Adapting Genome-wide Association Workflows for HPC Processing at Pawsey

Charlene Yang
Christopher Harris
Pawsey Supercomputing Centre
Technology Park, 26 Dick Perry Ave Kensington 6151, Australia

Sylvia Young
Grant Morahan
Harry Perkins Institute of Medical Research
6 Verdun St, Nedlands 6009, Australia

Abstract

As part of Bioinformatics Petascale Pioneers Program at Pawsey Supercomputing Centre, this project adapts a number of genome-wide association workflows to a unique Cray supercomputing environment, transforms the performance of these workflows and significantly boosts the research lifecycle. Pairwise and third order gene-gene interaction studies are sped up by hundreds of times through massive parallelization and GPU computing, laying the ground for larger datasets and higher order analysis in the future.

Introduction

Genome-wide association study involves exhaustive search across whole genome samples of many people to identify the genetic variations associated with certain diseases [1]. Such study greatly benefits human health by improved disease inheritance tracking and risk prediction. However, the amount of genetic variants in humans' DNA set, commonly seen as single nucleotide polymorphisms (SNPs), and their possible combinations result in extremely high computational complexity and hence unfeasibly long runtime for each analysis. Supercomputing holds the promise of performing quadrillions of operations in seconds, and this project demonstrates how it transforms the gene-gene interaction study and improves the performance by hundreds of times.

