BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 01/06 pp173–190
Volume 1, Number 3, September 2018
DOI: 10.26599/BDMA.2018.9020015

A Survey of SNP Data Analysis

Xiaojun Ding and Xuan Guo*

Abstract: Every person differs from every other person regarding their physical appearance, susceptibility to disease, response to medications, and so on. However, 99.9 percent of human DNA is the same. As such, differences in human genomes are very worthy of study. Single-Nucleotide Polymorphisms (SNPs) are the simplest form and most common source of genetic polymorphism. SNPs have been used to successfully identify defective genes that cause Mendelian diseases. However, most common human diseases are complex and are caused by multiple SNPs. Each SNP explains only a small fraction of genetic causes. Experiments on individual SNPs may reveal their non-detectable effects on complex diseases. Pathogenesis is a complicated topic, and it is difficult to correctly predict multiple SNPs. As such, the analysis of SNP data is a critical task in the study of genetic diseases. In this paper, we divide the methods for genome-wide SNP data analysis into two categories: single-trait Genome-Wide Association Studies (GWAS) in which pathology is mined from data of a single phenotype, and multiple-trait GWAS which identifies cross-phenotype associations. For single-trait GWAS, we review methods ranging from the simple to the complex, including TEAM, BOOST, AntEpiSeeker, SNPRuler, EDCF, HiSeeker, ORF, MLR-tagging, MSCD, and MIC. For multiple-trait GWAS, we describe methods in terms of their employed regression models, dimension-reduction methods, and meta-analysis methods. We also list the advantages and disadvantages of these methods. Finally, we discuss the future directions of SNP data analysis for genome-wide association.

Key words: SNP interactions; SNP combinations; GWAS; case-control study; disease association analysis; cross-phenotype association studies

1 Introduction

Most human genomes are similar with only a relatively few genetic differences between any two randomly selected human genomes[1]. However, these minor differences lead to a wide variety of human characteristics. Some people have black hairs and black eyes whereas others have blond hairs and blue eyes. Their responses to the same drug may also differ. The genetic differences between people merit careful investigation. A Single-Nucleotide Polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide differs in a pair of homologous chromosomes. SNPs are domain genetic differences[2]. The high throughput technique has generated a large volumn of SNP data. One goal of examining SNP data is to discover underlying rules hidden in the data. SNP data contains much information including causative mutation, evolutionary history, and population differences.

Biological and medical scientists are most interested in causal SNPs and have identified the genetic causes of many single-gene diseases. A single-gene disease

• Xiaojun Ding is with School of Computer Science and Engineering, Yulin Normal University, Yulin 537000, and School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China. E-mail: ding.xiaojun@foxmail.com.
• Xuan Guo is with Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203-5017, USA. E-mail: xuan.guo@unt.edu.
* To whom correspondence should be addressed.
Manuscript received: 2018-01-12; accepted: 2018-01-17

refers to a disease or pathological trait that is controlled by a pair of alleles. For example, Prescott et al.[3] identified a nonsynonymous SNP in ATG16L1 that is related to Crohn's disease. Seki et al.[4] reported that a functional SNP in cartilage intermediate layer protein is suspected as being related to lumbar disc disease. Zaimkohan et al.[5] reported that PCSK9 SNP rs11591147 is associated with risk of coronary artery disease in Iran.

As research continues, scientists will address more complex diseases. Complex diseases in the general population have a higher incidence (generally not less than 1%). A complex disease is often controlled by two or more alleles and is determined by genetic factors as well as environmental factors[6]. Also, there are often complicated non-linear relationships between genes and the environment. As such, pathogenesis is difficult to determine in normal individuals and in diseased individuals in one family. Currently, scientists can only determine the role of genetic factors in the occurrence of polygenic disease by studying a large number of patients. Age-related Macular Degeneration (AMD), for instance, is a complex disease and the leading cause of legal blindness in adults in the US. Klein et al.[7] analyzed the SNP data of the genotypes of 103 611 SNPs of 96 affected individuals and 50 healthy individuals and reported that two SNPs (rs380390 and rs1329428) are associated with AMD disease. Shen et al.[8] exhaustively evaluated all pairs of SNP-SNP interactions for 661 658 SNPs in 5269 cases and 5289 controls. The authors reported two SNPs, rs1105255 and rs651431, to be related to prostate cancer. These finding have inspired the development of methods for identifying causal SNPs.

In addition to examining one trait at a time, a significant number of candidate gene studies and Genome-Wide Association Studies (GWAS) have revealed that a genetic variant might be associated with more than one trait[9–19]. For example, many studies have shown that variants in the DSP gene are associated with chronic obstructive pulmonary disease status and lung function traits[20]; variants in protein tyrosine phosphatase non-receptor type 22 were identified as being associated with immune-related disorders, including rheumatoid arthritis[21], systemic lupus erythematosus[22], and type 1 diabetes[23]; and a genetic risk factor located in the human leukocyte antigen region was reported as being strongly associated with multiple diseases and traits, including type 1 diabetes mellitus, multiple sclerosis, and rheumatoid arthritis[21, 24]. The GWAS Catalog[25] is a manually curated resource of published GWAS and association results. In 2011, the authors of a study analyzing the GWAS Catalog concluded that 4.6% of SNPs are associated with more than one distinct traits[26]. The fact that genetic variants are associated with multiple and sometimes seemingly distinct trait is known as Cross-Phenotype (CP) association[27]. Recently, many methods have been developed for use in CP Association Studies (CPAS) which investigate at detecting genetic variants that affect multiple traits.

Traditional methods for analyzing genetic diseases are based on a hypothesis. First, scientists guess which candidate genes are associated with a particular disease, and then they conduct extensive experiments to test this hypothesis. This approach consumes lots of time and money. Modern methods are data-driven and can reveal the pathogenesis of whole genomes based on data from large populations. However, these methods have several challenges. First, the volume of data is tremendous. It is estimated that on average there is an SNP in every 300 bp in DNA sequences, with about 11 million SNPs in the whole human genome. Also, one of the commonly used electronic health records for measuring phenotypes has more than 13 000 codes. Considering the size of SNPs and the number of phenotypes, the number of SNPs and traits combinations to be tested is exponential. It is computationally impossible to investigate all SNP/trait combinations. Second, the data may be incomplete and contain a lot of noises due to technology limitations and unexpected observation errors. Third, robust and universal evaluative criteria have not yet been constructed because the pathogenesis of many complex diseases is still unknown. This review is intended to offer an overview of available methods for SNP data analysis, especially single-trait and multi-trait GWAS (CPAS), and their relative strengths and limitations.

2 Problem Statement

If mutations in some SNP loci lead to the establishment of different phenotypes, the distribution of SNP patterns must differ populations with different phenotypes. These SNP loci are as yet unknown. The job of scientists is to collect the SNP data of individuals with different phenotypes and identify the SNP loci where mutations are significantly associated with these phenotypes. Many strategies have been designed to

mine suspicious SNP combinations, and robust and powerful statistical tests have been created to extract significant combinations.

Because most SNPs are bi-allelic, removing triangular-allelic and quadrilateral-allelic SNPs can simplify the problem without loss of much useful information. The minor allele in a bi-allelic SNP is often denoted by a lowercase letter, and the major allele by an uppercase letter, such as a and A, respectively. In an SNP locus, there are three genotypes: the homozygous reference genotype (AA), the heterozygous genotype (Aa), and the homozygous variant genotype (aa). In raw data, they are usually encoded as 0, 1, and 2, respectively.

3 Analysis for Single Phenotype

In this section, we review existing methods for analyzing SNP data for a single trait, from the simple to the complex. Researchers have achieved great success in detecting single causal SNPs. Inspired by this, there has been a movement toward the detection of k-locus SNP. However, k-locus SNP is much more complicated and the difficulty is closely related to the size of the SNP combination. As this size increases, the number of disease models and SNP combinations increases exponentially and the disease models and algorithms for single SNP detection are not applicable. Researchers have developed various evaluation functions and methods to handle these challenges and there are many ways to categorize them. Here, we list these methods by their main underlying concepts and note that some combine several technologies.

3.1 Single SNP detection

Table 1 lists three disease models for identifying single causal SNPs including the dominant, recessive, and additive disease models.

Table 1 Disease models for single SNPs.

Dominant disease model			
Genotype	AA	Aa	aa
Risk	0	1	1
Recessive disease model			
Genotype	AA	Aa	aa
Risk	0	0	1
Additive disease model			
Genotype	AA	Aa	aa
Risk	0	0.5	1

In the dominant model, as long as there is a mutation on one chromosome, the risk of disease is denoted as 1. The risk is denoted 0 when there is no mutation on either chromosome. In short, the risk of illness will increase if a mutation appears. The risk is not affected by the number of mutation occurrences. In the recessive model, the risk of illness is greater than 0 only when the two chromosomes on the genetic loci have mutated. This means that only a homozygous "aa" can cause the disease. In the additive model, when one chromosome has mutated, the risk of illness is 0.5 and when both genes on the chromosome have mutated, the risk increases to 1. The risk of disease increases as the emergence of mutation increases.

Of course, the above three models are theoretical. In reality, although mutation can increase risk, the risk is also affected by environmental and other factors. Nor is the probability exactly 1 or 0.5, but a statistical test is often used in the evaluation. SNPs are regarded as suspicious pathogenic SNPs if their statistical test values are below a P-value threshold.

These three models are simple, intuitional, and widely recognized. Once a disease model is constructed, it remains to then investigate all the SNP data on the whole genome. The time complexity of this task is $O(n)$, where n is the number of SNPs. The performance of an exhaustive search is sufficient and since this process loses no useful information, it is considered to be the best method.

