Family-Based Genetic Association Analysis: Methods and Applications to Addiction Phenotypes

Camelia C. Minică

Family-Based Genetic Association Analysis: Methods and Applications to Addiction Phenotypes

Overige leden: Prof.dr. Michael C. Neale
Dr. Benjamin M. Neale
Prof.dr. Augustus B. Smit
Prof.dr. Henk Kelderman
Prof.dr. Eske M. Derks
Dr. Jouke-Jan Hottenga

Paranimfen: Dr. Hamdi Mbarek
Dr. Ştefan A. Minică

The investigations were supported by The European Research Council Starting Grant 284167

The SAS fellowship offered by the SAS Institute BV, Huizen, The Netherlands (licence agreement no. 97711, [175]), the travel grants offered by The Simons Foundation Fund, Nederlandse Vereniging voor Humane Genetica, Universitair Medisch Centrum, Utrecht, and Behavior Genetics Association, and the Talentgeld Research Grant offered by Vrije Universiteit Amsterdam, The Netherlands, are also gratefully acknowledged

Cover design: Camelia Minică, Vasile Ilieş, Olimpia Prislopean & Ştefan Minică

Front cover: *Test Fitting the Endless Column* (Constantin Brâncuşi, 1937), photo by Ştefan Georgescu-Gorjan, from the Gorjan Archive, used with permission from Sorana-Constantza Georgescu-Gorjan

Back cover: *Effects of Radiation on DNA's Double Helix* (detail), image credit NASA: `http://er.jsc.nasa.gov`, from *Radiation Biology Educator Guide*

LaTeX typesetting by Ştefan Minică

Printed and bound by Ipskamp Drukkers BV

ISBN: 978-94-028-0014-2

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C112051

VRIJE UNIVERSITEIT

Family-Based Genetic Association Analysis: Methods and Applications to Addiction Phenotypes

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op donderdag 18 februari 2016 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Camelia Claudia Minică

geboren te Jibou, Roemenië.

promotoren: prof.dr. D.I. Boomsma
prof.dr. J.M. Vink
prof.dr. C.V. Dolan

Memoriei mamei mele Maria Ilieş

Contents

1 Introduction 1

1.1 Introduction 1

1.2 Overview: Stages of Smoking Behavior 1

1.3 Overview: Early Stages of Cannabis Use 3

1.4 Statistical Power: An Important Consideration in GWAS 4

1.5 Family-Based Genetic Association Analyses: Methods and Applications to Addiction Behaviors 7

1.6 Outline 10

I Part I: Methods 13

2 Genetic Association in Multivariate Phenotypic Data: Power in Five Models 15

2.1 Introduction 17

2.2 Procedure 19

2.3 Study 1: Single common genetic factor 22

2.4 Study 2: Correlated genetic common factors 25

2.5 Study 3: Latent regression model 31

2.6 Study 4: Hybrid simplex (A,E)-factor (C) model 35

2.7 Study 5: Stationary double simplex (A,E) model 39

2.8 Discussion 39

2.9 Appendix: R code for calculating power. 44

3 The Use of Imputed Sibling Genotypes in Sibship-Based Association Analysis: on Modeling Alternatives, Power and Model Misspecification 47

3.1 Introduction 49

3.2 Methods 50

3.2.1 Models for sibship-based association 50
3.2.2 The full information model 50
3.2.3 The mixture approach 51
3.2.4 The dosage approach 53
3.3 Model fitting 54
3.4 Simulation details 54
3.5 Misspecification of the background covariance structure 54
3.6 Empirical illustrations 55
3.6.1 Height data 55
3.6.2 The analysis of smoking initiation 56
3.7 Results 57
3.7.1 The full information setting 57
3.7.2 The mixture and dosage approaches 58
3.7.3 Additional simulations: Explaining height results 66
3.7.4 The analysis of smoking initiation 66
3.8 Discussion 69

4 Sandwich Corrected Standard Errors in Family-Based Genome-wide Association Studies 73
4.1 Introduction 75
4.2 Methods 75
4.2.1 Family based model for genetic association 75
4.2.2 Estimation 76
4.2.3 Simulation details 78
4.3 Results 80
4.3.1 Correctly specified background model: type I and type II rates 80
4.3.2 Misspecified background model 81
4.3.3 FaST-LMM formulation of the ML sandwich correction . . 83
4.4 Discussion 85

5 MZ Twin Pairs or MZ Singletons in Population Family-Based GWAS? More Power in Pairs. 87
5.1 MZ Twin Pairs or MZ Singletons? 89
5.1.1 Supplementary Results: Type I Error Rate 92

6 The Weighting Is The Hardest Part: On The Behavior of the Likelihood Ratio Test and Score Test Under Data-Driven Weighting Scheme in Rare Variant Association Studies 95
6.1 Introduction 97
6.2 Material and Methods 98
6.2.1 Model formulation 98

6.2.2 Tests of variance components 99
6.2.3 Data simulation 100
6.2.4 Data-Driven search for optimal weights: exploring the misspecification space 101
6.2.5 Evaluating the type I error rates and power 102
6.2.6 Software 102
6.2.7 Empirical analysis: testing the constrained and the FMRP-Darnell gene sets for rare case mutations enrichment . . . 102
6.3 Results 105
6.3.1 Type I error 105
6.3.2 Power 105
6.3.3 Empirical analysis: testing the constrained and the FMRP-Darnell gene sets for rare case mutations enrichment . . . 112
6.4 Discussion 114
6.5 Supplemental Table and Figures 116

II Part II: Applications 119

7 Heritability, SNP- and Gene-Based Analyses of Cannabis Use Initiation and Age at Onset 121
7.1 Introduction 123
7.2 Materials And Methods 125
7.2.1 Phenotypes 125
7.2.2 Initiation of cannabis use ('ever/never') 125
7.2.3 Age at onset 126
7.2.4 Genotypes 126
7.3 Statistical analyses 127
7.3.1 Estimating the heritability of initiation 127
7.3.2 Power analysis 127
7.3.3 SNP-based association analysis of initiation 128
7.3.4 SNP-based survival analysis of age at onset 128
7.3.5 Gene-based analyses of initiation and age at onset 128
7.4 Results 129
7.4.1 Estimating heritability based on genetic relatedness 129
7.4.2 SNP- and gene-based analyses of initiation 129
7.4.3 SNP- and gene-based analyses of age at onset 133
7.5 Discussion 135

8 Genome-Wide Association Study of Cannabis Initiation Based on a Large Meta-Analytic Sample of 32,330 Subjects from The International Cannabis Consortium 139
8.1 Introduction 141

8.2 Materials and methods 143
8.2.1 Cohorts 143
8.3 Phenotype and covariates 144
8.3.1 Genotype imputation 144
8.4 Statistical analyses 144
8.4.1 GWAS in each discovery cohort 144
8.4.2 Meta-analysis of GWAS results 144
8.4.3 Estimation of SNP-based heritability and genetic overlap with nicotine use 145
8.4.4 Gene-based tests 145
8.5 Results 146
8.5.1 SNP-based heritability and genetic overlap with smoking initiation 146
8.5.2 Meta-analysis 146
8.5.3 Gene-based tests 148
8.6 Discussion 150
8.7 Conclusion 153

9 Survival Meta-Analysis of Age at Onset of Cannabis Use 155
9.1 Introduction 157
9.2 Materials And Methods 159
9.2.1 Sample 159
9.2.2 Phenotyping 159
9.2.3 Genotyping 159
9.2.4 Imputation 160
9.2.5 Quality checks prior to the meta-analysis 160
9.2.6 Statistical analysis 160
9.2.7 Meta-analysis 161
9.2.8 Gene-based tests 161
9.3 Results 162
9.3.1 Meta-analysis 162
9.3.2 Gene-based tests 167
9.4 Discussion 171

10 Pathways to Smoking: Biological Insights from the Tobacco and Genetics Consortium Meta-Analysis 173
10.1 Introduction 175
10.2 Materials and Methods 176
10.2.1 Sample description 176
10.2.2 Statistical analyses 176
10.3 Results 177
10.3.1 Gene-based tests 177
10.3.2 Pathway-based tests 178

10.4 Discussion 183

11 Summary 187

12 Discussion 191

12.1 Empirical Findings: Genes and Biological Pathways Implicated in Smoking Behaviors and Initiation of Cannabis Use 193

12.1.1 Conclusion 196

12.2 Means of Improving Statistical Power in GWAS 197

12.2.1 On the definition of the phenotype 197

12.2.2 In time of test, family is best 201

12.2.3 Set-based analyses: expedient in a genome-wide scan . . . 204

12.2.4 Conclusion 206

13 Chapters' Origin 207

14 Samenvatting 211

List of Tables 215

List of Figures 221

Bibliography 227

Chapter 1

Introduction

1.1 Introduction

The present thesis was written with the ultimate aim of finding genes and biological pathways implicated in addiction behaviors such as tobacco smoking (henceforth, smoking) and cannabis use. Notwithstanding the relatively high heritability of substance use phenotypes - ranging from about 40% for initiation of cannabis use (Vink, Wolters et al. [348]) to about 75% for smoking dependence (Vink, Willemsen et al. [347]) - the progress in mapping the associated genes has been slow. By 2012, when the current project started, few genetic loci were detected for tobacco smoking initiation, dependence and cessation. Several inconsistently replicated results were reported by candidate gene studies for cannabis dependence, while no genetic locus had been reported for initiation of cannabis use. In this chapter I first briefly review the current state of affairs with respect to the genetics of stages of tobacco smoking behavior and early stages of cannabis use. I then identify problems and promising solutions for improving statistical power which are to be explored in the thesis, especially as they apply to genetic association studies.

1.2 Overview: Stages of Smoking Behavior

Despite smoking prevalence decreasing in the past 30 years, there has been a steady increase in the absolute number of smokers, i.e., from 721 million in 1980 to 967 million in 2012 due to accelerated population growth (Ng, Freeman et al. [257]). Tobacco smoking kills nearly 6 million people yearly (World Health Organization, 2015) and it is well recognized as the world's leading cause of preventable disease and death. Regular tobacco use is a known risk factor for various disease traits such as lung cancer (Prevention 2008, Lee, Forey et al. [196]), leukemia (e.g., see Fircanis, Merriam et al. [125]), heart disease (e.g., see Huxley and Woodward [171]), chronic bronchitis and emphysema (see e.g., Forey, Thornton et al. [128]). Importantly, genetic factors account for about 32% - 55% of the variance in smoking initiation (in males and females, respectively; see e.g., Li, Cheng et al. [347], Vink, Willemsen et al. 2005), while the remaining variance is explained by shared (by the family members) and unique environmental factors. Note that whilst the shared environmental factors play an important role in early stages of smoking, their contribution to dependence is nil. That is, the individual differences in smoking dependence are largely explained by genetic factors, which account for about 56% to 75% of the variance (see e.g., Lessov, Martin et al. [198], Vink, Willemsen et al. [347], Broms, Silventoinen et al. [47]). The significant contribution of genetic factors to variability in smoking behaviors has been recently confirmed by a chip-based heritability study (Lubke, Hottenga et al. [220]), based on GCTA (Yang, Lee et al. [375]). According to Lubke et al. the currently typed (and tagged) common single nucleotide polymorphisms (SNPs) explain about 19% and about 24% of the variance in smoking initiation and current smoking, respectively. These estimates taken together with those derived from twin studies suggest that both common and rare variants contribute to individual differences in smoking behavior (given that GCTA focuses on common variants, while twin studies capture the effects of all genetic variants, common and rare). Genome-wide association studies and large meta-analyses conducted by consortia such as Tobacco and Genetics Consortium (TAG; 2010 [322]), the ENGAGE Consortium (Thorgeirsson, Gudbjartsson et al. [321]) and the Oxford-GlaxoSmithKline consortium (Liu, Tozzi et al. [214]) have located several common single nucleotide polymorphisms (SNPs) implicated in stages of smoking behavior. For instance, the TAG consortium (Tobacco and Genetics Consortium 2010 [322]) conducted collaborative analyses which combined three meta-analytic samples, namely the TAG, ENGAGE and the Oxford-GlaxoSmithKline samples (N=140,000 individuals) to follow-up 1,025 SNPs that passed the 10^{-4} threshold in the TAG sample. This consortium located 14 SNPs associated with different stages of smoking behaviors. Of these 14, 5 SNPs tagging the *CHRNA3-CHRNA5-CHRNB4* cluster of genes on chromosome 15q25, the non-coding RNA *LOC123688* region on 10q23, and the *EGLN2* gene on chromosome 19 were associated with quantity smoked, 8 SNPs tagging the *BDNF* gene on chromosome 11 were significantly associated with smoking initiation (i.e., ever/never smoking regularly), and one SNP tagging the *DBH* gene was significantly associated with smoking cessation (for details on the 14 SNPs that reached significance in the combined TAG analysis, see Table 2 on page 443, in Tobacco and Genetics Consortium 2010 [322]). The multiple genetic associations with quantity smoked at the 15q25 locus (i.e., including the *CHRNA3-CHRNA5-CHRNB4* cluster of genes; Saccone, Hinrichs et al. [288], Saccone, Wang et al. [287], Tobacco and Genetics Consortium 2010 [322]) were thoroughly interrogated and confirmed by fine-mapping by the Oxford-GlaxoSmithKline consortium (Liu, Tozzi et al. [214]). The ENGAGE Consortium meta-analysis (Thorgeirsson, Gudbjartsson et al. [321]) implicated in quantity smoked SNPs tagging the *CYP2A6* and *CYP2B6* genes, and SNPs tagging the nicotinic acetylcholine receptor subunits *CHRNB3* and *CHRNA6*.

1.3 Overview: Early Stages of Cannabis Use

Cannabis is among the drugs with the highest frequency of (ab)use. About 22.3% Europeans aged 16-64 reported to have experimented with cannabis use. In the United States the prevalence in ages 16-34 was estimated at 51.6% (European Monitoring Centre for Drugs and Drug Addiction, 2012). Early experimentation with cannabis (before age 18) has been shown to 'open the gate' to experimentation with other drugs (Agrawal, Grant et al. [13], Lynskey, Vink et al. [225]) and to escalated drug use over time (e.g., see Lynskey et al. [223], Lynskey, Heath et al. [224]). Importantly, about 9% of those who experiment with cannabis use develop dependence (Budney, Roffman et al. [49], Volkow, Baler et al. [356]). Regular cannabis use during adolescence predicts poor educational (Lynskey and Hall [223], Horwood, Fergusson et al. [163]) and professional attainment (Fergusson and Boden [122], Volkow, Baler et al. [356]), health problems (Joshi, Joshi et al. [178], Hall [153]) and may increase the risk of developing psychotic disorders (Di Forti, Sallis et al. [98], Di Forti, Marconi et al. [97]). According to results of twin and family studies genetic factors explain about 40% of the variance in liability to initiate cannabis use, while the remaining variance is accounted for by shared and unshared environmental factors (both about 30%; Kendler and Prescott [182], van den Bree, Johnson et al. [330], Verweij, Zietsch et al. [343], Vink, Wolters et al. [348], Distel, Vink et al. [101]).

Relative to tobacco smoking behaviors, attempts to identify genetic variants underlying the heritability of cannabis use behaviors are fewer, and have met with limited success. Research has been concerned mainly with behavior relevant to psychiatric diagnosis, such as cannabis dependence (e.g., see Agrawal et al. and Hopfer et al. Hopfer, Young et al. [162], Agrawal, Wetherill et al. [7]). Psychiatric diagnoses have the benefit of being more precise relative to self-reported measures of use (Burmeister, McInnis et al. [53]). They are therefore expected to confer advantages in terms of power in genetic association studies. Yet higher resolution phenotyping is typically too expensive in sufficiently large samples. Consequently, several insights into the genetic architecture of clinical diagnosis of cannabis dependence has come from small-sample candidate gene studies. The candidate-gene study by Hopfer et al. (Hopfer, Young et al. [162]) performed in a sample of 541 subjects (of whom 327 met one or more Composite International Diagnostic Interview criteria of substance abuse) reported association signals within the central cannabinoid receptor (*CNR1*) gene (SNP rs806380; P-value = 0.02). The signal was confirmed by e.g., Agrawal et al. (P-value = 0.009; Agrawal, Wetherill et al. [7]), although it was not replicated in other studies (Herman, Kranzler et al. [158], Hartman, Hopfer et al. [154]; see for details also

Table 2 in Agrawal et al. Agrawal and Lynskey [7]). Association signals within the fatty acid amide hydrolase (*FAAH*) gene were also reported in candidate-gene studies of cannabis dependence symptoms by e.g. Tyndale et al. (P-value < 0.05; Tyndale, Payne et al. [323]). However, the association was not consistently replicated (e.g., Haughey, Marshall et al. [156] see review by Agrawal et al. Agrawal and Lynskey [7]).

Despite the moderate heritability of about 40% based on twin and family studies (Verweij, Zietsch et al. [343], Vink, Wolters et al. [348]), no genetic locus has been implicated in the liability to initiate cannabis use. Among the attempts to locate the genes underlying the heritability of liability to initiate cannabis use, a linkage study by Agrawal and colleagues (Agrawal, Morley et al. [14]) failed to locate statistically significant associated genomic regions. Similarly, a meta-analysis (Verweij, Vinkhuyzen et al. [342]) combining the results of two genome-wide association studies failed to detect common SNPs associated with individual differences in the liability to initiation.

1.4 Statistical Power: An Important Consideration in GWAS

Statistical power is a key consideration in seeking genetic variants (GVs) associated with psychiatric traits such as substance use. The past ten years of GWAS have taught us that the psychiatric traits are highly polygenic, subjected to the influence of many GVs of small effect, each typically accounting for less than .1% of the phenotypic variance. It is well recognized that relatively large samples are needed to locate these individual GVs of small effect (Psychiatric GWAS Consortium Coordinating Committee [72], Visscher, Brown et al. [352]). The need for large samples is also due to the large number of tests (e.g., as many as 6 million genetic variants may be tested in a genome-wide scan), which requires an adapted α level. As proposed by Pe'er et al. (Pe'er, Yelensky et al. [264]) typically an α of 5×10^{-8} (corrected for multiple testing by taking into account that the tests are correlated) is used, to ensure the family-wise error rate does not exceed 0.05 (see Sham and Purcell [295] for more details on the burden of multiple testing in GWAS).

