Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 20, No. 3, September 2020, pp. 219-226
http://doi.org/10.5391/IJFIS.2020.20.3.219

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

A Fully Automated Parallel-Processing R Package for High-Dimensional Multiple-Phenotype Analysis Considering Population Structure

Gi Ju Lee[1], Sung Min Park[1], Junghyun Jung[2], and Jong Wha J. Joo[1]
[1]Department of Computer Engineering, Dongguk University, Seoul, Korea
[2]Department of Life science, Dongguk University, Seoul, Korea

ljfis

Abstract

A typical genome-wide association study is conducted through a single-phenotype analysis of the correlation between each phenotype and genotype one at a time. Alternatively, a multiple-phenotype analysis of the correlation between multiple phenotypes and a genotype often has many advantages over single-phenotype analysis. For example, statistical power in the association test may be increased in a multiple-phenotype analysis and thus may detect small effects that cannot be identified in a single-phenotype analysis. Of the several multiple-phenotype analytical methods that have been proposed, generalized analysis of molecular variance for mixed-model analysis (GAMMA) is used to analyze many phenotypes simultaneously while considering the population structure. This method shows higher accuracy than the other methods. However, GAMMA has not been widely used because no automated and user-friendly software is available; this is also the case with most other multiple-phenotype analysis methods. In addition, the lack of a parallel-processing option, which is essential in a genome-wide-association-studies analysis, is also prevalent in GAMMA. In this study, we propose an easy-to-use R package for GAMMA called GAMMA Renew (GAMMAR) that performs multiple-phenotype analysis using parallel processing. We evaluate GAMMAR using a recently published yeast dataset to locate trans-regulatory hotspots.

Keywords: GWAS, Multiple-phenotype analysis, Population structure, R package

Received: Jul. 21, 2020
Revised : Sep. 13, 2020
Accepted: Sep. 13, 2020

Correspondence to: Jong Wha J. Joo
(jwjjoo@dongguk.edu)

1. Introduction

Genome-wide association studies (GWAS) have successfully identified many genetic variants associated with a range of phenotypes and diseases. Unfortunately, it has been reported that these variants explain only a small portion of phenotypic variations. A typical GWAS analyzes the correlation between a phenotype and genotype one at a time, an approach referred to as a single-phenotype analysis. By contrast, multiple-phenotype analysis, in which multiple phenotypes are analyzed simultaneously, has many advantages over the single-phenotype analysis. This approach can increase the statistical power of an association test and detect causal variants that single-phenotype analyses miss because of small effect sizes [1]. In addition, analyzing multiple phenotypes together may be desirable. In microbiome data analysis, multiple-phenotype analysis is often preferable as networks between taxa are

complex, and determining whether a single nucleotide polymorphism (SNP) affects a specific taxon or many taxa jointly is difficult. In expression quantitative trait loci (eQTL) analysis, multiple-phenotype analysis may help to locate regulatory hotspots, which are variants that regulate thousands of genes. For these reasons, several multiple-phenotype analytical methods have been proposed, including generalized analysis of molecular variance for mixed-model analysis (GAMMA) [1], singular value decomposition (SVD) [2], multivariate distance matrix regression (MDMR) [3], probabilistic ANAlysis of genoMic dAta (PANAMA) [4], and linear mixed-effects model-expression heterogeneity (LMM-EH) [5].

One of the challenges with GWAS is to correct for the population-structure effects in association tests. Each population carries its own genetic and social history that produces genetic correlations between individuals and can cause false-positive associations in an association analysis. Particularly in multiple-phenotype analysis, the population structure can cause a major problem as the bias in the test statistics accumulates in each phenotype [1]. Although some multiple-phenotype analytical methods correct the bias caused by the population structure, they are not applicable to more than 10 phenotypes, as the computation time increases quadratically with the number of phenotypes. Unlike the others, GAMMA is a multiple-phenotype method that is applicable to numerous phenotypes when the population structure is considered.

