REVIEW ARTICLE

Statistical Methods and Software for Substance Use and Dependence Genetic Research

Tongtong Lan[1], Bo Yang[1], Xuefen Zhang[1,2], Tong Wang[1,*] and Qing Lu[2,*]

[1]*Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, Shanxi, China;* [2]*Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA*

ARTICLE HISTORY

Received: March 06, 2019
Revised: May 16, 2019
Accepted: May 24, 2019

DOI:
10.2174/1389202920666190617094930

Abstract: ***Background***: Substantial substance use disorders and related health conditions emerged during the mid-20th century and continue to represent a remarkable 21st century global burden of disease. This burden is largely driven by the substance-dependence process, which is a complex process and is influenced by both genetic and environmental factors. During the past few decades, a great deal of progress has been made in identifying genetic variants associated with Substance Use and Dependence (SUD) through linkage, candidate gene association, genome-wide association and sequencing studies.

Methods: Various statistical methods and software have been employed in different types of SUD genetic studies, facilitating the identification of new SUD-related variants.

Conclusion: In this article, we review statistical methods and software that are currently available for SUD genetic studies, and discuss their strengths and limitations.

Keywords: Substance dependence, linkage analysis, association analysis, interaction analysis, meta-analysis, GCTA.

1. INTRODUCTION

Substance use disorders present a significant global public health problem that places a substantial burden on individuals, health care systems, societies and economies [1, 2]. It is estimated that approximately 27 million people in the world suffered from substance use disorders in 2013, resulting in a remarkable global burden of disease that is expected to become more prevalent over time [3]. This burden is largely driven by the substance-dependence process, which is a complex disorder influenced by both genetic and environmental factors.

Substance dependence is defined as the syndrome of substance misuse that leads to adverse consequences and includes a cluster of symptoms such as tolerance, withdrawal, and inability to stop using (see DSM-IV substance dependence for the complete diagnostic criteria) [4]. The term substance often refers to products with addictive potential, such as tobacco, alcohol, cocaine and other licit and illicit drugs. The development of the substance-dependence process involves several steps: the initiation of substance use, the transition from experimental use to regular use, and the actual development of dependence [5]. Each step is influenced by both environmental and genetic factors. Twin studies have suggested a substantial genetic contribution to Substance Use and Dependence (SUD) [6]. Over the past few decades, we have successfully identified new genetic variants associated with SUD through linkage, candidate gene association, genome-wide association, animal and sequencing studies. Various statistical methods and software have played an important role in helping the identification of SUD-related variants. This article aims to give an overview of the widely used methods and software for SUD genetic data analysis, including data management, linkage analysis, association analysis and other analysis.

2. DATA MANAGEMENT

The quality-control process is an essential step in data analysis. It is used to identify and remove inaccurate information before the downstream analysis (*e.g.*, genetic association analysis). Researchers utilize different quality control filters and criteria. While some researchers have implemented stringent criteria in order to have high-quality data, others adopt flexible standards to keep as much information as possible.

Commonly used filters are based on the call rate, Minor Allele Frequency (MAF), and Hardy-Weinberg Equilibrium (HWE) test. A fixed call rate is commonly adopted in a study to remove low-quality SNPs and samples. For instance, if a researcher sets a threshold of 1% for sample call rate in a project, then samples with more than a 1% failed genotyping call are removed; likewise, if a researcher sets a threshold of 5% for SNP call rate, individual SNPs with 5% failed genotyping call are also removed [7]. The thresholds of sample call rate and SNP call rate do not have to be the same. Besides the call rate, filters based on MAF and the HWE test are also frequently used. HWE tests are often conducted on the control samples to exclude low-quality SNPs, while an MAF threshold (*e.g.*, MAF>1%) allows the study to focus on

*Address correspondence to these authors at the Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, Shanxi, China; Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA; Tel/ Fax: ++1-517-353-8623; E-mails: qlu@epi.msu.edu; wtstat1@sina.com

1389-2029/19 $58.00+.00

common SNPs and reduce the number of tests. Most of the above quality control procedures are implemented in the PLINK program [8]. PLINK offers several additional quality control procedures as well, such as detecting and removing duplicate samples.

Quality control procedures have also been developed for different types of studies. For population-based studies, individuals are required to be unrelated. PLINK and several other software offer a procedure to identify and remove individuals who are related. For family-based genetic data, statistical methods and software such as PREST [9] can be used to identify sample and pedigree structure errors based on the autosomal markers [10]. Misspecified familial relationships can also be detected by PREST [9], which infers the underlying relationships from the shared Identical By Descent (IBD). The reassigned family relationships can also be verified by using PREST [11, 12]. However, if the software fails to resolve the error, those families are subsequently treated as missing [13]. A Mendelian inconsistencies check is another commonly used quality control procedure for family data. Mendelian inconsistencies can be detected and set as missing by using the PedCheck [14] and Merlin [15] programs [11]. Markers with a high frequency of Mendelian segregation errors are often excluded from the analysis [13].

3. LINKAGE ANALYSIS

Linkage analysis uses family data to locate genetic regions that are in linkage with a given trait. It was one of the most commonly used tools to infer the location of trait-related genes on chromosomes before the age of Genome-Wide Association Studies (GWAS) [16]. Linkage analysis can be classified into model-based linkage analysis and model-free linkage analysis. While model-based linkage analysis assumes a particular genetic model (*e.g.*, the causal allele frequency, penetrance functions) [17], the model-free linkage analysis does not require such assumptions. These two analyses are sometimes referred to as parametric and non-parametric, although both of them require the estimation of parameters [18].

Model-based linkage analysis is powerful for scenarios with prior knowledge of the underlying genetic models, and has been implemented in several software packages (*e.g.*, Merlin, FASTLINK). However, for complex traits like SUD, we have limited knowledge about the underlying genetic model. Misspecification of model parameters in SUD model-based linkage analysis can lead to biased estimators and incorrect inferences.

Model-free linkage analysis is based on allele-sharing, which usually compares genetic similarity among affected relatives inferred from data to that from expectations. Because the model-free linkage methods do not require the specification of the inheritance parameters of a genetic model, they have been adopted in many genetic studies of SUD [17]. One of the most popularly used model-free linkage methods is the Haseman-Elston algorithm [19], which was later extended by Sham *et al.* [20]. The extended algorithm can perform genome-wide linkage analyses and has been implemented in the program Merlin [21]. GENEHUNTER [22], another linkage method for affected sib-pair analysis, has also been used by many SUD researchers [23]. The variance component method has also been widely used. It partitions the phenotype variations among family members into different variance components explained by the effects of the genes in the region of interest, additive genetic effects of other genes, and non-shared environmental determinants [24]. Software such as Merlin can be used to perform a variance components linkage analysis. All of the above programs can be used to conduct both two-point or multi-point linkage analyses.

Model-free linkage analysis is a robust but less powerful approach. Although including unaffected individuals generally provide more power for linkage analysis, there is a potential loss of power in the event of misclassification of affection status [23]. On the other hand, model-based linkage analysis requires knowledge of the parameters of the genetic model, and has advantages in terms of efficiency and a better location solution. Researchers, therefore, need to determine which methods to use based on the study purpose.