Statistical power in this context is defined as the probability of detecting the association of a GV with a given phenotype, given that the GV is truly associated, i.e., gives rise to individual differences in the studied trait. This involves testing the statistical association between the observed genetic variant and the phenotype in an appropriate regression model. For instance, if the phenotype $\mathbf{y}$ is continuous, we can use a linear regression model:

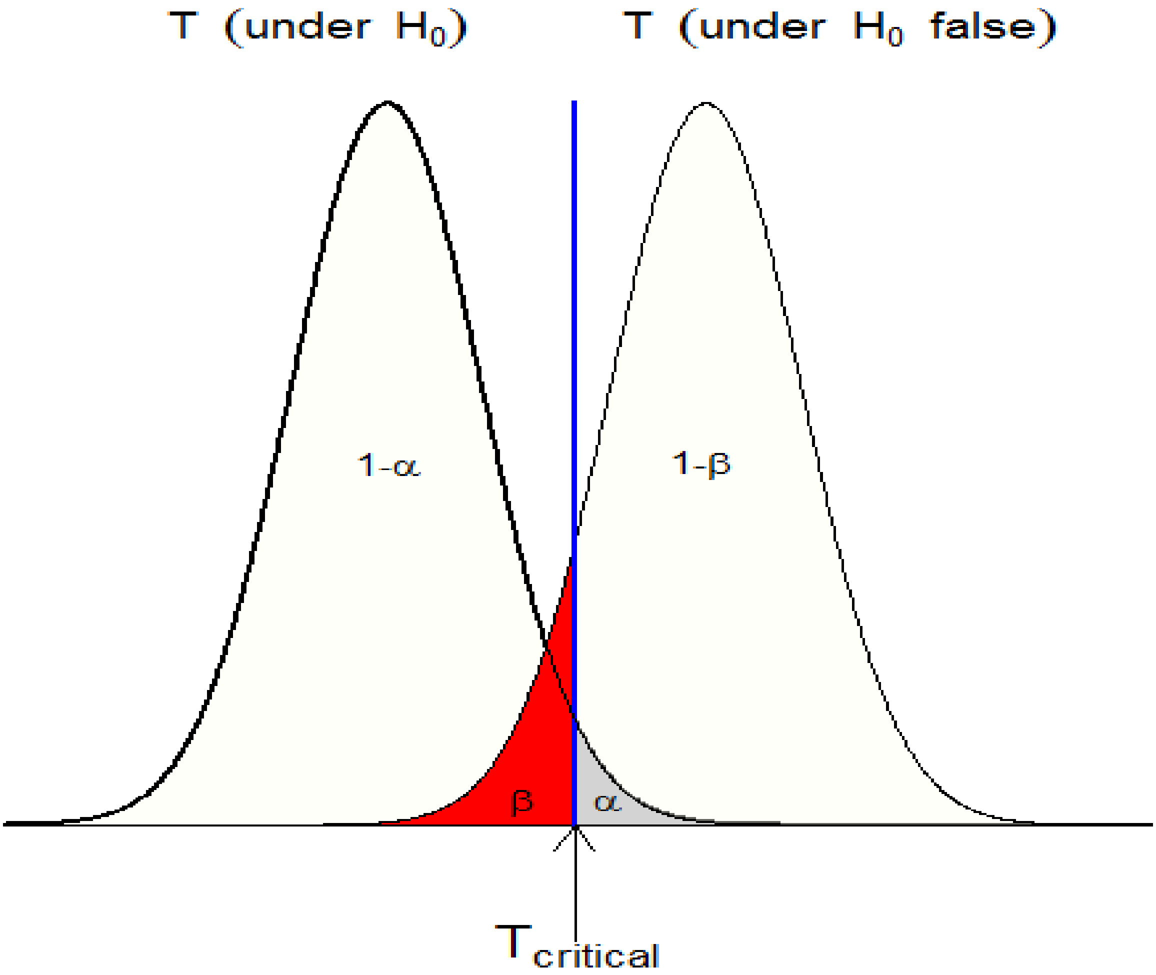
$$\mathbf{y}_j = \mathbf{b}_0 + \mathbf{b}_1 \times \mathbf{g}_j + \boldsymbol{\epsilon}_j \tag{1.1}$$

where j ($j = 1 \ldots n$) stands for individual, $\mathbf{y}_j$ is the phenotype score of individual

j, $\mathbf{b}_0$ is the intercept, $\mathbf{b}_1$ is the beta or the regression coefficient, $\mathbf{g}_j$ is the genotypic values at the tested locus (coded additively-codominant as 0, 1 or 2 to represent the number of effect alleles) and $\boldsymbol{\epsilon}_j$ is the random residual. The parameter of interest is the genetic effect, as one wants to determine whether the effect of the GV is unequal zero, i.e., reject the null hypothesis H_0: $\mathbf{b}_1 = 0$ in favor of the two-sided alternative hypothesis H_1: $\mathbf{b}_1 \neq 0$. By formulating a two-sided alternative hypothesis no prediction is made with respect to whether an additional copy of the tested allele has an increasing or a decreasing effect on the phenotypic mean.

To determine whether to reject the null hypothesis one computes a test statistic (T) with known distributions under the null and under the alternative hypotheses. The T calculated based on the sample at hand is then compared to the critical value of the test statistic on the null distribution corresponding to the chosen significance level, i.e., α. The chosen α level represents the probability that one accepts of committing a type I error, i.e., incorrectly rejecting the null hypothesis (i.e., prob($T \geq T_{critical} \mid H_0$ is true) $= \alpha$). Typically α is set at 0.05. In the context of GWAS the type I error probability is the probability of incorrectly concluding that the genetic variant is associated with the trait when in fact it is not (i.e., a false positive result). Related to this is the type II error probability – denoted by β – that refers to the probability of incorrectly accepting the null hypothesis of no genetic association when in fact the genetic variant has a genuine effect on the trait (i.e., prob($T < T_{critical} \mid H_1$ is true) $= \beta$), a false negative result. The probability of correctly rejecting the null hypothesis is $1 - \beta$. This probability, i.e., prob($T \geq T_{critical} \mid H_1$ is true), is called the power of the test. These probabilities and the relationships among them are illustrated in Figure 1.1 (for more details on statistical power in genetics see Dolan and van den Berg [103], Sham and Purcell [295]).

Figure 1.1: The sampling distributions of a test statistic under the null (H_0) and under the alternative hypothesis (H_0 false). For this illustration, we assumed the sampling distribution of the test statistic under H_0 is standard normal, and we set the critical value at $T_{critical}$ corresponding to a type I error of α (grey). β (red) represents the probability of type II error, $1 - \beta$ represents the statistical power, and $1 - \alpha$ represents the probability of correctly rejecting the null hypothesis when the tested variant has no effect on the trait.



1.5 Family-Based Genetic Association Analyses: Methods and Applications to Addiction Behaviors

The call for large samples to increase the statistical power to detect genetic association has led to the foundation of international consortia (e.g., see Sullivan [310]). Yet, increasing the sample size is but one means of increasing statistical power. As I will discuss in this thesis, refinement of statistical methodologies is another (complementary) means. The thesis has a strong focus on statistical modeling, aiming first at assessing and selecting from the pool of available modeling approaches the most powerful and computationally efficient ones in the GWAS context. Next, I use powerful analytic strategies to perform genome-wide searches for genes and biological pathways implicated in early stages of cannabis use and in smoking behaviors. Finally, I tie together the recommendations stemming from both the power studies and the empirical analyses into an overall strategy for improving statistical power in GWAS.

The answers I advance to the question 'How can statistical power be improved in GWAS while retaining the computational speed?' have been inspired by the data collected at twin registries (see Hur and Craig [170]) such as the Danish (Harvald, Hauge et al. [155]), the Swedish (Magnusson, Almqvist et al. [228]), the Brisbane (Gillespie, Henders et al. [142]), or the Netherlands Twin Register (see also `http://www.twinstudies.org/information/twinregisters/`). Twin registries contain a wealth of multivariate phenotypic data, relating to many different traits and often observed at multiple occasions. Advancement in genotyping and imputation techniques have resulted in the addition of measured genetic information to these databases. I had the opportunity to work with data collected at the Netherlands Twin Register (NTR) which includes more than 175,000 participants with phenotypic data; biological data are available in more than 12,000 participants. Detailed phenotyping has been performed at multiple time points via questionnaires and in laboratory studies since 1986 (see Boomsma, Vink et al. [38], Boomsma, De Geus et al. [40], Willemsen, De Geus et al. [361], Willemsen, Vink et al. [362] for details) and has included cognitive abilities, health and lifestyle, personality, and psychiatric traits such as major depressive disorder, anxiety and addiction. Many of the measured traits are multivariate, i.e., the phenotype vector contains several distinct but interrelated components. Repeated measurement naturally gives rise to multivariate data. Several approaches are possible given multivariate data: (a) univariate analyses of each component of the phenotype vector, (b) univariate analyses of a sum score or a factor score, or (c) multivariate analyses. Multivariate techniques take account of the correlations among the multiple dependent variables and this may increase power. However, these techniques give rise to tests with many degrees of freedom and this in turn may reduce power. Another strategy is to conduct univariate analyses in which each trait is tested individually. But obviously, this does not exploit the multivariate nature of the data. Alternatively, one may collapse the data into a univariate score (e.g., an average). This allows one to take a univariate approach which is a computationally easy alternative. However, dimension reduction techniques may discard information about individual differences and this may lower the power. These observations raise the question: In which circumstances a multivariate approach is more powerful than a univariate one? The complication arises that while the phenotypic covariance structure of the phenotypes might be known, this does not imply knowledge of the polygenic covariance structure or the exact role of the measured genetic variant therein. The answers to the above question pertain to phenotype definition in GWAS and hence, are among the key components of a strategy for improving the power to detect genetic association.

A further important issue relates to the genotyping resources. In addition to the phenotypic data, the Netherlands Twin Register includes genotypic data observed in part of the sample (see Willemsen, De Geus et al. [361]). Although the cost of genotyping has greatly decreased in the recent years, the genotyping of large numbers of samples still remains expensive. For genome-wide association studies not all family members are necessarily genotyped. Often genotypic data are limited to 1 family member. Limiting the analysis to the genotyped individuals, given the presence of phenotypic data in close relatives, may not be optimal. Previous research has shown that family-based imputed genotypes can boost statistical power. This alternative, however, comes at a computational cost and requires modeling choices. Hence, pertinent questions are: (a) does the power gain hinge upon factors such as the phenotypic correlations among the family members or minor allele frequency?; (b) which individuals, if genotyped, would be maximally informative about the missing genotypes in their relatives?; (c) which modeling approach should one prefer in such a circumstance?, and (d) what are the effects on power of misspecification of the familial covariance matrix?

The availability of data collected in families raises further questions regarding modeling the familial covariance matrix. Which analytic strategy is the most optimal to run family-based whole genome scans, given that in this context optimality is defined both in terms of power and computational tractability? Researchers in the field often use unweighted least squares (ULS) estimator as it admits a closed form solution, and so it has the advantage of being fast. However, despite the speedy computation, ULS ignores the familial clustering and requires a sandwich correction to arrive at correct standard errors. With this approach, the model for the familial covariance matrix is severely misspecified, yet the effect on power of such misspecification has not been characterized. Although computationally more demanding, maximum likelihood (ML) is an important alternative. However, ML (as currently implemented in dedicated GWAS software), employs a misspecified model for traits with a shared environmental component. These observations raise the practical question: in conducting a family-based analysis, should one use ULS, which is fast, robust, and requires no model to be specified for the background covariance matrix, or should one use ML, which is efficient and fast, provided one commits to a background model limited to additive genetic and unshared environmental effects?

Another issue that needs to be addressed relates to the power gains conferred by the inclusion into genome-wide analyses of data on monozygotic (MZ) twin pairs. Twin registries include phenotypic and genotypic data of large numbers of monozygotic (MZ) twins, yet often one MZ twin pair member is typically dropped from the analysis (e.g., Lowe, Maller et al. [219], Parsons, Lester et al. [263], Loukola, Wedenoja et al. [218], Psychosis Endophenotypes International Consortium, Wellcome Trust Case-Control Consortium et al. 2014). MZ twins are genetically identical individuals, hence including both MZ twins in an analysis is presumably seen as redundant. Clearly, reducing MZ pairs to singletons decreases the sample size, but does this affect the effective sample size and so the power? And if that is so, what modeling alternatives are suitable for family-based samples including MZ twin pairs?

Recently, with the increasing availability of exome/genome sequencing data, researchers have shifted their focus from common to rare variants (RV). The plethora of applications (e.g., Cohen, Boerwinkle et al. [70], Huyghe, Jackson et al. [172], Zhan, Larson et al. [378], Cruchaga, Karch et al. [81], Peloso, Auer et al. [265]) and methodological papers on rare variant tests (see e.g., Li and Leal [199], Madsen and Browning [227], Price, Kryukov et al. [276], Wu, Lee et al. [367], Lee, Wu et al. [196], Chen, Meigs et al. [64], Ionita-Laza, Lee et al. [176], Listgarten, Lippert et al. [211], Lippert, Xiang et al. [210], Svishcheva, Belonogova et al. [312]; see also Franić, Dolan et al. [129] for an overview) as well as on analytic strategies for rare variant association meta-analyses (see e.g., Liu, Peloso et al. [213]) demonstrate that the interest in interrogating the rare variants' contribution to psychiatric traits has intensified particularly over the past five years. This shift in focus has been motivated by the hypothesis that genetic variants conferring risk for psychiatric traits may be novel and rare, in so far as they are subject to strong purifying selection. Set-based tests (focusing on the effect of a set of GVs, rather than on the effects of single GVs) such as the sequence kernel association test (SKAT) are widely used in rare variant association analysis. SKAT is based on a random effects model, in which the betas of the RVs are assumed to be drawn from a zero mean distribution and common variance. The common variance follows from a transformation of the betas, by multiplying them with specific weights. These weights are typically assigned based on meta-information about the RVs, such as allele frequency and functional predictions (Kryukov, Pennacchio et al. [191], Madsen and Browning [227], Price, Kryukov et al. [276], Wu, Lee et al. [367]), with rarer and functional variants expected to have larger effects. Allele frequency, in particular, is an important weighting factor, as the rarer the variant is, the stronger the average purifying selection coefficient (Pritchard [277], Schork, Murray et al. [292]). If this assumption is true, the betas for rare variants will tend to be larger than for more common variants. Correct weighting is expected to boost statistical power, and yet the correct weights are generally unknown. Hence the critical questions arise: What the effect is of weight misspecification in SKAT? The two tests employed for hypothesis testing in SKAT – the likelihood ratio test and score test – are asymptotically equivalent, but do they behave similarly within the misspecification space? How robust is their performance in the presence of weighted neutral variation?

The application studies employ the most efficient analytic strategies to locate genes and pathways of genes implicated in early stages of cannabis use and smoking behaviors. The searches for the relevant genes and pathways are conducted genome-wide and are based on the Netherlands Twin Register sample (with data on lifetime cannabis use and age at initiation), and three large meta-analytic samples: the first two comprising respectively $N = 32,330$ individuals and $N = 24,222$ individuals (with data on lifetime cannabis use and age at onset, data from the International Cannabis Consortium) and the third one comprising $N = 74,053$ individuals (with data on smoking behaviors, data made publicly available from the Tobacco and Genetics Consortium). As tests focused on individuals SNPs are still underpowered with the current samples, complementary to the single variant analysis I will use the set as the unit of analysis (i.e., where the set is the gene, the pathway or the whole genome). The applications are aimed to contribute to the field by: (a) providing a heritability estimate for initiation of cannabis use by considering jointly the currently measured SNPs; (b) identifying genes associated with lifetime cannabis use and age at onset; (c) locating genes and biological pathways implicated in quantity smoked, ever smoking, smoking cessation and age at initiation.

1.6 Outline

In short, the aim of the present thesis is two-fold: first, to study and select from the pool of available statistical methods the most powerful ones (while retaining computational speed) for conducting common and rare variant association studies; second, using powerful approaches to identify genes and biological pathways associated with cannabis use initiation and smoking behaviors. Corresponding to these aims, the thesis has two parts. **Part I** comprises five chapters which address methodological issues related to the power of genome-wide association studies of common and rare variants. **Part II** comprises four chapters dedicated to applications of powerful analytic strategies to find genes and pathways of genes implicated in early-stages of cannabis use and smoking behaviors. **Chapter II** inquires the real-world factors affecting the power to detect genetic association when multivariate/longitudinal data are used in GWAS. **Chapter III** considers the circumstances in which family-based imputation of unobserved genotypes and the subsequent use of these data in GWASs increases the power to detect genetic association. **Chapter IV** focuses on the clustered nature of the phenotypic data collected at the twin registries and inquires which estimator ensures the most efficient use of family data in GWASs. **Chapter V** evaluates the power advantages conferred by the inclusion of MZ twin pairs in association analyses. **Chapter VI** evaluates the behavior of the likelihood ratio test and the score test under weight misspecification in rare variant association analysis and proposes a weighting method to increase the power to detect association with sets of SNPs. **Chapter VII** is based on an application aimed at evaluating the individual and collective contribution of genetic variants to early stages of cannabis use in data from NTR. **Chapters VIII** and **IX** extend the searches for common genetic variants and genes implicated in initiation of cannabis use and age at onset in the International Cannabis Consortium meta-analytic samples. **Chapter X** aims at finding genes and biological pathways implicated in smoking behaviors in the Tobacco and Genetics meta-analytic sample. **Chapter XI** contains a summary of the thesis. **Chapter XII** discusses the implications of the empirical findings and ties together the recommendations stemming from the results of the power studies into an overall strategy for improving statistical power in GWAS.

Part I

Methods

Chapter 2

Genetic Association in Multivariate Phenotypic Data: Power in Five Models

Abstract

This chapter concerns the power of various data analytic strategies to detect the effect of a single genetic variant (GV) in multivariate data. We simulated exactly fitting monozygotic and dizygotic phenotypic data according to single and two common factor models, and simplex models. We calculated the power to detect the GV in twin 1 data in an ANOVA of phenotypic sum scores, in a MANOVA, and in exploratory factor analysis (EFA), in which the common factors are regressed on the genetic variant. We also report power in the full twin model, and power of the single phenotype ANOVA. The results indicate that: (1) If the GV affects all phenotypes, the sum score ANOVA and the EFA are most powerful, while the MANOVA is less powerful. Increasing phenotypic correlations further decreases the power of the MANOVA. (2) If the GV affects only a subset of the phenotypes, the EFA or the MANOVA are most powerful, while sum score ANOVA is less powerful. In this case, an increase in phenotypic correlations may enhance the power of MANOVA and EFA. If the effect of the GV is modeled directly on the phenotypes in the EFA, the power of the EFA is approximately equal to the power of the MANOVA.

2.1 Introduction

Well-established twin registries, such as the Scandinavian twin registers, (Peltonen [266]), Netherlands Twin Register (Boomsma et al. [40]), the UK Adult Twin Register (Spector and Williams [301]), and the Brisbane Adolescent twin study (Wright and Martin [365]), contain a wealth of multivariate phenotypic data, relating to many different phenotypes, and often observed at multiple occasions. Developments in genotyping technology, have resulted in the addition of measured genetic information to these databases (Willemsen et al. [361], Boomsma, Busjahn and Peltonen [38]). The availability of genetic data has allowed researchers to shift their focus from family-based genetic covariance structure modeling [234, 253] to the detection of individual gene effects in linkage/association studies (e.g., [26,131,159,184,260,268,345,374]). Given the presence of multivariate phenotypic data, the question arises under which conditions a multivariate analysis is preferable to univariate analyses in studying the role of a given genetic variant (GV).

In linkage analyses, multivariate modeling was considered both for substantive reasons and for statistical power advantages that multivariate data conferred (e.g., [17,18,37,39,41,113,164,235]). To date, population-based association studies have focused mainly on the relationship between a measured GV and a univariate phenotype. In the case of psychological phenotypes, this phenotype is often a sum score (i.e., the sum calculated across all items of a phenotypic instrument), or a case-control affection status dichotomy. In genetic association studies, however, the power advantages of multivariate data are also of interest, especially as the contributions of individual genetic variants to the phenotypic variance are commonly assumed to be small (Evans [114], Gordon and Finch [143]). To date, three studies have addressed the question of the power to detect GVs using multivariate data. In this paper, we briefly discuss these studies, and we contribute to this area by examining the power to detect a GV in genetic covariance structures based on the single and two common factor models and models for repeated measures.

Ferreira and Purcell [123] considered the power of a multivariate test (MANOVA) based on Wilk's Lambda given varying number of phenotypes (5, 10, and 20), of which a varying number were affected by the GV. They also varied the positive intercorrelations between the phenotypes. They found that the multivariate test was more powerful than univariate tests, with (1) increasing correlations among the phenotypes and (2) increasing number of phenotypes affected (i.e., by the GV) increasing the power. However, they noted a sharp loss of power of the multivariate test when all phenotypes were affected by the GV. This loss in power is exacerbated by increasing phenotypic correlations. Their results are consistent with previous results obtained in linkage analysis [17, 113, 124] , and with the statistical literature on MANOVA (Cole, Maxwell, Arvey, and Salas [71]).

Medland and Neale [242] considered the single factor models with 3 or 5 indicators, in unrelated cases and in sib pairs [131]. They varied the effect of the GV in the factor model such that it was (1) part of the common factor, thus conveying its effect via the factor loadings on all variables; or (2) common to all phenotypes, but not conveyed via the factor; or (3) present only in a single phenotype, or in some (but not all) phenotypes; or (4) it was present in some phenotypes, but with opposite effects. Medland and Neale [242] studied the power to detect the GV in the factor model, in which the GV affected all phenotypes via the factor (one degree of freedom (DF) test), or directly affected all phenotypes (a DF = 3 or DF = 5 test). They also considered the power conferred by the univariate tests based on simple sum scores and factor scores (Lawley and Maxwell, 1971) [194]. They varied other important aspects such as the magnitude of the factor loadings and the degree of missingness. Based on their figures 1a and 1b (Medland and Neale, [242], p. 237), the main conclusion is that their combined multivariate approach (where the GV effect is conveyed via the common factor, or the GV affects the phenotypes directly) was almost universally as powerful as, or, depending on specific circumstances, more powerful than, the univariate tests using sum scores or factor scores.