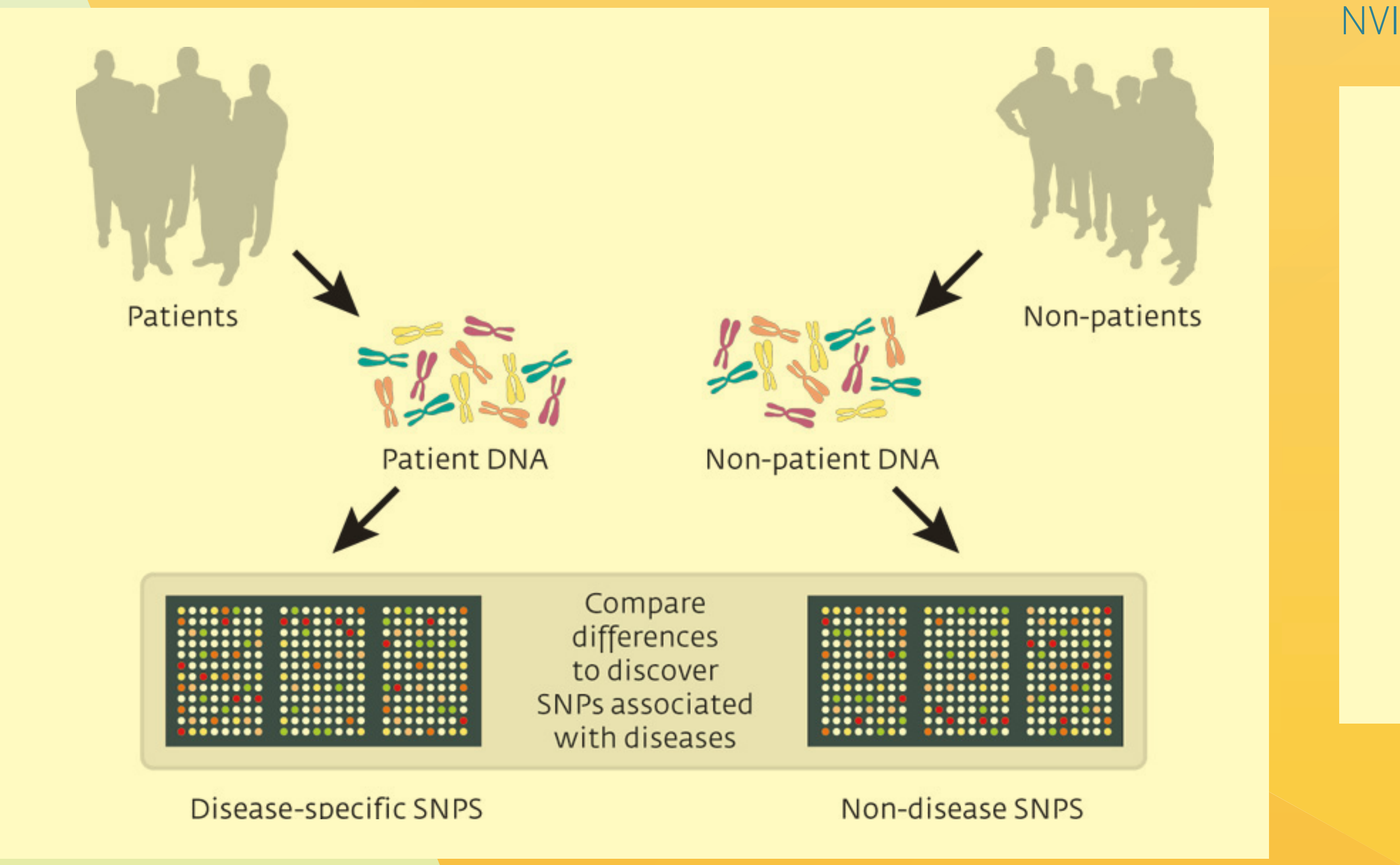

Figure 1. Genome-wide association study

Platform and Data

As part of the Bioinformatics Petascale Pioneers Program held at the Pawsey Supercomputing Centre, Western Australia, this project mainly uses Galaxy, a Cray XC30 machine, and Zeus, an SGI heterogeneous cluster, as its platform. It aims to adapt bioinformatics codes to the supercomputing environment, in particular the Cray software environment, so that at the end the project can be moved to the main power horse at Pawsey, Magnus, which is a Cray XC40 system.

Mangus

Cray XC40	Compute	Intel Xeon E5-2690V3 Haswell, 1,488 nodes, 35,712 cores, 1,097 TeraFLOPS
	Memory	DDR4-2133, 64GB per node, 93 TB total
