A$ for

'AA' SNPs and the region between the two lines for 'AB' SNPs. This yielded zero no call. Using the same training Xba HapMap sample (see parameter estimation of the GPDNN), an optimal value of $c = 3.5$ was selected to achieve the minimum error rate of all SNPs with three genotypes annotated against the HapMap gold-standard genotype. The same parameter $c = 3.5$ was then applied to all nontraining samples (including other platform of Affymetrix SNP arrays).

To compare the PICR with the CRLMM and the mean-intensity based genotype-calling method via the ROC curve, we assessed the performance of the genotype calling methods on heterozygous SNPs on all 90 HapMap Xba samples. The heterozygous SNPs annotated by the HapMap project on all 90 samples were taken as the positive samples, while the remaining homozygous SNPs were the negative samples. The percentage of the predicted 'AB' SNPs by the genotype calling method out of the total number of the heterozygous SNPs was the true positive (TP) rate, while the percentage of the predicted 'AB' SNPs out of the total number of the homozygous SNPs was denoted as the false positive (FP) rate. For the purpose of comparison, the ROC curve of the PICR was obtained by varying the slope parameter $c > 1$ in the clustering of the SNPs, and the ROC curve of the mean-intensity method was obtained similarly. Since the CRLMM calls the genotype of a given SNP by assigning it to the SNP cluster that has the minimum distance from its cluster center to the given SNP among the three genotypes 'AA', 'AB' and 'BB' with distances d_{AA}, d_{AB} and d_{BB}, respectively, the cutoff value of the ratio $\rho = d_{AB}/\min(d_{AA}, d_{BB})$ was used to call the heterozygous genotype for small ρ. The ROC curve was obtained by varying the cutoff for ρ from 0.1 to 10. The distances and the ratio ρ were calculated with the package 'oligo_1.4.0' of the CRLMM program in the BioConductor 2.2, available at: http://www.bioconductor.org/packages/2.2/bioc/html/oligo.html.

Comparison of genotype calling between PICR and CRLMM via HapMap samples

We used the R program 'crlmm' in the BioConductor 'oligo' package for genotype calling by the CRLMM method on the HapMap samples of 90 Xba arrays and 15 Nsp arrays with the HapMap standard genotype. This program has been trained previously and is ready to proceed with a built-in quantile normalization procedure (24), followed by genotype calling on multiple arrays. However, we noticed that the CRLMM program does not work for single array input, but requires at least two array CEL files to generate valid genotype calls. To ensure a fair comparison between the CRLMM method and our PICR-based single-array genotype-calling method, we conducted genotype calling with the CRLMM twice. The first time we used two arrays as input and repeated this procedure 45 times on the 90 Xba arrays, and similarly seven times on the 15 Nsp arrays (six pair-wise genotype calls and one three-array genotype call). The second time we used all 15 arrays as input and called the genotypes together. The genotyping accuracy was reported separately.

We conducted genotype calls of all 90 Xba arrays separately with the PICR-based genotype calling method without across-array normalization. We identified bright spots in a few arrays and excluded those probe intensities in the bright spots from the PICR regression model for quality control (see 'Results' section). We also noticed that excluding those probe intensities in the bright spots by setting them to missing values in the CRLMM genotype-calling procedure led to significantly lower accuracy. We thus kept all probe intensities in the CRLMM procedure for genotype calls. The accuracy was calculated by comparing the genotype calls against the gold-standard HapMap genotype.

To further examine the multiarray genotype-calling procedure by the CRLMM, we also assessed the performance of the CRLMM with multiple arrays for simultaneous genotype calling with a varying number of HapMap samples and our study samples on the Xba arrays. Accuracy of the CRLMM was assessed through the genotype calls on the HapMap samples, and the consistency of the CRLMM performance was assessed with the accuracy obtained by varying the number of HapMap samples.

RESULTS

The effect of HWMMN in Affymetrix SNP arrays

On careful examination, we found that probes of Affymetrix SNP arrays can contain non-negligible hybridization by sequences of the noncomplementary alleles with one or two mismatch nucleotides. We illustrate with homozygous SNPs. It is known that for a given homozygous SNP of genotype 'AA', all PA probes in the corresponding probe-set are perfect matches to the target sequences. Other probes, which have been deemed to reflect array background, bind the target sequences through hybridization with one or two mismatch nucleotides. MA probes have one mismatch nucleotide at position 13, PB probes have one mismatch nucleotide at position $13 + k$, and MB probes have two mismatch nucleotides at positions 13 and $13 + k$, where $k \neq 0$ is the nucleotide shift in the array design. Probes interrogating 'BB' SNPs work similarly. We examined scatter plots of probe intensities MM1 versus PM1, PM2 versus PM1, and MM2 versus PM1 in Figures S1(A), S1(B) and S1(C), respectively, in the Supplementary Data, from one sample of the Mapping 50K Xba 240 array to illustrate the critical effect of the HWMMN (see 'Materials and Methods' section for the notation of PM1, PM2, MM1 and MM2). Also shown in each plot are the Pearson correlation coefficient and a linear regression slope of the probe intensity in the plot. Similar effects were observed with other samples as well and are thus omitted.