Van der Sluis et al [338] discussed the power to detect the effects of GVs in uni- and multidimensional common factor models. They contrasted the power in these model to the sum score model, in the situation that the sum score is not a sufficient statistic (i.e., the univariate sum score entails a loss of information relative to the multivariate data). They showed that the use of the sum score generally entails a loss of power, except in specific circumstances. In addition, they discussed how violations of measurement invariance across multiple samples, or with respect to the GV itself, affect the power to detect GVs. Violations of measurement invariance with respect to the GV itself (i.e., direct effects of the GV on one or more phenotypes in the model, instead of GV effects that are common to all phenotypes and mediated by (genetic) common factors) resulted in notable loss of power in the sum score model and incorrectly specified factor models.

The present aim is to contribute to this work on the power to detect genetic association in multivariate data. We discuss five models that one may encounter in family-based genetic covariance structure modeling of MZ and DZ twin data (Neale and Cardon, [253]): genetic factor models with single or multiple genetic factors underlying the covariance among a set of phenotypes, and two variations on the simplex models, which have been used to analyze repeated measures (Eaves, Long, and Heath, [109]; Boomsma and Molenaar, [42]). In each model, the effect of the GV is specified as part of an additive genetic factor, so that its effect on the phenotypes is mediated by the additive genetic factor. We consider situations in which the GV affects all phenotypes, and situations in which the effect is limited to a subset of the phenotypes. The single common factor model has been considered previously in the studies by Medland and Neale (2010) [242] and van der Sluis et al (in press) [338]. The power to detect a GV in the other four models has not been considered so far.

We simulated data according to a full multivariate twin model in the five scenarios. We established the power to detect the GV in this true model, and we studied the power in four statistical models using only the data of the first twin members, i.e., in genetically uninformative samples. In the following sections, we describe the five study designs and the simulation procedures in more detail. Next, we present the results, and we end the paper with a discussion.

2.2 Procedure

To calculate the power to detect the GV effect, we generated conditionally multivariate normal (i.e., conditional on the GV) MZ and DZ twin data according to the five models of interest. Next, we computed the power to detect the GV effect in the full MZ and DZ twin data, and in three statistical models in which we used only the twin 1 data (i.e., the phenotypic data and the measured GV): a univariate ANOVA in which the sum of the phenotypic measures was regressed on the GV, MANOVA in which all phenotypes were regressed on the GV, and exploratory factor analysis (EFA), in which the common factors are regressed on the GV. We simulated multivariate data according to a multivariate ACE twin model, in which A, C, and E represent the additive genetic structure, shared, and specific environmental influences, respectively. The additive genetic structure included one or more additive genetic factors. To one of these, we added a single diallelic codominant GV (minor allele frequency of .2), and defined its effect of .25% of the variance of a given phenotype, which loaded directly on the genetic factor. Depending on the chosen additive genetic factor structure, the GV did (directly or indirectly), or did not exert an influence on any other phenotype.

The first model that we considered included a single additive genetic factor. The single factor model was considered implicitly by Ferreira and Purcell [123][1], and explicitly by Medland and Neale [242] and van der Sluis et al. [338]. In the present study, the GV was specified as a source of variation in the genetic factor, and so this factor mediated the relationship between the GV and the phenotypes (see Figure 2.1 below). We include it because the single factor model – as specified below – is an ideal, and because the comparison of the MANOVA and the EFA has yet to be made. The second and third models included two correlated additive genetic factors. In the second model, the GV was part of the first genetic factor, but exerted no influence on the second factor or on its indicators. These indicators are thus uninformative with respect to the effect of the GV. In the third model, the second factor was regressed on the first genetic factor. This implied that the GV of the first factor did exert an influence on the second factor, and thus on its indicators. This model may represent a latent phenotype-endophenotype relationship, in which the effect of the GV on the phenotype is mediated by the endophenotype (de Geus and Boomsma [92]; De Geus, Wright,

[1] Ferreira and Purcell chose the intercorrelations among the phenotypes to be equal, which is consistent with a single factor model.

Martin, and Boomsma [92]). Finally, we considered two hybrid simplex-factor models for repeated measures. These models have been applied mainly in genetic covariance structure modeling of twin data (Neale and Cardon, [253]; for a linkage application, see Eaves et al, [109] and Birley, et al., [34]). We considered an ACE model with the additive genetic and environmental autoregressions, and a common shared environmental factor, and a stationary AE simplex model. In the latter, the common shared environmental factor is omitted, and background influences of A and E are stable over time. In the former, shared environmental effects decline, and the genetic effects increase. We considered 4 repeated measures, and calculated the power to detect the GV effect given that it entered the model at occasions 1, 2, 3, or 4. We consider this to be of interest, as genetic innovation variance is often attributed to the action of new genetic effects (Eaves, Long, and Heath, [109]; Gillespie, Evans, Wright, and Martin, [141]), which may include the effects of measured GVs. The simplex-factor model is similar to twin models established in analysis of IQ data in children, with a decreasing role of shared environment and increasing genetic influences (e.g., Hoekstra, Bartels, and Boomsma, [160]). The stationary AE simplex model is consistent with results one would expect in twin studies of IQ conducted in young adults. Further details on the simulation settings are given in the tables and path diagrams below.

Given these models, we varied (1) the number of phenotypic measures, and (2) the parameter values that accounted for genetic and environmental contributions to phenotypic variance. The parameter values are supposed to be typical of results one may obtain in genetic covariance structure modeling. We provide these details below. Throughout we used exact data simulation (van de Sluis et al, [336])[2]. We simulated the data using MVRNORM in R (R-core development team, [316])[3], under the assumptions that mating is random, and the GV is in Hardy-Weinberg equilibrium. Given the diallelic GV, the total MZ and DZ sample sizes were distributed over 3 MZ groups (three pairs of identical genotypes) and 9 DZ groups (3 genotypes × 3 genotypes). The distribution of the total sample size over these groups depends on the minor allele frequency, which we set to equal .2 in all studies.

We first computed the power to detect the GV effect in (A) the full multivariate twin model. We calculated the power both in the model specified correctly with respect to the role of the GV (i.e., a 1 DF test), and in the model in which all present common genetic factors were regressed on the GV, i.e., an omnibus test, with DF equaling the number of genetic factors in the model. We added the power of the omnibus test because in practice one will not know the exact

[2] In exact data simulation, the simulated data fit the true model exactly, and lent themselves to power calculations as the likelihood ratio of the models with and without the GV effect equals the noncentrality parameter of the noncentral χ^2 distribution required to calculate the power.

[3] This is part of the MASS library. MVRNORM includes the facility for exact data simulation.

locus of the GV, and therefore will resort to the omnibus test. As mentioned, we did not consider the possibility that the GV affects a single phenotype (Medland and Neale, [242] did consider this possibility), we therefore limited our omnibus test to the common genetic factors. In the twin 1 phenotypic and GV data, we calculated the power in: (B) a univariate ANOVA, in which each univariate phenotype was regressed on the GV; (C) a univariate ANOVA in which the sum of the phenotypic measures was regressed on the GV; (D) MANOVA in which all phenotypes were regressed on the GV; and (E) an exploratory factor analysis (EFA), in which the phenotypic common factors were regressed on the GV. We fitted standard MANOVAs, subject to homogeneity of the conditional (i.e., on the GV) covariance matrices. We did not constrain these in the light of our information concerning the covariance structure. In specifying the EFAs, we did exploit this information to the extent that the specified dimensionality of the exploratory factor solution is consistent with the true model. We did not fit the exploratory factor model to the repeated measures data, as the autoregressive covariance structures are not compatible with an exploratory factor model (e.g., Mandys, Dolan, and Molenaar, [232]). As we considered only additive genetic effects, we included the GV as a covariate (rather than as a between-subject factor) in the analyses. Analyses A and E were done in MX (Neale et al., [254]), analyses B to D were done in R. We report the power of the tests of models B to E for N = 3000, and the power of the full twin model for NMZ = 1500 and NDZ = 1500, all given an α level of .01. In the case of the single phenotype ANOVA, we also report the power for the Bonferroni corrected alpha (i.e., .01 divided by the number of phenotypes). This correction is conservative, but the differences in power between the single phenotype test and the other tests are such that the choice of correction is unlikely to have any bearing on the conclusions. We note that a resample procedure such as permutation testing is unsuited as the data simulation is exact. The alpha of .01 is unrealistic given multiple testing. However, here we were interested solely in the differences between the tests in power, not in the absolute values. However, we report the non-centrality parameters (NCPs), so that the power of the tests of association can be computed for other total sample sizes and other α levels, if the reader so desires. R scripts that can be used to this end are provided in the Appendix. We report the power in the full twin model, as our simulation and testing procedure produces this result. However our main interest is in the sum score ANOVAs, MANOVAs, and EFAs. The comparison of the power in the full twin model with the power of the other tests is complicated by (1) the difference in number of individuals (a twin comprises two individuals), and (2) the differences in the expense of ascertainment (ascertaining twin pairs is usually more expensive than ascertainment of unrelated individuals). In the subsequent sections, we present the five studies in detail.

Table 2.1: Variance components in the 4 scenarios that were used to generate the data. The total variance of each phenotype, conditional on the GV, equaled one. We provide only 4 parameter values in each scenario, as we did not vary these parameter values over the phenotypes. For instance in scenario S3, conditional on the GV, 4 (or 8) tests loaded on the common A factor with loadings equal to $\sqrt{.5}$, the genetic residual is .1. The loadings on the common shared environmental factor equaled $\sqrt{.2}$. The unshared environmental residuals equaled .2. Therefore in scenario S3, the decomposition of phenotypic variance conditional on GV is $h^2 = .6$, $c^2 = .2$, and $e^2 = .2$.

Phenotypic correlations	Nr. of phenotypes	Scenario	Common A	Specific a_i	Common C	Specific e_i
.5	4/8	S1	.5	.1	0	.4
.2	4/8	S2	.2	.1	0	.7
.7	4/8	S3	.5	.1	.2	.2
.4	4/8	S4	.2	.1	.2	.5

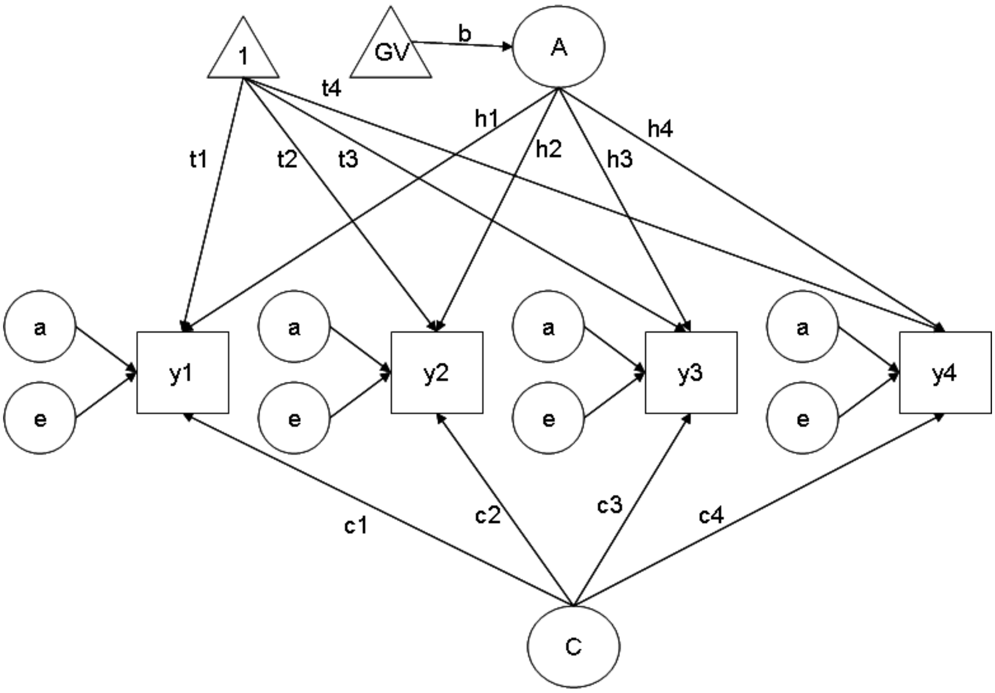
2.3 Study 1: Single common genetic factor

The objective of the first simulation study is to examine the power to detect a GV that affects all phenotypes via a common polygenic factor. Specifically, we examined how the sources of phenotypic correlations and the number of measured phenotypes affect the power to detect the GV effect. In this study we supposed that a single common genetic factor or, a common genetic factor and a shared environmental factor, account for the phenotypic correlations.

We simulated MZ and DZ phenotypic data which generate precisely the means and variances predicted by the common factor model shown in Figure 2.1. We specified either 4 (as depicted) or 8 phenotypes loading on the additive polygenic factor (A) and a shared environmental factor (C). Additional parameters are the phenotype-specific genetic (a_i) and unique environmental (e_i) factors. We added the GV to the common genetic factor (A), which thus affected all phenotypes (y_i). The GV accounted for .25% of the variance in the first phenotype (y_1). The chosen parameter values are given in Table 2.1. As we did not vary the parameters over the phenotypes, the effect size of .25% also holds with respect to the other phenotypes.

We simulated twin data, given eight scenarios in which we varied the role of the common factor A and C, and the specific environmental effect, as shown in Table 2.1. The heritability of the phenotypes ranged from $h^2 = .3$ (S1 and S3) to $h^2 = .6$ (S2 and S4). The influence of the common C was absent in scenarios S1 and S2, and present in scenarios S3 and S4 ($c^2 = .2$). As shown in Table 2.1,

Figure 2.1: Path diagram for the common factor model with 4 phenotypes. The triangles represent fixed regressors (i.e., the GV and the unit vector). The parameters t1 to t4 are intercepts, the parameter **b** is the effect of the GV on the common genetic factor. The GV enters the model via common genetic factor A and affects the indicators y_1 to y_4.



the implied correlations among the phenotypes were .5, .2, .7, and .4 in scenario S1, S2, S3 and S4, respectively. Table 2.2 contains the results.

Table 2.2 shows that in the single common factor model, the ANOVA of sum scores has the same power as the exploratory factor model. Due to the equality (over the phenotypes) of factor loadings and residual variances, the factor scores and the sum scores are perfectly correlated. Note that in Table 2.2, the information with respect to the single phenotype ANOVA is redundant because the number of phenotypes simulated is irrelevant in the analysis of a single phenotype. We included the power of the single phenotype ANOVA to ease comparison, and because the power associated with the Bonferroni corrected α varies as a function of the number of tests (4 vs. 8).

The single phenotype ANOVA with Bonferroni corrected alpha consistently has lowest power. The NCPs of the sum score ANOVA, the MANOVA and the EFA are comparable, and thus affected similarly by the differences in parameter configuration. The lower power of the MANOVA compared to the EFA stems from the differences in DF of the associated tests. Increasing the number of indicators resulted in a consistent increase in power of the sum score ANOVA

Table 2.2: The power, non-centrality parameter, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 1. In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/4$ (.0025; 4 phenotypes) or $\alpha = .01/8$ (.00125; 8 phenotypes). The power for the corrected alpha is displayed in italics.

Scenario	Nr. of phenotypes	True model	ANOVA Sum Scores	ANOVA single phenotype	MANOVA	EFA
N		2×1500	3000	3000	3000	3000
S1	4	.95 18.04 (1)	.81 12.02 (1,2998)	.56, .39 7.51 (1,2998)	.59 12.01 (4,2995)	.81 12.00 (1)
	8	.97 19.97 (1)	.85 13.35 (1,2998)	.56, .31 7.51 (1,2998)	.51 13.32 (8,2991)	.85 13.33 (1)
S2	4	.99 29.47 (1)	.96 18.78 (1,2998)	.56, .39 7.51 (1,2998)	.85 18.76 (4,2995)	.96 18.73 (1)
	8	.99 38.09 (1)	.99 25.04 (1,2998)	.56, .31 7.51 (1,2998)	.89 24.98 (8,2991)	.99 24.95 (1)
S3	4	.91 15.55 (1)	.70 9.69 (1,2998)	.56, .39 7.51 (1,2998)	.46 9.68 (4,2995)	.70 9.68 (1)
	8	.93 16.56 (1)	.73 10.18 (1,2998)	.56, .31 7.51 (1,2998)	.35 10.16 (8,2991)	.73 10.17 (1)
S4	4	.98 22.11 (1)	.86 13.66 (1,2998)	.56, .39 7.51 (1,2998)	.67 13.64 (4,2995)	.86 13.63 (1)
	8	.99 26.75 (1)	.91 15.81 (1,2998)	.56, .31 7.51 (1,2998)	.62 15.78 (8,2991)	.91 15.78 (1)

and EFA, as is to be expected as the increase in affected phenotypes in the one-dimensional model increases the GV signal. However, the increase in the number of phenotypes resulted in a decrease in power of the MANOVA in three cases, and slight increase in only scenario S2 (power .85 vs. .89). Overall the power of the MANOVA, sum score ANOVA, and EFA decreases with increasing phenotypic correlation (e.g., compare S1 and S3). Increasing the correlations increases the variance of the sum scores, and given the constant effect size, lowers the power of the test. Cole et.al [71] explained the role of the magnitude of phenotypic correlation on the power in the MANOVA given consistent effects on the dependent variables. Specifically they showed in two dimensions that the overlap between the 95% ellipsoids increases with increasing correlation (see Cole at al, 1994, Figure 1). This results in a loss of power (see also Ferreira and Purcell, 2009).

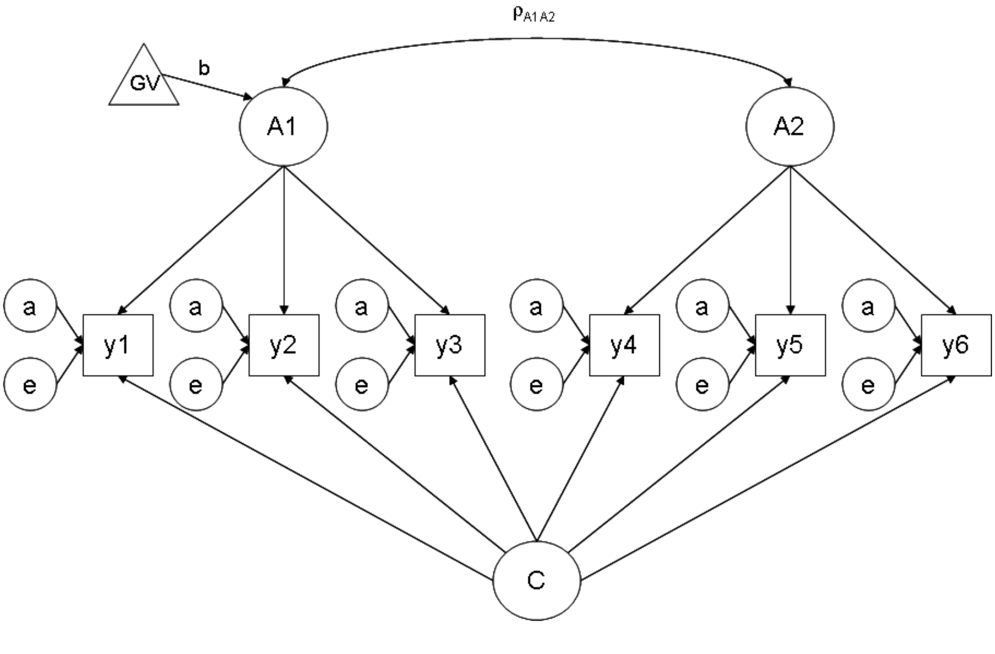
In conclusion, in this study, the methods of choice are the EFA or the sum score ANOVA. The power of these methods is equal because the factor loadings in the EFA are equal. We refer to Medland and Neale [242] and van der Sluis, et al. [338] for results obtained in the same setup, but with unequal loadings. The MANOVA fares relatively poorly because all phenotypes are affected by the GV and the phenotypic correlations are positive and relatively high (notably in scenario S3). The power of the EFA is relatively good because the test involves a single parameter, i.e., the latent mean difference between the genotypes on the common factor. The effect of the GV on the actual phenotypes is thus mediated by the common factor. Medland and Neale [242] also considered the EFA in which the GV has a direct effect on the phenotype. In that case, the NCP would be the same as shown in Table 2.2 for the EFA, but the DFs would equal 4 or 8 (i.e., the number of phenotypes). The power of this EFA based test would then equal that of the MANOVA.

