GpuGWAS

CUDA based Genome Wide Association Studies

- $1,000 Genome
- GWAS Statistical Methodology
- Implementation Description

tbi@gmu.edu
tbi1@jhmi.edu
703-200-6533

Thousand dollar Genome

- Rapid reduction in sequencing costs
- We are approaching a price point where DNA sequencing can be widely adopted

- Cost of sequencing the human genome:

Year	Cost
2003	$3 Billion
2007	$1 Million
2008	$60,000
2010	$19,500
2012	$4,000

Sequencing Innovation

Industrial sequencers

Portable desktop sequencers were introduced in 2011

MinION sequencer to be released in 2012

Clinical Sequencing

NATURE | NEWS FEATURE

Rare diseases: Genomics, plain and simple

A Pennsylvania clinic working with Amish and Mennonite communities could be a model for personalized medicine.

Trisha Gura

29 February 2012

Clinical Sequencing

- Clinic for Special Children in Pennsylvania
- Via Genetic data, reduced the medical costs of a bone marrow transplant for a new born baby
  - From $500,000 to $12,000 dollars
  - Early timing: genetic diagnosis before the patient is 24 hours old
  - Template for personalized medicine in the future

“Although genome sequencing is creeping into clinical care around the world, it has yet to become an integral part of everyday medical practice. ‘We've talked about **the thousand-dollar genome and the million-dollar interpretation**,’ says Eric Topol, a genomicist at the Scripps Research Institute in La Jolla, California.”

Genome Wide Association Studies

- A revolutionary new way for scientists to identify genes that influence health and disease
- Abreviation: GWAS
- Scans the SNP markers across the entire human genome seeking the genetic basis of diseases
- Genes discovered from GWAS provides the foundation for personalized medicine

Genome Wide Association Studies

- First studies in 2005 consisted of 100,000 SNPs derived from 150 patients

- Current studies analyzes 1 million SNPs from cohorts of 64,000 patients

Statistical Methodology

- GWAS via Logistic Regression is based on the following univariate model:

$$\text{logit}(\pi_i) \sim \beta_0 + \beta_1 \chi_i$$

π_i	: expected value of trait given genotype
β_0	: overall mean
β_1	: association parameter
χ_i	: genotype
H_0	: β_1 is 0

Statistical Methodology

$$\log p_{g_i}(x;\beta) = y_i \log p(x;\beta) + (1-y_i)\log(1-p(x;\beta))$$

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \log p_{g_i}(x_i;\beta) \\ &= \sum_{i=1}^{N} [y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))] \end{aligned}$$

$$\begin{aligned} p(x;\beta) &= Pr(G=1 \mid X=x) = \frac{\exp(\beta^T x)}{1+\exp(\beta^T x)} \\ 1-p(x;\beta) &= Pr(G=2 \mid X=x) = \frac{1}{1+\exp(\beta^T x)} \end{aligned}$$

$$l(\beta) = \sum_{i=1}^{N} \left[y_i \beta^T x_i - \log(1+e^{\beta^T x_i}) \right]$$

$$\begin{aligned} \frac{\partial l(\beta)}{\beta_{1j}} &= \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \frac{x_{ij} e^{\beta^T x_i}}{1+e^{\beta^T x_i}} \\ &= \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} p(x;\beta) x_{ij} \\ &= \sum_{i=1}^{N} x_{ij}(y_i - p(x_i;\beta)) \end{aligned}$$

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta))$$

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1-p(x_i;\beta))$$

$$\begin{aligned} &\frac{\partial l(\beta)}{\partial \beta_{1j} \partial \beta_{1n}} \\ &= -\sum_{i=1}^{N} \frac{(1+e^{\beta^T x_i}) e^{\beta^T x_i} x_{ij} x_{in} - (e^{\beta^T x_i})^2 x_{ij} x_{in}}{(1+e^{\beta^T x_i})^2} \\ &= -\sum_{i=1}^{N} x_{ij} x_{in} p(x_i;\beta) - x_{ij} x_{in} p(x_i;\beta)^2 \\ &= -\sum_{i=1}^{N} x_{ij} x_{in} p(x_i;\beta)(1-p(x_i;\beta)) \, . \end{aligned}$$

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta}$$

Computational Challenge

- Implementing GWAS requires a large amount of computational resources

- Exponentially dependent on both the amount of SNPs sequenced and the patient cohort size

Covariates

- Covariates are clinical or demographic factors that can have an confounding impact on the GWAS analysis

