34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Supercomputing Enabling Exhaustive Statistical Analysis of Genome Wide Association Study Data: Preliminary Results*

Matthias Reumann[¶], *Member, IEEE*, Enes Makalic[¶], Benjamin W. Goudey, Michael Inouye, Adrian Bickerstaffe, Minh Bui, Daniel J. Park, Miroslaw K. Kapuscinski, Daniel F. Schmidt, Zeyu Zhou, Guoqi Qian, Justin Zobel, John Wagner, John L. Hopper

Abstract—**Most published GWAS do not examine SNP interactions due to the high computational complexity of computing *p*-values for the interaction terms. Our aim is to utilize supercomputing resources to apply complex statistical techniques to the world's accumulating GWAS, epidemiology, survival and pathology data to uncover more information about genetic and environmental risk, biology and aetiology. We performed the Bayesian Posterior Probability test on a pseudo data set with 500,000 single nucleotide polymorphism and 100 samples as proof of principle. We carried out strong scaling simulations on 2 to 4,096 processing cores with factor 2 increments in partition size. On two processing cores, the run time is 317h, i.e. almost two weeks, compared to less than 10 minutes on 4,096 processing cores. The speedup factor is 2,020 that is very close to the theoretical value of 2,048. This work demonstrates the feasibility of performing exhaustive higher order analysis of GWAS studies using independence testing for contingency tables. We are now in a position to employ supercomputers with hundreds of thousands of threads for higher order analysis of GWAS data using complex statistics.**

I. Introduction

Genome Wide Association Studies (GWAS) are performed to identify genetic markers, i.e. single nucleotide polymorphisms (SNPs), for the analysis of biological traits and disease. These studies have been made possible by sequencing the human genome [1] and the completion of the subsequent human haplotype mapping project HapMap [2]. Since 2005 alone, over 2,700 GWAS have been conducted. With an average cost of $500,000 per study this amounts to $1.35 billion having been spend on generating data. Given the high costs involved in running a GWAS, there is clearly a great need to ensure that the information in the collected data is fully utilized. While GWAS have produced major advancements in the understanding of genetic basis of disease, the generated data is not extensively explored. For example, gene-gene and gene-environment interaction as well as complex networks of gene regulation replace the notion that a single gene is causative of a phenotypic trait or disease. Table 1 shows time estimates to perform 2-way and 3-way interaction studies using two standard analysis methodologies. The computation of 2-way interactions takes days if not years. The computation of 3-way interactions on standard sample sizes is estimated to take thousands if not millions of years [3]. This clearly demonstrates that higher order interaction studies cannot be carried out using commodity-computing resources but require clusters [4] or supercomputers [3].

TABLE I. Run time estimates for SNP interaction Analysis on single 3GHz CPU

# SNP (interaction)	IG method[a] [5]	BOOST[b] [6]
500,000 (2 way)	300 days	3.5 days
1,000,000 (2 way)	3.3 years	13.9 days
500,000 (3 way)	137,000 years	1,500 years
1,000,000 (3way)	1,095,000 years	12,600 years

a. IG – information gain

b. BOOST – Boolean operation based screening and testing

Most published GWAS do not examine SNP interactions due to: (a) the high computational complexity [3] of computing *p*-values for the interaction terms, and (b) the typically low power to detect significant interactions.

Thus, it is currently intractable to carry out any but simplistic analyses on these large data sets due to lack of computer power and memory and therefore the full utility of the resources and technology has not been realized.

We have been implementing and testing new methods and approaches in the field of genomics to make feasible the analysis of GWAS data using more complex models including machine learning approaches and systems biology.

Our aim is to utilize supercomputing resources to apply complex statistical techniques to the world's accumulating GWAS, epidemiology, survival and pathology data related

*This research was supported by the NHMRC APP1033452 grant and the Victorian Life Sciences Computation Initiative (VLSCI) an initiative of the Victorian Government, grant VR0550.

[¶]equal contribution

M. Reumann is from the IBM Research Collaboratory for Life Sciences-Melbourne, 187 Grattan Street, Carlton, VIC 3010, Australia and the Dept. Computing and Information Systems, University of Melbourne, Carlton, Australia (corresponding author phone: +61 3 9035 4432 e-mail: mreumann@ieee.org).

