The Annals of Applied Statistics
2012, Vol. 6, No. 2, 429–431
DOI: 10.1214/12-AOAS558

SPECIAL SECTION: STATISTICAL METHODS FOR NEXT-GENERATION GENE SEQUENCING DATA

BY KAREN KAFADAR

Indiana University

This issue includes six articles that develop and apply statistical methods for the analysis of gene sequencing data of different types. The methods are tailored to the different data types and, in each case, lead to biological insights not readily identified without the use of statistical methods. A common feature in all articles is the development of methods for analyzing simultaneously data of different types (e.g., genotype, phenotype, pedigree, etc.); that is, using data of one type to inform the analysis of data from another type.

In the first article of this section, Li et al. address the problem of multiple missing genotype data through a Bayesian hierarchical approach. The goal is to impute missing values in association studies between genotypes (as measured by single nucleotide polymorphisms, or SNPs, in DNA sequences) and phenotypes. Because missing SNP information is common, case-wise deletion is, at best, impractical, and often wasteful of valuable information when SNP information is available for the rest of the case. Li et al. develop a computationally-efficient approach to multiple imputation of many missing SNPs that uses all available phenotype information. They show that their Bayesian Association with Missing Data (BAMD) approach achieves the desired goal in that it enables efficient detection of SNPs that are highly associated with phenotypes.

Zhou and Whittemore propose likelihood-based methods to improve the accuracy of genotype calls using information on linkage disequilibrium (LD) and Mendelian pedigree information, particularly for multiple SNPs that exhibit high LD (SNPs for which the squared correlation coefficient between them is close to 1). Thus, use of LD or pedigree information can modulate the negative effects of errors in sequence reads and alignments and hence enable better inference. The approach is applied to data from both simulations and the "1000 Genomes Project" and is shown to improve the estimates of model parameters and hence the accuracy of genotype calling.

Shen and Zhang develop a change-point model based on a nonhomogeneous Poisson process (NHPP) to model sequence-read data on DNA copy number

Received April 2012.
Key words and phrases. High-throughput sequencing data, differential gene expression, phenotype, missing data, copy number variant, single nucleotide polymorphism, Bayesian hierarchical model, Bayesian model averaging.