Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Thompson and Charnigo, J Biom Biostat 2015, 6:1
DOI: 10.4172/2155-6180.1000e131

Editorial Open Access

Parallel Computing in Genome-Wide Association Studies

Katherine Thompson[1] and Richard Charnigo[1,2]*

[1]Department of Statistics, University of Kentucky, USA
[2]Department of Biostatistics, University of Kentucky, USA

Introduction

Technological advances have greatly reduced the cost and time needed to sequence an individual's genome (i.e., identify the DNA base pairs that constitute an individual's genome), leading to dramatic growth in the amount of genomic data to analyze. Although the development of novel sequencing technologies has been prolific, the growth in methods to detect associations between phenotypes (i.e., physical traits of interest) and locations along the genome, more specifically single nucleotide polymorphisms (SNPs), has been comparatively limited. Hindrances to methodological development include a lack of available computing power to analyze data in a feasible time frame (i.e., in a time frame useful to researchers). One technique used to speed up computation (i.e. make computation more efficient) is parallelization, where multiple tasks are performed simultaneously on multiple cores or threads within a machine [1]. Even with computational and methodological limitations, analysis of genome-wide association study (GWAS) data has led to the inference of connections among several phenotypes and SNPs [2,3].

For example, in a two-phase GWAS involving almost 400,000 SNPs and 1,363 samples, Sladek and colleagues identified four novel genomic locations associated with type 2 diabetes [2]. The study was complicated by the suspicion that there might be interactions between genetic and environmental risk factors. This GWAS would have been infeasible to perform on an individual desktop or laptop computer due to the amount of time that would have been required. In this case, researchers used a supercomputing facility to perform the data analysis.

Due to the size of GWAS data sets, computational efficiency and feasibility are of particular concern. Accordingly, the search for quantitative trait loci (i.e., locations along the genome connected to a quantitative trait of interest) has relied heavily on methods that are computationally feasible for data sets with up to thousands of individuals and hundreds of thousands of SNPs. Data sets can include samples from individuals with known familial history (pedigree-based samples) or individuals unrelated by familial history (population-based samples). This editorial focuses on association mapping, or the use of population-based samples, to find SNP-phenotype connections [4].

In this text, we briefly review association mapping methods, which includes comment on their respective computational abilities. Lastly, we discuss potential computational resource development that may enhance the feasibility of analyzing data using association mapping approaches.

Review

Current association mapping methods have demonstrated improved computational efficiency compared to initially-developed approaches. Although early methods such as TreeLD [5] were tested on fewer than 100 SNPs in no more than 250 individuals, more recent methods such as EMMAX can be readily applied to data from tens of thousands of individuals and as many as hundreds of thousands of SNPs [6]. Some of these more recent approaches are identified below.

Methods differ in their computational abilities due to variation in their complexities. Fundamental techniques that analyze SNPs marginally are easily parallelized since computations involving a single SNP do not affect computations from other SNPs. Those that analyze SNPs marginally include Single Marker Analysis [7] and the two-sample *t*-test. However, these methods do not allow for the analysis of covariates nor the accounting for multiple SNPs simultaneously. Approaches that do include regression-based methods such as EMMA [8], EMMAX [6], SNPAssoc [9], and PLINK [10]. Several implementations of regression-based techniques exist, and include but are not limited to SNPAssoc [9], EMMA [8], EMMAX [6], GEMMA [11], and pi-MASS [12]. By parallelizing the computations, total analysis time can be decreased by a large factor, depending upon the number of processors available [13].

Other more computationally intensive methods, such as tree-based approaches that account for heterogeneous correlation structures, could also benefit from parallelization. These techniques include those in [5,7,14-19], and whether or not they are already parallelized or could be parallelized is highly context and implementation dependent. However, these do not allow for external influences on a phenotype as do the less computationally intensive methods mentioned above. In exchange, tree-based approaches use information about evolutionary relationships among copies of particular SNPs to gain power in detecting SNP-phenotype associations. Although current implementations of some of these approaches are parallelized, speed of the other algorithms could be greatly improved by parallelization.

For a historical illustration of the benefits of parallel computing, we consider the 2001 work of Carlborg, Andersson-Eklund and Andersson. They studied the computational gain when using a regression-based method for quantitative trait mapping [13]. In this study, the relative increase in performance (analysis time on one processor divided by analysis time for multiple processors) was as large as 7.04 when the number of processors increased from 1 to 18. This is only a fraction of the number of processors currently available in many supercomputers. In addition, this study was performed by parallelizing the processes by chromosome rather than by amount of analyses, meaning that a more complicated parallelization would show an even larger improvement in computational time versus a single processor.

Techniques used in the 2001 study also point to the variation in time and knowledge needed to parallelize an analysis on a multi-core computer [13]. Tutorials such as that of [20,21] have been developed to aid an analyst using R or another coding software in employing multiple cores via free, open-source implementations of parallelized

***Corresponding author:** Richard Charnigo, Department of Biostatistics, University of Kentucky, USA, E-mail: rjcharn2@aol.com

Received May 11, 2015; **Accepted** May 12, 2015; **Published** June 11, 2015

Citation: Thompson K, Charnigo R (2015) Parallel Computing in Genome-Wide Association Studies. J Biom Biostat 6: e131. doi:10.4172/2155-6180.1000e131

data analysis techniques. Although some extra computer code may be required to parallelize techniques, running these analysis methods without parallelization on simple desktop computers has become infeasible due to the growing numbers of SNPs in modern data sets, as the analyses would not be completed in any reasonable time frame [22].

