International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 18, 2019, pp 87-98

Comparative Genome-Wide Association Studies using Canonical Correlation Analysis

Atul Chandra Singha[1,3*], Arafat Rahman[2], Md. Jahangir Alam[1] and Md. Nurul Haque Mollah[1*]

[1]Bioinformatics Lab, Department of Statistics, University of Rajshahi, Bangladesh
[2]Department of Microbiology, Noakhali Science and Technology University, Bangladesh
[3]Department of Statistics, Begum Rokeya University, Rangpur (BRUR), Bangladesh
* Correspondence should be addressed to Md. Nurul Haque Mollah (mollah.stat.bio@ru.ac.bd) and Atul Chandra Singha (atul@brur.ac.bd)

[Received June 10, 2019; Revised August 17, 2019; Accepted September 1, 2019]

Abstract

Genome-wide association studies (GWAS) are powerful tools for measuring the association between genotype-phenotype pairs in bioinformatics**.** Most of the human diseases and traits have a strong genetic architecture. GWAS is successful in identifying common genetic variants underlying complex traits or diseases like cancer, type-II diabetes, cardiovascular disease, schizophrenia and quantitative traits such as lipid levels and metabolomics. Now an important approach to GWAS is to test the association between multiple single nucleotid6e polymorphisms (SNPs) against multiple quantitative phenotypes. Canonical Correlation Analysis (CCA) is one of the most popular multivariate statistical techniques to test the association between multiple SNPs against multiple quantitative phenotypes. However, it is not robust against phenotypic contaminations. To overcome this problem, in this paper an attempt is made to robustify the CCA. To robustify the CCA, we consider some popular robust analyzers like Minimum Covariance Determinant (MCD), Minimum Variance Ellipsoid (MVE), Orthogonalized Gnanadesikan-Kettering (OGK) estimators including the Minimum β-divergence estimator. Using simulated data analysis, we observed that CCA based on Minimum β-divergence method (proposed) shows better performance than classical CCA as well as robust CCA based on MCD, MVE and OGK estimators in presence of outliers. Otherwise proposed method keeps equal performance to the classical CCA as well as robust CCA based on MCD, MVE and OGK estimators.

Keywords: SNPs, GWAS, CCA, Quantitative traits, Outliers, Minimum β-divergence method.
AMS Classification: 92D10.

1. Introduction

Genome-wide association studies (GWAS) is the most attractive area of research to demonstrate the layout of genotype-phenotype associations. It has developed the field of genetic component analysis for complex trait or disease over the past decade (Visscher et al., 2012; Visscher et al., 2017). GWAS study was first published in 2005 about the significant association of two SNPs with age-related macular degeneration problem (Klein et al., 2005). Recently, GWAS study plays an important role in identifying huge number of human disease and phenotypic traits that are strongly related to the genetic components. It has promising application in identifying genetic components for underlying complex traits or disease like cardiovascular disease (Deloukas et al., 2013), schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014), type-II diabetes (Mahajan et al., 2014), anorexia nervosa (Duncan, L., et al., 2017), major depressive disorder (Hyd, C. L., et al., 2016), cancers and subtypes of cancers (Milne, R. L., et al., 2017; Sud, A., et al., 2017), inflammatory bowel disease (de Lange, 2017), insomnia (Jasen, P. R., et al., 2019), body mass index(BMI) (Yengo et al., 2018), quantitative traits like lipid levels (Willer, C. J., et al., 2013; Surraka, I., et al., 2015) and metabolomics (Kettunen et al., 2012; Shin et al., 2014). Later the measure of association between multiple genotype and multiple phenotypes provides precise results Inouyeet al., 2012). Therefore, some complex genotype-phenotype correlations can be detected when testing several genetic components simultaneously (Marttinen et al., 2014). In 2009 canonical correlation analysis (CCA) was used to measure the association between single SNP and multiple phenotypes (Ferreira and Purcell 2009). In 2012 CCA was used to measure the associations between multiple SNPs and multiple phenotypes instead of considering the permutation test (Tang and Ferreira 2012). CCA, first developed by Harold Hotelling (1936) and showed the application to measures the relationship between two sets of multidimensional variables simultaneously by maximizing the correlation between their linear combinations. Here we robustify the CCA based on Minimum β-divergence method for more precise and robust results even in the phenotypic contaminated data. Therefore, in this study, we extended the CCA as a robust method in GWAS for identifying the relationships between multiple SNPs and multiple phenotypes and compared its performance with the other methods as well as the classical method.

