Published in final edited form as:

Methods Mol Biol. 2017 ; 1666: 455–467. doi:10.1007/978-1-4939-7274-6_22.

Cross-phenotype association analysis using summary statistics from GWAS

Xiaoyin Li^{1,*} and Xiaofeng Zhu²

¹Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, 44106, USA. xiaoyin.li@case.edu.

²Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, 44106, USA.

Abstract

For over a decade, genome-wide association studies (GWAS) have been a major tool for detecting genetic variants underlying complex traits. Recent studies have demonstrated that the same variant or gene can be associated with multiple traits, and such associations are termed cross-phenotype (CP) associations. CP association analysis can improve statistical power by searching for variants that contribute to multiple traits, which is often relevant to pleiotropy. In this chapter, we discuss existing statistical methods for analyzing association between a single marker and multivariate phenotypes, we introduce a general approach, CPASSOC, to detect the CP associations, and explain how to conduct the analysis in practice.

Keywords

Genome-wide association studies; cross-phenotype association; meta-analysis; multivariate phenotypes; pleiotropy; summary statistics

1. Introduction

In the last decade, Genome-wide association studies (GWAS) have identified more than 8500 associations with complex traits such as heart disease, diabetes, auto-immune diseases and psychiatric disorders, etc, and have provided a wealth of novel insights into the etiology of human diseases and traits (<https://www.genome.gov/gwastudies/>). More and more empirical evidence suggests that many detected genetic loci harbor variants that are associated with multiple, sometimes seemingly distinct, traits (1, 2). Such associations are termed cross-phenotype (CP) associations and are potential evidence for pleiotropy, i.e. that the same variants are associated with multiple traits. For example, a gene desert located on

*Corresponding author Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA, Phone: 216-368-4567, xiaoyin.li@case.edu.

³Notes

³The parameter SampleSize is a vector consisting of M sample sizes for obtaining the K summary statistics. It is used for weights in combining the summary statistics. The current version assumes sample sizes are the same for all SNPs. In our example with anthropometric traits, we chose the median sample size for each trait.

chromosome 8q24 is associated with multiple cancer types, and multiple GWAS have demonstrated that this region is associated with prostate, breast, colon, ovarian, and bladder cancers and chronic lymphocytic leukemia (3–7). Recent studies (8–11) also show that type I diabetes (T1D), rheumatoid arthritis, systemic lupus erythematosus (SLE) and Graves disease share a functional variant in *PTPN22*, which encodes the lymphoid PTP Lyp gene (12). Those CP associations indicate the sharing of the same susceptibility genes and alleles, suggesting a role for pleiotropic genetic effects (13).

In general, pleiotropy is defined as a genetic variant truly affecting multiple phenotypes and is one of the underlying causes for an observed CP association (13). A CP association can occur in the context of: 1) biological pleiotropy, a SNP or a gene affects multiple phenotypes; 2) phenotype causal relationships, often called mediated pleiotropy or referred to as the situation where the genetic effect on phenotype B is mediated by phenotype A, which is causally related to phenotype B; 3) spurious associations caused by misclassifications or ascertainment bias. CP association implies potential pleiotropy where a variant is associated with multiple traits regardless of the underlying cause. Thus, CP association is more general than pleiotropy.

To date, GWAS are usually performed one trait at a time, although multiple related phenotypes with a common physiological process are often collected and studied. When cross-phenotype effects exist, focusing on single traits will miss the opportunity to systematically integrate the phenome-wide data available for genetic association analysis. On the other hand, analyzing multiple phenotypes can further improve statistical power in detecting significant genetic loci and provide a biological interpretation. Statistical methods to detect CP effects have been developed from different angles. They can be broadly classified into multivariate analyses and univariate analyses (summarized in Table 1).

1.1 Multivariate approaches

The joint analysis of multiple phenotypes within a cohort has recently become popular for improving statistical power to detect genetic linkage and association. In such studies, correlated phenotypes, or a multivariate phenotype with several components, may be measured at the same time. For example, a hypertension study often measures systolic blood pressure (SBP), diastolic blood pressure (DBP), and hypertension status (HTN) (14, 15). Most multivariate methods are based on a multivariate regression framework and require both genotypes and phenotypes at the individual level, with an assumption of approximately multivariate normally distributed phenotypes. Linear mixed effects models (LME) are applied using fixed effects for the genetic variants and random effects to account for correlations among the multivariate phenotypes (16, 17). Extensions to allow for non-normally distributed phenotypes and categorical phenotypes have also been developed based on generalized estimating equations (GEEs), ordinal regression, and a Bayesian framework (18–20).

