Quantifying Human Genetic Variation At Scale

Research Computing Center, University of Chicago
Research II Allocation Request

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

Introduction

In our group, we develop and apply statistical methods to understand the genetic basis and biological mechanism of complex human phenotypes. The generation of massive data sets via next-generation genome sequencing efforts promises huge payoffs in the understanding of human phenotypic variation and human disease as it enables a transition from ad hoc hypothesis-driven experimentation to comprehensive, principled investigation, but it also creates new challenges in the scale of analysis that we hope to address in our work ahead.

We propose two high-impact computational projects on RCC's Midway cluster in which we apply Bayesian statistical methods developed in our group to full-scale genomic data sets. Each project uses a different mechanistic basis to extract new insight from this deluge of genomic data. The first project uses known biological pathways and existing, publicly available genome-wide association study (GWAS) summary statistics to improve our ability to resolve new genetic determinants of human phenotypes and disease. We request 600,000 service units (SUs) and will use our own storage assets on Midway to complete this project. The second project focuses on the ways in which genetic varation can specifically modulate gene expression; this work relies on the novel dataset currently being acquired by the Genotype Tissue Expression (GTEx) consortium. This project will require 190,000 SUs in addition to 1 TB of persistent storage. In total, we request 790,000 service units (SUs) and 1 TB of persistent storage.

Constructing an atlas of biological pathways and genetic variants

Research goals and potential impact

Recently, curated knowledge-bases of biological pathways have been created and shared [1]. Peer-reviewed pathways effectively organize functionally related genes that are involved in the development of certain complex traits. Leveraging these public assets in genome-wide association studies (GWAS) can help gain insight into the underlying biology of complex traits, which can then be applied to improve the health of people.

Motivated by this observation, we have proposed an efficient Bayesian approach [2] to integrate pathway enrichment analysis [3] with variant prioritization [4]. This method, however, is limited by the requirement of individual-level genotype and phenotype data, which are rarely accessible for large GWAS. Using a recently-developed Bayesian large-scale regression approach [5], we soften the requirement of [2]; the modified method relies solely on publicly available GWAS summary statistics. Hence, we can easily apply our methods on a wide range of human phenotypes including anthropometric traits (e.g. adult height), immune traits (e.g. Crohn's disease), metabolic phenotypes (e.g. blood lipid levels), and psychiatric diseases (e.g. schizophrenia). Our aim is to release an "atlas" of biological pathways and genetic variants across multiple complex human traits using GWAS summary statistics.

The atlas is likely to be useful for basic, translational, and clinical research in biology and medicine. First, our comprehensive results can help statisticians and computational biologists benchmark their methods for pathway analysis and/or variant prioritization. Second, our data-driven method links loci and

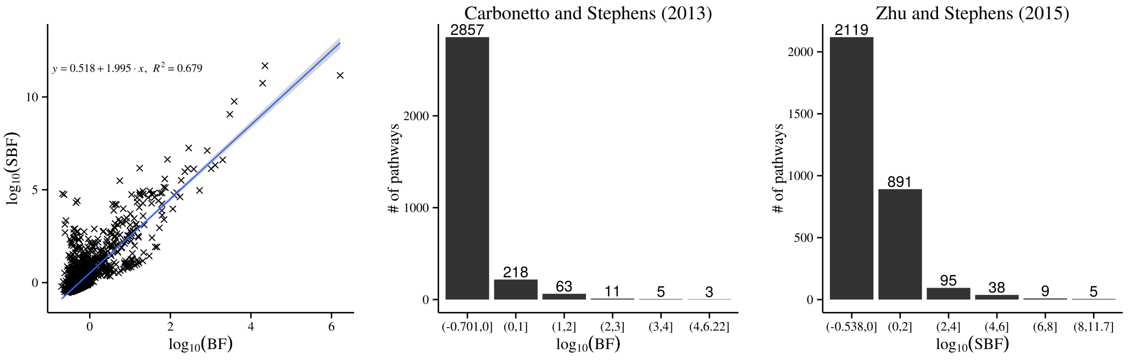

Figure 1. Comparison of [2] and [5] for Crohn's disease. BF and SBF measure the strength of pathway enrichment in [2] and [5] respectively. Both methods were run with 3,157 pathways.

pathways to pathogenesis jointly and therefore is statistically more powerful than conventional analysis (e.g. GWAS) for uncovering disease risk factors, which might suggest innovative treatment targets for experimental biologists and physicians. The benefit of pathway-based analysis has been empirically validated in complex disease studies. For example, the PI3K/RAS pathway is deregulated in 45% of 316 ovarian cancer cases [6] even though all seven genes in this pathway have relatively weak signals (0.5%–18%).

