The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Model Based Probe Fitting and Selection for SNP Array

Ling-Yun Wu[1,2,*] Xiaobo Zhou[1,†] Chung-Che Chang[3]
Stephen T.C. Wong[1]

[1]Center for Biotechnology & Informatics and Department of Radiology, The Methodist Hospital Research Institute, Weill Medical College, Cornell University, Houston, TX 77030, USA
[2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[3]Department of Pathology, The Methodist Hospital, Weill Medical College, Cornell University, Houston, TX 77030, USA

Abstract Recent advances of high-throughput SNP arrays such as Affymetrix's GeneChip Human Mapping 500K array set have made it possible to genotype large samples in a fast and cheap manner. A lot of algorithms were developed to call the genotypes from SNP array. When considering the low level preprocessing of SNP array, most algorithms just borrow the techniques from the gene expression microarray. As in the analysis of gene expression microarray, the low level preprocessing of SNP array, e.g. probe summarization, is very important in the analysis of SNP array such as genotyping, loss of heterozygous (LOH) inference, and copy number inference. In this paper, we present a model based method for probe fitting and selection of SNP array. This method exploits the abundant high quality genotypes such as HapMap data, which were genotyped and validated by several independent genotyping techniques. The new probe summarization method can be used with the existing genotyping methods to improve the accuracy.

Keywords SNP array; Genotyping; Probe Fitting

1 Introduction

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation between individuals. In the past few years, SNPs are increasingly being utilized in the genome-wide association studies (GWAS) [8], which aims to identify the genetic variations that attribute to phenotypic traits such as complex diseases. Recent advances of oligonucleotide microarray techniques have enabled the simultaneously genotyping of hundreds of thousands of SNPs. Great efforts have been made to improve the high-throughput genotyping technology. Currently the widely used SNP arrays are Affymetrix's GeneChip SNP arrays and Illumina's BeadChip SNP arrays [15]. Affymetrix's Genome-Wide Human SNP Array 6.0 includes probes for 906,600 SNPs and 946,000

*The work of first author is supported by National Natural Science Foundation under Grant No. 60503004 and the K. C. Wong Education Foundation, Hong Kong.
†Corresponding author. Email: XZhou@tmhs.org

non-polymorphic copy number probes. Illumina's High Density Human 1M-Duo chip can genotype more than 1 million SNPs. The rapid progress of high-throughput SNP genotyping technologies presents many statistical and computational challenges, such as the genotyping algorithm [16] and the copy number variation analysis [13]. In this paper, we focus on the genotyping algorithms for Affymetrix platform.

There are many genotyping algorithms have been developed for Affymetrix platform. Several algorithms were developed by Affymetrix as the official softwares of the SNP arrays: MPAM [12] for their first-generation 10K SNP arrays, DM [6] for the 100K SNP arrays, BRLMM [2] for the 500K SNP arrays, BRLMM-P [3] for the SNP Array 5.0, and Birdseed [1] for the latest SNP Array 6.0. A lot of genotyping algorithms were also developed by third party scientists and researchers, such as RLMM [16], PLASQ [9], GEL [14], CRLMM [4], SNiPer-HD [7], MAMS [17], CHIAMO [5]. These algorithms can be classified into two categories: single-array based method and multi-array based method. The single-array methods, which analyze one array at a time, include DM and GEL. In multi-array based method, the arrays are clustered into three genotypes for each SNP. Most existing genotyping algorithms are multi-array based.

The first and most critical step of SNP array analysis is the preprocessing of probe intensities. In the design of Affymetrix SNP platform, there are many for each SNP in the array, e.g. 24 or 40 probes in 500K arrays. The preprocessing procedure includes normalizing the probe intensities, background adjustment, and summarizing the probe intensities for each SNP. Most genotyping algorithms just borrow the techniques from the gene expression microarray, such as RMA [4], logarithmic additive model [2], median method [17], and so on. In this paper, we present a model based method for probe fitting and selection of SNP array. This method exploits the abundant high quality genotypes such as HapMap data, which were genotyped and validated by several independent genotyping techniques.