2. Materials and Methodology

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote genotype and phenotype matrices of dimensions $N\times G$ and $N\times P$ respectively, where N the number of samples, G and P the number of genotypic and phenotypic variables respectively. CCA (Hotelling 1936) provides a convenient statistical framework to simultaneously detect linear relationships between two groups of variable**s** $\boldsymbol{X} \in R^{N\times G}$ and $\boldsymbol{Y} \in R^{N\times P}$ where $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent two different views of the same objects. The objective is to find maximally correlated linear combinations of columns of each matrix. This corresponds to finding vectors $a \in R^{G}$ and $b \in R^{P}$ that maximize

$$\boldsymbol{r} = \frac{(\boldsymbol{Xa})^{\boldsymbol{T}}(\boldsymbol{Yb})}{\|\boldsymbol{Xa}\|\|\boldsymbol{Yb}\|} = \frac{\boldsymbol{a^T\Sigma_{XY}b}}{\sqrt{\boldsymbol{a^T\Sigma_{XX}a}}\sqrt{\boldsymbol{b^T\sum\Sigma_{YY}\,b}}}$$

The maximized correlation r is called canonical correlation between $\boldsymbol{X}$ and $\mathbf{Y}$. Finally, for G genotypes and P phenotypes the j=min(G,P) canonical correlations are then calculated as the square root of the j eigenvalues of the canonical correlation matrix $\boldsymbol{\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}}$ where, $\boldsymbol{\Sigma_{XX}}$ and $\boldsymbol{\Sigma_{YY}}$ are the G×G and P×P covariance matrices for genotypes and phenotypes respectively while $\boldsymbol{\Sigma_{XY}}$ and $\boldsymbol{\Sigma_{YX}}$ are the between G×P (or P×G) covariance matrices.

2.1 Covariance Matrix Determination Using Robust Method

2.1.1 Orthogonalized Gnanadesikan Kettenring (OGK)

Gnanadesikan and Kettenring (1972) was proposed positive definite, and approximately affine equivariant robust scatter matrices starting from any robust scatter matrix and then applied for robust covariance estimate the resulting of multivariate location and scatter estimates are called OGK.

The steps to estimate the OGK estimators are as follows,

Let $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m}] \in \boldsymbol{R^{m\times n}}$ be a data matrix with rows $x_i^T (i = 1,2, \dots, m)$ and columns $\boldsymbol{X}_j (j = 1,2, \dots, p)$.

Step-1: Let, m(.) and s(.) be robust univariate estimators of mean and variance.

Step-2: Construct $\boldsymbol{D} = \boldsymbol{diag}\left(\boldsymbol{s}(\boldsymbol{x_1}), \boldsymbol{s}(\boldsymbol{x_2}), \dots, \boldsymbol{s}(\boldsymbol{x_p})\right)$, and define $Y = XD^{-1}$

Step-3: Compute the correlation matrix U applying s(.) to the columns of Y:

$$\boldsymbol{U} = [\boldsymbol{u_{jk}}] = \begin{cases} \frac{\mathbf{1}}{\mathbf{4}}(\boldsymbol{s}(\boldsymbol{y_j} + \boldsymbol{y_k})^{\mathbf{2}} - \boldsymbol{s}(\boldsymbol{y_j} - \boldsymbol{y_k})^{\mathbf{2}}, & j \neq k \\ \mathbf{1}, & j = k \end{cases}$$

