LMM-22

An enhanced Linear Mixed Model (LMM) approach for Genome-wide Association Studies (GWAS) for the prediction of diseases and traits among humans from genomics data

Siddharth Sharma, 9th grade,
Basis Independent Silicon Valley School

Why individuals have different hair colors, eye colors or specific diseases?

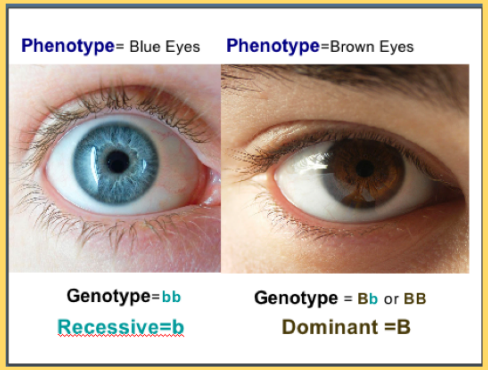
Genotype

+

Environment

+

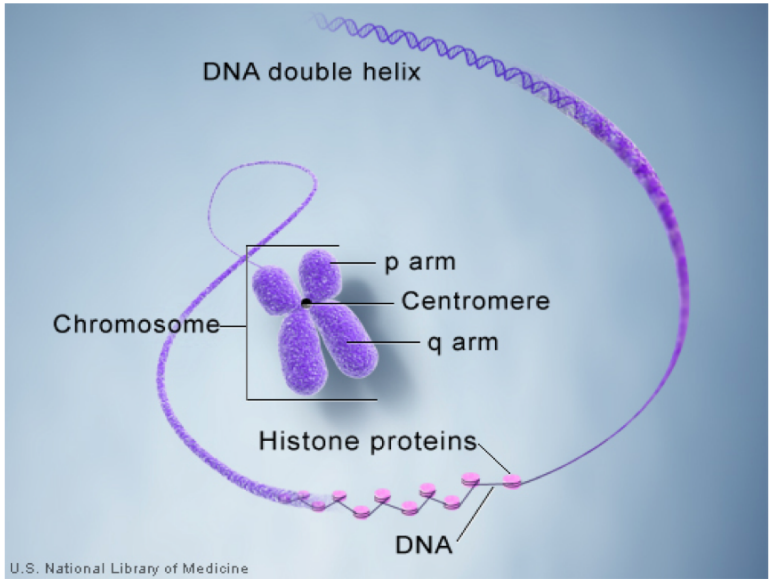
Genotype-Environment Interactions

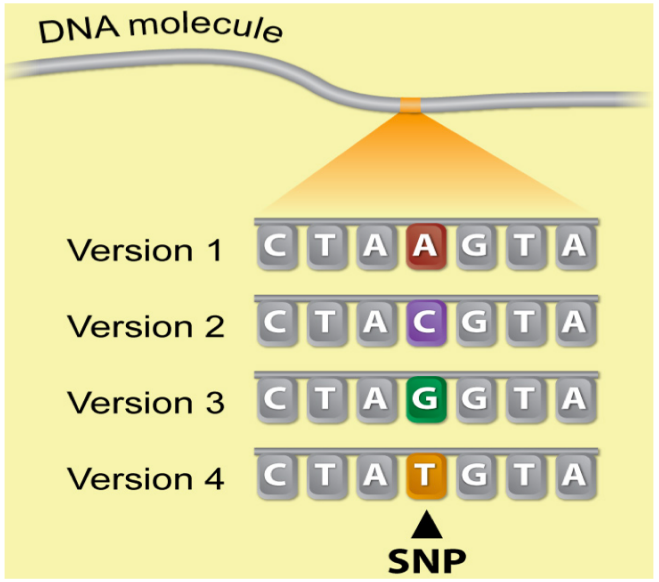


Genes associated with a phenotype or disease

https://www.ncbi.nlm.nih.gov/guide/howto/find-gen-phen/

Single Nucleotide Polymorphisms (SNP)





Single nucleotide polymorphisms, called SNPs are the most common type of genetic variation among humans. Each SNP represents a difference in a single DNA building block.