Other approaches have been developed based on a dimension reduction technique on the phenotypes, such as principal-component (PC) analysis and canonical correlation analysis (CCA) (21–23). Variable reduction approaches are in general only applicable to multivariate phenotypes consisting of all continuous phenotypes that are approximately normally

distributed. These approaches derive a single or a few new phenotypes that are linear combinations of the original phenotypes. For example, a few top PCs are constructed to preserve the multivariate trait variation, but this may not be able to capture sufficient association (24). In practice, typically not all subjects have been measured for all traits, and when different traits have been investigated in different studies, the individual subject data may not even be available.

1.2 Univariate approaches

An alternative way to analyze multivariate phenotypes is to perform univariate phenotype-genotype association tests for each phenotype individually and then combine the test statistics from the univariate analyses. A key feature of such approaches is the ability to derive the output statistics from SNP summary statistics, hence making it possible to perform systematic meta-analysis-type comparisons across multiple GWAS datasets.

The Fisher's combined P -value method is a step in this direction. However, it suffers from several limitations, such as an inability to provide the summary effect of a genetic variant and difficulty in addressing genetic heterogeneity. Moreover, it is designed for independent studies and therefore it is not straightforward when aggregating P -values of different but correlated phenotypes within the same cohort, which may result in inflated type I error.

The success of GWAS has shown that there is considerable benefit in being able to derive association tests on the basis of summary statistics through meta-analysis (25) and many association tests have been successfully extended to combine association across multiple phenotypes (26–28). The techniques developed for meta-analysis can be applied to summary data, thus diminishing the limitations that are imposed by restrictions on sharing individual-level data. Meta-analytical approaches aggregate summary statistics from individual studies into one statistic to test for CP effects, the evidence for association is combined across studies of multiple phenotypes. The null hypothesis is therefore that there is no association between a genetic marker and any of the phenotypes, and the alternative hypothesis is that the genetic marker is associated with at least one of the studied phenotypes.

The cross-phenotype meta-analysis (CPMA) statistic aims to detect multiple associations at a marker across different diseases that may share a common genetic background or involve a common biological process (27). The null hypothesis is that the expected P -values are uniformly distributed, and hence the values of $-\ln(P)$ are exponentially decaying with a decay rate $\lambda = 1$. The test statistic is expressed as a likelihood ratio test:

$$CPMA = -2 \frac{P[Data|\lambda = 1]}{P[Data|\hat{\lambda}]} \text{ and is asymptotically distributed as } \chi^2 \text{ with one degree of freedom.}$$

It determines evidence for the hypothesis that each independent SNP has multiple phenotypic associations, rather than directly evaluating the aggregated association evidence between a SNP and multiple phenotypes. When one of the traits is not associated with a SNP, CPMA will have less statistical power. Moreover, CPMA assumes that the P -values used for the individual traits come from different non-overlapping cohorts, and it does not allow for overlapping or correlated samples among studies.

Fixed and random effects meta-analysis methods can also be used to detect CP effects. Fixed effects meta-analysis assumes the genetic variant has the same effect across multiple phenotypes. Random effects meta-analysis allows a moderate level of effect heterogeneity, but nevertheless it is not well suited to situations where a genetic variant has opposite effects on different phenotypes, because of loss of power (13, 29). An extension of fixed effects meta-analysis is the subset-based meta-analysis, which allows for opposite effects and is able to test association to a subset of phenotypes (28). This method exhaustively searches all possible phenotype subsets and identifies the subset of traits with the strongest association, but at the cost of exponentially increased multiple tests. In addition, the method does not allow for heterogeneity either across cohorts or for the same phenotype.

O'Brien's linear combination test has been developed based on a weighted linear combination of the univariate test statistics (30). The power of O'Brien's method depends on the assumption of homogeneous genetic effect across phenotypes. Some authors have proposed extensions to O'Brien's approach to allow for heterogeneity among individual test statistics (31, 32). However, these approaches were specifically developed for correlated traits measured on the same individuals.