	Interconnect	Cray Aries, 72 Gb/s per node
	Topology	Cray Dragonfly, 56% populated
	Storage	Sonexion 1600, 3 PB, w/ 70 GB/s, Lustre filesystem
	Other specs	weight 1.7 tonnes/cabinet, power 50 kW/cabinet

The dataset being tested is a 156,850 SNPs by 703 samples dataset from the Type 1 Diabetes Genetics Consortium (T1DGC) [2]. It's not a particularly large dataset but provides enough information for later research. The testing is mainly conducted on Intel Xeon E5-2690V2 Ivy Bridge CPUs on Galaxy or NVIDIA Tesla K20 GPUs on Zeus.

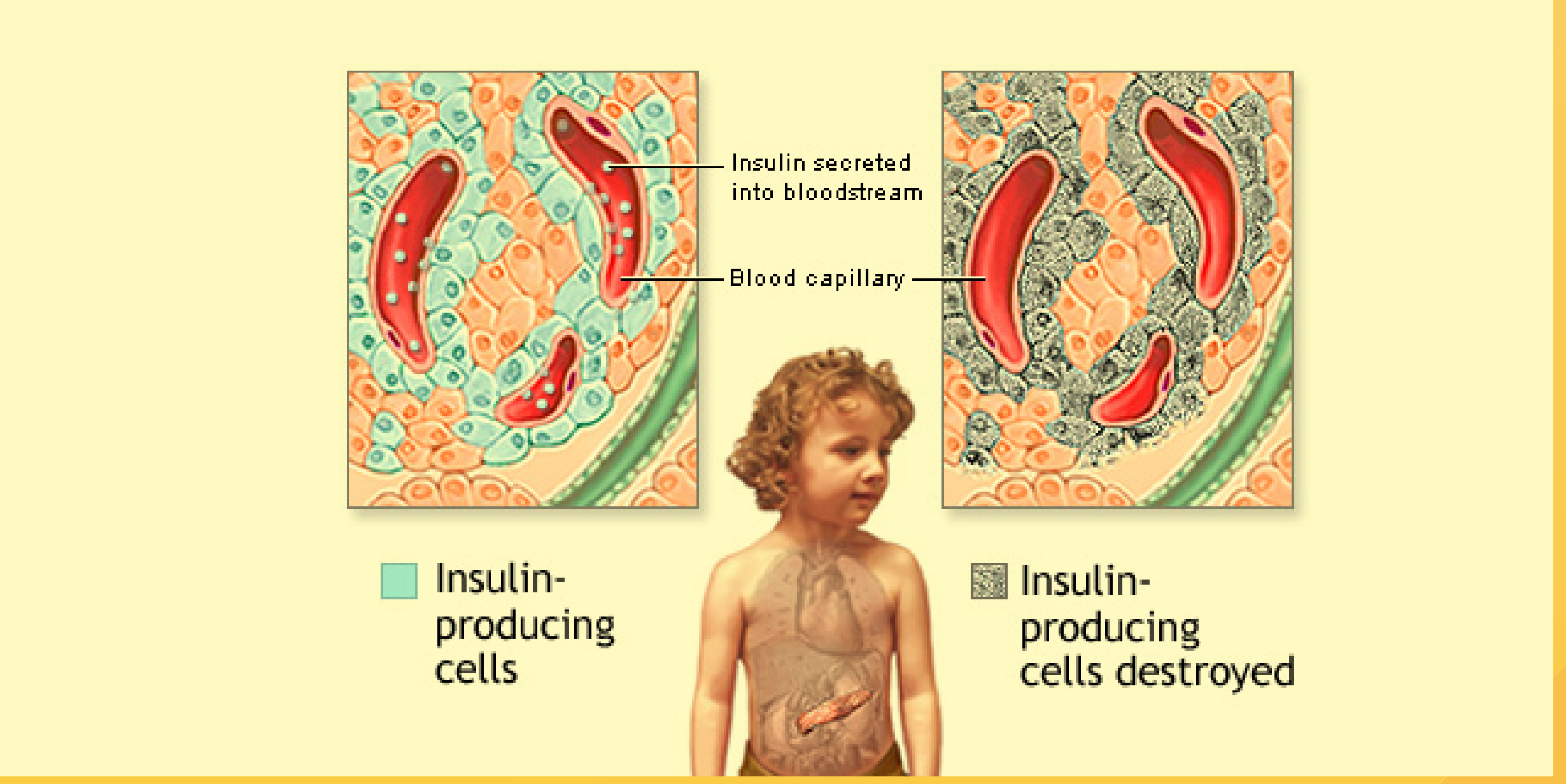
Insulin secreted into bloodstream
Blood capillary
Insulin-producing cells
▨ Insulin-producing cells destroyed

