frontiers in
GENETICS

SPECIALTY GRAND CHALLENGE ARTICLE
published: 24 February 2011
doi: 10.3389/fgene.2011.00005

Grand challenges in statistical genetics/genomics methodology

Hemant K. Tiwari*[1]* and *Nicholas J. Schork*[2]

[1] *Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA*
[2] *The Scripps Translational Science Institute, Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA, USA*
**Correspondence: htiwari@uab.edu; nschork@scripps.edu*

Recent developments in genomic technologies have provided researchers with an unprecedented ability to probe the genetic basis of complex biological processes as well as phenotypic expression. However, as impressive as these technologies are, the analysis and interpretation of the data they generate is exceptionally challenging due to the amount and sophistication of these data. In fact, it has been said of next-generation genomic technologies that they often result in situations in which researchers are "drowning in data, but thirsting for knowledge." *Frontiers in Statistical Genetics and Methodology (FSGM)* will provide a forum for theoretically minded statistical geneticists to publish methodology in an effort to help quench this thirst. In this context, as the senior editors of *FSGM* we feel that there are a number of important areas in modern statistical genomics that have immediate needs for methodological developments. We outline some of these below, but by no means want to convey the sense that *FSGM* will only publish papers dealing with these areas, as the field of statistical genomics is simply too broad and filled with too many challenges and needs to restrict attention of *FSGM* to a few thematic areas.

METHODOLOGIES FOR NEW DATA TYPES

Newer assays, especially DNA sequencing-based assays such as variant identification sequencing, RNA sequencing ("RNA-seq") for digital gene expression assays, chromatin immunoprecipitation sequencing ("ChIP-seq") for DNA–protein interaction analysis, antibody sequencing for immunological diversity analyses, methylation, and bisulfate sequencing ("Methyl-seq") for epigenomic assays, and many other assay types exploring the transcriptome, proteome, metabolome, etc., all require very diligent and careful analysis of the raw data that they produce. This more "upstream" analysis of modern genomic assay results is crucial if the more "downstream" analysis of the data (e.g., drawing inferences from the data) is to not suffer from the proverbial "garbage in, garbage out" principle. Better statistical analysis methods that take into account different sources of error during the processing stages of genomic assay data are sorely needed and will continue to be a needed as technologies are improved, extended, and replaced with more sophisticated ones.

APPROACHES TO INTEGRATION OF DIFFERENT DATA TYPES

Following the singular development, implementation, and application of specific genomic assays, is the combined use of those assays to address specific questions. For example, one may leverage DNA sequencing and transcriptomic or proteomic assays to identify "expression quantitative trait loci (eQTLs)" or "protein QTLs (pQTLs)." Drawing appropriate inferences from the combined set of assays is not trivial, as it requires not only a familiarity with models and methods for handling sources of error associated with each assay, but also an ability to model the relevant system as a "whole" above its assayable "parts." Such modeling can pose very tricky problems for the statistical geneticists.

MODELING POPULATION-LEVEL PHENOMENA

Population genetics is a ubiquitous biomedical science and not confined to the, e.g., ecological or genetic epidemiological sciences. For example, understanding how populations of cells harboring mutations contribute to tumorigenesis, how different antibody "species" defined by DNA sequence diversity contribute to immune responses, and how proteins harboring different amino acid sequences may influence fundamental molecular physiological interactions, all require a population perspective. Thus, statistical methodologies for assembling and studying networks, the flow, and transmission of information from, e.g., generation to generation, the hierarchical functioning of gene regulatory circuits, and related phenomena will become increasingly important.

PREDICTIVE MODELING

The availability of genetic assays results that pertain to phenotypic categories [e.g., genome wide association study (GWAS) findings with respect to clinical diagnoses; gene expression patterns that differentiate more or less aggressive tumors; selection studies to produce more fruitful livestock; proteomics or metabolite profiles indicative of drug response in cell line or clinical studies; etc.] create a need for the development of predictive models and classifiers for the phenotypic categories. Such models can be used by applications-oriented researchers in a wide variety of settings (e.g., selecting livestock for breeding; facilitating drug discovery; clinical biomarker use; etc.) but are complicated by a number of statistical issues, such as having more predictors/variables than units of observations, the need to accommodate potential confounding by covariates, dealing with mixed longitudinal data, etc.

METHODOLOGIES FOR CROSS-SPECIES DATA

Genomic analyses are, in a broad sense, rooted in evolutionary concepts and theory, given the relationships among species at the level of DNA. Explicitly exploiting this fact in research paradigms is an important area for statistical genetics concentration. For example, developing better analysis methods for measuring and leveraging conservation of DNA sequence in cross-species studies, developing statistical methods for leveraging pathway and genetic network data from one species in the study of another, and assessing homologies at levels beyond, but maybe dictated to some degree by, DNA sequence similarities (e.g., physiologic processes, phenotype profiles, etc.) are fast becoming target areas for researchers in the sequencing era.

DESIGN OF STUDIES PERTINENT TO GENETICS

The availability of cost–effective genomic technologies can often lull scientists into a "collect data first and ask questions later" mentality that might be detrimental to good science. Good study designs should never be eschewed for more data-generation horsepower. In fact, good study designs should go hand-in-hand with the use of more powerful technologies and assays. Thus, the need for the development of better field designs in agricultural genomics, the incorporation of contrasting strains of model organisms to assess genetic background effects in functional genomic studies, the design of efficient gene-based clinical trials in advancing "personalized medicine" and related activities should be emphasized by the community.

As noted, the foregoing problem areas are but a handful of the problem areas statistical geneticists currently face and will face in the future. We see these problems as both a call-to-action and a stimulating, yet seemingly daunting, intellectual challenge to the community and believe that *FSGM* will provide a premier vehicle for meeting this challenge.

Received: 04 January 2011; accepted: 10 February 2011; published online: 24 February 2011.

Citation: Tiwari HK and Schork NJ (2011) Grand challenges in statistical genetics/genomics methodology. Front. Gene. **2***:5. doi: 10.3389/fgene.2011.00005*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.