

High-Throughput Microarray-Based Genotyping

Geoffrey Yang, Ming-Hsiu Ho, Earl Hubbell
Affymetrix, Inc.
Geoffrey_yang@affymetrix.com

Abstract

A high throughput genotyping platform that scores over 10,000 single-nucleotide polymorphisms (SNPs) per individual on a single GeneChip® high-density oligonucleotide microarray has been developed to conduct studies to elucidate the genetic basis for complex diseases.

Currently, the genotyping of individual SNPs relies on the summary statistics (based on the observed intensities of probes on the microarray) for the entirety of the sample set. A classification scheme of those statistics across hundreds of individual DNA samples is obtained to make individual genotyping calls.

In contrast to the current approach, the method in this work makes individual genotyping call by finding the minimum residual, i.e. highest likelihood, among four possible states corresponding to three genotypes and no-call. Initially, the residual is calculated based on a given set of probe affinities for that individual sample. Auspiciously, the multitude of samples was utilized to iterate to refine the probe affinities, sample concentration, and background intensities. These refined parameters entail an improved call rate while maintaining high accuracy.

1. Introduction

A robust high-throughput genotyping platform that offers the ability to generate over 10,000 genotypes on a single microarray [2] using an innovative assay that eliminates the need for locus-specific PCR has been developed [1, 3]. The current genotype-calling approach assigns genotypes based on allele-specific hybridization intensities [4] with > 99.5% accuracy. Being highly accurate and readily scalable, this platform helps to map disease genes and understand disease susceptibility efficiently.

The current genotyping-calling algorithm relies on a classification method for relative allele signals

(RAS) [4]. RAS from sense and anti-sense strands for each sample form a data point on 2-D feature space. Ideally, the data points for one SNP across many samples show two to three clearly defined clusters, depending on the minor allele frequency of that SNP. Quality score is used to determine how compact those clusters are. Each data point is assigned with a genotype depending on which one of those three call zones it falls into.

The new algorithm in this work makes genotyping call by finding the minimum residual among four possible states (AA, AB, BB, and No Call) and it does not rely on a training set. Instead, the multitude of samples enables us to iterate to refine the probe affinities, sample concentration, and background intensities by model fitting.

2. Method

2.1. Array design

The current GeneChip® Mapping arrays include either 5 or 7 probe interrogation positions for both sense and antisense strands, with each position containing 4 probes -- perfect match (PM) and mismatch (MM) for alleles A and B. Each interrogation position involves a different shift from the center of the 25-mer probe sequence, i.e. -4, -2, -1, 0, +1, +3, and +4. Thus, up to 56 features are used for each SNP.

2.2. Model parameters

Each feature is assigned with two affinity terms -- true hybridization and alternative-allele hybridization. Affinity for probe x to target y is denoted as a_y^x , i.e. $a_A^{PM_A}$ and sample concentration for target y is denoted as c_y , i.e. c_A . See Figure 1 for details.

Affinity terms for all features are initialized before the iterations start. Probe-position specific background intensity terms are initialized to be the smallest of the

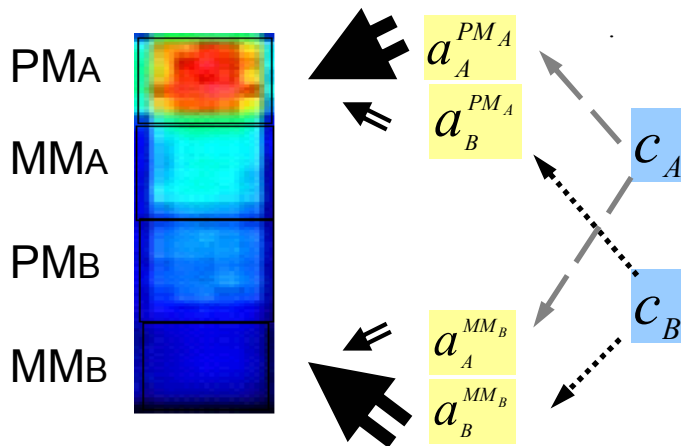


Figure1. Diagram showing notation of parameters for probe affinities and sample concentration, along with the relationship to a probe quartet in a probe interrogation position for a SNP.

four intensity values within that probe quartet.

2.3. Model fitting

Intensity (I) in each feature is modeled as the sum of background intensity (Bkg) and the linear product of probe affinity and sample concentration. Objective function is to minimize the sum of probe residuals across all probes and multiple samples, i.e., $R = \sum q \sum s \sum_{i=1 \text{ to } 4} (I - a_y^x C_y - Bkg)^2$ across samples (s) and probe positions (q). For each possible state, there is a residual value associated with its corresponding model parameters.

Concentration terms are calculated by solving the linear equation with the affinity and background terms fixed at that instant. Iterative processes are taken to update all the parameters, i.e., affinity terms are updated while keeping concentration and background terms fixed at that instant. At the end of each iteration, when all the parameters are updated once, probe residuals for all four states are calculated and summed. Genotype call for a SNP is assigned based on the state with the smallest summed residual, which is called the SNP residual. A “residual ratio” is defined for SNP residuals between consecutive iterations and it is compared against a convergence threshold. Analysis for a SNP will stop when convergence is reached or the number of iterations reaches the limit.

2.4. Statistical score for genotype call

Wilcoxon’s signed rank test [5] is applied at the last iteration to get a confidence measure of the final genotype call made. The details are as follows. Four matrices corresponding to each state are assembled. The contents of a matrix for state i ’s are the probe

residual differential, namely, the probe residual corresponding to the “state” minus the smallest of the remaining three probe residual. For example, with sample 1 on SNP1, $RDiff_{AA}(i) = R_{AA}(i) - \min(R_{AB}(i), R_{BB}(i), R_{NC}(i))$. Residual matrices for AB, BB, and NC are assembled in the same fashion. Wilcoxon’s test is then applied to these four matrices by comparing them to a zero matrix. The score from Wilcoxon’s test is then reported as the confidence measure of the genotype call made.

3. References

- [1] Dong, S., Wang, E., Hsie, L., Cao, Y., Chen, X. and Gingeras, T.R. Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res.*, (2001) 11, 1418–1424.
- [2] Fodor, S. P.; Read, J. L.; Pirrung, M. C.; Stryer, L.; Lu, A. T.; Solas, D. Light-directed, spatially addressable parallel chemical synthesis, *Science* (1991). 251(4995), 767-73.
- [3] Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.-m., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M.S., Boyce-Jacino, M.T., Fodor, S.P.A. and Jones, K.W. Large-Scale Genotyping of Complex DNA *Nature Biotechnology* (2003) 21, 1233 – 1237.
- [4] Liu, W-M., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T.B., Webster, T.A., Dong, S., Liu, G., Jones, K.W., Kennedy G.C., and Kulp, D. Algorithms for Large Scale Genotyping Microarrays. *Bioinformatics*, (2003) 19: 2397-2403.
- [5] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics*, (1945) 1, 80–83.