Firstly, the analysis identified a set of probes which have very low responses to the specific hybridization. These probes can be used to evaluate the quality of SNP array and do normalization between arrays, since their intensities are mainly due to non-specific hybridization and noise. Secondly, the probes are classified as high or low performance probes. The poorly performing probes with erratic hybridization behavior are discarded. When summarizing the probe intensities, the remaining probes are given weights according to their consistency among large samples. The new probe summarization method can be used with the existing genotyping methods to improve the accuracy. A simple genotyping method is developed based on the new summarization method. Although its simplicity, the new genotyping method can achieve very high accuracy and call rate compared with most existing genotyping algorithms. This concludes that the new probe summarization method reduces the difficulty of genotyping.

2 Methods

2.1 Motivations

Although Affymetrix has carefully selected the SNPs and probes for each SNP with high specificity and sensitivity, base on some rules [6], inevitably there remain some probes with high non-specific cross-hybridization phenomena. For example, as shown in Figure 1, for SNP_A-4199815, the probe 1,2,4,6 of allele A and probe 1,2,3,6 of allele B

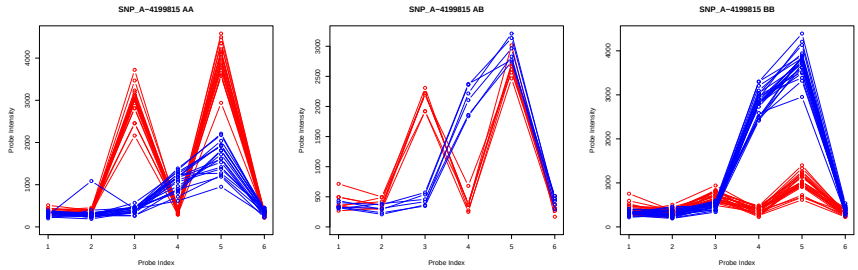

Figure 1: Probe intensities of SNP_A-4199815 in 48 HapMap samples, each figure for one genotype (from left to right): AA, BB, AB. Red lines for allele A and blue lines for allele B.

do not well respond to the true copy number of alleles. In the proposed method, we fit the probe intensities using the high quality HapMap data with known genotypes. Then the probes with unpredictable behaviors are selected and excluded from the consequent genotyping process.

2.2 Model

In this paper, only the perfect match probes are considered since there are only perfect match probes in the latest SNP array of Affymetrix. For each allele of certain SNP, the probe intensities are fitted using the following additive model:

$$y_{ij}^{(ks)} = b_i^{(ks)} + r_i^{(ks)}\theta_j^{(ks)} + \varepsilon_{ij}^{(ks)}$$

where s is the index of SNP in the array, $k = \{A,B\}$ is the index of allele for the SNP, $i = 1,2,\cdots,I^{(ks)}$ is the index of probes for the allele k of SNP s, $j = 1,2,\cdots,J$ is the index of arrays (samples). Here $y_{ij}^{(ks)}$ denotes the raw intensity of the ith probe in the jth array for the allele k of SNP s, $b_i^{(ks)}$ is the baseline response of the ith probe due to nonspecific hybridization, $r_i^{(ks)}$ is the rate of increase of the ith probe response to the sample, $\theta_j^{(ks)}$ is the copy number of this allele in the jth sample, and $\varepsilon_{ij}^{(ks)}$ is the random error that is assumed following normal distribution. Since each time we only consider one allele of one SNP, the superscripts (ks) are omitted hereafter.

We fit this model to the HapMap SNP arrays with known genotypes. We assume all HapMap samples are normal diploid, i.e. copy number equals to 2. The value of θ_j for each allele in the jth sample is calculated from the corresponding genotype of that SNP. In detail, if genotype is AA, $\theta_j^A = 2$ and $\theta_j^B = 0$; if genotype is BB, $\theta_j^A = 0$ and $\theta_j^B = 2$; if genotype is AB, $\theta_j^A = 1$ and $\theta_j^B = 1$. The SNPs in the non-pseudo-autosomal chromosome X region are treated specially for all male samples.

