# A High-Speed Algorithm for Genome-Wide Association Studies on Multi-GPU

Yen-Chen Chen, Su-Yun Huang (Advisor), Shih-Jen Tsai (Advisor), Albert C. Yang (Advisor), Weichung Wang (Advisor)

National Taiwan University, Academic Sinica, Taipei Veterans General Hospital, Harvard Medical School

## Genome-Wide Association Studies

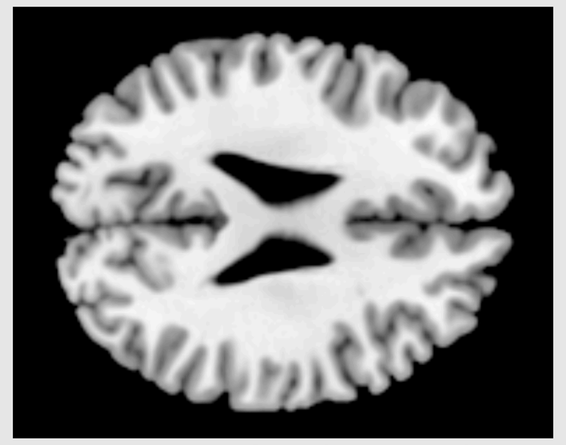


Figure 1: Brain fMRI example image

### GWAS

Genome-wide association studies (GWAS) have been a powerful tool for investigating the genetic architecture of human disease.

### SNP

Single nucleotide polymorphisms (SNPs) is a kind of commonly used genetic marker.

### Brain parameters

Although GWAS has been used for genetic researches for decades, how genetic markers are associated with brain function remain largely unknown.

## Difficulties

### Heavy computation

GWAS with tens of thousands of brain voxels involves intensive computation that is not feasible in conventional settings.

### Accuracy

GWAS study requests high accuracy for p value. And there is no robust rule for how small p value should be to count as important factor.

### Numerical integration

To calculate p value, we will have to deal with the unbounded numerical integration which is some heavy computation.

### Why compute p value

We decide to compute p value than taking a t value threshold because of these reasons.

1. Since we have 36 billion sets of t-tests, we have no robust theroy for what t value to choose.
2. There's no existing t table for extreme small p value with large degrees of freedom.
3. Due to the loss of SNP data for some patients, the degrees of freedom is not a fixed value.

## Objectives

### Problem

We want to know how all 653291 SNPs affect 55749 voxels of gray matter brain function parameters.

### Method

For each SNP to brain voxel pair, we will first solve a linear regression:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

and calculate a p-value for each of the t-tests:

$$t = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{ii}}}$$

$$p = \int_{-\infty}^{t} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \times \frac{1}{\sqrt{\nu\pi}} \times \frac{dx}{(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}}$$

$X$: Data matrix, $Y$: Vector of brain responses, $\sigma$: Standard deviation, $\nu$: Degrees of freedom, $\Gamma$: Gamma function.

### Parallelization

All linear systems are independent and thus perfect for parallelization. All we have to do is to improve the time consuming p value calculation.

## Reformation

In order to implement p value calculation in GPU, we reformed the integration.

$$p = C\int_{-\infty}^{t} \frac{dx}{(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}} = C\int_{-\frac{\pi}{2}}^{\arctan(\frac{t}{\sqrt{\nu}})} \cos^{\nu-1}\theta d\theta$$

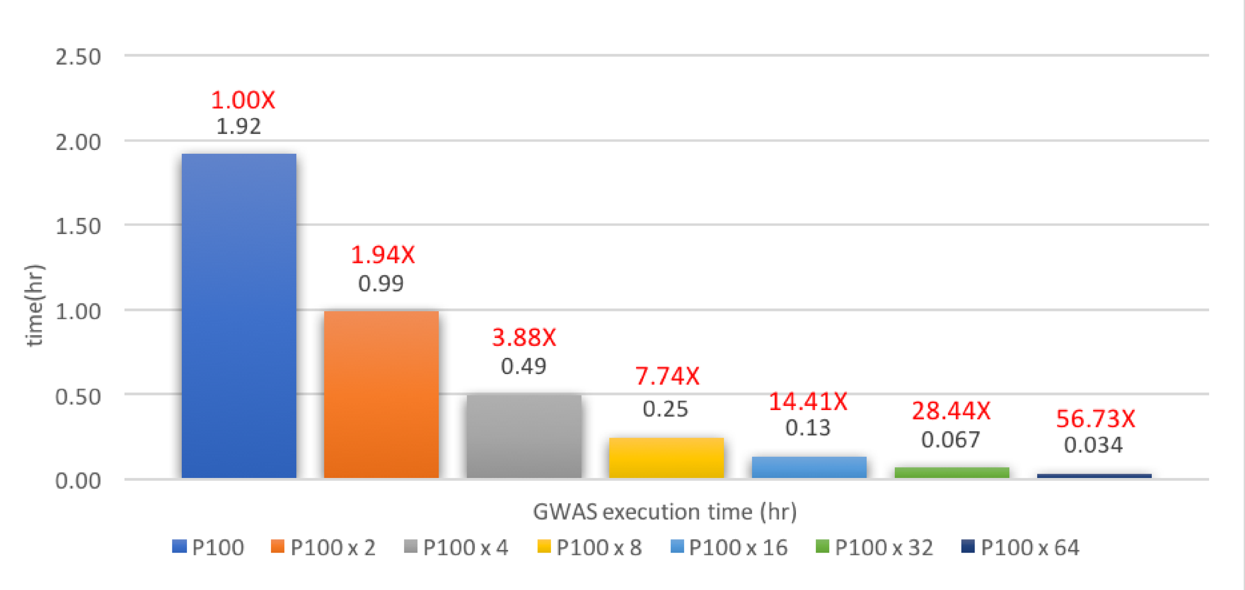
Simple calculus, significant effect.

