Bioinformatics, 32(13), 2016, 1981–1989
doi: 10.1093/bioinformatics/btw052
Advance Access Publication Date: 19 February 2016
Original Paper

OXFORD

Genetics and population analysis

metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis

Anna Cichonska[1,2,*], Juho Rousu[2], Pekka Marttinen[2], Antti J. Kangas[3], Pasi Soininen[3,4], Terho Lehtimäki[5], Olli T. Raitakari[6,7], Marjo-Riitta Järvelin[8,9,10,11], Veikko Salomaa[12], Mika Ala-Korpela[3,4,13], Samuli Ripatti[1,14,15] and Matti Pirinen[1,*]

[1]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Helsinki, Finland, [2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland, [3]Computational Medicine, University of Oulu, Oulu University Hospital and Biocenter Oulu, Oulu, Finland, [4]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland, [5]Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere School of Medicine, Tampere, Finland, [6]Department of Clinical Physiology and Nuclear Medicine, University of Turku and Turku University Hospital, Turku, Finland, [7]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku and Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland, [8]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment & Health, School of Public Health, Imperial College London, London, UK, [9]Centre for Life Course Epidemiology, Faculty of Medicine, University of Oulu, Oulu, Finland, [10]Biocenter Oulu, University of Oulu, Oulu, Finland, [11]Unit of Primary Care, Oulu University Hospital, Oulu, Finland, [12]National Institute for Health and Welfare, Helsinki, Finland, [13]Computational Medicine, School of Social and Community Medicine and the Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK, [14]Public Health, University of Helsinki, Helsinki, Finland and [15]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on 16 July 2015; revised on 4 December 2015; accepted on 19 January 2016

Abstract

Motivation: A dominant approach to genetic association studies is to perform univariate tests between genotype-phenotype pairs. However, analyzing related traits together increases statistical power, and certain complex associations become detectable only when several variants are tested jointly. Currently, modest sample sizes of individual cohorts, and restricted availability of individual-level genotype-phenotype data across the cohorts limit conducting multivariate tests.

Results: We introduce *metaCCA*, a computational framework for summary statistics-based analysis of a single or multiple studies that allows multivariate representation of both genotype and phenotype. It extends the statistical technique of canonical correlation analysis to the setting where original individual-level records are not available, and employs a covariance shrinkage algorithm to achieve robustness.

Multivariate meta-analysis of two Finnish studies of nuclear magnetic resonance metabolomics by *metaCCA*, using standard univariate output from the program SNPTEST, shows an excellent

agreement with the pooled individual-level analysis of original data. Motivated by strong multivariate signals in the lipid genes tested, we envision that multivariate association testing using *metaCCA* has a great potential to provide novel insights from already published summary statistics from high-throughput phenotyping technologies.

Availability and implementation: Code is available at https://github.com/aalto-ics-kepaco

Contacts: anna.cichonska@helsinki.fi or matti.pirinen@helsinki.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Most human diseases and traits have a strong genetic component. Genome-wide association studies (GWAS) have proven effective in identifying genetic variation contributing to common complex disorders, including type 2 diabetes (Mahajan *et al.*, 2014), cardiovascular disease (Deloukas *et al.*, 2013), schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), and quantitative traits, such as lipid levels (Global Lipids Genetics Consortium, 2013; Surakka *et al.*, 2015) and metabolomics (Kettunen *et al.*, 2012; Shin *et al.*, 2014).

A dominant approach to GWAS is to test one single-nucleotide polymorphism (SNP) at a time against one quantitative phenotype measure or a binary disease indicator. This univariate approach is unlikely to be optimal when millions of SNPs and a growing number of phenotypes, including serum metabolomic profiles (Kettunen *et al.*, 2012; Shin *et al.*, 2014), three-dimensional images (Wang *et al.*, 2013), and gene expression data (Ardlie *et al.*, 2015) become available simultaneously. Indeed, a recent comparison demonstrated that utilizing multivariate phenotype representation increases statistical power, and leads to richer findings in the association tests compared to the univariate analysis (Inouye *et al.*, 2012). Moreover, some complex genotype-phenotype correlations can be detected only when testing several genetic variants simultaneously (Marttinen *et al.*, 2014), and multi-genotype tests are common practice in rare variant association studies, where statistical power to detect any single variant is very small (Feng *et al.*, 2014; Lee *et al.*, 2014).

Unfortunately, restricted availability of complete multivariate individual-level records across the cohorts currently limits multivariate analyses. Often, only the univariate GWAS summary statistics, i.e. univariate regression coefficients with their standard errors, from individual cohorts are publicly available. Hence, a major question is how we can use these univariate association results to carry out a multivariate meta-analysis of GWAS (Evangelou and Ioannidis, 2013), which is crucial to increase the power to identify novel genetic associations.

Recently, two kinds of multivariate testing approaches operating on univariate summary statistics have been introduced: (i) one SNP against multiple traits (Stephens, 2013; van der Sluis *et al.*, 2013; Vuckovic *et al.*, 2015; Zhu *et al.*, 2015) and (ii) multiple SNPs against one trait (Feng *et al.*, 2014; Yang *et al.*, 2012). We propose a new framework, *metaCCA*, that unifies both of the existing approaches by allowing canonical correlation analysis (CCA) of multiple SNPs against multiple traits based on univariate summary statistics and publicly available databases.

CCA is a well-established statistical technique for identifying linear relationships between two sets of variables, and has been successfully applied to GWAS (Ferreira and Purcell, 2009; Inouye *et al.*, 2012; Marttinen *et al.*, 2013; Tang and Ferreira, 2012). Our *metaCCA* method extends CCA to the setting where original individual-level measurements are not available. Instead, *metaCCA* works with three pieces of the full data covariance matrix, and applies a covariance shrinkage algorithm to achieve robustness. We demonstrate the performance of *metaCCA* using SNP and metabolite data from three Finnish cohorts. In summary, this paper makes the following contributions.

- To our knowledge, we provide the first computational framework for association testing between multivariate genotype and multivariate phenotype, based on univariate summary statistics from single or multiple GWAS. Our implementation is freely available.
- We demonstrate how to accurately estimate correlation structures of phenotypic and genotypic variables without an access to the individual-level data.
- We avoid false positive associations by a covariance shrinkage algorithm based on stabilization of the leading canonical correlation.
- Our approach, *metaCCA*, is a general framework to conduct CCA when full data are not available, and therefore it is widely applicable also outside GWAS.