The comparison of MM1 intensity to PM1 intensity, PM2 intensity to PM1 intensity and MM2 intensity to PM1 intensity in Figure S1A–C indicates that MM1 intensity and PM2 intensity are comparable to PM1 intensity, indicated by a regression slope >0.2 and highly significant positive correlation; but MM2 is not comparable to PM1

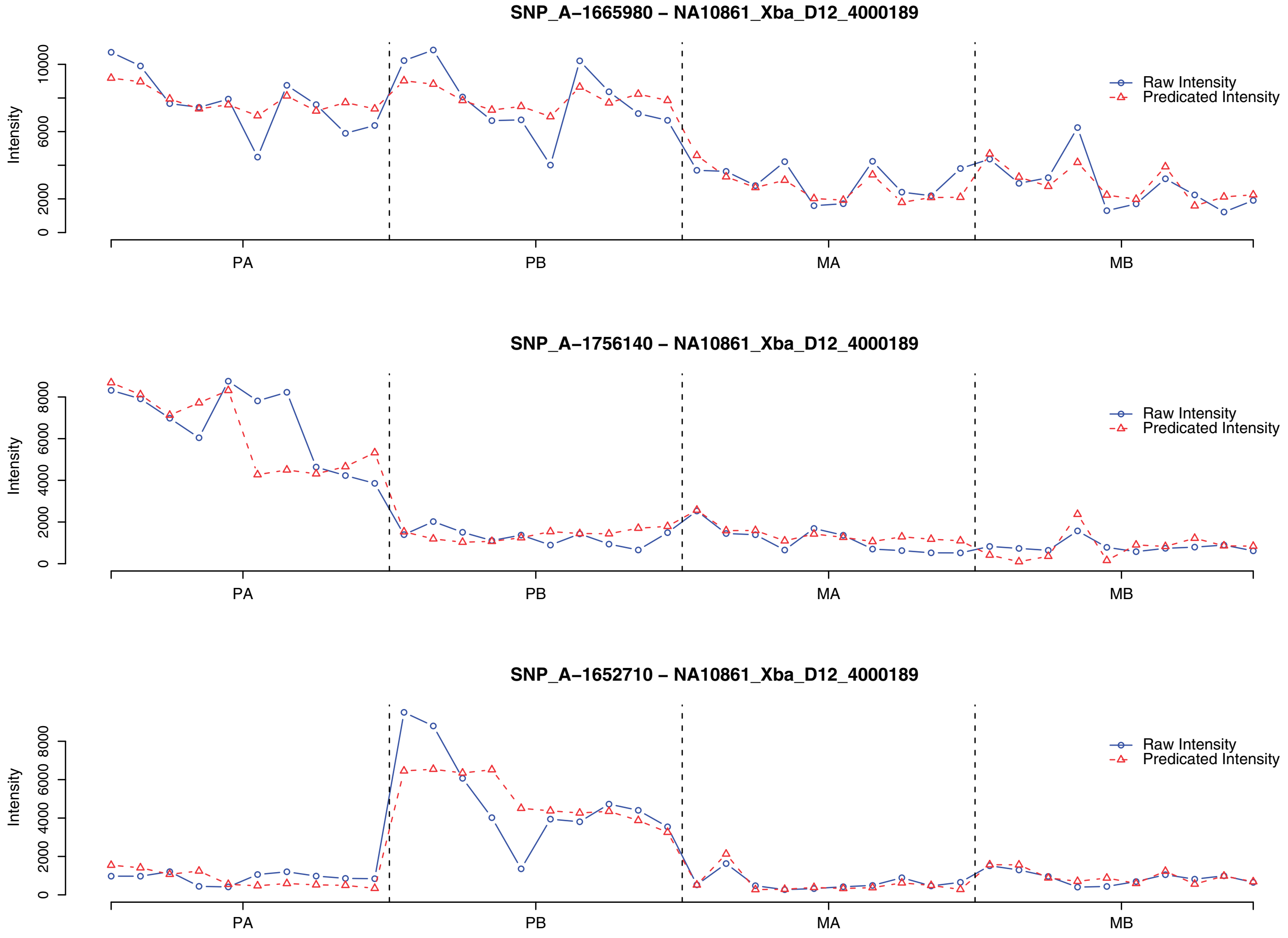

Figure 1. Comparison between raw intensities and predicted intensities by PICR of the 40 probes for three randomly selected SNPs of different genotypes on one randomly selected non-training Xba sample (NA10861_Xba_D12_4000189). (Top) SNP_A−1665980 of 'AB' type. (Middle) SNP_A−1756140 of 'AA' type. (Bottom) SNP_A−1652710 of 'BB' type.

with a regression slope <0.07 except that the slope was around 0.68 at the center position ($k = 0$), in which case there was only one mismatch nucleotide to the target sequence. This disparity resulted from lower intensities of MM2 probes because of one more mismatch nucleotide than the MM1 and PM2 probes. This result implies that PM2, MM1 and MM2 probe intensities are mainly attributable to specific-binding in HWMMN rather than background nonspecific-binding. Had the background nonspecific binding been dominant, the difference between the slopes of MM2 and MM1 or PM2 would not have been so large.

Accurate estimation of probe intensity and ACs with PICR

We developed the PICR model to characterize the complicated hybridizations of probes for Affymetrix SNP arrays (see 'Materials and Methods' section). Figure 1 illustrates the raw intensities of three randomly selected SNPs of genotypes 'AB', 'AA' and 'BB' from a nontraining sample compared to the estimated intensities by the PICR in Equation (4). It is shown that the PICR fitted the probe intensity data well. (For more examples, see Supplemental Data).

Figure 2 shows the estimated ACs of SNPs obtained by the PICR in a randomly selected nontraining sample (NA07056_Xba_A11_4000090). A total of 41 099 SNPs with HapMap annotation were plotted with genotypes in different colors. We found that 'AA' SNPs had a small estimated AC N_B relative to N_A with the mean around 0. A similar observation was true for 'BB' SNPs. Heterozygous SNPs had large positive values (close to each other) for both N_A and N_B. These observations indicate nearly unbiased estimation of ACs. The nearly equal distributions of the estimated total concentration among SNPs of different genotypes (Figure S2) indicate, at least partly, that the estimation of the total concentration was accurate.