Although linkage analysis can be used to map genes, the SUD-related chromosome regions identified by linkage analysis can be up to several million bases and contain thousands of genes. Moreover, family data especially those with large pedigrees are difficult to collect. Compared to linkage analysis, association analysis is generally more powerful and can be used for fine mapping. Thus, linkage analysis is used less often than association analysis in the recent SUD genetic research. To borrow the strengths from both analysis, some researchers [11] have adopted a multistage design that combines linkage analysis with association analysis to identify genetic variants associated with SUD.

4. ASSOCIATION ANALYSIS

Linkage analysis has been successfully implemented in many family studies, identifying genetic regions linked to SUD. However, linkage analysis is subject to a few limitations, including low power for detecting variants with small effects and being less useful for the fine mapping of causal variants. As stated above, some of these limitations can be overcome by association studies [25]. Association analysis is based on the concept of linkage disequilibrium. It can be classified into population-based association analysis and family-based association analysis. In other words, SUD-based association studies aimed at identifying genes related with SUD can be conducted by means of a population-based design (*e.g.*, case-control design) with unrelated individuals or a family-based design with related individuals [25]. This section provides an overview of the current statistical approaches and software for population-based and family-based SUD association studies.

4.1. Population-based Association Study

Population-based association studies are conducted on unrelated samples. The most commonly adopted population-based association studies are case-control studies, which compare the distribution of genetic variants in affected subjects with those in unaffected subjects. We summarize below methods for single-locus association analysis, multi-locus association analysis, and briefly discuss the issues of population stratification and multiple testing.

4.1.1. Methods for Single-locus Association Analysis

Assuming that a genetic variant is a Single Nucleotide Polymorphism (SNP) with three genotypes, AA, Aa and aa, where A is the minor frequent allele. One way to code the SNP is assuming an additive allelic effect (*i.e.*, AA=2, Aa=1, aa=0). The additive model, indicating that each additional copy of the A allele increases the disease risk [26], is commonly used in SUD association analysis [27]. In certain circumstances, where researchers have prior knowledge of the underlying mode of inheritance, other models can also be used. For instance, we could assume the SNP follows a dominant mode of inheritance (*i.e.*, AA=Aa=1, aa=0) or a recessive mode of inheritance (*i.e.*, AA=1, Aa=aa=0) [28]. Based on the coded genetic variable, classic statistical methods can be employed to link the SNP to different types of phenotypes.

Standard contingency table methods can be used to test the null hypothesis of no association of the SNP with a categorical phenotype. While the Z test and Chi-squared test are the typical methods, the Fisher's Exact Test can be adopted for studies with a small sample size. A contingency table method, such as Cochran-Armitage trend statistics [29] can be used to assess the association between a binary phenotype and an SNP with additive coding, which is equivalent to the score test in a logistic regression [30]. A logistic regression model is another popular method for analyzing data with a binary phenotype. It is more flexible than the Cochran-Armitage trend test and can accommodate covariates and model interaction effects [31]. A Manhattan plot is often used to visualize the results from a single-locus association analysis, and a Quantile-Quantile (Q-Q) plot is drawn to check systematic biases (*e.g.*, bias due to population stratification) [32].

Similar to logistic regression, linear regression can be used for testing an association between a genetic marker and a quantitative phenotype, allowing for covariates adjustment [33]. Linear regression assumes that the quantitative phenotype follows a normal distribution. For non-normally distributed phenotypes, it is necessary to transform the phenotypic data to normal before model-fitting or use a non-parametric method (*e.g.*, U statistics). Popular transformations include log transformation, Box-Cox transformation, and normal quantile transformation. Both linear regression and logistic regression are special cases of Generalized Linear Models (GLM), which have been implemented in most of existing statistical software packages.

4.1.2. Methods for Multi-locus Association Analysis

Complex diseases are often caused by multiple genetic variants and other factors (*e.g.*, environmental determinants) [34]. Since each genetic variant plays a small role in complex diseases, a joint association analysis of multiple SNPs (*e.g.*, all SNPs in a gene) is able to accumulate the effects from multiple SNPs and reduce the multiple testing issue [18, 31].

Methods used for single-locus analysis can also be extended for analyzing multiple SNPs. For instance, multiple logistic regression can be adopted for multi-locus association analysis. The corresponding score test is a generalization of the Armitage test and is related to the Hotelling-T^2 statistic [35].

Most of the methods introduced in the multiple-SNP analysis section do not consider Linkage Disequilibrium (LD) among SNPs. To take LD into account, haplotype-based analysis can be used to infer an association of a specific haplotype or haplotypes with a disease phenotype of interest. Before we perform a haplotype test, we need to first infer the haplotype phase based on genotypes. The likelihood-based approach and the Expectation-Maximization (EM) algorithm can be used to infer possible haplotype phases. A nice feature of the likelihood-based approach is that a likelihood-ratio test can be easily formed to test the association of the inferred haplotypes with the disease phenotype. Alternatively, we can also adopt a score test, which is computationally efficient and is widely used in SUD genetic research. The score test has been implemented in the program HAPLO.STATS [36, 37]. Other programs for haplotype phase estimation include Bayesian approaches, such as those implemented in PHASE and HAPLOTYPER [18]. The Bayesian approach and the partition-ligation (PL) algorithm implemented in the program PHASE [38] are reported to be more accurate in reconstructing haplotypes than the EM algorithm [39, 40].

Similar to multiple-SNP association analysis, a haplotype analysis can potentially reduce the multiple testing burden and be more powerful than a single-locus association analysis, especially when there are strong haplotype-specific effects. Nevertheless, a haplotype-based association analysis may also be subject to issues of uncertain phasing and low power due to a large number of inferred haplotypes [31]. There are several applications of haplotype-based analysis. For instance, Wang *et al.* found that both *CHRNA2* and *CHRNA6* were significantly associated with Nicotine Dependence (ND) [41] based on the haplotype-based analysis.

4.1.3. Issues in Population-based Association Analysis

Multiple testing is an important issue in association analysis, especially for a genome-wide association analysis involving millions of SNPs. Performing an association analysis on such a large number of SNPs without considering the issue of multiple testing could result in substantial false positive findings. The Bonferroni correction is an easy way to address this issue [42]. For genome-wide association data with 1 million SNPs, the Bonferroni correction at a family-wise error rate (FWER) of 5% results in a genome-wide significance level of 5×10^{-8} [42-44]. While the Bonferroni correction is easy to implement, it tends to be conservative and does not consider the LD among SNPs [45]. Nevertheless, the Bonferroni correction is still the most popular approach in SUD genetic research. Another popular approach for multiple testing correction is the False Discovery Rate (FDR) approach, which controls the expected proportion of false positives among all rejected hypotheses [46]. FDR tends to less conservative than the Bonferroni correction.

In population-based association studies, spurious associations could be caused by a variety of confounding factors. Confounding bias has been of concern in SUD genetic studies using samples from multiple ethnic groups (*i.e.*, population stratification), especially in large-scale national/international studies. For genome-wide association studies, Q-Q plot is a useful tool for visualizing a systematical bias due to factors such as population stratification. Deviation

Table 1. Summary information of statistical software for SUD genetic research.

