Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

# 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

* Siddharth Sharma, Intern at Stanford University Department of Biomedical Data Science, siddrrsh@gmail.com

***Abstract-*** Increasingly, genomics is being used for the prediction of specific traits and diseases (phenotypes) among humans. Wider availability of genomics data through multiple research projects (such as International HapMap Project[1] and 1000 Genomes[2]) has been a catalyst in that direction. With the recent advances in machine learning and big data analysis, data computation resources and data models needed for genomics data analysis are readily available. However, the prediction of traits and diseases has its own challenges in terms of computational requirements and computational analysis, statistical analysis (example: confounding variables), and limited quality of data collection. Linear Mixed Models (LMM, a type of linear regression) is a common approach for Genome-wide Association Studies (GWAS) for the prediction of common traits among humans using genomics. This paper researches the existing LMM-based approaches for Genome-wide Association Studies (GWAS), describes the experiment performed on FaST-LMM approach from Microsoft Research, and then proposes an enhanced approach (called LMM-22) on how to address computational and statistical issues. LMM-22 focuses on the parallelization of LMM computations and execution of LMM-22 on General Purpose Graphics Processing Units (GPU) as against CPUs to accelerate the LMM approach for GWAS studies.

***Index Terms-*** *Genomics, GWAS, phenotypes, SNP, LMM, Linear Mixed Models, GPU*

## Introduction

Genome-wide associations studies (GWAS) are emerging as commonly used method for scientists and medical researchers to identify genes involved in human diseases. This method searches the genomes for variations in single nucleotide polymorphisms or 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 certain disease. Linear Mixed Models (LMMs) has emerged as a common statistical and data science approach for performing GWAS studies.

## Background

### Gene

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA and act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people.

An **allele** is one of two or more versions of a gene. An individual inherits two alleles for each gene, one from each parent. If the two alleles are the same, the individual is homozygous for that gene. If the alleles are different, the individual is heterozygous.

### Genome and Nucleotide

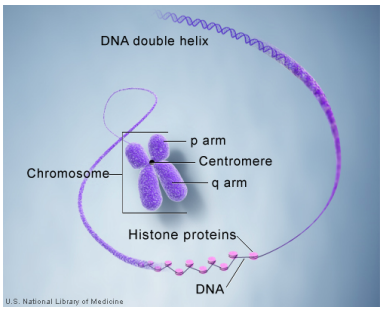
Genome is the genetic material of an organism. A genome is an organism's complete set of Deoxyribonucleic Acid (DNA), including all of its genes. Each genome contains all of the information needed to build and maintain that organism. In humans, a copy of the entire genome—more than 3 billion DNA base pairs—is contained in all cells that have a nucleus.

Nucleotides are organic molecules that form the DNA and RNA. A genome sequence is the complete list of the nucleotides (A, C, G, and T for DNA genomes) that make up all the chromosomes of an individual or a species. Within a species, the vast majority of nucleotides are identical

[1] https://ghr.nlm.nih.gov/primer/genomicresearch/hapmap
[2] 1000 Genomes Project: http://www.internationalgenome.org/

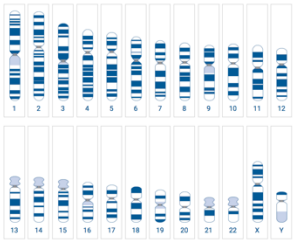
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

between individuals, but sequencing multiple individuals is necessary to understand the genetic differences between humans.



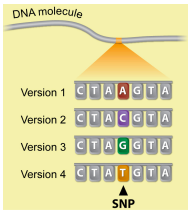

## Chromosomes

In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure. In humans, each cell normally contains 23 pairs of chromosomes[3], for a total of 46.



## Single Nucleotide Polymorphisms (SNP)

Single nucleotide polymorphisms, called SNPs (and pronounced as "snips"), are the most common type of genetic variation among humans. Each SNP represents a difference in a single DNA building block, called a nucleotide. 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. Most commonly, these variations are found in the DNA between genes. SNPs can act as biological markers, helping scientists locate genes that are associated with disease.




When SNPs occur within a gene or in a regulatory region near a gene, they can play a more direct role in disease by affecting the gene's function. Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual's response to certain drugs, and risk of developing particular diseases.

## Genotypes and Phenotypes

Genotype is an organism's full hereditary information expressed in terms of genomes. Genotype also refers to set of genes carried by an individual.

