**The University of Chicago**
**Department of Statistics**

**Seminars for Fourth Year Ph.D. Students**

---


## XIAOQUAN (WILLIAM) WEN

Department of Statistics
The University of Chicago


# Estimation of Allele Frequencies at Untyped Genetic Markers from Summary (or Pooled) Genotype Data

## THURSDAY, February 12, 2009 at 2:00 PM
## 110 Eckhart Hall, 5734 S. University Avenue

## ABSTRACT


Estimating missing genotype data in genetic studies has many important practical applications. Most recently it has been applied as an elegant approach to combining results from multiple association studies performed on different genotyping platforms. In brief, most current methods can be thought of as solving a supervised learning problem. First, they take a set of training data (the panel), consisting of $n$ individuals genotyped at a dense marker set $P$, and learn about patterns of correlations among these markers. Next they take a set of $m$ study samples genotyped at smaller marker set $O$ (a subset of $P$). They then use the observed data at markers $O$ to estimate ("impute") the values for the untyped genetic markers in the difference set $U = P - O$ for the study samples.

In this talk we consider extending these methods to deal with the situation where only summary-level data are available in the study sample. This situation is common in practice: for reasons of privacy or politics it is common for researchers to share summary level data with one another rather than individual-level data. Further, some types of experimental protocol (DNA pooling) yield only summary-level data, and not individual genotype data. Our approach is based on using multivariate normal distributions, with a sparse banded covariance matrix, to model the summary data. This leads to very simple linear-predictor rules for the estimation of the missing genotype frequencies. Results from both real data and simulation study show that this yields accurate imputation result, surprisingly similar to the accuracy obtainable from the individual genotype data. In the context of data from pooling experiments our approach can also reduce the error in estimated genotype frequencies at typed markers. We discuss the potential to extend this approach to perform association tests for untyped markers and meta-analyses of genome-wide association studies.


---

Information about building access for persons with disabilities may be obtained in advance by calling Kelly Macias at 773.834.5169 or by email (kmacias@galton.uchicago.edu).