

Statistical Genetics

Ina Hoeschele
VBI & Statistics

Statistical Genetics

- A field that resulted from the **merger of genetics and statistics** into a **quantitative/mathematical theory for the interpretation of genetic data**
- **It encompasses**
 - The search for and characterization of genes affecting human health and traits of interest in plants and animals
 - The study of the evolution of genes in natural populations
 - The study of the evolution of genomes and species
 - The analysis of DNA, RNA and protein sequence and structure (structural genomics)
 - The analysis of transcriptome, proteome and metabolome profiling data (functional genomics)
- **Modern Statistical Genetics** = the statistical contributions to genomics, transcriptomics, proteomics, and metabolomics
- **Handbook of Statistical Genetics** (2 volumes, Wiley, 2003)

Statistical Genetics & Genomics

Gene1 - DNA $\xrightarrow{\text{transcription}}$ mRNA1 $\xrightarrow{\text{translation}}$ protein1

Gene2 - DNA $\xrightarrow{\text{transcription}}$ mRNA2 $\xrightarrow{\text{translation}}$ protein2

MetaboliteA $\xrightarrow{\text{enzyme1}}$ MetaboliteB $\xrightarrow{\text{enzyme2}}$ MetaboliteC

Training Program and Research Group

- Research group:
 - Mix of graduate students (mostly PhD) and postdoctoral fellows
 - RAs and positions funded by NIH, NSF, companies
 - Fluctuates from 3 to 7 members
- Training Program:
 1. Bioinformatics track in Statistics PhD program (all required Stat PhD courses plus some biology courses)
 2. Statistics track in GBCB PhD program (Genetics, Bioinformatics and Computational Biology) (all or most required Stat PhD courses plus (required) biology and computer science courses)

Training Program and Research Group

- Training Program:
 - Background: BS / MS in Statistics, Genetics, Bioinformatics, Computer Science, Biology etc. (quantitative interests and experience)
 - Employment:
 - 12 graduate and 19 postdoctoral students
 - Faculty positions at universities and research institutes (Biostatistics, Bioinformatics, Human Genetics, Animal Science)
 - Companies: Pharmaceutical, Biotechnology, Statistical Software (Monsanto, GlaxoSmithKline, Pfizer, Insightful Corporation)

Courses

- Stat 5554 Variance components (Linear and nonlinear mixed models)
- Stat 5564 Statistical Genetics (mostly gene mapping)
- New (summer 2006): Microarray data analysis (initial focus on expression data)

Statistical Genetics Research

- **A major goal:** to identify **gene regulatory networks** underlying diseases, and to identify the best target (gene) for therapeutic **intervention**
- Various computational methods proposed (e.g., Bayesian networks, probabilistic Boolean networks, ...)
- **WE: causal modeling** for specific intervention experiments, linear (Structural Equation Models) and non-parametric
- Some related **statistical issues:**
 - High-dimensional data – dimension reduction (PCA, genetic PCA, PLS)
 - Multiple testing procedures (control of error rates based on proportion of false positives; dependency structures, classical and Bayesian)
 - Methods combining various sources of data (or various methods)
 - weighting schemes, Bayesian methods ...
 - Query, display and visualization of results