Shared Genomics: Developing an accessible integrated analysis platform for Genome-Wide Association Studies

David Hoyle, PhD, Mark Delderfield, Lee Kitching, Gareth Smith, Iain Buchan, MD, DPH.
North-West Institute for BioHealth Informatics (NIBHI), University of Manchester, Manchester, UK.

Abstract

Increasingly, genome-wide association studies are being used to identify positions within the human genome that have a link with a disease condition. The number of genomic locations studied means that computationally intensive and bioinformatic intensive solutions will have to be used in the analysis of these data sets. In this paper we present an integrated Workbench that provides user-friendly access to parallelized statistical genetics analysis codes for clinical researchers. In addition we biologically annotate statistical analysis results through the re-use of existing bionformatic Taverna workflows.

Introduction

Genome-wide genotyping technologies, such as Affymetrix SNP Chips or Illumina Bead technology are now allowing us to examine genetic variation between case and control groups on an unprecedented scale. This promising technology has come to the fore due to a number of large academic collaborative studies [1], carried out in specialized academic centres with expertise in statistical genetics, often utilizing large scale compute resources [2]. An inevitable reduction in unit cost will ultimately bring Genome-Wide Association (GWA) studies into the realm of mainstream clinical researchers and smaller, niche but richly characterized cohorts.

The analysis challenges posed by these new data sets will be far greater than that posed by the `flood' of post-genomic data within the last decade. Interpretation of results is more difficult, in part due to the complex etiology of the diseases, but also because for human diseases the huge variation in backgrounds, lifestyles and environmental exposures leads to many confounding signals that have the potential to confuse the analysis. Interpretation of analysis results requires input from clinical scientists who have built up expertise through a long history of studying and treating the disease. Yet analysis of genome-wide case-control study data will, we believe, require development of compute intensive and bioinformatic intensive solutions - two areas which are not traditionally within the expertise of clinical scientists.

The need for compute & bioinformatic intensive solutions.

Many genetic case-control calculations are of relatively low computational complexity. For example, calculation of a $\chi^2$, or trend test statistic, at a single locus. Within a genome-wide association study this process is repeated at the hundreds of thousands of loci afforded by current genotyping technologies. On top of this, the analysis process is often repeated many thousands of times, e.g. 10000 times, over randomized datasets in an attempt to obtain estimates of Type-I error rates. Obviously more complicated analysis algorithms or tasks, such as study of gene-gene or gene-environment interactions will require even more compute power if done on any large scale. Similarly, interpretation of ‘hits’ identified from the statistical analysis requires retrieval of multiple annotation data and/or running of secondary bioinformatic calculations.

Whilst scientific workflow authoring systems such as Taverna [3] and Kepler [4] aim to make the bioinformatic analyses more accessible, transparent, and re-useable by non-bioinformaticians, it is recognised that there is a long way to go to put workflows in the hands of `mere mortals' [5]. Similarly the use of High Performance Computing (HPC) clusters or compute Grids is currently a niche skill. Till then we are faced with the option of including extra layers of technical staff between the raw data and the clinical researcher. To enable genuine translational research it is better to put the analysis tools directly in the hands of the clinical scientist. We do not aim to achieve this by training medical specialists in HPC technologies or bioinformatics. Instead, we aim to develop a Workbench that allows clinical scientists to easily access HPC codes for analysis of case-control data, whilst hiding as much of the infrastructure detail as possible.

To help shape the development of our analysis system we chose to form a local user-group that would act both to inform us about user expectations and to
provide feedback on the usability of the system. We attempted to cover the spectrum of users, from those already familiar with statistical genetics analysis software to those who were not. Specifically our local user-group currently consists of two experienced statistical geneticists, two clinicians with specialist interest in the development of childhood asthma, and two clinicians with research interests in Type-2 Diabetes Mellitus.

From our meetings with our local user-group it became clear that,

- Any analysis platform should have a simple user-interface, similar in look-and-feel to other applications they commonly used.
- Aspects of the HPC infrastructure should be completely hidden (as much as is possible) from the user.
- SNP analysis results should be automatically annotated to display relevant biological information alongside the statistical analysis output.

In the following sections we describe the implementation of our analysis Workbench, from the user-interface and pipeline that a user follows, to the parallelized analysis codes that power the analysis.