3.2 Two-locus SNP detection

Figure 1 illustrates the natural idea to detect two-locus SNP combinations, that is to build the two-locus disease models and to investigate all the possible SNP combinations, as in the single SNP detection. Since there are three types for each SNP, there are

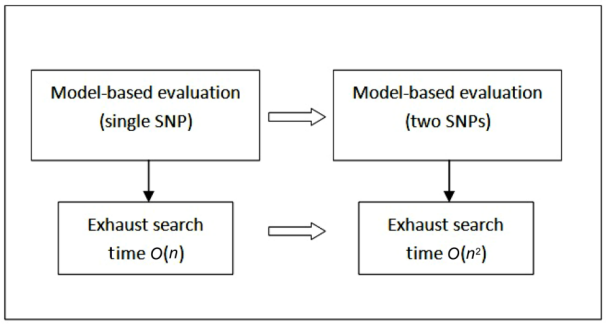


Fig. 1 From single SNP detection to two-locus SNP detection.

a total of $3^2 = 9$ genotypes for two-locus SNP and the relationship of these SNPs may be nonlinear. Researchers have divided them into marginal (main) effects (eME) and non-marginal effects (eNME)[28, 29] and have also introduced the parameters penetrance $p(D)$, odds ODD_{gi}, and $p(D|g_i)$ for the construction of these disease models, where $p(D|g_i)$ is the probability that an individual will be affected with a given genotype combination g_i. The relationships between penetrance $p(D)$, odds ODD_{gi}, and $p(D|g_i)$ are shown in Eqs. (1)–(3).

$$ODD_{gi} = \frac{P(D|g_i)}{P(\bar{D}|g_i)} = \frac{P(D|g_i)}{1 - P(D|g_i)} \tag{1}$$

$$p(D) = \sum_i p(D|g_i)p(g_i) \tag{2}$$

$$\frac{p(D|g_A)}{p(\bar{D}|g_A)} = \frac{\sum\limits_{g_B} p(D|g_A, g_B)p(g_B)}{\sum\limits_{g_B} p(\overline{D}|g_A, g_B)p(g_B)} \tag{3}$$

After determining the odds, a disease model can be described. There are many possible disease models. Table 2 lists two models with marginal effects, in which Model 1 is multiplicative and Model 2 has been used to describe handedness and the color of swine.

Once these models are constructed, the next step is much like the detection of a single SNP. The simplest way to detect causal SNPs is to exhaustively investigate all possible combinations in the disease models. However, the situation is more complicated than with that of a single SNP, in that there are many possible disease models. Velez et al.[30] generated 70 different penetrance functions that define a probabilistic relationship between a genotype and a phenotype. The number of SNP combinations also rapidly increases. For n SNPs, the number of all possible two-loci SNP combinations is C_n^2. When n is large, the computation costs are prohibitive.

Tree-based Epistasis Association Mapping (**TEAM**)[31] is a method by which all possible two-locus SNPs are exhaustively investigated. The authors who developed TEAM realized that there were many unnecessary and repetitive calculations in previous approaches. If two SNPs have the same genotypes on most samples, the computation of contingency tables can be shared by considering only samples with different genotypes[32]. To reduce the computational cost, TEAM utilizes a minimum spanning tree to maximize the shared computation of the contingency tables. TEAM controls false positives using a permutation test. The performance of TEAM is faster than the brute-force approach, but is still very slow for massive amounts of data.

Table 2 Odds tables of disease Models 1 and 2.

Model 1	BB	Bb	bb
AA	α	α	α
Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
aa	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$
Model 2	**BB**	**Bb**	**bb**
AA	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
Aa	$\alpha(1+\theta)$	α	α
aa	$\alpha(1+\theta)$	α	α

Boolean Operation-based Screening and Testing (**BOOST**) is a model-based exhaustive search method[33]. The authors of this method used a new Boolean representation to accelerate the collection of contingency tables; then, they used an upper bound for the likelihood ratio test based on log-linear models and the Kirkwood superposition approximation[34] for prune searching. Their model considers only those SNPs with no marginal effects. BOOST is very fast and takes less than 60 hours to complete an evaluation of all the pairs of roughly 360 000 SNPs on a standard 3.0 GHz desktop with 4 GB memory running the Windows XP system. Its code is available at http://bioinformatics.ust.hk/BOOST.html.

GBOOST, a tool that uses GPU parallel programming to speed up the BOOST[35], which achieves a 40-fold speedup compared with BOOST on a desktop computer equipped with a Nvidia GeForce GTX 285 display card. Its code is available at http://bioinformatics.ust.hk/BOOST.html#GBOOST.

3.3 *k*-locus SNP detection

When the size of combined SNPs increases, the number of possible disease models rapidly escalates and the number of potential SNP combinations increases exponentially. As such, investigating all combinations becomes impossible. Figure 2 exhibits the main ideas of k-locus SNP detection methods, researchers use an odds ratio, mutual information or χ^2 instead of disease models in the evaluation and they replace exhaustive search with a heuristic, stepwise search or feature selection method to accelerate the process. The details of these methods are introduced in the following paragraphs.

3.3.1 Heuristic search

Genetic Algorithm (GA), Ant Colony Optimization

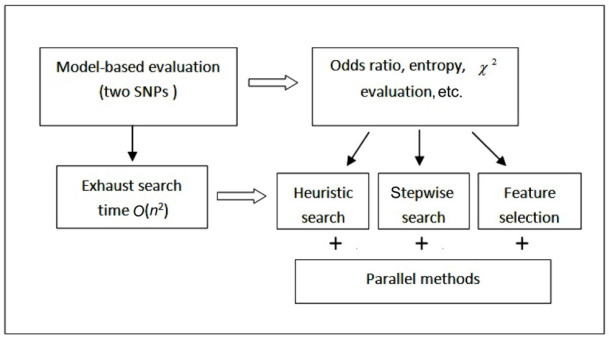


Fig. 2 From two-locus SNP detection to *k*-locus SNP detection.

(ACO), Particle Swarm Optimization algorithm (PSO), and Simulated Annealing algorithm (SA) are commonly used and powerful heuristic search methods. Actually, these are all metaheuristic methods, which make just a few or no assumptions about the problem being optimized and can search vast spaces for candidate solutions. As such, it is natural to adopt them in the search stage.

GA, first proposed by John Holland in 1975[36], is inspired by the process of natural selection. The main process of GA is as follows:

Step 1. A genetic representation of the solution domain and a fitness function are proposed for evaluating the solution domain.

Step 2. A large number of individuals are randomly generated as a population, with each individual corresponding to a possible solution.

Step 3. In a population, good individuals are likely to breed the next generation, in which the individuals are evaluated based on their fitness. In this process, a sub-population comprising good individuals is selected.

Step 4. The next generation is generated from the sub-population using a combination of genetic operators known as crossover and mutation.

Step 5. Steps 3 and 4 are repeated until a termination condition has been reached.

GA optimizes a problem by iteratively improving the populations of candidate solutions with regard to their fitness by selecting the superior and eliminating the inferior.

$$OR_T = \frac{ODD_T}{ODD_C} = \frac{\text{d0/h0}}{\text{d1/h1}} \tag{4}$$

GA is better at single-objective optimization and is not suitable for multiple disease models. Chuang et al.[37] adopted the odds ratio as a score function, which is calculated as shown in Eq. (4), where d0, d1, h0, and h1 are allele counting shown in Table 3. $OR_T = 1$ indicates no association between genotype and disease, $OR_T > 1$ indicates that the T allele is associated with the disease, and $OR_T < 1$ indicates that the T allele is protective. To explore SNP-SNP interactions in breast cancer, the authros designed a GA that analyzes multiple independent SNPs. They reported that SNP-SNP interactions with a high risk of breast cancer could be successfully predicted using the GA method. Chen et al.[38] and Yang et al.[39] also used the GA algorithm to analyze chronic dialysis susceptibility and breast cancer, respectively.

Similar to GA, the PSO, SA, and ACO methods also obtain the better solutions by iteratively improving existing solutions.

The ACO, first proposed by Dorigo and Gambardella[40], simulates the behaviors of real ant colonies. In the natural world, in the absence of pheromones, ants wander randomly. When they find food, they return to their colony while laying down pheromone trails. If a pheromone is present, ants will follow the path with the higher pheromone concentration. Ants that choose a shorter path traverse a distance more quickly, resulting in more pheromones being deposited along that path. Ultimately, most ants will follow the shortest path. This algorithm has been successfully used in many fields.

The **AntEpiSeeker** algorithm, developed by Wang et al.[41], uses a two-stage ACO algorithm to detect SNP interactions in case-control studies. χ^2 values are used as score functions to determine the association between multiple SNPs. This tool is available at http://nce.ads.uga.edu/ romdhane/AntEpiSeeker/index.html.

PSO is a computational method that optimizes a problem by identifying a population of candidate solutions, known as particles, and moving these particles around in the search-space according to simple mathematical formulae with respect to particle position and velocity. Each particle's movement is influenced by its local best-known location, but is also guided toward the best known positions in the search space, which are updated as better positions are identified by other particles. This results in the swarm moving toward the

Table 3 Allele counting.

Allele	Case	Control
T	d0	h0
C	d1	h1

best solutions.

Chuang et al.[42] used a PSO algorithm to identify the significant multi-SNP combinations.

Wu et al.[43] adopted a PSO algorithm to identify SNP interactions in renin-angiotensin system genes with respect to hypertension.

SA is a probabilistic technique for approximating the global optimum of a given function. Specifically, SA is often used when the search space is discrete. For problems in which finding an approximate global optimum is more important than finding a precise local optimum in a fixed amount of time, SA may be preferable to alternatives such as gradient descent.

Kim et al.[44] adopted an SA algorithm to identify relevant SNP sets and their experimental results showed that this method could obtain a new set of variants using a reduced number of variants, which improved prediction performance compared to other algorithms that use traditional feature selection.

These heuristic search algorithms obtain better solutions iteratively by improving upon existing solutions. They can only obtain the local optimal solution by searching within a small portion of the whole solution space. For small volumes of data, the local optimal solution may exhibit good performance. However, when the solution space is vast and the solutions are discontinuous, these algorithms lose many good solutions, and the local optimal solution obtained is not the high-probability global optimal solution.