1.3 CPASSOC

CPASSOC was first applied to the Continental Origins and Genetic Epidemiology Network (COGENT) African ancestry samples for three blood pressure traits (26). In this study CPASSOC identified four loci associated with HTN at genome wide significance level that were missed by a single trait analysis in the original report. All the four loci, *CHIC2*, *HOXA-EVX1*, *IGFBP1/IGFBP3* and *CDH17* have been shown by others to be associated with HTN-related traits in European ancestry GWAS (15, 33–37), suggesting that our method reliably identifies true signals. Six additional loci with suggestive association evidence ($P < 5.0 \times 10^{-7}$) were also observed, including *CACNA1D* and *WNT3*. CPASSOC was also applied to anthropometric traits, such as height, BMI, and waist-to-hip ratio adjusted for BMI (WHRadjBMI), where the summary statistics were obtained from the GIANT consortium (38). Those applications strongly suggest that analyzing multiple phenotypes can improve statistical power and that such an analysis can be executed with the summary statistics from GWAS.

CPASSOC uses summary-level data from single SNP-trait associations from GWAS to detect which variant is associated with at least one trait. This method improves statistical power by analyzing multiple phenotypes. CPASSOC provides two statistics, S_{Hom} and S_{Het} . S_{Hom} is similar to the fixed effect meta-analysis method (39) but accounting for the correlations of summary statistics among traits and among cohorts induced by correlated traits, potential overlapping or related samples.

In brief, we assume we have summary statistics of GWAS from J cohorts with K phenotypic traits. In each cohort, single SNP-trait association has been analyzed for each trait separately. Let T_{jk} be a summary statistic for a SNP, derived from the j^{th} cohort for the k^{th} trait. Let $T = (T_{11}, \dots, T_{J1}, \dots, T_{1K}, \dots, T_{JK})^T$ represent a vector of test statistics for testing the

association of a SNP with K traits. We use a Wald test statistic $T_{jk} = \frac{\hat{\beta}_{jk}}{\hat{s}_{jk}}$, where $\hat{\beta}_{jk}$ and $\hat{s}_{jk}$ are the estimated coefficient and corresponding standard error for the k^{th} trait in the j^{th} cohort. S_{Hom} is then defined as

$$S_{Hom} = \frac{e^T(RW)^{-1}T(e^T(RW)^{-1}T)^T}{e^T(WRW)^{-1}e}, \quad (1)$$

which follows a χ^2 distribution with one degree of freedom, where $e^T = (1, \dots, 1)$ has length $J \times K$ and W is a diagonal matrix of weights for the individual test statistics. We used the sample sizes for the weights, i.e., $w_{jk} = \sqrt{n_j}$ for the sample size n_j of the j^{th} cohort.

To define S_{Het} , we first let

$$S(\tau) = \frac{e^T(R(\tau)W(\tau))^{-1}T(\tau)(e^T(R(\tau)W(\tau))^{-1}T(\tau))^T}{e^TW(\tau)^{-1}R(\tau)^{-1}W(\tau)^{-1}e},$$

where $T(\tau)$ is the subvector of T satisfying $|T_{jk}| > \tau$ for a given $\tau > 0$, $R(\tau)$ is a sub-matrix of R representing the correlation matrix, and $W(\tau)$ is the diagonal submatrix of W , corresponding to $T(\tau)$. Here we let $w_{jk} = \sqrt{n_j} \times \text{sign}(T_{jk})$. Then the test statistic is $S_{Het} = \max_{\tau > 0} S(\tau)$.

The asymptotic distribution of S_{Het} does not follow a standard distribution but can be evaluated using simulation. S_{Het} is an extension of S_{Hom} that improves power when the genetic effect sizes vary for different traits. The distribution of S_{Het} under the null hypothesis can be obtained through simulations or approximated by an estimated beta distribution.

1.4 Discussion

We have reviewed several existing methods for detecting cross phenotype associations. Choosing an appropriate statistical approach depends on the study design, phenotypes, data availability and genetic heterogeneity. Systematic simultaneous analysis of multiple traits could improve the quality of inferences from analysis of outcomes that all relate to the biological construct of interest. Compared to multivariate approaches that model all outcomes simultaneously, combining statistics from the univariate analysis has the advantages of involving fewer assumptions about the relationships among individual phenotypes. For most published GWAS, obtaining summary data is substantially easier than accessing individual-level phenotype and genotype data. Compared with existing methods, CPASSOC has multiple advantages for identifying cross-phenotype associations. It can accommodate opposite risk effects and different types of phenotypic traits, either correlated, independent, continuous, or binary traits. It is well suited for overlapping or related subjects within and among different studies or cohorts. Since this approach accounts for the correlations of test statistics among traits or cohorts, S_{Hom} and S_{Het} are able to control the effect of cryptic relatedness occurring among cohorts.