Figure 2. Type 1 diabetes

Methodology and Results

The project first examines the second order SNP interactions and it compares the performance of a number of software that are commonly used in interaction study such as PLINK 1.9 [3], BOOST [4], GBOOST [5] and GWISFI [6]. With PLINK 1.9, the code scales nearly linearly across 40 cores, bringing the runtime down from hours to minutes (see the speedup in Figure 3). Since the evaluation of each SNP interaction is independent, the Cray-specific environment variable ALPS_APP_PE is employed and with the help of the --parallel option of PLINK 1.9, it's able to decompose the data into partitions and distribute the partitions onto different processors.

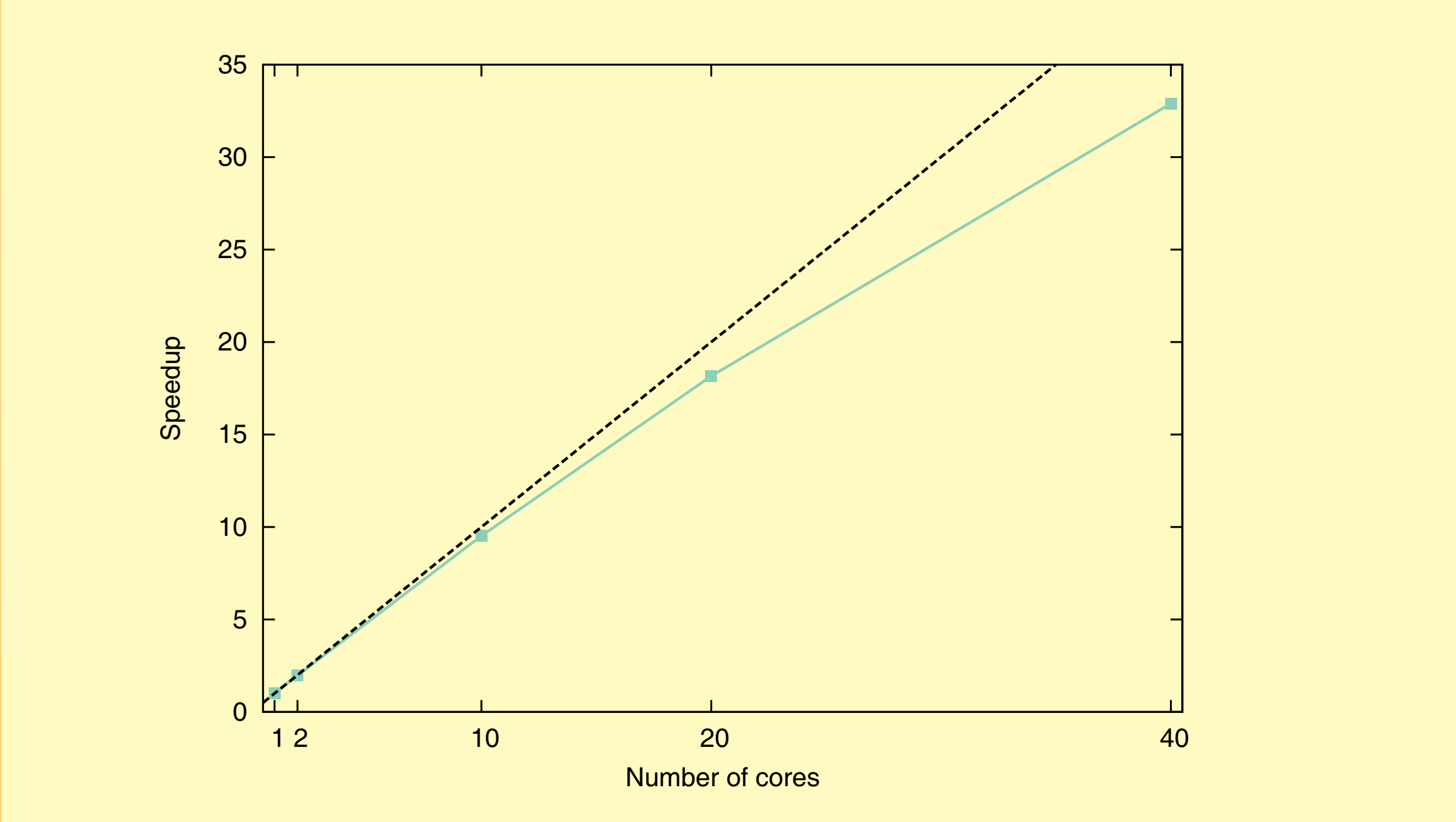

Figure 3. Scaling performance of PLINK 1.9 with boost

Even though the runtime has fallen into an acceptable range (minutes), we would want to speed it up further so that runtime won't be a problem when we process larger datasets and for higher-order interactions later on. GPU computing is introduced and as shown in Table 1, GBOOST and GWISFI achieves 54 times and 227 times of speedup compared to the original serial CPU implementation BOOST.

Table 1. Performance comparison of different software

Software	BOOST	PLINK 1.9	GBOOST	GWISFI
Runtime	4.5h	49.35m	4.9m	1.18m
Speedup	1	5.47	54.71	227.19

For third order methods, no GPU code has been developed yet but a reference serial code was produced in C++. Initial benchmarking indicates that the entire dataset of 156,850 pairs for the training population would require 91 days to process, and profiling reveals that 95% of the processing time was due to file access. The input method is then re-implemented to stage data in memory, resulting in a revised runtime of 2.8 days for the entire data set, an increase in efficiency by a factor of 32.

Since the algorithm is mostly 'pleasantly' parallel, a preliminary parallel implementation is trialed using the Cray environment variable ALPS_APP_PE. With the help of a wrapper bash script, this implementation runs multiple instances of the code with each processing a chunk of the data. Initial testing with a reduced dataset shows near-linear scaling but as each process is reading the entire dataset, the nodes on Galaxy do not contain enough memory to utilize every core.

To solve this memory issue, the Message Passing Interface (MPI) is employed to distribute the dataset between the processes. This distribution introduces some initial communication between processes. Specifically, the data for each previous order is broadcast from the process containing that data to all processes iteratively for each previous order. However, the resulting code is able to replicate the serial processing in hours using a few hundred cores, as shown in Figure 4.

The parallelized third order code is now extended to even higher order analysis and the optimization and parallelization is still continuing in order to conduct the multi-way interaction research in a timely manner.