The power of the full twin model was consistently high (>.90), but the NCPs display good variation (ranging from 15.55 to 38.09). For instance, retaining the sample sizes of 2×1500, but changing the alpha from $1E^{-2}$ to $1E^{-7}$, reduced the power to a low of .083 (S3, 4 phenotypes) and to a high of .801 (S2, 8 phenotypes).

2.4 Study 2: Correlated genetic common factors

In the second study, the model included two correlated genetic factors, of which only the first is affected by the GV. We examined how the sources of phenotypic correlations, i.e., the genetic correlation and the shared environmental factor, affect the power to detect the GV effect. In addition, we explored the impact of the number of phenotypic indicators (3 vs. 5 per factor) on the power. Figure 2.2 depicts the three indicator model. The covariances among the phenotypes are caused by two genetic correlated factors (A1 and A2) and a shared environmental factor (C). Additional parameters in the model are genetic specifics (a_i) and unshared environmental effects (e_i). The GV enters the model via the latent genetic factor A1, and so affects the indicators of A1, but affects neither A2, nor its indicators. The parameter values used to generate data for this study are given in Table 2.3. The GV explained .25% of the variance of the phenotype y_1. Given the parameter values, the GV explained the same amount of variance in the other indicators for the first genetic factor, but no variance in the indicators of the second genetic factor.

Figure 2.2: Path diagram of the oblique two common factor model (three indicator model). The triangle represents the GV as a fixed regressor. The unit vector, which is used to estimate intercepts is not included to avoid clutter (see Figure 2.1). The parameter **b** represents the effect of the GV. Note that the GV contributes to the variance of the first latent genetic factor A1 and affects its indicators (y_1-y_3), but does not affect the second common factor A2, or its indicators (y_4-y_6). The value of the correlation between A1 and A2 was varied. Parameters are not shown to avoid clutter.



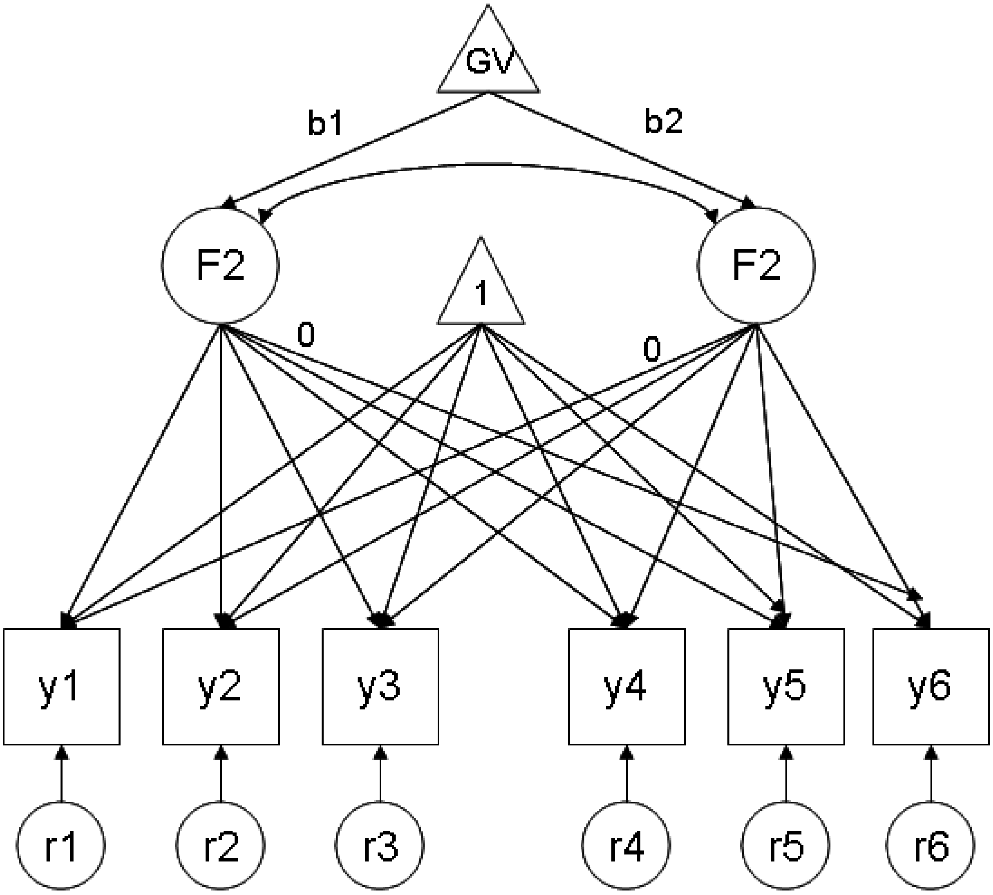
As we manipulated the correlations between the genetic factors (3 settings), the number of phenotypes per genetic factor (2 settings), and the parameter values (3 settings), we simulated data according to 18 scenarios. As above, we computed the power to detect the GV in the true multivariate twin model, in the univariate phenotype ANOVA, in the sum score ANOVA, in the MANOVA, and in the two factor oblique EFA. In fitting the EFA in MX, we identified the model by fixing the loading of the first phenotypic variable on the second factor, and the last phenotypic variable on the first factor to zero. All other loadings

Table 2.3: Variance components, conditional on GV, used to simulate data in study 2 (correlated genetic factors) and study 3 (regression of genetic factor 2 on genetic factor 1). The within (between) set phenotypic correlation is among phenotypes that load on the same (different) genetic factor (factors). For instance, in the 3 indicator S22 scenario, the phenotypes y_1 to y_3 (y_4 to y_6) loaded $\sqrt{.3}$ on the common A1 (A2) factor, each phenotype loading $\sqrt{.1}$ on the common C factor. The residual variance of each phenotype equaled .6 (.5 due to specific environment; .1 due to specific genes). So in scenario S22, the decomposition of phenotypic variance conditional on GV is $h^2 = .4$, $c^2 = .1$, and $e^2 = .5$.
* Within set correlation is among phenotypes that load on the same genetic factor (y_1-y_2), the between set correlation is among phenotypes that load on the different genetic factors (y_1-y_6).

Correlation coefficient	Scenario	Phenot. cor. within sets, between sets*	Nr. Ind.	Common A1, A2	Specific a_i	Common C	Specific e_i
$\rho_{A1A2} = .77$	S11	.8, .70	3/5	.3	.1	.5	.1
	S12	.4, .33	3/5	.3	.1	.1	.5
	S13	.3, .23	3/5	.3	.1	.0	.6
$\rho_{A1A2} = .47$	S21	.8, .64	3/5	.3	.1	.5	.1
	S22	.4, .24	3/5	.3	.1	.1	.5
	S23	.3, .14	3/5	.3	.1	.0	.6
$\rho_{A1A2} = .25$	S31	.8, .57	3/5	.3	.1	.5	.1
	S32	.4, .17	3/5	.3	.1	.1	.5
	S33	.3, .07	3/5	.3	.1	.0	.6

were estimated. The common factors were standardized and allowed to correlate. For the path diagram, see Figure 2.3. Other identifying constraints, but these constraints in the EFA does not affect the power of the omnibus test, in which all phenotypic common factors are regressed on the GV (McDonald, [240]; Dolan, et al., [104]). In the EFA, we regressed both common factors on the GV (a 2 DF test), that is, we did not exploit our knowledge of the locus of the GV in the model. We varied the number of indicators, the size of the genetic correlations, and the contribution of the common C factor. Table 2.4 contains results.

Figure 2.3: Exploratory (oblique) two common factor model as used in studies 2 and 3. Two factor loadings are fixed to zero (as depicted) to achieve rotational determinacy. The common factors are denoted F1 and F2, r1 to r6 represent the residuals. The triangles represent fixed regressors. The regression on the unit vector serves to estimate the intercepts, the regression on the GV estimates the effect of the GV (i.e., the parameters **b**1 and **b**2). Other parameters are not shown to avoid clutter.



The power of the sum score ANOVA is low, as expected because the phenotypes that are unaffected by the GV add only noise to the sum score. The single

Table 2.4: Power, non-centrality parameter, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 2. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which the 2 genetic factors are regressed on the GV (2 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/6$ (.0016; 6 phenotypes) and $\alpha = .01/10$ (.001; 10 phenotypes). The power for the corrected alpha is displayed in italics.

ρ_{A1A2}	Scenario	Nr. of indicators	True model	ANOVA Sum Scores	ANOVA single phenotype	MANOVA	EFA
N			2 × 1500	3000	3000	3000	3000
.77	S11	3	>.99,>.99 47.29 (1),(2)	.14 2.35 (1,2998)	.56, *.34* 7.51 (1,2998)	.97 30.23 (6,2993)	>.99 30.15 (2)
		5	>.99,>.99 57.40 (1),(2)	.15 2.38 (1,2998)	.56, *.29* 7.51 (1,2998)	.98 37.14 (10,2989)	>.99 37.02 (2)
	S12	3	>.99,>.99 29.57 (1),(2)	.28 4.03 (1,2998)	.56, *.34* 7.51 (1,2998)	.77 18.02 (6,2993)	.91 18.01 (2)
		5	>.99,>.99 39.06 (1),(2)	.31 4.40 (1,2998)	.56, *.29* 7.51 (1,2998)	.85 24.35 (10,2989)	.97 24.32 (2)
	S13	3	>.99,>.99 28.48 (1),(2)	.35 4.90 (1,2998)	.56, *.34* 7.51 (1,2998)	.74 17.35 (6,2993)	.89 17.34 (2)
		5	>.99,>.99 37.83 (1),(2)	.41 5.58 (1,2998)	.56, *.29* 7.51 (1,2998)	.83 23.60 (10,2989)	.97 23.58 (2)

Continued in Table 2.5.

Table 2.5: *Continued from Table 2.4.*

ρ_{A1A2}	Scenario	Nr. of indicators	True model	ANOVA Sum Scores	ANOVA single phenotype	MANOVA	EFA
N			2 × 1500	3000	3000	3000	3000
.47	S11	3	>.99,>.99 29.98 (1),(2)	.16 2.50 (1,2998)	.56, *.34* 7.51 (1,2998)	.79 18.62 (6,2993)	.92 18.61 (2)
		5	>.99,>.99 34.45 (1),(2)	.16 2.52 (1,2998)	.56, *.29* 7.51 (1,2998)	.78 21.55 (10,2989)	.95 21.54 (2)
	S12	3	.99,.97 23.25 (1),(2)	.32 4.50 (1,2998)	.56, *.34* 7.51 (1,2998)	.64 14.74 (6,2993)	.82 14.74 (2)
		5	>.99,>.99 28.07 (1)	.36 4.98 (1,2998)	.56, *.29* 7.51 (1,2998)	.66 18.07 (10,2989)	.91 18.07 (2)
	S13	3	.99,.97 23.61 (1),(2)	.41 5.62 (1,2998)	.56, *.34* 7.51 (1,2998)	.65 15.01 (6,2993)	.83 15.01 (2)
		5	>.99,>.99 28.86 (1),(2)	.49 6.54 (1,2998)	.56, *.29* 7.51 (1,2998)	.69 18.77 (10,2989)	.92 18.77 (2)
.25	S11	3	.99,.98 25.72 (1),(2)	.16 2.60 (1,2998)	.56, *.34* 7.51 (1,2998)	.68 15.56 (6,2993)	.85 15.55 (2)
		5	>.99,>.99 28.00 (1),(2)	.17 2.64 (1,2998)	.56, *.29* 7.51 (1,2998)	.61 16.91 (10,2989)	.88 16.92 (2)
	S12	3	.98,.96 21.42 (1),(2)	.35 4.83 (1,2998)	.56, *.34* 7.51 (1,2998)	.59 13.70 (6,2993)	.79 13.70 (2)
		5	>.99,.98 25.25 (1),(2)	.39 5.38 (1,2998)	.56, *.29* 7.51 (1,2998)	.59 16.31 (10,2989)	.87 16.31 (2)
	S13	3	.98,.96 22.37 (1),(2)	.46 6.15 (1,2998)	.56, *.34* 7.51 (1,2998)	.62 14.36 (6,2993)	.81 14.36 (2)
		5	>.99,.99 26.79 (1),(2)	.54 7.26 (1,2998)	.56, *.29* 7.51 (1,2998)	.64 17.57 (10,2989)	.90 17.58 (2)

(affected) phenotype ANOVA is more powerful (i.e., power based on the corrected alpha) than the sum score ANOVA when the phenotype intercorrelations were relatively large (e.g., S11). The NCPs of the MANOVA and the EFA are comparable, and affected similarly by the variation in parameters. However, as in study 1, the EFA has greater power due to the difference in DF of the associated tests. In comparison with study 1, the MANOVA fares relatively well, because the GV does not affect all the phenotypes (as in study 1; see also Ferreira and Purcell, [123]). Note that in this case (in contrast to study 1), the increase in the phenotypic correlation resulted in an increase in power (compare S11 and S13, or S11 and S31). The presence of phenotypes not affected by the GV has a beneficial effect in MANOVA, especially when the correlations among the phenotypes are relatively high (see also Ferreira and Purcell, [123]). Cole et al. [71] explained the role of the magnitude of phenotypic correlation on the power in the context of MANOVA, when some, but not all, dependent variables are affected (see Cole at al, [71], Figure 3). In general, power of all tests improved by increasing the number of phenotypic indicators (from 3 to 5). In conclusion, in this study, the methods of choice are the EFA or the MANOVA. The power of the EFA is relatively good because the test involves just two parameters, i.e., two latent mean differences. As in study 1, the effect of the GV on the phenotypes is mediated by the 2 common factors. Estimating the effect of the GV directly on the phenotypes in the EFA (a 6 or 10 DF test) would render the power of the EFA equal that of the MANOVA.

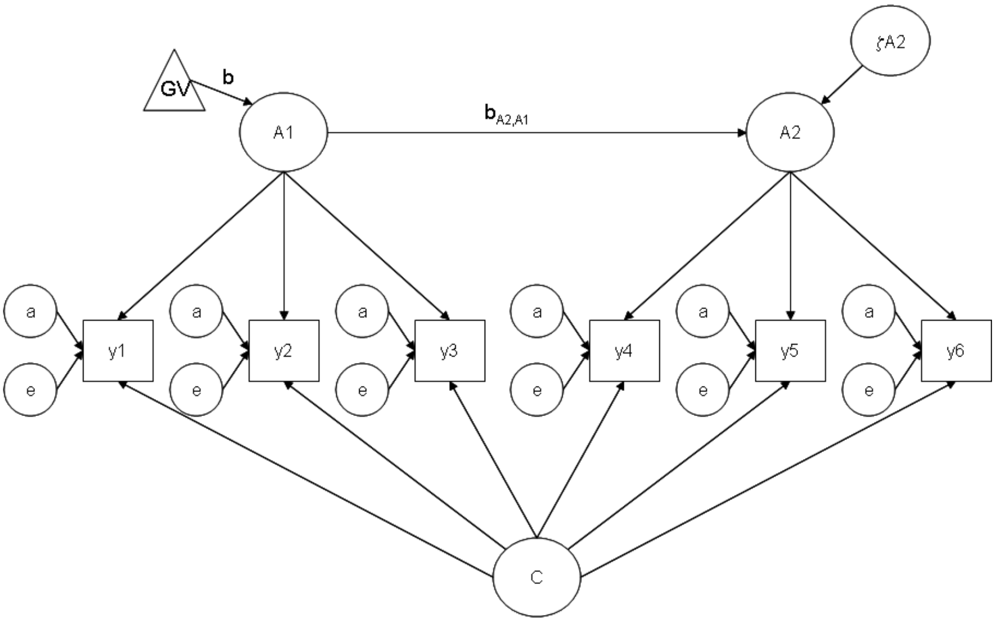
NMZ = 1500 and NDZ = 1500 afforded high power in the full twin model. But again the NCPs are quite variable. Changing the alpha from 1E^{-2} to 1E^{-7} reduced the power of the 1 DF test to a low of .24 (scenario S32, 2×3 phenotypes) and to a high of >.99 (scenario S11, 2×5 phenotypes).

2.5 Study 3: Latent regression model

In the third study, we specified a latent regression model with an independent (A1) and dependent common genetic factor (A2). The GV is introduced into A1, and exerts its influence both on the indicators of A1 (i.e., via A1), and on A2 and on its indicators. We included a common environmental factor, and varied the details of both the shared and unshared genetic and environmental effects. As in study 2, we considered both 3 and 5 indicators models. We simulated phenotypic data according to the model, as shown in Figure 2.4 (i.e., the three indicator model).

We chose parameter values such that the resulting correlations between the factors A1 and A2 equal the correlations of study 2. The other parameters in the model are additive genetic specifics (a_i) and unique environmental effects (e_i), which contribute to phenotypic variance. The GV effect is defined with respect to the first phenotype y_1, but the GV explained the same amount of variance (.25%)

Figure 2.4: Path diagram for the latent genetic regression model (3 indicator model). The triangle represents the GV as fixed regressor. The unit vector used to estimate intercepts is not included to avoid clutter (see Figure 2.1). The parameter **b** represents the effect of the GV. Note that the GV contributes to the variance of the first latent genetic factor A1 and affects its indicators (y_1-y_3). The GV contributes to A2 via the regression coefficient $\mathbf{b}_{A2A1}$, and so also affects the indicators y_4-y_6. The value of the parameter $\mathbf{b}_{A2A1}$ was varied. Parameters are not shown to avoid clutter.



in the other indicators of the first common genetic factor. Given $\rho_{A1A2} = .77$, the GV accounted for about .15% of the variance in the indicators of A2, the dependent genetic factor. The parameter values used in study 3 equaled those of study 2, and are shown in Table 2.3. As we manipulated the regressions between the genetic factors (3 settings), the number of phenotypes per genetic factor (2 settings), and the parameter values (3 settings), we simulated data according to 18 scenarios. Table 2.6 contains the results.

The present study resembles study 1 in that the effect of the GV is general. However, here the GV effect varied (e.g., .25% vs. .15% in S11), as did the intercorrelations among the phenotypes (see Table 2.3). Compared to study 2, the sum score fares well, especially when the phenotypic intercorrelations are relatively low, and the regression relationship of A2 and A1 is relatively strong: in scenarios S11, S12, and S13, the sum score ANOVA has the greatest power. However, given a weaker regression relationship the power of the EFA is greater than the power of the sum score ANOVA. The power of the MANOVA depends on (1) the differences in the GV effect on the phenotypes (general large effects in

Table 2.6: Power, non-centrality parameter, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 3. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which the 2 genetic factors are regressed on the GV (2 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/6$ (.0016; 6 phenotypes) and $\alpha = .01/10$ (.001; 10 phenotypes). The power for the corrected alpha is displayed in italics.

β_{A1A2}	Scenario	Nr. of indicators	True model	ANOVA Sum Scores	ANOVA single phenotype	MANOVA	EFA
N			2×1500	3000	3000	3000	3000
.77	S11	3	.95,.91 18.29 (1),(2)	.55 7.39 (1,2998)	.56,*.32*; .42,*.21* 7.51,4.50 (1,2998)	.33 8.80 (6,2993)	.54 8.81 (2)
		5	.96,.93 19.47 (1),(2)	.56 7.51 (1.2998)	.56,*.32*; .36,*.17* 7.51,4.50 (1,2998)	.27 9.27 (10,2989)	.57 9.29 (2)
	S12	3	.97,.95 20.97 (1),(2)	.83 12.68 (1,2998)	.56,*.32*; .42,*.21* 7.51,4.50 (1,2998)	.58 13.37 (6,2993)	.78 13.37 (2)
		5	.98,.97 23.46 (1),(2)	.87 13.87 (1,2998)	.56,*.32*; .36,*.17* 7.51,4.50 (1,2998)	.53 14.85 (10,2989)	.83 14.86 (2)
	S13	3	.99,.98 24.49 (1),(2)	.91 15.44 (1,2998)	.56,*.32*; .42,*.21* 7.51,4.50 (1,2998)	.70 16.05 (6,2993)	.86 16.05 (2)
		5	>.99,.99 27.86 (1),(2)	.94 17.58 (1,2998)	.56,*.32*; .36,*.17* 7.51,4.50 (1,2998)	.67 18.46 (10,2989)	.91 18.46 (2)

Continued in Table 2.7.