Phenotype is the observable physical or biochemical characteristics of an individual organism, determined by both genetic structure and environmental influences. Examples of phenotypes include height, eye color, IQ, genetic diseases (Prostate or colorectal cancer, breast cancer, Type-2 diabetes) etc. Most phenotypes are influenced by both one's genotype and by the unique environment that one has lived in. The genes contribute to a trait, and the phenotype is the observable expression of the genes (and therefore the genotype that affects the trait). The relationship between genotype and phenotype is expressed as follows:

```
genotype (G) + environment (E) + genotype &
environment interactions (GE) → phenotype (P)
```

An example of a phenotype is eye color, which is an inherited trait influenced by more than one gene, including OCA2 and HERC2. The interaction of multiple genes—and the variation in these genes between individuals—help to determine a person's eye color.

## 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 typically focus on associations between (SNPs) and traits such as major human diseases but can equally be applied to any other organism. When applied to

[3] Information about each chromosome and health implications of its genetic changes: https://ghr.nlm.nih.gov/chromosome

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

human data, 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.

For a GWAS study[4], researchers use two groups of participants: people with the disease being studied and similar people without the disease. Researchers obtain DNA from each participant. Each person's complete set of DNA or genome, is then purified from the blood or cells, placed on tiny chips and scanned on automated laboratory machines. The machines quickly survey each participant's genome for selected markers of genetic variation, which are SNPs. If certain genetic variations are found to be significantly more frequent in people with the disease compared to people without disease, the variations are said to be "associated" with the disease. The associated genetic variations can serve as pointers to the region of the human genome where the disease-causing problem resides.

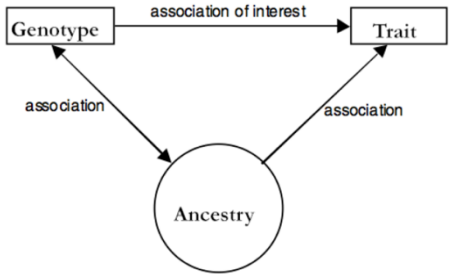
**International HapMap Project**

This project is a scientific effort to identify common genetic variations among people. The HapMap (short for "haplotype map") is a catalog of common genetic variants SNPs. Each SNP represents a difference in a single DNA building block, called a nucleotide. When several SNPs cluster together on a chromosome, they are inherited as a block known as a haplotype. The HapMap describes haplotypes, including their locations in the genome and how common they are in different populations throughout the world.

## Existing Approaches 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.

Statistical methods are becoming more common and widely adopted approach for GWAS studies. However, there are additional issues in the use of statistical methods for GWAS studies. Any observation in GWAS studies can be confounded by population structures, which are presence of subgroups in population with ancestry differences. In statistics, a confounding variable affects both dependent and independent variables causing spurious associations. See the diagram below[5] for an example of a confounding variable related to subgroups with ancestry differences. Ethnic groups often share similar dietary habits and lifestyle characteristics that lead to environmental factors that affect traits. For example, South Indians eat more rice than North Indians. Ignoring such ancestry differences among sample individuals can lead to false positives or incorrect associations. Furthermore, family relatedness (example: alleles transmitted from parents to children) can also cause confounding problems.

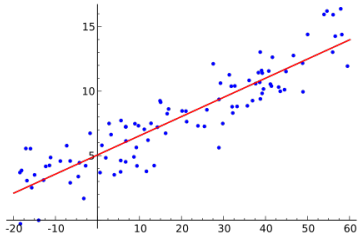



Initially, **Linear Regression Models** were used for GWAS studies. In linear regression, statistical analysis is done to model relationship between dependent variable and one or more independent variables. Linear regression models can be used to either a) fit a predictive model to an observed data set of y and X values, and the predict value of y, or b) quantify the strength of relationship between y and X. Linear regression is written in the vector form as:

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

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

The following diagram shows an example of linear regression of random data points.




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

[4] Detailed explanation of GWAS is here: https://www.genome.gov/20019523/

[5] http://faculty.washington.edu/tathornt/SISG2015/lectures/Taipei2015/Taipei2015session06.pdf

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

sophisticated 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,
$\beta$ is unknown vector of fixed effects,
$u$ is unknown vector of random effects,
and $\varepsilon$ is unknown vector of errors,
and X and Z are design matrices relating the observations y to $\beta$ and $u$ respectively.