Software Packages	URL	Platform	Programming Language	Limitations	Computational Cost	Public/ Commercial
PLINK	https://www.cog-genomics.org/plink2	Windows/ Linux/ Mac	C, C++	--	--	Public
PREST	http://www.utstat.toronto.edu/sun/ Software/Prest	Linux	C++		--	Public
PedCheck	https://watson.hgen.pitt.edu/register/ docs/pedcheck.html	Windows/ Linux/ Mac	C, C++	PedCheck performs single-locus analysis only.	--	Public
Merlin	http://csg.sph.umich.edu/abecasis/ Merlin/download/	Windows/ Linux/ Mac	C++	--	--	Public
FASTLINK	https://cran.r-project.org/web/packages /fastLink	Windows/ Linux/ Mac	R	FASTLINK is likely to have difficulty of producing high-quality matches when the overlap between two data sets is small.	--	Public
GENEHUNTER	http://www-genome.wi.mit.edu/ ftp/distribution /software/genehunter	Windows/ Linux/	C	It is not suitable for analyzing large multi-generational pedigrees.	--	Public
HAPLO. STATS	https://cran.r-project.org/web/packages/haplo.stats	Windows/ Linux/ Mac	R	--		Public
PHASE	http://www.stat.washington.edu/stephens/software.html	Windows/ Linux/ Mac	C++	When haplotype analysis is conducted on a large number of loci, many of the resulting haplotypes will have small frequencies. The power of haplotype analysis is hence significantly reduced due to the large number of degrees of freedom [18].	It is computationally slower than the HAPLOTYPER program.	Public

(Table 1) contd….

Software Packages	URL	Platform	Programming Language	Limitations	Computational Cost	Public/ Commercial
HAPLO-TYPER	http://www.people.fas.harvard.edu/~junliu/Haplo/click.html	Windows/ Linux/ Mac	C++	Efficiency may be lost because of incorrect prior information. When the sample size is small, the power of haplotype analysis is reduced [110].	It is computationally faster than the PHASE program.	Public
EIGENSTRAT	http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm	Linux	C++	Researchers should pay attention to the experimental design, matching the ancestry and laboratory treatment of cases and controls to the fullest extent possible [111].	--	Public
STRUCTURE	http://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/ structure.html	Windows/ Linux/ Mac	C, Java	--	--	Public
SKAT	http://cran.r-project.org/ web/packages/SKAT	Windows/ Linux/ Mac	R	--	--	Public
FBAT, PBAT	http://www.biostat.harvard.edu/~clange/default.htm	Windows/ Linux/ Mac	C, C++	FBAT are designed for analyzing nuclear families, but in practice general pedigrees are commonly collected [112].	It is computationally fast.	Public
METAL	http://www.sph.umich.edu/csg/abecasis/metal/	Windows/ Linux/ Mac/ Unix	C++	--	The analysis of 15 studies and 36 million association tests requires <6 min computing time.	Public

(Table 1) contd….

Software Packages	URL	Platform	Programming Language	Limitations	Computational Cost	Public/ Commercial
GWAMA	http://www.well.ox.ac.uk/GWAMA	Windows/ Unix /newer	C++, R, PERL	--	--	Public
GCTA	http://cnsgenomics.com/software/gcta/#Download	Windows/ Linux/ Mac	C, C++	GCTA only considers additive effects and SNP data while excluding other types of genetic variations (*e.g.*, rare mutations, CNVs) [113].	GCTA needs a few minutes to analyze a sample with n < 3000, but needs a few hours for n > 10,000 [93].	Public

Note: The software packages are arranged in the order where they appear in the article.

from the diagonal line indicates possible population stratification and inflation of spurious associations. One of the most common methods for controlling population stratification is the principal component analysis. In a principal component analysis, each individual receives scores on each principal component. These scores, representing continuous variations in race and ethnicity, can be used to control for the effects of population stratification [47]. The EIGENSTRAT can estimate Principal Components (PCs) on a sufficient number of SNPs, which can then be used as covariates in a regression model to control for population stratification [35]. Several other methods have also been used, including the program STRUCTURE. STRUCTURE first classifies the samples into different clusters and then tests the association within each cluster [48]. The genomic control approach, which estimates genomic inflation factor λ, is another fast and easy way to detect and adjust for population stratification [37]. Unfortunately, most of the above methods are designed to adjust for global ancestry, so they could be inadequate when local population structure is an important confounding factor.

4.2. Rare Variants Association Test

While many common variants have been identified to be associated with SUD, much of the genetic contribution to SUD remains unexplained. There is an increasing interest in studying SUD-related rare variants [49-51]. For such a purpose, methods for rare-variants analysis have been developed.

Methods developed for single-locus analysis can also be used to detect rare variants. However, due to high dimensional and low frequent features of rare variants, single-locus analysis is subject to low power [52]. Single-locus analysis is useful only if the sample size is large enough and/or the effect sizes are large [52].

Compared to single-locus tests, multiple-locus test is generally more powerful for rare variants analysis. Numerous multiple-locus methods and software have been developed for analyzing rare variants. Shrinkage methods, such as LASSO [53] and Bayesian Lasso [54], can be incorporated into the regression framework to select disease-related SNPs and to deal with the high-dimensional genetic data. In addition to these methods, collapsing methods, weighted sum methods and linear mixed methods (*e.g.*, SKAT) have also been developed to detect rare variants. While several SUD-related rare variants have been identified, overall they have a limited contribution to SUD [55].

4.3. Family-based Association Analysis

Compared to population-based association studies, family-based association studies provide robust protection against population stratification at the design stage. In a family-based association study, a typical Transmission Disequilibrium Test (TDT) compares the alleles that are transmitted to an affected child from parents to the alleles that are not transmitted. Therefore, it matches the ancestry background of samples within families, and provides robustness against population stratification at a locus-specific level. Many methods have been proposed for family-based association analysis. Some of these methods are extensions of population-based association methods while accounting for familiar correlations. We only review a few of methods commonly used for SUD genetic research. There are several popularly adopted family-based study designs, such as case-parent trio design and case-sibling design. We briefly introduce these designs and then summarize methods associated with these designs (Table **1**).

A case-parent trio design is one of the most popularly used family-based study designs, in which cases and their parents are recruited. The well-known method for case-parent trio design is the Transmission Disequilibrium Test (TDT) [56]. The classic TDT is essentially a McNemar test that compares the distribution of alleles transmitted from parents to affected offspring with that of the non-transmitted alleles [57]. The original form of TDT can only be applied to one affected child with known parental genotypes. It has been further extended to continuous phenotype, multiple offspring, and missing parental genotypes.

Case-sibling design, in which each proband is matched with one or more unaffected siblings, is an alternative design when the parental genotypes are missing. Case-sibling design requires at least one affected offspring and one unaffected offspring, and their genotypes can't be identical. Although case-sibling designs are generally robust to population stratification, they are not as powerful as unrelated case-control design because of the similarity of siblings [58].