3.3.2 Stepwise search methods

In 1994, Agrawal and Srikant[45] proposed the Apriori algorithm, which uses a "bottom-up" approach to generate candidate item sets of length k by extending one item from item sets of length $k-1$. Then it prunes candidates who have an infrequent subpattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. The algorithm terminates when no further successful extensions are found.

The Apriori algorithm has achieved great success in the data mining field. Since it was first proposed, researchers have proposed many stepwise search methods similar to this method.

The **SNPRuler**[46] is a stepwise search method that uses rules to describe the cause of a disease. For example, $s_i = 0 \wedge s_j = 2 \rightarrow disease$ is a rule, where s_i and s_j are the SNP values at loci i and j. The SNPRuler uses the χ^2 test value to measure the quality of a rule, which gives an upper bound of the χ^2 test value to replace the downward closure lemma. Then, it remains to extend candidate item sets until no further successful extensions are found. Although the SNPRuler can find high-order SNP combinations, there are two problems with this method. First, the upper bound is based on the assumption that the number of cases is larger than or equal to the number of controls. Second, the upper bound derived from the χ^2 formula is not a true upper bound and does not possess the anti-monotone property[47].

The Epistasis Detector based on the Clustering of relatively Frequent items (**EDCF**), proposed by Xie et al.[48], is a novel statistical method based on the clustering of relatively frequent items to detect multi-locus epistatic interactions in case-control studies. EDCF groups all genotype combinations into three clusters, representing frequent genotypes in cases, frequent genotypes in controls, and the remaining genotypes. In the three groups, items for higher-order interactions are constructed sequentially. The significance of the final partitions can be evaluated by Pearson's χ^2 test. EDCF first selects all significant two-SNP combinations and then extends them to k-locus SNP combinations until no more significant combinations can be found.

For interaction detection, **HiSeeker**[49] employs the χ^2 test and the logistic regression model to obtain candidate two-locus SNP combinations that have intermediate or significant associations with the phenotype. Then, two strategies (exhaustive and ACO-based search) are employed to detect high-order interactions by extending these candidate combinations. In two real case-control datasets, HiSeeker has detected several significant high-order combinations whose individual SNPs and pairwise interactions have no strong main effects or pairwise interaction effects[49].

Compared with exhaustive approaches, stepwise algorithms usually run much faster and can perform reasonably well for diseases with some marginal effects. However, the lemmas used in current stepwise search methods, which substitute for the downward closure lemma, are not entirely correct or complete, which means that SNP combinations with small or no marginal effects cannot be found.

3.3.3 Feature selection and search methods

Because there are a vast number of SNPs, it is impossible to investigate all SNP combinations. If a

method can identify the most likely causal pathogenic SNPs and then check all their combinations, it can significantly reduce the associated computation cost. Fortunately, several algorithms can make efficient feature selections, including linear regression, Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), and Deep Neural Network (DNN).

The Optimum Random Forest (**ORF**), proposed by Mao and Lee[50], uses a simple feature selection method that first sorts all SNPs and then identifies the m most significant disease-associated SNPs for a given threshold. Next, ORF randomly generates many classification trees based on the m SNPs and selects trees that can distinguish between cases and controls with high accuracy. The causal SNP combinations can then be extracted from these trees.

Mao and Lee[50] did not consider the relationships between SNPs when finding the m most significant SNPs. To address this issue, linear regression can be applied. The purpose of linear regression is to learn the relationship between the input and response variables. The linear regression model is expressed as shown in Eq. (5):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad (5)$$

$$Y = X\beta + \varepsilon \quad (6)$$

where y_i is the label of the i-th individual, x_{ij} is the genotype of the i-th individual on the j-th SNP locus, β_j is the regression coefficient, and ε_i is the model error. If all individuals are denoted as Y and all SNP data are denoted as X, the expression can be written as shown in Eq. (6). The goal of linear regression is to estimate the coefficients β to minimize the total error $||\varepsilon||$ for all individuals. The j-th SNP is strongly related to the disease state if β_j has a large absolute value.

MLR-tagging, a software package tool developed by He and Zelikovsky[51], uses multiple linear regression to select informative SNPs for examining their relationship with corresponding traits.

$$\beta^* = \arg\min_{\beta}\{\frac{1}{2}||Y - X\beta||_2 + ||\beta||_1\} \quad (7)$$

There may be many $|\beta_j|$ with absolute values higher than 0, so many candidate SNPs will be selected. To improve this approach, Feng et al.[52] used the Least Absolute Shrinkage and Selection Operator (LASSO) method and the sparse Least squares regression methods to select SNPs for the prediction of quantitative traits. The LASSO algorithm finds the β^* that satisfies Eq. (8). The strength of the LASSO penalty can be tuned to select a predetermined number of the most relevant SNPs. Wu et al.[53] used LASSO penalized logistic regression to analyze case-control data.

The Multi-SNP Combination set Detector (**MSCD**) is a feature selection method proposed by Ding et al.[54] The MSCD regards an individual's genotype data on a list of SNPs as a point with a unit of energy in a multi-dimensional space and finds a new coordinate system where the energy distribution difference between cases and controls is maximum. The energy difference on an axis α can be calculated as $|\alpha' \frac{HH'}{t-1}\alpha - \alpha' \frac{DD'}{r-1}\alpha|$, where H is a matrix in which each column corresponds to the sequenced SNP data of one individual. D is the matrix comprising the SNP data of healthy individuals. t is the number of controls, and r is the number of cases. The α value that satisfies Eqs. (8) and (9) is calculated. Each component of the axis corresponds to an SNP locus, and the absolute value of the component indicates the importance of the corresponding SNP locus in differentiating between cases and controls. SNPs with larger absolute values are selected as candidate SNPs. Then a pruning tree-search strategy is used to identify significant k-locus SNP combinations.

$$\alpha = \arg\max_{\alpha} |\alpha' \frac{HH'}{t-1}\alpha - \alpha' \frac{DD'}{r-1}\alpha| \quad (8)$$

and

$$\| \alpha \| = 1 \quad (9)$$

The Maximal Information Coefficient (**MIC**), proposed by Leem et al.[55], uses the mutual information of an interaction between two SNPs as the evaluation function. The mutual information is defined as shown in Eq. (10):

$$I(S_i, S_j|y) = \sum p(g_i, g_j, y) \log \frac{p(g_i, g_j|y)}{p(g_i|y)p(g_j|y)} \quad (10)$$

where S_i and S_j are the two SNP loci, y is the sample label, g_i is the genotype of i-th SNP, and $p()$ is the probability. Accordingly, the evaluation function of the high-order interaction is defined as follows:

$$I(S_i, ..., S_j|y) = \sum p(g_i, ..., g_j, y) \log \frac{p(g_i, ..., g_j|y)}{p(g_i|y)...p(g_j|y)} \quad (11)$$

$I(S_i, ..., S_j|y) > 0$ indicates that these SNPs may work together and have a close relationship with the disease.

MIC uses k-means algorithm to cluster all the SNPs characterized by a certain distance between the mutual information of two SNPs. In each cluster, the top d SNPs are selected based on their scores, which are the

sum of all the mutual information values between the selected SNP and the rest of the SNPs in the same cluster. After the selection of kd SNPs in k clusters, an exhaustive search is conducted to find significant k-locus SNP combinations. Authors reported that MIC made some interesting and meaningful observations on seven diseases in Wellcome Trust Case Control Consortium data. The MIC algorithm is very fast because the kd value is much smaller than the total number of SNPs.

4 Cross-phenotype Association Studies

We extend our focus from single-trait GWAS to the detection of CP effects and pleiotropy. In contrast to single-trait GWAS, multi-trait analyses differ considerably in the statistical methods they apply. CP associations may be due to many factors. One of the biological factors is pleiotropy[27,56], which is defined in genetics, as when a genetic variant affects multiple traits. The recently increased availability of GWASs with detailed phenotypic data from Electronic Health Records (EHR) and epidemiological studies has stimulated an increasing number of population-based CPAS. EHR data includes a subject's current vital signs and present and past health conditions, as well as any diagnostic procedures, laboratory profiles, and clinical interventions. One of the commonly used coding systems for EHR data is the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes, which contain about 13 000 codes.

A joint analysis of correlated phenotypes can explore the correlation between phenotypes and also has other advantages. (1) CP analysis may enable the detection of genetic variants with only small effects across multiple traits. (2) Joint study can avoid penalties for multiple tests associated with the individual analysis of phenotypes. (3) Significant CP outcomes may identify hidden connections between seemingly unrelated diseases and guide investigators to the potentially shared pathways of these disorders. Therefore, today, methods that can fully utilize information from multiple phenotypes in the detection of novel genetic loci are attracting more attention.

With respect to the degree of assessed genetic overlaps, CPAS methods can be broadly classified into three categories: genome-wide, regional, and single variant. Genome-wide CPAS methods often use an initial measurement of the genetic overlap between two or more traits. Region-based methods cluster variants into groups based on some criteria, such as LD-blocks or gene boundaries, and then test or estimate the CP effects within a group. Such approaches can increase their power by combining related information across biologically meaningful units. Variant-level methods test each variant individually before performing a combined analysis, and may fail to identify these CP effects unless all relevant variants pass the significance threshold. Regression modeling is the most common approach in CPAS. In the following subsections, we introduce several CPAS regression models that are used when subject-level phenotype and genotype data are available. Next, we discuss the dimension-reduction methods used in CPAS regression modeling. Instead of directly incorporating multiple phenotypes into the regression models, dimension reduction can reduce the number of phenotypes and use traditional GWAS methods on the combined phenotypes. We then describe CPAS summary statistics methods. Subject-level phenotype and genotype data from GWAS analyses are often not accessible to researchers, due to logistical and data confidentiality reasons, many meta-analysis methods that use univariate GWAS phenotype summary test statistics, which are typically available. Table 4 summarizes recently developed methods for detecting CP associations, any of which can due to pleiotropy. However, these methods cannot determine that the identified associations are genuinely caused by pleiotropy, that is, that the genetic markers directly affect all the multiple phenotypes. Therefore, we investigate methods for identifying pleiotropy.

