COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 10, No. 2, pp. 83-98, 2010

002

THE ANALYSIS OF BIASES OF COPY NUMBERS FROM AFFYMETRIX SNP ARRAYS*

LIN WAN†, YI XIAO†, QUAN CHEN†, MINGHUA DENG†, AND MINPING QIAN†

Abstract. Affymetrix SNP arrays are widely used for both genome-wide association and copy number variation (CNV) studies, both of which depend on accurate copy number estimation. However, depending on the method used to copy number estimation, distortions from the actual copy numbers can occur. Therefore, we demonstrate here several effects that can bias accurate copy number estimation, and we describe how some of these biases can be adjusted by existing methods, while others require further study.

1. Introduction. DNA microarrays are widely used in both molecular biology and medicine. Various prototypes of microarrays have been designed with different aims, such as gene expression measurement, transcription factor binding region identification, chromatin modification profiling, SNP genotyping, copy number variation scrutiny, and alternative splicing detection. Therefore, it becomes imperative to develop new methods of accurately preprocessing microarray data in order to meet the high resolution requirements of quantitative biology studies.

Specifically, human genetic variation studies [1] offer great promise in deciphering the genetics of complex diseases through genome-wide association (GWA) [2] and copy number variation (CNV) studies [3, 4]. Affymetrix SNP arrays, originally developed for SNP genotyping [5], are now being widely used for CNV analyses [3, 4, 6]. However, although numerous methods have been developed and have achieved high accuracy in SNP genotyping [7, 8, 9, 10], methods for CNV inference [11, 12, 13, 14, 15, 16] are still far from satisfactory [17, 18].

A key problem of current CNV studies is the lack of "gold standard" samples to directly evaluate copy number estimation by various methods [17, 18]. To compensate for this, methods for Affymetrix SNP arrays usually take the probe intensities, such as the mean intensities of perfect match probes of SNPs, as the estimated copy number [11, 15, 16, 6]. However, because of the complexities of hybridization, it is natural to ask whether such estimation is really proportional to the copy numbers of the target sequences. Up to now, most studies of SNP genotype calling and CNV inference have not elucidated the biases in estimated copy numbers for Affymetrix SNP arrays. On the other hand, many studies have reported on the effects of probe hybridization mechanisms of microarray that could distort the probe intensity from true copy number of target sequences. Although those studies were conducted on

*Dedicated to Michael Waterman on the occasion of his 67th birthday.

†School of Mathematical Sciences and Center for Theoretical Biology, Peking University, Beijing 100871, China. Email: wanlin@ctb.pku.edu.cn, {iamxiaoyi, chenquan1023}@126.com, {dengmh, qianmp}@math.pku.edu.cn

Affymetrix expression, tiling and exon arrays, we found that the basic principle underlying various arrays could be generally applied to Affymetrix SNP arrays. We summarize the possible effects as follows:

I. The binding affinities of probes differ according to the variability of probe sequences; therefore, intensities of probes for the same target sequences can vary greatly, depending on probe sequences [19, 20, 21, 22, 23, 24];
II. The background intensities of probes also have non-negligible variation from probe to probe [25];
III. Saturation of probe intensities is limited by the number of oligonucleotides for each probe [22, 26, 23];
IV. Cross-hybridization of probe to sequences other than the desired target sequence [25, 10, 27, 24];
V. Background noise in image scanning and computer data processing [28];
VI. Whole genome amplification (WGA) of sequences [12, 29].

With those effects in mind, we first explain how copy numbers estimated by traditional methods can remarkably distort the underlying real copy numbers of SNPs. Next, we show that such biases can prevent accurate SNP genotype calling and CNV inference. Finally, we explain how some existing algorithmic models can adjust these biases to reflect a truer expression of the underlying real SNP copy numbers. Especially, we emphasize a statistical correction method, termed probe intensity composite representation (PICR) model, which was proposed by the authors and Fu et al. in [24] which can efficiently adjust the effects I, II and IV (see Appendix B for a brief description of PICR).

2. Evidence of biases of copy numbers. As a means of estimating copy numbers, many studies have used probe intensities, such as the mean intensities of perfect match probes, as the copy number [11, 15, 16, 6]. This method assumes that probe intensity is approximately proportional to copy numbers of the target sequences. However, we found that the copy numbers estimated by mean intensity can, in fact, be greatly biased from the real copy number and thus seriously affect the accuracy of genotype calling and CNV inference. In this section, three cases are given to demonstrate such biases.[1]

2.1. DNA samples with known DNA copy number of chromosome X. We first utilized the samples of Affymetrix 100K SNP with known copy number of X chromosomes (1X to 5X) hybridized [13]. This sample was used as a benchmark to validate our assertions on copy number estimation. Here we use mean intensity of perfect match probes as the copy number estimation (Figure 1(a,c)), and we also use the PICR model for the purpose of comparison (Figure 1(b,d)).

[1] A brief description of the design of Affymetrix SNP array can be found in Appendix A. For data and method, please refer to Appendix C and D.