A Family-Based Association Test (FBAT) is the most commonly adopted method for SUD family-based association analysis. FBAT can be considered an extension of TDT, which allows for any type of genetic model, affected and unaffected siblings, qualitative and quantitative phenotypes and multiple markers [59-62]. FBAT has also been extended to incorporate haplotypes and gene-environment interaction analysis [63]. A software package for FBAT is freely available online. An association test implemented in the FBAT program has the option of computing the *p*-value from the *Z* statistic using Monte Carlo sampling under the null hypothesis of no linkage and no association [63]. *P*-value from FBAT can be computed either by asymptotic theory or by permutation test. FBAT has been successfully used in many studies. For instance, Shirley *et al.* identified 6 SNPs associated with alcohol dependence by using FBAT [64]. The TDT and FBAT approaches are usually applied to nuclear family data, but in practice, general pedigrees are commonly collected. A Pedigree Disequilibrium Test (PDT) is a method specifically designed for general pedigrees [65], and has been widely adopted in family-based association analysis of SUD. It is also worthwhile mentioning that the integrated software package PBAT (v. 3.5), which contains tools for the design of family-based association studies as well as tools for data analysis [66].

5. G-G/G-E INTERACTION ANALYSIS

Significant progress has been made toward the identification of genetic variants contributing to SUD. Despite these achievements, a large proportion of SUD phenotype variations remain unexplained. Gene-gene and gene-environment (G-G/G-E) interactions could play an important role in the SUD development process. The identification of these G-G/G-E interactions not only explain additional SUD variations, but also elucidate how genetic variants interplay with environmental determinants to cause SUD. A G-G/G-E interaction study, either using a population-based design or a family-based design, takes into account the complex relationship between genetic variants and environmental determinants, and could lead to novel G-G/G-E interactions contributing to SUD.

Most of G-G/G-E interaction studies incorporate a population-based design. While conventional methods, such as contingency-table-based methods and regression-based methods [67, 68], are very useful for G-G/G-E interaction analysis on a handful of variables, they are subject to low performance when there are a large number of genetic variants. Efforts have been made to extend the classic methods to high-dimensional genetic data, such as the development of stepwise logistic regression [69-71] and LASSO [72, 73], but the implementation of these methods on genome-wide data remains a great challenge [74]. One of the popular methods for SUD G-G/G-E interaction analysis is the Multifactor Dimensionality Reduction (MDR) [75]. The fundamental principle of MDR is pooling multi-locus genotypes into high-risk and low-risk groups based on case-to-control ratio, which effectively reduces the data dimensions. The Generalized Multifactor Dimensionality Reduction (GMDR) method further extends the original MDR method for handling various phenotypes, taking covariates into consideration [76]. MDR, as well as GMDR, employ an exhaustive search algorithm, which could be computationally expensive when considering high-order interactions or a larger number of genetic variants. GMDR-GPU, developed by Zhu *et al.*, not only allows for GWAS data but also runs much faster than the earlier version of the GMDR program by utilizing a more computationally efficient implementation [77, 78]. GMDR has been successfully used in detecting an ND-association interaction between *CHRNA4* and *CHRNB2* [79]. Other advanced methods, such as a neural network and random forest, have been used in the G-G/G-E interaction analysis, though none of them have been widely used in SUD G-G/G-E studies.

Methods have also been developed for family-based G-G/G-E interaction analysis. FBAT, which allows various pedigree structures, various phenotypes and multiple markers, has been extended to G-G/G-E interaction analysis [57]. Similarly, the extensions of MDR and GMDR, MDR-PDT [80] and PGMDR [81], have also been developed for family-based G-G/G-E interaction analysis. PGMDR allows for covariate adjustment and could attain more power than MDR-PDT.

6. META-ANALYSIS

GWAS allows us to identify SUD-associated variants with large or modest effect sizes. In order to discover SUD-associated variants with smaller effect sizes or low frequency, large-scale genetic studies with a much larger sample size are needed [32]. Meta-analysis has become popular because it does not require individual-level data and is capable of integrating information (*e.g.*, *p*-values and odd ratios) from multiple independent studies with a substantially increased sample size [82]. Based on a much larger sample, it increases the power of association studies, leading to the discovery of novel small-effect variants and low-frequency variants [28, 83]. Classic meta-analysis methods typically based on combining effect sizes (*i.e.*, fixed effects or random effects), *p*-values, or *Z* scores [28]. In this section, we summarize the classic methods available for SUD meta-analysis.

Prior to conducting a meta-analysis, we need to test heterogeneity to make sure that individual studies are similar to each other. I^2 is a statistic metric that has been widely used for evaluating heterogeneity. It measures the proportion of total variation between studies attributed to heterogeneity [84]. Many programs, such as METAL [85] and GWAMA [86], can estimate I^2 for the evaluation of heterogeneity.

P-value-based meta-analysis methods have been used for decades but have become less popular due to their limitations (*e.g.*, including weights is not a straightforward process) [87]. Among these methods, the Fisher's method [82] and the *Z*-score-based method are often adopted and are closely related to each other [34]. In a typical meta-analysis study, each site generates *p*-values or *Z* test statistics, and then uploads the results to a server. A meta-analysis method can be implemented to combine the information from all sites for the discovery of low-frequent variants or small-effect variants [82].

Methods for combining effect sizes can be based on two types of models: fixed effect models and random effect models. A mixed effect model, such as the DerSimonian and Laird model [88], is preferred when heterogeneity exists among studies. Otherwise, a fixed effect model should be used [32]. The major advantage of fixed effect models over random effect models is that they maximize discovery power [89]. Inverse variance weighting [90] and Cochran-Mantel-Haenszel [91] are commonly used approaches that provide similar results for fixed-effect-based meta-analyses [34]. Popularly packages for fixed-effect-based meta-analyses include METAL, GWAMA, R and PLINK, among which GWAMA and R can also be used for random-effect meta-analysis.

7. GENOME-WIDE COMPLEX TRAIT ANALYSIS

The genetic variants that have been discovered so far only account for a small fraction of SUD heritability [92]. We should point out that the heritability discussed in this section is the narrow-sense heritability, *i.e.*, the proportion of phenotypic variance due to additive genetic variance. Genome-wide Complex Trait Analysis (GCTA) can help us to study "missing heritability" and to identify which parts of a genome or categories of variants contribute to SUD heritability [93]. Yang *et al.* [94] developed a versatile tool, GCTA, to study heritability. GCTA is a user-friendly software with four main functions: data management, genetic relationships estimation from SNPs, mixed linear model analysis of variance and linkage disequilibrium structure estimation [93].

GCTA was developed based on a linear mixed method to address the "missing heritability" problem [94]. Yang *et al.* believe that "missing heritability" can be partially explained by small or medium-effect-size SNPs that don't reach the stringent genome-wide significance level. GCTA first calculates a Genetic Relationship Matrix (GRM), which estimates the genetic relationships between all pairs of individuals based on the genetic data. Given the calculated GRM, it further estimates the variance components by using a Restricted Maximum Likelihood (REML) algorithm from the Linear Mixed Model (LMM) [93], and then infers the heritability. In LMM, covariates (*e.g.*, sex, age, principle components) can also be easily adjusted. Using the actual genetic data, the estimated heritability from GCTA is expected to be close to that from twins or segregation studies. For example, the estimation of heritability of nicotine dependence is close to 50% [95, 96] provided by twin and family studies. GCTA can not only provide a better heritability estimate but also allows researchers to investigate which parts of the genome or categories of variants contribute to SUD [93].