A detailed discussion on the relationship between *metaCCA* and previously published multivariate association methods can be found in Supplementary Data.

2 Methods

This section is organized as follows. First, Section 2.1 explains univariate GWAS, the results of which, in the form of cross-covariance matrix, constitute an input to *metaCCA* described in Section 2.2; Section 2.3 demonstrates how a meta-analysis of several studies is conducted in our framework; Section 2.4 outlines a procedure for choosing SNPs representative of a given locus; finally, Section 2.5 introduces the data we used to test *metaCCA* in the meta-analytic setting.

2.1 Univariate GWAS

Let X and Y denote genotype and phenotype matrices of dimensions $N \times G$ and $N \times P$, respectively, storing the individual-level data; N the number of samples; G and P the number of genotypic and phenotypic variables, respectively. The columns of X and Y are standardized to have mean 0 and standard deviation 1.

Typically, univariate GWAS analysis of quantitative traits tests for an association between each pair of genotype $\boldsymbol{x}_g \in \mathbb{R}^N$ and phenotype $\boldsymbol{y}_p \in \mathbb{R}^N$ separately using a linear model:

$$\boldsymbol{y}_p = \alpha_{gp} + \boldsymbol{x}_g \beta_{gp} + \boldsymbol{\varepsilon}. \quad (1)$$

Coefficient β_{gp}, corresponding to the slope of the regression line, is the parameter of interest, since it depicts the size of the effect of the genetic variant $\boldsymbol{x}_g$ on the trait $\boldsymbol{y}_p$. Parameter α_{gp} is an intercept on the y-axis, and ε indicates a Gaussian error term or noise. The model is fit by the method of *least squares* that leads to a closed-form

estimate for the unknown parameter $\beta_{gp} = [\boldsymbol{x}_g^T \boldsymbol{y}_p][\boldsymbol{x}_g^T \boldsymbol{x}_g]^{-1} = [(N-1)s_{xy}][(N-1)s_{xx}]^{-1} = s_{xy}$, where s_{xy} is a sample covariance of $\boldsymbol{x}_g$ and $\boldsymbol{y}_p$, and $s_{xx} = 1$ is a sample variance of $\boldsymbol{x}_g$. Hence, the cross-covariance matrix Σ_{XY} between all genotypic and phenotypic variables is made of univariate regression coefficients β_{gp}:

$$\Sigma_{XY} = \frac{X^T Y}{N-1} = \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1P} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{G1} & \beta_{G2} & \cdots & \beta_{GP} \end{pmatrix}. \quad (2)$$

An important note is that if the individual-level datasets X and Y were not standardized before applying the linear regression, the standardization can be achieved afterwards by a transformation

$$\beta_{gp}^{\mathrm{STANDR}} = \frac{1}{\sqrt{N}\, SE_{gp}} \times \beta_{gp}, \quad (3)$$

where SE_{gp} indicates the standard error of β_{gp}, as given by GWAS software. (Typically, $\mathrm{SE}_{gp} \approx \sigma_p / (\sqrt{N}\sqrt{2f_g(1-f_g)})$, where σ_p is the standard deviation of the trait p, and f_g is the minor allele frequency of SNP g, but uncertainty in genotype imputation causes deviations from this expression.)

2.2 metaCCA

Conducting multivariate association tests requires estimates of the dependencies between genotypic and phenotypic variables, denoted Σ_{XX} and Σ_{YY}, respectively. Typically, they are calculated based on the individual-level measurements X and Y:

$$\Sigma_{XX} = \frac{X^T X}{N-1}, \quad (4)$$

$$\Sigma_{YY} = \frac{Y^T Y}{N-1}. \quad (5)$$

metaCCA operates on the cross-covariance matrix Σ_{XY} (Equation 2), and correlation structures $\hat{\Sigma}_{XX}, \hat{\Sigma}_{YY}$, estimated without an access to the individual-level data X and Y (Fig. 1A, B). To make the resulting full covariance matrix Σ a valid covariance matrix, *metaCCA* applies a shrinkage algorithm (Fig. 1C).

The rest of this section describes the details of *metaCCA* framework.

2.2.1 Estimation of genotypic correlation structure

Genetic variation is organized in haplotype blocks, whose structure is determined by mutation and recombination events, together with demographic effects, including population growth, admixture and bottlenecks (Wall and Pritchard, 2003). Hence, correlation structure of genetic variants differs between populations, such as, e.g. the Finns, Icelanders or Central Europeans. In *metaCCA*, $\hat{\Sigma}_{XX}$ is calculated using a reference database representing the study population, such as the 1000 Genomes database (1000 Genomes Project Consortium, 2012, www.1000genomes.org), or other genotypic data available on the target population. In the Section 3, we demonstrate that estimating $\hat{\Sigma}_{XX}$ from the target population (in our case, the Finns) leads to better results than utilizing the data comprising individuals across distinct populations (e.g. the Finns and other Europeans). However, since reference data on the target population may not always be at hand, we also present a robust but less powerful solution to multivariate association testing by simply using genotypes of all individuals from a certain broader geographical region (e.g. a continent) available under the 1000 Genomes Project.