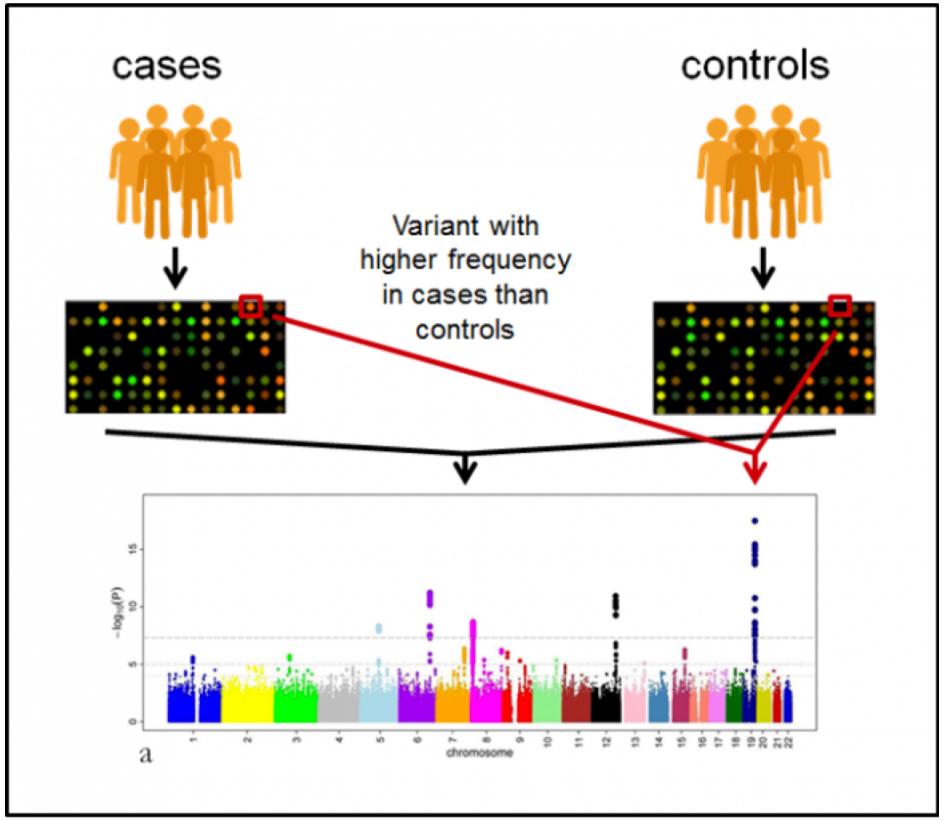
SNPs occur normally throughout a person's DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome.