4.1 Regression modeling

Regression modeling is a statistical approach for assessing the relationship between variables. For example, logistic regression is used to evaluate the relationship between genetic markers and disease status, and linear regression is used to assess an association with a continuous trait, such as blood lipid levels. Other regression models, such as generalized mixed effects models, generalized estimation equations, and frailty models are also often used to test the associations of genetic markers with continuous, categorical, or survival multivariate phenotypes. Here, we briefly introduce two to illustrate the basic concepts of applying regression modeling for CPAS.

Liu et al.[57] proposed an Extended Generalized

Table 4 Methods used in cross-phenotype association analyses.

Ref.	Number of SNPs	Trait type	P-value calculation	Implementation	Year
Regression modeling					
[57]	1	Any	Asymptotic theory	R	2009
[58]	1	Any	Permutation	R	2010
[59]	1	Continuous	Asymptotic theory	Perl & R	2011
[60]	$\geqslant 1$	Any	Asymptotic theory	R	2011
[61]	$\geqslant 1$	Continuous	Asymptotic theory	R	2012
[62]	1	Continuous	Asymptotic theory	R	2012
[63]	$\geqslant 1$	Continuous	Permutation	R	2014
[64]	$\geqslant 1$	Continuous	Asymptotic theory	R	2015
[65]	1	Any	Asymptotic theory	R	2015
[66]	$\geqslant 1$	Continuous	Asymptotic theory	C++	2015
[67]	$\geqslant 1$	Continuous	Asymptotic theory	R	2015
[68]	$\geqslant 1$	Continuous	Permutation	Matlab	2015
[69]	$\geqslant 1$	Any	Asymptotic theory	Python	2015
[70]	$\geqslant 1$	Continuous	Permutation	R	2016
[71]	$\geqslant 1$	Continuous	Asymptotic theory	R	2016
[72]	$\geqslant 1$	Continuous	Asymptotic theory	R	2017
[73]	$\geqslant 1$	Any	Asymptotic theory/ permutation	R	2017
Dimension reduction					
[74]	1	Continuous	Asymptotic theory	Fortran	2008
[75]	$\geqslant 1$	Continuous	Asymptotic theory	R	2010
[76]	1	Any	Asymptotic theory	R	2011
[77]	$\geqslant 1$	Continuous	Asymptotic theory	R	2012
[78]	$\geqslant 1$	Continuous	Asymptotic theory	R	2014
[70]	1	Continuous	Permutation	R	2016
[79]	1	Continuous	Asymptotic theory	R	2016
[80]	$\geqslant 1$	Continuous	Asymptotic theory	R	2017
Meta-analysis					
[81]	$\geqslant 1$	Any	Permutation	R	2013
[82]	$\geqslant 1$	Any	Asymptotic theory	Java	2014
[83]	$\geqslant 1$	Continuous	Asymptotic theory	R	2016
[84]	$\geqslant 1$	Any	Asymptotic theory	Matlab	2016
[85]	1	Any	Permutation	R	2016
[86]	$\geqslant 1$	Continuous	Asymptotic theory	Python	2016
[87]	$\geqslant 1$	Any	Permutation	R	2017
[88]	$\geqslant 1$	Any	Asymptotic theory	R	2017
[89]	$\geqslant 1$	Any	Asymptotic theory/ permutation	R	2017

Estimation Equation (EGEE) based approach to jointly test bivariate gene/phenotype association, which represents one continuous and one binary traits. Following the generalized linear model theory, the relationship between explanatory variables and two phenotypes is modeled as follows:

$$\begin{pmatrix} \mu_{i1} \\ \ln\left(\frac{\mu_{i2}}{1-\mu_{i2}}\right) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_i & 0 \\ 0 & \boldsymbol{x}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \tag{12}$$

where x_i is a column vector of explanatory variables, including genetic variants as well as other fixed effects (e.g., age, sex) for a subject i, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are regression parameter vectors of the same size for two traits, respectively, and μ_{i1} and μ_{i2} are the marginal means of the two phenotypes for subject i. Two steps, an estimation step and a testing step, are taken to test whether an SNP is associated with the continuous or binary trait. In the estimation step, the regression vector $\boldsymbol{\beta}$ is estimated in a set of estimation equations, as reported in the original study[90]. In the testing step, a Wald χ^2 statistic is used to determine if the considered SNP affects either of the two traits. The test statistic for the m-th SNP employed in the Wald test is as follows:

$$W = \left(\widehat{\beta}_{m1}, \widehat{\beta}_{m2}\right) \mathrm{var}\left(\widehat{\beta}_{m1}, \widehat{\beta}_{m2}\right)^{-1} \left(\widehat{\beta}_{m1}, \widehat{\beta}_{m2}\right)^{-1} \tag{13}$$

where $\widehat{\beta}_m$ is the regression coefficient corresponding to the m-th SNP and W follows χ^2 with two degrees of freedom. Analyses of empirical GWA data has confirmed the enhanced power of this bivariate analytical method.

Zhan et al.[73] proposed a Dual Kernel-based Association Test (DKAT) designed to measure the association between high-dimensional phenotypes with multiple genetic variants. DKAT is based on a Kernel Machine Regression (KMR) framework, which is a useful tool for assessing the gene/phenotype association of both common and rare variants[91–93]. In the DKAT concept, individual kernels are used for both the genetic variants and high-dimensional, structured traits. The DKAT statistic is defined as follows:

$$D = \frac{\mathrm{tr}\,(\boldsymbol{H}\boldsymbol{K}_G\boldsymbol{H}\boldsymbol{K}_Y)}{\sqrt{\mathrm{tr}\,(\boldsymbol{H}\boldsymbol{K}_G\boldsymbol{H}\boldsymbol{K}_G)\,\mathrm{tr}\,(\boldsymbol{H}\boldsymbol{K}_Y\boldsymbol{H}\boldsymbol{K}_Y)}} \tag{14}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, $\boldsymbol{H} = \boldsymbol{I}_n - \mathbf{1}\mathbf{1}'/n$ is a centering matrix, $\boldsymbol{I}_n$ is the n-th order identity matrix, $\mathbf{1}$ is an n-dimensional vector of ones, and $\boldsymbol{K}_G$ and $\boldsymbol{K}_Y$ are $n \times n$ matrices whose (i, j)-th elements are $k_g(\boldsymbol{G}_i, \boldsymbol{G}_j)$ and $k_y(Y_i, Y_j)$, respectively, in which $\boldsymbol{G}_i$ denotes the vector of genotypes, Y_i is the set of traits, and k_g and k_y are kernel functions. Two approaches have been proposed to accommodate multiple-candidate kernels. The first average-type strategy calculates an omnibus K^0 , which is usually a linear combination of all possible kernels, then applies the DKAT test. The second minimum-type approach selects the most significant kernel pair. To calculate the P-value, the DKAT test uses moment matching to approximate the empirical distribution of all $n!$ potential permuted DKAT statistics. In this way, to determine the P-value, we must only calculate the first three sample moments of these $n!$ permutations, which have closed-form expressions. A Pearson type III distribution is employed to approximate the permutation null distribution of the DKAT statistic by matching the first three moments. Compared with existing kernel association tests, such as the Multi-trait Sequence Kernel Association Test (MSKAT)[67] and the Gene Association with Multiple Traits (GAMuT)[94], DKAT improves statistical power by incorporating the inherently complex structure of the phenotypes using a phenotype/trait kernel. However, the addition of more noise traits not associated with the SNPs may result in loss of power. Thus, it is important to incorporate variable selection in DKAT to prioritize individual genetic variants/traits.

4.2 Dimension reduction

Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) are the conventional dimension-reduction methods used in the detection of SNP association with multiple correlated traits. Here, we discuss several PCA and CCA based methods to illustrate key CPAS concepts.

Tang and Ferreira[77] proposed a CCA-based algorithm to identify the correlation between a set of genotype data and a single trait or a set of phenotype data[95]. CCA is a method for identifying and measuring the associations between two multivariate sets of variables. Statistically, CCA results are inferred from cross-covariance matrices. Let $X = (X_1, \ldots, X_p)$ and $Y = (Y_1, \ldots, Y_q)$ denote p SNPs and q phenotypes. The i-th canonical correlations ρ_i is calculated as the square root of the i-th eigenvalue of the canonical correlation matrix $\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1/2}$, where Σ_X is the covariance matrix of X, Σ_Y is the covariance matrix of Y, and Σ_{XY} is the covariance matrix between X and Y. In Tang and Ferreira's method[77], Rao's F-approximation is used to test the significance of the canonical correlations and the test score is calculated.

$$F_{(df_1, df_2)} = \left(\frac{1-\lambda^{1/s}}{\lambda^{1/s}}\right)\left(\frac{df_2}{df_1}\right) \tag{15}$$

where

$$\begin{aligned} &\lambda = \Pi_{i=1}^{j}\left(1-\rho_i^2\right),\\ &s = \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}},\\ &df_1 = p\cdot q,\\ &df_2 = \left(n-1.5-\frac{p+q}{2}\right)s - \frac{pq}{2}+1 \end{aligned} \tag{16}$$

Note that this CCA-based method is a gene-based test in which the SNPs are from one gene, according to the annotation in the UCSC Genome Browser. Simulation results suggest that this method provides a robust test for the analysis of multiple quantitative traits without the need for permutation testing. Despite its fast computation speed, a limitation of this method is that it is not flexible, and therefore is unable to accommodate covariates.

Seoane et al.[78] proposed a modified CCA approach that uses an attribute selection strategy based on a GA to maximize the association between different phenotypes and genetic variants within a gene, pathway, or biologically relevant group. To identify sets of SNPs and traits that have a high correlation, the

GA-based optimization method is used, in which the association value is used as the fitness function. This optimization step has been formulated as an integer programming problem that can be solved using a binary GA to find an approximately satisfactory solution in a computationally tractable time. The binary encoding sets the feature to 1 if it is included in the analysis and to 0 otherwise. The authors used three population sizes of 100, 600, and 1000 for single gene/multiple phenotypes, multiple gene/single phenotype, and multiple gene/multiple phenotypes, respectively. For the multiple phenotypes, they found the mutation rate to be $1/82$, and for multiple genes, this rate was $1/3248$. This method was applied to the British Women's Heart and Health Study. A number of novel pleiotropic associations were identified between genetic variants and phenotypes, and previously reported genetic associations were also confirmed with improved statistical detection power.

