THE UNIVERSITY OF CHICAGO

Department of Statistics

STATISTICS COLLOQUIUM

---


KATHRYN ROEDER

Department of Statistics
Carnegie Mellon University


# A Method to Exploit the Structure of Genetic Ancestry Space to Enhance Case-Control Studies

MONDAY, April 20, 2015, at 4:00 PM
Eckhart 133, 5734 S. University Avenue
*Refreshments following the seminar in Eckhart 110.*

## ABSTRACT


In genetic studies of common and rare variants, considerable effort and expense is required to obtain a sample of control subjects matched by genetic ancestry to the case subjects. Alternatively, repositories like dbGaP already contain genetic data from tens of thousands of potential control samples. These data can be accessed, but only with considerable effort on the part of the research team. It should be possible to model these data and obtain allele frequency estimates, which would obviate the need for collecting additional large control samples. The task is challenging for two reasons: due to issues of privacy, genotype data can not be shared directly; and yet the control data must be chosen so that it is comparable in genetic ancestry to the particular case sample. Our proposed approach, the Universal Control Repository Network (UNICORN), aims to provide allele frequency information that is optimally matched to the case sample. To maintain the confidentiality of both cases and controls, no case genotype information is passed to UNICORN, nor will the controls available in the repository ever be accessible to external researchers. Instead we will use existing publicly available collections of control data to create a common genetic ancestry space onto which cases and control can be mapped independently. We use spectral clustering to construct ancestry spaces as well as to perform projections. The base space and projected controls are then used to estimate the allele frequency surface over the ancestry space. To identify small-scale frequency variation while also borrowing strength from the


---

entire data set we employ a combination of empirical Bayesian analysis across a hierarchical clustering of the controls and, for localized ancestry regions, a Gaussian process model of the minor allele frequency. We have examined the performance of UNICORN on two large, complex studies (Crohn's disease and autism) and obtained promising results. We believe that our proposed model will be of significant importance to researchers by enabling more powerful association studies with fewer resource expenditures. This is joint work with my PhD student, Corneliu Bodea.