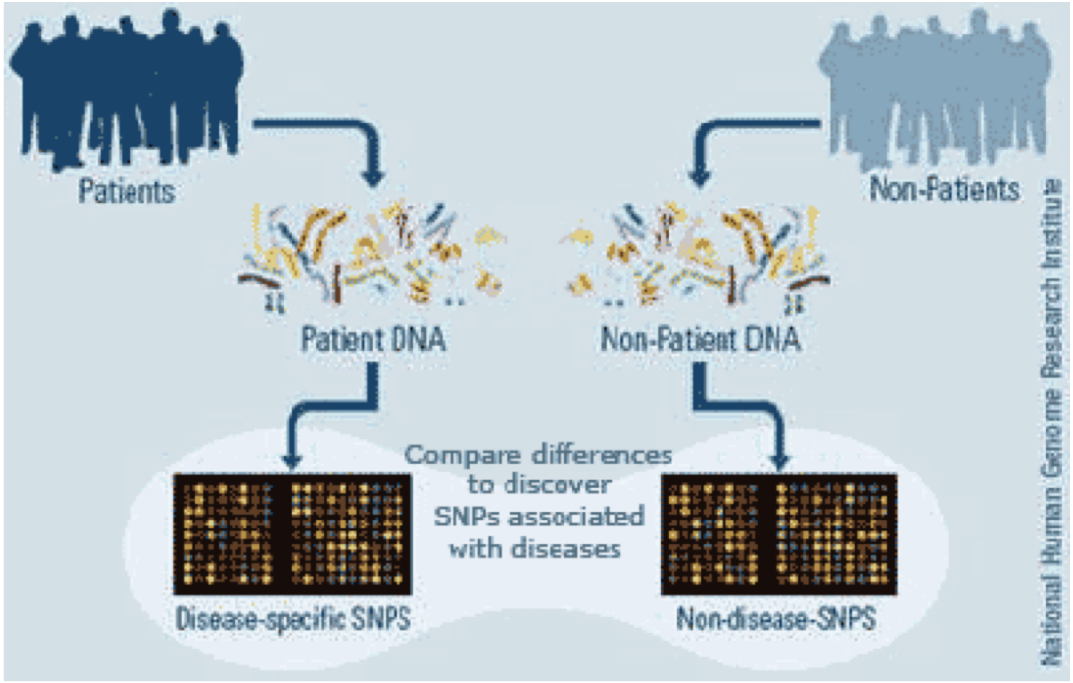
SNPs can act as biological markers, helping scientists locate genes that are associated with disease.

Genome-wide Association Studies

Genome-wide association studies (GWAS) have become a common way for scientists to identify genes involved in human diseases. This method searches the genome for small variations in SNPs that occur more frequently in people with a particular disease than in people without the disease.

Each study can look at hundreds or thousands of SNPs at the same time. Researchers use data from this type of study to pinpoint genes that may contribute to a person's risk of developing a particular disease.





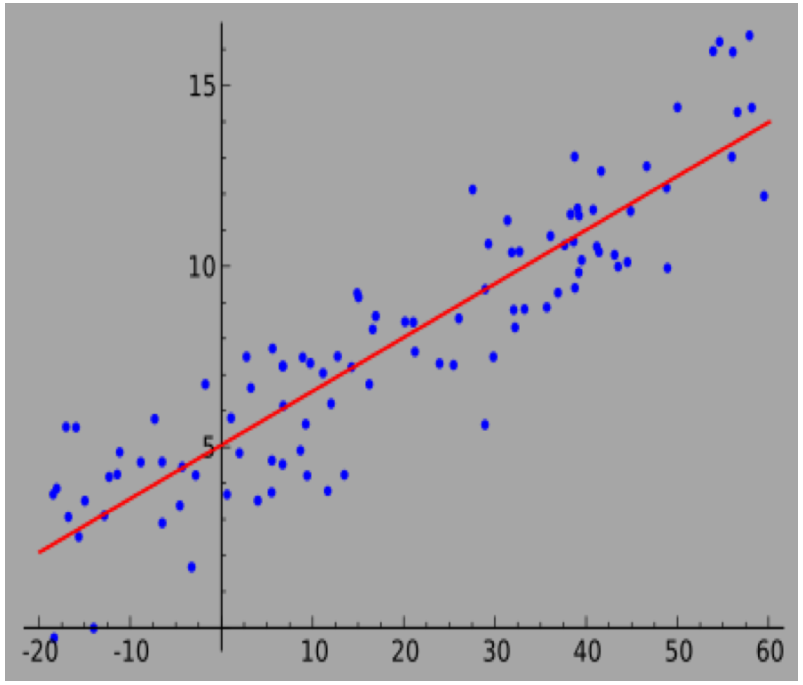
GWAS studies compare the DNA of participants having varying phenotypes for a particular trait or disease. GWAS studies have been used successfully to identify genetic variations that contribute to the risk of type 2 diabetes, Parkinson's disease, heart disorders, obesity, Crohn's disease and prostate cancer etc.