Despite the advantages of GAMMA, it has not been widely used. One of the main problems with multiple-phenotype analytical methods that utilize complex statistical models (including GAMMA) is that automated software is not provided. The complicated installation of multiple-phenotype analytical methods and the in-depth computational knowledge required to run them represent main bottlenecks in their usage. GAMMA requires multiple running steps. In addition, for each step, GAMMA uses a different programming language for which users with backgrounds in biology or genetics must manually install the required libraries with each version. Another problem with GAMMA is that it does not provide parallel processing. With the advent of high-throughput technology, the quantity of genetic data is growing every day, and parallel processing is essential for analyzing this large quantity. Moreover, GAMMA uses a permutation test to compute the p-value, which is often used in analyses with complex statistical models, and more than 10^4 permutations are impossible in genome-wide level analyses in practice because of the computational cost.

We developed a parallelism-enabled fully automated software in the widely used R language that we call GAMMA-Renew (GAMMAR).

2. Related Work

2.1 GWAS and eQTL

High-throughput technologies such as DNA microarray and next-generation sequencing technology have enabled us to conduct GWAS and eQTL analyses. These methods require an efficient analytical method to examine large datasets.

In genetics, eQTL and GWAS involve observational studies of a set of genetic variants in different individuals. When genetic variants are associated with a specific trait or gene expression, researchers generally focus on the association between SNPs and traits such as those found in major human genetic conditions and diseases.

2.2 Population-Structure Effects in GWAS

GWAS has reported the existence of a variety of hidden factors such as unobserved covariates, genetic associations, and environmental factors. These confounding factors can lead to complex dependencies between individuals and result in false positives. Many researchers have reported that population structures (one of the leading confounding factors) produce many false associations in GWAS [1, 6–15].

GWAS examines the association between the minor allele frequency of the SNP and the gene expression or disease condition to predict the association. However, not only do disease-causing SNPs produce differences in terms of the frequency of antagonistic genes, but SNPs affected by ancestry can also cause disease [6–15]. This is because the frequency of antagonistic genes varies from population to population due to the unique genetic and social histories of these populations.

2.3 GAMMA

A typical GWAS and eQTL analysis examines the correlation between one phenotype and one genotype at a time. However, a single-phenotype analysis can miss the unmeasured aspects of a complex biological network. Analyzing many phenotypes simultaneously may capture these unmeasured aspects and detect more variations. Although multiple-phenotype methods aim to detect variations associated with more than one phenotype, previous methods do not consider the effects of population structures, which can result in a significant number of false positives. GAMMA, which is an efficient and accurate

multiple-phenotype analysis method, considers the population structure and can be applied to numerous phenotypes.

3. Methods

GAMMAR is a fully automated program that automatically sets all necessary environmental variables when the R package is downloaded. The GAMMAR package uses the R language and requires version 3.5.0 or later. In addition to the default R packages, it also uses external packages such as lmmlite, doParallel, and foreach. These packages are automatically installed and configured when the GAMMAR program is installed.

3.1 Overview of the GAMMAR Program

GAMMAR is a fully automated multiple-phenotype analysis R package that is built on the GAMMA program and uses a linear mixed model to consider the population structure. It computes kinship, which contains the correlation of genotypes between samples, and estimates the variance components of the data by fitting them into a linear mixed model (details are given in the following sections). After correcting for the effects of population structure in the data, GAMMAR computes pseudo-f statistics to compute the associations between the given phenotypes and a single genotype. Because the results are pseudo-f statistics, GAMMAR uses permutations to compute the p-values (Figure 1 shows an overview of the GAMMAR package). It obtains genotypes, phenotypes, and user options such as the number of multiprocessors to run in parallel processing; otherwise, it identifies the number of permutations to perform as the input. First, it runs the Kinship function to calculate the kinship coefficients. It then runs the varComp function to calculate the variance components by fitting the data into a linear mixed model. Finally, it runs the GAMMAR function to return p- and f-values as the results. N_1 indicates the number of processors with which to run the program, and N_2 denotes the number of permutations used to compute the p-values.