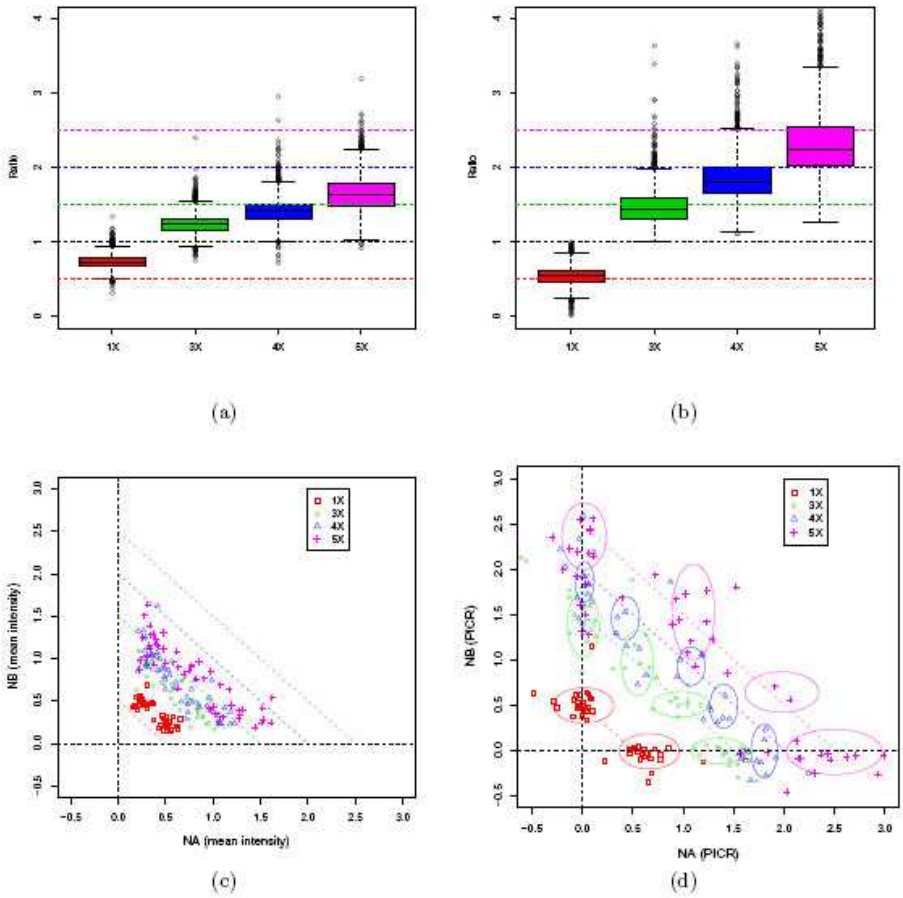

FIG. 1. *Copy number estimation with different number of Chromosome X. (a-b): Boxplots of ratios of the total copy numbers of the 2361 SNPs for 1X, 3X, 4X, and 5X samples relative to those in the 2X sample; dashed lines in the plots are from the real number of chromosome X of samples; (a) copy number estimated by mean of the all perfect match probe intensities for each SNP; (b) total copy number estimated by PICR model ($N_{total} = N_A + N_B$). (c-d) The allele-specific copy numbers of 50 randomly selected SNPs for 1X, 3X, 4X, and 5X samples relative to total copy numbers of those in the 2X sample; (c) copy number estimated by mean perfect match probe intensity for allele A and allele B, respectively; (d) allele-specific copy number estimated by PICR model; "NA" for copy number of allele A and "NB" for copy number of allele B.*

The ratio of total copy number of each SNP from the sample of interest to that of the reference sample was usually taken into account. Here, the normal sample with 2X was used as the reference compared to the 1X, 3X, 4X, and 5X samples. The estimated copy numbers for a total of 2361 SNPs on chromosome X in both Xba array and Hind array for 1X, 3X, 4X, and 5X samples relative to those in the 2X sample are shown in Figure 1(a). One can clearly observe that the median ratio of each SNP in each sample is greatly biased from its corresponding real ratio (dashed lines with

the same color in Figure 1(a)). For most samples of 4X and 5X, ratios calculated are below 2 and are obviously underestimated.

To show that such underestimation could confuse the inference of the subtle structure of genotype, we randomly selected 50 SNPs on chromosome X and plotted their allele-specific copy number estimated by mean intensity (Figure 1(c)). We found that the copy numbers estimated by mean intensities are not only seriously biased from the real copy numbers, but that their genotype patterns are also nearly mixed together, except for the 1X sample. However, after corrections by PICR, the accuracy of estimated copy numbers improved greatly such that the allelic copy number of most SNPs are much closer to their real number than that by the mean intensity (Figure 1(c,d)). Meanwhile, the different genotype groups become separately clustered and distinguishable (Figure 1(d)). However, about 15% to 25% of the SNPs are mixed with other groups among samples with 3X-5X, possibly resulting from the fact that saturation of probe intensity would occur with the increased copy number of target sequences.

2.2. DNA fragments with multiple SNPs. In the example above, we used the relative copy number with respect to copy number from reference samples. However, in many studies, well annotated reference samples are limited [17, 18] which makes it difficult for us to obtain the relative copy numbers. It is therefore important to study the biases caused by the multiplicative effects of probe intensity.

TABLE 1

Number of DNA fragments with multiple SNPs by SNP array prototype. DNA fragments with physical position annotated as "—" (unknown) in Affymetrix annotation files are not counted here.

