GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data

Karol Estrada1,2, Anis Abuseiris3,4,5,†, Frank G. Grosveld4, André G. Uitterlinden1,2, Tobias A. Knoch3,4,5,* and Fernando Rivadeneira1,2,∗
1Department of Internal Medicine, 2Department of Epidemiology, 3Biophysical Genomics & Erasmus Computing Grid, 4Department of Cell Biology, Erasmus MC, Dr. Molewaterplein 50, 3015GE Rotterdam, The Netherlands and 5Biophysical Genomics, Genome Organization & Function, BioQuant/German Cancer Research Center, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

Received on June 23, 2009; revised and accepted on August 13, 2009
Advance Access publication August 28, 2009

ABSTRACT

Summary: The current fast growth of genome-wide association studies (GWAS) combined with now common computationally expensive imputation requires the online access of large user groups to high-performance computing resources capable of analyzing rapidly and efficiently millions of genetic markers for tens of thousands of individuals. Here, we present a web-based interface—called GRIMP—to run publicly available genetic software for extremely large GWAS on scalable super-computing grid infrastructures. This is of major importance for the enlargement of GWAS with the availability of whole-genome sequence data from the 100 Genomes Project and for future whole-population efforts. Contact: ta.knoch@taknoch.org; f.rivadeneira@erasmusmc.nl

1 INTRODUCTION

By 2008 more than 150 associations between common genetic variants and human complex traits and disease have been successfully identified through the use of GWAS (Altshuler et al., 2008). It rapidly became evident that very large sample sizes are required to detect variants with modest genetic effects (e.g. a study requires ∼8000 samples to have 90% of power to find genetic variants with a frequency of 0.20, an odds ratio of 1.2 and a genome-wise significance of 10−8). Such study sizes are achieved by meta-analysis of data shared collaboratively in consortia analyzing 100’s of traits in greater than ∼40 000 individuals (e.g. Psaty et al., 2009). Since they use different genotyping platforms (e.g Affymetrix, Illumina), imputation of millions of markers from a reference (e.g. a HapMap population) is required (de Bakker et al., 2008; International HapMap Consortium et al., 2007). Statistical methods as linear or logistic regressions measure marker wise the actual association of the genetic variants with quantitative and dichotomous traits. Freely available software like MACH2QTL/DAT (Li et al., 2006), SNIPTEST (Marchini et al., 2007) or ProbABEL (Aulchenko et al., 2007) perform similarly well for these analyses and allow trivial parallelization for distributed computing: the computation time on a regular computer for one continuous trait (∼2.5 million markers, ∼6000 samples) is currently ∼6h. Assuming linear scaling future studies with ∼50 million markers from genome sequencing in 10^3–10^5 samples and even low (1%) allele frequencies can result in approximately ∼85–850 days of analysis. Thus, secure, fast accessible web services and scalable high-performance computing grid infrastructures as the Erasmus Computing Grid (de Zeeuw et al., 2008) or the German MediGRID (Krefting et al., 2008) are required to make this analysis feasible.

Here, we present a web-based interface and application to run publicly available genetic software for extremely large GWAS on such super-computing grid infrastructures. Consequently, we provide a solution to analyze GWAS in very large populations.

2 IMPLEMENTATION

To achieve high-speed result delivery, the work is split and distributed on different grid processors by trivial parallelization depending on the total data amount. The complete system consists of (i) the user remote access computer; (ii) a web server with user webservices and a data/application database; (iii) a submit machine with a job handler and a grid resource database; and (iv) grid resources with head nodes and execution nodes. The implementation consists of a hardened Linux system, which has a hardened apache2 web server and a PostgreSQL database. Php is used for the web site and the job-handler is scripted in Perl. Concerning security, data transmission is encrypted and complete user separation is applied. Currently, the system administrator manages user accounts and monitors user access, job status and statistics. He also uploads the GWA imputed data to all available grid head nodes for each genotyped cohort, since it is of large size and the same for all cohort phenotypes. Thus, only the phenotype information has to be uploaded by the GRIMP user to the system, which controls the detailed workflow (Fig. 1).

2.1 User package submission

After logging into the system the users manually specify the analysis details: they label the analysis and select a regression model (currently, linear and logistic models), dataset and optionally

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
The job handler checks every 5 min the database for sent jobs and with other and filler jobs. Speed delivery the individual jobs have highest priority compared in principle any grid driving middleware can be used here. For high-grid infrastructures. Currently, we use here the Globus toolkit, but grid middleware will handle the jobs of the package for these specific queued. Thereafter, the job handler creates a submit file and packages overflow, each head node has a predefined amount of jobs that can be into jobs to be distributed to an individual grid part. To avoid queue is sent.

head node including the uploaded package and a failure notification and the package on the head node is removed. In case of complete of a package are completed, the results are uploaded to the database failed job is resubmitted up to three times. After all individual jobs through the middleware on the specific grid head node. An individual verifies the current status of the individual jobs distributed to a CPU

2.5 Package post-processing and notification

Once all jobs of a package were finished, all individual result files are combined into one file together with additional marker annotations such as chromosome, position, allele frequency, sample size and quality of the imputed markers. The results are archived in the database for later analysis and the result files are compressed to save disk space. Depending on the choice of notification the user is now informed—e.g. by email.

3 RESULTS AND CONCLUSIONS

Through a web-based interface the successful implementation of GRIMP allows to use publicly available genetic software for very large GWAS on scalable super-computing grid infrastructures such as the Erasmus Computing Grid or the German MediGRID within an hour. The analysis of ∼2.5 million markers and ∼6000 samples now takes ∼12 min in contrast with ∼6 h. For 10^7 markers and ∼10^8 samples, we achieve ∼10–20 min, in contrast with ∼400 h, i.e. ∼17 days for a single CPU. Thus, GRIMP will improve the learning curve for new users and will reduce human errors involved in the management of large databases. Consequently, researchers and other users with little experience will largely benefit from the use of high-performance grid computing infrastructures. Since each Grid infrastructure has different middleware setups, adjustments might be needed for each particular GRIMP implementation. Currently, we have successfully setup GRIMP for the Rotterdam Study, an prospective population-based cohort study of chronic disabling conditions in >12 000 Dutch elderly individuals (http://www.epib.nl/ergo.htm; Hofman et al., 2007). Thus, with its user-friendly interface GRIMP gives access to distributed computing to primarily biomedical researchers with or without experience, but with extreme computational demands. This is of major importance for the enlargement of GWAS with the availability of whole-genome sequence data from the 1000 Genomes Project and for future whole-population efforts.

ACKNOWLEDGEMENTS

We thank Luc V. de Zeeuw, Rob de Graaf (Erasmus Computing Grid, Rotterdam, The Netherlands) and the National German MediGRID and Services@MediGRID, German D-Grid for access to their grid resources.


Conflict of Interest: none declared.

REFERENCES


2751


K.Estrada et al.

2752