An Imputation Approach for Oligonucleotide Microarrays

Ming Li[1], Yalu Wen[2], Qing Lu[2], Wenjiang J. Fu[2]*

1 Division of Biostatistics, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, United States of America, 2 Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, United States of America

Abstract

Oligonucleotide microarrays are commonly adopted for detecting and qualifying the abundance of molecules in biological samples. Analysis of microarray data starts with recording and interpreting hybridization signals from CEL images. However, many CEL images may be blemished by noises from various sources, observed as "bright spots", "dark clouds", and "shadowy circles", etc. It is crucial that these image defects are correctly identified and properly processed. Existing approaches mainly focus on detecting defect areas and removing affected intensities. In this article, we propose to use a mixed effect model for imputing the affected intensities. The proposed imputation procedure is a single-array-based approach which does not require any biological replicate or between-array normalization. We further examine its performance by using Affymetrix high-density SNP arrays. The results show that this imputation procedure significantly reduces genotyping error rates. We also discuss the necessary adjustments for its potential extension to other oligonucleotide microarrays, such as gene expression profiling. The R source code for the implementation of approach is freely available upon request.

Citation: Li M, Wen Y, Lu Q, Fu WJ (2013) An Imputation Approach for Oligonucleotide Microarrays. PLoS ONE 8(3): e58677. doi:10.1371/journal.pone.0058677

Editor: Jörg D. Hoheisel, Deutsches Krebsforschungszentrum, Germany

Received October 30, 2012; **Accepted** February 5, 2013; **Published** March 7, 2013

Funding: This work was partly supported by the Summer Oncology Scholarship from Michigan State University and the startup funds from University of Arkansas for Medical Sciences. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fuw@msu.edu

Introduction

Oligonucleotide microarrays have been commonly adopted in various biomedical researches, such as gene expression profiling, single nucleotide polymorphism (SNP) genotyping, and copy number estimation, etc [1-3]. Typically, a microarray is attached with millions of short immobilized nucleic acid sequences, known as probes, which are designed as sequences complementary to the nucleic acid molecules in biological samples, known as targets. The targets are usually labeled with fluorescent dyes and their abundance can be qualified by measuring fluorescent intensities yielded from their hybridization with the probes. The intensities are further stored in a CEL file, which becomes the raw data of a microarray experiment (See [4,5] for detail of cDNA chips). This technology is able to produce a large amount of data for thousands of genes or millions of SNPs simultaneously. However, the quality of a microarray may be affected by noises from various sources during the experiment. A series of pre-processing procedures are required before any subsequent analyses can be conducted, including image processing, background adjustment and data normalization [6].

All fluorescence images may have spatial defects to some extent, due to dust and debris, glass flaws, uneven distribution of fluids or surface coatings, etc [7]. Image processing is usually the first step to ensure the validity of downstream analyses. The intensities from a defect area will distort the scaling [8], which may further affect multiple samples through between array normalizations. The impact of image defects have been investigated by using a number of summarization packages available in bio-conductor, such as MAS5, RMA and GCRMA [9–11]. Suárez-Fariñas *et al.* studied the impact of blemished images on gene expression microarrays by directly applying MAS5 and GCRMA [7]. They concluded that a defect area of ~0.2% of a chip would artificially change the expression by two folds for 20 genes in MAS5, and for 3 genes in GCRMA. Reimers *et al.* conducted a similar comparison by using packages MAS5 and RMA [12]. They found that RMA was less affected than MAS5 when the blemished area was small, but its performance got worse with larger blemished regions [13].

To ensure the image quality, researchers were always recommended to visually examine all CEL images [14,15]. However, some defects cannot be easily recognized by naked eyes due to a large dynamic range of probe intensities [13]. In the past few years, several approaches have been proposed for detecting and removing the defect areas automatically. These approaches covered a wide variety of image defects. For example, Suárez-Fariñas *et al.* developed an approach, referred to as "harshlight", which was able to detect and mask three types of defects: localized blemishes affecting a few probes, diffuse defects affecting larger areas, and extended defects which may invalidate an entire chip [7]. Upton *et al.* used replications for identifying the abnormal spatial structures, referred to as "blob", "lines", "rectangular enhancement" and "coffee rings" [16]. Song *et al.* proposed a software package, referred to as Microarray Blob Remover (MBR), which allowed visualization, detection and removal of various 'blob' like defects [17].

