BMB
reports

BioSMACK: a linux live CD for genome-wide association analyses

Chang Bum Hong[1,#], Young Jin Kim[1,2,#], Sanghoon Moon[1,#], Young-Ah Shin[1,#], Min Jin Go[1], Dong-Joon Kim[1], Jong-Young Lee[1] & Yoon Shin Cho[1,]*

[1]Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Chungcheongbuk-do 363-951, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea

Recent advances in high-throughput genotyping technologies have enabled us to conduct a genome-wide association study (GWAS) on a large cohort. However, analyzing millions of single nucleotide polymorphisms (SNPs) is still a difficult task for researchers conducting a GWAS. Several difficulties such as compatibilities and dependencies are often encountered by researchers using analytical tools, during the installation of software. This is a huge obstacle to any research institute without computing facilities and specialists. Therefore, a proper research environment is an urgent need for researchers working on GWAS. We developed BioSMACK to provide a research environment for GWAS that requires no configuration and is easy to use. BioSMACK is based on the Ubuntu Live CD that offers a complete Linux-based operating system environment without installation. Moreover, we provide users with a GWAS manual consisting of a series of guidelines for GWAS and useful examples. BioSMACK is freely available at http://ksnp.cdc.go.kr/biosmack. [BMB reports 2012; 45(1): 44-46]

INTRODUCTION

A genome-wide association study (GWAS) is an approach that involves rapidly scanning markers across the whole genomes of many people to find genetic variations associated with a particular phenotype. Such studies are particularly useful in identifying genetic variations that contribute to common complex diseases, such as diabetes, cancer, and heart disease.

However, analyzing millions of genotype data points requires immense computing power and a highly skilled specialist for handling large amounts of data and series of analyses. To overcome the problems in conducting a GWAS, a variety of software has been developed for comprehensive analysis of millions of genotypes. Well known applications such as PLINK (1), Eigensoft (2), STRUCTURE (3), and SnpMatrix (4) provide researchers with a series of analytical tools for genome-wide scans. These tools give researchers the ability to be more creative in attempting complex and time-consuming analyses. However, as the number of software packages for GWAS grows, the complexity of pipelined software and the standardization of operating systems across computers also become more complex. Researchers often encounter these issues in the process of compiling, installing, and configuring software for their computers. Users have to configure the environmental parameters and library requirements according to the user's operating system and machine. The strategies required to solve this problem have become increasingly repetitive, error-prone, and time-consuming.

A possible solution to this problem is the development of an integrated system environment comprising an operating system, fully optimized software, and a configuration-free research environment. A live CD of such an integrated research environment would be extremely portable and bootable on any computer for trouble-free loading of software needed for a study. Previously, various packages have been introduced to provide the proper research environment and are built on existing Linux distributions such as BioLinux (NERC Environmental Bioinformatics Centre, http://nebc.nox.ac.uk/tools/bio-linux), Open Discovery (5), GRIMP (6), BioConductorBuntu (7), PhyLIS (8), and BioLinux. These packages were constructed for providing the usual bioinformatics software. BioconductorBuntu is a Linux distribution designed for simplifying implementation of microarray analysis using bioconductor packages which are a set of libraries containing an R statistics package. The purpose of PhyLis is to provide the research environment needed for Phylogenetics and Phyloinformatics. GRIMP consists of the software needed for handling large-scale genome-wide association analyses on imputed genotype data. However, no previous software has been developed for the integrated research environment of GWAS. Although GRIMP provides a set of tools for GWAS, GRIMP focuses on the grid computing needed for handling large amounts of im-

*Corresponding author. Tel: +82-43-719-8871; Fax: +82-43-719-8908; E-mail: yooncho33@korea.kr
#These authors contributed equally to this work and are listed in alphabetical order.
http://dx.doi.org/10.5483/BMBRep.2012.45.1.44

Received 5 August 2011, Revised 29 September 2011, Accepted 29 September 2011

Keywords: GWAS, Human genome, Linux, Live CD, SNP

puted genotype data and thus does not incorporate many of the packages that are standard tools for GWAS analysis. Because GWAS methods are currently experiencing rapid development, there is a need for a software package focusing specifically on GWAS.