- An example would be a GWAS conducted on lung cancer would be more accurate by the inclusion of the covariate for smoking-behavior

Covariates

- The addition of covariates can provide many benefits to a GWAS analysis at the expense of increasing computational cost:

$$\text{logit}(\pi_i) \sim \beta_0 + \beta_1 \chi_i + \beta_2 z_i$$

Computing GWAS

- The analysis for a large GWAS study is normally implemented on a computational cluster with a job queuing system to partition the GWAS analysis to the different servers

Computational Clusters

- Departments from Schools of Public Health typically supports a large cluster for genomic analysis

- Enigma (Johns Hopkins University)
  - 46 nodes, ~400 CPU cores

- Biostatistics Cluster (University of Michigan)
  - 23 nodes, ~300 CPU cores

Thousand-Dollar Genome, Million-Dollar Interpretation

- Different approaches to the data deluge challenge

- Advantages of Nvidia GPUs
  - GWAS is a data parallel computation
  - New GPU generations provides a data parallel improvement

GpuGWAS

- We utilized Nvidia GPUs to achieve a 10x improvement in implementing a GWAS analysis

- Components
  - PyCUDA, GenABEL, CUDA C

PyCUDA

- Open-source Python Toolkit
- Enables GPU programming with Python
- Directly access the CUDA computation API
- Convenient and full featured

GenABEL

- R library for GWAS analysis
- Compatible with the current GWAS data formats
- Contains functions for all aspects of GWAS (QC, GC, IBS, etc.)
- Integrates with PyCUDA via the RPy2 library

Dataset

- Phase III data of the International HapMap Project
- Central European subset
- 165 individuals of similar genetic ancestry
- 1.4 million SNPs

GpuGWAS Structure

- The program reads the GWAS inputs from the HapMap dataset
- Performs the GWAS logistic regression via a PyCUDA implmentation
  - Challenging process to parallelize
- Encapsulates the results as a GenABEL data object

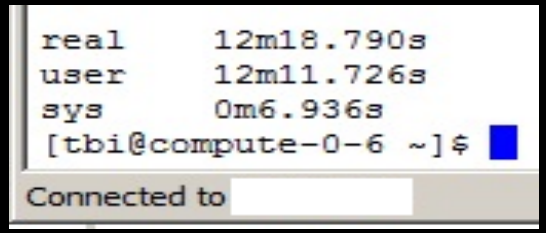
Compatibility

- Originally implemented on Ubuntu Linux 11.04

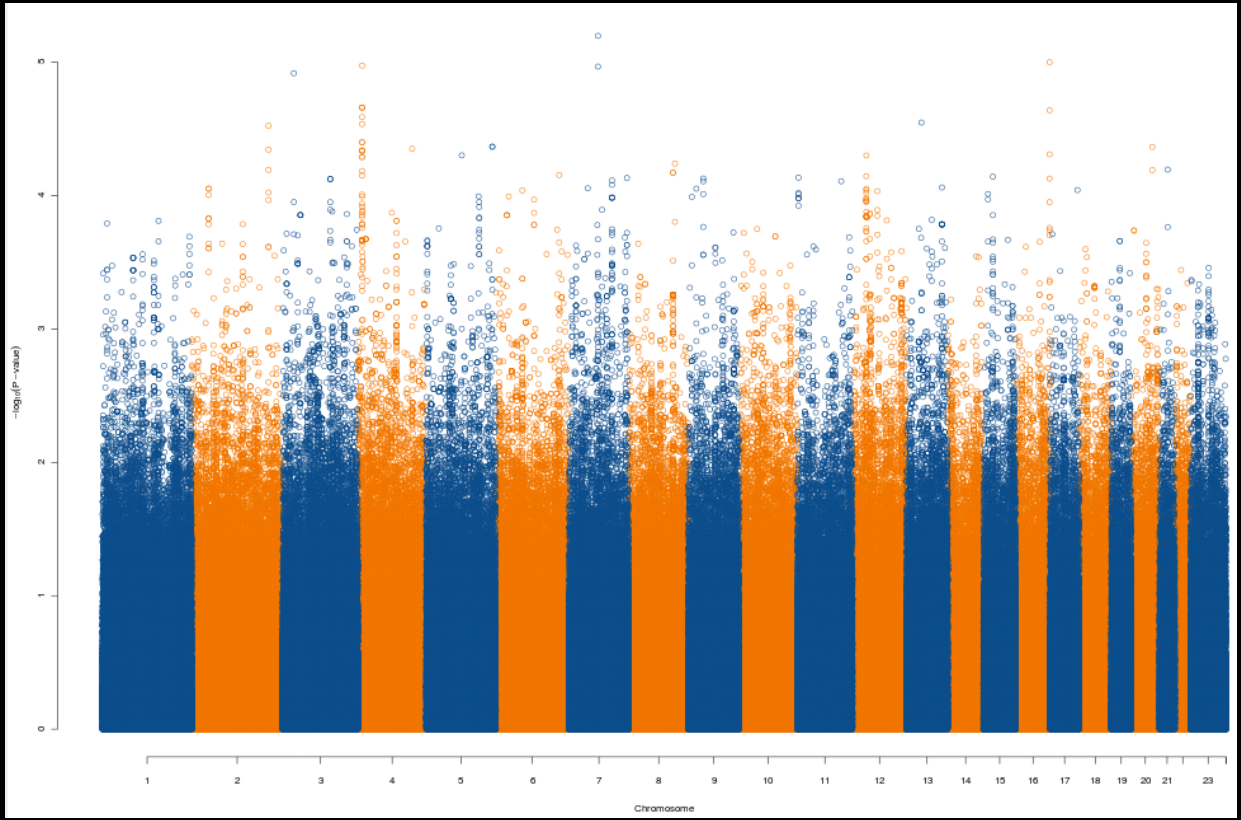
- Compatible across all platforms
  - Microsoft Windows
  - Apple OS X
  - Linux