Notice the large scale and large variance of the estimated ACs (N_A or N_B) within each SNP genotype (Figure 2). It should be noted that the estimated ACs reflect the quantities of DNA fragments after PCR amplification, rather than before the amplification, and are thus at a relatively larger scale than true CNs, and may still contain variation from the PCR amplification process in array preparation and cross-hybridization with off-target sequences, etc.

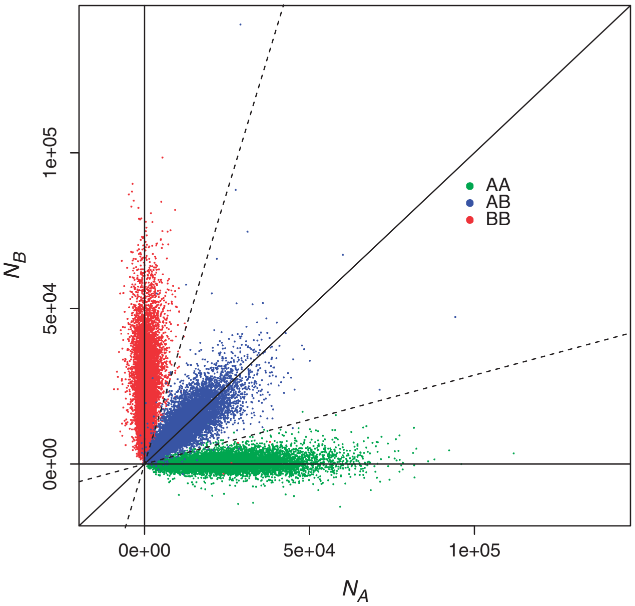

Figure 2. Scatter plot of the allelic concentration estimated by PICR of all SNPs on one randomly selected non-training sample (NA07056_Xba_A11_4000090) with colors indicating HapMap genotype annotation.

Comparison between total concentration and true CN and validation

To validate our estimation of the total concentration by comparing to the CN, we applied the PICR to the cell-line data of known true CNs with one to five copies of the X-chromosome, which were analyzed as a benchmark in the CARAT method of CN estimation (12).

After the across-array normalization on the total concentration (N_{total}) of SNPs to achieve the same median across samples (see 'Materials and Methods' section), we examined the agreement between the total concentration and the true CN through the Pearson correlation coefficient (r). Table 1 displays the percentiles of the correlation coefficient, and shows that 90% of the SNPs on the X-chromosome have a very high correlation ($r > 0.935$), indicating high agreement between the total concentration and the true CN.

In addition, we compared the total concentration to the true CN using sample 2X as the reference as in the CARAT study (12), and plotted the total concentration of all 2361 SNPs on the X-chromosome of the samples 1X, 3X, 4X and 5X in Figure 3. It can be seen that the estimated total concentrations are relatively close to the true fold of sample 2X (dashed lines), with slight underestimation bias for the samples of large CNs (4X and 5X). In contrast, the probe intensity-based CARAT method did not yield CNs as close to the true CN of the X-chromosome, even after the exclusion of about 20% of the SNPs that would not fit the CARAT model well [see Figure 1(a–d) and p3 of Huang *et al.* (12) for details].

Furthermore, we examined the ratio of the relative concentration [the total concentration of the sample at each SNP to that of the reference sample (2X)] to the true CN through boxplots of the log-ratios across all 2361 SNPs on the X-chromosome for each sample, and compared them between the PICR method and the mean intensity method. It is shown in Figure S3 (Supplementary Data) that in all samples (1X, 3X, 4X and 5X) the mid 50 percentiles (25–75 percentiles) of the log concentration ratio by the PICR method contain the true CN level (log-ratio = 0), while none of the samples of the log concentration ratio obtained by the mean intensity method does. While both methods were biased to overestimation for the small CN sample (1X) and to underestimation for large CN samples (3X, 4X and 5X), the PICR method had a much smaller bias—although it had a slightly larger variance than the mean intensity method. The known true CN further allows a comparison of the estimates between the PICR and the mean intensity methods through the mean squared error (MSE), where $\text{MSE} = (\sum_i (R_i - TR_i)^2)/N$, in which R_i is the relative concentration by the PICR or the mean intensity method, and TR_i is the true relative concentration for SNP i of each sample and the summation is over all N SNPs within the sample. Table 2 displays the known true CN, and the square-root of the MSE by the PICR and by the mean intensity method after excluding a small percentage of extreme values in the relative concentrations (the top and bottom 5%) from each sample, which left 2123 SNPs out of the total of 2361 SNPs on the X-chromosome from both Xba and Hind arrays. It is seen that the PICR had a smaller MSE and achieved overall better estimation of the CN through the use of relative concentration.

Table 1. Percentiles of the Pearson correlation (r) between the total concentration and true CNs[a]

Mean	5%	10%	25%	50%	75%	90%	95%
0.9688	0.8989	0.9351	0.9705	0.9868	0.9940	0.9971	0.9983

[a] The total concentrations were computed at each of all 2361 SNPs on the X-chromosome of samples 1X, 2X, 3X, 4X and 5X (both Xba and Hind arrays).

AC-based SNP genotype calling by PICR and comparison with other methods

We developed a genotype-calling method based on the accurate estimates of ACs N_A and N_B by the PICR. It is based on a statistical decision—whether one of the ACs is zero for a given SNP: a SNP is of 'AA' type if $N_B = 0$; 'BB' type if $N_A = 0$; or 'AB' type if both $N_A > 0$ and $N_B > 0$. To determine the clusters of the SNPs, the ACs were plotted as shown in Figure 2, and two lines were drawn through the origin (0, 0). The slopes of the two lines (c and $1/c$) were trained with one training sample to minimize the genotyping error rate for that sample (see 'Materials and Methods' section).