GWAS studies are difficult, time-consuming, and expensive

Data Science and statistical computations used for GWAS

GWAS studies require analysis of genomics data from tens of thousands of individuals. The human genome contains roughly 10 million SNPs. Hence, GWAS studies are difficult, time-consuming, and expensive to look at such large number of SNPs and then determine whether specific SNPs play a role in human disease.

Linear Regression for GWAS Studies



$$y = X\beta + \varepsilon$$

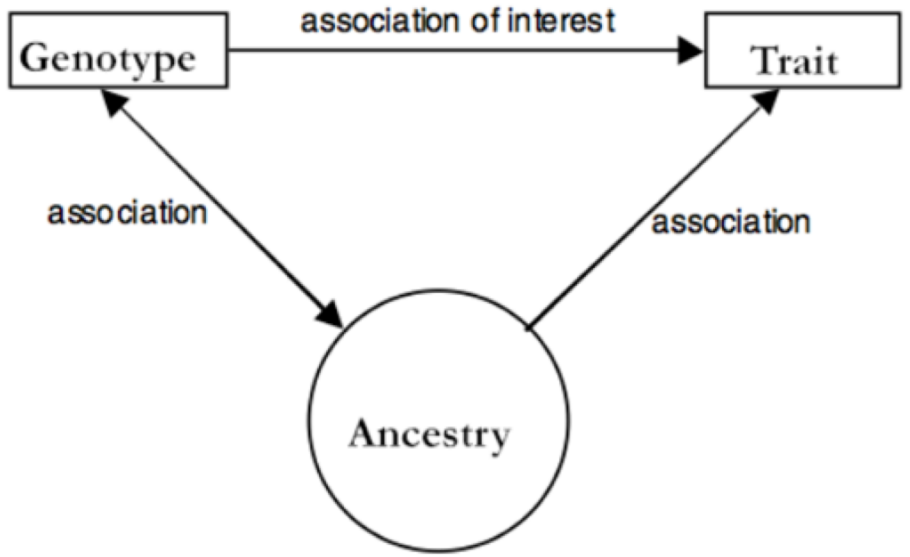
where y is vectors of dependent variables, ε is noise, β is parameter vector, and X is a matrix of independent variables.

However, the presence of confounding variables (such as population structure) in GWAS analysis requires more sophisticated models.

Confounding variables in GWAS Studies

Population structures--presence of subgroups in population with ancestry differences.

Family relatedness: alleles transmitted from parents to children



Linear Mixed Models

Linear Mixed Models (LMM) have emerged as a common statistical method. A LMM approach combines both fixed and random effects using a combination of fixed and random variables. LMM is represented as:

$$y = X\beta + Zu + \varepsilon$$

where y is a known vector of observations,
β is unknown vector of fixed effects,
u is unknown vector of random effects,
and ε is unknown vector of errors,
and X and Z are design matrices relating the observations y to β and u respectively.

10,000,000 SNPs per individual

x

100,000 individuals

+

Complex Matrix multiplications and transformations

=

Computationally complex problem in time and space

Research to enhance LMM for GWAS

Lippert, C. et al proposed to use only a subset of SNPs in the LMM. This approach relies on an estimate of the genetic similarity matrix (GSM), which encodes the pairwise similarity between every two individuals in the dataset. https://www.nature.com/articles/srep06874

Lippert, C et al showed how estimating the GSM from fewer SNPs than individuals leads to computations which are linear in time and memory instead of cubic and quadratic, respectively.

SNP selection approach can a) exclude SNPs that introduce noise, b) include SNPs that tag confounding structure, and c) include causal or tagging SNPs.

FaST-LMM from Microsoft Research

FaST-LMM (Factored Spectrally Transformed Linear Mixed Models) is a set of tools for performing efficient GWAS studies on large genomics data sets.

