Yuying Xie

Assistant Professor, CMSE; Department of Statistics
xyy@msu.edu | 517.353.7154 | 619 Red Cedar Rd., Room C413

BIO SKETCH

Yuying Xie recently joined the Michigan State University faculty with joint appointments in the Department of Computational Mathematics, Science, and Engineering, and the Department of Statistics. His research focuses mainly on the area of statistical machine learning and consists of two themes: the development of new statistical procedure for high dimensional network models using complex and noisy datasets, and the inference of association or causal relations among genes, external phenotypes, and intermediate phenotypes, such as gene expression levels and biomarker concentrations.

Dr. Xie obtained his B.S. in biology from Fudan University, China. He then attended the University of North Carolina at Chapel Hill, where he received his first Ph.D. in genetics under Professor David Threadgill (2010), and his second Ph.D. in statistics under Professors Yufeng Liu and William Valdar (2015). During his doctoral studies in statistics, he was also a graduate fellow at the Statistical and Applied Mathematical Sciences Institute.

RESEARCH INTERESTS

STATISTICAL GENETICS: genetic linkage and association study, QTL/eQTL Mapping, RNA-Seq and DNA methylation data analysis; STATISTICAL MACHINE LEARNING: graphical model, causal inference, and measurement error

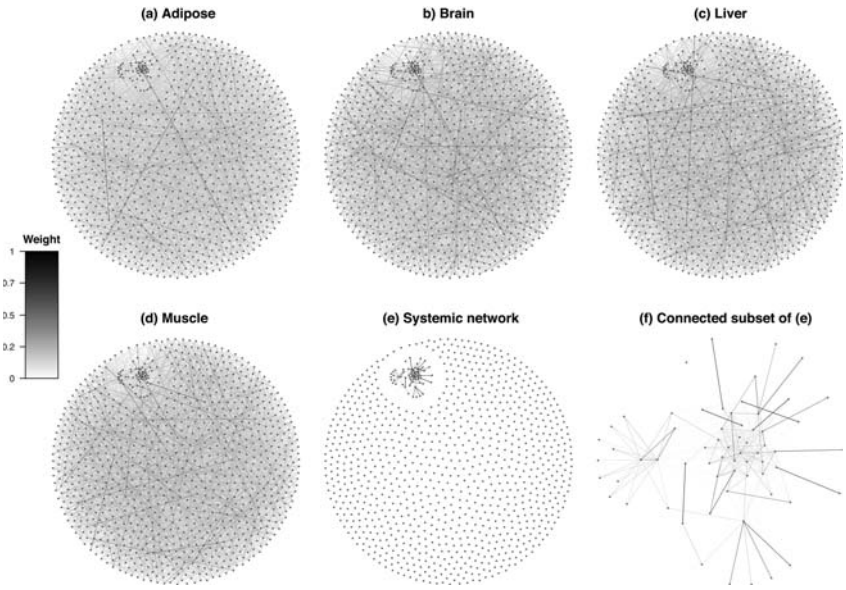

FIGURE 1. Topology of co-expression networks inferred by the EM method applied to measurements of the 1,000 genes with highest within-tissue variance in a population of F2 mice. Panels (a), (b), (c) and (d) display the category-specific networks estimated for adipose, hypothalamus, liver and muscle tissues respectively. Panel (e) shows the structure of the estimated systemic network, describing across-tissue dependencies, with panel (f) showing a zoomed-in view of the connected subset of nodes in this graph.

WEBSITE

http://www.stt.msu.edu/~xyy/

CURRENT RESEARCH FOCUS

The advances in high throughput biomolecular technology in recent years have resulted in a data explosion in biology. This enrichment of data offers the promise of solutions to many biological and medical goals, including detection of genes underlying complex diseases such as diabetes and hypertension. But, this promise will remain unfulfilled without the development of new statistical methodologies that can efficiently explore and make sense of the information within large, high-dimensional, and noisy data sets. Our research mainly focuses on the area of statistical machine learning and consists of two themes: (1) the development of new methodologies to estimate conditional dependence structure from noisy or dependent data sets, and (2) the inference of association or causal relations among genes, external phenotypes and intermediate phenotypes, such as gene expression levels and biomarker concentrations.

1. Joint estimation of multiple dependent Gaussian graphical models. Gaussian graphical models (GGM) are widely used to represent conditional dependence among sets of random variables. A common practical extension of GGM is the simultaneous estimation of multiple graphs that may share some common structure. An important assumption behind this extension, however, is that data collected from different categories are stochastically independent. While this assumption may hold, in other situations it is untrue; for example, gene expression across multiple tissues in the same individual is usually correlated. To deal with such data sets, we developed an elaboration that models such dependency by decomposing the problem into two graphical layers: the systemic layer, which is the network affecting all outcomes and thereby inducing cross-graph dependency, and the category-specific layer, which represents the graph-specific variation. We also developed a new graphical EM technique that estimates these two layers jointly. Furthermore, we also establish the estimation consistency and selection sparsistency of the proposed estimator, and apply the graphical EM technique to mouse genomic data to obtain some biologically plausible results as shown in Figure 1. At present, we are developing novel methodology to estimate high dimensional Directed acyclic graph from dependent data using gene expression data and other Omics datasets.