These approaches have showed great promise for image processing. However, the existing approaches commonly masked affected intensities with missing. Previous studies indicated that correction of the affected intensities could improve the reliability of data, and thus improve the reproducibility of the results [7]. However, relatively few strategies are available to impute the affected intensities. Upton *et al.* and Arteaga-Salas discussed possible correction for the affected intensities in some samples [16,18]. However, the correction relied on biological replicates which were not always available. Imputation approaches based on intensities from the same array are still in great need, especially when working with rare or expensive arrays [13].

In this article, we propose an imputation approach for the intensities from defect areas. The proposed approach is single-array-based, which does not require any biological replicate. It models the cross-hybridization between probes and targets, and imputes the intensities by the binding affinities between them. In the following, we first explain our approach and further examine its performance with Affymetrix high-density SNP arrays. The performance of imputation is evaluated by genotyping accuracies.

Methods

Among current platforms, Affymetrix DNA microarrays have been commonly used for gene expression profiling, SNP genotyping and copy number estimation at a relatively high resolution with low-cost. In this article, we introduce our approach by using Affymetrix high density SNP arrays. We used SNP arrays as an example, because the performance of imputation can be easily examined by genotyping accuracies.

The Affymetric SNP array is a major platform that was used in the international Haplotype Map (HapMap) projects [19]. The genotypes of HapMap samples are commonly viewed as a gold-standard for the assessment of many genotyping approaches. However, we still found a number of defect areas in a few HapMap arrays, showing bright spots in the CEL images. Though the defect areas are relatively small in HapMap samples, a large number of SNPs could be affected. The intensities from these defect areas are not biologically meaningful and may lead to large genotyping errors and bias in copy number estimation. Most of current genotyping approaches used multi-array training [20–24], and relied on the between array normalization to take care of the image defects [25]. It is rarely evaluated how the genotyping results are affected by these defect areas.

The design of Affymetrix SNP arrays

Affymetrix SNP arrays use multiple probe_sets to capture the property of each SNP. Here, we use Affymetrix Mapping 100 K arrays as an example to illustrate the design. In a 100 K array, ten quartets are used to interrogate a single dimorphic SNP site with alleles commonly denoted as *A* and *B*. Each quartet consists of 4 types of probes that are 25-mer in length, either perfectly matched to the target or mismatched at a particular SNP site for each allele: perfect match *A*, mismatch *A*, perfect match *B* and mismatch *B*, denoted respectively PA, MA, PB and MB for short. The quartets have different shifts (k) of the nucleotide (k may take the values -4, -3, -2, -1, 0, 1, 2, 3, 4) from the center of the probe sequence ($k = 0$ at position 13 of the 25-mer), see Figure 1 for detailed illustration (Figure 1 was adapted from Figure 1A of [26]). The Affymetrix Mapping 500 K SNP Array has a similar design with 100 K arrays, but only 6 quartets are used to interrogate each SNP instead of 10.

In order to account for potential spatial effect on an array, the probe_sets of each SNP are usually distributed evenly over the array. The intensities for these probe_sets are highly correlated and inherently determined by SNP alleles. Unless the majority of array is defective, most of the probe_sets will not be affected. Therefore, it is possible to impute the affected intensities with those that are unaffected.

PICR model for copy number estimation and genotype calling

In the past few years, a number of genotyping approaches, such as BRLMM and CRLMM, have been proposed for Affymetric high density SNP arrays [20,21]. Recently, a single-array approach was proposed for copy number estimation and SNP genotype calling, referred to as Probe Intensity Composite Representation (PICR) [27]. It was shown previously that PICR attained higher genotyping accuracies than a commonly adopted approach, CRLMM. The PICR approach is based on single array training, and estimates copy number abundance on a single SNP basis. It addresses the cross-hybridization through the binding free energies and affinities between probes and targets. The free energy is calculated based on either perfect match or mismatch binding. It should be noted that a target sequence with allele *A* could

