Research II allocation proposal (Multi-tissue methods and GTEx analysis)

Matthew Stephens Lab[1,2*]
1 Department of Statistics, University of Chicago, Chicago, IL, USA
2 Department of Human Genetics, University of Chicago, Chicago, IL, USA
∗ E-mail: mstephens@uchicago.edu

Research goals and potential impact

Variation in gene expression plays an essential role in the etiology of complex human disease. Understanding the genetic factors that underpin the quantitative levels of gene expression (known as eQTLs) provides intermediate insight into biological basis for disease associations identified in genetic mapping studies such as GWAS CITATION. To date, most efforts on eQTL discovery have focused on regulation of expression in a single type of cell or tissue. The importance of shared eQTL between human tissues has been well acknowledged CITATION, yet multi-cell and multi-tissue analyses remain challenging due to the lack of analytic tools that are both statistically powerful and computationally efficient.

Our group had previously developed a statistical framework that improved power to detect eQTL sharing among multiple tissues CITATION. However, a crucial drawback of the framework is that its computational time is proportional to 2^R, thus it becomes intractable to even consider over 10 tissues jointly. On the other hand, however, there is an immediate urge for being capable of performing analyses involving large number of tissues, as we are faced with the challenge of analyzing the Genotype Tissue Expression (GTEx) Project CITATION, a renowned global research effort aiming at understanding the role of regulatory variants in multiple tissues; at its current phase the project has data on 44 tissues.

Recently we have been devising two independent statistical approaches to jointly analyze large number of tissues. Once these new methods are fully developed, they will be applied to analyzing GTEx project for discovering novel patterns of gene regulations. Furthermore, our methods can potentially be applied to generic problems involving assessment of genetic factors across multiple data sources, a type of data integration problem of general interest in the emerging field of genetic data sciences.

Results from previous allocations and anticipated tasks for future allocations

Our previous approach to joint tissue analysis had been applied to the GTEx pilot phase involving 9 tissues. The result have contributed to the *Science* paper published in May, 2015 CITATION.

For large tissue problem posed by the current GTEx data we are developing two approaches to model, respectively, the qualitative and quantitative heterogeneity among tissues in lower dimensional spaces of probability measures, thus avoiding a growth in computation from 2^9 to 2^{44}. We are close to finishing up the theoretical part and a proof-of-concept computer implementation of the new methods. However development of high quality software for the new methods are still in progress and we are yet to perform comprehensive evaluation and comparison of the two approaches in simulation studies, and to complete genome-wide joint eQTL mapping in 44 GTEx tissues.

In preparing for analysis-ready data-sets, we have developed an efficient data management framework based on HDF5 file system. Approximately 50 billion lines of text files have been processed to created the data infrastructure for GTEx release in April, 2015. We are expecting a complete update of source data, and consequently the need to rebuild the data system, in October, 2015.

Estimate of requested resources

Based on SU consumption in the method development phase last quarter (70,000 SUs) we estimate a usage of 550,000 SUs and 1TB of storage upon completion of the project in the next 12 months. The SU requested will be dedicated to mainly 3 tasks: data processing, methods development and comparison and GTEx analysis.

For data set on each tissue it requires 450 SUs (16–20 CPUs on one node with 1GB memory per CPU) to perform genome-wide single tissue analysis and organizing the outcome into HDF5 system. It requires another 100 SUs (1 CPU on one node with 100 - 200GB memory) to merge and compute summary statistics from output across all tissues. Since the input information for the two methods we are developing are partially different, an additional 20% SUs is required to create companion data sets to provide information specifically required by different methods. In sum the data processing step required 24,000 SUs from previous allocation. The data to be released in October 2015 is a major update in GTEx project, approximately twice times in size compared to current release. We therefore estimate that required resource for data processing is 48,000 SUs. Our data is highly compressed (in `bz2` format with level 9 compression), taking about 300GB in space. We expect 600GB storage requirement for the October release.

For the quantitative model we are developing, it requires 200 SUs (16–20 CPUs on one node) to fit the model on 50,000 data points, which is about 0.05% of the total data set. In continued development of the method we will experiment with 0.05% data improve model parameter configurations (SUs required may be negligible). The ultimate goal is to apply the model on 50% data points which we believe will provide good representation for the entire data. The expected resource usage is 200,000 SUs.

For the qualitative model we need 1% data points to provide reliable experimental results. Depending on the accuracy required for convergence of algorithm it requires 1,000 – 3,000 SUs (16 – 20 CPUs on one node with 200G memory total) to complete the paralleled portion of computation involved. The rest part of the computational algorithm is not yet paralleled, but we estimate a requirement of 2,000 - 6,000 SUs in conducting it. In polishing the model and implementing its parallel version, we expect non-trivial SU consumption (about 20,000 SUs). Once we apply the model to 50% data points we expect usage of 170,000 – 470,000 SUs. Thus we request 300,000 SUs for data analysis under this model.

Substantial summary statistics for each gene and regulatory variant pair across genome will be generated in the joint GTEx data analysis. Given that intermediate output are constantly cleaned up from disk we expect about 300GB storage space for the genome-wide summary statistics from final size of summary statistics is about the same

References