Klei et al.[74] developed a method based on the Principal Component of Heritability (PCH) to reduce different phenotypes to a single trait with a higher heritability than any other linear combination of phenotypes. The authors defined the heritability attributable to an SNP as follows:

$$h_w^2 = \frac{\boldsymbol{w}' V_Q \boldsymbol{w}}{\boldsymbol{w}' V_P \boldsymbol{w}} \tag{17}$$

where V_Q is the genetic variance, V_P is the residual covariance, and $\boldsymbol{w}$ is used in a linear combination of the phenotypes $\boldsymbol{w}'\boldsymbol{y} = w_1 y_1 + \cdots + w_m y_m$. For any choice of vector $\boldsymbol{w}$, the linear association between $\boldsymbol{w}'\boldsymbol{y}$ and an SNP x can be modeled by $\boldsymbol{w}'\boldsymbol{y} = \mu + \beta x + \varepsilon$. The null hypothesis was $\beta = 0$. The test statistic is $T = b/\mathrm{se}(b)$, where b is the least-squares regression coefficient and $\mathrm{se}(b)$ is the standard error. T follows Normal $(d, 1)$ with a noncentrality parameter:

$$\delta \equiv \left(\frac{N h_w^2}{1 - h_{w^2}} \right)^{1/2} \tag{18}$$

where N is the number of subjects. $\boldsymbol{w}$ that maximizes heritability can be easily obtained by analyzing the eigenstructures of V_Q and V_R. In short, V_Q and V_R can be estimated empirically by sequentially fitting the linear model to all the phenotypes. Using Cholesky decomposition, V_R is decomposed as $V_R = LL'$. Setting $\boldsymbol{w} = (L^{-1})^t \boldsymbol{v}$ can maximize the heritability as defined above, where $\boldsymbol{v}$ is the eigenvalue of V_Q (Only one eigenvalue is non-zero because of the special structure of V_Q). The power of this PCH-based method is enhanced by taking a linear combination of phenotypes that reduces the overall variance and number of tests performed.

Ried et al.[79] developed a PCA-based approach to simultaneously capture the variation across multiple traits in a uniform manner across multiple studies. The authors performed PCAs on six anthropometric traits, including BMI, height, hip, waist, weight, and waist-to-hip ratio of 20 independent studies with non-overlapping, unrelated participants. In each study, they applied PCA to the standardized residuals of the traits adjusted for age and gender. The PCA results for each study were a set of six Principal Components (PCs) comprising orthogonal linear combinations of the six traits. Then, a combined average correlation matrix was derived. By using single-study correlation matrices, this correlation matrix was a weighted sum divided by the number of individuals. The associated PCs, termed AvPCs, represent combinations of different anthropometric traits. AvPCs can capture more complex body-shape phenotypes than can any single traits. Using the AvPCs, the authors performed regression to identify the associated genetic markers. Their experimental results showed that the AvPC was a robust CP representation that could be used in large-scale meta-analyses. However, PCA-derived results are often challenging to interpret and lose power when the weights for collapsing the multiple traits are inconsistent with the phenotype structure.

4.3 Meta-analysis using summary statistics

The summary test statistics of many GWAS analyses are more readily available than subject-level phenotype and genotype data. Therefore, there is great interest in jointly analyzing multiple phenotypes using only GWAS phenotype analysis summary statistics.

Ray and Boehnke[88] developed a unified association test, known as metaUSAT, what detects a single genetic variant associated with multiple traits using only summary statistics from existing GWAS studies. metaUSAT can test the genetic associations of categorical and/or continuous traits without subject-level data. It can also analyze a single trait across multiple studies with overlapping samples. The proposed metaUSAT statistic is as follows:

$$T_\omega = \omega T_{\text{metaMANOVA}} + (1 - \omega) T_{SSU} \tag{19}$$

where

$$\begin{gathered} \omega \in [0, 1], \\ T_{\text{metaMANOVA}} = \boldsymbol{Z}' \widehat{\boldsymbol{R}}^{-1} \boldsymbol{Z}, \\ T_{SSU} = \sum_{k=1}^{K} Z_k^{\gamma} \end{gathered} \tag{20}$$

$\boldsymbol{Z} = (Z_1, \cdots, Z_k)$ are summary statistics, and $\widehat{\boldsymbol{R}}$ is the $K \times K$ estimated correlation matrix of the original traits. The P-value p_w of T_w can be calculated by many different algorithms because T_w is approximately distributed as a linear combination of chi-squared variables under H_0. metaUSAT is defined as the weighted combination with the most significant P-value, as follows:

$$T_{metaUSAT} = \min_{\omega \in [0,1]} p_\omega \tag{21}$$

It demonstrated that the correlation matrix of univariate summary statistics is the same as the trait correlation matrix of the null hypothesis of no association[96]. If two cohorts have overlapping study subjects, the correlation matrix of the univariate summary statistics is calculated as follows:

$$\operatorname{corr}\left(Z_i, Z_j\right) = \frac{\sum_l \left(Z_{il} - \mu_i\right)\left(Z_{jl} - \mu_j\right)}{\sqrt{\sum_l \left(Z_{il} - \mu_i\right)^2 \left(Z_{jl} - \mu_j\right)^2}} \tag{22}$$

where μ_i and μ_j are test statistic means for many independent SNPs across the genome. If two cohorts have no overlapping subjects, Z_i and Z_j are independent. The results of simulation experiments demonstrated that metaUSAT had a low Type-I error and has comparable and sometimes higher power in detecting associations than existing methods, such as $minP$[97], S_{Hom}[96], and SPU[98].

Liu and Lin[89] proposed the use of univariate summary test statistics for the detection of homogeneous and heterogeneous genetic effects on multiple phenotypes by considering the correlation between these summary statistics. The authors theoretically justified that under the null hypothesis, the correlation matrix of the univariate summary test statistics does not depend on the genotype, and is the same as the correlation matrix of the original multiple phenotypes, conditional on the covariates. Thus, they reported that the correlation matrix of the univariate summary test statistics can be estimated using a large number of independent null SNPs over the whole genome. To detect both homogeneous and heterogeneous effects, they defined the following linear mixed model for the summary statistics:

$$\boldsymbol{Z} = \mu_0 \boldsymbol{J} + \boldsymbol{b} + \epsilon, \epsilon \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{23}$$

where $\boldsymbol{Z} = (Z_i, \cdots, Z_k)'$ is a vector of the univariate Wald-type statistics of k phenotypes, $\boldsymbol{J} = (1, 1, ..., 1)'$, μ_0 is a scalar denoting the shared common effect size, and $b_k = \mu_k - \mu_0$ denotes a departure from the common effect of a genetic variant on the k-th phenotype. b_k is assumed to follow an arbitrary distribution F with mean 0 and variance τ and be mutually independent. Using the above model, the authors found the joint testing of H_0: $\mu = 0, \tau = 0$ to be equivalent to testing for the associations between a genetic variant and k phenotypes. Under H_0, the scores of μ and τ are calculated as follows:

$$\begin{aligned} \boldsymbol{U}_{\mu_0} &= \boldsymbol{J}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}, \\ \boldsymbol{U}_{\tau_0} &= \left(\boldsymbol{Z} - \widehat{\mu}_0 \boldsymbol{J}\right)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{Z} - \widehat{\mu}_0 \boldsymbol{J}\right) \end{aligned} \tag{24}$$

where $\widehat{\mu}_0$ is the MLE of μ_0 under τ_0. The authors used three approaches, including the inverse-variance weighting scheme, inverse standard deviation weighting schemes, and adaptive procedure, to find an optimal 0 to maximize power in the following linear combination:

$$T_\phi = \phi U_{\mu_0}^2 + (1 - \phi) U_{\tau_0} \tag{25}$$

In addition to combining two testing statistics, Fisher's and Tippett's procedures were also used to combine the two corresponding independent P-values of U_{μ_0} and U_{τ_0}. Simulation studies showed that these mix-type tests are robust and can identify more significant SNPs than the Wald test.

4.4 Identifying pleiotropy

It is well known that causal conclusions cannot strictly be drawn from mere statistical associations between distinct phenotypes, such as low serum cholesterol levels and cancer, unless all possible confounders of the association are identified, measured, and adjusted. Solovieff et al.[27] categorized pleiotropy with regard to complex traits into three broad groups: biological pleiotropy in which causal variants of different traits to fall into the same gene or regulatory unit, mediated pleiotropy in which a variant directly affects one trait which affects another trait, and spurious pleiotropy in which different causal variants may be tagged by the same variant. Thus, CP associations may arise due to one or more pleiotropy cases. Many methods have been proposed to distinguish biological from mediated and spurious pleiotropy. We discuss two of these methods.