Step-4: Compute the eigen decomposition: $\boldsymbol{U} = \mathbf{E\Delta E}^{\boldsymbol{T}}$

Step-5: Project the data onto the basis eigenvectors: $\boldsymbol{Z = YE}$

Step-6: Estimate the variances in the coordinate directions:

$$\boldsymbol{\Gamma = diag\left(s(z_1)^2, s(z_2)^2, \dots, s\left(z_p\right)^2\right)}$$

Step-7: The estimated covariance matrix is then,

$$\boldsymbol{\hat{\Sigma} = D^2 E \Gamma E^T}$$

2.1.2 Minimum Covariance Determination Estimators (MCD)

Rousseeuw (Rousseeuw 1985) introduced the minimum covariance determinant estimator (MCD) method to estimate the mean vector and covariance matrix along with outliers in multidimensional data. This method considers all subsets and then compute the determinant of the covariance matrix for each subset. The subset with the smallest determinant is used to calculate the usual mean vector, and corresponding covariance matrix, these estimators are called minimum covariance determinant estimators.

Let $\boldsymbol{X = [x_1, x_2, \dots, x_m] \in R^{m \times n}}$ be a data matrix. We define the mean vector $\boldsymbol{\mu = (\mu_1, \mu_2, \dots, \mu_n)^T}$ and covariance matrix $\sum_{n \times n}$. Usually, MCD method attempts to find out the h subset data (where $\frac{m}{2} \leq h < m$) whose sample covariance matrix determinant is minimum. Consider all $\binom{m}{h}$ subsets are $h \times n$ submatrix of $\boldsymbol{X}$ denoted by $\boldsymbol{X_H}$. The mean and covariance matrix for all the subsets $\boldsymbol{X_H}$ is defined as,

$$\boldsymbol{\mu(X_H) = h^{-1}(X_H)^T I_h}$$

and $\boldsymbol{\Sigma(X_H) = h^{-1}(X_H - \mu(X_H))^T(X_H - \mu(X_H))}$

Then MCD method aims to minimize the determinant of $\boldsymbol{\Sigma(X_H)}$ from all subsets.

i.e., $\boldsymbol{h^{MCD} = \mathbf{argmin}_{h \in X_h} det(\Sigma(X_H))^{1/n}}$

Therefore, the covariance matrix estimated by this way is called MCD estimator.

2.1.3 Minimum Volume Ellipsoid Estimator (MVE)

The Minimum Volume Ellipsoid (MVE) estimator (Rousseeuw 1985) has been studied extensively and used in the detection of multivariate outliers. The method attempts to estimate the ellipsoid of minimum volume that contains a subset of at least $\boldsymbol{h}$ data points. Subsets of size $\boldsymbol{h}$ are called half sets because $\boldsymbol{h}$ is often chosen to be just more than half of the data points. Then the location estimate is defined as the center of this ellipsoid and the covariance estimate is provided by its shape.

Let $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m}] \in \boldsymbol{R^{m\times n}}$ be a data matrix. We define the mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \dots, \boldsymbol{\mu_n})^{\boldsymbol{T}}$ and covariance matrix $\sum_{n\times n}$. Usually, MVE method attempts to find out the h subset data (where $\frac{m}{2} \le h < m$) whose volume of ellipsoid is minimum. Consider all $\binom{m}{h}$ ellipsoids are determined from subsets $h \times n$ submatrix of $\boldsymbol{X}$ **(**denoted by $\boldsymbol{X_H}$**).** Then the MVE estimator of mean and covariance matrix for all subsets $\boldsymbol{X_H}$ is defined as,

$$\{H: (\boldsymbol{X_H} - \boldsymbol{\mu}(\boldsymbol{X_H}))^{\boldsymbol{T}}\boldsymbol{\Sigma}(\boldsymbol{X_H})^{-\boldsymbol{1}}(\boldsymbol{X_H} - \boldsymbol{\mu}(\boldsymbol{X_H})) \le \boldsymbol{k^2}\} \ge \boldsymbol{h} \qquad \textbf{(i)}$$