Table 2.7: *Continued from Table 2.6.*

β_{A1A2}	Scenario	Nr. of indicators	True model	ANOVA Sum Scores	ANOVA single phenotype	MANOVA	EFA
N			2 × 1500	3000	3000	3000	3000
.47	S11	3	.96,.93 19.55 (1),(2)	.38 5.24 (1,2998)	.56,*.09*; .42,*.04* 7.51,1.50 (1,2998)	.41 10.16 (6,2993)	.62 10.17 (2)
		5	.98,.95 21.09 (1),(2)	.39 5.33 (1,2998)	.56,*.09*; .36,.03 7.51,1.50 (1,2998)	.34 10.88 (10,2989)	.66 10.90 (2)
	S12	3	.96,.93 19.73 (1),(2)	.68 9.43 (1,2998)	.56,*.09*; .42,*.04* 7.51 & 1.50 (1,2998)	.54 12.55 (6,2993)	.74 12.55 (2)
		5	.98,.96 22.70 (1),(2)	.74 10.43 (1,2998)	.56,*.09*; .36,.03 7.51,1.50 (1,2998)	.51 14.40 (10,2989)	.81 14.41 (2)
	S13	3	.98,.96 22.57 (1),(2)	.80 11.78 (1,2998)	.56,*.09*; .42,*.04* 7.51,1.50 (1,2998)	.64 14.63 (6,2993)	.82 14.63 (2)
		5	>.99,.98 26.41 (1),(2)	.86 13.69 (1,2998)	.56,*.09*; .36,*.03* 7.51,1.50 (1,2998)	.63 17.40 (10,2989)	.89 17.41 (2)
.25	S11	3	.97,.95 20.66 (1),(2)	.29 4.11 (1,2998)	.56,.03; .42,*.013* 7.51,.49 (1,2998)	.47 11.23 (6,2993)	.68 11.24 (2)
		5	.98,.96 22.29 (1),(2)	.29 4.19 (1,2998)	.56,*.03*; .36,*.008* 7.51,.49 (1,2998)	.39 12.02 (10,2989)	.72 12.03 (2)
	S12	3	.96,.93 19.64 (1),(2)	.57 7.63 (1,2998)	.56,*.03*; .42,*.013* 7.51,.49 (1,2998)	.53 12.52 (6,2993)	.74 12.52 (2)
		5	.98,.97 22.76 (1),(2)	.63 8.51 (1,2998)	.56,*.03*; .36,*.008* 7.51,.49 (1,2998)	.51 14.52 (10,2989)	.82 14.52 (2)
	S13	3	.99,.96 22.05 (1),(2)	.70 9.71 (1,2998)	.56,*.03*; .42,*.013* 7.51,.49 (1,2998)	.62 14.24 (6,2993)	.81 14.24 (2)
		5	>.99,.98 26.04 (1),(2)	.79 11.46 (1,2998)	.56,*.03*; .36,*.008* 7.51,.49 (1,2998)	.62 17.15 (10,2989)	.89 17.15 (2)

S11, S12, and S13 in contrast to S31, S32, and S33), and (2) the intercorrelations among the tests (generally low in S13, S23, and S33; generally high in S11, S21, and S31) (see, Cole et al., 1994). The greatest power is observed in S13 (6 phenotypes), i.e., a general effect, but low phenotypic intercorrelations (.70). The lowest power is in S11 (10 phenotypes), i.e., general effects and high phenotypic correlations (.27). In this scenario, the single phenotype ANOVA happens to be more powerful (.36). The NCP of the MANOVA equals that of the EFA, so it is again the difference in DF that determine the difference in power. Conducting the EFA with GV effect directly on the phenotypes (rather than being mediated by the common factors) would render the power of the EFA equal to that of the MANOVA.

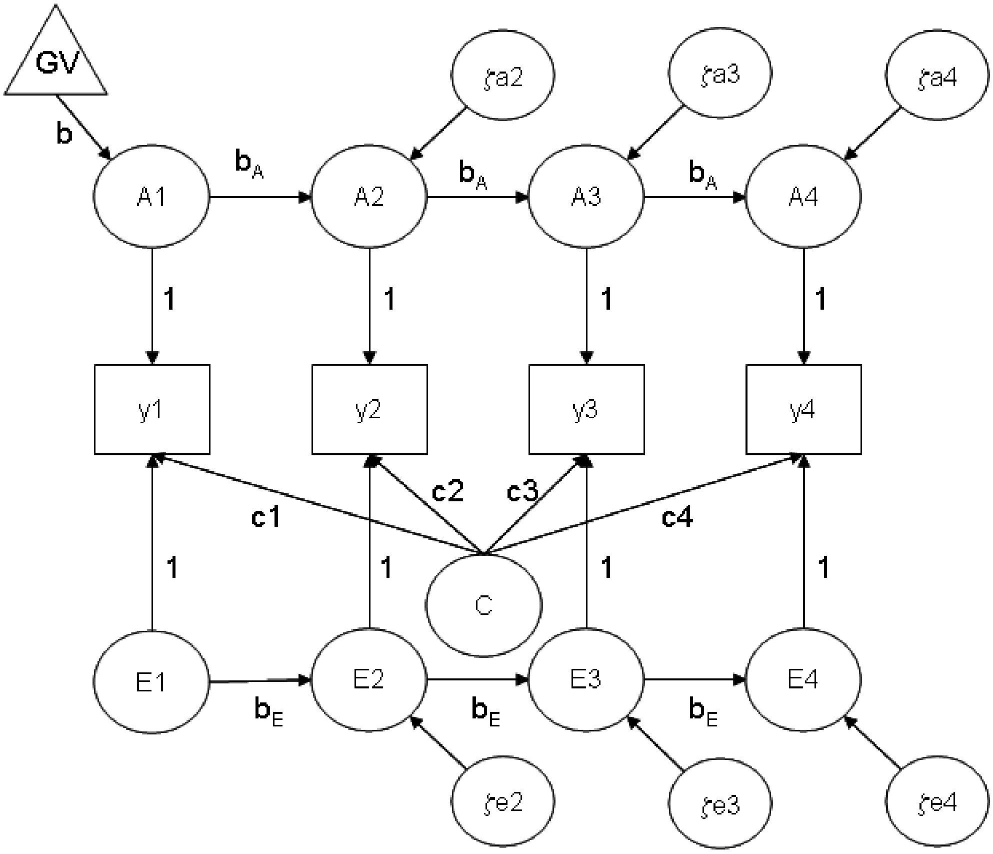
The power of the 1 DF test in the full twin model is high (>.96). Changing the alpha from $1E^{-2}$ to $1E^{-7}$, reduced the power to a low of .15 (S11, 3 indicators) and to a high of .65 (S13, 5 indicators).

2.6 Study 4: Hybrid simplex (A,E)-factor (C) model

In the fourth study, we considered a hybrid factor-simplex model for four occasions. We varied the occasion (t) at which the GV entered the model as part of the genetic factor (A(t)). In this model, which is shown in Figure 2.5, the phenotype y(t) was regressed on a latent genetic factor A(t), environmental influences common to all phenotypes C(t), and specific environmental influences E(t): y(t) = A(t) + C(t) + E(t). The stability of the phenotypic individual differences depended on the common shared environmental factor, and on the autoregressive coefficients in the genetic and unshared environmental simplexes, i.e., β_A and β_E. The parameters β_A and β_E equal 1 and .7, respectively. The other parameters in the model are the residual variances, $\sigma^2_{\zeta_A}$ ($\sigma^2_{\zeta_A}$ =.1) and $\sigma^2_{\zeta_E}$ ($\sigma^2_{\zeta_E}$=.204), representing the amount of variance in the genetic factors A(t) that is not explained by the independent factors A(t-1).

The GV was added to the genetic factor (A) at occasion t, and its effect is defined as .25% of the variance in the phenotype y(t) that depends directly on A(t). Because of the genetic autoregression, the GV effect entering at occasion t is transmitted to the phenotypes measured at the subsequent occasions. For example in Figure 2.5, the GV also affects y_2, y_3 and y_4. The parameter values in the model are given in Table 2.8, along with the expected phenotypic covariances. As shown in Table 2.8, the variance due to the common shared environmental factor decreased over time (c^2 decreases from .3 to 0), while the variance due to the genetic factor increased through time (h^2 increases from .3 to .6). We calculated the power in the full multivariate twin model, in the univariate ANOVA of the sum scores, i.e., the sum of the phenotypes observed at the different occasions; in the univariate ANOVA of each individual phenotype, and in the MANOVA. As mentioned above, we did not fit the exploratory factor model on these longi-

Figure 2.5: Path diagram for the hybrid simplex-factor model. The triangle represents the GV as fixed regressor. The unit vector used to estimate intercepts is not included to avoid clutter (see Figure 2.1). In this model the GV enters at occasion 1. We also considered the cases in which the GV enters at occasions 2, 3, or 4.



tudinal data as the autoregressive covariance structure is not compatible with an exploratory factor model. Table 2.9 contains the results.

The results are consistent with the results of the preceding studies. First, when the GV affected all phenotypes (enters at occasion 1), the ANOVA of sum scores was the most powerful test of association (.71). Its power decreased from .71 to .03, as the GV entered the model at a progressively later occasion. This is expected as the GV signal in the sum score is weakened by the presence of unaffected phenotypes. The power of MANOVA followed a reverse pattern: it was the lowest when all phenotypes were associated with the GV (.51). Given the relatively large phenotypic correlations, this is consistent with Cole at al. ([71]; Figure 1; see also Ferreira and Purcell, [123]), and with the results of study 1 and study 3. The power is high when the GV entered at a later occasion, ranging from .95 (occasion 2) to .88 (occasion 4). The power of the single (affected) phenotype

Table 2.8: Variance components in study 4 at the 4 occasions, and the implied phenotypic covariance matrix, conditional on GV. The model comprises a simplex for A and E, and a common C factor. The factor loadings on the common C factor are $\sqrt{.3}$, $\sqrt{.2}$, $\sqrt{.1}$, and 0. The C factor loading decreases, the additive genetic variance increases, and the unshared environmental variance remains constant. Consequently the total phenotypic variance, conditional on the GV, remains 1 at each occasion. The autoregressive parameters β_A and β_E equal 1 and .7, respectively. The residual variances equal $\sigma^2_{\zeta_A} = .1$ and $\sigma^2_{\zeta_E} = .204$.

occasion	h^2	c^2	e^2
t1	.3	.3	.4
t2	.4	.2	.4
t3	.5	.1	.4
t4	.6	.0	.4
phenotypic covariance matrix			
1			
.825	1		
.669	.821	1	
.437	.596	.780	1

Table 2.9: Power, non-centrality parameters, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 4. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which all 4 genetic factors are regressed on the GV (4 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/4$ (.0025; 4 phenotypes). The power for the corrected alpha is displayed in italics.

Model	True model	ANOVA Sum scores	ANOVA Single phenotypes at 4 occasions	MANOVA
	2×1500	3000	3000	3000
GV at t1	.93,.79 16.48 (1),(4)	.71 9.80 (1,2998)	.56 & .56 & .56 & .56 *.38 & .38 & .38 & .38* 7.51 & 7.51 & 7.51 & 7.51 (1,2998)	.51 10.45 (4,2995)
GV at t2	>.99,>.99 40.65 (1),(4)	.41 5.51 (1,2998)	.01 & .56 & .56 & .56 *.0025 & .38 & .38 & .38* 0 & 7.51 & 7.51 & 7.51 (1,2998)	.95 24.36 (4,2995)
GV at t3	>.99,>.99 39.10 (1),(4)	.15 2.45 (1,2998)	.01 & .01 & .56 & .56 *.0025 & .0025 & .38 & .38* 0 & 0 & 7.51 & 7.51 (1,2998)	.93 23.33 (4,2995)
GV at t4	>.99,>.99 32.82 (1),(4)	.03 .61 (1,2998)	.01 & .01 & .01 & .56 *.0025 & .0025 & .0025 & .38* 0 & 0 & 0 & 7.51 (1,2998)	.88 19.82 (4,2995)

ANOVA had a constant value of .38.

The first column in Table 2.9 contains the power of the full true multivariate twin model. As in the MANOVA, the power of this model was lowest when the GV entered at occasion 1 (.93). It increased to >.99 when the GV entered at later occasions. The differences in power of the 1 DF test are more pronounced given an alpha of $1E^{-7}$: .10 (t1), .85 (t2), .82 (t3), .66 (t4).

2.7 Study 5: Stationary double simplex (A,E) model

In the fifth simulation study, we considered a stationary double-simplex model (A, E) with a single phenotype measured at four occasions. Common environmental effects were absent. As in study 4, the GV was added to the genetic factor (A) at occasion t, and its effect is defined as .25% of the variance in the phenotype y(t) that depends directly on A(t). Due to autoregression, the GV that enters the model at occasion t affects the phenotype at the subsequent occasions. The path diagram of this model is the same as that in Figure 2.5, except for the absence of C. The genetic autoregression coefficient (β_A) equals .9 and the environmental autoregressive coefficient (β_E) equals .7. The residual genetic and environmental (innovation) variances equal $\sigma^2_{\zeta_A} = .114$ and $\sigma^2_{\zeta_E} = .204$, respectively. These parameters resulted in a stationary model, in which the h^2 and e^2 at each occasion equal .6 and .4, respectively. The phenotypic correlations equal .82 (t1-t2, t2-t3, t3-t4), .68 (t1-t3, t2-t4), and .57 (t1-t4). Table 2.10 contains the results.

The results resemble those of study 4. We find that the sum score ANOVA is most powerful when all phenotypes were affected (GV entered at the first occasion), and that the power of this ANOVA declines progressively as the GV entered at later occasion. The power of the MANOVA was lowest when all phenotypes were affected, and increased sharply when the GV entered at a later occasion (see Cole, et al. [71]; Ferreira and Purcell, [123]).

The power of the full multivariate twin model resembled that of the MANOVA: the power was relatively low when the GV entered at t1 (.82), but increased sharply when the GV enter at t2 or later (>.99). Given an alpha of $1E^{-7}$, the power of the 1 DF test ranges from .03 (GV enters at occasion 1) to .83 (GV enters at occasion 2).

2.8 Discussion

In this paper, we considered the power of tests of genetic association using multivariate phenotypic data. Our main interest was in power of tests based on sum score ANOVAs, MANOVAs and EFAs in phenotypic data of unrelated subjects. We also reported the power of single phenotype ANOVAs, and the power of the likelihood ratio test in the full MZ & DZ twin model. Based on the results of

Table 2.10: Power, non-centrality parameters, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 5. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which all 4 genetic factors are regressed on the GV (4 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/4$ (.0025; 4 phenotypes). The power for the corrected alpha is displayed in italics.

Model	True model	ANOVA Sum scores	ANOVA Single phenotypes at 4 occasions	MANOVA
	2×1500	3000	3000	3000
GV at t1	.81, .60 12.10 (1),(4)	.52 6.94 (1,2998)	.56 & .45 & .36 & .28 *.39 & .29 & .21 & .15* 7.51 & 6.08 & 4.93 & 3.99 (1,2998)	.35 7.80 (4,2995)
GV at t2	>.99, >.99 38.91 (1),(4)	.30 4.31 (1,2998)	.01 & .56 & .45 & .36 *.0025 & .39 & .29 & .21* .0 & 7.51 & 6.08 & 4.93 (1,2998)	.93 23.32 (4,2995)
GV at t3	>.99, >.99 38.73 (1),(4)	.13 2.12 (1,2998)	.01 & .01 & .56 & .45 *.0025 & .0025 & .39 & .29* .0 & .0 & 7.51 & 6.08 (1,2998)	.93 23.18 (4,2995)
GV at t4	>.99, .99 38.43 (1),(4)	.03 .58 (1,2998)	.01 & .01 & .01 & .56 *.0025 & .0025 & .0025 & .39* .0 & .0 & .0 & 7.51 (1,2998)	.93 22.94 (4,2995)

factor model-based studies (1, 2, and 3), we conclude that overall the EFA is the most powerful model to detect association. The factor model was also found to be powerful to detect linkage by using IBD mapping in sibs (Boomsma, [39]; Boomsma and Dolan, [41]). Medland and Neale [242] and van der Sluis et al. [338] also found this approach to be powerful to detect factor level association in single factor models. However, note that in the present paper the success of the EFA in studies 1, 2, and 3 hinges on the fact that the GV effect on the phenotypes is mediated (or conveyed) by common factors, i.e., that the factor model is measurement invariant with respect to the GV (van der Sluis et al. [338]). This reduces the number of parameters that are estimated to accommodate the mean differences, and so increases the power. Van der Sluis et al. [338] demonstrated that violation of this invariance (i.e., direct effects of the GV on one or more phenotypes in the model) may greatly reduce the power. We noted that in studies 1, 2, and 3, the NCP of the MANOVA and the EFA were approximately equal, and the differences in power are solely a function of the number of estimated parameters. Modeling direct effects of the GV on the phenotypes in the EFA (as studied by Medland and Neale, [242]) renders the power of the likelihood ratio test asymptotically equal to the power in the MANOVA.

The power of the MANOVA (and so of the EFA) depends on whether the GV affects all phenotypes, or only a subset, and on the intercorrelations among the phenotypes (as noted by Ferreira and Purcell, [123]). If all phenotypes are affected, the power is relatively low, especially if the phenotypes are relatively highly correlated. If the GV affects a subset of phenotypes, increasing phenotypic correlations can be beneficial. We refer to Cole et al. [71] for a graphical explanation of these mixed effects.

In the special case of study 1, the phenotypic sum score is a sufficient statistic, in the psychometric IRT sense: the sum scores contain the same amount of the information as the constituent phenotypic test scores. Under these specific circumstances the tests based on the sum score ANOVA and EFA (subject to measurement invariance) are equally powerful. However, even if the sum score is not sufficient, the sum score ANOVA may still fare well, i.e., if the GV effect is present in all phenotypes, as shown in Study 3, 4, and 5. However, the power decreases with increasing phenotypic correlations, as higher phenotypic correlations result in larger phenotypic (sum score) variance. Also the power in the sum score ANOVA decreases as the variation in the GV effect over the phenotypes increases (study 3, S31 to S33; see also Medland and Neale, [242]).

The results of the repeated measures studies 4 and 5 are consistent with the results of studies 1 to 3. Specifically, if the GV entered at the first occasion and so affected all phenotypes, the power of the MANOVA was relatively low, while the power of the sum score ANOVA was relatively large. The power of the sum scores ANOVA decreased and the power of the MANOVA increased as the GV entered at a later occasion. For instance, as shown in Table 2.9, when the GV enters occasion 4, the power of the sum score ANOVA and the MANOVA are

.03 and .88, respectively. We note that the striking differences in the covariance structures of the repeated measure models (increasing h^2 in study 4, constant h^2 in study 5) had little bearing on the power. We did not consider the EFA, as this model is not consistent with repeated measures (Mandys, Dolan, and Molenaar, [232]). This is not to say that the factor analytic approach to repeated data is necessarily suboptimal, but the identification of the exact conditions in which an EFA of repeated measures conferred relatively good power to detect a GV is beyond the present scope.

We note the following limitations of the present power study. First, we have chosen configurations of parameter values that we deemed plausible. Many other configurations are possible. For instance, low broad sense heritability (say, .10) does not rule out the presence of quantitative trait loci of relatively large effect. Second, we have limited our analyses to 3 factor models and 2 univariate simplex models. Other models such as multivariate simplex models, or growth curves model may be of interest, depending on the available data. Third, although we reported the power of the true full multivariate twin model, we have made no effort to compare and discuss the power of this model with the power of the other tests ((M)ANOVAs and EFAs), as the study of twins and the study of unrelated subjects differ in sampling requirements (given that about one person in 50 is a twin). In terms of sample sizes, we retain an equal number of cases (3000), but a case in a twin sample naturally consists of two individuals. To arrive at an equal number of individuals, the power in the full twin model could be recalculated for NMZ = 750 and NDZ = 750 using the R code in the Appendix (these results are available on request). However, if twin data are available, genetic association analysis performed in the context of a genetically informative design is very powerful. In addition, the DZ sibpairs provides a within-family test of association that guards against stratification (see Medland and Neale, [242]; Fulker, et al., [131]). Fourth, we have limited our study to multivariate normally distributed data. Multivariate modeling of discrete data is an important issue that remains to be addressed. Fifth, we have limited the phenotypic covariance structure modeling to the exploratory factor model. Confirmatory modeling is often a viable option, is more parsimonious, and may possibly confer greater power. Sixth, in the factor models, the effect of the GV on the phenotypes was conveyed via the common genetic factors. This is in keeping with the notion that a polygenic genetic factor represents the aggregated effects of many loci. However, one cannot discard the possibility that a measured genetic locus may have a direct effect on a given phenotype (see Medland and Neale, [242], and van der Sluis et al, [338]). As studied by Medland and Neale [242] the GV effect may vary in sign from phenotype to phenotype.