We assessed the performance of the AC-based genotype-calling method by the PICR with the same slopes for clustering the SNPs of all 90 Xba HapMap samples and 15 Nsp HapMap samples using the gold-standard HapMap annotation. Table 3 displays the summary

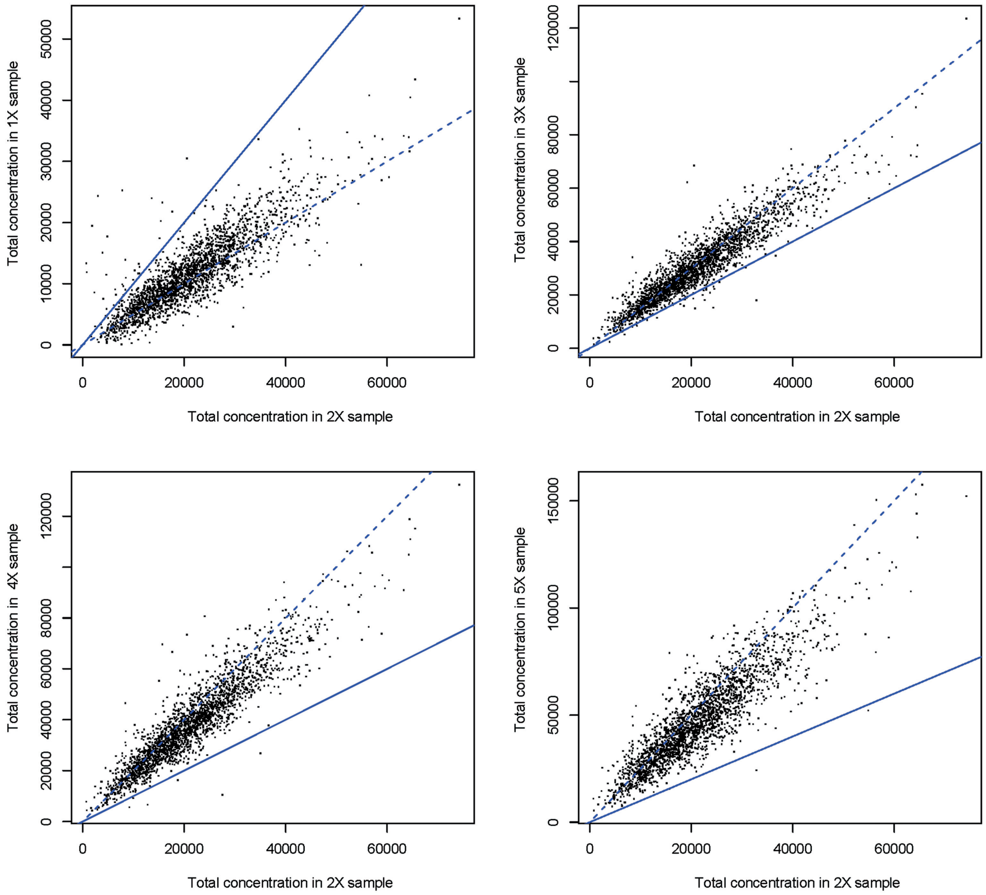

Figure 3. Scatter plot of the total concentration estimated by PICR of all SNPs on the X-chromosome. (Upper left) Sample 1X versus sample 2X. (Upper right) Sample 3X versus sample 2X. (Lower left) Sample 4X versus sample 2X. (Lower right) Sample 5X versus sample 2X. The solid line in each panel represents the diagonal true concentration of sample 2X, and the dotted line represents the theoretical concentration of the sample (1X, 3X, 4X or 5X) relative to sample 2X. The closeness of the estimated concentration to the theoretical concentration illustrates the validity of this copy number estimation method by PICR. A slight bias to underestimation of the concentration was observed in the large copy number samples (4X, 5X).

statistics of the genotype-calling accuracy separately for the Xba arrays and the Nsp arrays. This genotype-calling method yielded zero no-calls and high accuracy with mean 99.66% (SD = 0.0034) correct genotyping over 90 Xba arrays and mean 99.16% (SD = 0.0014) over 15 Nsp arrays. The consistently high accuracy across arrays, laboratories (two different data sets), and array platforms (100K and 500K arrays) indicates the robustness of this method, particularly as it was achieved by training on a single Xba array.

We also examined the sensitivity and specificity of the PICR genotype-calling method using the ROC curve. We took the 90 HapMap samples on the Xba arrays and used the heterozygous SNPs by the HapMap annotation as the positive set and the homozygous SNPs by the HapMap annotation as the negative set to examine the sensitivity and specificity. In comparing with the mean-intensity genotype-calling method and the CRLMM, we found that the PICR achieved very high sensitivity and specificity, as shown in Figure 4A and B.

We further compared the PICR method with other SNP genotype calling methods using the 90 HapMap Xba arrays and 15 Nsp arrays. Among the current genotype calling methods, the CRLMM method (17) was developed based on RLMM (14) and further achieved improvement by taking into consideration the probe sequence effect (the nucleotide effect and position effect) and the target sequence effect (the fragment length and GC content), and was reported to be robust and more accurate than RLMM and BRLMM (6). The Birdseed genotype-calling method (5) requires a fairly large number (> 44) of arrays for genotype calling. We thus excluded the RLMM, BRLMM and Birdseed methods in this comparison study.

Table 3 shows the comparison of the PICR and the CRLMM methods with summary statistics of the genotype-calling accuracy against the gold-standard HapMap genotype calls. Figure 5 shows a comparison of the error rates of genotype calling of each Xba array between the PICR method and the CRLMM method by pair-wise genotype calls. It is seen that the PICR method outperformed the CRLMM method with pair-wise genotype calling in all 90 Xba arrays except for two, and reduced the error rate of the CRLMM method by more than 60% on average.