This approach relies on an estimate of the genetic similarity matrix (GSM), which encodes the pairwise similarity between every two individuals in the dataset.

FaST-LMM from Microsoft Research

The idea behind this approach is that linkage disequilibrium among the SNPs mitigates the need to use all of them.

FaST-LMM takes the generalized LMM models and reduces the complexity from $O(MN^3)$ to $O(MNK)$ for testing M SNPs on K number of similarity matrix. It simplifies the matrix computation of $G.G^T$ for similarity matrices used in LMM.

Experiment: LMM for GWAS

Objective: Understand the implementation of existing LMM algorithms for GWAS studies and then based on my analysis, identify enhancements that I can propose and apply for LMM algorithms for GWAS.

I conducted an experiment by performing GWAS data analysis and predictions using existing Linear Mixed Models approach (I have chosen FaST-LMM) for the prediction of traits on an existing genomic dataset, namely International HapMap project. I have selected a subset of variants on chromosome 22. To measure the accuracy of phenotype predictions, I split individuals in my sample dataset into training and validation groups. The LMM approach will fit covariate and SNP effects based on the individuals in the training set, then generate predictions based on never-before-seen individuals in the validation set.

Hypothesis

LMM algorithms are easy to setup and apply for GWAS studies. To validate this hypothesis, I will analyze how easy it is to setup, perform and analyze LMM-based algorithm in my own computing environment.

LMM algorithms can be applied to different genomics datasets. In this project, I will apply LMM algorithms to both real and synthetic genomics data.

Hypothesis [2]

Computational complexity and computing resources (from both time and computing cost perspective) can be reduced for LMM algorithms by making changes to LMM implementation to perform parallel computations of matrixes/vectors.

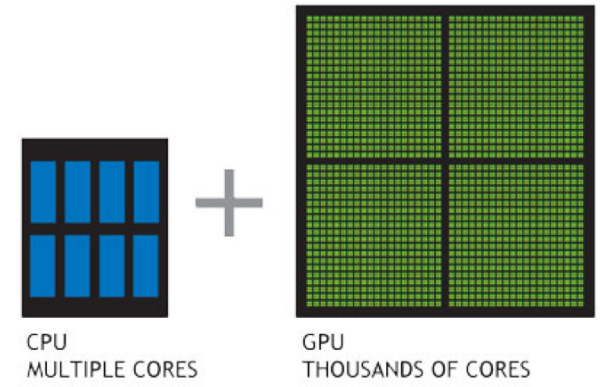
LMM algorithms can be accelerated in terms of time complexity by using GPU-based computing resources instead of general-purpose CPU computing resources.