In conclusion, the power studies to date have produced useful information concerning the power to detect the effects of GVs using multivariate data. We note that in the scenarios considered here (see also Medland and Neale, [242]), a multivariate approach is almost always more powerful than a univariate (i.e., single phenotype) approach. However, multivariate data require modeling choices. The reasons for collecting multivariate data depend on the nature of the phenotype(s) of interest (Hottenga and Boomsma, [164]). For instance, if the phenotypes are psychometric indicators, a well fitting common pathway model (e.g., McArdle and Goldsmith, [238]; Neale and Cardon, [253]), or a model involving a single common genetic factor plus relatively small genetic residuals would justify the use of EFA (and in special cases the use of sum scores). However, a set of phenotypes may be viewed as a system of related variables, rather than as a set of psychometric indicators. Huberty and Morris [167] describe such a system as a "collection of conceptually interrelated variables that, at least potentially, determine one or more meaningful underlying variates" (p. 304). Clearly this is sufficiently vague to justify the specific advice that one should carry out power analyses tailored to the theoretical and empirical knowledge of the (genetic) covariance structure at hand[4], rather than rely on general advice.

[4] The R and MX scripts used in this study are available on request. These can be tailored to one's own requirements.

2.9 Appendix: R code for calculating power.

```r
# start power chi2 test
rm(list=ls(all=TRUE))  # wise

powchi=function(alpha,df,NCP) {
  cv=qchisq(alpha,df,lower.tail=F)
  pchisq(cv,df=df,ncp=NCP,lower.tail=F)
}
#
alpha1=.01            # Input Type I error probability
df=1                  # Input Degrees of freedom
N1=7000               # Input The sample size N
NCP1=132.6            # Input NCP

power1=powchi(alpha1,df,NCP1)

print(c(alpha1,NCP1,power1))

N2=3000                 # Input new N
NCP2=N2*(NCP1/N1)

power2=powchi(alpha1,df,NCP2)

print(c(alpha1,NCP2,power2))

alpha2=1E-7             # Input new alpha

power3= powchi(alpha2,df,NCP2)

print(c(alpha2,NCP2,power3))
```

Listing 2.1: R code for computing the power of likelihood ratio test statistic. The input are the non-centrality parameter (NCP), the sample size (N), the degrees of freedom (DF), and the alpha (alpha). N and alpha can be varied. The actual input in this code is arbitrary.

```
# start power one way ANOVA (1df test)
rm(list=ls(all=TRUE))   # wise

powanova=function(alpha,df1,df2,NCP) {
  cv=qf(alpha,df1,df2,lower=F)
  pf(cv,df1,df2,ncp=NCP,lower=F)
}

#
alpha1=.01      # Input type I error probability
N1=3000         # Input sample size
NCP1=6.94       # Input non-centrality parameter
df1=1           # Hypothesis degrees of freedom
df2=N1-2        # Error degrees of freedom

power1=powanova(alpha1,df1,df2,NCP1)

print(c(alpha1,NCP1,df1,df2,power1))

N2=1000        # input new N
df2=N2-2
NCP2=(NCP1/N1)*N2

power2=powanova(alpha1,df1,df2,NCP2)

print(c(alpha1,NCP2,df1,df2,power2))

alpha2=.01/4    # input new alpha

power3=powanova(alpha2,df1,df2,NCP2)

print(c(alpha2,NCP2,df1,df2,power3))
```

Listing 2.2: R code for computing the power of ANOVA. The input are the non-centrality parameter (NCP), the alpha (alpha), the sample size (N). N and alpha can be varied.

```r
#start power MANOVA
rm(list=ls(all=TRUE))   # wise

powmanova=function(alpha,df1,df2,NCP) {
  fcrit = qf(alpha, df1, df2, lower=F)
  pf(fcrit, df1, df2, ncp=NCP, lower=F)
}

#
alpha1=.01              # Input alpha
nv=4                    # Input number of tests
N1=3000                 # Input sample size
NCP1=10.45              # Non-centrality parameter
df1=nv                  # Hypothesis degrees of freedom
df2=N1-nv-1             # Error degrees of freedom

power1=powmanova(alpha1,df1,df2,NCP1)

print(c(alpha1,NCP1,df1,df2,power1))

N2=6000                # input new N
NCP2=(NCP1/N1)*N2
df2=N2-nv-1

power2=powmanova(alpha1,df1,df2,NCP2)

print(c(alpha1,NCP2,df1,df2,power2))

alpha2=1E-5             # input new alpha

power3=powmanova(alpha2,df1,df2,NCP2)

print(c(alpha2,NCP2,df1,df2,power3))
```

Listing 2.3: R code for computing the power of MANOVA. The input are the alpha (alpha), the non-centrality parameter (NCP), the sample size (N), the number of tests (nv), the hypothesis degrees of freedom (df1), the error degrees of freedom (df2). N and alpha can be varied.

Chapter 3

The Use of Imputed Sibling Genotypes in Sibship-Based Association Analysis: on Modeling Alternatives, Power and Model Misspecification

Abstract

When phenotypic, but no genotypic data are available for relatives of participants in genetic association studies, previous research has shown that family-based imputed genotypes can boost the statistical power when included in such studies. Here, using simulations, we compared the performance of two statistical approaches suitable to model imputed genotype data: the mixture approach, which involves the full distribution of the imputed genotypes and the dosage approach, where the mean of the conditional distribution features as the imputed genotype. Simulations were run by varying sibship size, size of the phenotypic correlations among siblings, imputation accuracy and minor allele frequency of the causal SNP. Furthermore, as imputing sibling data and extending the model to include sibships of size two or greater requires modeling the familial covariance matrix, we inquired whether model misspecification affects power. Finally, the results obtained via simulations were empirically verified in two datasets with continuous phenotype data (height) and with a dichotomous phenotype (smoking initiation). Across the settings considered, the mixture and the dosage approach are equally powerful and both produce unbiased parameter estimates. In addition, the likelihood-ratio test in the linear mixed model appears to be robust to the considered misspecification in the background covariance structure, given low to moderate phenotypic correlations among siblings. Empirical results show that the inclusion in association analysis of imputed sibling genotypes does not always result in larger test statistic. The actual test statistic may drop in value due to small effect sizes. That is, if the power benefit is small, that the change in distribution of the test statistic under the alternative is relatively small, the probability is greater of obtaining a smaller test statistic. As the genetic effects are typically hypothesized to be small, in practice, the decision on whether family-based imputation could be used as a means to increase power should be informed by prior power calculations and by the consideration of the background correlation.

3.1 Introduction

Increasingly twin and family registries include both phenotypic data and genotypic data measured in family members (Boomsma et al. [40]; Willemsen et al. [361]). However, due to specific design or resources, the genotypic data may be limited to a subset of the family members, such as a single sibling. It is well recognized that limiting association analysis to 'the complete data participants', i.e., discarding relatives whose data are limited to phenotypic measures, is wasteful. As demonstrated by Visscher and Duffy [349] and by Chen and Abecasis [66] the genetic relations among the relatives can be used to impute genotypes of relatives lacking observed genotypic data. Subsequently including the relatives in the association study will increase the power to detect association, although actual increase depends on the phenotypic correlations among the relatives (Visscher and Duffy [349]) and on the accuracy of the imputations (Chen and Abecasis [66]).

The goal of this article is to further investigate the factors affecting power following family-based imputation. We consider imputation of up to 3 sibling genotypes given a single genotyped sibling or a single genotyped sibling and one parent. Within these imputation setups we carry out an extensive comparison of the performance of the two statistical approaches, namely, the mixture model, which involves the full distribution of the imputed genotypes and the dosage approach, in which the mean of the conditional distribution features as the imputed genotype. The comparison is performed for two minor allele frequencies (MAF) and a range of background correlations. Sibling data only are included into the association analysis, where the sibships vary from 1 (the genotyped sib) to 4 (1 observed, 3 imputed genotypes). To check the validity of our simulation program and the power calculations, we also report the power in the full information model, as an indication of the maximum power, attainable when all siblings in a sibship have observed genotypes.

Secondly, we examined the effect on power of misspecification of the background covariance structure in family-based association analysis. Imputing genotypes and extending the model to include sibships of size two and greater does require modeling the background covariance matrix. Such modeling may be of interest substantively, or as a means to reduce the parameter space. As the calculation of power to detect a measured (imputed) genetic effect will require some choice of background covariance structure, one may ask whether misspecification will affect the statistical power. To address this question, we simulate sibling phenotypes according to an additive genes/unique environment (AE) model and next, we fit two alternative models to these data: a correctly specified AE model, consistent with the model used for simulation, and a misspecified common environment/unique environment (CE) model. We compare the observed powers of the two models, with and without the misspecification. As model misspecification is of interest regardless of whether genotypes are imputed or not, we study its effect on power both in the 'all genotypes observed' setting (i.e., the full information setting) and in the setting in which some genotypes were imputed (i.e., the dosage setting).

Finally, we illustrate empirically the results obtained using simulations. In one empirical dataset we sought to quantify the power gains conferred by family-based imputation when the trait of interest is assessed on a continuous scale. This analysis aims to replicate 112 of the 180 height single nucleotide polymorphisms (SNPs) reported by Lango Allen et al. [16] in a Netherlands Twin Register (NTR) dataset consisting of 5910 siblings with observed and imputed genotypes. We explore the mixed results by means of the analysis of simulated data. The second illustration considers tests of association between observed and kinship-based imputed SNPs and a discrete trait – smoking initiation. Specifically, in a dataset comprising of 5981 observed and imputed sibling genotypes we reran the analysis of Vink et al. [346] for 20[1] of the 41 SNPs associated with smoking initiation in their discovery sample. Both analyses used solely sibling data and were carried out first in the 'complete data' samples and then by extending the samples to include the imputed sibling genotypes.

3.2 Methods

3.2.1 Models for sibship-based association

We simulated genotypic and phenotypic data for nuclear families with four siblings. In the full information setting, we computed the power to detect genetic association using the complete information, i.e., 1 to 4 sibling genotypes and phenotypes. Next, we limited the genotypic information to 1 sibling, or to 1 sibling and 1 parent, and, conditional on this information we calculated the missing genotype distribution in the remaining siblings. In this limited information setting we considered the power of the mixture model and of the dosage approach. Below we provide the details of the three modeling approaches and of our simulations.

3.2.2 The full information model

We considered a diallelic locus with alleles A and a, and frequencies $p(A)$ and $q = 1 - p(a)$, observed in nuclear families with four siblings. Let $\mathbf{g}_i$ denote the vector of genotypes of m (1 to 4) sibs in family i, where possible elements of $\mathbf{g}_i$ are AA, Aa, and aa (Falconer and Mackay [116]). Throughout, the locus has an additive effect on the phenotype, so we can assign the values $d, 0$ and $-d$ to the three possible genotypes, where the value of d is dictated by the minor allele frequency and our effect size. Letting the allele A be increaser allele, we code 1 for the genotype AA, 0 for the genotype Aa, and -1 for the genotype aa. Let $\mathbf{x}_i$ denote the vector of genotype indicators $(-1, 0$ or $1)$. We regressed

[1] Of the 41 SNPs 20 were available in the current sample

the phenotypes $\mathbf{x}_i$ observed in m sibs in family i (i.e., $\mathbf{y}_i^t = [y_{i1} \dots y_{im}]$, where t denotes transposition) on the indicators:

$$\mathbf{y}_i = b_0 + b_1 \times \mathbf{x}_i + \mathbf{e}_i \tag{3.1}$$

where $\mathbf{b}_0$ is an m vector containing the intercept (e.g., for $m = 4$, the elements of $\mathbf{b}_0$ are $\mathbf{b}_0^t = [\ b_0\ b_0\ b_0\ b_0\ ]$), b_1 is the scalar parameter of main interest, and $\mathbf{e}_i$ is the m vector of residuals. Conditional on genotype, the means are $\mu_1 = b_0 + b_1$, $\mu_2 = b_0$, or $\mu_3 = b_0 - b_1$, and the residuals are distributed $\mathbf{e} \mid \mathbf{x} \sim N(0, \mathbf{S}_0)$, where $\mathbf{S}_0$ is the $m \times m$ positive definite covariance matrix. In the OpenMx specification, the background covariance matrix was estimated using the decomposition $\mathbf{S}_0 = \mathbf{D}\mathbf{D}^t$, where $\mathbf{D}$ is an unconstrained lower triangular matrix.

We refer to this model as the full information model, as this model is based on the complete genotype information measured in all siblings in the sibship, i.e., all elements of $\mathbf{x}_i$ are observed. In this setting, the power analyses were based on both exact data simulations (Van der Sluis et al. [336]) and on the standard Monte Carlo procedure. In the latter, power was computed as the proportion of analyses in which minus twice the difference in the log likelihoods the two models – with and without the genotypic effect – is greater than a critical value associated with the chosen alpha (i.e., $c_\alpha = 6.64$ given $\alpha = .01$). The Monte Carlo procedure was employed for consistency: in the mixture approach, we do not have sufficient statistics and therefore cannot conduct exact power calculations.

3.2.3 The mixture approach

We considered the situation in which phenotypic data have been collected in sibships of sizes 2, 3, and 4, while genotypic data are limited to 1 sibling, or to 1 sibling and 1 parent.

Conditional on sib 1 genotype (g_{i1}), we calculated the probability of the sibling j ($j = 2 \dots m$) genotype (g_{ij}) as:

$$prob(g_{ij} \mid g_{i1}) = \frac{prob(g_{ij}\ \&\ g_{i1})}{prob(g_{i1})} \tag{3.2}$$

(Chen and Abecasis [66]). The probabilities $prob(g_{ij}\ \&\ g_{i1})$ and $prob(g_{i1})$ can be derived from Mather and Jinks ([237], ch. 7). Given m sibs, we calculate 3^{m-1} conditional probabilities given the sib 1 genotype. This procedure is followed for size 3 and 4 sibships, where conditionally on the genotypic information within a family, the siblings 2 to 4 genotypes are independent events.

Equation 2 can be extended to include parental genotype (g_p) if this is available additionally to the sib 1 genotype. Thus, more accurate conditional probabilities of the sib j ($j = 2 \dots m$) genotype are obtained as:

$$prob(g_{ij} \mid g_{i1}\ \&\ g_p) = \frac{prob(g_{ij}\ \&\ g_{i1}\ \&\ g_p)}{prob(g_{i1}\ \&\ g_p)} \tag{3.3}$$

Table 3.1: Posterior probabilities of the sibling 2 (s2) genotype AA, Aa, or aa, conditional on the observed genotype in a single sib (s1) or in a single sib and a single parent (p1), and given MAF = .2. The Hardy Weinberg (H-W) probabilities are the unconditional probabilities. The GPI is Kinghorn's genetic probability index, a distance measure (ranging from 0 to 100) of the imputed probabilities from the H-W probabilities.

Observed	Posterior probabilities of the s2 genotype			GPI
	AA	*Aa*	*aa*	
None (H-W)	.04	.32	.64	0
s1 *AA*	.36	.48	.16	49.33
s1 *Aa*	.06	.58	.36	38.67
s1 *aa*	.01	.18	.81	47.29
s1 *AA* and p1 *AA*	.60	.40	.00	68.92
s1 *Aa* and p1 *AA*	.10	.90	.00	86.67
s1 *AA* and p1 *Aa*	.30	.50	.20	45.33
s1 *Aa* and p1 *Aa*	.10	.50	.40	28.65
s1 *aa* and p1 *Aa*	.05	.50	.45	26.69
s1 *Aa* and p1 *aa*	.00	.60	.40	41.38
s1 *aa* and p1 *aa*	.00	.10	.90	72.27

Again the relevant probabilities can be derived from Mather and Jinks [237]. To provide an indication of the values of the posterior probabilities, these are shown in Table 3.1 for MAF of .2. Table 3.1 includes the unconditional Hardy-Weinberg (H-W) probabilities and the genetic probability index (GPI; Kinghorn [187]), which is a measure of the distance of the imputed probabilities to the H-W probabilities. The measure ranges from 0 (H-W probabilities) to 100 (genotype observed). We return to this measure in the discussion. For instance, given aa observed in sib 1, the genotype probabilities of AA, Aa, and aa are .01, .18, and .81, respectively. Given aa observed in sib 1 and in the parent, these probabilities are .0, .10, and .90.

To test for association, we fitted a mixture model that incorporates the regression model defined in Equation (1). That is, we regressed the observed phenotypes on the possible elements of x_{ij} (i.e., 1, 0, -1), and we weighted the associated densities by the conditional probabilities calculated conditional on sib 1, or on sib 1 and parent 1.

The mixture fitted to the data is a 3^{m-1} component mixture, where the proportion of sibpair genotypes within each component of the mixture is determined by the conditional probabilities (i.e., the finite mixture proportions). For example, consider a sibship of size 2, where we have at our disposal the phenotypes observed in both siblings y_{i1} and y_{i2}, the genotype observed in sib 1 (g_{i1}) and 3 probabilities based on g_{i1} (and on parental genotype (g_p), if available). Conditional on the sib 1 observed genotype (and possibly g_p) the distribution of the vector $\mathbf{y}_i$ of the observed phenotypes is assumed to follow a three component bivariate normal mixture. This mixture distribution can be expressed as the sum of 3 component distributions weighted by the fixed mixing proportions p_k (i.e., the probabilities, conditional on the observed genotype, of impute genotype AA, Aa, or aa) of sib-pairs in each component:

$$\mathbf{f}(\mathbf{y}_i; \mathbf{p}, \mathbf{S}, \boldsymbol{\mu}) = \sum_{k=1}^{3} p_k N_k(\mathbf{y}_i; \mathbf{S}_0, \boldsymbol{\mu}_k) \tag{3.4}$$

in which $\mathbf{S}$ equals $[\ \mathbf{S}_0\ \mathbf{S}_0\ \mathbf{S}_0\ ]$, the matrix $\boldsymbol{\mu}$ contains the means vectors of each component $[\ \boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2,\ \boldsymbol{\mu}_3\ ]$ (possible elements of $\boldsymbol{\mu}_k$ are $b_0 + b_1, b_0$, and $b_0 - b_1$), $\mathbf{p} = [\ p_1,\ p_2,\ p_3\ ]$ where p_k represents the fixed mixing proportions of sib-pairs within the k-component distribution, and $N_k(\mathbf{y}_i; \mathbf{S}_0, \boldsymbol{\mu}_k)$ represents the k-variate normal density function within each component. The number of components conditional on the sib 1 genotype is 3^{m-1}, hence in the case of 3 (4) siblings, we have 9 (27) components. As in the full information setting, in the specification in OpenMx, we modeled the background covariance matrix using the decomposition $\mathbf{S}_0 = \mathbf{D}\mathbf{D}^t$. We imposed no additional constraints on $\mathbf{D}$.