SNPs on signal fragment	50K Xba	50K Hind	250K Nsp	250K Sty
8	2	0	1	2
7	2	3	6	5
6	14	18	29	26
5	57	75	155	134
4	305	328	893	797
3	1647	1685	5375	4696
2	7784	7704	31026	28296
1	36488	34654	179469	163484

In the processing of Affymetrix SNP arrays, DNA samples are first digested with a restriction enzyme (e.g., XbaI) and ligated with adapters into DNA fragments before whole genome amplification. After the amplification, those DNA fragments containing the SNPs are then interrogated with SNP-specific probe-sets on the array through hybridization [5]. It was pointed out that the amplification may have different effi-

ciencies for different fragment sequences with different length and GC content [12]. Thus, we investigated the SNPs on same fragments which are supposed to have the same copy number of target sequences and found that the estimated copy numbers of SNPs on these DNA fragments are not always the same. That is, about 40% of SNPs on the 50K Xba array are located on fragments with multiple SNPs (Table 1). Here, we took 305 DNA fragments from the 50K Xba array having 4 SNPs on each as an example. For about 150, 50% out of 305 fragments, we found that the ratio of the maximum to the minimum copy numbers (mean intensities) of the 4 SNPs on each fragment was greater than 2. Strikingly, across 90 HapMap samples, the pattern of the copy numbers (mean intensities) of these 4 SNPs is similar, indicating that such difference in mean intensities (estimated copy number) of SNPs is by no means solely a factor of randomness. Figure 2(a) demonstrates 4 SNPs (SNP_A-1697318, SNP_A-1683386, SNP_A-1737711 and SNP_A-1730973) located on the same fragment with systematic copy number differences across 90 HapMap samples .

Based on this line of evidence, we asked whether such difference was due to the bias from probe-dependent hybridization, such as effects I-IV, or probe dependent hybridization. To address this question, PICR, which is an efficient correction for biases from I, II and III, was applied. But the corrected copy numbers of SNPs on the same DNA fragment can still not be considered as equal (data not shown).

2.3. Systematic biases in image scanning and computer data processing. When exploring how the estimated copy numbers of SNPs on the same DNA fragments differ so dramatically, Wan et al. studied the raw data (CEL files) of these arrays and found systematic biases in image scanning and computer data processing [28]. Figure 2(c) shows 3 randomly selected CEL images (generated by the `dChip` software before normalization) from the 90 HapMap samples. Each image consists of 1600×1600 grids representing the intensities and probe locations on the array. It is observed that the probe intensities have surprising patterns of bright and dark horizontal bands consistently across the arrays. This type of pattern was systematically observed in all 90 arrays [28]. These patterns were not designed on purpose by Affymetrix for probe intensities of SNP array.

Comparing Figures 2(a,b) with Figure 2(c), we observed that the array intensities have strong position-dependence, and, thus, the estimation of copy numbers must also be position-dependent. For example, probes of SNP_A-1683386 (green) are mainly located in the lower half of the array while probes of SNP_A-1737711 (blue) are mainly located in the upper half of the array, which has lower intensities than the lower part. Therefore, the estimated copy number values are positively correlated with the intensity variation along the vertical direction of the array with Pearson correlation coefficient 0.29 (p-`value` $< 2.2e-16$) [28]. Although the examples in the illustration here were selected from DNA fragments of multi-SNPs, such intensity bias