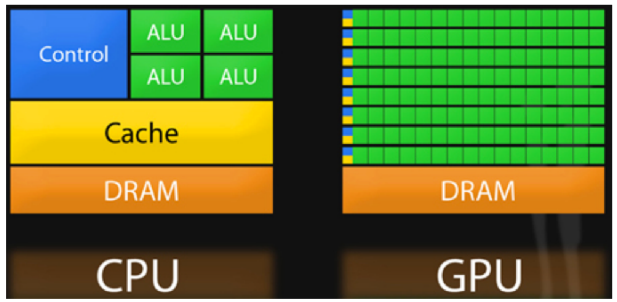
LMM-22 Approach

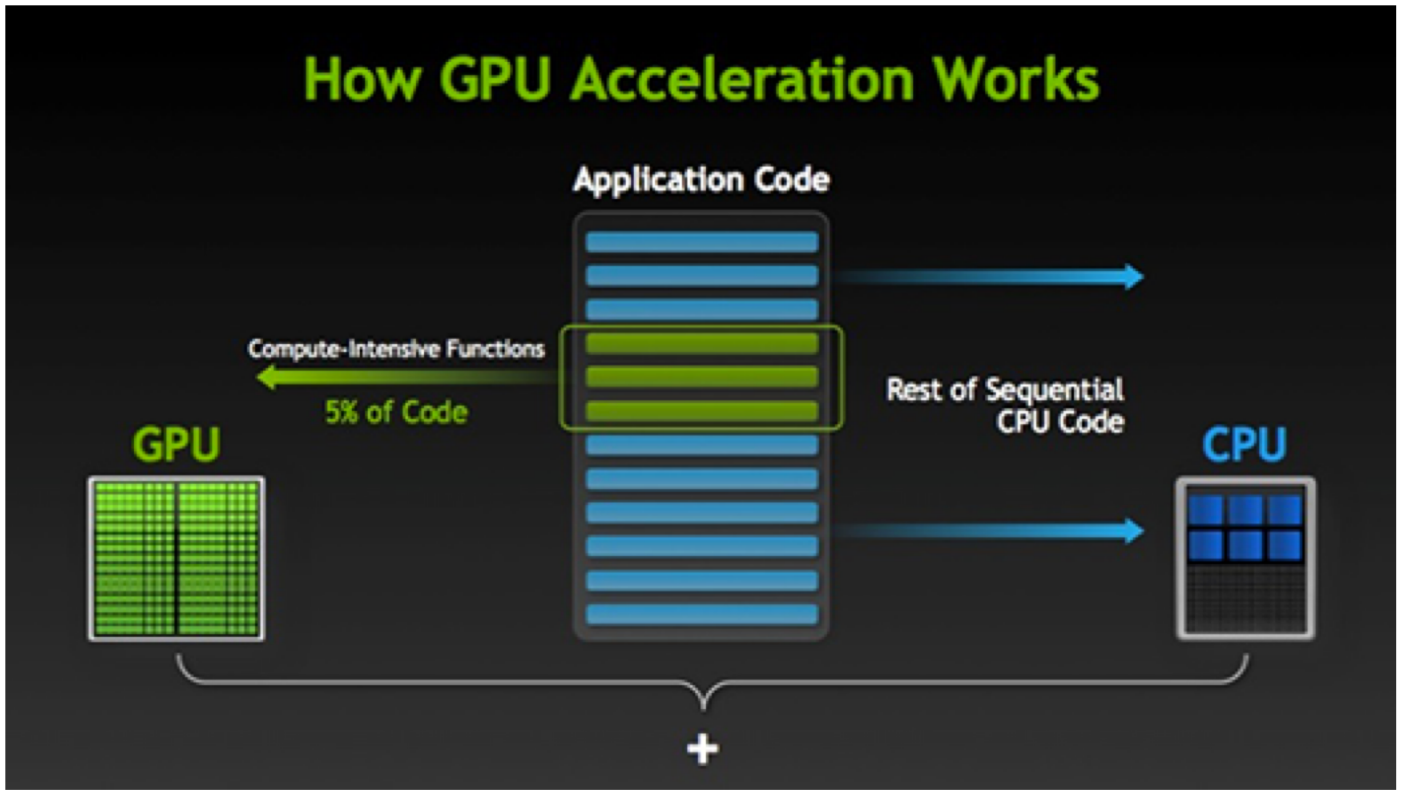
Use of parallelization for LMM execution on GPU clusters.

Python implementation of FaST-LMM that changes the implementation to map matrix computations to stages that can be executed in parallel on GPUs

Parallel Programming using GPU







GPUs are much faster in doing matrix multiplication and running parallel tasks

Running Python Programs on GPUs

Python is one of the most popular programming languages today for science, engineering, data analytics and deep learning applications

Anaconda distribution
https://www.anaconda.com/distribution/

Install the latest version of the Numba package.