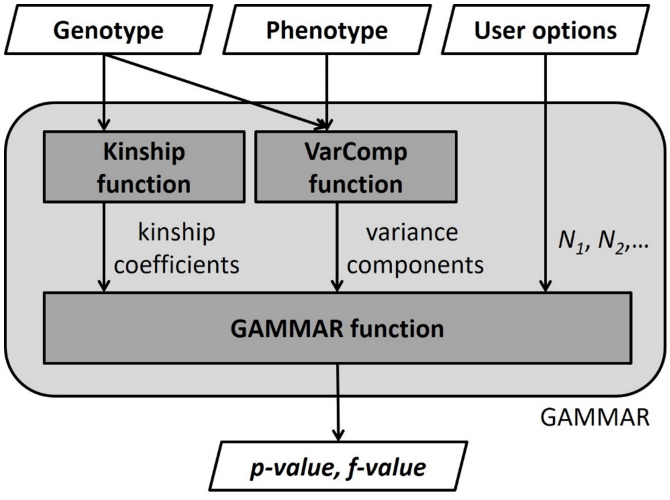


Figure 1. Overview of the GAMMAR package.

3.2 Population-Structure Correction Model in GAMMA

It is widely known that the population structure confounds the association analysis in GWAS, thus inducing spurious associations [1, 6–16]. A typical GWAS uses the following linear model to test the associations:

$$y_j = X_i\beta_j + e_j. \tag{1}$$

Let n be the number of individuals and m be the number of genes. In Eq. (1), y is an $n \times m$ matrix, where each column vector y_j is a vector of length n with jth phenotype values, X_i is a vector of length n containing ith SNP values, β_j is a value that includes the effect of the ith SNP on the jth phenotype, and e_j is a vector of length n with independent and identically distributed (i.i.d.) residual errors of the jth phenotype. Here, $e_j \sim N(0, \sigma^2_{g_j} I)$, where I is an n by n identity matrix with unknown magnitude $\sigma^2_{e_j}$. Under the assumption of a linear model, each phenotype follows a multivariate normal distribution with mean and variance given as $y_j \sim N(X_i\beta_j, \Sigma_j)$.

Recently, the linear mixed model has emerged as a powerful tool in GWAS that considers the population structure in the association test as follows:

$$y_j = X_i\beta_j + u_j + e_j, \tag{2}$$

where u_j is a vector of length n that contains the effects of the population structure of the jth phenotype ($u_j \sim N(0, \sigma^2_{g_j} K)$), K is the kinship matrix that encodes the relatedness between the individuals, and ${\sigma_{g_j}}^2$ is the variance of the phenotype accounted for by the genetic variation under the linear mixed model $y_j \sim N(X_i\beta_j, \Sigma_j)$, where $\Sigma_j = \sigma^2_{g_j} K + \sigma^2_{e_j} I$.

GAMMA computes variance components for each phenotype and uses median values of $\widehat{\sigma}_{g_j}$ and $\widehat{\sigma}_{e_j}$ to compute $\widehat{\Sigma} = \widehat{\sigma}^2_{g_j} K + \widehat{\sigma}^2_{e_j} I$. The square root of $\widehat{\Sigma}$ is used to transform the genotypes and phenotypes ($\widehat{\Sigma} y_j \sim N(\widehat{\Sigma}^{-1/2} X_i\beta_j, \sigma^2 I)$) to make the data i.i.d. (Figure 2).

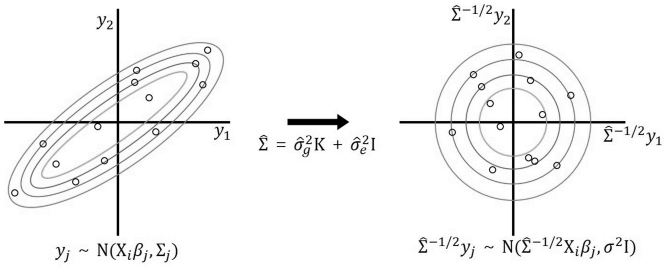


Figure 2. Population stratification correction. The left graph shows when the data containing a population structure affects variance $\hat{\Sigma}$ The data are i.i.d. after the genotypic and phenotypic values are transformed using $\hat{\Sigma}^{-1/2}$.