3.2.4 The dosage approach

In this approach, we calculated the expected value of the genotype indicator based on the conditional probabilities estimated as defined by the equations (2) or (3). That is, conditional on the sib 1 genotype (g_{i1}), and given our coding of 1 (AA), 0 (Aa), and -1 (aa), the average indicator is calculated as:

$$x^{\bullet}_{ij} = prob(g_{ij} = AA \mid g_{i1}) - prob(g_{ij} = aa \mid g_{i1}) \tag{3.5}$$

where $x^{\bullet}_{ij}$ represents the vector of expected number of increaser alleles in sib j, and $\mathbf{x}^{\bullet}_i = [\ x^{\bullet}_{i1},\ x^{\bullet}_{i2} \ldots x^{\bullet}_{im}]$ in family i. We specify the regression model for the observed phenotype $\mathbf{y}$ in family i as follows:

$$\mathbf{y}_i = \mathbf{b}_0 + \mathbf{x}^{\bullet}_i \times b_1 + \mathbf{e}_i \tag{3.6}$$

where, as above, $\mathbf{b}_0$ is an m vector containing the intercept, b_1 is a scalar parameter, and $\mathbf{e}_i$ is the m vector of residuals. The residuals are distributed approximately as $\mathbf{e} \mid \mathbf{x} \sim N(0, \mathbf{S}_1)$. The subscript serves to indicate that the conditional covariance matrices – $\mathbf{S}_0$ and $\mathbf{S}_1$ – are not expected to be exactly equal, as the variance of $e_{ij} \mid x^{\bullet}_{ij}$ is slightly lower than the variance of $e_{ij} \mid x_{ij}$ (Visscher and Duffy [349]; Chen and Abecasis [66]). As in the previous approaches, the expected background covariance matrix is modeled using the Cholesky decomposition. We computed the power to detect genetic association in this model by the means of the Monte Carlo procedure.

3.3 Model fitting

We implemented the three models in OpenMx (R package version 1.0.5; Boker et al. [36]). The full information model and the dosage model were also implemented in the R-nlme package (using the lme function; Pinheiro et al. [270]). This implementation is identical to the OpenMx implementation, except that the conditional covariance matrix was constrained as $\mathbf{S}_1 = \mathbf{J}\sigma_A^2\mathbf{J}^t + \sigma_e^2\mathbf{I}$, where $\mathbf{J}$ is the $m \times 1$ unit vector, and $\mathbf{I}$ is the $m \times m$ identity matrix. This specification is consistent with the simulation in the full information model, but slightly misspecified in the dosage model: as mentioned above, the variance of $e_{ij} \mid x_{ij}^{\bullet}$ $(j = 2 \ldots m)$ is lower than the variance of $e_{ij} \mid x_{ij}$. We expect this misspecification to be trivial, as the effect size of the QTL is small (1%; see below). In all cases the models were fitted by means of maximum likelihood estimation.

3.4 Simulation details

1000 genotypic and phenotypic datasets comprising 500 nuclear families with 4 siblings were simulated in R (R development core team [317]). We first simulated parental genotypes at a single diallelic locus in Hardy-Weinberg equilibrium and, given random mating, we used these to generate the sibling genotypes. We assumed the diallelic genotype explained 1% of the phenotypic variance. As mentioned above, we varied the minor allele frequencies (.2 and .5) and the background phenotypic correlations among siblings (.2 to .8, by 2). Note that the effect size was 1% regardless of MAF. We calculate and report the increase in power relative to an association analysis which includes only the subjects with observed genotypes, given the α of .01. All simulations were carried out using the R software package (`http://www.r-project.org/`) and were run on the Genetic Cluster Computer (`http://www.geneticcluster.org`).

3.5 Misspecification of the background covariance structure

Next, we studied the effect of misspecification of the background covariance matrix (i.e., more serious than the difference between $\mathbf{S}_0$ and $\mathbf{S}_1$) in family-based association analysis. Sibling phenotypes were simulated according to an AE model, which included: a) a SNP with equally frequent alleles, accounting for 1% of the phenotypic variance, b) background heritability of .8, .45 and .15, and c) unshared environmental effects. We considered nuclear families with sibship size 2 (pairs of monozygotic (MZ) and dizygotic (DZ) twins) and 4 (pairs of twins and 2 siblings), where genotypes as well as phenotypes were observed in all siblings in the sibship (the full information setting). Furthermore, we simulated the limited information setting, where some genotypes were missing (the dosage setting). That is, in this latter setting, 50% genotypes were missing among parent 1 and parent 2 and 50% genotypes were missing among each sibling in the sibship. To model association, two alternative models were fitted to the AE simulated data: a) the correctly specified AE model, and b) a CE model, where the background correlations among siblings in the sibship were (incorrectly) constrained to be equal. To obtain empirical estimates of the power, we carried out 10,000 replications and we computed the proportion of datasets in which the genetic effect was detected, given four levels of significance ($\alpha = 10^{-2}$, 10^{-3}, 10^{-4} and 10^{-7}). In addition, we verified the type I error rates, in both settings, when fitting the model with and without the misspecification. For this, sibling data were simulated under the null model of no association given the conditions described above; we then evaluated the effect of background misspecification on the type I error rates at alpha levels of 10^{-2}, 10^{-3} and 10^{-4} by examining 10,000 replicates (100,000 replicates for the $\alpha = 10^{-4}$ cell).

3.6 Empirical illustrations

3.6.1 Height data

We illustrated empirically the results obtained using simulations. The first analysis examines the power advantages conferred by family-based imputation when the trait tested for association is continuous. First we performed family-based imputation of 112 of the 180 SNPs previously associated with height (Lango Allen et al. [16]) and next, we carried out a sibship-based association analysis. We ran the analysis with and without the imputed sibling genotypes, and we assessed the association signals in the two samples.

The data set used for this illustration consisted of 2164 Dutch nuclear families from the NTR, where observed or self-reported height data were available for 5910 siblings born between 1914 and 1991 (N = 3667 females with a mean height of 169.89 cm and SD = 6.43 cm, and N = 2243 males with a mean height of 183.16 cm and SD = 7.07 cm). Families were included if at least one member had observed genotypes. Height was measured in adults at 18 years or older, and data of individuals with multiple measurements available underwent consistency checks (i.e., 236 siblings, representing 1.3% of the 17,195 siblings who formed the initial phenotypic sample were discarded due to differences larger than 5 cm between multiple measures). As imputation exploits biological relationships within nuclear families, we also excluded self-reported half-siblings and non-biological parents (N = 108 individuals, .5% of the phenotypic sample). Genotypic data were limited to 2410 siblings and 1437 parents. Conditional on the observed genotypes we imputed 3500 siblings who had height but no genotype data. To impute missing sibling genotypes we used our own R script (CSIBPROB, see http://www.psy.vu.nl/nl/over-de-faculteit/medewerkers-alfabetisch/medewerkers-m-p/minica-c-c/index.asp).

In the next step, we carried out a linear mixed association analysis (Visscher, Benyamin and White [351]), first by limiting the sample to the observed genotypes and second, by extending the sample to incorporate imputed siblings genotype data by using genotype dosages. Height was regressed on the genotype indicator variable and on the observed covariates (sex and birth cohort) modeled as fixed effects. As the sample included monozygotic twins (i.e., N = 656 MZ twin pairs) and full siblings, we modeled the background covariance structure by an AE model. Like in the simulations, the association analysis was limited to the sibling data.

3.6.2 The analysis of smoking initiation

The second empirical example illustrates the power gains obtained by the inclusion into an association analysis of imputed sibling genotypes when the phenotype of interest is dichotomous. Specifically, we reran the association analysis conducted by Vink et al. [346]) for 201 SNPs of the 41 SNPs associated with smoking initiation (at p-values $< 10^{-4}$) in their discovery sample. The original analysis was ran in unrelated individuals (N = 3497), while the present one is sibship-based, performed by implementing the above described two-step approach (i.e., imputation of missing sibling genotypes, which are subsequently incorporated in an association analysis).

Measured phenotypes were available for 17,641 siblings in 10,200 Dutch nuclear families from the NTR. Based on self-report, half-siblings (N = 78) and non-biological parents (N = 192) were excluded (representing .9% of the initial phenotypic sample). As in the previous empirical example, solely families with at least one parental or sibling observed genotype were retained for the analysis. There were 2210 families that met this criterion. In these families 2458 siblings and 1420 parents had observed genotypic data which were exploited to impute siblings with measured phenotypes but lacking genotypic data. The final phenotypic sample comprised of 3125 controls (never smoked tobacco) and 2856 cases (ever smoked tobacco); the siblings were born between 1914 and 1993 (mean age = 42.62 years, SD = 11.61) and 61% of the sample were females. There were 86 siblings with observed genotypes but no smoking initiation data.

To model association we used an AE generalized mixed effects model, fitted firstly to the sample limited to the 'complete data' siblings, and then to the sample incorporating siblings with imputed genotypes by using dosages. Sex and age were included as covariates. Model fitting was performed by using the MASS package (the function glmmPQL, Venables and Ripley [340]) and the nlme package for R (Pinheiro et al. [270]).

3.7 Results

3.7.1 The full information setting

We first evaluated the power to detect association in the full information setting to obtain an indication of the maximum power given maximum information (i.e., all siblings in the sibship have measured genotypes and phenotypes). This verifies the validity of our simulation program and our subsequent power calculations. The results are displayed in Figure 3.1 and the exact calculations and numerical values are shown in Table 3.2. As mentioned, the effect size was chosen to equal 1% regardless of MAF, so that these results apply equally to MAF = .2 and MAF = .5.

Figure 3.1: The expected power in the full information setting for various background correlations, given $\alpha = .01$, MAF = .2 and an effect size of 1%. Left: 500 families with 1, 2, 3 and 4 siblings. Right: 500 genotyped siblings regardless of sibship size (i.e., 500 singletons, 250 size 2 sibships, 166 size 3 sibships, and 125 size 4 sibships).

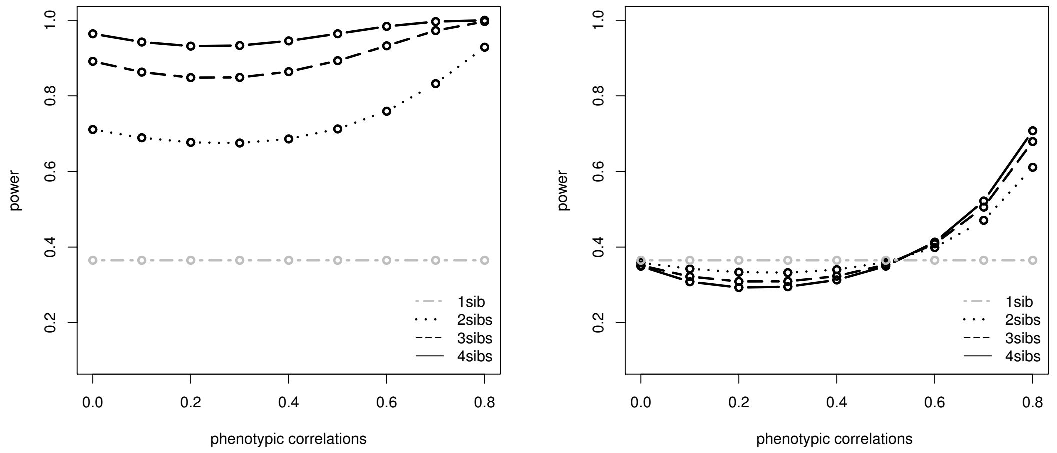


Figure 3.1 (left) demonstrates the effect of the background correlation on power, in 500 families comprising size 1, 2, 3, or 4 sibship. The differences in power between the sibships sizes are expected given the differences in sample sizes (500 singletons confer less power than do 500 size 4 sibships). This is of little concern as we are interested in the change in power associated with the use of imputed genotypes within each sibship size. However, merely for comparison, we also calculated the power for a constant number of individual cases, specifically, 125 size 4 sibships, 166 size 3 sibships, 250 size 2 sibships, and 500 singletons. These results are shown in Figure 3.1 (right). As Visscher, Andrew and Nyholt [350] noted, power suffers when related individuals are included into analysis for small

Table 3.2: Power in the full information model given an effect size of 1%, $\alpha = .01$ and $N = 500$ families. Power is shown as a function of the sibship size (nsib) and background correlation (ρ_{bg}). In the case of a singleton (nsib = 1), the background correlation is not relevant.

	background correlation			
nsib	$\rho_{bg} = .2$	$\rho_{bg} = .4$	$\rho_{bg} = .6$	$\rho_{bg} = .8$
4	.93	.95	.98	.99
3	.85	.86	.93	.99
2	.68	.69	.76	.93
1	.37	.37	.37	.37

to moderate phenotypic correlations. However, for larger phenotypic correlations, the power of a family based design exceeds the power of an association analysis conducted in unrelated individuals, given constant genotyping resources.

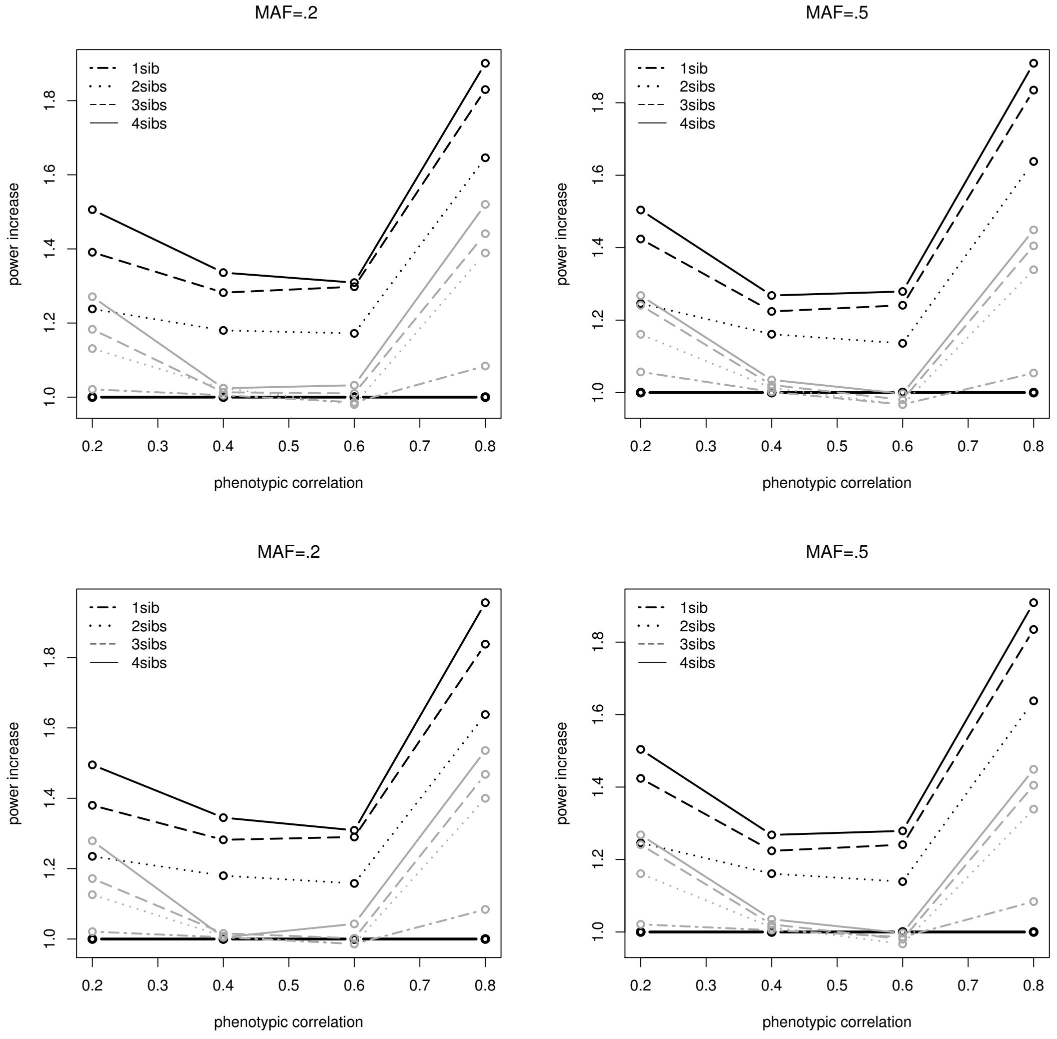
3.7.2 The mixture and dosage approaches

Next, we considered the genotypic sample consisting of both observed and imputed sibling genotypes, and within this setting we examined the power and the estimation precision of the mixture model and of the dosage approach. Figure 3.2 depicts the results of the power analyses.

We plotted the power relative to the expected power afforded by a sample size of 500 singletons, given the α of .01. The actual power in this case is .37 (Table 3.2), but this is scaled to equal 1, and the observed power is divided by this .37. Across all settings considered here, there was no difference in the observed powers of the mixture model and the dosage approach. We found the power of the two approaches was similarly affected by three factors: the phenotypic correlation (see also Visscher and Duffy [349]), the sibship size (2 to 4), and the accuracy of the imputation (based on 1 sibling or on 1 sibling and 1 parent).

When the imputation was based on 1 genotyped sibling, appreciable increase in power is observed only given relatively strong or weak background phenotypic correlations among the sibs. That is, when the background correlations were either small (i.e., less than .4) or high (i.e., greater than .6) imputing siblings increased power by about a factor of 1.2 to 2 relative to 'no imputation analysis'. Phenotypic correlations had a similar, albeit weaker effect, on the power given imputation based on 1 sibling and 1 parent genotypes. Within this setting, the association analysis including imputed sibling genotypes had greater power given low and high phenotypic correlations and it had reduced power for moderate phenotypic correlations. However, even for phenotypic correlations in this range this analysis was about a factor of 1.2 to 1.3 more powerful than the 'no imputation' analysis.

Figure 3.2: The empirical power of the Dosage model (top) and the Mixture model (bottom), relative to the expected power afforded by 500 singletons (the black bolded line), given $\alpha = .01$. The grey lines: the empirical power afforded by sibships sizes 2, 3 and 4 when imputation is based on 1 genotyped sibling. The black lines: the empirical power afforded by sibships sizes 2, 3 and 4 when imputation is based on 1 sibling and 1 parental genotypes. Power calculations are based on 1000 datasets comprising 500 families, each dataset with a simulated genetic variant explaining 1% of the phenotypic variance, regardless of MAF.



The power also increased with increasing sibship size. Apart from the moderate correlations condition, power was always larger in larger sibships, where, for instance, a size 4 sibship was about 10% more powerful than a size 2 sibship.

Furthermore, as is to be expected, the design in which imputation was based on 1 parent and 1 sibling genotypes was found consistently more powerful than

Table 3.3: Average estimates of the genetic effect b_1 and the associated standard deviations (in parenthesis) for the Mixture models, for MAF = .2. The true parameter value is $b_1 = .1767$ (1000 replicates)

Models	Sibship size	Background correlations .2	.4	.6	.8
Observed genotypes	1	.176 (.079)	.176 (.077)	.175 (.078)	.180 (.076)
Conditional probabilities given 1 sibling genotype (g_{i1})	2	.176 (.075)	.176 (.077)	.175 (.078)	.180 (.070)
	3	.176 (.073)	.176 (.076)	.175 (.078)	.180 (.067)
	4	.177 (.073)	.176 (.076)	.176 (.077)	.180 (.064)
Conditional probabilities given 1 sibling & 1 parent genotypes (g_{i1} & g_p)	2	.176 (.070)	.176 (.073)	.175 (.071)	.178 (.062)
	3	.176 (.067)	.176 (.070)	.176 (.067)	.177 (.057)
	4	.176 (.064)	.175 (.068)	.176 (.066)	.177 (.053)

the design in which the imputation was based on 1 sibling genotype only, with average power gains of 10-15%, across all conditions. In this setting, the additional information about parental genotype allowed an increase in the accuracy of the imputed genotypes, an increase that resulted in greater precision of estimating the genetic effect, and therefore was associated with greater power.

Tables 3.3 and 3.4 display the mean and the standard deviation of the estimate of b_1 for MAF = .2 obtained in the mixture model and in the dosage model, as fitted in OpenMx (MAF = .5 produced comparable results). The averages of the estimate of the genetic effect b_1 are close to their true value both when the analysis is limited to the observed genotypes and when it additionally includes imputed siblings. The variation in the standard deviation of the parameter estimate reflects the variation in power. Including siblings with missing genotypes yields unbiased estimates of the genetic effect and, as it leads to an increase in the sample size, it allows for higher estimation accuracy.

The results obtained using the dosage model implemented in OpenMx and nlme are quite similar (results not shown), notwithstanding that the background covariance matrix is highly constrained in nlme, but unconstrained in OpenMx[2]. This is expected as in the nlme specification the model for the background covariance matrix is almost completely consistent with the data generating model (the minor difference stemming from the differences between $\mathbf{S}_0$ and $\mathbf{S}_1$, as mentioned above).

