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

Supervised Genotype Calling for Affymetrix SNP Arrays

Introduction

SNPs 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 SNPs for case and control samples. Although next generation sequencing technology provides cheap and whole-genome sequences for genotyping SNPs, SNP arrays are still more cost-effective for specific studies (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. The computational challenge is to correctly assign each (I_A, I_B) data point to a cluster and call its genotype accordingly.

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.

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 then compares them with the means obtained from training data to call genotypes.

These methods can call genotypes very accurately, but are parameter sensitive and computationally intensive. As the number of probes on SNP arrays 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 I_A and I_B of a SNP for a number of samples are available, one can find boundaries that separate different genotypes from each other as a function of I_A and I_B. 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. For SNP arrays, which generate allele intensities following bivariate Gaussian distributions, the most suitable type of separator is ellipse.

Once ellipses that 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 ellipses can be solved efficiently. And once the ellipses are found, future genotype calls can be done in linear time. Also, outlier effect can be controlled by specifying different criteria for finding the ellipses, a feature not provided by simply fitting a Gaussian distribution.

References

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.