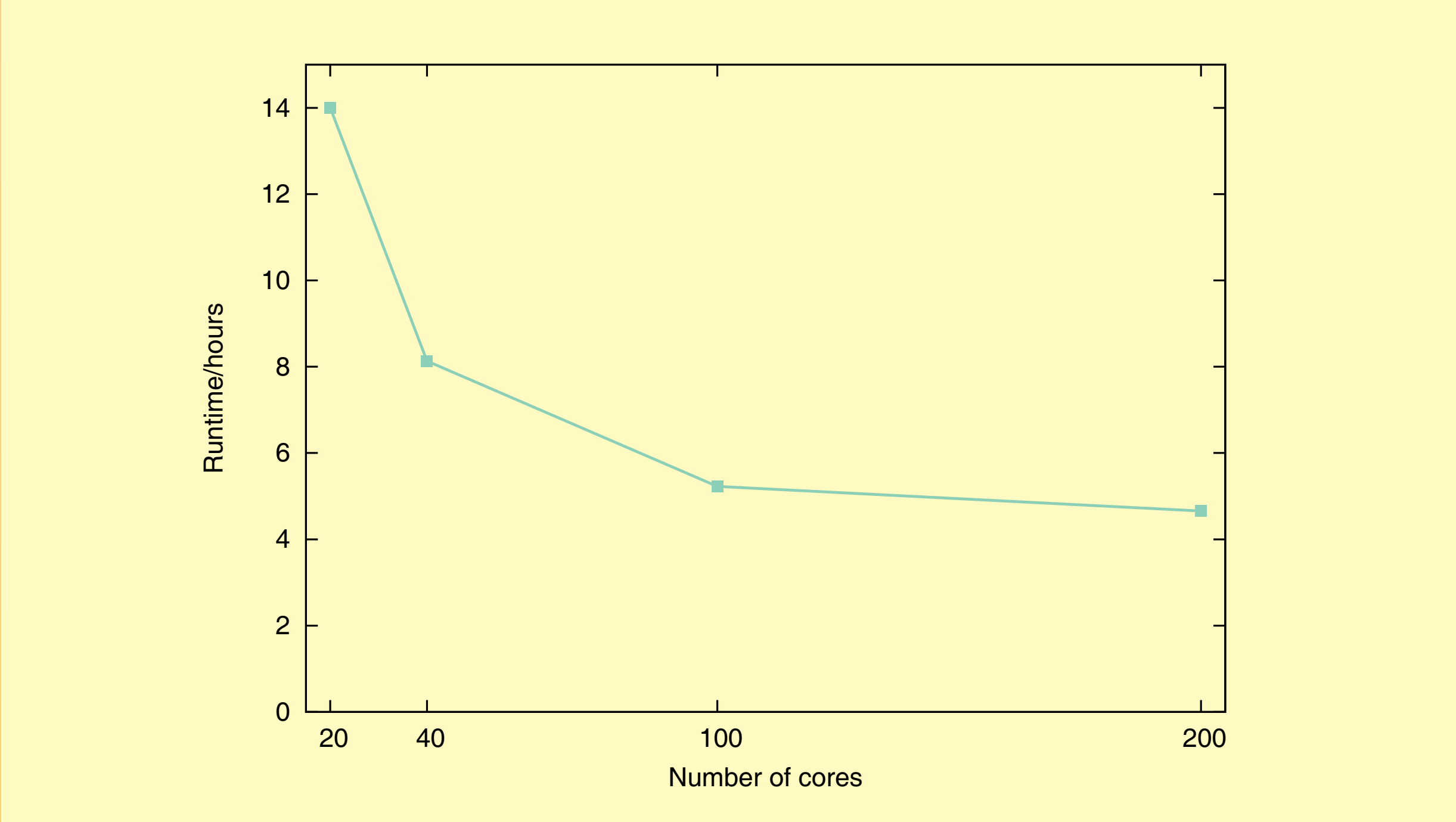

Figure 4. Scaling performance of MPI parallelized third order code

Conclusions

With advanced technologies such as parallel computing and GPU computing, this project has been accelerated by hundreds of times in terms of the analysis turnaround. While the presented workflows are yet to be applied to larger datasets, this project has illustrated that high order interaction studies are becoming feasible on modern supercomputers and advanced computing resources such as these at Pawsey Supercomputing Centre are capable of enabling research opportunities that may be unachievable otherwise.

Acknowledgements

This work uses data provided by Type 1 Diabetes Genetic Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development, and Juvenile Diabetes Research Foundation International (JDRF).

This work is supported by Pawsey Supercomputing Centre in Western Australia through the use of advanced computing resources.

References

[1] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 6(2): 95-108, 2005.
[2] T1DGC. https://www.niddkrepository.org/studies/t1dgc.
[3] PLINK 1.9. https://www.cog-genomics.org/plink2.
[4] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson LS Tang, and Weichuan Yu. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics, 87(3): 325-340, 2010.
[5] Ling Sing Yung, Can Yang, Xiang Wan, and Weichuan Yu. Gboost: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics, 27(9): 1309-1310, 2011.
[6] Qiao Wang, Fan Shi, A. Kowalczyk, R.M. Campbell, B. Goudey, D. Rawlinson, A. Harwood, H. Ferra, and A. Kowalczyk. Gwis_Fi: A universal GPU interface for exhaustive search of pairwise interactions in case-control GWAS in minutes. In 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 403-409, Nov 2014.

GOVERNMENT OF WESTERN AUSTRALIA
An Australian Government Initiative
National Collaborative Research Infrastructure Strategy
CSIRO
Curtin University
ECU EDITH COWAN UNIVERSITY
Murdoch UNIVERSITY
THE UNIVERSITY OF WESTERN AUSTRALIA
HARRY PERKINS INSTITUTE OF MEDICAL RESEARCH
THE UNIVERSITY OF MELBOURNE