Where, $\boldsymbol{k}$ is the fixed constant which explains the magnitude value of covariance matrix determinant and

$$\boldsymbol{\mu}(\boldsymbol{X_H}) = \boldsymbol{h^{-1}}(\boldsymbol{X_H})^{\boldsymbol{T}}\mathbf{I}_{\boldsymbol{h}},$$

and $$\boldsymbol{\Sigma}(\boldsymbol{X_H}) = \boldsymbol{h^{-1}}(\boldsymbol{X_H} - \boldsymbol{\mu}(\boldsymbol{X_H}))^{\boldsymbol{T}}(\boldsymbol{X_H} - \boldsymbol{\mu}(\boldsymbol{X_H}))$$

The standard MVE method find out the ellipsoids determined by the covariance matrix which consists (q+1) observations of $\boldsymbol{X}$**.** i.e., the index of each subset of size (q+1) is defined by $H = \{1,2, \dots, (q+1)\} \subset \{1,2, \dots, m\}$**.**

2.1.4 Minimum β- Divergence method (Proposed)

To estimate the robust covariance matrix using maximumβ-likelihood estimator (Mollah et al., 2010) we used the maximum β-likelihood estimators for the mean vector $\boldsymbol{\mu_{n\times 1}}$ and the covariance matrix $\boldsymbol{\Sigma_{n\times n}}$ obtained iteratively as follows:

$$\mu_{t+1} = \frac{\sum_{j=1}^{n} \varphi_\beta(x_j/\mu_t, \sum_t)\,(x_j - \mu_t)x_j}{\sum_{j=1}^{n} \varphi_\beta(x_j/\mu_t, \sum_t)}$$

$$\text{And} \sum_{t+1} = \frac{\sum_{j=1}^{n} \varphi_\beta(x_j/\mu_t,\sum_t)(x_j-\mu_t)(x_j-\mu_t)^T}{(1+\beta)^{-1}\sum_{j=1}^{n} \varphi_\beta(x_j/\mu_t,\sum_t)}$$

where, $\varphi_\beta(x_j; \mu_t, \Sigma_t) = exp\left\{-\frac{\beta}{2}(x-\mu_t)^T{\Sigma_t}^{-1}(x-\mu_t)\right\}$, be the β-weight function (Mollah et al., 2010).]. It produces almost zero weight for contaminated data points. The notations $\boldsymbol{\mu_{t+1}}$ and $\sum_{\boldsymbol{t+1}}$ are the update of $\boldsymbol{\mu_t}$ and $\sum_{\boldsymbol{t}}$ in the (t+1)-th iteration respectively. It should be noted here that the proposed iterative estimation of the mean vector and covariance matrix reduces to the non-iterative classical mean vector and covariance matrix defined as

$$\mu = \frac{1}{n}\sum_{j=1}^{n} x_j\ ,\ \Sigma = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)(x_j - \mu)^T$$

For β tends to zero.

2.2 Data Simulation

To investigate the performance of the proposed method (RCCA) in ~~a~~ comparison to the traditional method, we have simulated the SNP data for a hypothetical gene by considering 2000 individual populations and the phenotype data like age, sex etc. For generating the SNP data, we consider the coalescent-based approach called GENOME (Liang et al., 2007) and some phenotypes are generated by associating with SNP. Then the generated datasets are considered in a two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ for SNPs and phenotypes respectively.

3. Result and Discussions

To investigate the performance of our proposed method in comparison to classical CCA method as well as robust methods like MVE, MCD and OGK, we used our simulated dataset. The analysis results (**Table 2**) shows that the proposed method and all other robust methods including classical method performs are almost same