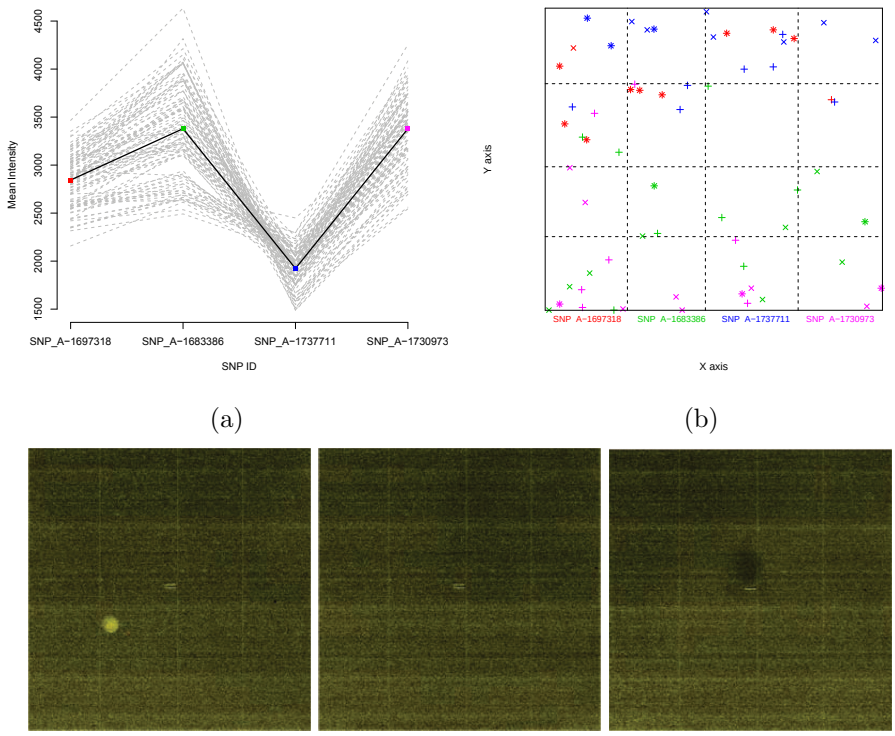

FIG. 2. *The copy number bias and systematic biases in image scanning and computer data processing. (a) Mean intensities of 4 SNPs (SNP_A-1697318, SNP_A-1683386, SNP_A-1737711 and SNP_A-1730973) in all 90 samples. The mean intensities of the 4 SNPs in all 90 samples are plotted in gray color. The colored points represent the mean of each SNP in the 90 samples. Y-axis indicates the mean intensity. (b) The probe positions of the 4 SNPs on the Mapping 50K Xba 240 array. The colored points in the plot correspond to the perfect match probes of the SNP with SNP ID denoted in the same color at the bottom of the figure. The symbols "+" and "×" in the plot represent perfect match probes for allele A and allele B, respectively. (c) Raw images of 3 CEL files randomly selected from the 90 HapMap samples; each image consists of* 1600×1600 *grids representing the intensities and locations of probes on the array.*

also influences the copy number estimation of fragments with single SNP.

This phenomenon was also observed in other prototypes of Affymetrix SNP arrays, including both 100K arrays and 500K arrays generated from various laboratories (data not shown). Therefore, without correcting the above intensity bias, copy number estimation might be subject to systematic bias. Interestingly, it has recently been noticed that such systematic bias also exists in Affymetrix gene expression arrays [30]. This makes it a general imperative that bias correction be conducted before further analysis.

It is further demonstrated that such bias within array has a multiplicative effect (or additive effects in the log scale) to probe intensities. This means that such bias can be adjusted by within array by certain normalization approach. Based on this observation, Wan et al. developed an efficient method to adjust the systematic bias within an array. After such correction, the estimated copy number and intensity variations are not correlated with Pearson correlation coefficient 0.039 (p-value = 0.082) [28]. However, even after adjustment of image scanning effects on probe level by method of [28] and then applying PICR, the copy numbers of SNPs on the same DNA fragment can still not be considered as equal (data not shown), and we speculate that effect VI is the most plausible reason (see section 3.3 for more details).

2.4. Biases of nonspecific hybridization. To illustrate the biases from nonspecific hybridization, we utilize the information of Y chromosome of female samples. For female samples, the copy numbers of SNPs located on the Y chromosome should be zero. Thus, the probes of SNPs located on Y chromosome can be mostly considered as nonspecific hybridization, while probes of SNPs located on other chromosomes can mostly be considered to have both specific and nonspecific hybridization, but with specific hybridization as the domain. We used samples from Genome-Wide Human SNP 6.0 Data Set since the SNP 6.0 array has over 900 SNPs located on Y chromosome for our statistical analysis. To our surprise, a non-negligible portion of probes with pure nonspecific hybridization can have intensities as high as the probes with specific hybridization (see Table 2). This shows that background intensities (nonspecific hybridization) of probes also have non-negligible variation from probe to probe. Studies have also shown that the nonspecific hybridization is probe sequence-dependent [25].

TABLE 2

Nonspecific binding effects. Two female samples from the Genome-Wide Human SNP 6.0 Data Set were used as an illustration. The probe intensities of SNP located on Y chromosome (Y) and non-Y chromosomes ($\bar{Y}$) are summarized.

Samples	Probe Intensity	Min	1st Qu.	Median	Mean	3rd Qu.	Max
NA12892	Y	84	234	476.5	1392	1492	20350
	$\bar{Y}$	64	473	973	1229	1705	64880
NA06985	Y	49	126	235.5	811.9	808.8	13420
	$\bar{Y}$	38	247	543	723.1	1005	20100

3. Methods for bias corrections.

3.1. The biases of the probe from binding affinities in hybridization, background intensity, and saturation. Probe intensities not only vary from copy numbers of target sequences but also strongly depend on binding affinities of probe sequences. Since the binding affinities of probe sequences can vary greatly from probe

to probe, the intensities of probes for a SNP cannot be treated as repeat samples of its copy number. Therefore, instead of simply using mean or median intensity as the estimation, we need to adjust the effects of probe binding affinities to accurately estimate copy numbers.

Several groups have constructed the probe binding models and indicated how binding affinities depend on the probe sequences [19, 20, 21, 22, 23, 31, 32, 33, 34]. Zhang et al. proposed a simple approximation of binding free energy using a positional-dependent-nearest-neighbor (PDNN) model with good accuracy for perfect matches [19, 20]. Other investigators have developed similar models [21, 22]. The PDNN model is formulated in Equation (1), where E is the binding free energy of perfect match, ω_l is a weight factor that depends on the position of consecutive bases along the oligonucleotide; b_l is the l-th nucleotide of probe sequence, and λ is the stacking energy of the pair of nearest-neighbors along the probe sequence.

$$E = \sum_{l=1}^{24} \omega_l \lambda(b_i, b_{i+1}), \tag{1}$$

where the λ characterizes the effect of the probe sequence to free energy by nearest-neighbor of nucleotides. Thus, the PDNN model significantly improves the accuracy from models based only on single nucleotide contents.