E. Makalic, M. K. Kapuscinski, M. Bui, A. Bickerstaffe, D. F. Schmidt and J. L. Hopper are from the Melbourne School of Population Health, University of Melbourne, Parkville, Australia.

M. Inouye is from the Dept. of Pathology and the Dept. Microbiology and Immunology, University of Melbourne, Parkville, Australia.

D. Park is from the Dept. of Pathology, University of Melbourne, Parkville, Australia.

B. W. Goudey, Z. Zhou and J. Wagner are from the IBM Research Collaboratory for Life Sciences-Melbourne, Carlton, Australia.

Guoqi Qian is from the Dept. of Mathematics and Statistics, University of Melbourne, Parkville, Australia.

J. Zobel is from the Dept. of Computing and Information Systems, University of Melbourne, Parkville, Australia; and NICTA VRL.

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

to breast and prostate cancers so as to uncover more information about genetic and environmental risk, biology and aetiology.

To achieve this goal, we define three objectives:

- Develop analytical and computational approaches applicable to GWAS and ancillary data on disease and health-related conditions using high performance supercomputing.
- Apply these methods to provide new insights into the genetic and environmental causes, aetiology and biology of breast and colorectal cancers.
- Develop and maintain an expert workforce skilled in the statistical, computing and content areas necessary to achieve Objectives 1 and 2, so as to be able to apply these approaches to other diseases and health-related conditions.

In this article we demonstrate a computation framework and show preliminary results on a pseudo data set to address the issue of high computational complexity of computing *p*-values for interaction terms and to facilitate objective 1 by leveraging supercomputing resources for the analysis of higher order interactions of GWAS data.

II. Methods

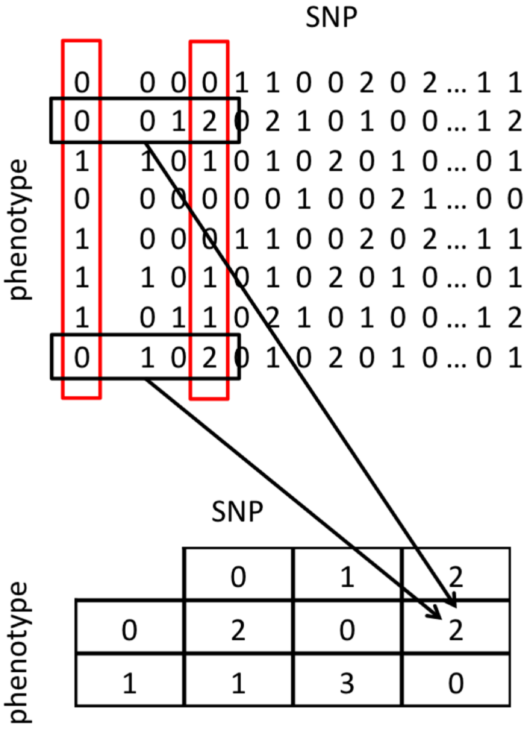
A. Genome Wide Association Study

A GWAS typically involves using high-throughput genotyping technologies to measure hundreds of thousands of genetic markers (single nucleotide polymorphism (SNPs)) across the genome for both cases (affected) and controls (unaffected), providing a novel way to identify candidate markers/genes related to a wide range of phenotypes, i.e. traits (e.g., height, weight) and diseases (e.g., breast cancer, asthma) [7]. In GWAS, a phenotype is defined by 0 (non-disease) or 1 (disease). Similarly, a SNPs are represented by 0 (majority allele), 1 (minority allele), 2 (heterozygous) or 3 (undefined).

B. Bayesian Posterior Probability test

We chose to build a computation framework around a statistic that computed the independence of contingency tables has a long history and is well established [8]. A contingency table is assembled (Fig. 1) by counting pairs of phenotype and SNP, i.e. how many pairs of [0, 0], [1, 0], [1, 0], [1, 1] etc. are given for a particular SNP and the corresponding phenotype.