2. Methods

In this section, we will illustrate the procedure for conducting CPASSOC analysis with the sex-specific summary statistics of the three traits: height, BMI and WHRadjBMI (38), which were downloaded from the GIANT consortium website (https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files).

The data we downloaded include sex-specific GWAS meta-analysis summary statistics from the discovery phase, including the SNP estimated effect sizes and their corresponding standard errors. For the discovery stage in sex-specific studies, 46 studies (up to 60,586 males, 73,137 females) were included on height and BMI, and 32 studies (up to 34,629 males, 42,969 females) were included on WHR. To apply CPASSOC for combining all the sex-specific summary statistics on height, BMI and WHRadjBMI, we chose 2,723,278 SNPs available for all three traits.

2.1 Step 1. Estimating the correlation matrix

The CPASSOC package is downloaded from <http://hal.case.edu/zhu-web/>. To perform CPASSOC analysis, a correlation matrix is required to account for the correlations among phenotypes or induced by overlapping or related samples from different cohorts. The correlations can be estimated from the summary statistics for all the independent SNPs in a genome-wide association study. Zhu et al. (26) suggested using a set of SNPs in linkage equilibrium to estimate the correlation coefficients. Table 2 gives an example of the estimated correlations among the three sex-specific traits: height, BMI and WHRadjBMI. Since only summary statistics are available in GIANT, we borrowed the linkage disequilibrium (LD) pattern from the ARIC European American (EA) data (downloaded from dbGaP <http://www.ncbi.nlm.nih.gov/gap>). The SNP set is obtained from applying pairwise LD pruning with $r^2=0.2$ ARIC EA data using the software PLINK (<http://pengu.mgh.harvard.edu/purcell/plink/>). SNPs with large effect sizes may represent true association, and consequently may inflate correlations among summary statistics. Therefore, we removed SNPs whose summary statistics Z scores were greater than 1.96 or less than -1.96 (See Note 1). Below is an example of R code showing how to estimate the correlation matrix.

```
# read the data: the summary statistics data file ("HeightBMIWHRadjBMI.txt")
# for CPASSOC includes seven columns (SNP ID, height Z-score for male, height
# Z-score for female, BMI Z-score for male, BMI Z-score for female, WHRadjBMI
# Z-score for male, WHRadjBMI Z-score for female. The SNP list file after
# pruning based on ARIC data is "ARICdot2.prune.in".
Dt<-read.table("HeightBMIWHRadjBMI.txt",header=T)
# extract the SNP set obtained from ARIC EA data.
pruned.lst<-read.table("ARICdot2.prune.in")
```

¹When estimating the correlation matrix, SNPs with large effect sizes may represent true association, and consequently may inflate correlations among summary statistics. Therefore, we suggest removing SNPs with large effect sizes from the correlation matrix estimation.

```

Dtcorr<-Dt[Dt$MarkerName%in%pruned.lst$V1,]
# Remove SNPs with Z-score larger than 1.96 or smaller than -1.96, assuming
the summary statistics for 6 sex-specific traits are named Z1,Z2,...,Z6.
Index<-which(abs(Dtcorr$Z1)>1.96 | abs(Dtcorr$Z2)>1.96 |abs(Dtcorr$Z3)>1.96
| abs(Dtcorr$Z4)>1.96 |abs(Dtcorr$Z5)> 1.96 | abs(Dtcorr$Z6)> 1.96)
Dtcorr<-Dtcorr[-Index,]
corMatrix<-diag(x=1,ncol=6,nrow=6)
for(i in 1:5){
for(j in (i+1):6)
corMatrix[i,j]=corMatrix[j,i]=cor(Dtcorr[,i+1],Dtcorr[,j+1])
}

```

2.2 Step 2. Perform CPASSOC analysis

The input file for CPASSOC software is an $M \times K$ matrix of summary statistics, where M is the number of SNPs and K is the number of summary statistics to be combined, with $M=2,723,278$ and $N=6$ in our example with anthropometric traits. Each column consists of summary statistics of M SNPs for one trait, SNP with missing values should be excluded (see Note 2). If multiple traits have been analyzed in a cohort, each phenotype's summary statistics will be put in one column. Each row represents a SNP. An example of the input file is presented in Table 3.