The specific hybridization intensity (I_s) of probe has also been modeled with copy numbers (N) and binding affinity (E) through the Langmuir-like absorption principle [19, 20] ($I_s = N/(1 + e^E)$). Thus, the probe intensity I can be formulated as Equation (2), where E is the binding free energy of the hybridization, N is the copy number, or concentration, of sequences in binding, I_{bg} is the background intensity, and ε is the measurement error of intensity [19, 20].

$$I = I_s + I_{bg} + \varepsilon = N \frac{1}{1 + e^E} + I_{bg} + \varepsilon. \tag{2}$$

It is noticed that for Affymetrix SNP arrays, the binding intensities of perfect match probes not only include the contribution from the their perfect match target sequences from one allele, but also potentially include the binding of sequences from their opposite allele with one mismatch site (the cross-hybridization of probe to off-target allele sequence). This kind of binding affinity with 1-2 mismatch sites is modeled by generalized PDNN (GPDNN) model [24].

Based both on the GPDNN model and Langmuir-like absorption principle, a statistical regression model, probe intensity composite representation (PICR)[24] model, considers the complex binding of both perfect match of probe with its target sequences and specific binding to other target sequences (maybe in opposite allele) with 1-2 mismatch sites. Thus, the intensity of SNP probe can be described as

$$I = I_s^A + I_s^B + I_{bg} + \varepsilon \tag{3}$$

where I_s^A is the specific hybridization intensity from allele A and I_s^B is the specific hybridization intensity from allele B (see Appendix B for more details). The PICR model was successfully applied to Affymetrix SNP array for SNP genotyping and copy number variation detections [24]. Furthermore, the PICR model can also reduce the effect of the background intensity I_{bg} in Equation (2) via the regression approach. The idea of PICR model is briefly described in Appendix B.

If copy numbers estimated by PICR are used for the example of X chromosome, it is interesting to observe that ratios of copy numbers of SNPs for the 4 samples (1X, 3X, 4X, and 5X) relative to those in sample 2X (Figure 1(c)) are much closer to the true ratios than the traditional method and that the different genotype groups become separately clustered and distinguishable (Figure 1(d)). On the other hand, in Figure 1(d), one can also see that the saturation problem still exists after correction by PICR. In fact, [19, 20] the effect from a limited number of oligonucleotides for each probe is not considered. Hence, the result based on [19, 20] does not eliminate the saturation problem. Abdueva et al. [26] did consider the saturation problem, but their free energy model is not as accurate as [19, 20]. Held et al. [22] proposed a more complete model recognizing concentrations of oligonucleotide for each probe, but there is still lack of feasible statistical algorithm based on this model. Neither can it be applied to real SNP arrays. It should also be noted that Binder et al. have also conducted a systematic survey and analysis of both specific and nonspecific hybridization of gene expression arrays [31, 32, 33, 34].

3.2. Biases from cross-hybridization of probe to sequences other than the desired original target sequence. In the interrogation of heterozygous SNPs, we found that there would be non-negligible cross-hybridization in Affymetrix SNP arrays because 1) the intensity of perfect match probes not only contributed by perfect match binding with one allele, but also from mismatch binding with the other allele with one mismatch nucleotide and 2) mismatch binding can be influential and comparable to perfect match binding and thus should not be regarded as background nonspecific binding. This explains why most methods do not have performances in the genotyping of heterozygous SNP (e.g., AB) as good as those for homozygous SNPs (e.g., AA or BB). Bengtsson et al. [10] used a statistical approach to solve this kind of cross-hybridization, but it depends on a large training sample. PICR is more efficient for adjustment of cross-hybridization in SNP array [24] since it qualitatively characterizes both perfect and mismatch hybridization by physicochemical principle and needs only each sample (array) itself to estimate the copy number. For homozygous SNPs, one can see that copy number estimated by PICR of null-allele is closer to zero (Figure 1(d)), while copy number estimated by the mean intensity is not (Figure 1(c)). For heterozygous SNPs, the allele-specific copy numbers by PICR are closer to the real copy numbers (Figure 1(d)).

Although perfect match probes are designed to be perfectly complementary to target sequences, they may also share sequence similarity to other sequences existing in the genome, thus resulting in another source of probe cross-hybridization. Such cross-hybridization is, however, much more difficult to adjust. Johnson et al. provided an adjustment method for Affymetrix tiling array using probe mappings to genomic sequences [25], and Karpur et al. [27] provided an adjustment method for Affymetrix Exon array using probe mappings to off-target transcripts. Approaches similar to [25, 27] can also be used for Affymetrix SNP array to adjust cross-hybridization of probes having many copies on whole genome with a few mismatches.