2.3 Model Fitting

We fit the model to estimate the parameters for each probe. The probe and array outliers are identified iteratively. Specifically, we first fit the model to the probe set for each allele of SNPs. The probes with large residual standard error (more than three times

as large as the median standard error of all probes) are marked as outliers. The probes with negative estimate of r_i or negative adjusted R-squared value are also regarded as outliers. After removing outlier probes, the model is applied to estimate the values of θ_j for each array. Similarly, the arrays with large residual standard error are marked as outliers. This procedure is repeated until there is no anymore new probe or array outlier. Since the variances of probe errors are varied, the weighted linear regression is used to fit the model for each array so that the probes with small error variances are given larger weights. We use the adjusted R-Squared value of each probe as the weight. By this way, the probes are given weights according to their consistency among large samples.

2.4 Normalization

The normalization is necessary for properly fitting model. We use two normalization methods, one for the initial iteration and the other for the rest iterations. In the first run, all arrays are normalized by simple linear scaling method so that the mean intensity of all perfect match probes of SNPs in autosomal chromosomes is equal to 1000. After the first run, a set of probes that do not response to the specific alleles are identified by the following criteria: the p-value of coefficient r_i larger than 0.1, r_i smaller than 50. In other words, the intensities of these probes mainly attribute to the non-specific hybridizing and have lower correlation with the sample. By using this probe set as an invariant set, a linear regression is applied to normalize each array to a selected reference array.

2.5 Genotyping

Actually, all existing genotyping algorithms can be applied to call genotypes based on the probes after fitting and selection. The selection of probes may improve the accuracy of genotyping algorithms. Here we developed a fast method to call the genotypes from the fitted model for each SNP. The method is based on the contrast value, which is defined as follows:

$$C_j = \frac{\theta_j^A - \theta_j^B}{\theta_j^A + \theta_j^B}$$

Then the genotype call G_j is inferred from C_j by predetermined thresholds:

$$G_j = \begin{cases} AA, & \text{if } C_j \geq 1/3, \\ BB, & \text{if } C_j \leq -1/3, \\ AB, & \text{otherwise.} \end{cases}$$

3 Results

3.1 Data Preparation

The training data used to fit the probe intensities are 270 HapMap 500K SNP arrays, which are downloaded from the HapMap websites. The genotype calls for these arrays are mainly produced from HapMap genotypes, which are detected by using several independent genotyping methods and believed to be very accurate [16, 7, 11, 10]. First, 1875 SNPs whose alleles are obviously flipped in HapMap data are corrected. Then the SNP that is NoCall in HapMap data is assigned BRLMM call or DM call, if they are consistent, i.e. they are equal or one of them is NoCall.

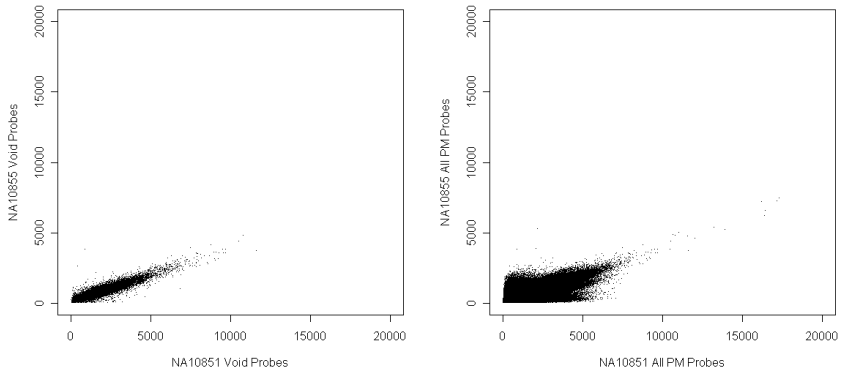

Figure 2: Scatter plot of void probes of two HapMap arrays (left), compared with that of all perfect match probes (right).

3.2 Void Probes

As shown in Figure 1, there are some probes that do not response to the target SNP, i.e. the intensities of these probes show no correlation with the genotypes of SNP. We call these probes as *Void Probes*. The void probe is identified when the t-test p-value of the estimated coefficient r_i is insignificant (larger than 0.1). By this way, about 3% probes ($\sim 82,000$) are found as void probes. These void probes can also be used as invariant set to do normalization between arrays, and to measure the quality of each array. Figure 2 show a scatter plot of void probes of two arrays compared with that of all perfect match probes. Obviously, there are strong correlations between the void probes of two arrays. The Pearson correlation of the void probes is 0.95, while that of all perfect match probes is 0.78.