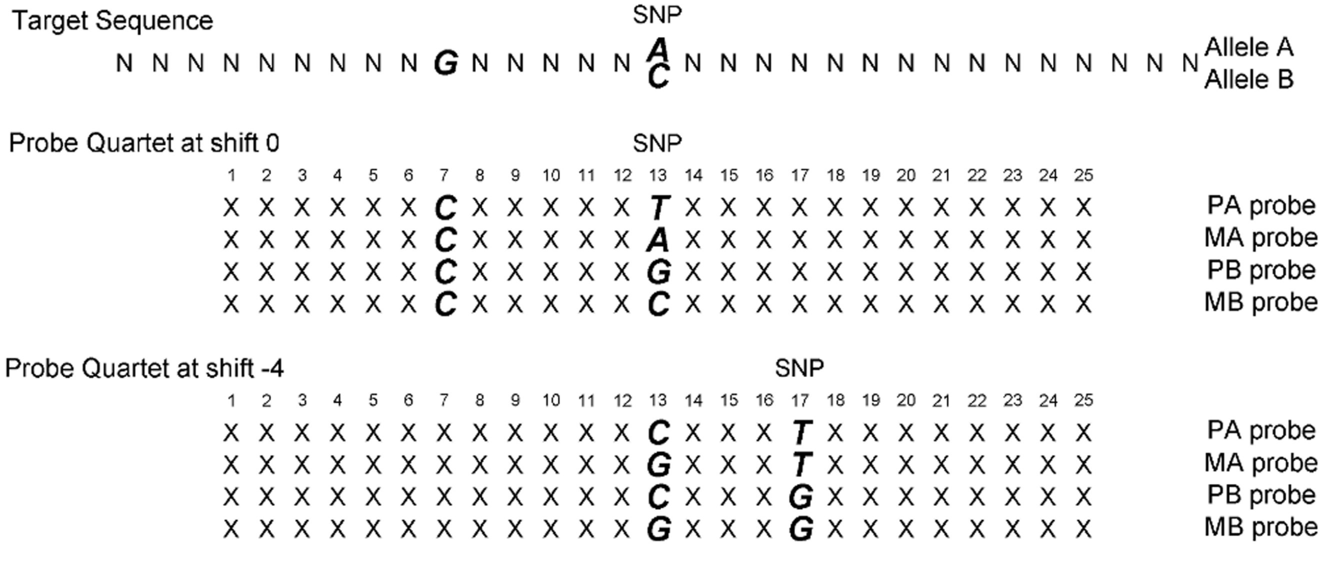
Target Sequence
SNP
N N N N N N N N N G N N N N N A/C N N N N N N N N N N N N N N N
Allele A
Allele B
Probe Quartet at shift 0
SNP
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
X X X X X X C X X X X X T X X X X X X X X X X X X
X X X X X X C X X X X X A X X X X X X X X X X X X
X X X X X X C X X X X X G X X X X X X X X X X X X
X X X X X X C X X X X X C X X X X X X X X X X X X
PA probe
MA probe
PB probe
MB probe
Probe Quartet at shift -4
SNP
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
X X X X X X X X X X X X C X X X T X X X X X X X X
X X X X X X X X X X X X G X X X T X X X X X X X X
X X X X X X X X X X X X C X X X G X X X X X X X X
X X X X X X X X X X X X G X X X G X X X X X X X X
PA probe
MA probe
PB probe
MB probe

Figure 1. Twenty-five-mer oligonucleotides which are perfectly matched or mismatched to the target sequence with SNP allele A or B.
doi:10.1371/journal.pone.0058677.g001

hybridize to PA probe through perfect match binding and hybridize to PB, MA or MB probes through mismatch binding. Importantly, the hybridization to mismatch probes can have two mismatch nucleotides when shift $k \neq 0$ (Figure 1). Given the quartet (PA, MA, PB, MB) with shift k, the corresponding binding free energy is calculated with a positional-dependent nearest neighbor (PDNN) model [28,29].

Following the same notation with Wan *et al.*, we first briefly introduce PICR model, the detail of which can be found elsewhere [27]. A target sequence TA with allele A could hybridize to PA probe through perfect match binding, the binding free energy of which is :

$$E(\mathrm{TA},\mathrm{PA}) = \sum_{p=1}^{24} \omega_p \lambda(b_p^{\mathrm{PA}}, b_{p+1}^{\mathrm{PA}}) \tag{1}$$

where $\mathrm{PA} = (b_1^{\mathrm{PA}}, b_2^{\mathrm{PA}}, \cdots, b_{25}^{\mathrm{PA}})$ is the 25-mer PA probe, ω_p is a weight factor that depends on the position of consecutive bases along the probe, and λ represents stacking energy depending on nearest neighbor along the probe.