To better model known disease-related variants, Yang *et al.* developed a GCTA-COJO module by treating the effects of known variants as fixed effects and excluding these variants from GRM [97]. This module has been applied to SUD data. Clarke *et al.* used the GCTA-COJO module for a stepwise conditional analysis, in which disease-associated SNPs are gradually added to the model until there is no SNP significantly associated with SUD [98]. Other functions, such as the one computing LD scores and the one calculating the eigenvectors of GRM, are also included in GCTA [93, 99, 100].

8. POST-GENOMICS AND INTEGRATIVE ANALYSIS

The advent of high-throughput technologies makes it feasible to study multi-level omic data, which includes not only genotype data but also other types of data, such as gene expression and DNA methylation data.

An expression Quantitative Trait Loci (eQTL) is genetic loci that regulate gene expression [101]. eQTLs mapping is a useful tool to find genomic locations likely regulating transcript expression. Methods of eQTL mapping have been developed to dissect genetic variants affecting gene expression [102]. The statistical methods, such as single-marker regression analysis, interval mapping, and multiple interval mapping [103], have been widely used for eQTL mapping. The main difference between the eQTL mapping and the QTL mapping method is that a large number of gene expressions are analyzed simultaneously in eQTL mapping. Therefore, the issue of multiple testing need to be considered. Typically, *p*-values corresponding to the peaks of LOD score curves from each transcript are obtained and are adjusted for multiple testing by using FDR [104]. However, this approach only takes LOD score peaks into account [105]. In order to address this limitation, statistical methods, such as Empirical Bayes methods, are developed to control the overall FDR [106]. eQTL mapping provides a list of transcripts that can regulate transcriptional expression. The identification of the hot spots is usually the next task. The easiest way to determine eQTL hot spots is to directly count the number of localized transcripts. However, it can lead to spurious identifications or "ghost" hotspots under a number of situations. Therefore, statistical tests, such as a Poisson-based test, are used to determine the true hot spots [104]. Additionally, Kendziorski *et al.* summarize evidence across every transcript to identify significant eQTL hot spots [106].

In addition to eQTL analysis, integrative analysis combining genetic data with other data sources (*e.g.* transcriptional, epigenetic, and metabolite data) can provide a comprehensive view of disease biological mechanism [107]. For example, by integrating GWAS and Epigenome-Wide Association Study (EWAS) data, we can not only identify variants at the DNA and epigenetic level but also identify the modulation caused by epigenetic change of genetic variants leading to gene expression alterations. Given the rapid development in both technology and analytical capability, we anticipate that the integrative approach will grow rapidly to provide novel means to study SUD [108, 109].

CONCLUSION

This paper provides an overview of statistical methods and software for data management, linkage analysis, association analysis and other analysis involved in SUD genetic research. The quality control process is used to identify and remove inaccurate information before the downstream analysis. Stringent quality control ensures the results from the downstream analysis are reliable. In this paper, we summarized some filters commonly used in SUD genetic research and discussed various quality control filters and criteria used by researchers.

The initial understanding of the genetic influences of SUD is assisted by linkage mapping, resulting in the identification of several SUD-associated regions. However, regions identified through linkage analysis could harbor many genes, which makes fine-mapping of the disease-related gene difficult. This problem can be overcome by an association analysis [25]. Depending on the study design, association analysis be classified into population-based association analysis and family-based association analysis. Population-based association studies have many advantages, such as the easy collection of study samples and high power. Nevertheless, population-based association studies are subject to population stratification issues. Although association analysis has been widely used in SUD genetic research, the identified SUD-related variants only explain a small fraction of the heritability [93]. GCTA is a versatile and user-friendly tool for estimating the heritability explained by all SNPs. The application of GCTA to the SUD genetic dataset suggests that heritability can be explained by small- to medium-effect variants and their possible interactions [94]. With the increased sample size, we are able to detect these small- to medium-effect variants. Meta-analysis is an attractive strategy to combine information from different studies to detect small- to medium-effect variants [32]. Recently, the development of sequencing and other omics technologies have promoted the development of advanced statistical methods and software. There is a pressing need for the development of new analytical tools to keep pace with fast-growing technology, and the increasing amount of SUD omic data. We anticipate that more advanced methods and software will be developed in the future, facilitating the SUD genetic discovery process.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

This study was supported, in part, by the National Institute on Drug Abuse under Award Number R01DA043501 and the National Library of Medicine under Award Number R01LM012848.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

REFERENCES

[1] Peiper, N.C.; Ridenour, T.A.; Hochwalt, B.; Coyne-Beasley, T. Overview on prevalence and recent trends in adolescent substance use and abuse. *Child Adolesc. Psychiatr. Clin. N. Am.*, **2016**, *25*(3), 349-365.

[2] Bevilacqua, L.; Goldman, D. Genes and addictions. *Clin. Pharmacol. Ther.*, **2009**, *85*(4), 359-361.

[3] Prom-Wormley, E.C.; Ebejer, J.; Dick, D.M.; Bowers, M.S. The genetic epidemiology of substance use disorder: a review. *Drug Alcohol Depend.*, **2017**, *180*, 241-259.

[4] Laura Jean, B. Genetic vulnerability and susceptibility to substance dependence. *Neuron*, **2011**, *69*(4), 618-627.

[5] Wang, J.C.; Kapoor, M.; Goate, A.M. The Genetics of Substance Dependence. *Annual Review of Genomics & Human Genetics*, **2012**, *13*(1), 241.

[6] Vink, J.M.; Willemsen, G.; Boomsma, D.I. Heritability of smoking initiation and nicotine dependence. *Behav. Genet.*, **2005**, *35*(4), 397-406.

[7] Neil, C.; Fangyi, G.; Nilanjan, C.; Jin, S.C.; Kai, Y.; Meredith, Y.; Constance, C.; Kevin, J.; William, W.; Maria, T.L. Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One*, **2009**, *4*(2), e4653.

[8] Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **2007**, *81*(3), 559-575.

[9] Mcpeek, M.S.; Sun, L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.*, **2000,** *66*(3), 1076-1094.

[10] Gizer, I.R.; Ehlers, C.L.; Vieten, C. Seaton-Smith, K.L.; Feiler, H.S.; Lee, J.V.; Segall, S.K.; Gilder, D.A.; Wilhelmsen, K.C. Linkage scan of nicotine dependence in the University of California, San Francisco (UCSF) Family Alcoholism Study. *Psychol. Med.*, **2011,** *41*(4), 799-808.

[11] Shizhong, H.; Bao-Zhu, Y.; Kranzler, H.R.; David, O.; Raymond, A.; Farrer, L.A.; Joel, G. Linkage analysis followed by association show NRG1 associated with cannabis dependence in African Americans. *Biol. Psychiatry*, **2012**, *72*(8), 637-644.

[12] Bao-Zhu, Y.; Shizhong, H.; Kranzler, H.R.; Farrer, L.A.; Joel, G. A genomewide linkage scan of cocaine dependence and major depressive episode in two populations. *Neuropsychopharmacology*, **2011**, *36*(12), 2422-2430.

[13] Gizer, I.R.; Ehlers, C.L.; Vieten, C.; Seaton-Smith, K.L.; Feiler, H.S.; Lee, J.V.; Segall, S.K.; Gilder, D.A.; Wilhelmsen, K.C. Linkage scan of alcohol dependence in the UCSF Family Alcoholism Study. *Drug Alcohol Depend.*, **2011**, *113*(2), 125-132.