CPASSOC performs two tests: S_{Hom} and S_{Het} . S_{Hom} is more powerful when heterogeneity is not present, while S_{Het} allows for trait heterogeneity effects. The two tests can be performed using R code, as described below:

1. Perform S_{Hom} test.

```

source("FunctionSet.R")
# Extract the MxK matrix of summary statistics.
Sumstat<-subset(Dt,select=c("Z1","Z2","Z3","Z4","Z5","Z6"))
## Because each SNP may have a different sample size in the GIANT consortium
data, we used the median sample size across SNPs for each trait (see Note 3).
Samplesize<-c(60491.2,73046.2,58603.5,67938.6,34596.1,42730.2)
Test.shom<-SHom(Sumstat,Samplesize,corMatrix)
# obtain P-values of  $S_{Hom}$ 
p.shom<-pchisq(Test.shom, df = 1, ncp = 0, lower.tail = F)
res.shom<-data.frame(Dt[, "MarkerName"],p.shom)
names(res.shom)<-c("MarkerName","p.Shom")

```

2. Perform S_{Het} test. The empirical distribution of S_{Het} is approximated by a gamma distribution with a mean shift and we use simulations to estimate this gamma distribution. The three parameters: shape **k**, scale **theta** and shift **c**, are

²The current version of CPASSOC assumes no missing summary statistics. Therefore, SNPs with missing values should be removed before performing CPASSOC.

dependent on the trait correlations and the corresponding summary statistics. When there are no missing values, they can be estimated by calling the function `EstimateGamma`. When the dimension of `CorrMatrix` is large, directly using the empirical distribution is desirable (see Note 4). If we deal with a large dataset, it will be very useful to split the data into small pieces first.

```
#Create 20 equally size folds
folds <- cut(seq(1,nrow(Sumstat)),breaks=20,labels=FALSE)
res.shet=list()
#estimate parameters of gamma distribution.
para = EstimateGamma (N = 1E4, Samplesize, corMatrix) ;
for(i in 1:20){
#Segement your data by fold using the which() function
Indexes<- which(folds==i,arr.ind=TRUE)
Test.shet<-SHet(Sumstat,Samplesize,corMatrix)
# obtain P-values of  $S_{Het}$  using the estimated gamma parameters
p.shet = pgamma(q = Test.shet-para[3], shape = para[1], scale = para[2],
lower.tail = F);
res.shet[[i]]<-data.frame(Sumstat[Indexes,"MarkerName"],p.shet)
}
res.shet <-do.call(rbind.data.frame,res.shet)
names(res.shet)<-c("MarkerName","p.Shet")
```

The output of both tests is a vector of P -values for M SNPs. Because CPASSOC combines multiple traits, we can apply the significant level $P = 5 \times 10^{-8}$ as in GWAS. We examined loci that reached genome-wide significance ($P < 5 \times 10^{-8}$) by CPASSOC from sex-specific data, as presented in Fig 1. To do this, for a SNP reaching $P < 5 \times 10^{-8}$ by either S_{Hom} or S_{Het} we examined the region within 500 kb of each side of the SNP. The SNP was considered to be identified only by CPASSOC if no SNPs reached $P < 5 \times 10^{-8}$ by the conventional meta-analysis from the discovery phase data in the 1.0 Mb region, and the SNP is not in LD with sentinel SNPs from the GIANT studies.

Acknowledgment

This work was supported by a grant from National Heart Genome Research Institute (HG003054).

⁴.The empirical distribution of S_{Het} is approximated by a gamma distribution. However, the gamma distribution may not work well when the dimension of `CorrMatrix` is large; the `EstimateGamma` function may fail. In this case, directly using the empirical distribution is desirable. The function `EmpDist` can be used here. For example,

```
Stat=EmpDist (N = 1E6, Samplesize, CorrMatrix);
```

where `Stat` is an array of the empirical distribution. The length of `Stat` is by default $N=1E6$, which can be changed.