We note that the CRLMM has been reported to yield accurate genotype calls if multiple arrays are genotyped simultaneously. To further investigate the effect of the number of arrays in genotype calling by the CRLMM, we first genotyped the 15 HapMap Nsp arrays together by the CRLMM as shown in Table 3, and second genotyped a pool of 8 Xba arrays of HapMap samples and our study samples as shown in Table 4. We examined the quality of the seven study arrays and did not see any quality issue, as shown in Figure S4 in the Supplementary Data. We observed that although the genotype-calling accuracy of the CRLMM improved with 15 Nsp arrays pooled together (Table 3), its accuracy also varied from 99.0% to 99.8% with the number of HapMap Xba arrays in the pool (Table 4). It is known that the CRLMM was trained with the HapMap samples and thus its accuracy should not be examined solely on the HapMap samples, to avoid the double-dipping problem (using the same data set for both model training and testing). The pooling of arrays from different sources avoided the double-dipping problem and discovered the varying accuracy of the CRLMM.

Revealing subtle genotype structure of SNPs

In the previous section, we showed that the PICR yielded accurate estimation of the total concentration, whose ratio to the reference sample was consistent with the true CN. Here we further demonstrate that the PICR can reveal subtle genotype structure of SNPs through the multiple copies of the X-chromosome data.

We obtained the relative ACs of each SNP by taking the ratio of the ACs of each SNP to the corresponding total concentration of the reference sample (2X), and then plotted the relative ACs of all 1204 SNPs on the X-chromosome of the Xba array. Figure 6 compares the clusters of all 1204 SNPs with the scatter plot of the relative ACs of each sample (1X, 3X, 4X or 5X) by the PICR method in the right panels and by the mean intensity method in the left panels. The clustering of the SNPs, indicated by different colors, was achieved by assigning each SNP's ACs (N_A, N_B) to the closest possible genotype through the Euclidean distance. For example, assuming no insertion and no deletion, the SNPs on the sample 4X have a total of five possible genotypes 'AAAA', 'AAAB', 'AABB', 'ABBB' and 'BBBB', corresponding to the relative true ACs: (2, 0), (1.5, 0.5), (1, 1), (0.5, 1.5) and (0, 2) to the reference sample (2X). The relative true ACs having the smallest distance to the SNP at $(N_A/N_{total,2}, N_B/N_{total,2})$ determines the SNP genotype, where $N_{total,2}$ is the total concentration of the reference sample 2X. Figure 6 shows that through the estimated ACs by the PICR method, most SNPs were assigned into clusters of different genotypes. In contrast, the mean intensity method yielded severely biased results and the SNPs were hardly distinguishable except for the 1X sample. This demonstrated that the PICR model may reveal subtle structure through accurate estimation of the ACs.

Table 2. Square root of Mean Squared Error (RtMSE) of the relative concentration estimated by PICR and mean intensity methods by sample of X-chromosomes

Sample	1X	3X	4X	5X
True CN ratio	0.5	1.5	2	2.5
RtMSE By PICR	0.1053	0.1766	0.2738	0.4730
RtMSE by mean intensity	0.2250	0.2613	0.5650	0.8467

We also noticed that the relative ACs for samples 3X, 4X and 5X to sample 2X were underestimated by both the mean intensity method and the PICR method, compared to the true CNs, and more severely by the former than the latter. This also confirmed the previous observation in the boxplot of the log-ratio of the relative concentration in Figure S3. We believe that the large quantity of the large CN X-chromosome samples (4X, 5X) produced a high concentration of the target sequences and may have led to saturation and loss of efficiency of probe binding to some extent. We are working on a model for correcting the saturation in array hybridization.

We also noticed that the ACs estimated by the PICR method had a larger variance than those estimated by the mean intensity method, which can be explained by the fact that the mean intensity has a variance reduced by a factor $(1/n)$, where $n = 10$ is the number of PM probes of each allele on the Xba arrays. We believe that more sophisticated statistical techniques may help to reduce the variance of the PICR estimates.

DISCUSSION

Hybridization with mismatch nucleotides by off-target sequences to array probes has been noticed to decrease the accuracy of probe intensity-based CN estimation in various platforms of microarrays (18,25,26). Correct modeling of HWMMN is thus critical and can improve the accuracy of copy number estimation. In this work, we observed a strong effect of HWMMN and have provided a quantitative characterization. We studied oligonucleotide sequence binding through binding free energy with a GPDNN model and binding affinity, based on Zhang's affinity function, and characterized probe intensities in

Table 3. Comparison between PICR and CRLMM in genotype-calling accuracy against the gold-standard HapMap genotype

Sample	Genotype-calling method	Mean	SD[a]	Median	5%[b]	95%[c]
90 Xba arrays	PICR (single array)	0.9966	0.0034	0.9975	0.9920	0.9987
	CRLMM (45 pairwise)	0.9923	0.0021	0.9924	0.9886	0.9954
15 Nsp arrays	PICR (single array)	0.9916	0.0014	0.9921	0.9894	0.9930
	CRLMM (six pairs + one triplet)	0.9806	0.0032	0.9813	0.9747	0.9844
	CRLMM (all arrays together)	0.9962	0.0003	0.9963	0.9958	0.9967

[a] Standard deviation.
[b] 5th percentile.
[c] 95th percentile.