Note that this integration is **bounded** and much simpler.

## Runtime Comparison

### High scalability

The algorithm has high scalability and runs incredibly fast.




The same problem that takes 3 months to do by 12 nodes MATLAB code can now be done in 2 minutes by using 64 NVIDIA P100 GPUs.

### Spec (Reedbush-H)

- Intel Xeon E5-2695 v4 x 2, Broadwell-EP 2.10 GHz, 18 cores, 45 MB L3-cache
- 256 GB RAM (DDR4-2400, 153.6 GB/sec)
- InfiniBand EDR 4x (100 Gbps/Processor)
- NVIDIA Tesla P100 GPU x 2, 16 GB Memory

## Numerical Results

To find the smallest p value, we usually look at $-\log p$ values instead.

We can sum up $-\log p$ values for all SNP to observe which part of brain is affected most.

### Standard brain

Map the results back to the standard brain

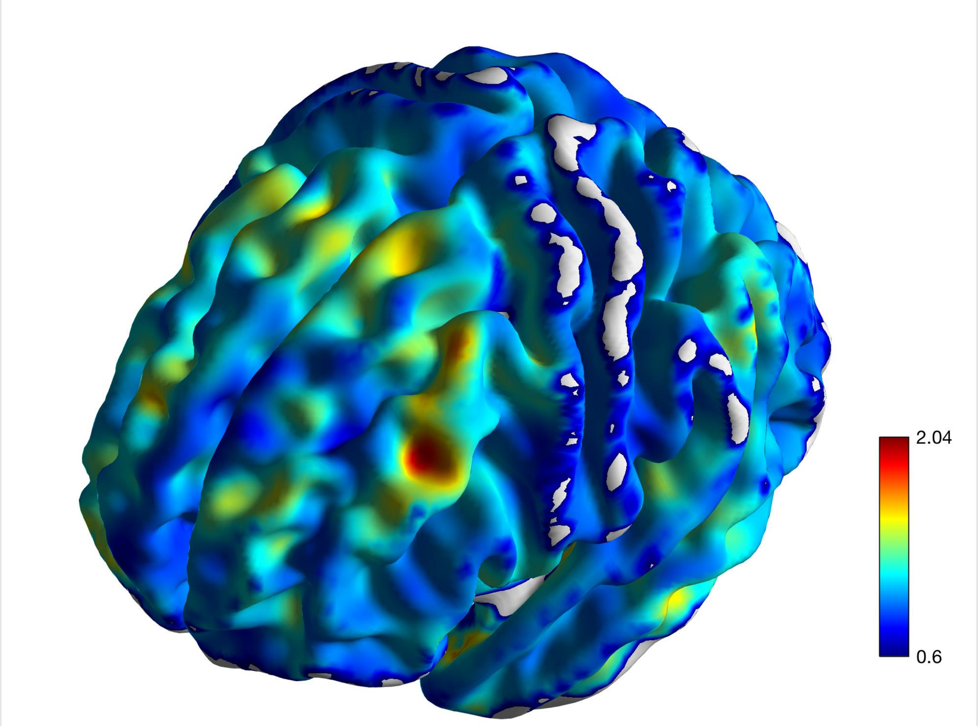



Figure 2: Example heat map of a standard brain

### Conclusion

We develop a method which makes large scale GWAS analysis possible. This algorithm can deal with extreme large scale GWAS problems while computes accurate p values at the same time.

## Acknowledgement

We are grateful to Professor Kengo Nakajima for helpful discussions and the Information Technology Center of The University of Tokyo for the usage of the Reedbush-H system.