3.3 Genotyping Results of HapMap Data

We used the proposed method to genotype another set of 39 HapMap 500K arrays provided in the Affymetrix' s website. The genotypes downloaded from HapMap website are used as the gold standard. The new method achieves accuracy 99.65832% at call rate 96.77113%. Compared with BRLMM, which obtains accuracy 99.8454% at call rate 99.65887%, the performances of new method is competitive and acceptable. One of the reasons that the call rate of new method is lower than BRLMM is that there are many missing genotypes in the training data. We will try to establish a better training dataset by integrating the HapMap genotypes with the results of other genotyping softwares such as BRLMM.

3.4 3.4 Genotyping Results of Tumor Data

In most of literature, only normal samples such as HapMap samples are used to evaluate the genotyping algorithms. However, the tumor samples may be very different from the normal samples since there are much more regions of DNA copy number gain and loss in the tumor samples than the normal samples. The variation of DNA copy number will

definitely influence the genotyping results. In order to test the performance of developed genotyping methods on the tumor data, we use several myelodysplastic syndromes (MDS) samples with significant copy number changes. The results of one patient are depicted in Figure 3. In this example, the tumor tissue is from the blast of patient MDS-8 while the normal reference tissue is from the lymphoid of the same patient. For the normal disomy chromosome 8 (bottom of Figure 3), both probe fit model and BRLMM algorithm show very good consistency between two samples (all spots distribute closely to the diagonal). The SNPs with genotype AA, AB, BB cluster at (1,1), (0,0), (-1,-1) respectively. For the monosomy chromosome 7 (top of Figure 3), since there is only one copy of chromosome, we expect there are no heterozygous genotypes. In the results of probe fit model (top left), the SNPs with genotypes AB in normal samples disperse to two groups, one centers at (-1, 0), the other at (1, 0), which represent LOH from AB to BB and AA respectively. The results of BRLMM also exhibits same pattern but two LOH groups are not separated clearly. Considering the genotype calls, the heterozygous calls in this chromosome should be caused by genotyping error. The heterozygous rates of the genotyping methods based on probe fit model is 5.7% respectively, while BRLMM calls 11.7% heterozygous SNPs. That is, the error rate of probe fit method is one fold less than BRLMM in this monosomy chromsome.

4 Conclusion

The probe summarization is critical step of the genotype calling algorithm based on SNP array. The existing genotyping algorithms borrow the probe summarization techniques used in gene expression microarray analysis, including simple summation, median, and log additive model. All of them do not exploit the availability of high quality genotypes such as HapMap data, which does not exist in the field of gene expression microarray. In this paper, a novel probe summarization and selection method is presented, which can be used for downstream genotype calling, LOH inference, copy number analysis, and so on. By using the HapMap gentoypes as gold standard training data, the systematic differences in intensity between probes are described by several parameters. The poorly performing probes are easily identified and the probes are weighted according to their performance, i.e., the most informative and responsive probes are given high weights. The proposed probe summarization method can be used with any existing genotyping methods to improve the accuracy. The new probe summarization method reduces the difficulty of genotyping. A simple genotyping algorithm is developed to illustrate this point. Although its simplicity, the new genotyping algorithms can achieve very high accuracy and call rate compared with most existing genotyping algorithms.

References

[1] Affymetrix. Birdseed algorithm.

[2] Affymetrix. BRLMM: An improved genotype calling method for the GeneChip Human Mapping 500K array set. Affymetrix white paper, Affymetrix, 2006.

[3] Affymetrix. BRLMM-P: A genotype calling method for the SNP 5.0 array. Affymetrix white paper, Affymetrix, 2007.

