Relating Genotypes to Phenotypes in Complex Environments: Generalized Linear Model (GLM) Based Quantitative Trait Locus (QTL) Analysis

Gregory M. Striemer, Ali Akoglu, David Lowenthal, Peter Bradbury, Liya Wang, Matthew Vaughn, Stephen Goff

gmstrie@email.arizona.edu+, akoglu@ece.arizona.edu+, dkl@cs.arizona.edu*, pjb39@cornell.edu++, wangli@cshl.edu**, vaughn@cshl.edu**, sgoff@iplantcollaborative.org#

+Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, *Department of Computer Science, University of Arizona, Tucson, AZ
++USDA Agricultural Research Service, Ithaca, NY, **Cold Spring Harbor Laboratory, NY, #iPlant Collaborative, University of Arizona, Tucson, AZ

Motivation and Goals

- Elucidating the relationship between plant genotypes and the resultant phenotypes in complex (e.g., non-constant) environments is one of the foremost challenges in plant biology (National Research Council, 2008).
- Plant phenotypes are determined by often intricate interactions between genetic controls and environmental contingencies. In a world where the environment is undergoing rapid, anthropogenic change, predicting altered plant responses is central to studies of plant adaptation, ecological genomics, crop improvement activities (ranging from international agriculture to biofuels), physiology (photosynthesis, stress, etc.), plant development, and many many more.
- Natural phenotypic variation within a species or population is largely quantitative, polygenic, and controlled by the interaction of environmental and genetic factors. A major goal of any phenotype to genotype linkage is to be able to rapidly identify and predict the causal genetic variation underlying the phenotypic variation.
- Current technological advances in phenotyping and genotyping are allowing increasingly rapid advances in determining the genetic architecture underlying diverse phenotypes.
- However both genomics and phenotyping systems have begun to generate vastly more data than can be easily interrogated on local systems or by non-expert laboratories.
- As such, our goal is to develop GPU implementation of the General Linear Model (GLM) to statistically link genotype to phenotype and dramatically decrease the execution time for GLM analyses.
- Results of this study will enable larger, more intensive genetic mapping analyses to be conducted.

Background and Problem Statement

The **Plant Science Cyberinfrastructure Collaborative (PSCIC)** program is intended by NSF to create a new type of organization – a cyberinfrastructure collaborative for the plant sciences - that would enable new conceptual advances through integrative, computational thinking.

The **"iPlant Collaborative" (iPC, http://www.iplantcollaborative.org/)** utilizes new computer, computational science and cyberinfrastructure solutions to address an evolving array of grand challenges in the plant sciences.

***iPG2P: Relating Genotype to Phenotype in Complex Environments* is one of the current projects of the iPC to understand the link between the genetic variation with the physical variation through the combined and integrated efforts of specialists in functional-, quantitative-, and computational genetics/genomics, bioinformatics, modelers, physiologists, computer scientists.**

Given
- A particular species of plant (*e.g.* maize, rice, soybean)
- genetic description of an individual (*Genotype*)
- growth environment
- trait of interest (flowering time, yield, or any of hundreds of others)

Predict, in **non-constant** environments
- The quantitative result (*Phenotype*)

Reverse problem: What genotype will yield the desired result in a given environment?

Challenge

Single-SNP test: a couple of minutes

1000-replicate bootstrap: a few hours

Runtimes only gets larger (months to years) for more combinatorial analyses

6.5 million markers = Two Arabidopsis-sized genomes @ 5% diversity

38,963 expression phenotypes: # transcripts in Arabidopsis measured by UHTS

Terminology

Single-nucleotide polymorphism (SNP; pronounced "snip"): a single base pair within a DNA sequence that can vary among individuals. An example of a SNP is the change from A to T in the sequences AATGCT and ATTGCT.

Genetic variation in a DNA sequence occurs when a single nucleotide in a genome is altered. SNPs are usually considered to be point mutations that have been evolutionarily successful enough to recur in a significant proportion of the population of a species.

Quantitative Trait Locus (QTL) analysis is a statistical method that links two types of information—phenotypic data (trait measurements) and genotypic data (usually molecular markers)—in an attempt to explain the genetic basis of variation in complex traits. QTL analysis allows researchers in fields as diverse as agriculture, evolution, and medicine to link certain complex phenotypes to specific regions of chromosomes. The goal of this process is to identify the action, interaction, number, and precise location of these regions.

Mapping QTL is to identify genomic loci that associate with the phenotype and to estimate their genetic effects.