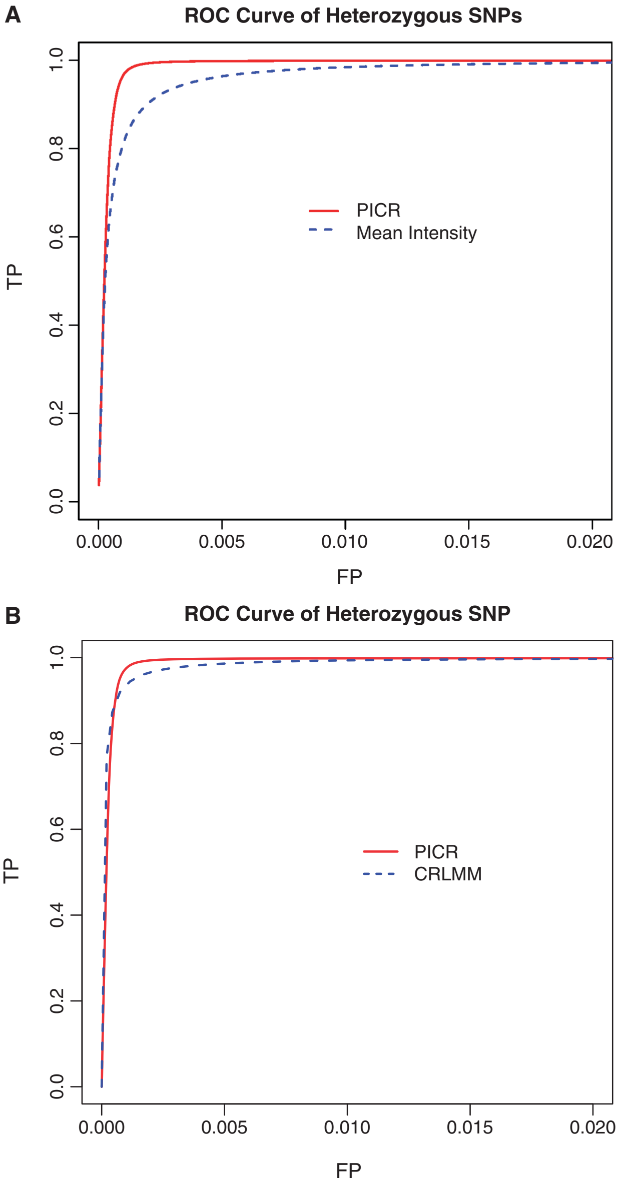

Figure 4. The ROC curves for genotype calling on heterozygous SNPs versus homozygous SNPs by the PICR, the mean intensity and the CRLMM methods. The positive set contains all heterozygous SNPs by the HapMap annotation on the 90 HapMap Xba samples, while the negative set contains all homozygous SNPs on the same arrays. The ROC curves were obtained by varying the slopes c and $1/c$ of the lines used for clustering the SNPs into heterozygous or homozygous SNPs in the PICR and the mean intensity method, and by varying the cutoff value of the distance ratio ρ from the given SNP to the genotype centers in the CRLMM method (see 'Materials and Methods' section). (**A**) Comparison of ROC curve between the PICR and the mean intensity method. (**B**). Comparison of ROC curve between the PICR and the CRLMM method.

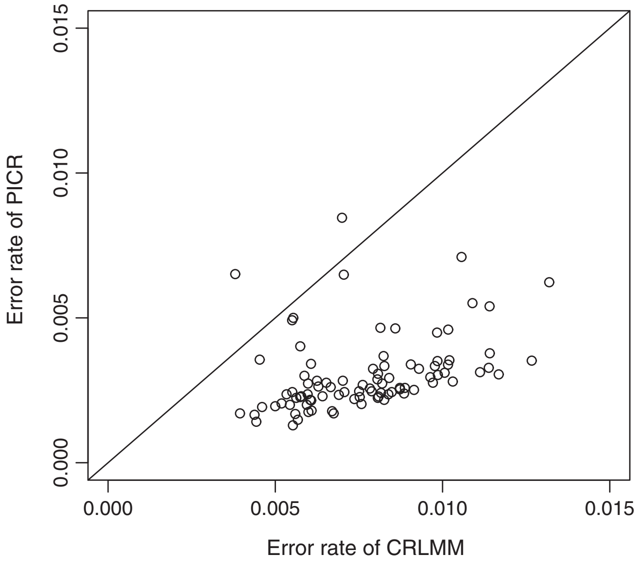

Figure 5. Comparison between the PICR method and the CRLMM method in genotype calling error rate using the HapMap genotype as the gold standard on 90 Xba arrays of the HapMap sample. The genotype calling was conducted with a pair of arrays each time by the CRLMM method, and with a single array by the PICR method.

Table 4. Genotyping accuracy by CRLMM on HapMap samples in simultaneously calling 8 Xba arrays with varying number of HapMap samples

HapMap + Study samples	1+7	2+6	3+5	4+4	5+3	6+2	7+1
Mean accuracy (%) On HapMap samples	99.0	99.2	99.5	99.6	99.7	99.7	99.8

both perfect match hybridization and HWMMN through the PICR model. We then developed a method of AC estimation based on the PICR through a statistical regression. We further developed a genotype-calling method based on the estimated ACs from the PICR. The consistent accuracy of our AC-based genotype-calling method across different laboratories and different array platforms suggests that the PICR accurately characterizes the complex oligonucleotide sequence hybridization (see Figures 1 and S1) and yields nearly unbiased estimation of AC.

Former belief has held that perfect match probe intensities are proportional to the CNs, and has led to their use as a surrogate in many studies. However, this notion has been questioned for its accuracy, and a correction has been suggested using the CN and probe-binding affinity (8,19,20,27). In fact, the PICR is the first model that rigorously formulates the relationship of intensities among perfect match and 'mismatch' probes with AC, sequence-specific binding affinity and background nonspecific binding. More importantly, the PICR potentially provides a general framework to uncover the hidden and biologically meaningful relative ACs by transforming noisy probe intensity data into unbiased estimation of relative AC data, which can then be used for subsequent analysis, such as genotype calling, as shown in this paper.

