

## **Genetic profiling with imaging data: concepts and methods**

Speaker: Nicholas J. Schork, Ph.D.

Polymorphism Research Laboratory, Department of Psychiatry, UCSD

There are differences between research questions related to genetics and those related to genomics. Genetics research focuses on inheritance and the algebraic laws (such as Mendel's laws) that govern the transmission of chromosomes (and therefore genes) from parents to offspring. In contrast, genomics focuses on the structure, molecular impact (e.g., expression patterns), and evolution (e.g., conservation and homology) of genes. The statistical modeling issues of genetics research are therefore in some way distinct from those that arise in genomics. Ideally, however, future biological and medical research will seek to integrate the research field of genetics and genomics. Currently, most of the inference and statistical questions in genetics and genomics research relates to how one can deal with massive amounts of data. The Human Genome Project, the International HapMap Project, the coordinated dissemination of genetic and genomic information via web-accessible databases, and related initiatives, have provided genetics researchers access to enormous quantities of information. In addition, these initiatives have motivated recent advances in high-throughput, data-intensive molecular genetic and phenotyping technologies for DNA sequencing, genotyping, microarray-based gene expression analysis, imaging, multiplex clinical characterization, etc. Consequently, Dr. Schork wants to emphasize the need for the integration of the massive amount of data generated from studies in both genetic and genomics. To achieve this, his idea is that most genetic and genomic data should be analyzed as a "whole" or as "profiles" of the individuals assayed and can hence be treated with sophisticated multivariate analysis techniques that focus on the similarity and dissimilarity of these wholes or profiles. According to Dr. Schork, many multivariate analysis techniques, such as traditional cluster analysis, have been developed and applied to high-dimensional genetic data. However, many are problematic for various reasons. Non-cluster-based analysis methods that exploit similarity or "distance" between individual units of observation with respect to the massive amounts of data collected on them (profiles) are particularly well suited for such data, since they can be used to evaluate appropriate hypotheses without necessarily examining each data point in isolation and they can be rooted in fundamental statistical genetic concepts. As per Dr. Schork, similarity-based methods are particularly useful to visualize the influence of genes in brain function. For example, correlation matrices can generate a graphical representation of data in a two-dimensional map where the variable values are represented by colors ("heat maps"). Another example are Tree representations, which can graph deviations from the centroid vs. interpoint distances ("Neighbor-joining tree") and have been used successfully in fMRI studies on the effect of COMT polymorphisms in the cognition of schizophrenic subjects, among others.