Linear Mixed Models (LMMs) have emerged as a common approach for identifying causal features and predicting phenotypes. As with a standard linear model, LMMs include fixed effects for each genomic feature and any recorded covariates (also called feature vectors), such as age or sex. LMMs also include random effects: in the context of genomic models, these random effects are correlated between individuals on the basis of their genetic similarity. These random effects can account for heritable differences in phenotype that are not reflected by genomic features or covariates.

LMM approaches for GWAS have been refined in the research community over years to address issues related to confounding variables and computational complexity. Mathematical computation (for example: matrix and vector multiplications, variable computation) for massive amount of data for SNPs (10M SNPs per individuals and 10s of thousands of individuals) are both complex in time and memory and are expensive in terms of computational capacity needs. For example, 100s of computers may be needed over multiple weeks to perform such computations.

Yu et al [1] researched the use of unified mixed-model method for association mapping that accounts for multiple levels of relatedness. To reduce computational complexity and number of SNPs to be processed, Lippert, C. et al[6] [2] 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. Lippert, C et al [2] also 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. In a related approach [2], SNPs are chosen such that they are roughly equally spaced across the genome. The idea behind this approach is that linkage disequilibrium (non-random association of alleles at different loci in a given population) among the SNPs mitigates the need to use all of them. When the number of selected SNPs is less than the sample size of the data, then the computation of $P$ values becomes linear in sample size, rather than quadratic. This further reduces computational complexity.

Another approach [3] [4] for reducing computational complexity is based on a mathematical equivalence between the LMM and linear regression. An LMM is equivalent to a form of linear regression in which the SNPs that determine the GSM in the LMM view are covariates in the linear-regression view. The linear-regression view suggests including in the GSM only those covariates that are correlated to the phenotype. This form of subsetting is referred to as SNP selection, and can a) exclude SNPs that introduce noise, b) include SNPs that tag confounding structure, and c) include causal or tagging SNPs.

Based on the above research, researchers at Microsoft have developed an open source algorithm called FaST-LMM[7]. FaST-LMM (Factored Spectrally Transformed Linear Mixed Models) is a set of tools for performing efficient GWAS studies on large genomics data sets.

## Cloud Computing Platforms

A parallel trend has been the easy availability of scalable compute resources and capacity through Cloud Computing platforms such as Amazon Web Services (AWS)[8] and Microsoft Azure. Research scientists can easily get on-demand (there is no need to first buy expensive hardware resources) access to large number of server resources for running computations for GWAS studies. They can then scale this capacity up or down based on the computing needs of algorithm and datasets. Next, researchers have to only pay for compute resources actually used. Cloud Computing platforms have changes the cost economics of GWAS studies and made it easier and less costly for research community.

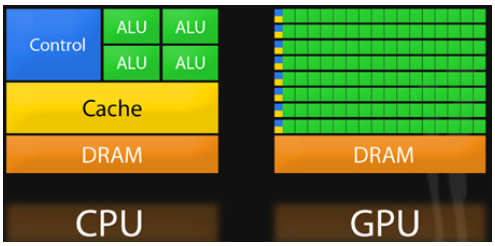
## Graphics Processing Units (GPU)

Given LMM algorithms for GWAS studies need to perform computation on large blocks of SNPs data and matrices/vectors, GPUs have emerged as a more common

[6] https://www.nature.com/articles/srep06874
[7] Fast-LMM from Microsoft Research https://www.microsoft.com/en-us/research/project/fast-lmm-software-papers/
[8] Amazon Web Services: https://aws.amazon.com

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

approach for parallel data processing than general purpose CPUs. Initially, GPUs were designed for graphics and image processing. However, the ability of GPUs to do parallel processing of data (specific those involving vectors and matrices) has made them ideal choice for parallel computation of massive amount of data for applications such as machine learning, high performance computing, genomics etc.




Cloud Computing platforms offer general-purpose GPU compute instances in an on-demand and scalable basis. For example, GPU compute instances from AWS, Amazon EC2 P3 instances[9], offer up to 8 NVIDIA Volta GV100 GPUs.

## Experiment Setup

The objective of my experiment is to first 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 chose FaST-LMM algorithm from Microsoft Research for my experiment for the prediction of traits on an existing genomic dataset, namely International HapMap project. The reasons from my choice of FaST-LMM was its easy availability as an open source implementation and good documentation.

