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.

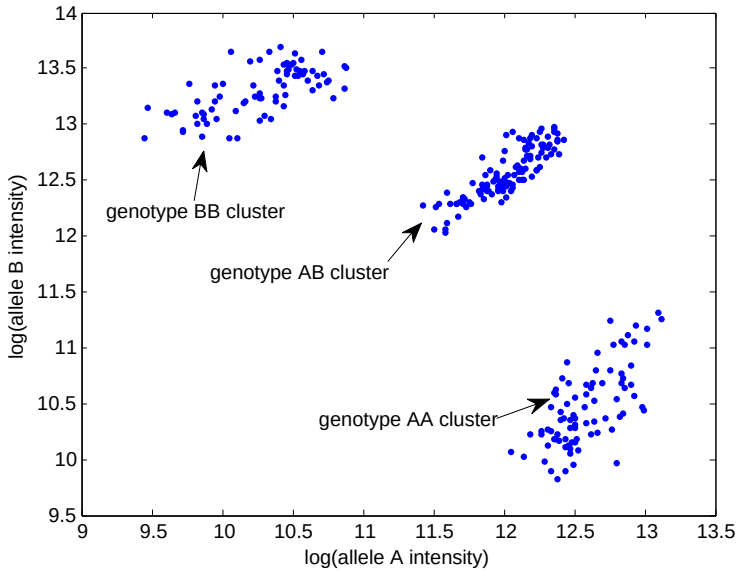

Figure 1: Genotype clusters obtained from Affymetrix SNP array allele intensity values