Lastly, the cost of parallelization can be highly variable, contingent upon type of machine and the number of threads or cores the machine contains. In the academic community, channels allowing researchers to share supercomputing resources for specific tasks (see those in [22], for example) are being developed so that researchers can pool resources for GWAS data analysis.

Discussion

In recent years, GWAS analysis methods have been heavily criticized, partly due to their lack of power in detecting SNP-phenotype associations [23]. However, methods such as those in [15,16] have shown promise in improving detection power, subject to increased computational cost. Although some of them entail computations that are not directly parallelizable, others either use or could benefit from parallel computing. This might yield a substantial gain in computation speed, even with the added cost of using extra information in the data. In addition, computational hardware has been drastically improved in recent years [1], meaning that we have the computational power necessary to parallelize computations, even in the analysis of large genomic data sets.

In addition to gains in computational speed, more effective use of computational resources would also enhance the capacity of analysis methods to handle more complex data structures. As an example from recent history, pleiotropic SNPs (i.e., single SNPs affecting multiple traits) associated with human diseases have been reviewed in [24]. In addition, cases where multiple SNPs affect a single trait have been detected (e.g., see [3]). Also, interactions among genetic and external factors are prevalent; for instance, undesirable cardiovascular traits such as elevated serum cholesterol can be particularly sensitive to an external component such as diet when the genetic component is unfavorable. Situations like these give rise to complicated data types that will require more sophisticated methodology. Parallel computing may be used to offset the computational intensity imposed by such complicated data structures.

Recent advances in computational hardware have been advantageous in analyzing large data sets in many fields, including genetics, environmental sciences, chemistry, physics, and engineering [25]. Association mapping method development would highly benefit from parallelizing of computations, due to added computational speed in detecting associations among SNPs and phenotypes. Since the number of SNPs in each data set has grown so large that detecting SNP-phenotype associations could take years in the absence of parallelization, computational speed increases have become necessary. In fact, some recent analyses of large data sets that identified SNP-phenotype links would have been infeasible without advanced computational resources (e.g., see [2]). Although the growth in available data has not been an issue due to reduced data storage costs, supercomputing facilities are becoming a necessity for analyzing GWAS data [22]. Computational equipment developed so that academic researchers can share these resources, along with developing cloud technologies (for example, from Google and Amazon), have begun to provide additional computing power at reduced cost [22]. The growing accessibility of these resources evidences their promise in providing the computational power to analyze existing and newly-collected data in the coming years. With this computational power, methods that use information about the evolutionary history present within each SNP could realize improvements in both computational speed and detection power for GWAS emerging from complicated, medically-relevant scenarios.

References

1. Berger B, Peng J, Singh M (2013) Computational solutions for omics data. Nature Reviews Genetics 14: 333-346.
2. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.
3. Linnen CR, Poh YP, Peterson BK, Barrett RDH, Larson JG, et al. (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339: 1312-1316.
4. Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics 10: 565-577.
5. Zöllner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169: 1071-1092.
6. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348-354.
7. Zhang Z, Zhang X, Wang W (2012) HTreeQA: Using semi-perfect phylogeny trees in quantitative trait loci study on genotype data. G3: Genes, Genomes, Genetics 2: 175-189.
8. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.
9. González JR, Armengol L, Solé X, Guinó E, Mercader JM, et al. (2007) SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23: 644-645.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559-575.
11. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat. Genet 44: 821-824.
12. Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5: 1780-1815.
13. Carlborg Ö, Andersson-Eklund L, Andersson L (2001) Parallel computing in interval mapping of quantitative trait loci. Journal of Heredity 92: 449-451.
14. Mailund T, Besenbacher S, Schierup MH (2006) Whole genome association mapping by incompatibilities and local perfect phylogenies. BMC Bioinformatics, 7: 454.
15. Besenbacher S, Mailund T, Schierup MH (2009) Local phylogeny mapping of quantitative traits: Higher accuracy and better ranking than single-marker association in genomewide scans. Genetics 181: 747-753.
16. Thompson KL, Kubatko LS (2013) Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. BMC Bioinformatics 14: 200.
17. Minichiello MJ, Durbin R (2006) Mapping trait loci by use of inferred ancestral recombination graphs. American J Hum Genet 79: 910-922.
18. McClurg P, Pletcher TM, Wiltshire T, Su AI (2006) Comparative analysis of haplotype association mapping algorithms. BMC Bioinformatics 7: 61.
19. Pan F, McMillan L, Pardo-Manuel de Villena F, Threadgill D, Wang W (2009) TreeQA: Quantitative genome wide association mapping using local perfect phylogeny trees. Pacific Symposium on Biocomputing 415-426.
20. Arends D, Prins P, Broman KW, Jansen RC (2014) Tutorial-Multiple-QTL Mapping (MQM) Analysis for R/qtl.
21. Wang L, Ware D, Lushbough C, Merchant N, Stein L (2014) A genome-wide association study platform built on iPlant cyber-infrastructure. Concurrency and Computation: Practice and Experience 27: 420-432.
22. Stillman JH, Armstrong E (2015) Genomics are transforming our understanding of responses to climate change. BioScience 65: 237-246.

23. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five Years of GWAS Discovery. Am J Hum Genet 90: 7-24.

24. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607-618.

25. Barney B (2015) Introduction to parallel computing. Lawrence Livermore National Laboratory.