Table 1: Canonical Correlation Analysis (CCA) for contaminated data

Correlation pairs	Classical	MVE	MCD	OGK	**Proposed** (RCCA)
r_1	0.34 (0.000)	0.34 (0.000)	0.30 (0.000)	0.31 (0.000)	0.31 (0.000)
r_2	0.13 (0.568)	0.18 (0.000)	0.17 (0.000)	0.18 (0.000)	0.23 (0.000)
r_3	0.12 (0.756)	0.14 (0.287)	0.14 (0.287)	0.17 (0.000)	0.19 (0.000)
r_4	0.11 (0.853)	0.13 (0.568	0.13 (0.568)	0.14 (0.293)	0.15 (0.159)
r_5	0.10 (0.951)	0.10 (0.453)	0.12 (0.653)	0.09 (0.503)	0.12 (0.460)
r_6	0.08 (0.985)	0.089 (0.762)	0.069 (0.832)	0.087 (0.782)	0.089 (0.876)
r_7	0.061 (0.979)	0.054 (0.803)	0.034 (0.872)	0.074 (0.840)	0.07 (0.865)

NB: The values within the parenthesis indicates the p.value.

in the absence of contamination (**Figure 1**). Again, to examine more robustness of the proposed method in comparison to all other robust methods including the

classical method, we contaminated (5% only) the datasets. Then we apply the classical method and other robust methods including the proposed method. The analysis results (**Table 1**) shows that our proposed method performs better than classical method as well as robust methods (**Figure 2**).

Table 2: Canonical Correlation Analysis (CCA)

Correlationpairs	Classical	MVE	MCD	OGK	**Proposed** (RCCA)
r_1	0.31 (0.000)	0.34 (0.000)	0.30 (0.000)	0.31 (0.000)	0.31 (0.000)
r_2	0.23 (0.000)	0.21 (0.000)	0.21 (0.000)	0.22 (0.000)	0.23 (0.000)
r_3	0.18 (0.000)	0.17 (0.000)	0.18 (0.000)	0.17 (0.000)	0.20 (0.000)
r_4	0.15 (0.005)	0.15 (0.000)	0.16 (0.000)	0.16 (0.000)	0.15 (0.009)
r_5	0.11 (0.552)	0.10 (0.453)	0.12 (0.653)	0.09 (0.503)	0.12 (0.549)
r_6	0.08 (0.812)	0.089 (0.762)	0.069 (0.832)	0.087 (0.782)	0.089 (0.802)
r_7	0.061 (0.831)	0.054 (0.803)	0.034 (0.872)	0.074 (0.840)	0.08 (0.819)

NB: The values within the parenthesis indicates the p.value.

Table 3: Power Analysis for contaminated data

Methods	Classical	MVE	MCD	OGK	**Proposed** (RCCA)
Power (for each canonical correlation pair(r)	1.00	1.00	1.00	1.00	1.00
	1.00	1.00	1.00	1.00	1.00
	1.00	0.97	0.97	1.00	1.00
	0.99	0.94	0.94	0.97	0.99
	0.89	0.72	0.89	0.61	0.89
	0.60	0.60	0.35	0.57	0.60
	0.68	0.58	0.50	1.00	1.00

