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

Supervised Genotype Calling for Affymetrix SNP Arrays

Introduction

Single nucleotide polymorphisms (SNPs), genomic positions at which a single nucleotide differs between individuals, are responsible for many phenotypic variations among individuals. To study the associations between SNPs and phenotypes, one must be able to accurately call genotypes of case and control samples at these SNPs. Although next generation sequencing (NGS) technology provides cheap and whole-genome sequences for genotyping SNPs, SNP arrays are still more cost-effective for specific experiments (Rho et al., 2010).

In Affymetrix SNP arrays, oligonucleotide probes are used to measure intensities I_A and I_B for alleles A and B of SNPs in samples. If a sample has genotype AA or BB for a SNP, one observes higher value of I_A or I_B respectively. For genotype AB, one observes similar values of I_A and I_B. If one plots (I_A, I_B) of a SNP for a number of samples, normally 3 clusters can be observed, one for each genotype (Figure 1). The computational challenge is to correctly assign each (I_A, I_B) data point to a cluster and call its genotype accordingly.

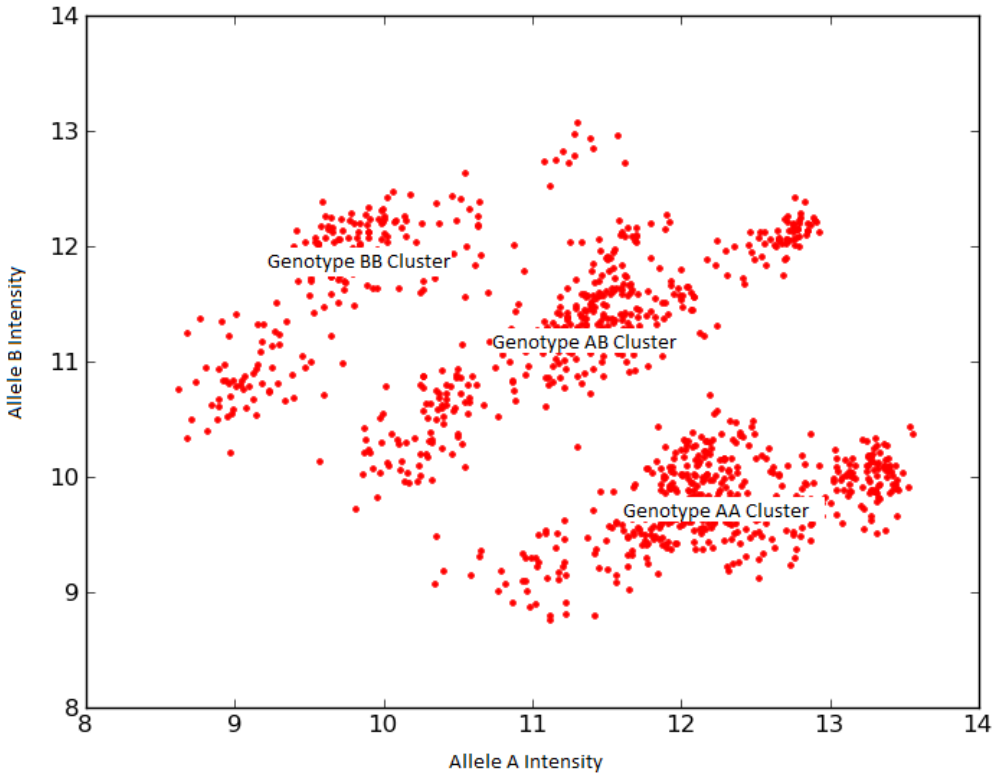

Figure 1

Motivation

Many genotype calling algorithms use model-based unsupervised clustering methods to identify clusters and then assign genotypes to each cluster. For instance, Norlén et al. (2008) proposed a Gaussian mixture model for clustering; Lin et al. (2008) proposed a statistical model and a modified K-means algorithm to incorporate pedigree information in clustering. These methods use EM algorithm to estimate model parameters, which is sensitive to starting parameters and slow to converge (Wu, 1983).

A few other genotype calling algorithms exploit reference genotype calls. For instance, Rabbee et al. (2005), proposed the RLMM algorithm, which estimates allele intensities of SNPs on a chip by fitting a linear mixed model and compares them with the means obtained from training data to call genotypes. This method models probe effect in the linear mixed model and is able to make more accurate genotype calls. However, fitting a linear mixed model can be computationally intensive, and may involve optimizing a non-convex function (Zhou et al., 2014).

As the number of probes on a SNP array and the number of individuals involved in association studies continue to increase, both fast and accurate genotype calling algorithms need to be developed.

Proposal

If reference genotype calls and allele intensities of each SNP for a number of samples are available, for each SNP one can find boundaries that separate different genotypes from each other. One can then use the boundaries to call the same SNP in different samples with unknown genotypes.

Various types of separator for pattern separation exist. If one assumes that allele intensities generated by each of the three genotypes of a SNP follow bivariate Gaussian distribution, then the most suitable type of separator is ellipsoid. Once the ellipsoids that best separate the 3 genotypes of a SNP are found, one can use them as decision regions to call genotypes.

The advantage of this approach is that the problem of finding separating ellipsoids can be formulated as a convex programming problem and can be solved efficiently (Vandenberghe et al., 1999; Glineur, 1998). Once the ellipsoids are found, future genotype calls can be done in linear time. Also, outlier effect can be controlled using different criteria for constructing the ellipsoids.

References

Glineur F. (1998). Pattern separation via ellipsoids and conic programming. (MS Thesis). Faculté Polytechnique de Mons, Mons, Belgium.

Lin, Y., Tseng, G. C., Cheong, S. Y., Bean, L. J., Sherman, S. L., & Feingold, E. (2008). Smarter clustering methods for SNP genotype calling. Bioinformatics, 24, 2665-2671.

Norlén, H., Pettersson, E., Ahmadian, A., Lundeberg, J., & Sundberg, R. (2008). Classification of SNP genotypes by a Gaussian mixture model in competitive enzymatic assays. Mathematical Statistics Stockholm University Research Report, 3, 1-26.

Rabbee, N., & Speed, T. P. (2005). A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics, 22, 7-12.

Rho, S. W., Abell, G. C., Kim, K., Nam, Y., & Bae, J. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends in Biotechnology, 28, 291-299.

Vandenberghe, L., & Boyd, S. (1999) Applications of semidefinite programming. Applied Numerical Mathematics, 29, 283-299.

Wu, C. F. (1983). On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11, 95-103.

Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods, 11, 407-409.