Further, a target sequence TA with allele A could hybridize to MA, PB, MB ($k=0$) probes through mismatch binding with one unique mismatch. For example, the binding free energy between TA and MA is

$$E_1(\mathrm{TA},\mathrm{MA}) = \sum_{\substack{l=1 \\ l \neq 12,13}}^{24} \theta_l \lambda(b_l^{\mathrm{MA}}, b_{l+1}^{\mathrm{MA}}) + \delta(\{b_{12}^{\mathrm{MA}} b_{13}^{\mathrm{MA}} b_{14}^{\mathrm{MA}}\}, \\ \{b_{12}^{\mathrm{TA}} b_{13}^{\mathrm{TA}} b_{14}^{\mathrm{TA}}\}) \tag{2}$$

where λ is the same stacking energy in Equation (1) and θ_p is a position weight factor for mismatch. $\delta$is the corresponding stacking energy for the triplets at mismatch position.

Finally, a target sequence TA with allele A could hybridize to MB probes through mismatch binding with two mismatches when shift $k \neq 0$. The mismatch would occur at both the center and shift k of the probe. The binding free energy is:

$$E_2(\mathrm{TA},\mathrm{MB}^k) = E_1(\mathrm{TA},\mathrm{MB}^0) + \xi^k(\{b_{12+k}^{\mathrm{MB}^k} b_{13+k}^{\mathrm{MB}^k} b_{14+k}^{\mathrm{MB}^k}\}, \\ \{b_{12+k}^{\mathrm{TA}} b_{13+k}^{\mathrm{TA}} b_{14+k}^{\mathrm{TA}}\}), \tag{3}$$

The term $E_1(\mathrm{TA},\mathrm{MB}^0)$ in Equation (3) is the binding free energy if there is only one unique mismatch at the center, and the term ξ^k is the binding free energy of the triplets at the second mismatch.

Eventually, a linear regression model was used to model the probe intensities of quartets:

$$\begin{aligned}
&\vdots \\
&I_{\mathrm{PA}^k} = N_A \varphi(E(\mathrm{TA},\mathrm{PA}^k)) + N_B \varphi(E_1(\mathrm{TB},\mathrm{PA}^k)) + b + \varepsilon_{\mathrm{PA}^k} \\
&I_{\mathrm{PB}^k} = N_A \varphi(E_1(\mathrm{TA},\mathrm{PB}^k)) + N_B \varphi(E(\mathrm{TB},\mathrm{PB}^k)) + b + \varepsilon_{\mathrm{PB}^k} \\
&I_{\mathrm{MA}^k} = N_A \varphi(E_1(\mathrm{TA},\mathrm{MA}^k)) + N_B \varphi(E_{t_k}(\mathrm{TB},\mathrm{MA}^k)) + b + \varepsilon_{\mathrm{MA}^k} \\
&I_{\mathrm{MB}^k} = N_A \varphi(E_{t_k}(\mathrm{TA},\mathrm{MB}^k)) + N_B \varphi(E_1(\mathrm{TB},\mathrm{MB}^k)) + b + \varepsilon_{\mathrm{MB}^k} \\
&\vdots
\end{aligned} \tag{4}$$

where $t_k = 2 \ if \ k \neq 0;\ t_k = 1 \ if \ k = 0;$

where $(I_{\mathrm{PA}^k}, I_{\mathrm{PB}^k}, I_{\mathrm{MA}^k}, I_{\mathrm{MB}^k})$ is the probe intensities of a quartet with shift k. E is the binding free energy between the probe and the target. $\varphi(x) = 1/(1+\exp(x))$ gives the corresponding binding affinity in the form of Langmuir adsorption [28,30]. The coefficients N_A and N_B are copy number abundance of target sequences with allele A and allele B, respectively. The model intercept b represents the baseline intensity. The error terms $(\varepsilon_{\mathrm{PA}}^k, \varepsilon_{\mathrm{MA}}^k, \varepsilon_{\mathrm{PB}}^k, \varepsilon_{\mathrm{MB}}^k)$ are independent measurement errors of intensities, each following a normal distribution with mean 0 and a constant variance. All the parameters are trained by a single HapMap sample, and are entirely determined by probe sequences. Empirically, we found the parameters were robust across different samples and different platforms, providing unbiased estimation of copy number abundance [27]. For a randomly selected HapMap sample, the linear model had a median R-square value of 0.764.