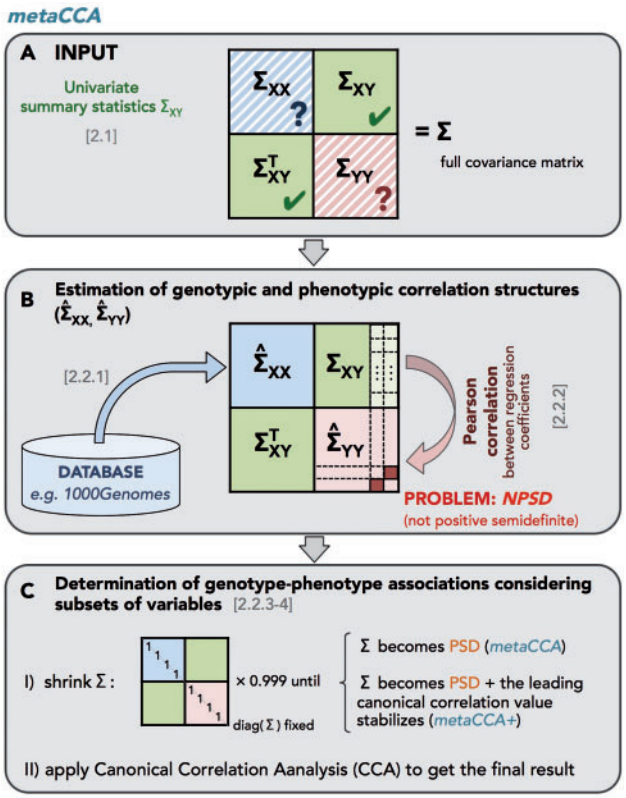

Fig. 1. Schematic picture showing an overview of *metaCCA* framework for summary statistics-based multivariate association testing using canonical correlation analysis. (**A**) *metaCCA* operates on three pieces of the full covariance matrix Σ: Σ_{XY} of univariate genotype-phenotype association results, Σ_{XX} of genotype-genotype correlations, and Σ_{YY} of phenotype-phenotype correlations. (**B**) $\hat{\Sigma}_{XX}$ is estimated from a reference database matching the study population, e.g. the 1000 Genomes, and phenotypic correlation structure $\hat{\Sigma}_{YY}$ is estimated from Σ_{XY}. (**C**) A covariance shrinkage algorithm is applied to add robustness to the method. Numbers in brackets refer to subsections in Methods. Meta-analysis of several studies is performed by pooling covariance matrices of the same type, before step (C), as described in Section 2.3. The data reduction achieved by *metaCCA* can be seen in Supplementary Figure S1

2.2.2 Estimation of phenotypic correlation structure

In our framework, phenotypic correlation structure $\hat{\Sigma}_{YY}$ is computed based on Σ_{XY}. Each entry of $\hat{\Sigma}_{YY}$ corresponds to a Pearson correlation between two column vectors of Σ_{XY} - univariate regression coefficients of two phenotypic variables s and t across G genetic variants:

$$\hat{\Sigma}_{YY}(s,t) = \frac{\sum_{g=1}^{G} (\beta_{gs} - \mu_s)(\beta_{gt} - \mu_t)}{\sqrt{\sum_{g=1}^{G} (\beta_{gs} - \mu_s)^2} \sqrt{\sum_{g=1}^{G} (\beta_{gt} - \mu_t)^2}}, \quad (6)$$

where μ_s and μ_t are the mean values $\mu_s = \frac{1}{G}\sum_{g=1}^{G} \beta_{gs}$ and $\mu_t = \frac{1}{G}\sum_{g=1}^{G} \beta_{gt}$. (The detailed justification is provided in Supplementary Data.) In Supplementary Table S2, we demonstrate that the higher the number of genotypic variables G, the lower the error of the estimate. Thus, $\hat{\Sigma}_{YY}$ should be calculated from summary

statistics of all available genetic variants, even if only a subset of them is taken to the further analysis.

2.2.3 Canonical correlation analysis

CCA (Hotelling, 1936) is a multivariate technique for detecting linear relationships between two groups of variables $X \in \mathbb{R}^{N\times G}$ and $Y \in \mathbb{R}^{N\times P}$, where X and Y constitute two different views of the same object. The objective is to find maximally correlated linear combinations of columns of each matrix. This corresponds to finding vectors $\boldsymbol{a} \in \mathbb{R}^G$ and $\boldsymbol{b} \in \mathbb{R}^P$ that maximize

$$r = \frac{(X\boldsymbol{a})^T(Y\boldsymbol{b})}{\|X\boldsymbol{a}\|\|Y\boldsymbol{b}\|} = \frac{\boldsymbol{a}^T\Sigma_{XY}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T\Sigma_{XX}\boldsymbol{a}}\sqrt{\boldsymbol{b}^T\Sigma_{YY}\boldsymbol{b}}}. \quad (7)$$

The maximized correlation r is called *canonical correlation* between X and Y. We provide the technical details of the method, as well as its extension to subsequent canonical correlations and their significance testing in Supplementary Data.

2.2.4 Shrinkage

At this point, we have three covariance matrices, namely Σ_{XY}, $\hat{\Sigma}_{XX}$, and $\hat{\Sigma}_{YY}$. However, in many cases, the resulting full covariance matrix

$$\Sigma = \begin{pmatrix} \hat{\Sigma}_{XX} & \Sigma_{XY} \\ \Sigma_{XY}^T & \hat{\Sigma}_{YY} \end{pmatrix}$$

is not positive semidefinite (PSD), and therefore its building blocks cannot be just plugged into the CCA framework (Equation 7). To overcome this problem, in *metaCCA*, we apply shrinkage to find a nearest valid Σ (Ledoit and Wolf, 2003). We use an iterative procedure where the magnitudes of the off-diagonal entries are being shrunk towards zero until Σ becomes PSD (Algorithm 1).

Algorithm 1

```
while Σ notPSD
    Σ = 0.999 × Σ;
    diag(Σ) = 1;
```

Assuring the PSD property of the full covariance matrix is necessary, although, as we demonstrate in the Section 3, not sufficient to obtain reliable results of the association analysis when the estimate $\hat{\Sigma}_{XX}$ (and/or $\hat{\Sigma}_{YY}$) is noisy. In order to address this issue, we propose a variant of *metaCCA*, called ***metaCCA+***, where the full covariance matrix Σ is shrunk beyond the level guaranteeing its PSD property. A challenge, however, is to find an optimal shrinkage intensity. Shrinkage applied without any stopping criterion would lead to gradual removal of all dependencies between genotypic and phenotypic variables. Ledoit and Wolf (2003) introduced an analytic approach for determining the optimal shrinkage level but it requires the individual-level datasets X and Y. In *metaCCA+*, we monitor the leading canonical correlation value r, and we continue the shrinkage of the full covariance matrix Σ until r stabilizes. Specifically, we track the percent change pc of r between subsequent shrinkage iterations, and we determine an appropriate amount of shrinkage using an elbow heuristic, similar to the criterion for finding the number of clusters, frequently used in the literature (Tibshirani *et al.*, 2001). The idea is that the slope of the graph should be steep to the left of the elbow, but stable to the right of it. We find the elbow, and thus the appropriate number of shrinkage iterations, by taking the point closest to the origin of the plot of pc versus iteration number, as schematically shown in Supplementary Figure S2.

Building blocks $\hat{\Sigma}_{XY}, \hat{\Sigma}_{XX}, \hat{\Sigma}_{YY}$ of the resulting full covariance matrix Σ, shrunk until it became PSD or beyond, are then plugged into the CCA framework to get the final genotype-phenotype association result. In practice, in order to protect from false positive signals, the shrinkage mode of *metaCCA+* should be applied whenever $\hat{\Sigma}_{YY}$ is estimated from summary statistics of a small number of genetic variants, and/or $\hat{\Sigma}_{XX}$ is calculated using a generic reference population.