The hypotheses of my experiment are as follows:

1) *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. Hence, I chose an environment (Linux server on AWS Cloud platform) that is different from that outlined (Microsoft Azure Cloud) in the FaST-LMM documentation.
2) *LMM algorithms can be applied to different genomics datasets.* In this project, I will apply LMM algorithms to both real and synthetic genomics data.
3) *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.
4) *LMM algorithms can be accelerated in terms of time complexity by using GPU-based computing resources instead of general-purpose CPU computing resources.* This is because of a better ability of GPUs (than CPUs) to do parallel processing of data, specific those involving vectors and matrices.

I started with the existing FaST-LMM algorithm and its open source implementation to understand the existing LMM approaches for GWAS studies. *Note that I mostly followed the steps of FaST-LMM as-is to build an understanding of the current state of LMM algorithms.* I used a simulated phenotype data set. First, I analyzed the open source implementation of FaST-LMM on github[10] and tried to understand the logic of its implementation. Next, I did the following steps as outlined in the documentation of the FaST-LMM on github project. I performed these steps on AWS cloud platform instead of Azure. [see appendix A for commands that I used]:

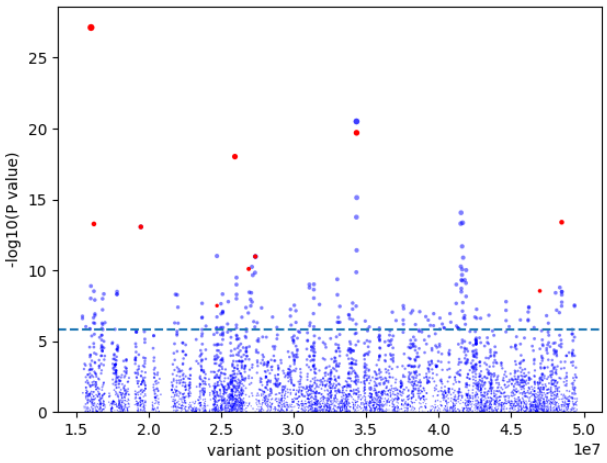
1) I setup a Linux EC2 server on AWS and then did remote ssh to that Linux server.
2) I setup python 2.7 environment.
3) Next, I installed `numpy`, `pysnptools` and `fastlmm` packages for python.
4) FaST-LMM uses four input files containing (1) the SNP data to be tested, (2) the SNP data used to determine the genetic similarity matrix (GSM) between individuals, (3) the phenotype data, and (4) a set of covariates.
5) I downloaded genotype data from International HapMap project[11] to simulate phenotypes and perform GWAS analysis. The HapMap3 dataset, available in PLINK format from the HapMap Project website, contains genotypes from 1,184 persons.
6) I used `PLINK` to select relatively common variants (those whose minor allele has frequency >5% in this dataset) on chromosome 22.
7) I used the `single_snp()` function to perform single-variant association testing using LMM.

Manhattan plots are typically used to visualize the results of GWAS analysis. The following Manhattan plot shows the

[9] AWS P3 GPU Instances: https://aws.amazon.com/ec2/details/
[10] FaST-LMM open source project on github: https://github.com/Azure/Cortana-Intelligence-Gallery-Content/tree/master/Resources/Phenotype-Prediction
[11] International HapMap project: https://hapmap.ncbi.nlm.nih.gov/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

scatter plots of -*log p* vs. genomic position for each variant. Causal variants and their neighbors form peaks above a background of unassociated variants.




In this example, the p-values for all true causal variants pass the threshold for significance: their p-values are sufficiently low that they lie above the threshold line on the -\|*log p* axis. The plot also reveals two common types of "false positives", i.e. variants which are not causal that nonetheless have significant p-values.

Next, to assess the accuracy of phenotype predictions as per FaST-LMM documentation, I split the 1,184 persons in my dataset into training and validation groups. FaST-LMM fits covariate and SNP effects based on the individuals in the training set, then generates predictions based on never-before-seen individuals in the validation set. Then, I trained a FaST-LMM model using the randomly-chosen training set. Note that FaST-LMM extracts the phenotype and covariate information for training set individuals from the provided text files, which contain information on all individuals.

Finally, I performed predictions on the validation set and compared these to the true simulated phenotypes. I used the coefficient of determination ($R^2$), i.e. the proportion of variance in phenotype accounted for by the predictions, as my metric for accuracy.