Mixed effect model for multiple SNPs

The PICR model is a single-SNP, single-array approach, which was shown to have greater genotyping accuracies than a commonly used approach, CRLMM [27]. However, it also has a few limitations. First, the PICR model assumes the probe_sets of each SNP are mutually independent, which may not always be valid. Second, the PICR model utilized a linear regression framework with a very limited sample size (40 for 100 K array, 24 for 500 K array). It is less robust with outliers, such as probe intensities from defect areas. In this study, we extended the PICR model to a multi-SNP setting with a mixed effect model. In the extended model, we take into account the correlation among probe_sets for each SNP. Further, by adopting a mixed effect model, we borrow strength from multiple SNPs and improve its robustness to outliers. Finally, we are able to impute the intensities from defect areas by using a large number of SNPs that are not affected.

Assuming we have a SNP array with a large number of N SNPs, each has M probe_sets ($M=40$ for 100 K SNP arrays). Each probe_set has binding affinities $f_{ij,\mathrm{A}}$ and $f_{ij,\mathrm{B}}$ for target sequences TA and TB, respectively. The binding affinities can be calculated based on the equations described in Section 2.2. For example, if the j-th probe is a PA probe with shift k, then$f_{ij,\mathrm{A}} = \varphi(E(\mathrm{TA},\mathrm{PA}^k))$, $f_{ij,B} = \varphi(E_1(\mathrm{TB},\mathrm{PA}^k))$.We model the intensity values for the j-th probe_set of the i-th SNP with a mixed effect model:

$$I_{ij} = \beta_0 + \beta_{\mathrm{A}} f_{ij,\mathrm{A}} + \beta_{\mathrm{B}} f_{ij,\mathrm{B}} + \alpha_{i0} + \alpha_{i,\mathrm{A}} f_{ij,\mathrm{A}} + \alpha_{i,\mathrm{B}} f_{ij,\mathrm{B}} + \varepsilon_{ij};$$

where I_{ij} is the intensity for the j-th probe of SNP i ; $i = 1, 2 \ldots\ldots N$; $j = 1, 2, \ldots\ldots 40$; and $f_{ij,\mathrm{A}}$, $f_{ij,\mathrm{B}}$ are the binding affinities of the j-th probe of SNP i with respect to two allele A and B, respectively. Here, β_0, β_{A}, β_{B} are the fixed effect for the baseline intensity, copy number abundance for allele A and copy number abundance for allele B, respectively; $\alpha_{i0}, \alpha_{i,\mathrm{A}}, \alpha_{i,\mathrm{B}}$ are the corresponding random effect subjected to SNP-wise variability with $\alpha_{i0} \sim N(0,\sigma_0^2)$, $\alpha_{i,\mathrm{A}} \sim N(0,\sigma_{\mathrm{A}}^2)$, and $\alpha_{i,\mathrm{B}} \sim N(0,\sigma_B^2)$; ε_{ij} is a random measurement error with $\varepsilon_{ij} \sim N(0,\sigma^2)$. By using the above mixed effect model, the intensity values are assumed to be independent across different SNPs, but correlated within the same SNP. The fixed effect β_{A}, β_{B} has an interpretation as the average copy number abundance of the N SNPs of interest. Further, $\beta_{\mathrm{A}} + \alpha_{i,\mathrm{A}}, \beta_{\mathrm{B}} + \alpha_{i,\mathrm{B}}$ has an interpretation as the copy number abundance for the i-th SNP. The random effect $\alpha_{i,\mathrm{A}}, \alpha_{i,\mathrm{B}}$ is a constant for all the available probes of SNP i, but varies across different SNPs, the variance of

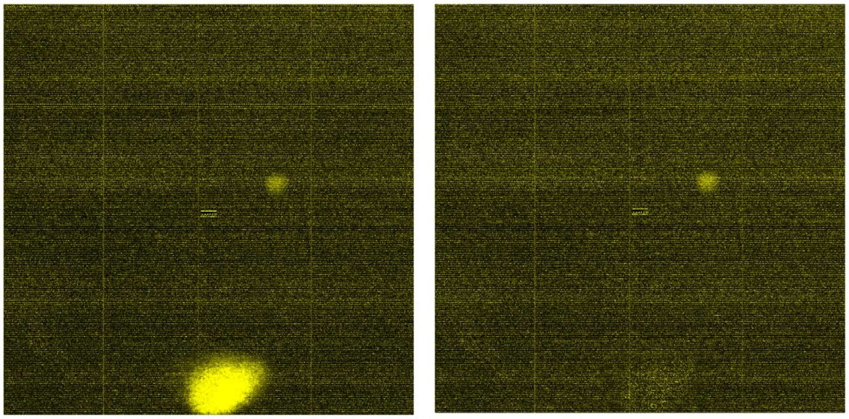

Figure 2. CEL images before and after imputation. Left: before, Right: after.
doi:10.1371/journal.pone.0058677.g002

which are estimated by all probe intensities of N SNPs in the model.