References

- (1). Welter D, MacArthur J, Morales J et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42: D1001–1006 [PubMed: 24316577]
- (2). Wagner GP, and Zhang J (2011) The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature reviews. Genetics* 12: 204–213
- (3). Ghoussaini M, Song HL, Koessler T et al. (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer I* 100: 962–966
- (4). Gudmundsson J, Sulem P, Manolescu A et al. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics* 39: 631–637 [PubMed: 17401366]
- (5). Tomlinson I, Webb E, Carvajal-Carmona L et al. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* 39: 984–988 [PubMed: 17618284]
- (6). Turnbull C, Ahmed S, Morrison J et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature genetics* 42: 504–U547 [PubMed: 20453838]
- (7). Grisanzio C, and Freedman ML (2010) Chromosome 8q24-Associated Cancers and MYC. *Genes & cancer* 1: 555–559 [PubMed: 21779458]
- (8). Begovich AB, Carlton VEH, Honigberg LA et al. (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* 75: 330–337 [PubMed: 15208781]
- (9). Bottini N, Musumeci L, Alonso A et al. (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature genetics* 36: 337–338 [PubMed: 15004560]
- (10). Kyogoku C, Langefeld CD, Ortmann WA et al. (2004) Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. *Arthritis Rheum* 50: S258–S258
- (11). Smyth D, Cooper JD, Collins JE et al. (2004) Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes* 53: 3020–3023 [PubMed: 15504986]
- (12). Siminovitch KA (2004) PTPN22 and autoimmune disease. *Nature genetics* 36: 1248–1249 [PubMed: 15565104]
- (13). Solovieff N, Cotsapas C, Lee PH, Purcell SM, and Smoller JW (2013) Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 14: 483–495 [PubMed: 23752797]
- (14). Ehret GB, and Munroe PB, and Rice KM et al. (2011) Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 478: 103–109 [PubMed: 21909115]
- (15). Franceschini N, and Fox E, and Zhang Z et al. (2013) Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. *Am J Hum Genet* 93: 545–554 [PubMed: 23972371]
- (16). Zeger SL, and Liang KY (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42: 121–130 [PubMed: 3719049]
- (17). Zhou X, and Stephens M (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 11: 407–+ [PubMed: 24531419]
- (18). O'Reilly PF, Hoggart CJ, Pomyen Y et al. (2012) MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PloS one* 7:
- (19). Zhang HP, Liu CT, and Wang XQ (2010) An Association Test for Multiple Traits Based on the Generalized Kendall's Tau. *J Am Stat Assoc* 105: 473–481 [PubMed: 20711441]
- (20). Lange C, Silverman SK, Xu X, Weiss ST, and Laird NM (2003) A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 4: 195–206 [PubMed: 12925516]
- (21). Ott J, and Rabinowitz D (1999) A principal-components approach based on heritability for combining phenotype information. *Hum Hered* 49: 106–111 [PubMed: 10077732]

- (22). Klei L, Luca D, Devlin B, and Roeder K (2008) Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol* 32: 9–19 [PubMed: 17922480]
- (23). Ferreira MA, and Purcell SM (2009) A multivariate test of association. *Bioinformatics* 25: 132–133 [PubMed: 19019849]
- (24). Aschard H, Vilhjalmsdottir BJ, Greliche N, Morange PE, Tregouet DA, and Kraft P (2014) Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet* 94: 662–676 [PubMed: 24746957]
- (25). Panagiotou OA, Willer CJ, Hirschhorn JN, and Ioannidis JP (2013) The power of meta-analysis in genome-wide association studies. *Annu Rev Genomics Hum Genet* 14: 441–465 [PubMed: 23724904]
- (26). Zhu X, Feng T, Tayo BO 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 [PubMed: 25500260]
- (27). Cotsapas C, Voight BF, Rossin E et al. (2011) Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* 7: e1002254 [PubMed: 21852963]
- (28). Bhattacharjee S, Rajaraman P, Jacobs KB et al. (2012) A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. *Am J Hum Genet* 90: 821–835 [PubMed: 22560090]
- (29). Han B, and Eskin E (2011) Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. *Am J Hum Genet* 88: 586–598 [PubMed: 21565292]
- (30). O'Brien PC (1984) Procedures for comparing samples with multiple endpoints. *Biometrics* 40: 1079–1087 [PubMed: 6534410]
- (31). Xu X, Tian L, and Wei LJ (2003) Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* 4: 223–229 [PubMed: 12925518]
- (32). Yang Q, Wu H, Guo CY, and Fox CS (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic epidemiology* 34: 444–454 [PubMed: 20583287]
- (33). Wain LV, and Verwoert GC, and O'Reilly PF et al. (2011) Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat Genet* 43: 1005–1011 [PubMed: 21909110]
- (34). Heald AH, Siddals KW, Fraser W et al. (2002) Low circulating levels of insulin-like growth factor binding protein-1 (IGFBP-1) are closely associated with the presence of macrovascular disease and hypertension in type 2 diabetes. *Diabetes* 51: 2629–2636 [PubMed: 12145180]
- (35). Rajwani A, Ezzat V, Smith J et al. (2012) Increasing circulating IGFBP1 levels improves insulin sensitivity, promotes nitric oxide production, lowers blood pressure, and protects against atherosclerosis. *Diabetes* 61: 915–924 [PubMed: 22357965]
- (36). Ganesh SK, Chasman DI, Larson MG et al. (2014) Effects of long-term averaging of quantitative blood pressure traits on the detection of genetic associations. *Am J Hum Genet* 95: 49–65 [PubMed: 24975945]
- (37). Warren HR, Evangelou E, Cabrera CP et al. (2017) Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet* 10.1038/ng.3768:
- (38). Park H, Li XY, Song YE, He KRY, and Zhu XF (2016) Multivariate Analysis of Anthropometric Traits Using Summary Statistics of Genome-Wide Association Studies from GIANT Consortium. *PloS one* 11: e0163912 [PubMed: 27701450]
- (39). Willer CJ, Li Y, and Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26: 2190–2191 [PubMed: 20616382]
- (40). Yang JJ, Li J, Williams LK, and Buu A (2016) An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC bioinformatics* 17: 19 [PubMed: 26729364]
- (41). Kavvoura FK, and Ioannidis JPA (2008) Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum Genet* 123: 1–14 [PubMed: 18026754]