DNA CN estimation is essential for human genetic variation studies. In particular, accurate CN estimation and genotype calling play a critical role in CNV studies, in which CNV detection becomes more challenging with the growing number of microarray platforms (28,29). While most CNV studies depend on genotype information

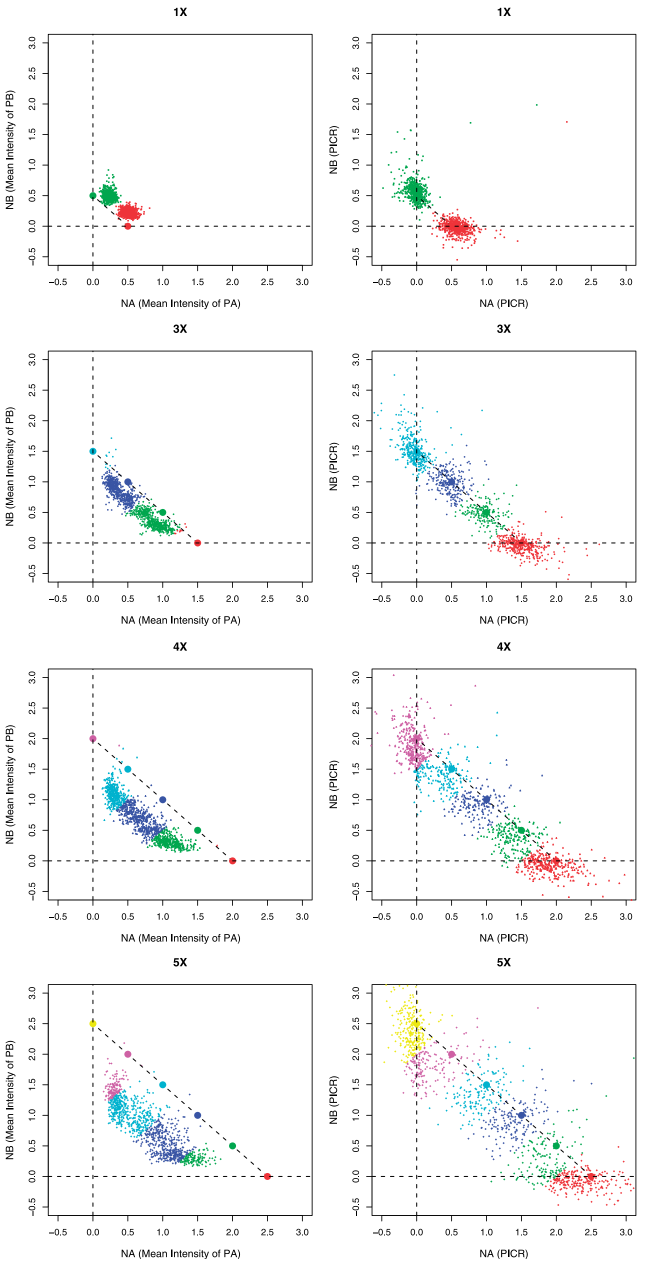

Figure 6. Estimated relative ACs of all 1204 SNPs located on the X-chromosome on the Xba array and their clusters. The ACs are relative to the corresponding SNP total concentration in the reference sample 2X. 'NA' is the AC of allele A, and 'NB' the AC of allele B. The colored dots represent the theoretical centers of the clusters of possible genotypes in each sample. (Left) Mean intensity method. (Right) the PICR method. The clustering was conducted by assigning each SNP to the genotype with the closest theoretical center. For example, in sample 4X relative to sample 2X, the relative true ACs at the theoretical centers are (2,0), (1.5, 0.5), (1,1), (0.5, 1.5) and (0,2), corresponding to the genotypes 'AAAA', 'AAAB', 'AABB', 'ABBB' and 'BBBB', respectively.

and require predetermination or simultaneous determination of SNP genotypes, many rely on large sample multi-array training, and therefore may not be suitable for cross-laboratory and small sample studies, particularly so if the given SNP is strongly associated with the phenotype of interest. Since large-scale genome-wide studies often involve collaboration among laboratories, the development of robust methods of CN estimation becomes crucial.

Across-array normalization has been widely used in microarray studies to remove the variability from artifacts in array processes that confound biological differences. However, this practice tends to introduce undesired variability into SNP genotyping, in that an individual's genotype then depends on the data of other individuals, irrespective of the fact that full individual information is available in one single array. Our AC-based genotype-calling method through the PICR does not require across-array normalization because this variability is modeled by the intercept, and consequently does not affect the estimation of the relative values of the ACs N_A and N_B for any given SNP within the same array. Therefore, an SNP genotype is fully determined by the individual's data alone.

Recent studies on the genome-wide CNVs have identified genomic wave—a special pattern in normalized intensities that presents spatial autocorrelation along the chromosomes (30,31). It has been shown that genomic wave is observed consistently across array media (CGH arrays, Affymetrix arrays and Illumina arrays), and is highly correlated with the GC content of the genomic segment. Furthermore, adjustment for genomic wave in the CNV algorithms improves the performance of CNV detection (31). Following a reviewer's suggestion, we examined the effect of the genomic wave artifact through the PICR model and observed the genomic wave in the background term of the PICR, positively correlated with the GC content through the 15 HapMap Nsp arrays (data not shown). Since genomic wave is a complicated phenomenon, we believe more work needs to be done to study it with the PICR model.