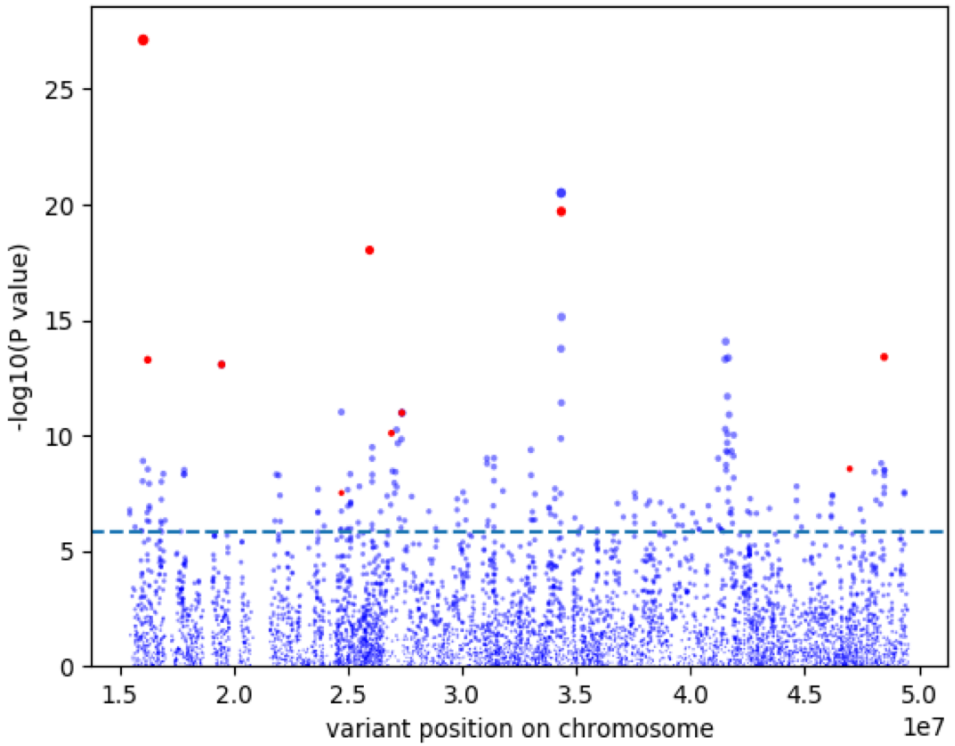
CUDA Python from Nvidia
https://developer.nvidia.com/how-to-cuda-python

Results and Project Status

I have conducted a detailed research of the existing approaches of LMM for GWAS studies. I now understand the limitations and issues with the existing approaches. I was able to run FaST-LMM on a new environment (I chose AWS EC2 compute) from scratch with synthetic phenotype data and genomics data from HapMap project. I have analyzed the python open source implementation of FaST-LMM from Microsoft Research. Presently, I am working on making changes to the implementation to introduce parallelization and testing it on a GPU cluster for performance improvements.

Results of GWAS Study using FaST-LMM

Manhattan plots are typically used to visualize the results of GWAS analysis. The following Manhattan plot shows the scatter plots of *-log p* vs. genomic position for each variant. Causal variants and their neighbors form peaks above a background of unassociated variants.