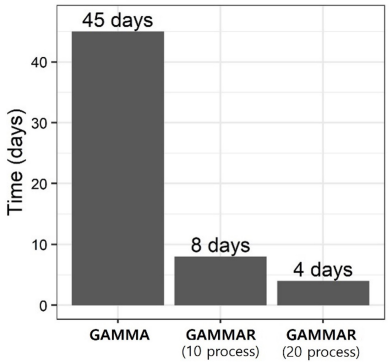


Figure 3. Performance comparison of GAMMA and GAMMAR in terms of computational time.

3.3 GAMMAR Implementation

GAMMAR was written in R language version 3.5.0 and is offered under the GNU Affero General Public License version 3 (AGPL-3.0; https://www.gnu.org/licenses/why-affero-gpl.html).

3.3.1 Variance components estimation

Variance components estimation involves implementing the linear mixed model as given in Eq. (2) and computing the variance components of the data (σ_g and σ_e). After estimating the variance components by fitting the data into the linear mixed model, we use the variance components to correct for the effects of population structure by transforming the genotypes and phenotypes as previously described.

3.3.2 Parallel processing

GAMMA performs a permutation test to compute p-values, which consumes considerable time. To reduce the running time, GAMMA uses an adaptive permutation. However, this still consumes considerable time, and performing permutations of more than 10^4 at the genome-wide level in practice is impossible, even when running on a high-performance server. In other words, GAMMA cannot provide p-values of less than 10^{-4}, which is sufficient for GWAS considering the fact that standard GWAS requires p-values of less than 10^{-8}. GAMMAR allows multiprocessing in a user-friendly manner that does not require that the data be divided into small windows for distribution into clusters. Users may specify the number of processes to use for parallel processing.

3.3.3 Genomic control

Genomic control is a commonly used statistical method to correct the confounding effects of population stratification in genetic association studies. We provide this as a function in GAMMAR.

4. Results

4.1 Performance Analysis

We evaluated the performance of GAMMAR by comparing it with GAMMA in analyzing a previously generated yeast dataset that contains $1,012$ segregants with $5,720$ genes and $42,052$ SNPs [17]. Figure 3 shows that GAMMAR reduced the execution time significantly, whereas GAMMA required much more time to configure the environments and to run the multiple-step burden. Approximately 45 days were required to analyze the data, not including the additional time required to set up the environment and pre- and post-process the data for running and transferring the data in each step. Thus, when running the programs on the same system, GAMMAR was approximately five and more than 10 times faster than GAMMA when using 10 and 20 processes, respectively. The result was based on 10^4 permutations.

4.2 GAMMAR Analysis using the Yeast Dataset

The genome of eukaryotes was first decoded in budding yeast, which is a single-celled organism [18]. For many years, yeast has been widely studied in genetics and physiology as a eukaryote model system, as it has 23% homologous genes to humans and a short lifecycle. It also is a well-annotated genome [19]. We evaluated GAMMAR with a yeast dataset to identify trans-

Table 1. Regulatory hotspot identification of the yeast dataset

eQTL Hotspot	Number of eQTLs	Number of eGenes	Putative regulators
III:150000*	9	5	**RER1,** PMP1, SLM5, NPP1, RHB1
III:190000*	14	8	**PHO87, MATALPHA1, MATALPHA2,** POF1, RPS14A, MAK31, SNT1, RRP43
XII:650000*	31	12	**LCB5,** NDL1, CCC1, AAT2, TEN1, YPS13 TIS11, MSS51, TMA7, RED1, HSP60, YPT6
XIV:360000	28	17	YSF3, SRV2, ASI3, FPR1, EAF7, NRK1, TOM22, FYV6, HRB1, CPT1, KRE33, ELA1, WHI3, PGA1, MPP6, TCB2 POL1
XIV:440000*	14	4	**TOP2,** TCB2, YPT53, RHO2
XIV:470000*	21	13	**TPM1, SAL1,** NSG2, EOS1, YAF9, SNN1, INP52, GA2, RIO2, KSH1, TOM22, NIS1, RPL9B

Putative regulators identified by previous studies are denoted in boldface; * indicates a previously discovered hotspot.