Architecture & Implementation of an analysis Workbench

The Shared Genomics platform comprises a Workbench [6] that has been implemented with C#.NET Windows Forms. The user navigates through the Workbench by clicking on one of five main menu items; Upload File Set, Create Data Set, Submit Job, Analyse Results and Browse Annotation Sessions – an example of which is shown in Figure 1. To complete a piece of analysis from start to finish the user must operate the forms in this order. However, they can go back at any point in the Workbench to a previous form and try out different analysis parameters or source data. Any File Sets, filtered Data Sets, Analysis Results and Viewed Annotation History generated by the user are saved to a database, thereby allowing users to shutdown the Workbench without losing important reusable analysis components.

A schematic of the analysis pipeline that a user follows is shown in Figure 2. Starting with the ‘Upload File Set’ Form, the user supplies the Workbench with a genotype data file (exported as text, e.g. from Illumina BeadStudio or BC SNP Max), phenotype data and schema. Once the user has successfully logged in, validated there are no errors with their schema and chosen to resolve any missing markers, they can then upload their file set to the server via Microsoft Background Intelligent Transfer System (BITS) to only consume idle bandwidth. The user is alerted by email when their file set has been uploaded and parsed on the server. They can then utilise functionality within the ‘Create Data Set’ Form to make appropriate filter choices, e.g. adherence to Hardy-Weinberg equilibrium or minimum levels for the minor allele frequency, as part of their Quality Assurance process. The filter routines are processed in parallel on the HPC infrastructure, speeding up the time that the filter results and metrics are returned, allowing the user to experiment in real time with different filter thresholds. Once generated, a suitable filtered data set can be saved, ready to be analysed in the ‘Submit Job’ form. Details about the data set are provided to act as a reminder about its contents and to aid selection. The user can then select the job parameters and submit it for analysis on the HPC cluster and will receive an email alert when it completes. The ‘Analyse Results’ Form enables job history to be reviewed and to open job outputs in a grid view, a close-up of which is shown in Box 7 of Figure 2. In each row of the grid an rs SNP identifier is displayed along with its associated statistical analysis results, e.g. test-statistic values together with the large sample approximation and empirical p-values.

![Figure 1. Example of the Workbench user-interface. Data set preparation and statistical analysis is performed by following the pipeline indicated by the menu items on the left.](image-url)
Figure 2: An operational overview of the Shared Genomics analysis platform detailing the main processes, data stores, data flows and user interaction points. The process numbers indicate the normal sequence of operation and each process is marked as originating either on the client workbench or on the server infrastructure.
As well as displaying statistical results for each rs identifier we also display biological annotation data, such as precise genomic location, gene name if the SNP resides within a gene, existing biomedical literature for that gene, biochemical pathways involving the gene etc. These are viewed by right-clicking on the green square within each cell of the grid. The close-up in Box 7 of Figure 2 shows an example of the menu of annotation data available for a particular locus. Clicking on one of the annotation menu items will display the annotation, for example taking the user directly to the UCSC Genome Browser or the specific KEGG pathway. For the repetitive task of annotating the statistical results our aim was to retrieve relevant biological annotation relevant to each particular SNP in an XML format through the invocation of a high volume of Taverna workflows [3]. After each workflow completes, the XML output would be parsed and displayed in real time to the user. An example Taverna workflow used is shown in Figure 3. As well as retrieving annotation data directly associated with the SNP or its immediate genomic region, use of the Taverna enactment engine allows us to easily chain workflows and services together to enhance the annotation data we present to the user. For example, literature abstracts retrieved are passed to a text-mining web-service to automatically highlight technical and scientific terms within them. We also retrieve, from OMIM [7], known diseases or conditions associated with any gene in which the SNP resides, as well as any other genes and their pseudonyms that are also associated with those diseases. In this way we make it easier for the clinical researcher to identify novel patterns and possible points of intersection with other conditions, data sets or studies.

![Figure 3. Example of a Taverna workflow used to annotate the statistical results within the Workbench.](image)

Although each row of the results grid provides information on a single SNP, data from multiple SNPs can also be analysed. For example, the Workbench allows for the calculation of the frequency of KEGG pathway terms within a user-selected set of SNPs from the results grid. In this way enrichment of particular pathways in SNPs statistically significantly associated with clinical outcome can be detected, aiding identification of relevant biochemical pathways and physiological processes. This approach can easily be extended to other controlled vocabularies, e.g. to detect enrichment of Gene Ontology terms within statistically significant SNPs.