2.2.5 Types of the multivariate association analysis

We consider the following two types of the multivariate analysis.

1. *Univariate genotype – multivariate phenotype*
   One genetic variant tested for an association with a set of phenotypic variables (matrix $\hat{\Sigma}_{XX}$ not needed).
2. *Multivariate genotype – multivariate phenotype*
   A set of genetic variants tested for an association with a set of phenotypic variables.

The first type corresponds to a standard multi-trait analysis. The second type takes into account the effects across genomic variants on multiple traits, which are ignored when analyzing only a single SNP or a single trait at a time.

2.3 Meta-analysis

metaCCA allows to conduct summary statistics-based multivariate analysis of one or multiple GWAS. In the meta-analytic setting, covariance matrices $\Sigma_{XY}^{(i)}, \hat{\Sigma}_{XX}^{(i)}$, and $\hat{\Sigma}_{YY}^{(i)}$ corresponding to $i = 1, \ldots, M$ independent studies on the same topic are pooled using a weighted average:

$$\Sigma_{XY} = \frac{(N_1 - 1)\Sigma_{XY}^{(1)} + \cdots + (N_M - 1)\Sigma_{XY}^{(M)}}{N - M}, \quad (8)$$

where N_i denotes the number of samples in the ith cohort, and $N = N_1 + \cdots + N_M$. This step is performed before applying the shrinkage to the full covariance matrix. As is typical for a fixed-effects meta-analysis, the weighted average is used in order to account for the varying precision of the estimates. The formulas for $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YY}$ are analogous to (8). However, if all cohorts included in the meta-analysis have the same underlying population, only one genotypic correlation estimate is needed.

2.4 Choosing SNPs representing a locus

When analyzing multiple genetic variants together, we use a procedure for selecting from a given locus a set of SNPs that jointly capture a maximal amount of genetic variation in the locus, as measured by a linkage disequilibrium (LD) score.

In each iteration, a SNP g that maximizes LD-score, which we define as $\sum_k \hat{r}_{gk}^2\sigma_k^2$, is selected, where the sum is over all SNPs k that have not yet been chosen; $\hat{r}_{gk}$ denotes a partial correlation between SNPs g and k; σ_k^2 indicates empirical variance of the residuals for SNP k after the effects of the selected SNPs have been regressed out. The residual variance σ_k^2 gets smaller, if the SNP has already been well explained by the previously chosen ones; hence, highly correlated SNPs will not be selected together. In the first iteration, $\hat{r}_{gk}$ is the Pearson correlation coefficient between SNPs g and k, and $\sigma_k^2 = 1$, meaning that the starting SNP is the one capturing the highest amount of genetic variation in the region. For each locus, we

select the smallest number of SNPs that explain, at median, over 95% of the variance of the remaining SNPs in the locus.

2.5 Datasets

In order to test our approach, we used genotypic and phenotypic data from three Finnish population cohorts: the Cardiovascular Risk in Young Finns Study (YFS, $N_1 = 2390$; Raitakari *et al.*, 2008), the FINRISK study survey of 1997 ($N_2 = 3661$; Vartiainen *et al.*, 2010), and the Northern Finland Birth Cohort 1966 (NFBC, $N_3 = 4702$; Rantakallio, 1969). The detailed description of the cohorts can be found in Supplementary Data.

Our phenotype data consist of 81 lipid measures (Supplementary Table S1) from a high-throughput nuclear magnetic resonance (NMR) platform (Soininen *et al.*, 2009, 2015). As a pre-processing step, within each cohort, each trait was quantile normalized, and the effects of age, sex and ten leading principal components of the genetic population structure were regressed out using a linear model. All cohorts were genotyped using Illumina arrays, and imputed by IMPUTE2 (Howie *et al.*, 2009) using the 1000 Genomes Project reference panel (1000 Genomes Project Consortium, 2012). In the analyses, we included 455 521 SNPs on chromosome 1 and, additionally, the SNPs in the following 5 genes:

- *APOE* (apolipoprotein E), 259 SNPs on chr 19;
- *CETP* (cholesteryl ester transfer protein), 387 SNPs on chr 16;
- *GCKR* (glucokinase (hexokinase 4) regulator), 160 SNPs on chr 2;
- *PCSK9* (proprotein convertase subtilisin/kexin type 9), 265 SNPs on chr 1;
- *NOD2* (nucleotide-binding oligomerization domain containing 2), 145 SNPs on chr 16.

We expected that this set of genes would provide a comprehensive spectrum of associations with our phenotypes, since *APOE*, *CETP*, *GCKR*, and *PCSK9* have well-known associations to lipid levels, whereas *NOD2* is not known to have such an association (NHGRI GWAS catalogue, Hindorff *et al.*, 2011, www.genome.gov/gwastudies). All SNPs used were of good quality: IMPUTE2 info ≥ 0.8 (Marchini and Howie, 2010), and minor allele frequency ≥ 0.05.

For multi-SNP models, we compared the results from Finnish genotype data with those obtained by estimating the genotypic correlation structure $\hat{\Sigma}_{XX}$ from the 1000 Genomes Project data on 503 European individuals (release 20130502).

For each cohort, genotypic and phenotypic correlation structures computed based on $X^{(i)}$ and $Y^{(i)}$, as shown in the Equations (4) and (5), can be found in Supplementary Figures S3 and S4.

3 Results

3.1 Performance assessment

The purpose of this section is to validate that *metaCCA* applied to summary statistics produces similar results to the standard CCA (MATLAB function *canoncorr*) applied to the individual-level data.

For *metaCCA*, we always use $\hat{\Sigma}_{YY}$ estimated by the method described in Section 2.2.2 using summary statistics of the entire chromosome 1.

We focus on the effects of (i) the amount of shrinkage applied to the full covariance matrix (*metaCCA*/*metaCCA*+) and (ii) estimating $\hat{\Sigma}_{XX}$ from the population underlying the analysis (here, Finnish), or from a more heterogeneous panel (here, European individuals from the 1000 Genomes database).

3.1.1 Univariate genotype – multivariate phenotype

We conducted a meta-analysis of the three cohorts (YFS, FINRISK and NFBC) by testing associations between each SNP in the five genes (as listed in Section 2.5; 1 216 SNPs in total) with different numbers of traits, ranging from 2 to 50. Multi-trait analyses are most useful for correlated traits (Stephens, 2013). To reflect this, for each SNP, we started with a randomly selected trait, and at each step of the analysis, added the trait mostly correlated with the already chosen ones, excluding correlations with absolute values above 0.95. For each SNP, we repeated the procedure three times with different starting lipid measures.