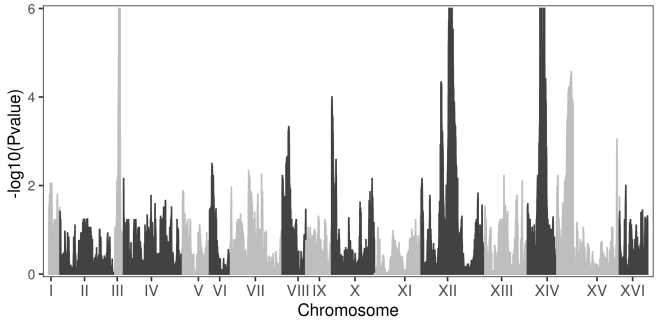


Figure 4. Regulatory hotspots in yeast. The x and y axes correspond to the SNP positions and GAMMAR p-value, respectively.

eQTL hotspots [17]. After adjusting for batch effects by using the ComBat method with the growth covariate [20], we performed GAMMAR analysis using expression levels of 5,720 genes with 42,052 SNPs. A total of 692 SNPs was determined based on the GAMMAR p-value$< 5 \times 10^{-5}$ (Figure 2). We next divided the whole yeast genome into 603 20-kb bins, and then SNPs with the smallest GAMMAR p-values were selected in each bin for comparison with the previous yeast eQTL studies [21, 22]. We determined that 117 *trans*-eQTLs had 59 eGenes on three chromosomes. In eQTL studies, genes with *cis*-acting SNP effects are referred to as eGenes [23]. Information on these eGenes was obtained from the original yeast study [1]. Collectively, we defined the six bins as *trans*-regulatory hotspots and 59 eGenes as putative regulators. Of the 59 total eGenes, nine had been previously identified [21, 22, 24] (Table 1). In four previous studies, MATing type protein ALPHA 1; III:190000 (MATALPHA1) was reported to be a casual regulator [21, 25–27], and Killer toxin REsistant 33; XIV:360000 (KRE33) was recently identified as a putative causal regulator [24].

5. Conclusion

Although multiple-phenotype analysis is advantageous over single-phenotype analysis in many respects, it has not been widely used because of certain inconveniences when applying it. GAMMA is a representative multiple-phenotype analytical method that is applicable to high-dimensional data and can correct for population-structure effects in GWAS and eQTL studies. GAMMA has flaws. For example, it requires that the necessary libraries for executing python and R programs be manually installed along with the required older versions. Another problem with the usage of GAMMA is that it lacks a parallel-processing option. It uses a permutation test for computing the p values for which parallel processing is essential. Thus, even when using high-performance servers, GAMMA cannot run more than 10^4 permutations in practice.

In this study, we provided a fully automated and easy-to-use R package called GAMMAR to solve various inherent problems with GAMMA. When the GAMMAR package is installed, all of the necessary environments are automatically installed and configured. In addition, GAMMAR allows parallel processing, which significantly increases the allowed number of permutations to reach the standard GWAS threshold of 10^{-8}. With the advent of data collection technologies, the amount of genome data has been growing daily, and many researchers have focused on multiple-phenotype analysis. We believe GAMMAR provides an efficient and user-friendly means of conducting multiple-phenotype analyses in the new era.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2017R1C1B5017497) and by the R&D program for Advanced Integrated-Intelligence for Identification (AIID) through the NRF funded by the Ministry of Science and ICT (No. 2018M3E3A1057288).