Results from previous allocations

Our previous approach [2] requires individual-level genotypes and phenotypes. To circumvent the limited availability of individual-level data, our new method is built on a Bayesian regression model [5] that only requires summary data. We apply the two methods to the Wellcome Trust Case Control Consortium (WTCCC) GWAS data for Crohn's disease [7]. The summary statistics-based method obtains results comparable to the analysis using individual-level data (Figure 1).

Our new method uses a parallel algorithm based on mean-field variational approximation [8]. Hence, compared with the serial implementation in [2], the new method is more scalable. Table 1 reports the timing results of two methods with varying number of CPUs on two experiments.

Experiment	Partition	Number of CPUs	Wall time (s)	
			Carbonetto-Stephens (2013)	Zhu-Stephens (2015)
1	`westmere`	1	403.364	387.661
1	`westmere`	2	402.207	282.357
2	`bigmem`	1	1416.164	2083.683
2	`westmere`	2	1209.082	1035.426
2	`westmere`	4	1166.437	720.544

Table 1. Scaling results of two methods. Experiment 1 consists of 11,861 genetic variants on two chromosomes. Experiment 2 consists of 30,939 genetic variants on four chromosomes.

Estimate of requested resources

To create the pathway-phenotype atlas, we will retrieve 4,000 curated biological pathways from web databases and then apply our new method on 30 full sets of publicly available GWAS summary statistics, each including 1–2 $\times 10^6$ genetic variants. For each GWAS dataset, our analysis will consist of two phases. Computation in Phase I will be performed on all genetic variants across the genome, using a multi-threaded implementation. Phase II will only use genetic variants assigned to pathways, and, depending on the pathway size, the number of variants can be in the 10^2 – 10^4 range. Both Phase I and II can be further decomposed into multiple concurrent jobs on Midway.

Based on our pilot experiments performed under our previously accepted proposals on Midway, analyzing a typical set of GWAS summary statistics requires 20,000 SUs, broken down as follows. Phase I requires 16–20 CPUs on one node (`sandyb` or `ivyb`), 1–3 GB memory per CPU, and 600–650 hours of wall time. Phase II will require 2–4 CPUs on one node (`sandyb` or `ivyb`), 0.5–5 GB memory per CPU, and 1,500–1,800 hours of wall time. In addition, use of a `bigmem` node is needed to pre-process the data set, which requires 1 CPU with 150–200 GB memory and 20–30 hours of wall time.

Applying this computational workflow to the 30 publicly available sets of GWAS summary statistics justifies the request for 600,000 SUs to generate the proposed atlas of biological pathways and genetic variants. All of the GWAS data sets as well as the results of this workflow can be stored in existing space we have purchased via the Cluster Partnership Program, so we are not requesting additional persistent storage for this project.

Applying multi-tissue methods to GTEx expression data

Research goals and potential impact

Variation in gene expression plays a crucial role in the etiology of complex human disease. Understanding the genetic factors that underpin the quantitative levels of gene expression (known as expression quantitative trait loci, or eQTLs) provides intermediate insight into biological basis for disease associations identified in genetic mapping studies such as GWAS [9]. To date, most efforts on eQTL discovery have focused on a given genetic variant's contribution to gene expression within a single cell type or tissue. Although the importance of shared eQTL between human cells and tissues has been well acknowledged [10], multi-cell and multi-tissue analyses remain challenging due to the lack of adequate data sources and of analytic tools that are both statistically powerful and computationally efficient.