CPU Results

- CPU analysis computed on a Dell PowerEdge server with dual Xeon E5420 quad-core processors
  - completed in 12m18.79s



CPU Manhattan Plot

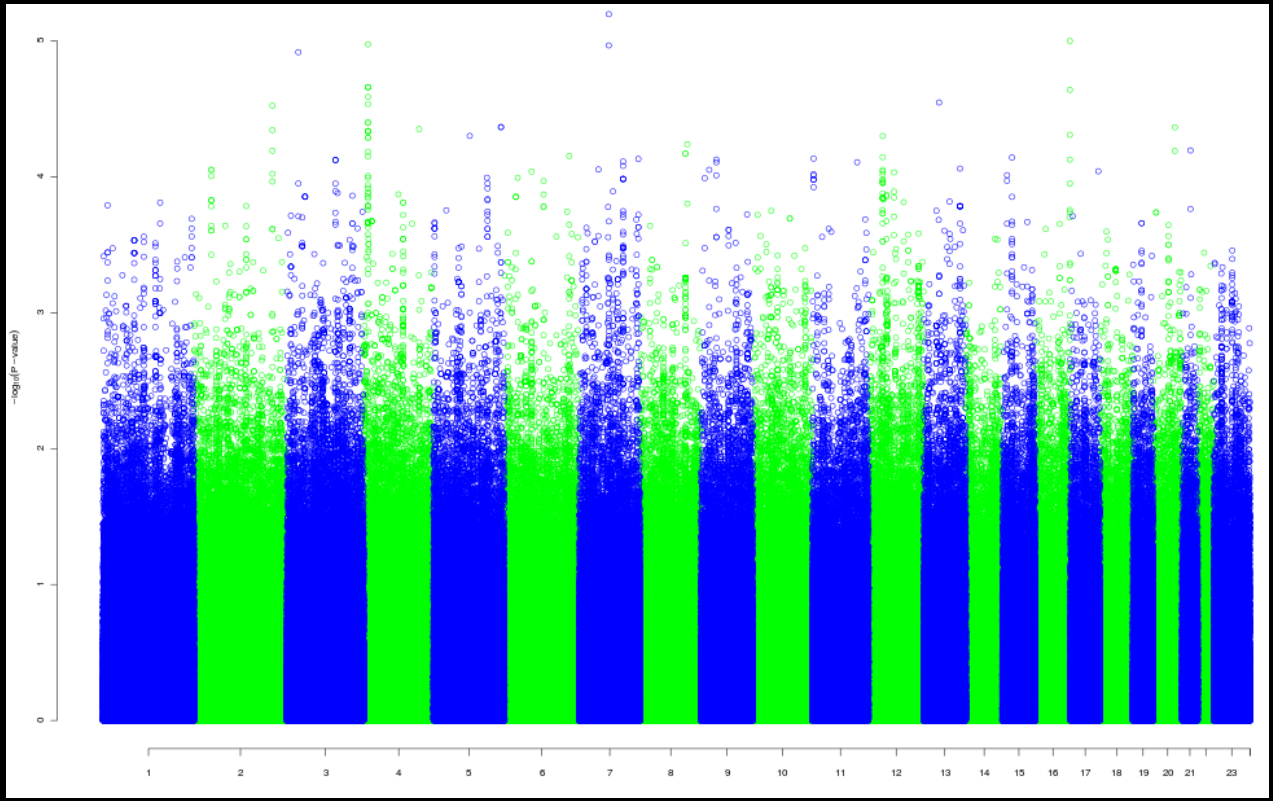


GPU Results

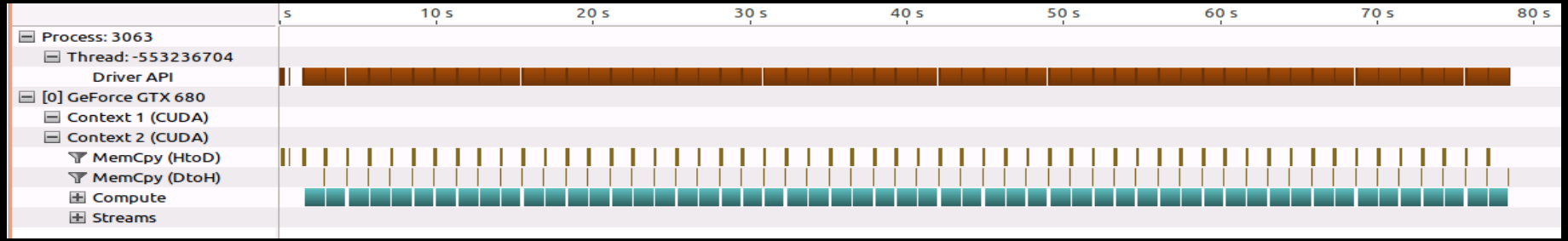
- The results from the GPU analysis matched the results of the CPU analysis
  - completed in 1m18.90s

```
real    1m18.898s
user    1m17.300s
sys     0m1.490s
```

GPU Manhattan Plot



GPU Optimization



- Currently in the process of improving GpuGWAS

- GpuGWAS analyzed with the Nvidia Visual Profiler

GPU Optimization

- Identified the computation of matrix dot products as a major point for improvement

- Aiming to improve performance by incorporating the PyCUDA sparse library, the CULA library wrapper, and improving the memory transfer

Multi-GPU Implementation

- In the process of integrating the GpuGWAS algorithm to run on a GPU cluster

- Since there are no communication between the GPU instances, each additional GPU translates to 100% parallel efficiency

Tip of the iceberg in performance

- Achieved a 10x speedup for a cohort size of 165 patients
- CPU programs scale poorly with matrix size

TABLE 1 Run times required for a 60-ms simulation such as that presented in Fig. 2 using both Monte Carlo and moment-closure approaches

	Δt (μs)	N	Time (min)
Monte Carlo	0.01	100	50