The scatter plot in Figure 2a shows that *metaCCA* applied to the cohort-wise summary statistics provides an excellent agreement with the standard CCA of the pooled individual-level data. Thus, in this one-SNP–multi-trait analysis, due to the reliable $\hat{\Sigma}_{YY}$ estimate used, we can base the inference on *metaCCA*, and put less weight on *metaCCA*+ (Fig. 2b) that, as expected, produces conservative *P*-values.

The wide range of the observed $-\log 10$ *P*-values (0–88) shows that multivariate association tests can be very powerful in realistic settings, and that our example assesses the performance of *metaCCA* throughout the range that is important in practical analyses. Supplementary Figure S5 further refines the behaviour of *metaCCA* within the range most encountered in genome-wide association studies (0–10).

3.1.2 Multivariate genotype – multivariate phenotype

When both genotype and phenotype are multivariate, genotypic correlation structure $\hat{\Sigma}_{XX}$ needs to be estimated in addition to $\hat{\Sigma}_{YY}$. We conducted the meta-analysis of two study cohorts (YFS and NFBC), and computed $\hat{\Sigma}_{XX}$ either from FINRISK (FIN) or from a more generic population of the 1000 Genomes European individuals (1000G). (Supplementary Table S3 shows errors of $\hat{\Sigma}_{XX}$ estimates.) We analyzed together between 2 and 10 highly correlated lipid measures, chosen sequentially as in the single-SNP tests in Section 3.1.1. For each of the five genes, we analyzed together between 2 and 10 SNPs that were chosen to be approximately uncorrelated to cover a large proportion of genetic variation within the gene. Each set of SNPs was tested for an association with each group of correlated lipid measures. We repeated the procedure ten times for each gene, with different starting phenotypes and SNPs.

The results are summarized in Figure 2c–f. Figure 2c shows that when genotypic correlation $\hat{\Sigma}_{XX}$ is estimated from the target population, *metaCCA* produces highly consistent results with the standard CCA based on the individual-level data. When $\hat{\Sigma}_{XX}$ is estimated from a less well matching population (Fig. 2e), the accuracy is reduced, and some $-\log 10$ *P*-values become clearly overestimated. In both cases, further shrinkage by *metaCCA*+ removes, almost completely, any overestimation (Fig. 2d, f). This property is expected to be important in genome-wide association studies, where *metaCCA*+ can protect from false positives when genotypic correlation structure cannot be accurately estimated. *metaCCA*+ has less statistical power than the individual-level CCA, but it is still able to detect strong true associations.

3.2 Application to summary statistics from SNPTEST

In the genetics community, established software packages like SNPTEST (Marchini and Howie, 2010) are used to perform univariate genome-wide tests. In this section, we conduct a meta-analysis of univariate results from standard SNPTEST runs on NFBC and YFS cohorts by *metaCCA*. These cohorts have been meta-analyzed