[14] O'Connell, J.R.; Weeks, D.E. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **1998**, *63*(1), 259-266.

[15] Abecasis, G.A.R.; Cherny, S.S.; Cookson, W.O.; Cardon, L.R. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **2002**, *30*(1), 97-101.

[16] Schnell, A.H.; Sun, X. Model-based linkage analysis of a quantitative trait. In: *Statistical Human Genetics. Methods in Moecular Biolgy;* Elston, R.C.; Satagopan, J.M.; Sun, S. Eds.; Humana Press, **2012**; Vol. *850*, pp. 263-283.

[17] Xu, W.; Bull, S.B.; Mirea, L.; Greenwood, C.M.T. Model-free linkage analysis of a binary trait. In: *Statistical Human Genetics. Methods in Moecular Biolgy;* Elston, R.C.; Satagopan, J.M.; Sun, S. Eds.; Humana Press, **2012**; Vol. *850*, pp. 317.

[18] Lu, Q.; Song, Y.; Gray-Mcguire, C. *Software for Genetics/Genomics*, John Wiley & Sons: New York, **2013**.

[19] Haseman, J.K.; Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, **1972**, *2*(1), 3-19.

[20] Sham, P.C.; Purcell, S.; Cherny, S.S.; Abecasis, G.R. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am. J. Hum. Genet.*, **2002**, *71*(2), 238-253.

[21] Joel, G.; Carolien, P.; Roger, W.; Kathleen, B.; James, P.; Michael, K.; Lindsay, F.; Kranzler, H.R. Genomewide linkage scan for nicotine dependence: identification of a chromosome 5 risk locus. *Biol. Psychiatry*, **2007**, *61*(1), 119-126.

[22] Kruglyak, L.; Lander, E.S. Faster multipoint linkage analysis using Fourier transforms. *J. Comput. Biol.*, **1998**, *5*(1), 1-7.

[23] Gelernter, J.; Panhuysen, C.; Weiss, R.; Brady, K.; Hesselbrock, V.; Rounsaville, B.; Poling, J.; Wilcox, M.; Farrer, L.; Kranzler, H.R. Genomewide linkage scan for cocaine dependence and related traits: significant linkages for a cocaine-related trait and cocaine-induced paranoia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **2010**, *136B*(1), 45-52.

[24] Amy, W.; Lind, P.A.; Jelger, K.; Feiler, H.S.; Smith, T.L.; Schuckit, M.A.; Kirk, W. The investigation into CYP2E1 in relation to the level of response to alcohol through a combination of linkage and association analysis. *Alcoholism Clin. Exp. Res.*, **2011**, *35*(1), 10-18.

[25] Nielsen, D.A.; Kreek, M.J. Common and specific liability to addiction: approaches to association studies of opioid addiction. *Drug Alcohol Depend.*, **2012**, *123*(1), S33-S41.

[26] Balding, D.J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **2006**, *7*(10), 781-791.

[27] Mccarthy, M.; Abecasis, G.; Cardon, L.; Goldstein, D.; Little, J.; Ioannidis, J.; Hirschhorn, J. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **2008**, *9*(5), 356-369.

[28] Wang, M.H.; Cordell, H.J.; Steen, K.V. Statistical methods for genome-wide association studies. *Semin. Cancer Biol.*, **2018**, *55*, 53-60.

[29] Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics*, **1955**, *11*(3), 375-386.

[30] Agresti, A. *Categorical Data Analysis*, 2nd ed.; Wiley & Sons: New York, 2003.

[31] Balding, D.J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, **2006**, *7*, 781-791.

[32] Zeng, P.; Zhao, Y.; Qian, C.; Zhang, L.; Zhang, R.; Gou, J.; Liu, J.; Liu, L.; Chen, F. Statistical analysis for genome-wide association study. *J. Biomed. Res.*, **2015**, *29*(4), 285-297.

[33] Langefeld, C.D.; Fingerlin, T.E. Association Methods in Human Genetics. In: *Topics in Biostatistics*, Ambrosius, W.T. Ed.; Humana Press: Totowa, NJ, **2007**; pp. 431-460.

[34] Camastra, F.; Di, T.M.; Staiano, A. Statistical and computational methods for genetic diseases: an overview. *Comput. Math. Methods Med.*, **2015**, *2015*, 1-8.

[35] Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; David, R. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **2006**, *38*(8), 904-909.

[36] Schaid, D.J.; Rowland, C.M.; Tines, D.E.; Jacobson, R.M.; Poland, G.A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **2002**, *70*(2), 425-434.

[37] Levran, O.; Londono, D.; O'Hara, K.; Nielsen, D.A.; Peles, E.; Rotrosen, J.; Casadonte, P.; Linzy, S.; Randesi, M.; Ott, J. Genetic susceptibility to heroin addiction: a candidate gene association study. *Genes Brain Behav.*, **2010**, *7*(7), 720-729.

[38] Stephens, M.; Donnelly, P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **2003**, *73*(5), 1162-1169.

[39] Stephens, M.; Smith, N.J.; Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **2001**, *68*(4), 978-989.

[40] Niu, T.; Qin, Z.S.; Xu, X.; Liu, J.S. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **2002,** *70*(1), 157-169.

[41] Wang, S.; A, D.v.d.V.; Xu, Q.; Seneviratne, C.; Pomerleau, O.F.; Pomerleau, C.S.; Payne, T.J.; Ma, J.Z.; Li, M.D. Significant associations of CHRNA2 and CHRNA6 with nicotine dependence in European American and African American populations. *Hum. Genet.*, **2014**, *133*(5), 575-586.

[42] Itsik, P.E.; Roman, Y.; David, A.; Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.*, **2010**, *32*(4), 381-385.

[43] Mckay, J.D.; Hung, R.J.; Han, Y.; Zong, X.; Carreras-Torres, R.; Christiani, D.C.; Caporaso, N.E.; Johansson, M.; Xiao, X.; Li, Y. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.*, **2017**, *49*(7), 1126.

[44] Risch, N.; Merikangas, K. The future of genetic studies of complex human diseases. *Science*, **1996**, *273*(5281), 1516-1517.

[45] Treutlein, J.; Rietschel, M. Genome-wide association studies of alcohol dependence and substance use disorders. *Curr. Psychiatry Rep.*, **2011**, *13*(2), 147-155.

[46] Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **1995**, *57*(1), 289-300.

[47] Bierut, L.J.; Arpana, A.; Bucholz, K.K.; Doheny, K.F.; Cathy, L.; Elizabeth, P.; Sherri, F.; Louis, F.; William, H.; Sarah, B. A genome-wide association study of alcohol dependence. *Proc. Natl. Acad. Sci. U.S.A.*, **2010**, *107*(11), 5082-5087.

[48] Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, **2000,** *155*(2), 945-959.

[49] Rivas, M.A.; Beaudoin, M.; Gardet, A.; Stevens, C.; Sharma, Y.; Zhang, C.K.; Boucher, G.; Ripke, S.; Ellinghaus, D.; Burtt, N.; Fennell, T.; Kirby, A.; Latiano, A.; Goyette, P.; Green, T.; Halfvarson, J.; Haritunians, T.; Korn, J.M.; Kuruvilla, F.; Lagace, C.; Neale, B.; Lo, K.S.; Schumm, P.; Torkvist, L.; Dubinsky, M.C.; Brant, S.R.; Silverberg, M.S.; Duerr, R.H.; Altshuler, D.; Gabriel, S.; Lettre, G.; Franke, A.; D'Amato, M.; McGovern, D.P.; Cho, J.H.; Rioux, J.D.; Xavier, R.J.; Daly, M.J. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, **2011**, *43*(11), 1066-1073.