	0.01	1000	794
	0.01	10,000	8755
Moment closure	0.01	—	95
	1	—	0.9

- GPU programs scale better
- GPU analysis could potentially enable larger-scaled and more accurate GWAS implementations

Experimental dataset

- Started collaborating with Dr. Dhananjay Vaidya to analyze an experimental GWAS dataset
- Consists of 2.4 million SNPs and 2,000 individuals
- Looking forward to a substantial improvement of the program performance
- Opportunity to validate the GPU results in a scientific GWAS study

Benefits of GpuGWAS

- Extendable to GPU clusters
- Integrated with an established GWAS library
- Potential speedup for larger datasets

Goals

- Encapsulate GpuGWAS as a R-package available via the CRAN or Bioconductor repositories
- Implement a large scale analysis on a cluster of GPU servers
- Extend GpuGWAS to other widely used aspects of genetic analysis

Clinical Analysis

- Possible utilization of the GPUs for clinical genomic applications
  - a single GPU can substitute for a number of servers
  - scalable from small clinics to large hospitals

- Enable a distributed system of genomic analysis were each clinic is empowered to analyze genomic data locally

Thank you

- George Mason University
  - Dr. Jeffery Solka
  - Dr. M. Saleet Jafri
- Johns Hopkins University / GeneSTAR
  - Dr. Dhananjay Vaidya
  - Dr. Diane Becker