SoCal: supervised genotype calling via ellipsoidal separation for Affymetrix SNP microarray

Huwenbo Shi (603-778-363) shihuwenbo@ucla.edu

1 Introduction

SNP microarray is a cost–effective approach to genotype samples for specific association studies. In Affymetrix SNP microarrays, oligonucleotide probes are first used to bind DNA fragments containing SNPs. Then, for each SNP, a fluorescence scanner quantifies perfect match (PM) and mismatch (MM) for each of the two alleles, denoted by A and B, on each strand of the DNA fragment. The genotype calling procedure for SNP microarray consists of two steps. In the first step, information from microarray is summarized to obtain the intensities, θ_A and θ_B, of the two alleles of each SNP. In the second step, SNPs are classified into genotype AA, AB, or BB based on the allele intensities they generate. The focus of this article is on the second step of the genotype calling procedure—genotype classification using summarized allele intensities.

For a specific SNP, if a sample has genotype AA or BB, the intensity, θ_A or θ_B, will be higher respectively. If a sample has genotype AB, the intensities, θ_A and θ_B, will be similar. If one plots $log(\theta_A)$ versus $log(\theta_B)$ of a SNP for a number of samples, normally 3 ellipsoidal clusters are observed, one for each genotype, as shown in Figure 1. Many genotype calling algorithms use model–based unsupervised clustering methods to identify clusters and then assign genotypes to each cluster. To estimate model parameters, these methods use the EM algorithm, which is sensitive to starting parameters and slow to converge. Rabbee et al. proposed the RLMM algorithm, a supervised genotype calling method that uses reference genotype calls to form Gaussian decision boundaries for each genotype. This method involves fitting a linear mixed model, which can be computationally intensive.

As the number of probes on SNP microarrays and the number of individuals involved in association studies continue to increase, both fast and accurate genotype calling algorithms are needed.

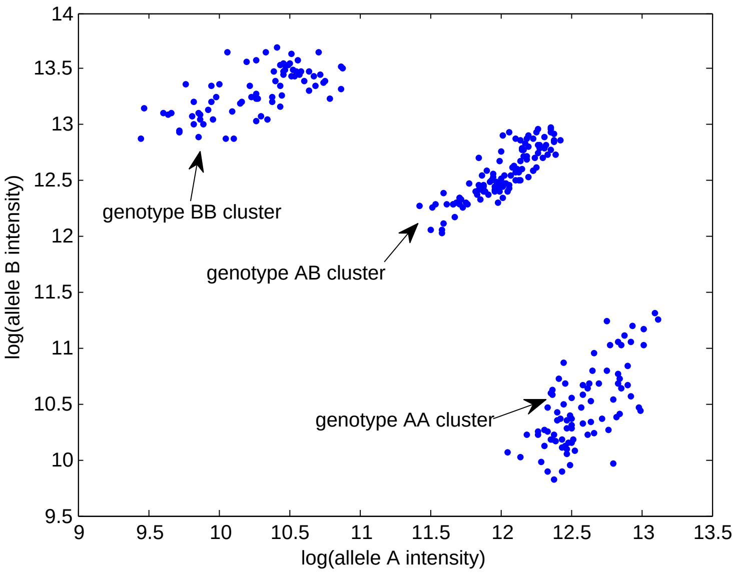

Figure 1: Genotype clusters obtained from Affymetrix SNP array allele intensity values

2 Method

SNP allele intensities are first summarized from raw microarray data using SNPRMA. After this step, SoCal calls genotypes in two steps. In the first step, SoCal finds ellipsoidal regions containing each of the genotype of a SNP using reference genotype calls. In the second step, SoCal classifies samples with unknown genotypes using minimum distance classification. The organization of this section is as follows. First, I introduce the problem of pattern separation by ellipsoids and show how SoCal forms ellipsoidal regions for each genotype of a SNP. Then, I show how SoCal uses these ellipsoids to call genotypes.

2.1 Pattern separation by ellipsoids

An ellipsoid $\mathcal{E} \subseteq \mathbb{R}^n$ can be expressed as $\mathcal{E} = \{x \in \mathbb{R}^n | (x-c)^T E(x-c) \leqslant 1\}$, where c is the center of the ellipsoid, and E a positive semidefinite matrix. Let a_i be the points to be included in an ellipsoidal region, and b_j be the points to be excluded, the problem is to find a c and E that separate a_i from b_j.

2.2 Forming decision regions for each genotype

Instead, COMB provides a flexible way to control outlier effect using an approach that combines the previous two methods with an additional l_1-norm regularization. The method of COMB can be expressed as a conic programming problem:

$$\begin{aligned}
\text{minimize } & -\beta_1 k + \beta_2 trace(T) + \beta_3 \|u - 1\|_1 \\
\text{subject to } & (1, a_i)^T \tilde{E} (1, a_i) \leqslant u_i \ \forall i \\
& (1, b_j)^T \tilde{E} (1, b_j) \geqslant k \ \forall j \\
& \tilde{E} = \left[\begin{array}{cc} s & v^T \\ v & F \end{array} \right] \succeq 0 \\
& \left[\begin{array}{cc} F & I \\ I & T \end{array} \right] \succeq 0
\end{aligned}$$

where $\beta_i > 0$ are the weights assigned to each sub-objectives: maximizing separation ratio, minimizing ellipsoid volume, and controlling outliers. The final ellipsoidal decision region $\mathcal{E}^*$ is defined as $\mathcal{E}^* = \{x \in \mathbb{R}^n | (x - c^*)^T E^* (x - c^*) \leqslant 2(1+k)\}$, where $c^* = -F^{-1}v$, $E^* = F/(1 - s + {c^*}^T F c^*)$.

2.3 Missing clusters

Talk about how SoCal handles missing clusters

2.4 Genotype calling and outlier detection

Talk about how classification and outlier detection is done