Figure 1. Manhattan plots of S_{Hom} and S_{Het} for combining three sex specific traits (38). The loci reaching genome-wide significance by S_{Hom} and S_{Het} but not by the conventional meta-analysis, are marked with corresponding SNP names. (A) S_{Hom} (B) S_{Het} .

Table 1.

Methods for detecting CP associations.

Methods	References	Input	Allow Overlapping Subjects	Combine data across multiple studies	Account for correlation	Allow heterogeneity effects
GEE	(16, 17)	individual-level data	Yes	No	Yes	No
PC analysis	(21, 22)	individual-level data	Yes	No	Yes	No
CCA	(23)	individual-level data	Yes	No	Yes	No
Fisher's p value	(40)	P value	No	Yes	No	Yes
CPMA	(27)	P value	No	Yes	No	Yes
Fixed and random effects meta-analysis	(41)	Summary statistics	No	Yes	No	No (fixed effects) Moderate level (Random effects)
subset-based meta-analysis	(28)	Summary statistics	No	Yes	No	Yes
Extensions to O'Brien's method	(31, 32)	individual-level data	Yes	No	Yes	Yes
CPASSOC	(8, 26, 38)	Summary statistics	Yes	Yes	Yes	Yes

Table 2.
Correlations for sex-specific cohorts for height, BMI, and waist-to-hip ratio adjusted for BMI.

The SNP sets for correlation estimation include 81,322 SNPs for height, 82,012 SNPs for BMI, and 81,130 SNPs for WHRadjBMI from the GIANT consortium studies.

		Height		BMI		WHRadjBMI	
		Male	Female	Male	Female	Male	Female
Height	Male	1	0.128	−0.038	−0.007	−0.004	−0.002
	Female	0.128	1	−0.012	−0.056	0.007	−0.003
BMI	Male	−0.038	−0.012	1	0.096	0.010	−0.007
	Female	−0.007	−0.056	0.096	1	0.011	0.001
WHRadjBMI	Male	−0.004	0.007	0.010	0.011	1	0.031
	Female	−0.002	−0.003	−0.007	0.001	0.031	1

Table 3.
The input file for CPASSOC

This is the start of the $M \times K$ matrix of summary statistics ($M = 2,723,278$ and $N = 6$). Each column represents the summary statistics from M SNPs for one trait. Each row represents a SNP.

rs10	0.404545	-0.00455	0.326087	-0.0087	-0.0087	-1.04348
rs1000000	1.486486	0.362319	-0.67532	-0.51351	-0.525	-0.60759
rs10000010	0.491803	-0.07018	0.3125	-0.72581	1.911765	2.575758
rs10000012	2.247191	0.313253	1.914894	1.222222	-0.61	0.387755
...						