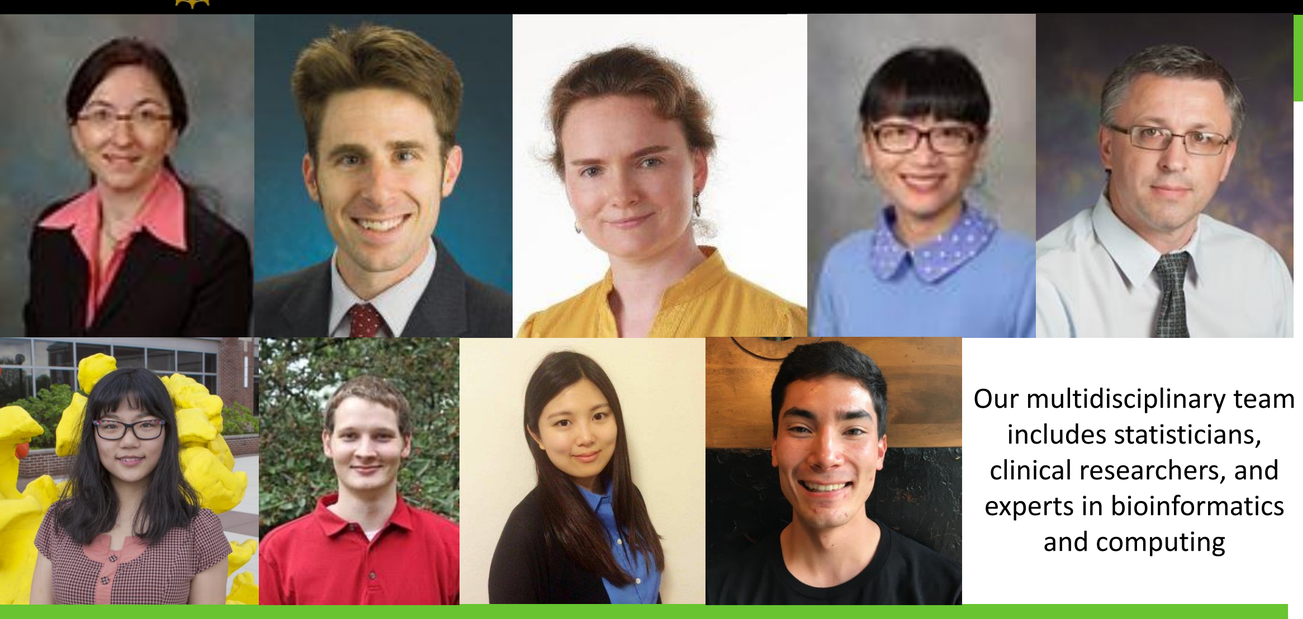
# SCALING THE COMPUTATION OF EPISTATIC INTERACTIONS IN GWAS DATA

M. Allen, A. Chen, N. Ertekin-Taner, J. Heldenbrand, V. Kindratenko, A. Lipka, L. S. Mainzer, X. Nie, K.R. Wagner, L. Wang, C. Younkin

AN NSF INDUSTRY/UNIVERSITY COOPERATIVE RESEARCH CENTER



Our multidisciplinary team includes statisticians, clinical researchers, and experts in bioinformatics and computing

## NEED AND INDUSTRIAL RELEVANCE

Most GWAS data collected by academia or industry have been analyzed only for associations between phenotypes and polymorphisms at single loci. These associations account for only a small proportion of the phenotype heritability.
**Full analysis of epistatic effects** would greatly increase usefulness of the data.