References

[1] J. W. J. Joo, E. Y. Kang, E. Org, N. Furlotte, B. Parks, F. Hormozdiari, A. J. Lusis, and E. Eskin, "Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure," *Genetics*, vol. 204, no. 4, pp. 1379-1390, 2016. https://doi.org/10.1534/genetics.116.189712

[2] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10101-10106, 2000. https://doi.org/10.1073/pnas.97.18.10101

[3] N. J. Schork and M. A. Zapala, "Statistical properties of multivariate distance matrix regression for high-dimensional data analysis," *Frontiers in Genetics*, vol. 3, article no. 190, 2012. https://doi.org/10.3389/fgene.2012.00190

[4] N. Fusi, O. Stegle, and N. D. Lawrence, "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies," *PLoS Computational Biology*, vol. 8, no. 1, article no. e1002330, 2012. https://doi.org/10.1371/journal.pcbi.1002330

[5] J. Listgarten, C. Kadie, E. E. Schadt, and D. Heckerman, "Correction for hidden confounders in the genetic analysis of gene expression," *Proceedings of the National Academy of Sciences*, vol. 107, no. 38, pp. 16465-16470, 2010. https://doi.org/10.1073/pnas.1002425107

[6] R. A. Kittles, W. Chen, R. K. Panguluri, C. Ahaghotu, A. Jackson, C. A., Adebamowo, et al., "CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification?," *Human Genetics*, vol. 110, no. 6, pp. 553-560, 2002. https://doi.org/10.1007/s00439-002-0731-5

[7] M. L. Freedman, D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, N. Patterson, et al., "Assessing the impact of population stratification on genetic association studies," *Nature Genetics*, vol. 36, no. 4, pp. 388-393, 2004. https://doi.org/10.1038/ng1333

[8] J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly, "The effects of human population structure on large genetic association studies," *Nature Genetics*, vol. 36, no. 5, pp. 512-517, 2004. https://doi.org/10.1038/ng1337

[9] C. D. Campbell, E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman, C. Groop, D. Altshuler, K. G. Ardlie, and J. N. Hirschhorn, "Demonstrating stratification in a European American population," *Nature Genetics*, vol. 37, no. 8, pp. 868-872, 2005. https://doi.org/10.1038/ng1607

[10] A. Helgason, B. Yngvadottir, B. Hrafnkelsson, J. Gulcher, and K. Stefansson, "An Icelandic example of the impact of population structure on association studies," *Nature Genetics*, vol. 37, no. 1, pp. 90-95, 2005. https://doi.org/10.1038/ng1492

[11] A. P. Reiner, E. Ziv, D. L. Lind, C. M. Nievergelt, N. J. Schork, S. R. Cummings, et al., "Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study," *The American Journal of Human Genetics*, vol. 76, no. 3, pp. 463-477, 2005. https://doi.org/10.1086/428654

[12] B. F. Voight and J. K. Pritchard, "Confounding from cryptic relatedness in case-control association studies," *PLoS Genetics*, vol. 1, no. 3, article no. e32, 2005. https://doi.org/10.1371/journal.pgen.0010032

[13] M. Berger, H. H. Stassen, K. Kohler, V. Krane, D. Monks, C. Wanner, et al., "Hidden population substructures in an apparently homogeneous population bias association studies," *European Journal of Human Genetics*, vol. 14, no. 2, pp. 236-244, 2006. https://doi.org/10.1038/sj.ejhg.5201546

[14] M. F. Seldin, R. Shigeta, P. Villoslada, C. Selmi, J. Tuomilehto, G. Silva, J. W. Belmont, L. Klareskog, and P. K. Gregersen, "European population substructure: clustering of northern and southern populations," *PLoS Genetics*, vol. 2, no. 9, article no. e143, 2006. https://doi.org/10.1371/journal.pgen.0020143

[15] M. Foll and O. Gaggiotti, "Identifying the environmental factors that determine the genetic structure of populations," *Genetics*, vol. 174, no. 2, pp. 875-891, 2006. https://doi.org/10.1534/genetics.106.059451

[16] J. Flint and E. Eskin, "Genome-wide association studies in mice," *Nature Reviews Genetics*, vol. 13, no. 11, pp. 807-817, 2012. https://doi.org/10.1038/nrg3335

[17] F. W. Albert, J. S. Bloom, J. Siegel, L. Day, and L. Kruglyak, "Genetics of trans-regulatory variation in gene expression," *Elife*, vol. 7, article. e35471, 2018. https://doi.org/10.7554/eLife.35471

[18] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, et al., "Life with 6000 genes," *Science*, vol. 274, no. 5287, pp. 546-567, 1996. https://doi.org/10.1126/science.274.5287.546