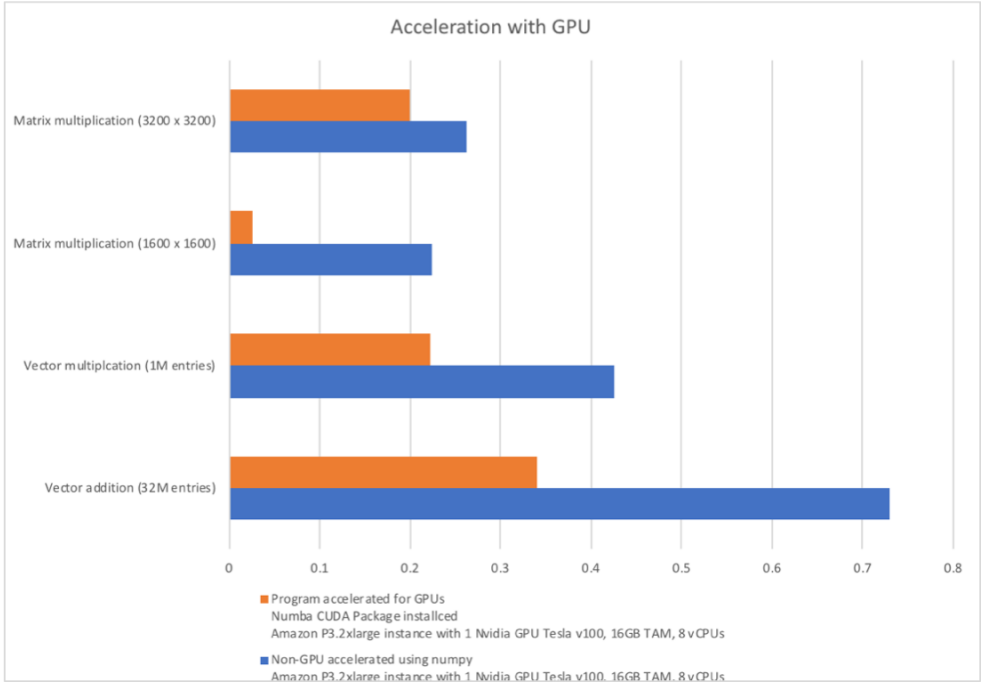
GPU Acceleration for LMM

I compared how these primitive operations can be accelerated by using GPUs. I compared the performance of these primitive operations across two environments on AWS EC2:

1) Non-GPU accelerated computations using *numpy* on Amazon P3.2xlarge instance, and
2) GPU accelerated computations using Numba CUDA package installed on Amazon P3.2xlarge with Nvidia Tesla v100. The following table shows the relative acceleration by using GPU.

Depending on the operation type, GPU acceleration is 24%-89%. This shows that parallelization and multi-threading on GPUs can be used to accelerate LMM computations for GWAS studies.

	Non-GPU accelerated using numpy Amazon P3.2xlarge instance with 1 Nvidia GPU Tesla v100, 16GB TAM, 8 vCPUs	Program accelerated for GPUs Numba CUDA Package installced Amazon P3.2xlarge instance with 1 Nvidia GPU Tesla v100, 16GB TAM, 8 vCPUs	Performance Improvements
Vector addition (32M entries)	0.7302	0.3398	53%
Vector multiplcation (1M entries)	0.4266	0.2229	48%
Matrix multiplication (1600 x 1600)	0.2238	0.025546	89%
Matrix multiplication (3200 x 3200)	0.261642	0.199748	24%
	all times in seconds	all times in seconds	



Learnings and Challenges in my Project

Statistical models and methods behind GWAS studies are more complex than I had originally assumed. Specific I had to understand matrix transformations and deeper statistical concepts such as confounding variables, level-2 regularization methods.

I had to use synthetic data for phenotypes instead of any real data. Also, genomics data I could use for my experiment from HapMap project was limited.

The cost of running LMM models on a GPU compute cluster on AWS was more than I had originally budgeted for the project.

Further Research and Next Steps

1. Try LMM-22 on a massive amount of synthetic and real SNP and phenotype data. Ideally, such data should be for ~10,000 individuals with segmented set of SNPs = ~100,000. Such data should also account for confounding variables such as family relatedness and population structure

2. Take massive amount of synthetic and real SNP and phenotype data, and then perform LMM-based GWAS study across two environments: a) cluster of CPU-based computing server instances on AWS cloud, and b) cluster of GPU-based computing server instances. Then I need to compare the relative speed-up of computing on b) as compared to a) for similar LMM analysis and data set.

References

- Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–5 (2011).
- Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–8 (2006).
- Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep.* **3**, 1815; 10.1038/srep01815 (2013).
- Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb).* **91**, 47–60 (2009).