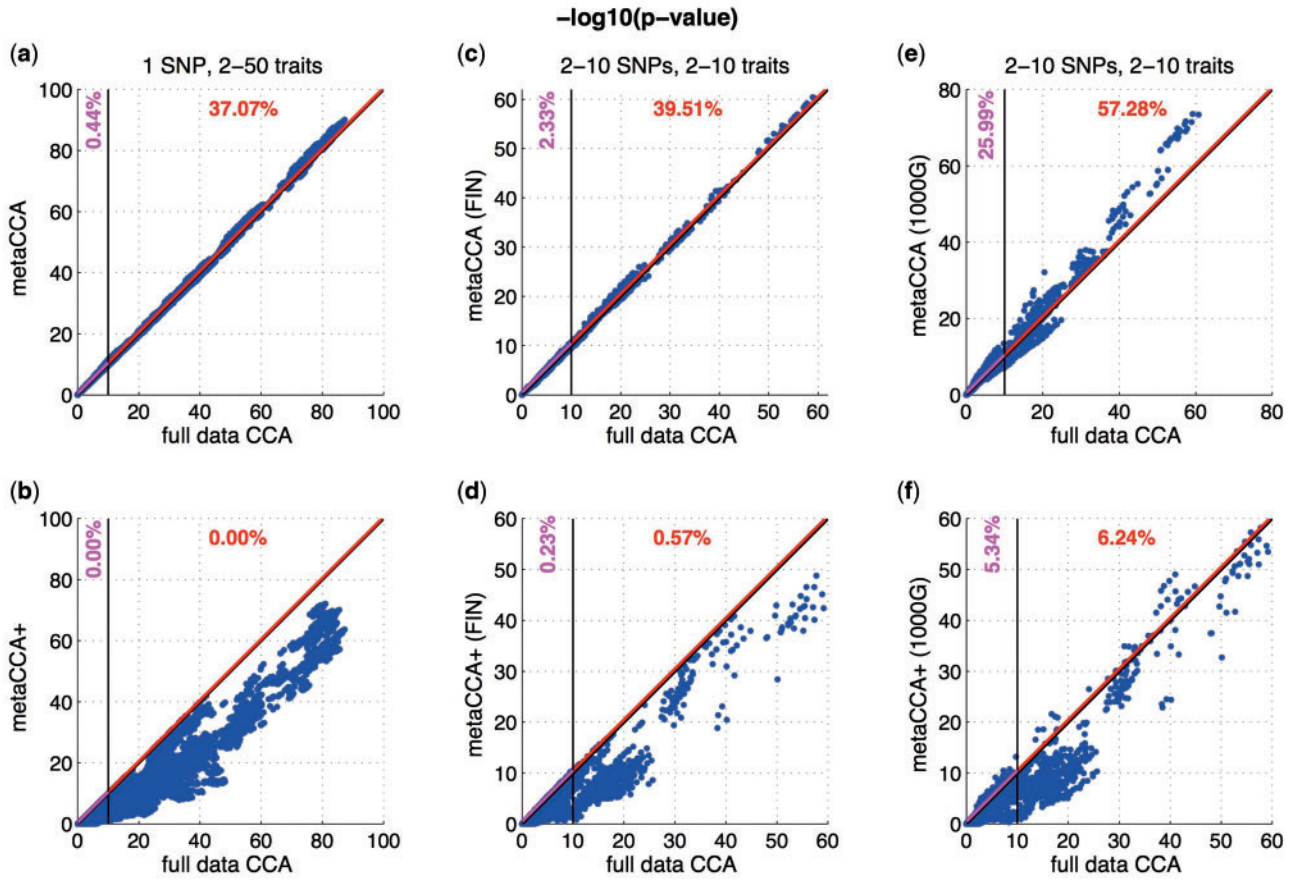

Fig. 2. Scatter plots of −log 10 *P*-values between the pooled individual-level analysis of original datasets (*full data CCA*) and *metaCCA* (first row), *metaCCA*+ (second row). (**a**, **b**) *Univariate genotype – multivariate phenotype*; meta-analysis of NFBC, FINRISK and YFS cohorts; (**c–f**) *Multivariate genotype – multivariate phenotype*; meta-analysis of NFBC and YFS cohorts; *metaCCA*/*metaCCA*+ was used with $\hat{\Sigma}_{XX}$ computed from FINRISK (FIN; c, d), or from the 1000 Genomes database (1000G, 503 EUR individuals; e, f) In all the cases, lipid correlation structure $\hat{\Sigma}_{YY}$ was calculated from univariate summary statistics of SNPs from the entire chromosome 1. Single point corresponds to the result of one out of (a–b) 178 752, (c–f) 4050 multivariate tests. Numbers at the top of each plot indicate percentages of at least 0.5 unit overestimated *metaCCA*'s/*metaCCA*+'s −log 10 *P*-values in the ranges [0, 10] (purple) or (10, max(−log 10 *P*-value)] (red). This threshold is represented by purple and red lines. Supplementary Figure S5 shows these results restricted to the *x*-axis range of [0, 10], and Supplementary Figure S6 illustrates the impact of the number of genotypic and phenotypic features included in the analysis on the accuracy of *metaCCA*/*metaCCA*+

previously using standard CCA applied to pooled individual-level genotypes and the same serum metabolomic profiles that we consider here (Inouye *et al.*, 2012). This single-SNP–multi-trait GWAS highlighted candidate genes for atherosclerosis, and demonstrated the power of incorporating multiple related traits into the analysis. Here, we show that by *metaCCA* we obtain those same results without the access to the individual-level data, and, in addition to that, we can also analyze multiple SNPs jointly by using only summary statistics from the original studies.

We wanted to choose a set of correlated traits for the joint analysis, and therefore we proceeded as follows. By an agglomerative hierarchical clustering (average linkage) of Σ_{YY} (81 traits), we identified groups of related lipid measures. From the largest of 6 distinct clusters, we selected a set of traits in such a way that no pair exhibited correlation above 0.95. We ended up with a group of 9 lipid measures related to 8 VLDL particles of different sizes and one HDL particle (highlighted in blue in Supplementary Table S1).

We conducted two types of meta-analyses of NFBC and YFS:

1. *Univariate genotype – multivariate phenotype*
   Each SNP from chromosome 1 tested for an association with the set of 9 correlated lipid measures.
2. *Multivariate genotype – multivariate phenotype*
   For each of the 5 genes (APOE, CETP, GCKR, PCSK9, NOD2), the smallest set of SNPs that explained, at median, over 95% of the variance of the remaining SNPs is chosen (see Section 2.4), and tested for an association with the set of 9 correlated lipid measures.

The input summary statistics for *metaCCA* were obtained by performing univariate tests for each SNP-trait pair separately using SNPTEST applied to the individual-level data, and transforming the resulting regression coefficients using (3). The correlation structure of analyzed traits, $\hat{\Sigma}_{YY}$, was estimated from summary statistics of SNPs across the entire genome. The genotypic correlation structure for multi-SNP analyses, $\hat{\Sigma}_{XX}$, was calculated from the FINRISK cohort.

We compared the results of *metaCCA* and *metaCCA*+ with the pooled individual-level CCA of original datasets. Figure 3 shows scatter plots of − log 10 *P*-values for 455 521 SNPs from chromosome 1. The results of *metaCCA* demonstrate an excellent agreement with the original *P*-values, validating that *metaCCA* can conduct reliable multivariate meta-analysis from standard univariate GWAS software output. As anticipated, *metaCCA*+ produces

conservative *P*-values. Here, *metaCCA* is indeed the method of choice in practice, due to the high quality of covariance estimate used. Manhattan plots illustrating *P*-values along the chromosome are shown in Supplementary Figure S7. Genome-wide significant associations (at the threshold of $P = 5 \times 10^{-8}$ standard in the field) are located within two regions: *USP1/DOCK7* and *FCGR2A/3A/2C/3B*, which are known to be associated with lipid metabolism (NHGRI GWAS catalogue, Hindorff *et al.*, 2011). *metaCCA* identified both regions, and *metaCCA+* found the stronger out of the two signals (*DOCK7/USP1*). For top-SNP in *FCGR2A/3A/2C/3B*, *metaCCA+*'s −log 10 *P*-value is 6.11, compared to 7.73 produced by CCA on the individual-level data.

Figure 4 summarizes the results of the multi-SNP–multi-trait meta-analysis, and shows the performance of *metaCCA* when different numbers of SNPs, from 2 up to 25, representing a gene, are tested jointly for an association with the group of 9 related lipid traits. Numbers of SNPs that are chosen by our approach (Section 2.4) are marked with **x**. Figure 4 validates that by using this protocol, a gene is described well, since when adding more SNPs no clear power gain is observed. Both *metaCCA* and *metaCCA+* (Fig. 4, Supplementary Table S4) produced very accurate *P*-values. For the largest signals (*APOE*, *CETP*), −log 10 *P*-values are less than one unit overestimated by *metaCCA*, and underestimated by *metaCCA+*. These differences would be unlikely to lead to false inferences when a reference significance level in a gene-based analysis was set to $0.05/20000 = 2.5 \times 10^{-6}$, i.e. 5.61 on − log 10 scale, based on there being about 20 000 protein-coding genes in the human genome. At this level, both *metaCCA* and *metaCCA+* found an association between *APOE*, *CETP*, *GCKR* and the network of VLDL and HDL particles studied. For *APOE* and *CETP*, gene-based signals are clearly higher than the univariate ones, even before accounting for different numbers of tests. Moreover, in case of *APOE*, the multi-SNP–multi-trait signal is nearly 4.5 units higher than the single-SNP–multi-trait one. Note that *NOD2* has no (known) association with metabolic traits, and therefore it serves as a negative control Figure 4 and Supplementary Table S4.