Affymetrix SNP 6.0 arrays have been launched to enter the market and contain more than 906 600 SNPs and more than 946 000 probes for the detection of CNVs (4,5). Although this array has dropped the mismatch probes and usually has six perfect match probe pairs for each SNP, the distortion of HWMMN as summarized above still remains as a major challenge. We have found that estimating the background term with the nonspecific-binding hybridization of the PDNN model makes the PICR work well (data not shown). With the assistance of proper statistical techniques, the PICR is expected to provide an alternative method for accurate CN estimation and SNP genotype calling for SNP 6.0 arrays without requiring a relatively large number of arrays for genotype calling. Furthermore, the Affymetrix GeneChip 500K arrays have been used in a number of large-scale GWASs, including the Wellcome Trust GWAS on seven common disorders (1). Given the large scale of these studies, the data obtained with 500K arrays will continue to remain an important resource for studies on human diseases, and the PICR will remain a viable approach to the studies. We anticipate that the PICR may provide a new tool for further mining the GWAS data to produce more biological findings.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr. Li Zhang for providing the parameters of the PDNN model and his kind technical guidance on SNP array design and annotation. This study was initiated and partly conducted during the visit of LW and MQ to the Department of Epidemiology, Michigan State University.

FUNDING

The National Natural Science Foundation of China (No.10871009, No.10721403), the National High Technology Research and Development of China (No. 2006AA02Z331, No. 2008AA02Z306), the National Key Basic Research Project of China (No. 2009CB918503), the U.S. Public National Center for Research Resources (Health Service Resource grant RR03655), the National Institute of General Medical Sciences (GM28356); and the National Cancer Institute (Cancer Center Support Grant P30CAD43703).

Conflict of interest statement. None declared.

REFERENCES

1. Wellcome Trust Case-Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
2. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
3. Weir,B.A., Woo,M.S., Getz,G., Perner,S., Ding,L., Beroukhim,R., Lin,W.M., Province,M.A., Kraja,A., Johnson,L.A. *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893–898.
4. McCarroll,S.A., Kuruvilla,F.G., Korn,J.M., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M.H., de Bakker,P.I., Maller,J.B., Kirby,A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
5. Korn,J.M., Kuruvilla,F.G., McCarroll,S.A., Wysoker,A., Nemesh,J., Cawley,S., Hubbell,E., Veitch,J., Collins,P.J., Darvishi,K. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.
6. Affymetrix. (2006) BRLMM: an improved genotyping calling method for the GeneChip Human Mapping 500K Array Set. *White Paper*.
7. Di,X., Matsuzaki,H., Webster,T.A., Hubbell,E., Liu,G., Dong,S., Bartell,D., Huang,J., Chiles,R., Yang,G. *et al.* (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.
8. LaFramboise,T., Weir,B.A., Zhao,X., Beroukhim,R., Li,C., Harrington,D., Sellers,W.R. and Meyerson,M. (2005) Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput. Biol.*, **1**, e65.
9. 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.
10. Slater,H.R., Bailey,D.K., Ren,H., Cao,M., Bell,K., Nasioulas,S., Henke,R., Choo,K.H. and Kennedy,G.C. (2005) High-resolution identification of chromosomal abnormalities using oligonucleotide arrays containing 116,204 SNPs. *Am. J. Hum. Genet.*, **77**, 709–726.
11. Abdueva,D., Skvortsov,D. and Tavare,S. (2006) Non-linear analysis of GeneChip arrays. *Nucleic Acids Res.*, **34**, e105.
12. Huang,J., Wei,W., Chen,J., Zhang,J., Liu,G., Di,X., Mei,R., Ishikawa,S., Aburatani,H., Jones,K.W. *et al.* (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. *BMC Bioinformatics*, **7**, 83.
13. Nicolae,D.L., Wu,X., Miyake,K. and Cox,N.J. (2006) GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics*, **22**, 1942–1947.
14. Rabbee,N. and Speed,T.P. (2006) A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, **22**, 7–12.
15. LaFramboise,T., Harrington,D. and Weir,B.A. (2007) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics*, **8**, 323–336.
16. Xiao,Y., Segal,M.R., Yang,Y.H. and Yeh,R.F. (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, **23**, 1459–1467.
17. Carvalho,B., Bengtsson,H., Speed,T.P. and Irizarry,R.A. (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.
18. Bengtsson,H., Irizarry,R., Carvalho,B. and Speed,T.P. (2008) Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, **24**, 759–767.
19. Zhang,L., Wu,C., Carta,R. and Zhao,H. (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.*, **35**, e18.
20. Zhang,L., Miles,M.F. and Aldape,K.D. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol.*, **21**, 818–821.
21. Held,G.A., Grinstein,G. and Tu,Y. (2006) Relationship between gene expression and observed intensities in DNA microarrays–a modeling study. *Nucleic Acids Res.*, **34**, e70.
22. Held,G.A., Grinstein,G. and Tu,Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl Acad. Sci. USA*, **100**, 7575–7580.
23. Kennedy,G.C., Matsuzaki,H., Dong,S., Liu,W.M., Huang,J., Liu,G., Su,X., Cao,M., Chen,W., Zhang,J. *et al.* (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233–1237.
24. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
25. Kapur,K., Jiang,H., Xing,Y. and Wong,W.H. (2008) Cross-hybridization modeling on Affymetrix exon arrays. *Bioinformatics*, **24**, 2887–2893.
26. Johnson,W.E., Li,W., Meyer,C.A., Gottardo,R., Carroll,J.S., Brown,M. and Liu,X.S. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.
27. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
28. Scherer,S.W., Lee,C., Birney,E., Altshuler,D.M., Eichler,E.E., Carter,N.P., Hurles,M.E. and Feuk,L. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
29. Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.
30. Diskin,S.J., Li,M., Hou,C., Yang,S., Glessner,J., Hakonarson,H., Bucan,M., Maris,J.M. and Wang,K. (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.*, **36**, e126.
31. 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.