Based on the above mixed effect model, an imputation procedure can be conducted as follows: 1) First, defect areas are identified by examining a CEL image or using existing image processing software, such as harshlight and BMR. The intensities from defect areas will be set as missing. 2) Second, for each affected SNP, a large number of unaffected SNPs (e.g. 100 SNPs) will be randomly selected to fit a mixed effect model. 3) The missing values can be imputed by the predicted values based on the mixed effect model.

Results

We considered 90 Affymetrix Mapping_100 K_Xba arrays from the HapMap study. D-ChIP was first used to examine the CEL image for each sample [31,32], and 12 samples were identified with potential defect areas. Here, we only presented the CEL image for one sample (NA12812) as an example. The CEL images of the other 11 samples can be found in Figures S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, and S11. The CEL image before and after imputation were given in Figure 2. Before imputation, a bright spot was identified at lower part of the image (Figure 2). Further examination indicated that over 20,000 SNPs were affected. However, for most of the SNPs, the number of affected probes is less than 10. We used a rectangle area to cover the bright spot and set the intensities from the defect area as missing. The data was then imputed according to the procedure described above. After imputation, the damaged area was recovered (Figure 2).

In order to examine the quality of imputation, PICR was applied to conduct genotyping for both the original CEL and the imputed CEL. Figure 3 showed the estimation of allelic copy number abundance before and after imputation. PICR model conducted genotype calling by clustering all SNPs into three groups, corresponding to three possible genotypes, AA, AB and BB. The genotyping error rate was calculated as the discordant rate between PICR-genotype-calls and HapMap gold standards. Before imputation, applying PICR yielded a genotyping error rate of 2.54%, while the average genotyping error rate was 0.4% among all 90 samples. After imputation, the genotyping error rate was reduced by over 5 folds, to a normal level of 0.44%. The genotyping error rates for all 12 problematic samples were listed in Table 1. The results showed that the error rates were reduced for all 12 samples, especially for those with an error rate greater than 1% before imputation. Further examination showed the estima-

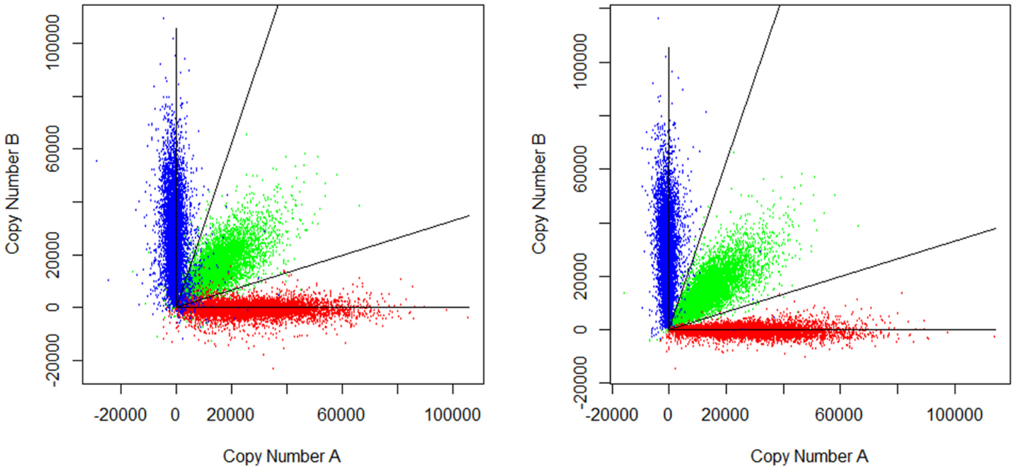

Figure 3. Allelic copy number abundance before and after imputation. Left: before, Right: after; Red: genotype AA, Blue: genotype BB, Green: genotype AB
doi:10.1371/journal.pone.0058677.g003

Table 1. Genotyping error before and after imputation.