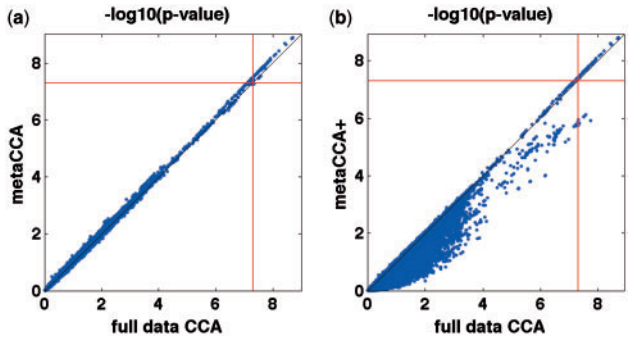

Fig. 3. Scatter plots of −log 10 *P*-values from the pooled individual-level CCA of NFBC and YFS and (**a**) *metaCCA*, (**b**) *metaCCA+*. Each point corresponds to one genetic variant from the chromosome 1, tested for an association with the group of 9 correlated lipid measures. In total, 455 521 SNPs were analyzed. Red lines indicate the significance level of 5×10^{-8} (7.301 on −log 10 scale)

4 Discussion

The advantage of multivariate testing of genetic association is well reported in the literature (Inouye *et al.*, 2012; Stephens, 2013), and also demonstrated in our results (e.g. *CETP* in Supplementary Table S4 that has multivariate *P*-value 13 orders of magnitude smaller than any of the univariate *P*-values). Optimal use of correlated traits is becoming increasingly important as high-throughput phenotyping technologies are being more widely applied to individual study cohorts and large biobanks (Soininen *et al.*, 2015).

We introduced *metaCCA*, a computational approach for the multivariate meta-analysis of GWAS by using univariate summary statistics and a reference database of genetic data. Thus, our framework circumvents the need for complete multivariate individual-level records, and tackles the problem of low sample sizes in individual cohorts by a built-in meta-analysis approach. To our knowledge, *metaCCA* is the first summary statistics-based framework that allows multivariate representation of both genotypic and phenotypic variables.

In large meta-analytic efforts, the ability to work with summary statistics is beneficial, even when there is an access to the individual-level data. For example, with a study design of the Global Lipids Genetics Consortium (2013), we estimate that the reduction in the size of input data between *metaCCA* and standard CCA could be over 750-fold (Supplementary Figure S1).

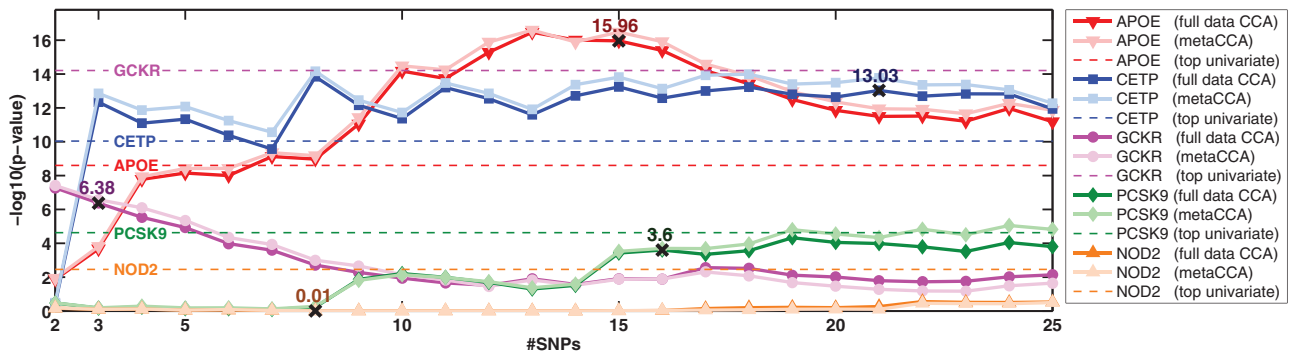

Fig. 4. *Multi-SNP–multi-trait* analysis: −log 10 *P*-values of CCA on pooled individual-level datasets (NFBC + YFS), and the meta-analyses conducted using *metaCCA*, as a function of the number of SNPs representing a gene. Sets of 2–25 SNPs were tested for an association with the group of 9 related lipid measures. In practice, the smallest number of SNPs that explain, at median, over 95% of the variance of the remaining SNPs would be chosen to represent a gene, and is marked with **x**. The evolution of the median variance explained versus the number of SNPs is shown in Supplementary Figure S8. For each gene, the largest −log 10 *P*-value from single-SNP–single-trait tests (*top univariate*) is represented by a dashed line. The largest single-SNP–multi-trait −log 10 *P*-values are 11.54 for *APOE*, 23.77 for *CETP*, 9.64 for *GCKR*, 6.58 for *PCSK9* and 0.97 for *NOD2*. The values are summarized with details in Supplementary Table S4. The number of tests in each gene is 1 for multi-SNP, *G* for single-SNP–multi-trait, and $9 \times G$ for single-SNP–single-trait tests, where *G* is the number of SNPs in that gene

We provided two variants of the algorithm: *metaCCA* and *metaCCA+*. Based on our results, *metaCCA* is the method of choice when the accuracy of estimated correlation matrices $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{YY}$ is good, i.e. $\hat{\Sigma}_{XX}$ estimated from genetic data on the target population, and $\hat{\Sigma}_{YY}$ estimated from at least one chromosome. In such cases, *P*-values from *metaCCA* were very accurate, meaning that false positive and false negative rates are close to those of standard CCA applied to the individual-level data. We emphasize that *metaCCA* should not be used when the quality of $\hat{\Sigma}_{XX}$ and/or $\hat{\Sigma}_{YY}$ estimates is reduced, i.e. when a generic reference population and/or summary statistics of only a small number of genotypes are available. In such cases, *metaCCA+* proved useful to protect from an increase of false positive associations (Fig. 2 and Supplementary Figure S9). This is important in GWAS context, where false positives could lead to considerable waste of resources in subsequent experimental and functional studies. A topic for future work would be to further develop our current heuristic stopping criterion of *metaCCA+* to decrease its false negative rate without sacrificing its good false positive rate.

We derived the framework assuming that all traits within each cohort have been measured on the same number of individuals (N). We note that the distribution of the test statistic depends on N (Supplementary Data), as do the effect size transformation (Equation 3) and meta-analysis approach (Section 2.3). While a small proportion of missing data for each trait could be handled by statistical imputation methods, further work is required to study how *metaCCA* should be used when the sample sizes between the traits vary considerably. However, with high-throughput phenotyping technologies, we believe that *metaCCA* can be applied to many existing and forthcoming studies.