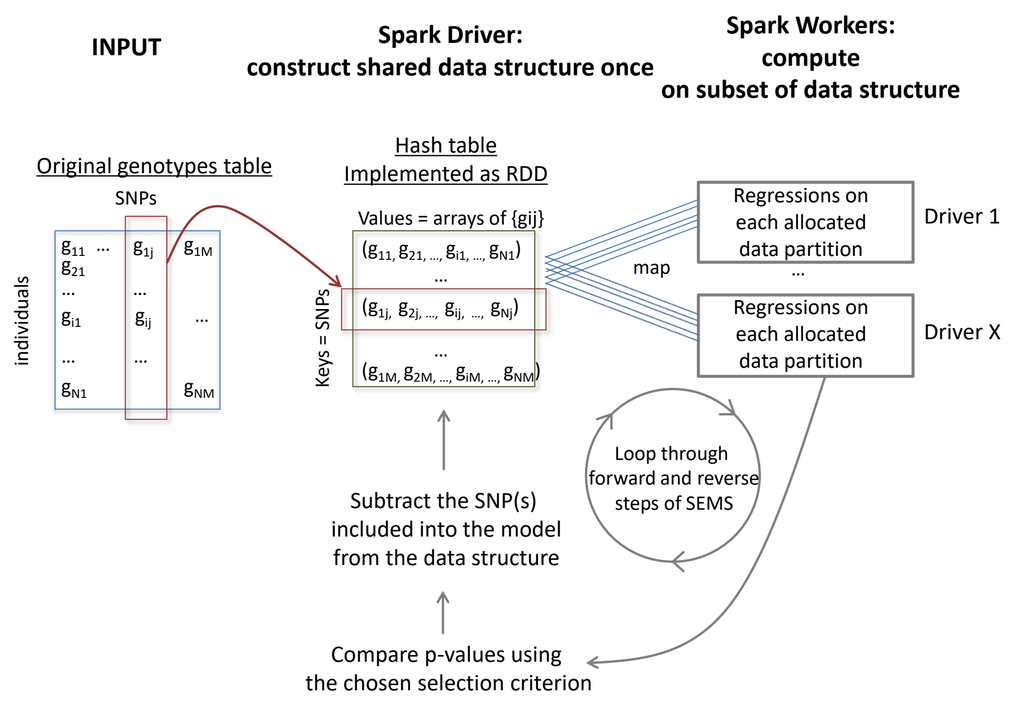
**The aim of our project** is to develop software in which a complete model for additive and epistatic interactions of multiple orders are calculated efficiently for continuous-trait GWAS . This massively parallelized solution will have options to use prior biological information to focus the model construction on relevant pathways.

The procedures and software developed by the project will immediately benefit the research in neurodegenerative diseases and plant biomass. Those products will be equally applicable to other **complex traits of medical and agricultural importance**, such as obesity, autism, and plant disease resistance.

**Cloud deployment of the scalable software** will further facilitate adoption in areas where multidimensional datasets comprised of transcriptomics and metabolomics as well as in vitro or in vivo phenotypes pose severe computational challenge.

## APPROACH

### A possible data model for Spark implementation




**Preliminary C/MPI code models:** https://github.com/lsmainzer/EpiQuant

| | Pros | Cons |
|---|---|---|
| **Spark with Scala API** | 1. Built-in fault tolerance<br>2. Code is concise<br>3. Scala supports both OO and functional programming paradigms. | 1. Slower<br>2. Requires specialized infrastructure |
| **C with MPI/ OpenMP** | 1. Fast<br>2. Compatible with traditional HPC | 1. Longer development time<br>2. Code is verbose<br>3. Lack of fault tolerance |

## PROJECT GOALS AND OBJECTIVES

(1) Experiment with different **statistical approaches** for analysis of eGWAS or other phenotypically complex GWAS data, with an emphasis on detection of epistatic interactions among SNPs. Combine their strongest elements for the selection of the best model that includes additive effects, as well as two-way and higher order epistatic interactions. Include biological information to focus the model search.
- Stepwise Epistatic Model Selection (SEMS).
- Least Absolute Shrinkage and Selection Operator (LASSO).
- LD, biomolecular pathways, gene sets.

(2) Design a scalable, efficient and easy to deploy **software** for building the above model on diverse kinds of eGWAS data.
- Extensively compare, test and improve software components.
- Generate algorithm and memory models for SEMS and LASSO.
- Determine which computational components can utilize GPU vs CPU.
- Identify when it is appropriate to use SPARK vs. MPI.
- Design an optimized application that runs on appropriate platform.
- Enable input/output compatibility with PLINK.
- Enable plugin development for data pre- and post-processing.
- Develop a containerized method of deploying this application, to facilitate adoption in a variety of environments.