[19] D. Botstein, S. A. Chervitz, and M. Cherry, "Yeast as a model organism," *Science*, vol. 277, no. 5330, pp. 1259-1260, 1997. https://doi.org/10.1126/science.277.5330.1259

[20] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118-127, 2007. https://doi.org/10.1093/biostatistics/kxj037

[21] R. E. Curtis, S. Kim, J. L. Woolford Jr, W. Xu, and E. P. Xing, "Structured association analysis leads to insight into Saccharomyces cerevisiaegene regulation by finding multiple contributing eQTL hotspots associated with functional gene modules," *BMC Genomics*, vol. 14, article no. 196, 2013. https://doi.org/10.1186/1471-2164-14-196

[22] L, Lin, Q. Chen, J. P. Hirsch, S. Yoo, K. Yeung, R. E. Bumgarner, Z. Tu, E. E. Schadt, and J. Zhu, "Temporal genetic association and temporal genetic causality methods for dissecting complex networks," *Nature Communications*, vol. 9, article no. 3980, 2018. https://doi.org/10.1038/s41467-018-06203-3

[23] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, et al., "The genotype-tissue expression (GTEx) project," *Nature Genetics*, vol. 45, no. 6, pp. 580-585, 2013. https://doi.org/10.1038/ng.2653

[24] E. R. Jerison, S. Kryazhimskiy, J. K. Mitchell, J. S. Bloom, L. Kruglyak, and M. M. Desai, "Genetic variation in adaptability and pleiotropy in budding yeast," *Elife*, vol. 6, article no. e27167, 2017. https://doi.org/10.7554/eLife.27167

[25] G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak, "Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors," *Nature Genetics*, vol. 35, pp. 57-64, 2003. https://doi.org/10.1038/ng1222

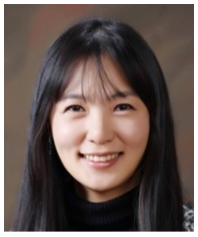
[26] J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt, "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nature Genetics*, vol. 40, no. 7, pp. 854-861, 2008. https://doi.org/10.1038/ng.167

[27] S. I. Lee, A. M. Dudley, D. Drubin, P. A. Silver, N. J. Krogan, D. Pe'er, and D. Koller, "Learning a prior on regulatory potential from eQTL data," *PLoS Genetics*, vol. 5, no. 1, article no. e1000358, 2009. https://doi.org/10.1371/journal.pgen.1000358

Gi Ju Lee received the M.S. degrees from Dongguk University, Seoul, Korea, in 2019. Currently, he has been under the Ph.D. degree candidate at the Department of Computer Science and Engineering, Dongguk University, since 2019. His research areas include developing bioinformatics tools, genome privacy and security .
E-mail: beartange3@gmail.com

Sung-min Park received the B.S. degrees from Dongguk University, Seoul, Korea, in 2019. Currently, he has been under the M.S. degree candidate at the Department of Computer Science and Engineering, Dongguk University, since 2019. His research areas include bioinformatics and life science.
E-mail: 9904trs@naver.com

Junghyun Jung received his Ph.D.in 2020 from Dongguk University, Seoul, Korea. Currently, he is a postdoctoral scholar at the Department of Clinical Pharmacy, USC School of Pharmacy, University of Southern California, since 2020. His research areas include developing bioinformatics tools and analyzing functional genomic data related to the immune system.
E-mail: junghyunjj219@gmail.com



Jong Wha J. Joo received the B.S. degree in computer science and engineering from Seoul National University, Seoul, Korea, in 2005, the M.S. degree in computer science from Brown University, Providence, RI, USA, in 2007, and the Ph.D. degree in bioinformatics from the University of California, Los Angeles, CA, USA, in 2016. She is currently an Assistant Professor with the Department of Computer Science and Engineering, Dongguk University, Seoul. Her research interests include developing efficient computational methodologies and algorithms for genome-wide association studies and expression quantitative trait loci studies.
E-mail: jwjjoo@dongguk.edu