**Implementation of Statistical Analysis Routines**

Ultimately, powering the Workbench is a set of statistical analysis routines from which the user selects and which run on a chosen cluster or Grid compute resource. For the statistical analysis codes we took as our starting point an existing codebase for the analysis of genome-wide case-control data. In this instance PLINK was chosen as our initial our codebase [8]. PLINK provides a number of routines for analysis of case-control data, is written in C++ and is freely available.

Adaptation of the PLINK codebase proceeded via a two stage process. Firstly, a thorough review of the PLINK codebase was undertaken to identify areas where the initial codebase could be improved before parallelization. Secondly, parallelization of our modified PLINK codebase was performed by inclusion of appropriate Message Passing Interface (MPI) calls. Given MPI shares memory/data between nodes, a simple memory management model was preferred over C++/STL. Therefore the decision was taken to write our modified PLINK codebase in ANSI C. By writing in ANSI C we also gain an increase in portability of code, a distinct advantage when we ultimately wish to deploy the modified code over multiple clusters. Code was further modified by simplification of the application logic and I/O code whilst retaining the bulk of the code relating to the statistical analysis algorithms. Consequently we have achieved approximately a one-third reduction in code footprint and a 2x speed-up compared to the original PLINK codebase even before parallelisation. Results from PLINK and our codes were checked to assess consistency and yielded results that were identical to within machine precision. We have also achieved numerical consistency with PLINK for both 32-bit and 64-bit platforms.

For the most part parallelisation of the genome-wide calculations is straight forward, with parametric
sweeps over SNPs or pairs of SNPs providing a obvious starting point for parallelisation. The scientific calculation is thus, “embarrassingly” parallelisable, with virtually no communication between nodes required [9]. Consequently near linear speed-up of calculations should be achievable when running on a cluster. For example, the improved single-core performance of our modified code is apparent in Figure 4, which shows performance timings of our parallelised code when running an epistasis calculation for a number of SNPs, on different numbers of cores. The timings shown in Fig.4 are for a single data set. To obtain empirical Type-1 error estimates would take proportionately longer, and so it is clear that running the original single-core PLINK code over any sizeable number of SNPs would be prohibitive.

**Figure 4.** Performance of our HPC analysis codes, bencharked against the original PLINK code and run on multiple cores.

**Discussion & Conclusions**

The advent of large scale genetic data sets in the form of genome-wide association studies presents new informatics challenges. The primary stakeholders for these new data sets, i.e. clinical researchers, will be comparative newcomers to bioinformatics, let alone HPC. Rather than increase the need for extra technical expertise on any research team that wants to utilise the emerging genome-wide association technology, we would argue that it is better to develop analysis solutions that can used directly by clinicians in a simple and intuitive manner. The Workbench we have developed provides such a solution, through user-friendly access to a parallelized version of an existing GWA software package. Whilst

the need for familiarity, on the part of the clinician, with some aspects of SNP analysis cannot be completely removed, by providing automatic biological annotation of statistical analysis results we hope to ease the interpretation of the statistical analysis hits and ultimately enhance the generation of novel hypotheses or putative causal mechanisms. The Workbench client can be downloaded from the Shared Genomics website [6] and the HPC services accessed from a fixed IP address. We also aim to publish our modified PLINK source code through suitable open source code repositories.

**Acknowledgements**

We would like to thank Microsoft for funding of the Shared Genomics project.

**References**

1. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007:447:661-678.
2. D.A. Trégouët et al. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nature Genetics*, 2009:41:283.
3. T. Oinn et al. Taverna: Lessons in creating a workflow environment for the life sciences. Concurrency and Computation: Practice & Experience. 2006:18:1067-1100.
4. B. Ludäscher et al. Scientific workflow management and the Kepler system. Concurrency and Computation: Practice & Experience. 2006:18:1039-1065.
5. T. McPhillips, S. Bowers, D. Zinn, and B. Ludäscher. Scientific workflow design for mere mortals. Future Generation Computer Systems. 2008. doi:10.1016/j.future.2008.06.013.
6. [www.nibhi.org.uk/sharedgenomics/default.aspx](http://www.nibhi.org.uk/sharedgenomics/default.aspx)
7. OMIM : Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim/
8. S. Purcell et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 2007: 81:559-575.
9. I. Foster. Designing and building parallel programs: concepts and tools for parallel software engineering. Addison-Wesley, Reading, MA,USA.1995.