3.3. Biases from whole genome amplification. Although procedures have been taken to correct biases of Affymetrix SNP array by PICR incorporating the adjustment of image scanning effects, it has been found that the copy numbers of SNPs on the same DNA fragment cannot yet be considered equal (data not shown). Meanwhile, studies indicate that whole genome amplification (WGA) can introduce many artifacts of CNV [29]. Some authors account for the biases introduced by WGA by the length and GC content of DNA fragments amplified by PCR [29, 12]. However, length and GC content cannot explain our cases of biases of copy number estimation of SNPs on the same DNA fragment. It is also known that PCR is taken with a single primer for all DNA sequences during the WGA procedures and that can fail if the primer is bound to the internal regions of DNA sequences [35, 36]. Therefore, further studies are needed to determine whether such biases are from PCR failure caused by binding of primer to the internal regions of DNA fragments.

3.4. Discussion. Although we showed here that the accuracies of copy number estimations may be impeded by various biases of probe intensity, it is still possible to obtain highly reliable results with Affymetrix SNP array at relatively high resolution with low-cost if proper statistical methods are applied. Affymetrix SNP 6.0 array is now the standard protocol. Although Affymetrix SNP 6.0 drops the mismatch probes and usually returns 6 perfect match probes for each SNP, the biases we summarized above still remain a major challenge, especially the problem of cross-hybridization. And our recent study indicate that by replacing general term of background with the probe-dependent background estimated by PDNN model of non-specific hybridization, PICR can still work well (in preparation).

Acknowledgments. This work is supported by 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), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES

[1] E. Pennisi. Breakthrough of the year. human genetic variation. Science, 318:5858(2007), pp. 1842–3.

[2] wellcome^trust. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 447:7145(2007), pp. 661–78.

[3] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. *Global variation in copy number in the human genome.* Nature, 444:7118(2006), pp. 444–54.

[4] B. A. Weir, M. S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhim, W. M. Lin, M. A. Province, A. Kraja, L. A. Johnson, K. Shah, M. Sato, R. K. Thomas, J. A. Barletta, I. B. Borecki, S. Broderick, A. C. Chang, D. Y. Chiang, L. R. Chirieac, J. Cho, Y. Fujii, A. F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B. E. Johnson, M. G. Kris, A. Lash, L. Lin, N. Lindeman, E. R. Mardis, J. D. McPherson, J. D. Minna, M. B. Morgan, M. Nadel, M. B. Orringer, J. R. Osborne, B. Ozenberger, A. H. Ramos, J. Robinson, J. A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M. R. Spitz, M. S. Tsao, D. Twomey, R. G. Verhaak, G. M. Weinstock, D. A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M. F. Zakowski, Q. Zhang, D. G. Beer, II Wistuba, M. A. Watson, L. A. Garraway, M. Ladanyi, W. D. Travis, W. Pao, M. A. Rubin, S. B. Gabriel, R. A. Gibbs, H. E. Varmus, R. K. Wilson, E. S. Lander, and M. Meyerson. *Characterizing the cancer genome in lung adenocarcinoma.* Nature, 450:7171(2007), pp. 893–8.

[5] G. C. Kennedy, H. Matsuzaki, S. Dong, W. M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, M. T. Boyce-Jacino, S. P. Fodor, and K. W. Jones. *Large-scale genotyping of complex dna.* Nat Biotechnol, 21:10(2003), pp. 1233–7.

[6] S. A. McCarroll, F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby, A. L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P. J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K. W. Jones, R. Rava, M. J. Daly, S. B. Gabriel, and D. Altshuler. *Integrated detection and population-genetic analysis of snps and copy number variation.* Nat Genet, 40:10(2008), pp. 1166–74.

[7] X. Di, H. Matsuzaki, T. A. Webster, E. Hubbell, G. Liu, S. Dong, D. Bartell, J. Huang, R. Chiles, G. Yang, M. M. Shen, D. Kulp, G. C. Kennedy, R. Mei, K. W. Jones, and S. Cawley. *Dynamic model based algorithms for screening and genotyping over 100 k snps on oligonucleotide microarrays.* Bioinformatics, 21:9(2005), pp. 1958–63.

[8] N. Rabbee and T. P. Speed. *A genotype calling algorithm for affymetrix snp arrays.* Bioinformatics, 22:1(2006), pp. 7–12.

[9] Y. Xiao, M. R. Segal, Y. H. Yang, and R. F. Yeh. *A multi-array multi-snp genotyping algorithm for affymetrix snp microarrays.* Bioinformatics, 2007.

[10] H. Bengtsson, R. Irizarry, B. Carvalho, and T. P. Speed. *Estimation and assessment of raw copy numbers at the single locus level.* Bioinformatics, 24:6(2008), pp. 759–67.

[11] H. R. Slater, D. K. Bailey, H. Ren, M. Cao, K. Bell, S. Nasioulas, R. Henke, K. H. Choo, and G. C. Kennedy. *High-resolution identification of chromosomal abnormalities*

using oligonucleotide arrays containing 116,204 snps. Am J Hum Genet, 77:5(2005), pp. 709–26.

[12] Y. Nannya, M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaishi, M. Kurokawa, S. Chiba, D. K. Bailey, G. C. Kennedy, and S. Ogawa. *A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays.* Cancer Res, 65:14(2005), pp. 6071–9.

[13] J. Huang, W. Wei, J. Chen, J. Zhang, G. Liu, X. Di, R. Mei, S. Ishikawa, H. Aburatani, K. W. Jones, and M. H. Shapero. *Carat: a novel method for allelic detection of dna copy number changes using high density oligonucleotide arrays.* BMC Bioinformatics, 7:83, 2006.