RESULTS AND DISCUSSION

We implemented a research environment for GWAS in BioSMACK, including 4 software packages for quality control and population stratification analysis (PLINK, EIGENSOFT, STRUCTURE, and R), 2 packages for meta-analysis (R-meta and METAL), and 2 packages for genotype imputation (IMPUTE and MACH) (Table 1). Moreover, BioSMACK supports the user with a research environment wrapped in a user-friendly graphical interface. Our modified Linux distribution has the advantages of portability and feasibility. The extreme portability of BioSMACK makes it particularly useful for educational purposes and simple analyses conducted on the fly without installation and configuration. For example, we were able to use BioSMACK on various kinds of laptops and netbooks in the 5th workshop of the Asian Institute in Statistical Genetics and Genomics. Moreover, the fully functional research environment of GWAS can be set up on any computer within a couple of hours. Although the computing performance depends on the power and types of processors in the machine in use, users can analyze GWAS data from tens of thousands of individuals if sufficient memory space in available.

Our first release mainly focuses on building a genome-wide association research environment. Issues concerning computing power still need to be resolved. Nonetheless, this problem can be overcome via a cloud computing system which is supported by the Ubuntu Server Edition with a built-in open source cloud. BioSMACK can be used in a cloud system as an on-demand virtual system with fully implemented genome-wide analysis software. BioSMACK is extremely portable and provides a complete research environment. BioSMACK will be updated regularly with up-to-date software.

To reduce the burden for the geneticist conducting GWAS, we developed BioSMACK to provide a research environment for GWAS with no configuration required and easy usage. BioSMACK is based on the Ubuntu Live CD, which offers a complete Linux-based operating system environment without installation. Furthermore, it contains a series of analysis tools such as PLINK, EIGENSTRAT, STRUCTURE, and the SnpMatrix package for genome-wide scanning of data. We also provide users with a GWAS manual that consists of a series of guidelines for GWAS and useful examples. Through the integrated research environment provided by BioSMACK, users can complete GWAS without certain difficulties and a large commitment of time. Our application will be very useful for investigators interested in GWAS.

MATERIALS AND METHODS

Implementation

BioSMACK is entirely based on open-source software which is free to use and later, redistribute under a GNU General Public License (GPL). BioSMACK is based on the popular and user-friendly Ubuntu Linux distribution (based on Ubuntu v5.5), and is fully functional on conventional desktop PC systems. The distribution comes with most commonly used GWA software pre-compiled, installed and configured, which allows software to be executed by simply typing the appropriate command (Table 1) or Java swing based graphic user interface (Fig. 1). BioSMACK also contains scripting languages such as Perl, Python, and R.

Installation

BioSMACK is capable of booting from USB flash drives or a live CD, and can be installed on a hard disk. First, it can be booted from CD/USB flash drives without making changes to the underlying operating system. Second, BioSMACK can be installed from the live CD. The live CD allows for a graphical installer that allows for a new installation (erasing the previous operating system). The BioSMACK has been tested on most 32-bit and 64-bit PCs and the Apple MacBook (Apple computers that use Intel processors).

Software packages

PLINK: The majority of genome-wide association studies (GWAS) have been performed using PLINK, which is a basic, open-source C/C++ toolset used to perform large-scale whole-genome analy-

Table 1. GWA software packages and commands used to call functions

Software package	Command	Software URL
PLINK	Plink	http://pngu.mgh.harvard.edu/~purcell/plink/
EIGENSOFT	Eigenstrat	http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm
STRUCTURE	Structure	http://pritch.bsd.uchicago.edu/software.html
R(RMETA)	R	http://cran.r-project.org/web/packages/rmeta/index.html
METAL	Metal	http://cran.r-project.org/web/packages/rmeta/index.html
IMPUTE	Impute	https://mathgen.stats.ox.ac.uk/impute/impute.html
MACH	Mach1	http://www.sph.umich.edu/csg/yli/mach/tour/imputation.html