[50] Gudmundsson, J.; Sulem, P.; Gudbjartsson, D.F.; Masson, G.; Agnarsson, B.A.; Benediktsdottir, K.R.; Sigurdsson, A.; Magnusson, O.T.; Gudjonsson, S.A.; Magnusdottir, D.N.; Johannsdottir, H.; Helgadottir, H.T.; Stacey, S.N.; Jonasdottir, A.; Olafsdottir, S.B.; Thorleifsson, G.; Jonasson, J.G.; Tryggvadottir, L.; Navarrete, S.; Fuertes, F.; Helfand, B.T.; Hu, Q.; Csiki, I.E.; Mates, I.N.; Jinga, V.; Aben, K.K.; van Oort, I.M.; Vermeulen, S.H.; Donovan, J.L.; Hamdy, F.C.; Ng, C.F.; Chiu, P.K.; Lau, K.M.; Ng, M.C.; Gulcher, J.R.; Kong, A.; Catalona, W.J.; Mayordomo, J.I.; Einarsson, G.V.; Barkardottir, R.B.; Jonsson, E.; Mates, D.; Neal, D.E.; Kiemeney, L.A.; Thorsteinsdottir, U.; Rafnar, T.; Stefansson, K. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.*, **2012**, *44*(12), 1326-1329.

[51] Jonsson, T.; Atwal, J.K.; Steinberg, S.; Snaedal, J.; Jonsson, P.V.; Bjornsson, S.; Stefansson, H.; Sulem, P.; Gudbjartsson, D.; Maloney, J.; Hoyte, K.; Gustafson, A.; Liu, Y.; Lu, Y.; Bhangale, T.; Graham, R.R.; Huttenlocher, J.; Bjornsdottir, G.; Andreassen, O.A.; Jonsson, E.G.; Palotie, A.; Behrens, T.W.; Magnusson, O.T.; Kong, A.; Thorsteinsdottir, U.; Watts, R.J.; Stefansson, K. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature*, **2012**, *488*(7409), 96-9.

[52] Lee, S.; Abecasis, G.R.; Boehnke, M.; Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **2014**, *95*(1), 5-23.

[53] Tong, W.T.; Fang, C.Y.; Trevor, H.; Eric, S.; Kenneth, L. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **2009**, *25*(6), 714-721.

[54] Jiahan, L.; Kiranmoy, D.; Guifang, F.; Runze, L.; Rongling, W. The Bayesian lasso for genome-wide association studies. *Bioinformatics*, **2011**, *27*(4), 516-523.

[55] Asimit, J.; Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.*, **2010**, *44*, 293-308.

[56] Spielman, R.S.; Mcginnis, R.E.; Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **1993**, *52*(3), 506-516.

[57] Wen, Y.; Lu, Q. *Analysis of Gene-Gene Interactions Underlying Human Disease,* John Wiley & Sons: New York, **2014**.

[58] Teng, J.; Risch, N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.*, **1999**, *9*(3), 234-241.

[59] Nan, M.L.; Lange, C. Family- based methods for linkage and association analysis. *Adv. Genet.*, **2008**, *60*(4), 219-252.

[60] Rabinowitz, D.; Laird, N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.*, **2000**, *50*(4), 211-223.

[61] Lange, C.; Van, S.K.; Andrew, T.; Lyon, H.; Demeo, D.L.; Raby, B.; Murphy, A.; Silverman, E.K.; Macgregor, A.; Weiss, S.T. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat. Appl. Genet. Mol. Biol.*, **2004**, *3*(1), 1-17.

[62] Sungho, W.; Wilk, J.B.; Mathias, R.A.; O'Donnell, C.J.; Silverman, E.K.; Kathleen, B.; O'Connor, G.T.; Weiss, S.T.; Christoph, L. On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet.*, **2009**, *5*(11), e1000741.

[63] Steve, H.; Xin, X.; Lake, S.L.; Silverman, E.K.; Weiss, S.T.; Laird, N.M. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet. Epidemiol.*, **2004**, *26*(1), 61-69.

[64] Hill, S.Y.; Jones, B.L.; Zezza, N.; Stiffler, S. ACN9 and alcohol dependence: family-based association analysis in multiplex alcohol dependence families. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **2015**, *168b*(3), 179-187.

[65] Martin, E.R.; Monks, S.A.; Warren, L.L.; Kaplan, N.L. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet.*, **2000**, *67*(1), 146-154.

[66] Laird, N.M.; Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet*, **2006**, *7*(5), 385-394.

[67] James, G.W. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.*, **2002**, *155*(5), 478-484.

[68] Wang, S.; Zhao, H. Sample size needed to detect gene-gene interactions using association designs. *Am. J. Epidemiol.*, **2003**, *158*(9), 899-914.

[69] Cordell, H.J.; Clayton, D.G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.*, **2002**, *70*(1), 124-141.

[70] Josephine, H.; Jurg, O. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.*, **2003**, *4*(9), 701-709.

[71] Marchini, J.; Donnelly, P.; Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **2005**, *37*, 413-417.

[72] Dahinden, C.; Parmigiani, G.; Emerick, M.C.; Bühlmann, P. Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, **2007**, *8*(1), 476.

[73] Li, M.; Romero, R.; Fu, W.J.; Cui, Y. Mapping haplotype-haplotype interactions with adaptive LASSO. *BMC Genet.,* **2010**, *11*(1), 79.

[74] Li, M.; Lou, X.Y.; Lu, Q. On epistasis: a methodological review for detecting gene-gene interactions underlying various types of phenotypic traits. *Recent Pat. Biotechnol.*, **2012**, *6*(3), 230-236.

[75] Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, **2001**, *69*(1), 138-147.

[76] Lou, X.Y.; Chen, G.B.; Yan, L.; Ma, J.Z.; Zhu, J.; Elston, R.C.; Li, M.D. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.*, **2007**, *80*(6), 1125-1137.

[77] Zhu, Z.; Tong, X.; Zhu, Z.; Liang, M.; Cui, W.; Su, K.; Li, M.D.; Zhu, J. Development of GMDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. *PLoS One*, **2013**, *8*(4), e61943.

[78] Jiekun, Y.; Ming, D.L. Association and interaction analyses of 5-HT3 receptor and serotonin transporter genes with alcohol, cocaine, and nicotine dependence using the SAGE data. *Hum. Genet.*, **2014**, *133*(7), 905-918.

[79] Li, M.D. Detection of gene-gene interaction among CHRNA4, CHRNB2, BDNF and NTRK2 in nicotine dependence. *Biol. Psychiat.*, **2008**, *64*(11), 951-957.

[80] Martin, E.R.; Ritchie, M.D.; Hahn, L.; Kang, S.; Moore, J.H. A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genet. Epidemiol.*, **2006**, *30*(2), 111-23.