Sample ID	# of SNP affected	Ave. # of probe affected/ SNP	Error Rate Before Imputation	Error Rate After Imputation
NA12812	20075	5.67	2.54%	0.45%
NA10835	20925	3.76	1.56%	0.32%
NA12239	12758	3.34	1.05%	0.26%
NA12144	7330	3.22	0.95%	0.24%
NA12005	4404	3.01	0.83%	0.65%
NA12056	7490	3.28	0.87%	0.85%
NA12146	7657	3.16	0.73%	0.25%
NA12155	9394	3.18	0.71%	0.71%
NA07056	2026	2.94	0.70%	0.54%
NA12236	8071	3.19	0.66%	0.65%
NA12813	4788	3.15	0.61%	0.25%
NA10863	5050	3.07	0.56%	0.55%

doi:10.1371/journal.pone.0058677.t001

tion of allelic copy number abundance was substantially improved after imputation.

According to the number (proportion) of affected probes, we further classified all SNPs into four groups, including 0 affected probe, 1–4 affected probes (<10%), 5–8 affected probes (10%–20%) and >8 affected probes (>20%). The result was summarized in Table 2. Most of the affected SNPs had 1–4 affected probes (<10%). Before imputation, the error rates increased dramatically with the number of affected probes, from 3.3% (1–4 affected probes) to 9.19% (>8 affected probes). After imputation, the error rates were significantly reduced for all groups. While the error rate was still considerably lower for SNPs in <10% group, no significant difference was found between 10–20% group and >20% group. The imputation appeared to be effective for all groups, and was able to reduce the error rate by as high as 13 folds. For Affymetrix 500 K SNP arrays, we expected the number of affected probes to be proportional. A similar defect area would affect a lot more SNPs, but less number of probes for each SNP.

Discussion

In this study, we have proposed a multi-SNP approach for imputing the intensities from defect areas on Affymetrix SNP arrays. The approach can be viewed as an extension of previously proposed PICR model. Similar to PICR, it is a single-array approach, which does not require any biological replicate or between-array normalization. It would be especially helpful for small studies with limited sample sizes. Further, the results have showed that this approach is effective to impute intensities from defect areas. The genotyping error rates were significantly reduced.

In this article, we have focused on Affymetrix SNP arrays, because the performance can be easily examined by genotyping error rates. However, our approach models the fundamental mechanism of physical binding between DNA nucleotides. It can potentially be extended to other oligonucleotide microarrays with a similar design. For example, the gene expression arrays have similar PM/MM probe_sets, with the number of pairs ranging from 11–16. To extend the proposed approach to gene expression microarrays, the binding free energy will need to be re-calculated. Zhang *et al.* have investigated the sequence-specific binding and non-specific binding for gene expression data, and provided estimation for the corresponding binding free energy [28,29]. Further adjustment will also be necessary to accommodate the diverse number of probe_sets for each gene/transcript, and the competing hybridization process of DNA sequences from various experimental conditions.

In our study, we also found that the genotyping error rates for a few arrays remained similar after imputation, e.g. NA10853. One reason was that the defect area was relatively small and only a small number of SNPs were affected. Given the large number of total SNPs on the array, it did not have a great impact on the overall genotype calling. Another possible reason is that the genotyping errors can be caused by multiple sources. For those arrays, there might be other possible factors affecting genotyping accuracies.

Table 2. Genotyping error rates stratified by number of affected probes.

# of affected probes	0	1–4 (<10%)	5–8 (10%–20%)	>8 (>20%)
# of SNPs	415417 (84.5%)	65458 (13.3%)	8313 (1.7%)	2688 (0.5%)
# of errors before imputation	1871	2165	504	247
Error rates before imputation	0.45%	3.33%	6.06%	9.19%
# of errors after imputation	1871	379	74	19
Error rates after imputation	0.45%	0.58%	0.89%	0.71%
Fold Change of Error rates	1	5.7	6.8	13

doi:10.1371/journal.pone.0058677.t002

One remaining problem of the proposed approach is to efficiently identify the boundary for defect areas. The defect areas on an array may have irregular shapes. In this article, we have used a rectangle area to cover the defect area for imputation. However, these rectangles were slightly larger than the bright spots and some unaffected intensities are set to missing as well. It will be helpful to identify the boundary of defect areas effectively and accurately. In addition, other studies have shown that some outliers can not be easily identified by checking CEL images. Several computational algorithms have emerged as promising tools for detecting defect areas on a CEL image, such as 'harshlight' and 'BRB'. These available packages are potentially helpful to identify the outliers automatically.