We compute the Bayesian factor to determine the posterior probability for each contingency table. For our purpose of demonstrating the analysis framework on a supercomputer, however, we arbitrarily chose the computation of the Bayesian Posterior Probability (BPP) test for $a \times b$ contingency tables
as described in [8]. It includes the calculation of a Monte Carlo sum to determine the intrinsic priors that are needed to perform the BPP test. Each test is independent which allows the BPP test to be replaced by other statistics like e.g. the Pearson χ^2 test. It is beyond this manuscript to discuss the differences between statistics that can be used to compute the independence of contingency tables and we refer to [6].



0 | 0 0 0 1 1 0 0 2 0 2 … 1 1
0 | 0 1 2 0 2 1 0 1 0 0 … 1 2
1 | 1 0 1 0 1 0 2 0 1 0 … 0 1
0 | 0 0 0 0 0 1 0 0 2 1 … 0 0
1 | 0 0 0 1 1 0 0 2 0 2 … 1 1
1 | 1 0 1 0 1 0 2 0 1 0 … 0 1
1 | 0 1 1 0 2 1 0 1 0 0 … 1 2
0 | 1 0 2 0 1 0 2 0 1 0 … 0 1

	0	1	2
0	2	0	2
1	1	3	0

Figure 1. Example of how a *2 ×3* contingency table is constructed. The phenotype–SNP pairs are summed up over all samples that then make up the values of a two dimensional matrix.

C. Computing Resources

Our team has access to an IBM Blue Gene supercomputer at the Victorian Life Sciences Computation Initiative (VLSCI). However, all methods and principles described in this work can be applied to any distributed memory, parallel computer. The VLSCI Peak Computing Facility (PCF) currently offers access to two racks of IBM Blue Gene/P (BG/P) with 1,024 quad core chips (850 MHz) per rack, i.e. 8,192 cores, and a total of 8TB memory. While we currently can only present our results for analysis that were performed on up to 4,096 cores on BG/P, our methodology is applicable for the future BG/Q supercomputer with 65,536 cores (1.6 GHz) and 262,144 hardware threads that is currently being installed at VLSCI and will be available in July 2012.

D. Parallelization and Load Balancing

The parallelization and load balancing exploits the fact that the computation of the BPP test can be computed on a single contingency table independent of the test being performed on any other contingency table. Thus, once we have created the contingency tables for all SNP interactions, we can split the data across all processing cores and carry out the computation in pleasantly parallel fashion.

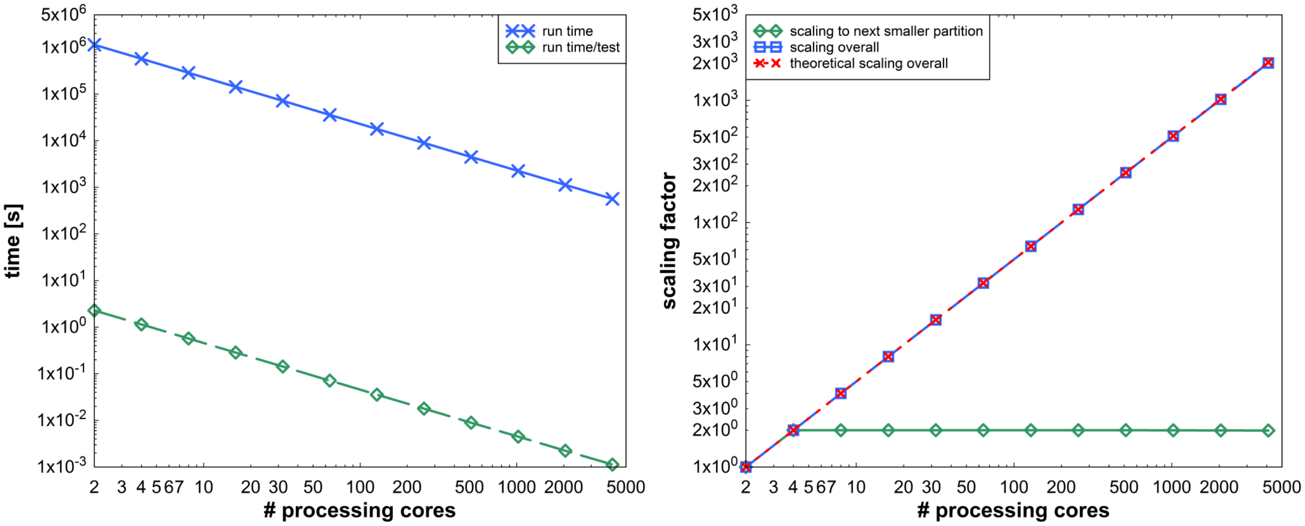


Fig. 2 Performance results: maximum run times (left) and scaling (right) between partition N = 2 and N = 4,096 processing cores