## PRELIMINARY RESULTS

### Modeling epistasis as two-way interaction terms

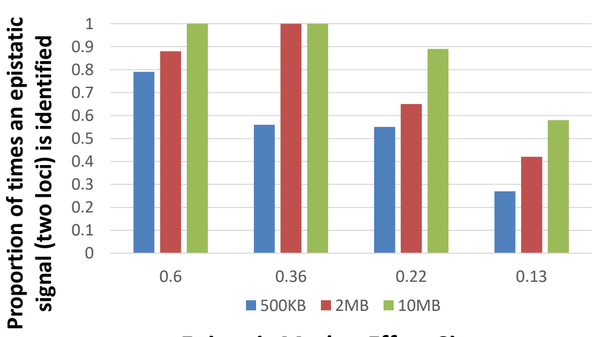
$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i$$

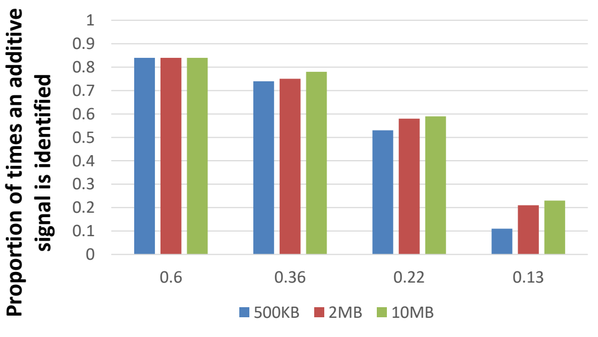
- $Y_i$: Phenotype of $i^{th}$ individual
- $\mu$: Grand Mean
- $\sum_{j \in I} \beta_j X_{ij}$: Main (additive) effects of genomic markers
- $\sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv}$: Two-way interaction (epistatic) effects between genomic markers markers
- $\varepsilon_i$: Random error term

- $I$ is a subset of markers with additive effects in model
- $U$ is a subset of markers with two-way epistatic effects in model
- Determining the optimal model:
  - AIC, BIC, mBIC
  - Permutation procedure

Bogdan et al. (2004)

SEMS (now multithreaded Java code) has been validated on 1,106 SNPs drawn from a dataset of 1,555 Maize NAM individuals to conduct a simulation study. At higher heritabilities, SEMS successfully detects additive and epistatic QTN regardless of the effect sizes of the simulated QTN, with very low misidentification. Figures below illustrate this for heritability of 0.5, for markers within a given window of each QTN. At heritability of 0.9 SEMS detects nearly 100% of the signal. Horizontal axis denotes QTN effect sizes.




## TEST CASES, USE CASES

**Case 1: Identifying genetic interactions that influence brain gene expression and Alzheimer's disease.** Zou *et a*l. (2012) Brain Expression Genome-Wide Association Study Identifies Human Disease-Associated Variants. PLoS Genet 8(6): e1002707.

- eGWAS: ~200 AD and ~200 non-AD samples,
- temporal cortex and cerebellum tissue,
- ~300,000 GWAS SNP genotypes and gene expression measures,
- 18,401 genes (24,526 probes).

**Case 2: Plant GWAS.**

- Maize NAM panel: 1,106 SNPs, 1,555 individuals – a small test case
- Scale-out target:
  - 100,000s – millions SNPs,
  - 10s of thousands of individuals ,
  - dozens of phenotypes

## IMPACT, MILESTONES, DELIVERABLES

Developing a streamlined method for multi-locus epistasis analysis will
- enable identification of hitherto unknown combinations of genetic loci that impact important phenotypes, such as gene expression patterns in AD, elucidating the immediate therapeutic targets;
- lead to improved understanding of the biological pathways and cellular mechanisms that affect trait variance, such as crop biomass;
- inform therapeutic strategies aimed at disease;
- guide crop and livestock breeding programs for increased performance.

Year 1: We will have a fully developed software design for submission of large GWAS analyses, with appropriate choices of the hardware and cluster architectures, and a selection of statistical tools and options.

Year 2: We will have a production-grade software developed, containerized, internally tested, and made opensource for download and deployment in HPC or cloud environment.

CENTER FOR COMPUTATIONAL BIOTECHNOLOGY AND GENOMIC MEDICINE

ILLINOIS MAYO CLINIC