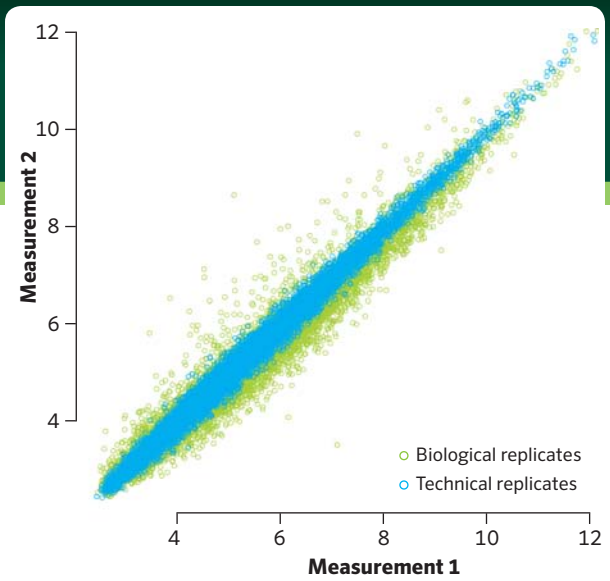

FIGURE 2. Scatter plot for effect of measurement errors. Each point represents a gene, *x* and *y* axes are gene expression levels measured by microarray.

2. Estimation of Gaussian graphical model from noisy data. A notable drawback of the existing methods for estimating GGMs is that they ignore the existence of measurement error. Measurement error is both common and varied in biological data. An example of such error is illustrated in Figure 2 depicting microarray data. Each blue dot represents two measurements of the expression level of a single gene, in the same individual; i.e., a pair of technical replicates. Each green dot represents two measurements of the expression level of a single gene, but in different individuals; i.e., a pair of biological replicates. As shown in Figure 2, a large proportion of the total variation among patients is from measurement error. Moreover, it has been shown that RNA-seq could only reliably quantify expression of only 30% of the genes with a relative error less than 20% of the total variance. Many factors could introduce variation to microarray and RNA-Seq results, including RNA degradation rates during sample collection, amplification efficiency in the PCR step, and molecular constitution and secondary structure of the RNA among genes. With the existence of measurement error, the graph for the outcome variables becomes more connected, and has weaker signals than the true underlying biological network. To address this issue, we proposed a new experimental design using technical replicates, and developed a new methodology to efficiently estimate the sparse GGM while taking account the measurement error. Moreover, the asymptotic properties of the proposed method in high dimensional settings were established. Numerical results suggested that the new experimental design and algorithm are superior to existing methods in both estimation and selection accuracy.

3. Statistical genetics. Genetic crosses in model organisms are widely used to understand the heritable architecture of medically relevant phenotypes. To facilitate the large-scale interdisciplinary research, several more sophisticated experimental designs including the Collaborative Cross (CC), the Diversity Outbred (DO) cross and the CC Recombinant Inbred Cross (CC-RIX) have been developed. Data obtained from each separate population reveals different information about the genetic architectures of a target trait.

We are developing statistical methods to study complex human diseases, such as diabetes and obesity, using data from these populations separately or jointly.

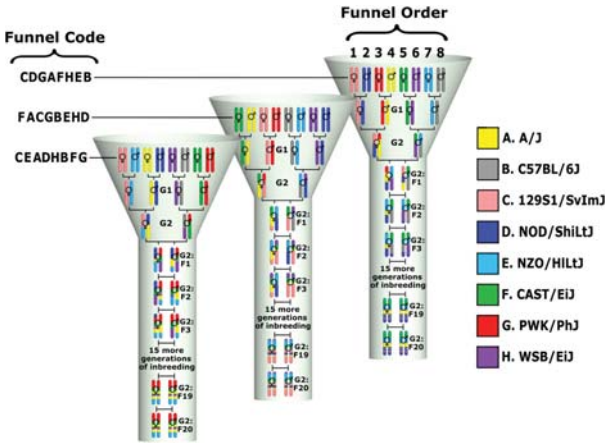

FIGURE 3. The Collaborative Cross-breeding scheme. Adapted from "The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population" by Collaborative Cross Consortium, *Genetics*. 2012; 190(2): 389–401.

RECENT PUBLICATIONS

J.J. Crowley, V. Zhabotynsky, W. Sun, S. Huang, I.K. Pakatci, Y. Kim, J. Wang, A.P. Morgan, J.D. Calaway, D.L. Aylor, Z. Yun, T.A. Bell, R.J. Buus, M.E. Calaway, J.P. Didion, T.J. Gooch, S.D. Hansen, N.N. Robinson, G.D. Shaw, J.S. Spence, C.R. Quackenbush, C.J. Barrick, R.J. Nonneman, K. Kim, J. Xenakis, Y. Xie, W. Valdar, A.B. Lenarcic, W. Wang, C.E. Welsh, C. Fu, Z. Zhang, J. Holt, Z. Guo, D.W. Threadgill, L.M. Tarantino, D.R. Miller, F. Zou, L. McMillan, P.F. Sullivan, F.P. de Villena, "Corrigendum: analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance," *Nature Genetics*, 47(4):353–60 (2015).

J. Phillippi, Y. Xie, D. Miller, T. Bell, Z. Zhang, A. Lenarcic, D. Aylor, S. Krovi, D. Threadgill, F. de Villena, W. Wang, W. Valdar, J. Frelinger, "Using the emerging Collaborative Cross to probe the immune system," *Genes and Immunity* 15(1): 38–46.

C.D. Eversley, Y. Xie, R.S. Pearsall, D.W. Threadgill, "Mapping six new susceptibility to colon cancer (Scc) loci using a mouse interspecific backcross," *G3: Genes, Genomes, Genetics,* 2(12): 1577–84.

X. Zhang, F. Pan, Y. Xie, F. Zou, W. Wang, "COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study," *Journal of Computational Biology*, 17(3): 401–15.