Didelez and Sheehan[99] developed a framework for causal inference based on Mendelian randomization, which defines an Instrumental Variable (IV) and tests for, or estimates, the causal effect of a phenotype on another phenotype[100]. Let X be the cause, such as cholesterol level, Y be the response, such as coronary heart disease, G be the instrument (the genetic marker in GWAS), and U be an unobservable variable that represents the confounding aspect between X and Y. For a valid IV, G must satisfy three assumptions: first,

that G is associated with X; second, that G is not associated with U; and third, that G is not associated with Y when conditional on X. In the simplest case where the dependencies among variables Y, X, G, and U are linear, the following models can be defined:

$$E(Y|X=x,U=u)=\alpha+\beta_1 x+\beta_2 u,$$
$$E(X|G=g,U=u)=\gamma+\delta_1 g+\delta_2 u \quad (26)$$

The Average Causal Effect (ACE) is defined as the difference in expectations for different settings of X. In this framework, β_1 is the causal parameter of interest since $\mathrm{ACE}(x_1,x_2)=\beta_1(x_1-x_2)$. β_1 can be consistently estimated as the following ratio,

$$\widehat{\beta}_1=\frac{r_{Y|G}}{r_{X|G}} \quad (27)$$

where $r_{X|G}$ is the consistent estimate of θ_1 and $r_{Y|G}$ is the regression coefficient of G in a linear least squares regression of Y on G. However, typically, there is an uncertainty regarding the third assumption of the IV, which will not hold if G affects both X and Y (i.e., if the variants are truly pleiotropic). Many approaches have been proposed to overcome this limitation in Mendelian randomization settings[101–103]. For example, MR-Egger, proposed by Bowden et al., is an adaption of the Egger regression, which can detect some violations of the standard IV assumptions and provides an effect estimation that is not subject to third violations[103].

Han et al.[96] developed a statistical test, known as BUHMBOX, to distinguish between whole-group pleiotropy and subgroup heterogeneity in CP associations. Whole-group pleiotropy means that the sharing of risk alleles across different traits are driven by all the subjects, whereas subgroup heterogeneity is when these risk alleles are driven by a subset of subjects. Subgroup heterogeneity can occur for many reasons, such as in a case group that includes subjects with atypical clinical symptoms for a different disease. The BUHMBOX concept is that when there is subgroup heterogeneity, known risk alleles for the first-disease-associated loci will be enriched in a subgroup of second-disease cases, which produces positive correlations between these known alleles from independent loci. In contrast, known risk alleles for the first-disease-associated loci will be uniformly distributed in whole-group pleiotropy. So the proposed BUHMBOX test statistic compares the risk allele frequency of variants associated with the first disease in second-disease cases. BUHMBOX tests whether the known risk alleles are enriched in a subgroup of cases or whether they are evenly distributed across all cases. Subgroup heterogeneity can lead to a significant BUHMBOX test statistic. In contrast, a lack of actual subgroup heterogeneity or high type II error can lead to a non-significant BUHMBOX test statistic. The results of systematic experiments showed that BUHMBOX achieved 81.7% power and a 4.3% false positive rate in detecting heterogeneity at $P<0.05$. When using BUHMBOX, users must specify the two traits of interest, one of which must have a list of SNPs associated with known risk alleles.

5 Summary and Outlook

To date, single causal SNP detection has achieved successful results. Although the exhaustive search method is slow, its computation cost is affordable. For two-locus SNP detection, the exhaustive search method may not finish in an acceptable amount of time. Also, many complex diseases can be caused by multiple SNPs. For complex disorders, k-locus SNP detection is mainly used for data analysis. As yet, no one method performs consistently better than others in all scenarios. For example, heuristic search methods, such as GA, PSO, and ACO, can obtain better solutions for high-order SNP combinations, but are not suitable for large SNP data due to the excessuve computation cost. These methods may also miss many significant associations by searching only a limited feasible solution space. Stepwise search methods are relatively more powerful than heuristic search methods. For example, the SNPRuler has detection power for SNPs with no marginal effects and requires a reasonable time on large-scale datasets. However, it is weak in the detection of SNPs with marginal effects and requires a huge amount of memory. The main challenge in stepwise search methods is to construct a correct and complete lemma to replace the downward closure lemma in the apriori algorithm. As yet, there is no lemma that retains completeness for all kinds of diseases. Feature selection methods are faster than other methods because the number of selected features (candidate SNPs) is much smaller than that in other methods. However, the main problem in feature selection methods is that they do not analyze how much useful information remains in the selected SNPs, so it is unclear whether important SNPs are filtered out in the feature selection progress. In k-locus SNP detection, stepwise search and feature selection methods are the best choice for genome-wide association studies.

As data becomes more available, Cross-Phenotype Association Study (CPAS) is an emerging research area in human genetics. This review extended the focus from single-trait GWAS to the detection of CP effects and pleiotropy. In contrast to single-trait GWAS, where the analysis approaches are now somewhat standardized, multi-trait methods vary considerably in their applied statistical procedures. Here, we broadly introduced and discussed commonly used regression models, dimension-reduction methods, and summary statistics in CPAS. A fundamental challenge in CPAS is to assure comparability and replicability of the results. As the field moves toward large-scale sequencing-based association studies based on widely available electronic medical records, a critical step will be to develop fast and efficient algorithms that have excellent scalability. These practical and powerful approaches will increase our understanding of the shared genetics among traits and reveal that phenotypes are a set of related indicators of biological mechanisms rather than isolated manifestations. As our knowledge of the molecular links between diseases increases, these insights will facilitate the establishment of the foundation for drug design and personalized medicine development.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61662028).

References

[1] B. S. Shastry, SNP alleles in human disease and evolution, *J. Hum. Genet.*, vol. 47, no. 11, pp. 561–566, 2002.

[2] Z. P. Cai, H. Sabaa, Y. N. Wang, R. Goebel, Z. Q. Wang, J. F. Xu, P. Stothard, and G. H. Lin, Most parsimonious haplotype allele sharing determination, *BMC Bioinformatics*, vol. 10, p. 115, 2009.

[3] N. J. Prescott, S. A. Fisher, A. Franke, J. Hampe, C. M. Onnie, D. Soars, R. Bagnall, M. M. Mirza, J. Sanderson, A. Forbes, et al., A nonsynonymous SNP in *ATG16L1* predisposes to Ileal Crohn's disease and is independent of *CARD15* and *IBD5*, *Gastroenterology*, vol. 132, no. 5, pp. 1665–1671, 2007.

[4] S. Seki, Y. Kawaguchi, K. Chiba, Y. Mikami, H. Kizawa, T. Oya, F. Mio, M. Mori, Y. Miyamoto, I. Masuda, et al., A functional SNP in CILP, encoding cartilage intermediate layer protein, is associated with susceptibility to lumbar disc disease, *Nat. Genet.*, vol. 37, no. 6, pp. 607–612, 2005.

[5] H. Zaimkohan, M. Keramatipour, S. M. H. Ghaderian, J. Tavakkoly-Bazzaz, A. Tahooni, M. Piryaei, N. M. Ghahhari, M. M. Golchin, and M. Ahani, PCSK9 SNP RS11591147 association study with coronary artery disease risk in Iran, *Acta Med. Mediterr.*, vol. 31, p. 1435, 2015.

[6] X. Guo, N. Yu, F. Gu, X. J. Ding, J. X. Wang, and Y. Pan, Genome-wide interaction-based association of human diseases—A survey, *Tsinghua Sci. Technol.*, vol. 19, no. 6, pp. 596–616, 2014.

[7] R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, et al., Complement factor H polymorphism in age-related macular degeneration, *Science*, vol. 308, no. 5720, pp. 385–389, 2005.

[8] J. W. Shen, Z. Q. Li, Z. J. Song, J. H. Chen, and Y. Y. Shi, Genome-wide two-locus interaction analysis identifies multiple epistatic SNP pairs that confer risk of prostate cancer: A cross-population study, *Int. J. Cancer*, vol. 140, no. 9, pp. 2075–2084, 2017.

[9] M. J. Simmonds and S. C. L. Gough, The HLA region and autoimmune disease: Associations and mechanisms of action, *Curr. Genomics*, vol. 8, no. 7, pp. 453–465, 2007.

[10] H. Ueda, J. M. M. Howson, L. Esposito, J. Heward, H. Snook, G. Chamberlain, D. B. Rainbow, K. M. D. Hunter, A. N. Smith, G. Di Genova, et al., Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease, *Nature*, vol. 423, no. 6939, pp. 506–511, 2003.

[11] L. A. Criswell, K. A. Pfeiffer, R. F. Lum, B. Gonzales, J. Novitzke, M. Kern, K. L. Moser, A. B. Begovich, V. E. H. Carlton, W. T. Li, et al., Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: The *PTPN22* 620W allele associates with multiple autoimmune phenotypes, *Am. J. Hum. Genet.*, vol. 76, no. 4, pp. 561–571, 2005.

[12] A. Zhernakova, C. C. Van Diemen, and C. Wijmenga, Detecting shared pathogenesis from the shared genetics of immune-related diseases, *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 43–55, 2009.

[13] R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. W. De Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, et al., Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels, *Science*, vol. 316, no. 5829, pp. 1331–1336, 2007.

[14] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, D. R. Cox, D. A. Hinds, L. A. Pennacchio, A. Tybjaerg-Hansen, A. R. Folsom, et al., A common allele on chromosome 9 associated with coronary heart disease, *Science*, vol. 316, no. 5830, pp. 1488–1491, 2007.

[15] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Baker, A. Palsson, et al., A common variant on chromosome 9p21 affects the risk of myocardial infarction, *Science*, vol. 316, no. 5830, pp. 1491–1493, 2007.

[16] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. J. Dixon, T. Meitinger, P. Braund, H. E. Wichmann, et al., Genomewide association analysis of coronary artery disease, *N. Engl. J. Med.*, vol. 357, no. 5, pp. 443–453, 2007.

[17] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, Potential etiologic and functional implications of genome-

wide association loci for human diseases and traits, *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 23, pp. 9362–9367, 2009.

[18] X. Guo, J. Zhang, Z. P. Cai, D. Z. Du, and Y. Pan, Searching genome-wide multi-locus associations for multiple diseases based on Bayesian inference, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 14, no. 3, pp. 600–610, 2017.

[19] X. Guo, J. Zhang, Z. P. Cai, D. Z. Du, and Y. Pan, Dam: A Bayesian method for detecting genome-wide associations on multiple diseases, in *Proc. 11^{th} Int. Symp. Bioinformatics Research and Applications*, Norfolk, VA, USA, 2015, pp. 96–107.