[81] Lou, X.Y.; Chen, G.B.; Yan, L.; Ma, J.Z.; Mangold, J.E.; Zhu, J.; Elston, R.C.; Li, M.D. A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am. J. Hum. Genet.*, **2008**, *83*(4), 457-467.

[82] Evangelou, E.; Ioannidis, J.P. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, **2013**, *14*(6), 379-389.

[83] Sagoo, G.S.; Little, J.; Higgins, J.P. Systematic reviews of genetic association studies. *Hum. Genome Epidemiol. Network. PLoS Med.*, **2009**, *6*(3), e28.

[84] Higgins, J.P.; Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.*, **2002**, *21*(11), 1539-1558.

[85] Willer, C.J.; Abecasis, G.R.; Li, Y. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **2010**, *26*(17), 2190-2191.

[86] Magi, R.; Morris, A.P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, **2010**, *11*, 288.

[87] Eleftheria, Z.; Ioannidis, J.P.A. Meta-analysis in genome-wide association studies. *Pharmacogenomics*, **2009**, *10*(2), 191-201.

[88] Dersimonian, R.; Nan, L. Meta-analysis in clinical trials. *Control. Clin. Trials*, **1986**, *7*(3), 177-188.

[89] Pereira, T.V.; Patsopoulos, N.A.; Salanti, G.; Ioannidis, J.P. Discovery properties of genome-wide association signals from cumulatively combined data sets. *Am. J. Epidemiol.*, **2009**, *170*(10), 1197-1206.

[90] Kavvoura, F.K.; Ioannidis, J.P. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum. Genet.*, **2008**, *123*(1), 1-14.

[91] Mantel, N. Chi-square tests with one degree of freedom; extensions of the mantel- haenszel procedure. *J. Am. Stat. Assoc.*, **1963**, *58*(303), 690-700.

[92] Goldstein, D.B. Common genetic variation and human traits. *N. Engl. J. Med.*, **2009**, *360*(17), 1696-1698.

[93] Yang, J.; Lee, S.H.; Goddard, M.E.; Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **2011**, *88*(1), 76-82.

[94] Jian, Y.; Beben, B.; Mcevoy, B.P.; Scott, G.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **2010**, *42*(7), 565-569.

[95] Vrieze, S.I.; McGue, M.; Miller, M.B.; Hicks, B.M.; Iacono, W.G. Three mutually informative ways to understand the genetic relationships among behavioral disinhibition, alcohol use, drug use, nicotine use/dependence, and their co-occurrence: twin biometry, GCTA, and genome-wide scoring. *Behav. Genet.*, **2013**, *43*(2), 97-107.

[96] Palmer, R.H.; McGeary, J.E.; Heath, A.C.; Keller, M.C.; Brick, L.A.; Knopik, V.S. Shared additive genetic influences on DSM-IV criteria for alcohol dependence in subjects of European ancestry. *Addiction*, **2015**, *110*(12), 1922-1931.

[97] Yang, J.; Ferreira, T.; Morris, A.P.; Medland, S.E.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; Weedon, M.N.; Loos, R.J.; Frayling, T.M.; McCarthy, M.I.; Hirschhorn, J.N.; Goddard, M.E.; Visscher, P.M. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **2012**, *44*(4), 369-375.

[98] Clarke, T.K.; Adams, M.J.; Davies, G.; Howard, D.M.; Hall, L.S.; Padmanabhan, S.; Murray, A.D.; Smith, B.H.; Campbell, A.; Hayward, C. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). *Mol. Psychiatr*., **2017**, *22*(10), 1376-1384.

[99] Otto, J.M.; Gizer, I.R.; Ellingson, J.M.; Wilhelmsen, K.C. Genetic variation in the exome: Associations with alcohol and tobacco co-use. *Psychol. Addict. Behav.*, **2017**, *31*(3), 354-366.

[100] Brazel, D.M.; Jiang, Y.; Hughey, J.M.; Turcot, V.; Zhan, X.; Gong, J.; Batini, C.; Weissenkampen, J.D.; Liu, M.; Barnes, D.R.; Bertelsen, S.; Chou, Y.L.; Erzurumluoglu, A.M.; Faul, J.D.; Haessler, J.; Hammerschlag, A.R.; Hsu, C.; Kapoor, M.; Lai, D.; Le, N.; de Leeuw, C.A.; Loukola, A.; Mangino, M.; Melbourne, C.A.; Pistis, G.; Qaiser, B.; Rohde, R.; Shao, Y.; Stringham, H.; Wetherill, L.; Zhao, W.; Agrawal, A.; Bierut, L.; Chen, C.; Eaton, C.B.; Goate, A.; Haiman, C.; Heath, A.; Iacono, W.G.; Martin, N.G.; Polderman, T.J.; Reiner, A.; Rice, J.; Schlessinger, D.; Scholte, H.S.; Smith, J.A.; Tardif, J.C.; Tindle, H.A.; van der Leij, A.R.; Boehnke, M.; Chang-Claude, J.; Cucca, F.; David, S.P.; Foroud, T.; Howson, J.M.M.; Kardia, S.L.R.; Kooperberg, C.; Laakso, M.; Lettre, G.; Madden, P.; McGue, M.; North, K.; Posthuma, D.; Spector, T.; Stram, D.; Tobin, M.D.; Weir, D.R.; Kaprio, J.; Abecasis, G.R.; Liu, D.J.; Vrieze, S. Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. *Biol. Psychiatr.*, **2018**, *85*(11), 946-955.

[101] Liu, C. Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases. *Neurosci. Bull.*, **2011**, *27*(2), 123-133.

[102] Gaffney, D.J.; Veyrieras, J.B.; Degner, J.F.; Pique-Regi, R.; Pai, A.A.; Crawford, G.E.; Stephens, M.; Gilad, Y.; Pritchard, J.K. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.*, **2012**, *13*(1), R7.

[103] Cookson, W.; Liang, L.; Abecasis, G.; Moffatt, M.; Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **2009**, *10*(3), 184-194.

[104] Kendziorski, C.; Wang, P. A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome*, **2006**, *17*(6), 509-517.

[105] Storey, J.D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*(16), 9440-9445.

[106] Kendziorski, C.M.; Chen, M.; Yuan, M.; Lan, H.; Attie, A.D. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, **2006**, *62*(1), 19-27.

[107] Sun, Y.V.; Hu, Y.J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.*, **2016**, *93*, 147-90.

[108] Sun, L.; Dimitromanolakis, A. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. *BMC Proc.*, **2014**, *8*(Suppl 1), S23.

[109] Kruglyak, L.; Daly, M.J.; Reeve-Daly, M.P.; Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **1996**, *58*(6), 1347-1363.

[110] Li, S.S.; Cheng, J.J.; Zhao, L.P. Empirical *vs.* Bayesian approach for estimating haplotypes from genotypes of unrelated individuals. *BMC Genet.*, **2007**, *8*, 2.

[111] Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **2006**, *38*(8), 904-909.

[112] Horvath, S.; Xu, X.; Lake, S.L.; Silverman, E.K.; Weiss, S.T.; Laird, N.M. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet. Epidemiol.*, **2004**, *26*(1), 61-69.

[113] Mayhew, A.J.; Meyre, D. Assessing the heritability of complex traits in humans: methodological challenges and opportunities. *Current genomics*, **2017**, *18*(4), 332-340.