TABLE II. RUN TIME [S], SCALING FACTOR AND PARALLEL EFFICIENCY FOR BAYESIAN POSTERIOR PROBABILITY (BPP) TEST[a]

# CPUs	Data decomposition	Compute contingency table	BPP test	Run time	Run time per test	Scaling factor	Overall scaling	Parallel efficiency [%]
2	1.0 x 10^{-6}	4.2171	1141280.23	1141286.13	2.2826	N/A	N/A	N/A
4	2.0 x 10^{-6}	2.1348	570839.25	570843.08	1.1417	1.9993	2.00	99.96
8	2.0 x 10^{-6}	1.0688	285445.33	285448.11	0.5709	1.9998	4.00	99.96
16	2.0 x 10^{-6}	0.5334	142714.21	142716.47	0.2854	2.0001	8.00	99.96
32	2.0 x 10^{-6}	0.2670	71358.13	71360.12	0.1427	1.9999	15.99	99.96
64	2.0 x 10^{-6}	0.1336	35682.45	35684.30	0.0714	1.9998	31.98	99.95
128	2.0 x 10^{-6}	0.0668	17845.32	17847.11	0.0357	1.9994	63.95	99.92
256	2.0 x 10^{-6}	0.0334	8924.38	8926.13	0.0178	1.9994	127.86	99.89
512	2.0 x 10^{-6}	0.0168	4462.36	4464.10	0.0089	1.9995	255.66	99.87
1,024	5.520 x 10^{-4}	0.0076	2234.82	2236.55	0.0045	1.9960	510.29	99.67
2,048	6.250 x 10^{-4}	0.0037	1120.43	1122.15	0.0022	1.9931	1017.05	99.32
4,096	3.010 x 10^{-3}	0.0019	563.14	564.86	0.0011	1.9866	2020.47	98.66

[a] 500,000 SNPs and 100 samples

The idea is simple with the following approach:

- Determine how many contingency tables need to be computed
- Divide the number by the number of processing cores available – the absolute number determines the number of tests each processor has to perform
- Split the list of contingency tables accordingly and distribute to each processor.

This strategy will lead to some processing cores without computation, i.e. they will be idle. However, if one was to apply a more sophisticated method, then all processors might do work but the difference between how many tests each processor performs is one test. Since the overall run time depends on the processor that takes longest to perform its task, our simpler method will run as fast as other sophisticated methods.

In this study we distribute the GWAS data set to all compute tasks, calculate the contingency tables for all tests and perform the Bayesian Posterior Probability test on a 2 × 4 contingency table, i.e. for single SNP testing, where a SNP can have values 0, 1, 2 and 3. Data decomposition and communication was implemented using the Message Passing Interface (MPI) and C/C++.

E. Strong scaling experiments

We perform strong scaling experiments to investigate the performance of the proposed strategy for a framework to carry out complex statistical analysis of GWAS. We define an artificial GWAS data set with 500,000 SNPs of 100 samples. This leads to 500,000 tests to be performed. We set the SNPs to all 1's as well as the phenotypic trait is 1. While this test does not contain any information, it allows us to test performance.

We performed the test on 2 to 4,096 processing cores on BG/P with processor numbers incrementing by 2. The problem size was kept constant. We estimated this simulation would take over four weeks on a single processing core. Computing resource constraints make unfeasible for us to perform the simulation on a single processing core. Thus, the overall speedup was defined with respect to the run time on 2 processing cores.

All simulations were carried out in virtual node mode, i.e. each core was associated with one single threaded MPI task.