Recently, the Genotype Tissue Expression (GTEx) Project, a renowned global research effort aiming at understanding the role of regulatory variants in multiple tissues, have completed its Pilot Phase [11]. Our group, as part of the Statistical Genetics Committee in GTEx, previously developed a statistical framework that improved power to detect eQTL sharing among multiple tissues [12]. However, a critical drawback of the framework is that its computational time is proportional to 2^R where R is the number of tissues; it becomes intractable to even consider over ten tissues jointly. On the other hand, however, the GTEx consortium will release eQTL data on 44 tissues in approximately 900 post-mortem human donors by the end of 2015, a rich and expensively acquired data set that demands a correspondingly rich analysis in the near future.

Results from previous allocations

Our previous statistical framework for joint tissue analysis was applied to the GTEx pilot phase involving nine tissues. Computations were performed on the Midway cluster. The result contributed to the *Science* paper published in May 2015 [11].

We are actively engaged in devising two independent statistical approaches to jointly analyze large numbers of tissues. Once these new methods are fully developed, we will apply them to the full GTEx data set upon its release. Furthermore, our methods can potentially be applied more generally to problems involving assessments of genetic factors across multiple data sources, a type of data integration problem of broader interest in the emerging field of genetic data science. By using lower dimensional spaces of probability measures, we will avoid the exponential growth in the computational cost of joint multi-tissue analysis that hampered the scaling of our previous methods.

In preparing for analysis-ready data-sets, we have developed an efficient data management framework based on an HDF5 file format. In a previous allocation on Midway, we processed approximately 5×10^{10} lines of text to prepare the smaller April 2015 GTEx release for analysis. 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

Our request for the GTEx eQTL project is dedicated to three tasks: ingesting the GTEx data into a format suitable for joint tissue analysis and the development and production application of two statistical methods.

Preparing the upcoming October 2015 release of GTEx data for analysis will involve a workflow that is directly scaled up from our previous work on the April 2015 release, which contained about half as many volunteers. Changes in the consortium's pre-release data processing means that we will need to re-process the full data set. To ingest the raw data as released by GTEx into a form amenable for analysis will require 1000 SUs (16 CPUs on one sandby or ivb node with 1 GB memory per CPU for approximately 60 hours of wall time, broken into several concurrent jobs) to perform genome-wide single tissue analysis and to organize this data into our HDF5 system. We will need another 200 SUs per tissue on a bigmem node (1 CPU with 100–200 GB of memory for approximately 200 hours per tissue) to merge and compute summary statistics. This work will thus require 50,000 SUs for the upcoming data processing task. Storing the results of this processing will require 600 GB of peristent storage.

For the first "quantitative" model we are developing, we will need approximately 200 SUs (16 CPUs on one sandyb or ivyb node for 12 hours of wall time) to fit the model on a sampling of 50,000 data points, which is about 0.05% of the total data set. As we complete development of this model, we will work at this scale to improve model parameter configurations, requiring a negligible amount of compuational effort. The limits on the expressiveness of the quantitative model means that we will ultimately only need to sample 10% of the total data points in our production work, requiring 40,000 SUs.

For the second "qualitative" model, we will develop the model at a scale of 1% of the full GTEx data set and scale it to cover 10% of the data set in production. Based on experimentation, the first part of this model will require approximately 20,000 SUs (16 CPUs on one bigmem node with 200 GB memory for 1250 hours of total wall time). We are engaged in parallelizing the remaining part of the qualitative model but estimate that it will consume 60,000 SUs on sandyb or ivyb nodes. In tuning the model and implementing its parallel version, we expect non-trivial computational needs (about 20,000 SUs) as we bring the methods up to scale. Thus, this second model will consume approximately 100,000 SUs.

Storing the artifacts of these two models will require a further 400 GB of persistent storage, justifying the remainder of our request of 190,000 SUs and 1 TB of persistent storage for this project.

References

1. Emek Demir et al. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, 2010.

2. Peter Carbonetto and Matthew Stephens. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. *PLoS Genetics*, 9(10):e1003770, 2013.

3. Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010.

4. Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

5. Xiang Zhu and Matthew Stephens. Bayesian large-scale regression with GWAS summary statistics; (Program #1334F). In *Presented at the 60th Annual Meeting of The American Society of Human Genetics, Baltimore, MD*, 2015.

6. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

7. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

8. Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

9. Frank W. Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, February 2015.

10. Antigone S Dimas et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, 325(5945):1246–50, September 2009.

11. K. G. Ardlie et al. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

12. Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, 9(5):1–8, 2013.