Dear *Bioinformatics* Editors,

Enclosed is our manuscript "A convex optimization framework for gene-level tissue network estimation with missing data and its application in disease architecture", which we are submitting as an original paper to Bioinformatics.

There has recently been a lot of interest in understanding the tissue shared architecture of gene expression (*GTEx Consortium 2017 Nature, Urbut et al 2019 Nat Genet, Dey et al 2017 PLoS Genetics, Pierson et al 2015 PLoS Comp Biol*).

We are interested in genes showing high correlation in expression in multiple tissue-pairs across individuals in the Genotype Tissue Expression (GTEx) project. The biggest challenge here stems from the extensive missing entries in the data matrix resulting from subjects contributing only a subset of tissues. Our paper makes the following contributions.

- We propose Robocov as a novel optimization-based approach for sparse estimation of the correlation and causal network structure of variables from a data matrix with missing entries. Our simulation experiments show that for missing data problems, Robocov correlation estimator outperforms the standard correlation estimator and Robocov causal network estimator outperforms other competing methods like CLIME (Cai et al 2011) and GLASSO (Friedman et al 2008).
- Genes with high average Robocov correlation or partial correlation across tissue-pairs are enriched for pathways related to immune system, heat stress factors and circadian clock. Specifically expressed genes in blood (Finucane et al 2018 Nature Genet) are enriched in these genes; however contrary to expectation, housekeeping genes show no enrichment.
- Top genes prioritized by average Robocov correlation/partial correlation across tissue pairs provide unique autoimmune disease information in a stratified LD score regression (Finucane et al 2015 Nat Genet) analysis. In comparison, genes prioritized by standard correlation based approach provide no disease information.

We present a new method, Robocov, for sparse estimation of correlation and causal structure for missing data problems, and we show the merit of this method over standard approach from both methodological, biological and disease information perspective. We hope the reviewers will find the method and its application interesting. We look forward to your reply.

Sincerely,

Kushal K. Dey
Department of Epidemiology,
Harvard School of Public Health

Rahul Mazumder,
Sloan School of Management,
Massachusetts Institute of Technology