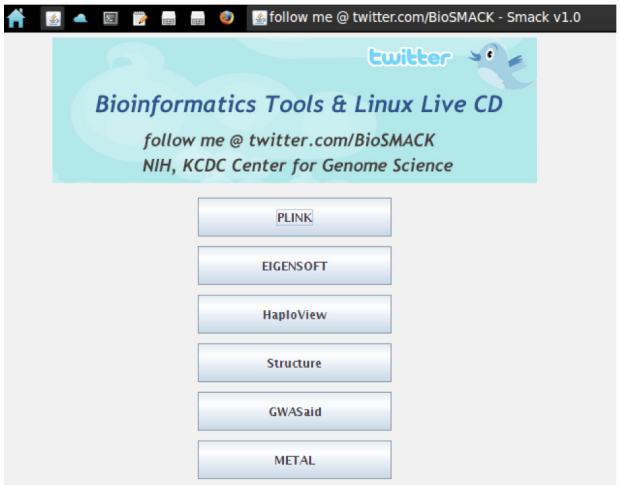

Fig. 1. Java swing based graphic user interface.

ses in a computationally efficient manner (1).

SnpMatrix: SnpMatrix is an R package that provides data classes and methods for facilitating the analysis of genome-wide association studies (4).

EIGENSTRAT: EIGENSTRAT is software used to detect the sample structure and correct spurious associations (i.e., false positives) derived from population stratification in genome-wide association studies, and has been primarily used in "GWAS". The method is based on principal components analysis (PCA) to explicitly model ancestry differences between cases and controls. EIGENSTRAT is implemented as part of the EIGENSOFT package (2).

Structure: Structure is a free program used to investigate population structure (3).

R (RMETA): R (http://www.r-project.org/) is used to facilitate data manipulation, calculation, and graphical display (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, and clustering etc.). R is available as free software in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems, including FreeBSD, Linux, Windows, and MacOS. RMETA in the R package is used for simple fixed and random effect meta-analysis in 2-sample comparisons and cumulative meta-analyses (http://CRAN.R-project.org/package=rmeta).

METAL: METAL provides an efficient computational tool for meta analysis of genome-wide association scans. METAL allows analyses of very large data sets and supports a variety of input file formats (9).

IMPUTE: IMPUTE is a program used to estimate the genotype of SNPs that were not observed in a Genome-wide association study (10).

MACH: MACH is also a program used to infer the genotype of unobserved SNPs in a Genome-wide association study (11).

Manual of BioSMACK

All procedures ranging from genotyping to imputation for GWAS are described in the BioSMACK manual, along with necessary options for executing each program. The manual included in BioSMACK briefly shows running examples.

Acknowledgements

This work was supported by grants from the Korea Centers for Disease Control and Prevention (090-091-4800-4845-301) and an intramural grant from the Korea National Institute of Health (2010-N73002-00), the Republic of Korea.

REFERENCES

1. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575.
2. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet* **38**, 904-909.
3. Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*. **155**, 945-959.
4. Clayton, D. and Leung, H. T. (2007) An R package for analysis of whole-genome association studies. *Hum. Hered.* **64**, 45-51.
5. Vetrivel, U. and Pilla, K. (2008) Open discovery: An integrated live Linux platform of Bioinformatics tools. *Bioinformation* **3**, 144-146.
6. Estrada, K., Abuseiris, A., Grosveld, F. G., Uitterlinden, A. G., Knoch, T. A. and Rivadeneira, F. (2009) GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. *Bioinformatics* **25**, 2750-2752.
7. Geeleher, P., Morris, D., Hinde, J. P. and Golden, A. (2009) BioconductorBuntu: a Linux distribution that implements a web-based DNA microarray analysis server. *Bioinformatics* **25**, 1438-1439.
8. Thomson, R. C. (2009) PhyLIS: a simple GNU/Linux distribution for phylogenetics and phyloinformatics. *Evol. Bioinform. Online* **5**, 91-95.
9. Willer, C. J., Li, Y. and Abecasis, G. R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191.
10. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906-913.
11. Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol* **34**, 816-834.