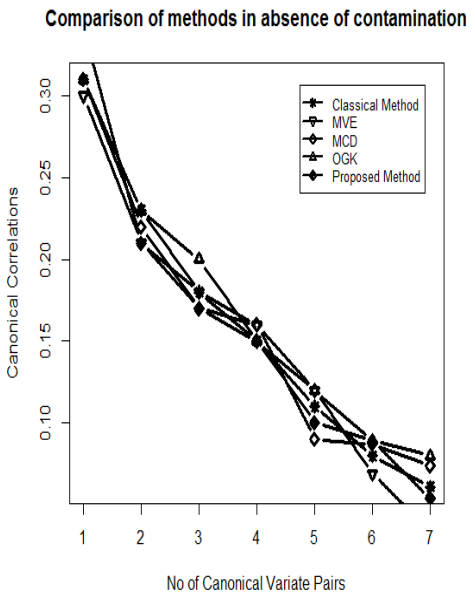

Figure 1: Comparison of canonical correlations by different methods

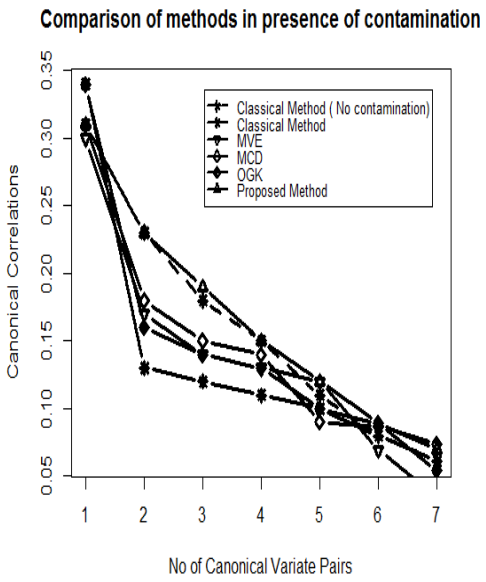

Figure 2: Comparison of canonical correlations by different methods

We also identified the power analysis for each canonical correlation coefficient. This power analysis represents the probability of getting true results that means higher value of power is the lower probability of type-2 error. The power analysis results for all methods (Table 3) in the presence of contaminated data show that proposed method (RCCA) represents better power than the other methods. We do not perform the power analysis for uncontaminated data due to almost similar performance for all methods.

4. Conclusion

Canonical correlation analysis (CCA) is an efficient and powerful tool for measuring the association between multiple genotypes and phenotypes in GWAS studies. This paper discusses a highly robust CCA approach using minimum β-divergence based covariance matrix. To investigate the performance of the proposed method in comparison to other robust method including the classical method, we generated two synthetic datasets (e.g. Contaminated and

uncontaminated). The analysis results show that the classical method, MVE, MCD, OGK and proposed method performs equally in absence of contamination. But in the presence of contamination, the proposed method performs better than the classical method as well as MVE, MCD and OGK method. Finally, we hope that our work helps to extend the application area of CCA in the field of both genetics and outside genetics.

Acknowledgments: We would like to thank both reviewer and editor for their valuable comments and suggestions that help us to improve the manuscript.

Reference

[1] Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. American journal of human genetics, 90(1), 7–24.

[2] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics, 101(1), 5–22.

[3] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. Science (New York, N.Y.), 308(5720), 385–389.

[4] Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B. A., Stirrups, K., König, I. R., Cazier, J. B., Johansson, A., Hall, A. S., Lee, J. Y., Willer, C. J., Chambers, J. C., Esko, T., … Samani, N. J. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nature genetics, 45(1), 25–33.

[5] Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421–427.

[6] Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., Horikoshi, M., Johnson, A. D., Ng, M. C., Prokopenko, I., Saleheen, D., Wang, X., Zeggini, E., Abecasis, G. R., Adair, L. S., … Morris, A. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nature genetics, 46(3), 234–244.

[7] Duncan, L., Yilmaz, Z., Gaspar, H., Walters, R., Goldstein, J., Anttila, V., Bulik-Sullivan, B., Ripke, S., Eating Disorders Working Group of the Psychiatric Genomics Consortium, Thornton, L., Hinney, A., Daly, M., Sullivan, P. F., Zeggini, E., Breen, G., and Bulik, C. M. (2017). Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. The American journal of psychiatry, 174(9), 850–858.

[8] Hyde, C. L., Nagle, M. W., Tian, C., Chen, X., Paciga, S. A., Wendland, J. R., Tung, J. Y., Hinds, D. A., Perlis, R. H., and Winslow, A. R. (2016). Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. Nature genetics, 48(9), 1031–1036.

[9] Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., Jiang, X., Rostamianfar, A., Finucane, H., Bolla, M. K., McGuffog, L., Wang, Q., Aalfs, C. M., ABCTB Investigators, Adams, M., Adlard, J., … Simard, J. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nature genetics, 49(12), 1767–1778.

[10] Sud, A., Kinnersley, B., and Houlston, R. S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. Nature reviews. Cancer, 17(11), 692–704.

[11] de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S. G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., Tremelling, M., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nature genetics, 49(2), 256–261.