## LMM-22 Approach

Based on my analysis of various LMM approaches, the key challenges in applying LMM to GWAS studies are related to 1) spurious associations and false positive caused by the presence of confounding variables (such as population structure and family relatedness) in the genomics data set, and 2) computational complexity of LMM in terms of time and computing resources needed to perform computation on very large datasets. These LMM computations involve operations on matrixes and vectors on a massive amount of data, for example 10M SNPs (which can be reduced further by subsetting and selection approaches) per individuals and 10s of thousands of individuals.

While the first problem of confounding variables has been targeted by multiple researchers (as I explained in the earlier sections), the second problem of reducing computational time complexity is still an open research area. My proposed approach (called LMM-22 as humans have 22 pairs of autosomal chromosomes and one pair of sex chromosomes) looks at how statistical and matrix computations for LMM algorithms can be parallelized and then executed on GPU clusters to reduce both time and computational needs.

I analyzed the codebase of FaST-LMM and associated papers. 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.

My LMM-22 approach is focused on the use of parallelization of LMM execution on GPU clusters. The general idea is to take python implementation of FaST-LMM and then change the implementation to map the computation of matrix computation to stages that can be executed in parallel on GPUs.

## Results

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 LMM approaches related to confounding variables and computational complexities.

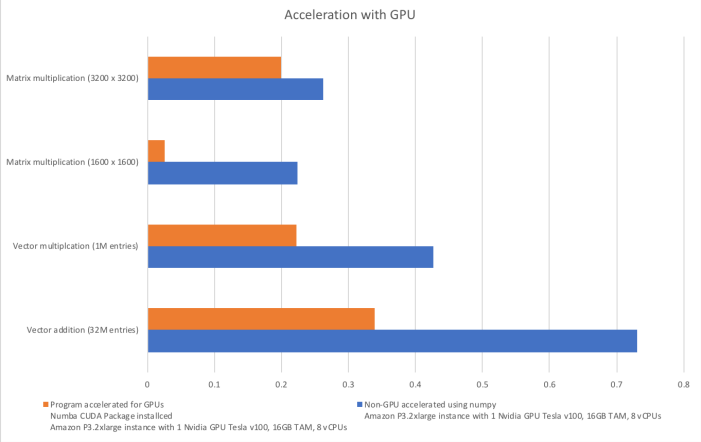
I was able to run FaST-LMM on a new environment (I chose AWS EC2 compute) from scratch with synthetic phenotype data and genomics data from HapMap project. I have analyzed the python open source implementation of FaST-LMM from Microsoft Research. Presently, I am working on making changes to the implementation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

I analyzed the primitive operations (vector addition and multiplications, matrix multiplication) that are used in LMM algorithm. Next, I compared how these primitive operations can be accelerated by using GPUs. I compared the performance of these primitive operations across two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0154.v1

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[12] 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 | |




## Conclusion

Linear Mixed Models and its implementations address the issues related to confounding variables in GWAS studies. While approaches such as FaST-LMM have reduced the computational complexity by using SNP subsetting and Similarity matrix, the time complexity of Linear Mixed Models (LMM) approach to GWAS can be further reduced by use of parallelization and GPUs for accelerating vector/matrix multiplication and statistical computations. GPUs accelerate matrix operations by 24-89%.

## Learning and Challenges

The 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

One of the stated objectives and hypothesis of my project is to try existing and proposed LMM algorithm 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, as I discussed in the background. I couldn't get to this part of my project during the current time duration. I plan to carry such analysis as part of the next steps for my research.

Another hypothesis for my project is that LMM algorithms can be accelerated in terms of time by using GPU-based computing resources instead of general-purpose CPU computing resources. To validate this hypothesis, I plan to 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. Given the $ cost associated with provisioning CPU and GPU servers on AWS cloud platform, I couldn't perform this experiment as part of my project given the limited $ budget I had allocated to my project.

## Acknowledgment

I will like to acknowledge and thanks Mr. Kevyn Adams, my project mentor. I would also like to thank the research team at Microsoft who developed FaST-LMM for the contributions to the field of genomics.

## References

[1] Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–5 (2011).

[2] 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).

[12] Nvidia NUMBA CUDA package for GPU acceleration
https://developer.nvidia.com/how-to-cuda-python

[3] 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).

[4] 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).

## Authors

**Siddharth Sharma** – Researcher and Intern at Stanford University Department of Biomedical Data Science
.