Supporting Information

Figure S1 The CEL image before and after imputation for sample NA07056.
(TIF)

Figure S2 The CEL image before and after imputation for sample NA10835.
(TIF)

Figure S3 The CEL image before and after imputation for sample NA10863.
(TIF)

Figure S4 The CEL image before and after imputation for sample NA12005.
(TIF)

Figure S5 The CEL image before and after imputation for sample NA12056.
(TIF)

Figure S6 The CEL image before and after imputation for sample NA12144.
(TIF)

Figure S7 The CEL image before and after imputation for sample NA12146.
(TIF)

Figure S8 The CEL image before and after imputation for sample NA12155.
(TIF)

Figure S9 The CEL image before and after imputation for sample NA12236.
(TIF)

Figure S10 The CEL image before and after imputation for sample NA12239.
(TIF)

Figure S11 The CEL image before and after imputation for sample NA12813.
(TIF)

Acknowledgments

We thank four anonymous reviewers for their helpful suggestions which improved this manuscript.

Author Contributions

Conceived and designed the experiments: ML WF. Performed the experiments: ML WF. Analyzed the data: ML YW WF. Contributed reagents/materials/analysis tools: ML YW QL WF. Wrote the paper: ML QL WF.

References

1. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.
2. Hacia JG, Fan JB, Ryder O, Jin L, Edgemon K, et al. (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet 22: 164–167.
3. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23: 41–46.
4. Brown CS, Goodwin PC, Sorger PK (2001) Image metrics in the statistical analysis of DNA microarray data. Proc Natl Acad Sci U S A 98: 8944–8949.
5. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, et al. (2002) Fully automatic quantification of microarray image data. Genome Res 12: 325–332.
6. Wu Z (2009) A review of statistical methods for preprocessing oligonucleotide microarrays. Stat Methods Med Res 18: 533–541.
7. Suarez-Farinas M, Pellegrino M, Wittkowski KM, Magnasco MO (2005) Harshlight: a "corrective make-up" program for microarray chips. BMC Bioinformatics 6: 294.
8. Naef F, Lim D, Patil N, Magnasco MO (2001) From features to expression: High density oligonucleotide array analysis revisited. Technical Report 1: 1–9.
9. Hubbell E, Liu WM, Mei R (2002) Robust estimators for expression analysis. Bioinformatics 18: 1585–1592.
10. Wu Z, Irizarry RA, Gentleman R, Murillo FM, Spencer F (2004) A model based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc 99: 909–917.
11. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
12. Reimers M, Weinstein JN (2005) Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. BMC Bioinformatics 6: 166.
13. Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJ, Harrison AP (2008) An overview of image-processing methods for Affymetrix GeneChips. Brief Bioinform 9: 25–33.
14. Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (2003) The analysis of gene expression data: methods and software. In Statistics for Biology and Health. New York Springer. pp. 455.
15. Affymetrix Inc. (2004) Gene Chip Expression Analysis: Data Analysis Fundamentals. Affymatrix.
16. Upton GJ, Lloyd JC (2005) Oligonucleotide arrays: information from replication and spatial structure. Bioinformatics 21: 4162–4168.
17. Song JS, Maghsoudi K, Li W, Fox E, Quackenbush J, et al. (2007) Microarray blob-defect removal improves array analysis. Bioinformatics 23: 966–971.
18. Arteaga-Salas JM, Harrison AP, Upton GJ (2008) Reducing spatial flaws in oligonucleotide arrays by using neighborhood information. Stat Appl Genet Mol Biol 7: Article29.
19. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
20. Affymetrix Inc.. (2006) BRLMM: An improved genotype calling method for the GeneChip Human Mapping 500K Array set. Affymetrix.
21. Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 8: 485–499.
22. Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, et al. (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. Bioinformatics 21: 1958–1963.
23. Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22: 7–12.
24. Xiao Y, Segal MR, Yang YH, Yeh RF (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. Bioinformatics 23: 1459–1467.
25. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
26. Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, et al. (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Res 14: 414–425.
27. Wan L, Sun K, Ding Q, Cui Y, Li M, et al. (2009) Hybridization modeling of oligonucleotide SNP arrays for accurate DNA copy number estimation. Nucleic Acids Res 37: e117.
28. Zhang L, Miles MF, Aldape KD (2003) A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol 21: 818–821.
29. Zhang L, Wu C, Carta R, Zhao H (2007) Free energy of DNA duplex formation on short oligonucleotide microarrays. Nucleic Acids Res 35: e18.

30. Vainrub A, Pettitt BM (2002) Coulobm blockage of hybridization in two-dimensional DNA arrays. Physics Review E 66.
31. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98: 31–36.
32. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, et al. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20: 1233–1240.