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.

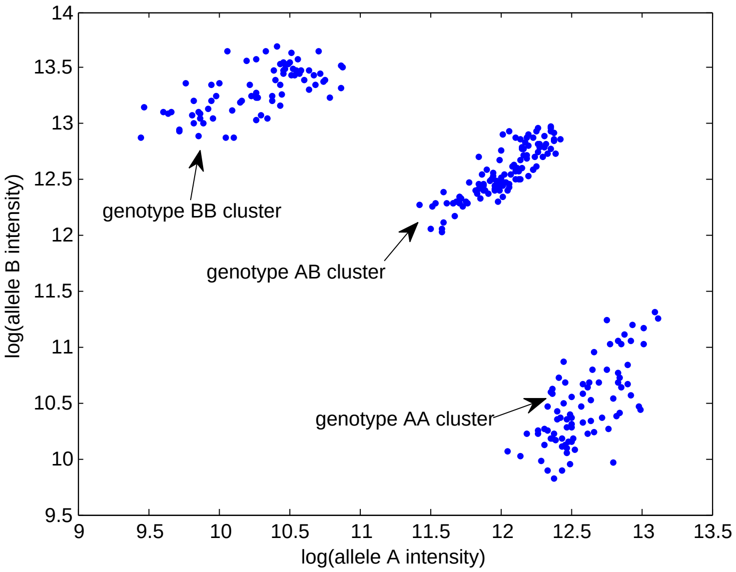

Figure 1: Genotype clusters obtained from Affymetrix SNP array allele intensity values

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 ellipsoid and show how SoCal forms ellipsoidal regions for each genotype of a SNP. Then, I show how SoCal uses these ellipsoidal regions to call genotypes.

2.1 Pattern separation by ellipsoid

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 definite matrix denoting the orientation of the ellipsoid. Let a_i be the points to be included in an ellipsoid, and b_j be the points to be excluded, the problem of ellipsoidal separation is to find c and E such that $(a_i - c)^T E(a_i - c) \leqslant 1 \, \forall i$ and $(b_j - c)^T E(b_j - c) > 1 \, \forall j$.

2.2 Forming ellipsoidal decision regions for each genotype

Let $G = \{AA, AB, BB\}$ be the set of genotypes of a SNP, and J_{AA}, J_{AA}, J_{BB} the index set of samples with the corresponding genotype. Let $X = \{(log(\theta_A), log(\theta_B))_i | i = 1, \cdots, |J_{AA}| + |J_{AB}| + |J_{BB}|\}$ be the set of log transformed allele intensities of all the samples, and $X_{AA} = \{x_j | x_j \in X, j \in J_{AA}\}$, $X_{AB} = \{x_j | x_j \in X, j \in J_{AB}\}$, $X_{BB} = \{x_j | x_j \in X, j \in J_{BB}\}$ the set of log transformed allele intensities of the corresponding genotype.

To find the ellipsoid that includes X_{AA} and excludes X_{AB} and X_{BB}, one sets $\{a_i\} = X_{AA}$ and $\{b_j\} = X_{AB} \cup X_{BB}$, and solves the following conic programming problem. For the sake of space, detailed derivation of the problem formulation is not presented here.

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

In the problem formulation above $\beta_i > 0$ are the weights assigned to each sub-objectives of finding the ellipsoid—maximizing separation ratio, minimizing ellipsoid volume, and controlling outliers.

Let $\tilde{E}^* = \left[\begin{array}{cc} s & v^T \\ v & F \end{array} \right]$ be the optimal solution to the problem above. The separating ellipsoid $\mathcal{E}^*$ is defined as $\mathcal{E}^* = \{x \in \mathbb{R}^n | (x - c^*)^T E^* (x - c^*) \leqslant \beta_4 (1 + k)\}$, where $c^* = -F^{-1} v$, $E^* = \frac{F}{(1 - s + c^{*T} F c^*)}$. Here, β_4 is a positive constant controlling the size of the ellipsoid. In SoCal, β_1,β_2,β_3,β_4 are empirically set to 1,10, 100, and 30 respectively.

Similary, to find the ellipsoid that includes X_{AB} and excludes X_{AA} and X_{BB}, one sets $\{a_i\} = X_{AB}$ and $\{b_j\} = X_{AA} \cup X_{BB}$, and solves the above conic programming problem. The same procedure also applies to finding the ellipsoid that includes X_{BB} and excludes X_{AA} and X_{AB}.

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