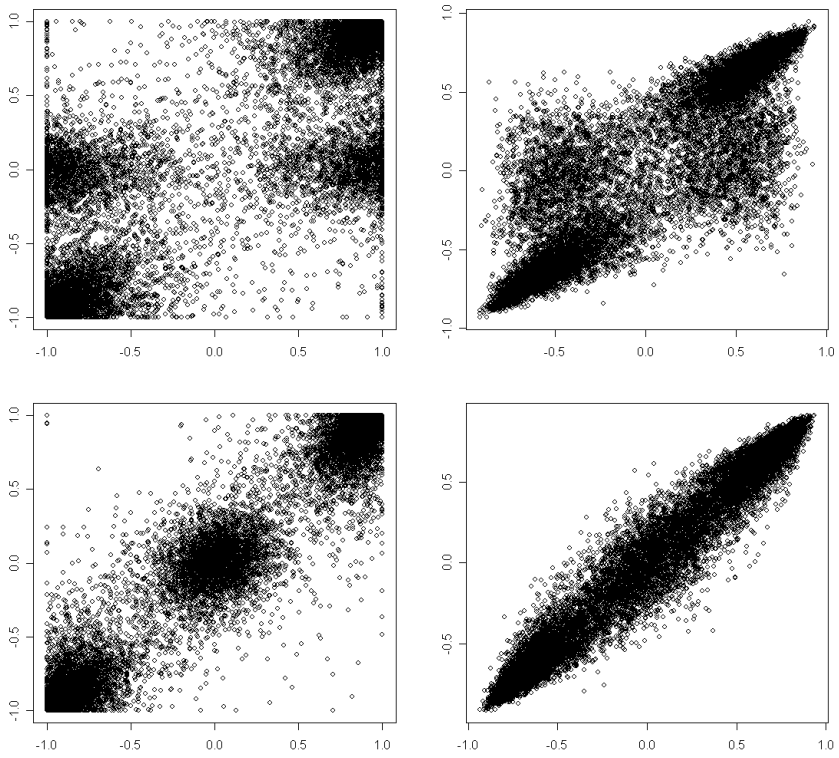

Figure 3: Scatter plots of contrast values of paired tumor/normal samples: MDS-8 Blast and MDS-8 Lymphoid. Axis y is the contrast of normal sample and axis x is the contrast of tumor sample. Both the results in monosomy chromosome 7 (top) and disomy chromosome 8 (bottom) are shown. Left: the results of probe fit model, right: the results of BRLMM.

[4] B. Carvalho, H. Bengtsson, T. P. Speed, and R. A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8(2):485–99, 2007.

[5] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 2007.

[6] 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 100k SNPs on oligonucleotide microarrays. *Bioinformatics*, 21(9):1958–63, 2005.

[7] J. Hua, D. W. Craig, M. Brun, J. Webster, V. Zismann, W. Tembe, K. Joshipura, M. J. Huentelman, E. R. Dougherty, and D. A. Stephan. SNiPer-HD: Improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics*, 23(1):57–63, 2007.

[8] D. J. Hunter, P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. B. Hayes, M. Tucker, D. S. Gerhard, Jr. Fraumeni, J. F., R. N. Hoover, G. Thomas, and S. J. Chanock. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, 39(7):870–4, 2007.

[9] 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):323–36, 2007.

[10] P. Lamy, C. L. Andersen, F. P. Wikman, and C. Wiuf. Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res*, 34(14):e100, 2006.

[11] S. Lin, B. Carvalho, D. J. Cutler, D. E. Arking, A. Chakravarti, and R. A. Irizarry. Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays. *Genome Biol*, 9(4):R63, 2008.

[12] W. M. Liu, X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, T. B. Ryder, T. A. Webster, S. Dong, G. Liu, K. W. Jones, G. C. Kennedy, and D. Kulp. Algorithms for large-scale genotyping microarrays. *Bioinformatics*, 19(18):2397–403, 2003.

[13] 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):6071–9, 2005.

[14] D. L. Nicolae, X. Wu, K. Miyake, and N. J. Cox. GEL: A novel genotype calling algorithm using empirical likelihood. *Bioinformatics*, 22(16):1942–7, 2006.

[15] J. Perkel. SNP genotyping: Six technologies that keyed a revolution. *Nat Methods*, 5(5):447–453, 2008.

[16] N. Rabbee and T. P. Speed. A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics*, 22(1):7–12, 2006.

[17] Y. Xiao, M. R. Segal, Y. H. Yang, and R. F. Yeh. A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics*, 23(12):1459–67, 2007.