[14] T. LaFramboise, D. Harrington, and B. A. Weir. *Plasq: a generalized linear model-based procedure to determine allelic dosage in cancer cells from snp array data.* Biostatistics, 8:2(2007), pp. 323–36.

[15] C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and M. E. Hurles. *A robust statistical method for case-control association testing with copy number variation.* Nat Genet, 40:10(2008), pp. 1245–52.

[16] J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler. *Integrated genotype calling and association analysis of snps, common copy number polymorphisms and rare cnvs.* Nat Genet, 40:10(2008), pp. 1253–60.

[17] N. P. Carter. *Methods and strategies for analyzing copy number variation using dna microarrays.* Nat Genet, 39:7 Suppl(2007), pp. S16–21.

[18] S. W. Scherer, C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler, N. P. Carter, M. E. Hurles, and L. Feuk. *Challenges and standards in integrating surveys of structural variation.* Nat Genet, 39:7 Suppl(2007), pp. S7–15.

[19] L. Zhang, M. F. Miles, and K. D. Aldape. *A model of molecular interactions on short oligonucleotide microarrays.* Nat Biotechnol, 21:7(2003), pp. 818–21.

[20] L. Zhang, C. Wu, R. Carta, and H. Zhao. *Free energy of dna duplex formation on short oligonucleotide microarrays.* Nucleic Acids Res, 35:3(2007), pp. e18.

[21] G. A. Held, G. Grinstein, and Y. Tu. *Modeling of dna microarray data by using physical properties of hybridization.* Proc Natl Acad Sci U S A, 100:13(2003), pp. 7575–80.

[22] G. A. Held, G. Grinstein, and Y. Tu. *Relationship between gene expression and observed intensities in dna microarrays–a modeling study.* Nucleic Acids Res, 34:9(2006), pp. e70.

[23] N. Ono, S. Suzuki, C. Furusawa, T. Agata, A. Kashiwagi, H. Shimizu, and T. Yomo. *An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays.* Bioinformatics, 24:10(2008), pp. 1278–85.

[24] L. Wan, K. L. Sun, Q. Ding, Y. H. Cui, M. Li, Y. L. Wen, R. Elston, M. P. Qian, and W. J. Fu. *Hybridization modeling of oligonucleotide snp arrays for accurate dna copy number estimation.* Nucleic Acids Research, 37:17(2009), pp. e117.

[25] W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, M. Brown, and X. S. Liu. *Model-based analysis of tiling-arrays for chip-chip.* Proc Natl Acad Sci U S A, 103:33(2006), pp. 12457–62.

[26] D. Abdueva, D. Skvortsov, and S. Tavare. *Non-linear analysis of genechip arrays.* Nucleic Acids Res, 34:15(2006), pp. e105.

[27] K. Kapur, H. Jiang, Y. Xing, and W. H. Wong. *Cross-hybridization modeling on affymetrix exon arrays.* 2008.

[28] L. Wan, W.J. Fu, M. Deng, and M. Qian. *A method to correct systematic bias in affymetrix snp arrays.* Proceeding of The International Conference on BioMedical Engineering and Informatics, pp. 442-6, 2008.

[29] T. J. Pugh, A. D. Delaney, N. Farnoud, S. Flibotte, M. Griffith, H. I. Li, H. Qian, P. Farinha, R. D. Gascoyne, and M. A. Marra. *Impact of whole genome amplification on analysis of copy number variants.* Nucleic Acids Res, 36:13(2008), pp. e80.

[30] K. E. Bjork and K. Kafadar. *Systematic order dependent effect in expression values, variance, detection calls and differential expression in affymetrix genechips(r).* Bioinformatics, 2007.

[31] H. Binder and S. Preibisch. *Specific and nonspecific hybridization of oligonucleotide probes on microarrays.* Biophys J, 89:1(2005), PP. 337–52.

[32] H. Binder, K. Krohn, and S. Preibisch. *"hook"-calibration of genechip-microarrays: Chip characteristics and expression measures.* Algorithms Mol Biol, 3(2008), pp. 11.

[33] H. Binder and S. Preibisch. "hook"-calibration of genechip-microarrays: Theory and algorithm. *Algorithms Mol Biol*, 3:12, 2008.

[34] H. Binder, J. Brucker, and J. Burden C. *Nonspecific hybridization scaling of microarray expression estimates: A physicochemical approach for chip-to-chip normalization.* J Phys Chem B, 2009.

[35] F. Z. Sun and M. S. Waterman. *Whole genome amplification and branching processes.* Adv Appl Prob, 29(1997), pp. 629–88.

[36] R. Andreson, T. Mols, and M. Remm. *Predicting failure rate of pcr in large genomes.* Nucleic Acids Res, 36:11(2008), pp. e66.

Appendix

Appendix A. Brief description of Affymetrix SNP array design.