III. Results

Table 1 shows the results for timing and strong scaling. The run times decrease linearly with the number of cores for all compute partitions (Fig. 2). On two processing cores, the run time is 317h, i.e. almost two weeks. On 4,096 processing cores, the run time is reduced to less than 10 minutes. The corresponding speedup factor is 2,020. This is very close to the theoretical speedup factor of 2,048. Indeed, the scaling is just below 2 between for N vs. 2N processing cores. Only at N = 2,048 and N = 4,096 we see that the scaling is getting slightly smaller with 1.9931 and 1.9866, respectively. The parallel efficiency also shows near perfect load balancing and scaling. It is 99.96% for the smaller partitions N = 2, 4, 8, 16, 32 and goes down to 98.66% on N = 4,096.

The run time is determined by the BPP test with the building of the contingency table only taking up a very small fraction of time. It is also worth noting that the run time per BBP test is decreasing linearly with the number of processors used.

IV. Discussion

A. Scaling simulations

The simulations scale as expected. Since each test is an independent test on a single contingency table and because each contingency table is small in memory, the simulations are pleasantly parallel. A speedup of 2 weeks on two processing cores down to fewer than ten minutes on 4,096 processing cores demonstrates that exhaustive analysis of GWAS data becomes feasible.

Further, the high parallel efficiency on even the largest number of processing cores used shows that massively parallel, distributed memory supercomputers are suitable to perform the task of computing independence in contingency tables for a large set of tests.

B. Limitations and future work

There are a few limitations to the current state of this work. The pseudo data set that we generated is not representative of real GWAS studies. We note that the computation of the Monte Carlo Sum differs given a random distribution of 0's and 1's in the SNP data. However, our results show the worst-case scenario since the maximum number of iterations has been used for computing the Monte Carlo sum. We also show the statistic for univariate analysis only. Therefore the number of tests performed is 500,000. For two way interaction, the number of tests increases to $N(N - 1)/2$ with N being the number of SNPs. However, since the calculation of a test is deterministic, we can estimate accurately the run times for these higher order interaction studies. Roughly 125 billion tests would have to be performed for a two-way interaction analysis of a GWAS data set that measured 500,000 SNPs per sample. This is ~250,000 times the number of tests performed in the presented study which would be over 41 thousand hours. This leads to the observation that future work must include speeding up of the test statistic. By choosing other test statistics than the BPP would enable carrying out more tests per second per processing core. An ad hoc implementation of the χ^2 test [4] showed an increase of four orders of magnitude of the number of tests that can be performed on a single processing core per second. Assuming scalability as shown in our results, this would make the analysis possible in 4.1 hours. Extrapolating our parallel performance to the full 4 rack Blue Gene/Q system at VLSCI with 262,144 hardware threads assuming the same performance on the full system, the run time would be reduced to under 0.06 seconds. Compared to previously estimated 1.2 years [3] in 2008 and less than 9 hours [4] in 2011 this would be a speedup of over 630 billion and 500,000, respectively. Considering that future analysis will be performed on imputed GWAS data with millions of SNPs, such a speedup and further optimization is required for exhaustive analysis of many way SNP interaction studies.

V. Conclusion

This work demonstrates the feasibility of performing exhaustive higher order analysis of GWAS studies using independence testing for contingency tables. The problem is pleasantly parallel that allows the use of massively parallel, distributed memory supercomputers. We are now in a position to employ supercomputers with hundreds of thousands of threads for higher order analysis of GWAS data using complex statistics.

References

[1] Consortium International Human Genome Sequencing. Initial sequencing and analysis of the human genome. *Nature* 2001;409 (6822):860–921

[2] International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426(6968):739.

[3] Ma L, Runeshae HB, Dvorkin D, et al. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. *BMC Bioinformatics* 2008;9:315

[4] Wang Z, Wang Y, Tan K-L, et al. eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study. *Bioinformatics* 2011;27(8):1045–1051

[5] Wan X., Yang C., Yang Q. et al. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am J Hum Genet.* 2010;87:325–340

[6] McCarthy M. I., Abecasis G. R., Cardon L. R. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 2009;9:356–369

[7] Casella G., Moreno E. Assessing robustness of intrinsic tests of independence in two-way contingency tables. *Journal of the American Statistical Association*. 2009;104(487):1261–1271