[20] B. D. Hobbs, K. De Jong, M. Lamontagne, Y. Bossé, N. Shrine, M. S. Artigas, L. V. Wain, I. P. Hall, V. E. Jackson, A. B. Wyss, et al., Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis, *Nat. Genet.*, vol. 49, no. 3, pp. 426–432, 2017

[21] R. M. Plenge, L. Padyukov, E. F. Remmers, S. Purcell, A. T. Lee, E. W. Karlson, F. Wolfe, D. L. Kastner, L. Alfredsson, D. Altshuler, et al., Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: Association of susceptibility with *PTPN22*, *CTLA4*, and *PADI4*, *Am. J. Hum. Genet.*, vol. 77, no. 6, pp. 1044–1060, 2005.

[22] C. Kyogoku, W. A. Ortmann, A. Lee, S. Selby, V. E. H. Carlton, M. Chang, P. Ramos, E. C. Baechler, F. M. Batliwalla, J. Novitzke, et al., Genetic association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE, *Am. J. Hum. Genet.*, vol. 75, no. 3, pp. 504–507, 2004.

[23] J. A. Todd, N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, R. Bailey, S. Nejentsev, S. F. Field, F. Payne, et al., Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes, *Nat. Genet.*, vol. 39, no. 7, pp. 857–864, 2007.

[24] W. S. Bush, M. T. Oetjens, and D. C. Crawford, Unravelling the human genome-phenome relationship using phenome-wide association studies, *Nat. Rev. Genet.*, vol. 17, no. 3, pp. 129–145, 2016.

[25] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, et al., The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog), *Nucleic Acids Res.*, vol. 45, no. D1, pp. D896–D901, 2017.

[26] S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell, Abundant pleiotropy in human complex diseases and traits, *Am. J. Hum. Genet.*, vol. 89, no. 5, pp. 607–618, 2011.

[27] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, Pleiotropy in complex traits: Challenges and strategies, *Nat. Rev. Genet.*, vol. 14, no. 7, pp. 483–495, 2013.

[28] Y. Zhang and J. S. Liu, Bayesian inference of epistatic interactions in case-control studies, *Nat. Genet.*, vol. 39, no. 9, pp. 1167–1173, 2007.

[29] W. Li and J. Reich, A complete enumeration and classification of two-locus disease models, *Hum. Hered.*, vol. 50, no. 6, pp. 334–349, 2000.

[30] D. R. Velez, B. C. White, A. A. Motsinger, W. S. Bush, M. D. Ritchie, S. M. Williams, and J. H. Moore, A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, *Genet. Epidemiol.*, vol. 31, no. 4, pp. 306–315, 2007.

[31] X. Zhang, S. P. Huang, F. Zou, and W. Wang, Team: Efficient two-locus epistasis tests in human genome-wide association study, *Bioinformatics*, vol. 26, no. 12, pp. i217–i227, 2010.

[32] Y. Wang, G. M. Liu, M. L. Feng, and L. Wong, An empirical comparison of several recent epistatic interaction detection methods, *Bioinformatics*, vol. 27, no. 21, pp. 2936–2943, 2011.

[33] X. Wan, C. Yang, Q. Yang, H. Xue, X. D. Fan, N. L. S. Tang, and W. C. Yu, Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies, *Am. J. Hum. Genet.*, vol. 87, no. 3, pp. 325–340, 2010.

[34] H. Matsuda, Physical nature of higher-order mutual information: Intrinsic correlations and frustration, *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, vol. 62, no. 3, pp. 3096–3102, 2000.

[35] L. S. Yung, C. Yang, X. Wan, and W. C. Yu, GBOOST: A GPU-based tool for detecting gene-gene interactions in genome-wide case control studies, *Bioinformatics*, vol. 27, no. 9, pp. 1309–1310, 2011.

[36] J. H. Holland, *Adaptation in Natural and Artificial Systems*: *An Introductory Analysis with Applications to Biology*, *Control*, *and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

[37] L. Y. Chuang, M. C. Lin, H. W. Chang, and C. H. Yang, Odds ratio-based genetic algorithm for prediction of snp-snp interactions in breast cancer association study, presented at the 26th Int. Conf. Advanced Information Networking and Applications Workshops (WAINA), Fukuoka, Japan, 2012, pp. 920–925.

[38] J. B. Chen, L. Y. Chuang, Y. D. Lin, C. W. Liou, T. K. Lin, W. C. Lee, B. C. Cheng, H. W. Chang, and C. H. Yang, Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility, *Mitochondrial DNA*, vol. 25, no. 3, pp. 231–237, 2014.

[39] C. H. Yang, Y. D. Lin, L. Y. Chuang, and H. W. Chang, Evaluation of breast cancer susceptibility using improved genetic algorithms to generate genotype SNP barcodes, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 10, no. 2, pp. 361–371, 2013.

[40] M. Dorigo and L. M. Gambardella, Ant colonies for the travelling salesman problem, *Biosystems*, vol. 43, no. 2, pp. 73–81, 1997.

[41] Y. P. Wang, X. Y. Liu, K. Robbins, and R. Rekaya, AntEpiSeeker: Detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm, *BMC Res. Notes*, vol. 3, p. 117, 2010.

[42] L. Y. Chuang, M. C. Lin, H. W. Chang, and C. H. Yang, Analysis of SNP interaction combinations to determine breast cancer risk with PSO, presented at the 11th

Int. Conf. Bioinformatics and Bioengineering (BIBE), Taichung, China, 2011, pp. 291–294.

[43] S. J. Wu, L. Y. Chuang, Y. D. Lin, W. H. Ho, F. T. Chiang, C. H. Yang, and H. W. Chang, Particle swarm optimization algorithm for analyzing SNP-SNP interaction of renin-angiotensin system genes against hypertension, *Mol. Biol. Rep.*, vol. 40, no. 7, pp. 4227–4233, 2013.

[44] D. H. Kim, S. Uhmn, and J. Kim, Finding relevant SNP sets and predicting disease risk using simulated annealing, *Int. J. Softw. Eng. Appl.*, vol. 6, no. 3, pp. 81–88, 2012.

[45] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, in *Proc. 20th VLDB Conf.*, Santiago, Chile, 1994, pp. 487–499.

[46] X. Wan, C. Yang, Q. Yang, H. Xue, N. L. S. Tang, and W. C. Yu, Predictive rule inference for epistatic interaction detection in genome-wide association studies, *Bioinformatics*, vol. 26, no. 1, pp. 30–37, 2010.

[47] Y. Wang, G. M. Liu, M. L. Feng, and L. Wong, Response: An empirical comparison of several recent epistatic interaction detection methods, *Bioinformatics*, vol. 28, no. 1, pp. 147–148, 2012.

[48] M. Z. Xie, J. Li, and T. Jiang, Detecting genome-wide epistases based on the clustering of relatively frequent items, *Bioinformatics*, vol. 28, no. 1, pp. 5–12, 2012.

[49] J. Liu, G. X. Yu, Y. Jiang, and J. Wang, Hiseeker: Detecting high-order SNP interactions based on pairwise SNP combinations, *Genes*, vol. 8, no. 6, p. 153, 2017.

[50] W. D. Mao and J. Lee, A combinatorial analysis of genetic data for Crohn's disease, presented at the 1st Int. Conf. Bioinformatics and Biomedical Engineering, Wuhan, China, 2007, pp. 1031–1034.

[51] J. W. He and A. Zelikovsky. Multiple linear regression for index SNP selection on unphased genotypes, presented at the 28th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 2006, pp. 5759–5762.

[52] Z. Z. Feng, X. J. Yang, S. Subedi, and P. D. McNicholas, The lasso and sparse least squares regression methods for SNP selection in predicting quantitative traits, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 2, pp. 629–636, 2012.

[53] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, Genome-wide association analysis by lasso penalized logistic regression, *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[54] X. J. Ding, J. X. Wang, A. Zelikovsky, X. Guo, M. Z. Xie, and Y. Pan, Searching high-order SNP combinations for complex diseases based on energy distribution difference, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 3, pp. 695–704, 2015.

[55] S. Leem, H. H. Jeong, J. Lee, K. Wee, and K. A. Sohn, Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure, *Comput. Biol. Chem.*, vol. 50, pp. 19–28, 2014.

[56] J. Hodgkin, Seven types of pleiotropy, *Int. J. Dev. Biol.*, vol. 42, no. 3, pp. 501–505, 1998.

[57] J. F. Liu, Y. F. Pei, C. J. Papasian, and H. W. Deng, Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations, *Genet. Epidemiol.*, vol. 33, no. 3, pp. 217–227, 2009.

[58] Q. Yang, H. S. Wu, C. Y. Guo, and C. S, Fox, Analyze multivariate phenotypes in genetic association studies by combining univariate association tests, *Genet. Epidemiol.*, vol. 34, no. 5, pp. 444–454, 2010.

[59] J. Huang, A. D. Johnson, and C. J. O'donnell, PRIMe: A method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies, *Bioinformatics*, vol. 27, no. 9, pp. 1201–1206, 2011.

[60] M. D. Yuan and G. Q. Diao, Joint association analysis of bivariate quantitative and qualitative traits, *BMC Proc.*, vol. 5, no. S9, p. S74, 2011.

[61] A. Maity, P. F. Sullivan, and J. I. Tzeng, Multivariate phenotype association analysis by marker-set kernel machine regression, *Genet. Epidemiol.*, vol. 36, no. 7, pp. 686–695, 2012.

[62] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M. R. Jarvelin, and L. J. M. Coin, MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS, *PLoS One*, vol. 7, no. 5, p. e34861, 2012.

[63] P. Marttinen, M. Pirinen, A. P. Sarin, J. Gillberg, J. Kettunen, I. Surakka, A. J. Kangas, P. Soininen, P. O'Reilly, M. Kaakinen, et al., Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression, *Bioinformatics*, vol. 30, no. 14, pp. 2026–2034, 2014.

[64] Y. F. Wang, A. Y. Liu, J. L. Mills, M. Boehnke, A. F. Wilson, J. E. Bailey-Wilson, M. M. Xiong, C. O. Wu, and R. Z. Fan, Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models, *Genet. Epidemiol.*, vol. 39, no. 4, pp. 259–275, 2015.