We briefly describe the design of Affymetrix SNP arrays by taking the 100K Mapping Array as an example. The Affymetrix SNP 100K Mapping Array consists of one pair of Mapping 50K Xba array and 50K Hind array, each using 10 quartets to interrogate a single dimorphic SNP site of alleles A and B, for example. Each quartet consists of 2 pairs (a perfect match probe and a mismatch probe) of probes, one pair for allele A and the other for allele B. Each probe presents a 25-mer oligonucleotide sequence designed to either perfectly match the target sequence or mismatch at the SNP site. The 10 quartets are designed to take a different shift (k) of nucleotide on the probe sequence (k may take -4, -3, -2, -1, 0, 1, 2, 3, 4) surrounding the center nucleotide of the probe sequence ($k = 0$ at position 13 of the 25-mer) and may also take sense or antisense strands. It is important to note that mismatch probes have one sure mismatch nucleotide at the center position ($k = 0$) and may have another mismatch at a shift $k \neq 0$. Similar to the 100K array design, the Affymetrix Mapping 500K SNP Array consists of one pair of Mapping 250K Nsp array and 250K Sty array, but only 6 quartets are used to interrogate each SNP, while the SNP 6.0 array has fewer probes for each SNP and removess the mismatch probes.

Appendix B. Brief introduction of GPDNN and PICR model.

Based on previous work [24], we concluded that probe intensities for most oligonucleotide arrays vary with copy numbers of target sequences and strongly depend on

binding affinities of probe sequences. Thus far, however, the binding free energies of only perfect match sequences have been characterized quantitatively, through a positional-dependent-nearest-neighbor (PDNN) model, while probe intensities have been modeled with copy numbers and binding affinities using the Langmuir-like absorption principle [19, 20] (also see section 3.1). The situation of hybridization is especially complex for the Affymetrix SNP array: we showed that probes of Affymetrix SNP array are subjected to cross-hybridization by binding with off-target allele sequences, which can, in turn, greatly affect the accuracies of both genotype calling and copy number estimation [24] (also see section 3.2).

To address the biases arising from both binding affinity and cross-hybridization of probes, we studied the hybridization properties of sequence binding and developed the generalized PDNN (GPDNN) model for the binding free energy involving both perfect match and cross-hybridization bindings. We then developed the probe intensity composite representation (PICR) model based on both the Langmuir-like adsorption principle [19, 20] and the GPDNN model. In PICR, the intensities of each probe of a given SNP are decomposed into 4 terms: 2 terms for specific binding of the two alleles, 1 term for background nonspecific binding, and an error term (see Equation (4)).

$$
\begin{cases}
\vdots \\
I_i = N_A\varphi(E^{i,A}) + N_B\varphi(E^{i,B}) + I^i_{bg} + \varepsilon \\
\vdots \\
I_j = N_A\varphi(E^{j,A}) + N_B\varphi(E^{j,B}) + I^j_{bg} + \varepsilon \\
\vdots
\end{cases} \tag{4}
$$

where $\varphi(x) = 1/(1 + e^x)$; N_A is the copy number of allele A, and N_B is the copy number of allele B; $E^{i,A}$ is the binding free energy of probe i to the sequences of allele A, and $E^{i,B}$ is the binding free energy of probe i to the sequences of allele B; I^i_{bg} and I^j_{bg} are the background terms of probe i and j; ε is an error term. These binding free energies like the $E^{i,A}$ and $E^{i,B}$ will be first calculated by the GPDNN model , and then PICR model leads to a statistical linear regression based of Equation (4) that yields consistent estimation of copy numbers utilizing the probe intensities for each SNP.

For more information about PICR, please refer to [24] or the website at `http://ctb.pku.edu.cn/~wanlin/PICR`.

Appendix C. Data. The SNP array samples of the HapMap trio dataset were downloaded from the Mapping 100K dataset (`http://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx`) and the Mapping 500K dataset (`http://www.affymetrix.com/support/technical/sample_data/500k_data.affx`). All the annotation files of the corresponding Affymetrix SNP ar-

ray were downloaded from the Affymetrix website. Data of the Affymetrix 100K SNP arrays hybridized with samples of 1 to 5 copies of the X-chromosome (1X to 5X) were obtained from the Affymetrix Sample Data Sets for Copy Number Analysis (`http://www.affymetrix.com/support/technical/sample_data/copy_number_data.affx`). Data of the Affymetrix SNP 6.0 Array were from the Genome-Wide Human SNP 6.0 Data Set, which can be ordered from Affymetrix.

Appendix D. Data processing and copy number estimations. Two methods were utilized for copy number estimations for Affymetrix SNP array.

- **Copy number from mean intensity of perfect match probes.** A probe-level `quantile` normalization is taken across samples. For each sample and given SNP, copy number for allele A is taken as mean intensity of perfect match probes of allele A; copy number for allele B is taken as mean intensity of perfect match probes of allele B [11, 15, 16, 6]. The total copy number of each SNP is taken as copy number of allele A plus copy number of allele B.
- **Copy number estimated by the PICR model [24].** The PICR model can estimate allele-specific copy number (N_A and N_B in Equation (4)) with adjustment of the biases from probe binding affinity, background intensity, and cross-hybridization from allele-specific target bindings for a single sample. For each sample, the total copy number is taken as the sum of two allele-specific copy numbers estimated by PICR of each SNP. We then normalized copy number estimated between samples by multiplying a sample-specific constant to make the median copy number of each sample the same across different samples.