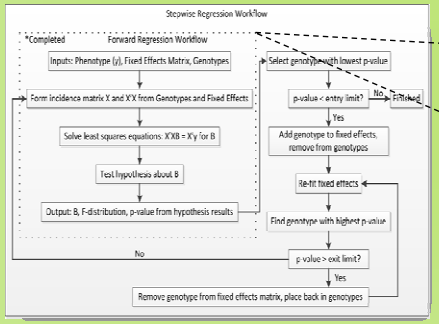


Figure 1. Stepwise regression workflow
GPU implementation of Forward Regression is signified

GLM Based QTL Mapping

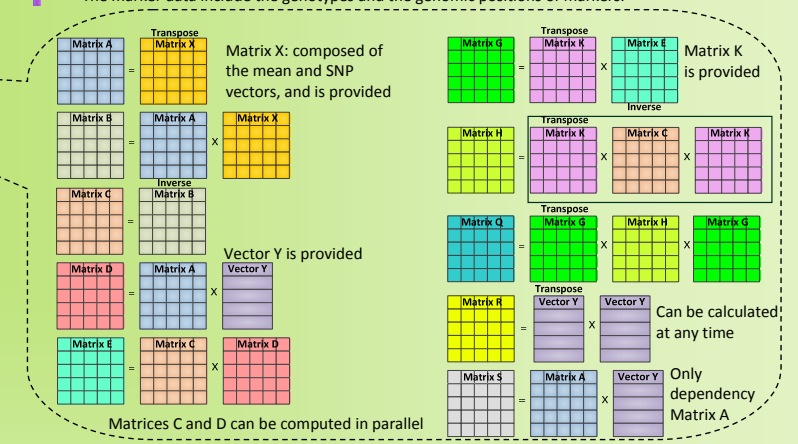
To begin a QTL analysis, scientists require two sets of data.

1. Two or more strains of organisms that differ genetically with regard to the trait of interest. For example, they might select lines fixed for different alleles influencing egg size (one large and one small).
2. Genetic markers (SNPs) that distinguish between these parental lines. Molecular markers are preferred for genotyping, because these markers are unlikely to affect the trait of interest.

To carry out the QTL analysis the parental strains are crossed, resulting in heterozygous (F1) individuals, and these individuals are then crossed. Finally, the phenotypes and genotypes of the derived (F2) population are scored. Markers that are genetically linked to a QTL influencing the trait of interest will segregate more frequently with trait values (large or small egg size in our example), whereas unlinked markers will not show significant association with phenotype.

Figure 1 illustrates the workflow for forward regression (enclosed in dotted box **along with its associated sequence of matrix intensive calculations**), and how it is integrated with the stepwise regression workflow. Outside of forward regression, the stepwise regression consists of seven different steps.

GLM of multiple QTL: We consider experimental crosses derived from two inbred lines (for example, F2, backcross and recombinant inbred lines). Observed data in QTL studies consist of phenotypic values of a complex trait, genetic markers across the genome, and/or some relevant environmental factors (covariates). The marker data include the genotypes and the genomic positions of markers.



Results

- **CPU Configuration:** Intel Xeon @2.33GHz with 2GB RAM running serial Matlab version of the Forward Regression
- **GPU Configuration:** NVIDIA Tesla C1060 @1.3GHz running parallel version
- Data set investigates the role of D2a gene in the host chromosome
- Testing carried out using chromosome 10 map information which lists the genetic and physical positions of the markers.
- Performance evaluated based on 100K SNPs each of length 191

SNP Size	GPU Time (msec)	CPU Time (sec)	Speedup
5K	8		
10K	11		
15K	13		
30K	**17**		
50K	30		
100K	**62**	**11**	**177x**

- **Beyond 30K** threading power is fully utilized on the GPU
- Forward regression achieved **177x** speedup over the Matlab version.
- GPU performance includes data transfer to the GPU and back to the host

Conclusion and Future Work

- The routines within forward regression (Figure 1) have been highly tailored to the GPU architecture. New versions of matrix transpose, multiplication, and inversion have been engineered to deal specifically with the problem of regression in the context of SNPs.
- Our next step is to modify the forward regression CUDA algorithm to handle the addition of extra genotype data to the fixed effects matrix based on regression results of the genotypes.
- On the GPU, finer granularity produces higher performance. Therefore, each of the major matrix routines will be further modified to automatically decrease the granularity as the effects matrix grows. This will require multiple threads performing the regression on a single SNP in a genotype, rather than a single thread assigned to a single SNP.
- The multi-GPU implementation of the Stepwise Regression flow will be evaluated using the 128-NVIDIA Quadro Plex S4 based system at the Texas Advanced Computer Center. This effort will require features to automatically distribute the workload based on problem size, as well as mechanisms for inter-GPU communication when regression is performed on a single genotype across multiple GPUs.
- The GPU to GPU communication feature will be required so that p-values and F-distributions can be shared in order to determine which SNPs should be taken from the genotype and added to the effects matrix, which will be the same on all GPUs.

Acknowledgements

This work is supported by the iPlant Collaborative. The iPlant Collaborative is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#EF-0735191).