For multivariate phenotype data, several types of association tests are possible. Natural question is which one should we prefer in practice. It is evident that single-SNP–multi-trait tests can detect much stronger signals at some SNPs than any of the univariate tests separately (e.g. *CETP* in Supplementary Table S4), and identify associations not found by univariate approach (Inouye *et al.*, 2012). On the contrary, for some other SNPs, the highest univariate signal may be clearly higher than the multi-trait one, even after accounting for the increase in the number of tests. For example, in *GCKR* (Supplementary Table S4), the top SNP's (rs1260326) association was explained already by one of the traits individually (M.VLDL.FC). Given the difference in degrees of freedom of the tests, this led to a 4.6 units higher $-\log 10$ *P*-value in the univariate test compared to the multivariate one. Thus, for single-SNP analysis, univariate and multivariate tests complement each other and neither should be excluded from consideration.

When also genotypes are multivariate, even more possibilities for association testing emerge. To illustrate our multi-SNP approach, we proposed a procedure for selecting, for each gene, the smallest number of SNPs that explained, at median, over 95% of the variance of the remaining SNPs in the locus. We demonstrated that testing multiple SNPs jointly can be more powerful than single-SNP–single-trait (*APOE*, *CETP* in Fig. 4 and Supplementary Table S4) and single-SNP–multi-trait tests (*APOE* in Supplementary Table S4). Moreover, *metaCCA* could equally well incorporate any other way of choosing the SNPs, for example, motivated by functional annotations (ENCODE Project Consortium, 2012), known expression effects (Ardlie *et al.*, 2015) or previous GWAS results on other traits (Hindorff *et al.*, 2011). A topic for further research could be to extend the covariance matrix-based analyses from CCA to dynamic approaches that learned from the data the set of variants and traits to be considered together. This would circumvent the need to restrict the subset of variables before the analysis.

We envision that multivariate association testing using *metaCCA* has a great potential to provide novel insights from already published summary statistics of large GWAS meta-analyses on multivariate high-throughput phenotypes, such as metabolomics and transcriptomics. Finally, we hope that our work helps extending the application area of CCA to summary statistics data also in other data-rich fields outside genetics.

Funding

Funding: This work was financially supported by the Helsinki Doctoral Education Network in Information and Communications Technology (HICT) to A.C., the Academy of Finland [257654 and 288509 to M.P.; 251170 to the Finnish Centre of Excellence in Computational Inference Research COIN; 259272 to P.M.; 251217 and 255847 to S.R.] and the Sigrid Juselius Foundation to S.R., A.J.K., P.S. and M.A.K. S.R. was supported by EU FP7 projects ENGAGE (201413), BioSHaRE (261433), the Finnish Foundation for Cardiovascular Research and Biocentrum Helsinki. A.J.K., P.S. and M.A.K. were supported by Strategic Research Funding from the University of Oulu and by Novo Nordisk Foundation.

Conflict of Interest: A.J.K., P.S. and M.A.K. are shareholders of Brainshake Ltd., a company offering NMR-based metabolite profiling. A.J.K. and P.S. report employment relation for Brainshake Ltd.

References

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Ardlie,K.G. *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Sci.*, **348**, 648–660.

Deloukas,P. *et al.* (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.*, **45**, 25–33.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Evangelou,E. and Ioannidis,J.P. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, **14**, 379–389.

Feng,S. *et al.* (2014) RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics*, btu367.

Ferreira,M.A. and Purcell,S.M. (2009) A multivariate test of association. *Bioinformatics*, **25**, 132–133.

Global Lipids Genetics Consortium (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.

Hindorff,L.A. *et al.* (2011) A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies. (July 2015, date last accessed).

Hotelling,H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321–377.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLos Genet.*, **5**, e1000529.

Inouye,M. *et al.* (2012) Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLos Genet.*, **8**, e1002907.

Kettunen,J. *et al.* (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.*, **44**, 269–276.

Ledoit,O. and Wolf,M. (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, **10**, 603–621.

Lee,S. *et al.* (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.

Mahajan,A. *et al.* (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.*, **46**, 234–244.

Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.

Marttinen,P. *et al.* (2013) Genome-wide association studies with high-dimensional phenotypes. *Stat. Appl. Genet. Mol. Biol.*, **12**, 413–431.

Marttinen,P. *et al*. (2014) Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics*, **30**, 2026–2034.

Raitakari,O.T. *et al*. (2008) Cohort profile: the Cardiovascular Risk in Young Finns Study. *Int. J. Epidemiol*., **37**, 1220–1226.

Rantakallio,P. (1969) Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scand*., **193**, 193.

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

Shin,S.Y. *et al*. (2014) An atlas of genetic influences on human blood metabolites. *Nat. Genet*., **46**, 543–550.

Soininen,P. *et al*. (2009) High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst*, **134**, 1781–1785.

Soininen,P. *et al*. (2015) Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circul.: Cardiovasc. Genet*., **8**, 192–206.

Stephens,M. (2013) A unified framework for association analysis with multiple related phenotypes. *PLoS One*, **8**, e65245.

Surakka,I. *et al*. (2015) The impact of low-frequency and rare variants on lipid levels. *Nat. Genet*., **47**, 589–597.

Tang,C.S. and Ferreira,M.A. (2012) A gene-based test of association using canonical correlation analysis. *Bioinformatics*, **28**, 845–850.

Tibshirani,R. *et al*. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, **63**, 411–423.

van der Sluis,S. *et al*. (2013) TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLos Genet*., **9**, e1003235.

Vartiainen,E. *et al*. (2010) Thirty-five-year trends in cardiovascular risk factors in Finland. *Int. J. Epidemiol*., **39**, 504–518.

Vuckovic,D. *et al*. (2015) MultiMeta: an R package for meta-analyzing multiphenotype genome-wide association studies. *Bioinformatics*, btv222.

Wall,J.D. and Pritchard,J.K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet*., **4**, 587–597.

Wang,Y. *et al*. (2013) Random forests on Hadoop for Genome-Wide Association Studies of multivariate neuroimaging phenotypes. *BMC Bioinf*., **14**, 1–15.

Yang,J. *et al*. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet*., **44**, 369–375.

Zhu,X. *et al*. (2015) Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet*., **96**, 21–36.