[65] D. Ray, J. S. Pankow, and S. Basu, USAT: A unified score-based association test for multiple phenotype-genotype analysis, *Genet. Epidemiol.*, vol. 40, no. 1, pp. 20–34, 2016.

[66] F. P. Casale, B. Rakitsch, C. Lippert, and O. Stegle, Efficient set tests for the genetic analysis of correlated traits, *Nat. Methods*, vol. 12, no. 8, pp. 755–758, 2015.

[67] B. L. Wu and J. S. Pankow, Sequence kernel association test of multiple continuous phenotypes, *Genet. Epidemiol.*, vol. 40, no. 2, pp. 91–100, 2016.

[68] D. D. Lin, J. Y. Li, V. D. Calhoun, and Y. P. Wang, Detection of genetic factors associated with multiple correlated imaging phenotypes by a sparse regression model, presented at the 12th Int. Symp. Biomedical Imaging (ISBI), New York, NY, USA, 2015, pp. 1368–1371.

[69] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P. R. Loh, L. Duncan, J. R. B. Perry, N. Patterson, E. B. Robinson, et al., An atlas of genetic correlations across human diseases and traits, *Nat. Genet.*, vol. 47, no. 11, pp. 1236–1241, 2015.

[70] Z. C. Wang, Q. Y. Sha, and S. L. Zhang, Joint analysis of multiple traits using "optimal" maximum heritability test, *PLoS One*, vol. 11, no. 3, p. e0150975, 2016.

[71] J. P. Sun, K. Oualkacha, V. Forgetta, H. F. Zheng, J. B. Richards, A. Ciampi, C. M. T. Greenwood, and U.

Consortium, A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects, *Eur. J. Hum. Genet.*, vol. 24, no. 9, pp. 1344–1351, 2016.

[72] S. Lee, S. Won, Y. J. Kim, Y. Kim, B. J. Kim, and T. Park, Rare variant association test with multiple phenotypes, *Genet. Epidemiol.*, vol. 41, no. 3, pp. 198–209, 2017.

[73] X. Zhan, N. Zhao, A. Plantinga, T. A. Thornton, K. N. Conneely, M. P. Epstein, and M. C. Wu, Powerful genetic association analysis for common or rare variants with high-dimensional structured traits, *Genetics*, vol. 206, no. 4, pp. 1779–1790, 2017.

[74] L. Klei, D. Luca, B. Devlin, and K. Roeder, Pleiotropy and principal components of heritability combine to increase power for association analysis, *Genet. Epidemiol.*, vol. 32, no. 1, pp. 9–19, 2008.

[75] H. Mei, W. Chen, A. Dellinger, J. He, M. Wang, C. Yau, S. R. Srinivasan, and G. S. Berenson, Principal-component-based multivariate regression for genetic association studies of metabolic syndrome components, *BMC Genet.*, vol. 11, p. 100, 2010.

[76] I. Mukhopadhyay, S. Saha, and S. Ghosh, Integrating binary traits with quantitative phenotypes for association mapping of multivariate phenotypes, *BMC Proc.*, vol. 5 Suppl 9, p. S73, 2011.

[77] C. S. Tang and M. A. R. Ferreira, A gene-based test of association using canonical correlation analysis, *Bioinformatics*, vol. 28, no. 6, pp. 845–850, 2012.

[78] J. A. Seoane, C. Campbell, I. N. M. Day, J. P. Casas, and T. R. Gaunt, Canonical correlation analysis for gene-based pleiotropy discovery, *PLoS Comput. Biol.*, vol. 10, no. 10, p. e1003876, 2014.

[79] J. S. Ried, M. J. Jeff, A. Y. Chu, J. L. Bragg-Gresham, J. Van Dongen, J. E. Huffman, T. S. Ahluwalia, G. Cadby, N. Eklund, J. Eriksson, T. Esko, et al., A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape, *Nat. Commun.*, vol. 7, p. 13357, 2016.

[80] N. Lin, Y. Zhu, R. Z. Fan, and M. M. Xiong, A quadratically regularized functional canonical correlation analysis for identifying the global structure of pleiotropy with NGS data, *PLoS Comput. Biol.*, vol. 13, no. 10, p. e1005788, 2017.

[81] A. Derkach, J. F. Lawless, and L. Sun, Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests, *Genet. Epidemiol.*, vol. 37, no. 1, pp. 110–121, 2013.

[82] S. Van Der Sluis, C. V. Dolan, J. Li, Y. Song, P. C. Sham, D. Posthuma, and M. X. Li, MGAS: A powerful tool for multivariate gene-based genome-wide association analysis, *Bioinformatics*, vol. 31, no. 7, pp. 1007–1015, 2015.

[83] J. Kim, Y. W. Zhang, and W. Pan, Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data, *Genetics*, vol. 203, no. 2, pp. 715–731, 2016.

[84] A. Cichonska, J. Rousu, P. Marttinen, A. J. Kangas, P. Soininen, T. Lehtimäki, O. T. Raitakari, M. R. Järvelin, V. Salomaa, M. Ala-Korpela, et al., metaCCA: Summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis, *Bioinformatics*, vol. 32, no. 13, pp. 1981–1989, 2016.

[85] X. Y. Liang, Z. C. Wang, Q. Y. Sha, and S. L. Zhang, An adaptive fisher's combination method for joint analysis of multiple phenotypes in association studies, *Sci. Rep.*, vol. 6, p. 34323, 2016.

[86] B. C. Brown, C. J. Ye, A. L. Price, and N. Zaitlen, Transethnic genetic-correlation estimates from summary statistics, *Am. J. Hum. Genet.*, vol. 99, no. 1, pp. 76–88, 2016.

[87] I. Y. Kwak and W. Pan, Gene- and pathway-based association tests for multiple traits with GWAS summary statistics, *Bioinformatics*, vol. 33, no. 1, pp. 64–71, 2016.

[88] D. Ray and M. Boehnke, Methods for meta-analysis of multiple traits using GWAS summary statistics, *Genet. Epidemiol.*, vol. 42, no. 2, pp. 134–145, 2018.

[89] Z. H. Liu and X. H. Lin, Multiple phenotype association tests using summary statistics in genome-wide association studies, *Biometrics*, vol. 74, no. 1, pp. 165–175, 2018.

[90] D. B. Hall, On the application of extended quasi-likelihood to the clustered data case, *Can. J. Stat.*, vol. 29, no. 1, pp. 77–97, 2001.

[91] M. C. Wu, S. Lee, T. X. Cai, Y. Li, M. Boehnke, and X. H. Lin, Rare-variant association testing for sequencing data with the sequence kernel association test, *Am. J. Hum. Genet.*, vol. 89, no. 1, pp. 82–93, 2011.

[92] I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. H. Lin, Sequence kernel association tests for the combined effect of rare and common variants, *Am. J. Hum. Genet.*, vol. 92, no. 6, pp. 841–853, 2013.

[93] X. Zhan, S. Giririrajan, N. Zhao, M. C. Wu, and D. Ghosh, A novel copy number variants kernel association test with application to autism spectrum disorders studies, *Bioinformatics*, vol. 32, no. 23, pp. 3603–3610, 2016.

[94] K. A. Broadaway, D. J. Cutler, R. Duncan, J. L. Moore, E. B. Ware, M. A. Jhun, L. F. Bielak, W. Zhao, J. A. Smith, P. A. Peyser, et al., A statistical approach for testing cross-phenotype effects of rare variants, *Am. J. Hum. Genet.*, vol. 98, no. 3, pp. 525–540, 2016.

[95] H. Hotelling, Relations between two sets of variates, *Biometrika*, vol. 28, nos. 3&4, pp. 321–377, 1936.

[96] B. Han, J. G. Pouget, K. Slowikowski, E. Stahl, C. H. Lee, D. Diogo, X. Hu, Y. R. Park, E. Kim, P. K. Gregersen, et al., A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases, *Nature Genetics*, vol. 48, no. 7, pp. 803–810, 2016.

[97] K. N. Conneely and M. Boehnke, So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests, *Am. J. Hum. Genet.*, vol. 81, no. 6, pp. 1158–1168, 2007.

[98] J. Kim, Y. Bai, and W. Pan, An adaptive association test for multiple phenotypes with GWAS summary statistics, *Genet. Epidemiol.*, vol. 39, no. 8, pp. 651–663, 2015.

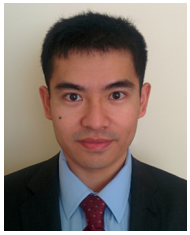
[99] V. Didelez and N. Sheehan, Mendelian randomization as an instrumental variable approach to causal inference, *Statistical Methods in Medical Research*, vol. 16, no. 4, pp. 309–330, 2007.

[100] J. Pearl, *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge, UK: Cambridge University Press, 2009.

[101] M. F. Del Greco, C. Minelli, N. A. Sheehan, and J. R. Thompson, Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome, *Stat. Med.*, vol. 34, no. 21, pp. 2926–2940, 2015.

[102] J. Bowden, M. F. Del Greco, C. Minelli, G. Davey Smith, N. Sheehan, and J. Thompson, A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization, *Stat. Med.*, vol. 36, no. 11, pp. 1783–1802, 2017.

[103] J. Bowden, G. D. Smith, and S. Burgess, Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression, *Int. J. Epidemiol.*, vol. 44, no. 2, pp. 512–525, 2015.



Xuan Guo received the BSc degree in computer science from Southwest University in 2009 and MSc degree in computer science from Wuhan University in 2011. He received the PhD degree in computer science from Georgia State University in 2015. He was an post-doctoral research associate at Oak Ridge National Laboratory from 2015 to 2017. He is currently an assistant professor at the University of North Texas. His research interests include bioinformatics, computational biology, and high performance computing.

Xiaojun Ding received the BS, MS, and PhD degrees in computer science and technology from Central South University, China, in 2001, 2008, and 2015, respectively. He is currently a lecturer in Yulin Normal University. His research interests include bioinformatics, machine learning, and artificial intelligence.