[12] Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., de Leeuw, C. A., Benjamins, J. S., Muñoz-Manchado, A. B., Nagel, M., Savage, J. E., Tiemeier, H., White, T., 23andMe Research Team, Tung, J. Y., Hinds, D. A., Vacic, V., Wang, X., Sullivan, P. F., van der Sluis, S., … Posthuma, D. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nature genetics, 51(3), 394–403.

[13] Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., Visscher, P. M., and GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. Human molecular genetics, 27(20), 3641–3649.

[14] Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H. Y., Demirkan, A., Den Hertog, H. M., Do, R., Donnelly, L. A., Ehret, G. B., Esko, T., Feitosa, M. F., … Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nature genetics, 45(11), 1274–1283.

[15] Surakka, I., Horikoshi, M., Mägi, R., Sarin, A. P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S., Kettunen, J., Pirinen, M., Karjalainen, J., Thorleifsson, G., Hägg, S., Hottenga, J. J., Isaacs, A., Ladenvall, C., Beekman, M., Esko, T., … ENGAGE Consortium (2015). The impact of low-frequency and rare variants on lipid levels. Nature genetics, 47(6), 589–597.

[16] Kettunen, J., Tukiainen, T., Sarin, A. P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L. P., Kangas, A. J., Soininen, P., Würtz, P., Silander, K., Dick, D. M., Rose, R. J., Savolainen, M. J., Viikari, J., Kähönen, M., Lehtimäki, T., Pietiläinen, K. H., Inouye, M., McCarthy, M. I., Jula, A., … Ripatti, S. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nature genetics, 44(3), 269–276.

[17] Shin, S. Y., Fauman, E. B., Petersen, A. K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T. P., Walter, K., Menni, C.,

Chen, L., Vasquez, L., Valdes, A. M., Hyde, C. L., Wang, V., Ziemek, D., Roberts, P., Xi, L., … Soranzo, N. (2014). An atlas of genetic influences on human blood metabolites. Nature genetics, 46(6), 543–550.

[18] Inouye, M., Ripatti, S., Kettunen, J., Lyytikäinen, L. P., Oksala, N., Laurila, P. P., Kangas, A. J., Soininen, P., Savolainen, M. J., Viikari, J., Kähönen, M., Perola, M., Salomaa, V., Raitakari, O., Lehtimäki, T., Taskinen, M. R., Järvelin, M. R., Ala-Korpela, M., Palotie, A., and de Bakker, P. I. (2012). Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. PLoS genetics, 8(8), e1002907.

[19] Marttinen, P., Pirinen, M., Sarin, A. P., Gillberg, J., Kettunen, J., Surakka, I., Kangas, A. J., Soininen, P., O'Reilly, P., Kaakinen, M., Kähönen, M., Lehtimäki, T., Ala-Korpela, M., Raitakari, O. T., Salomaa, V., Järvelin, M. R., Ripatti, S., and Kaski, S. (2014). Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. Bioinformatics (Oxford, England), 30(14), 2026–2034.

[20] Ferreira, M. A., and Purcell, S. M. (2009). A multivariate test of association. Bioinformatics (Oxford, England), 25(1), 132–133.

[21] Tang, C. S., and Ferreira, M. A. (2012). A gene-based test of association using canonical correlation analysis. Bioinformatics (Oxford, England), 28(6), 845–850.

[22] Hotelling, H. (1936) Relations between two sets of variates. Biometrika, **28**, 321-377.

[23] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point, Mathematical Statistics and Applications, Vol. B (Grossmann et al., eds.), 283-297, Reidel, Dordrecht.

[24] Mollah, M. N., Sultana, N., Minami, M., and Eguchi, S. (2010). Robust extraction of local structures by the minimum beta-divergence method. Neural networks : the official journal of the International Neural Network Society, 23(2), 226–238.

[25] Liang, L., Zöllner, S., and Abecasis, G. R. (2007). GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics (Oxford, England), 23(12), 1565–1567.