Table 3.4: Average estimates of the genetic effect b_1 and the associated standard deviations (in parenthesis) for the Dosage models, for MAF = .2. The true parameter value is $b_1 = .1767$ (1000 replicates)

Models	Sibship size	Background correlations .2	.4	.6	.8
Observed genotypes	1	.176 (.079)	.176 (.077)	.175 (.078)	.180 (.076)
Dosage	2	.176 (.075)	.176 (.077)	.175 (.078)	.180 (.070)
conditional on 1 sibling	3	.176 (.074)	.176 (.077)	.175 (.078)	.180 (.067)
genotype (g_{i1})	4	.177 (.073)	.177 (.076)	.176 (.077)	.181 (.066)
Dosage conditional	2	.176 (.070)	.176 (.073)	.176 (.071)	.179 (.063)
on 1 sibling & 1 pa-	3	.176 (.067)	.176 (.071)	.176 (.067)	.178 (.059)
rent genotypes (g_{i1} & g_p)	4	.176 (.065)	.176 (.069)	.177 (.067)	.178 (.056)

The effects on power of misspecification of the background covariance structure

Figure 3.3 (left) indicates that in the full information setting the observed power of the misspecified CE model was in good agreement with the power of the correctly specified AE model for weak to moderate background correlations. With an increase in the background correlations we noted a slight discrepancy among the powers of the two models (Figure 3.3, right). The discrepancy is higher (up to about 9%) for the size 2 sibship than for the size 4 sibships. Results for the dosage model were similar (data not shown).

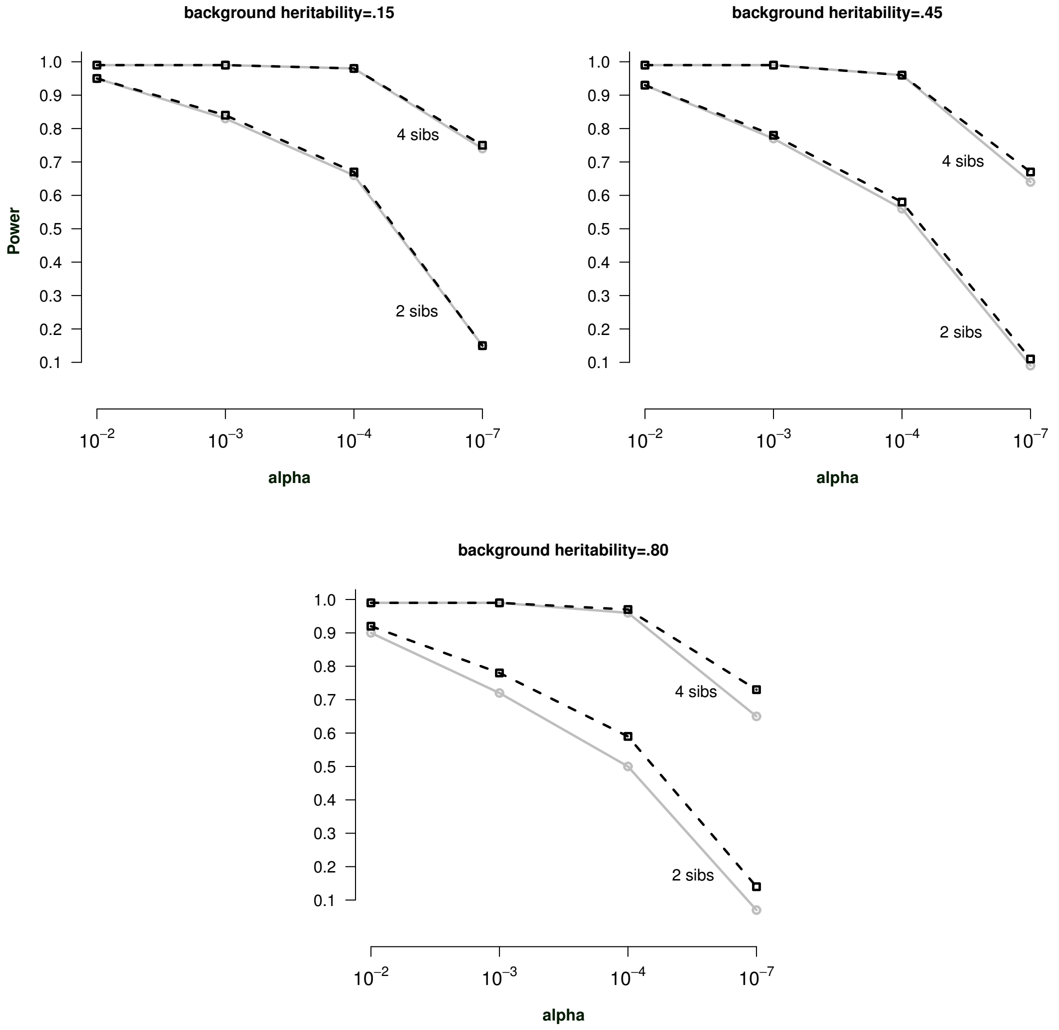
We also assessed the empirical type I error rates. Results for both the full information setting and the dosage model are given in Table 3.5. As can be seen in Table 3.5 results were akin in the two settings: they indicate that for low and moderate background correlations, the misspecification of the background covariance structure yields empirical type I error rates that are consistent with the specified alpha levels. With these settings, the likelihood-ratio test in the linear mixed model appears to be robust to the degree of misspecification of the family structure considered here. However, one can note that when the background correlations are high (MZ correlation = .80), in the incorrect model the rate of type I errors is higher than expected. This effect is stronger in the size 2 sibship than in the size 4 sibship where the misspecification pertains to a single element of a 4 × 4 covariance matrix. Finally, we note that given the scenarios considered, the full information model and the dosage approach yielded similar results, confirming that imputation per se does not affect the type I error rates (see also Chen and Abecasis [66]).

[2] Fitting the constrained model in Mx produced identical results

Table 3.5: Type I error rates in the Full information and in the Dosage settings, in the correctly specified model (AE background) and in the misspecified model (CE background, results displayed in italics). We simulated sibling phenotypes for 500 monozygotic and 500 dizygotic families and a SNP having a MAF = .5 and explaining 1% of the phenotypic variance. We varied the sibship size and the magnitude of the MZ background correlations (10,000 simulations/cell for the cells $\alpha = 10^{-2}$ and $\alpha = 10^{-3}$; 100,000 replicates for the $\alpha = 10^{-4}$ cell).

Sibship size	Background correlations	Level of significance	No missing genotypes AE/CE	Observed and imputed genotypes AE/CE
2	.15	$\alpha = 10^{-2}$	.010/.010	.009/.009
		$\alpha = 10^{-3}$	.001/.001	.001/.0009
		$\alpha = 10^{-4}$	.0001/.0001	.00007/.00008
	.45	$\alpha = 10^{-2}$	.010/.012	.010/.012
		$\alpha = 10^{-3}$	.001/.001	.0008/.001
		$\alpha = 10^{-4}$	.00007/.0001	.00009/.0001
	.80	$\alpha = 10^{-2}$	.009/.01	.01/ .01
		$\alpha = 10^{-3}$	.001/.002	.001/ .002
		$\alpha = 10^{-4}$	.0001/.0004	.0001/ .0002
4	.15	$\alpha = 10^{-2}$	.010/.010	.009/.009
		$\alpha = 10^{-3}$	.0009/.001	.001/.0009
		$\alpha = 10^{-4}$	.0001/.0001	.0001/.0001
	.45	$\alpha = 10^{-2}$	.008/.011	.008/.010
		$\alpha = 10^{-3}$	.001/.001	.0009/.001
		$\alpha = 10^{-4}$	.00009/.0001	.00007/.0001
	.80	$\alpha = 10^{-2}$	.01/.01	.009/ .01
		$\alpha = 10^{-3}$	.001/.002	.001/ .001
		$\alpha = 10^{-4}$	.0001/.0002	.0001/ .0002

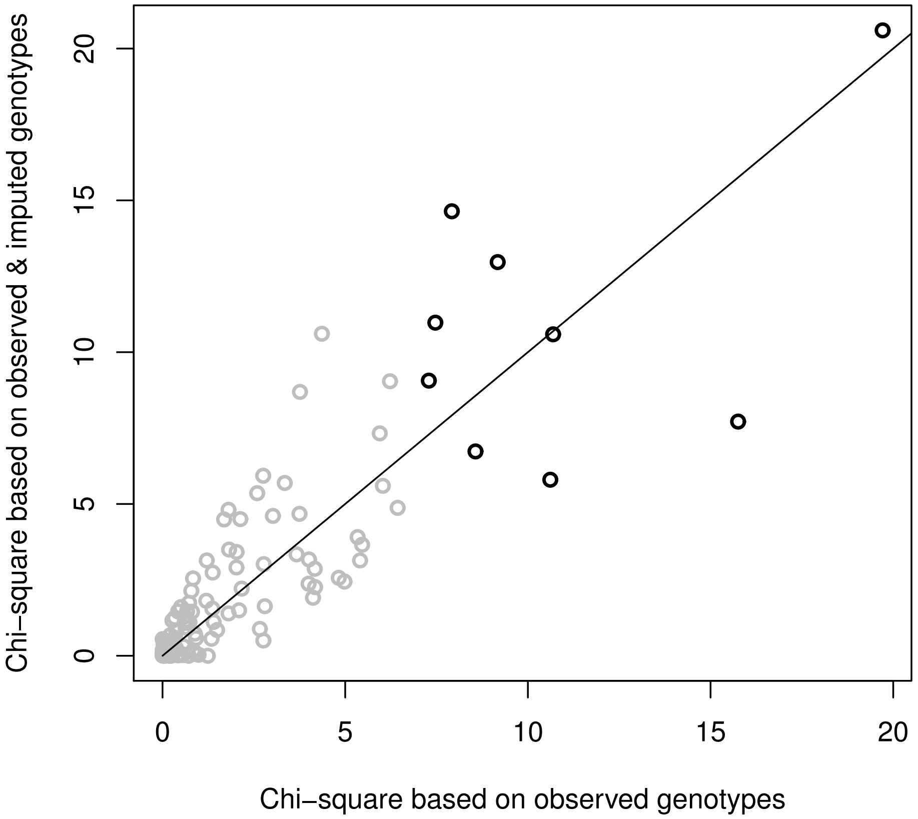
Figure 3.3: The empirical power to detect a genetic variant with a MAF = .5, that explains 1% of the trait variance in the correctly specified AE linear mixed model (the grey line) and in the misspecified CE linear mixed model (the black dashed line). In the correct model the background covariances among identical twins were specified as twice larger than in fraternal twins. In the incorrect model the background covariance matrix was estimated subject to equal covariances. The empirical power was computed for 10,000 datasets (100,000 datasets for the 10^{-7} cell) consisting of 500 MZ and 500 DZ families with sibships of size 2 and 4.



Application: Height data

The results of the sibship-based association analysis aimed at replicating 112 height SNPs in the NTR sample are illustrated in Figure 3.4. Imputation enhanced the association signal at some loci, notwithstanding that the sibling cor-

Figure 3.4: Chi-square values obtained in the analysis that incorporates 3500 imputed sibling genotypes along with the 2410 observed genotypes relative to the chi-square values obtained in the "no imputation analysis". In the latter analysis the sample is limited to the 2410 observed sibling genotypes. 112 SNPs were tested for association with height. Shown in black are the 9 hits at $\alpha = .01$ based on the observed data. Points below the diagonal are due to drop in test statistic following imputation.



relations are in the region where the power gains are lowest (i.e., siblings are correlated about .45 for height, e.g., Visscher et al. [353]). To provide an illustration, we show in Table 3.6 the markers - associations with p-values $< 10^{-2}$ based on the observed data - for which we obtained the largest increase in χ^2 by including into analysis imputed sibling genotypes.

One SNP only – rs1351394 – reached a significant association with height (p-value $< .01/112$), and clearly the association signal was stronger in the sample that included imputed siblings genotypes, i.e., $\chi^2 = 20,599$ versus $\chi^2 = 19,711$ in

Table 3.6: Increase in χ^2 obtained in a family-based association analysis that includes 2410 observed and 3500 imputed sibling genotypes, relative to an association analysis limited to the observed genotypes. The first 4 SNPs are hits at $\alpha = .01$, the SNP rs1351394 is a Bonferroni significant result.

SNP	χ^2 (no imputation analysis)	χ^2 (imputed siblings included)	χ^2 increase
rs1351164	7,467	10,972	1.47
rs724016	9,174	12,967	1.41
rs4282339	7,289	9,063	1.24
rs7759938	7,918	14,640	1.85
rs1351394	19,711	20,599	1.05

the no imputation analysis, respectively. In addition, we report the associations with a p-value $< .01$, as the present sample comprising 5910 observed and imputed sibling genotypes was underpowered to yield more significant Bonferroni[3] corrected results. These results indicate that imputation increased the power to detect association, which is consistent with our simulation results. That is, for some SNPs the χ^2 as obtained when all sibling data are used is up to a factor of 1.85 larger than the χ^2 as obtained when the analysis is limited to siblings with observed genotypes. The χ^2 averaged over the 112 SNPs was $\chi^2 = 2.499$ in the imputed sample, a value larger than the average χ^2 obtained based on the 'observed sample' ($\chi^2 = 2.285$). Importantly, the results also indicate that the value of test statistic may drop following imputation[4] (i.e., the points below the diagonal in Figure 3.4). We conjectured that this drop in value is due to the small effect sizes given that the 180 SNPs identified explain only about 10% of the height variance (Lango Allen et al. [16]). While power is increased by the imputation, the actual test statistic may still drop in value, as it remains a single realization of the distribution of the test statistic. This is more likely to occur if the gain in power is relatively small. To test this, we carried out additional simulations.

[3] For convenience we have chosen the Bonferroni method to correct for multiple testing, although this procedure can be conservative (Laird and Lange [192]). However, in Figure 3.4 we plot the values of the noncentrality parameter of the likelihood ratio test, as these values do not depend on the chosen alpha, or the correction for multiple testing. They are illustrative of the variation in power - before and following imputation – given various effect sizes.

[4] As an additional check, the analysis of height data was repeated in Merlin (Abecasis et al. [4]), and this analysis produced similar results (results not shown).

3.7.3 Additional simulations: Explaining height results

Genotypes and phenotypes of a trait with heritability of 80% (provided that the heritability of height has been estimated at about 80%, Silventoinen et al. 2003) were simulated for 100 samples consisting of 500 MZ and 500 DZ families with size 4 sibships. The effect sizes of the genetic variants were varied such that they explained .1%, .5% and 1% variance in the phenotype. To mimic the height data we also varied the percent of missingness among the observed parental and sibling genotypes: 50% genotypes were missing among parent 1 and parent 2 and 25%, 60%, 90% and 95% genotypes were missing among sibling 1, sibling 2, sibling 3 and sibling 4, respectively.

In the first step, we imputed the missing sibling genotypes conditional on the observed genotypic data. We then ran the association analyses in each of the three samples: the full information sample, where all siblings (N = 4000) had complete phenotype and genotype data, the imputed sample, consisting of siblings with observed (N = ∼1600) and imputed genotypes (N = ∼2400), and the limited sample, where missing genotypes were not imputed (N = ∼1600 genotypes observed). Figure 3.5 displays the results.

The χ^2 trend as obtained in the three samples was expected to decrease as the genotypic information decreases, with the imputed sample yielding a χ^2 value that is intermediate between those obtained in the full information setting and in the limited sample. We found that, for the .1% effect size case, we observed this trend in only 39% of the analyses, these results are shown in black in Figure 3.5a. However, an increase in the size of the effect was accompanied by an increase in the proportion of results consistent with the expected rank ordering of the χ^2 values; that is, in the .5% (1%) effect size case the trend was monotonically decreasing in 67% (80%) of the analyses (Figures 3.5b and 3.5c). It follows from these results that the most likely explanation for the drop in test statistic following imputation is the small effect sizes of the 112 SNPs accompanied by large standard errors of the relevant parameter. That the effect sizes are small, in fact too small to be detected in the present sample is evident in the fact that there were only 9 hits based on the observed data - displayed as black points in Figure 3.4 - at the very liberal alpha of .01.

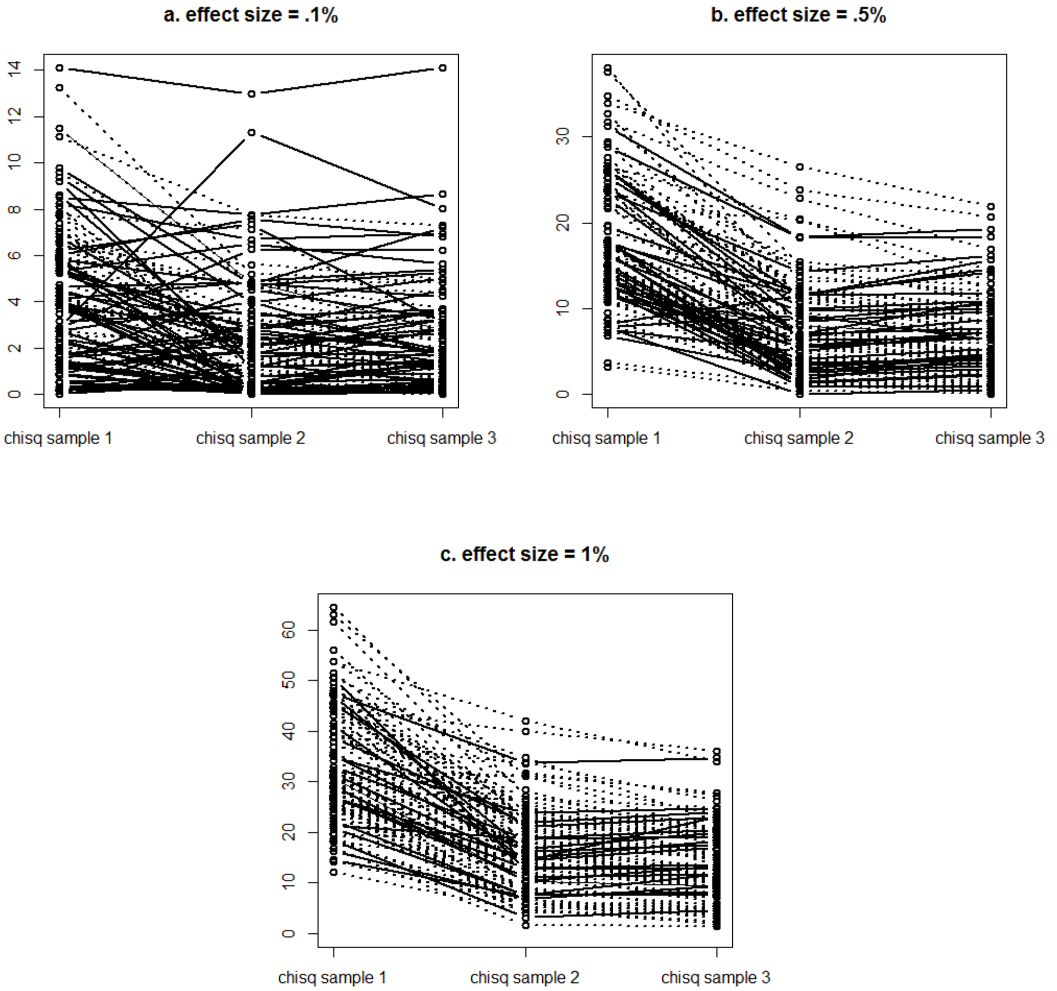
3.7.4 The analysis of smoking initiation

Results for the association analysis of smoking initiation are given in Table 3.7. The results are comparable to those obtained in the previous analysis of height: some SNPs, but not all, showed an increase in the test statistic with the inclusion into analysis of imputed sibling genotypes. Notable is the increase in power obtained at SNP rs3949478, whose association signal approached the significance threshold of 4E-04 based on the observed data and reached a p-value of 7.27E-06 by including the imputed genotypes. This SNP, located in the *ENTPD1* gene,

Table 3.7: Results of 20 tests of genetic association with smoking initiation, ran in the 'complete data' sample (N = 2458) and in the sample that includes additionally imputed siblings genotypes (N = 5981). Sibling data only were included into analysis. The background covariance matrix was modeled by an AE model. The model was fitted by means of quasi-likelihood and provided Wald-type tests of effects, which, for consistency, were converted to chi-square values.

CHR	SNP	No imputation analysis		Imputed siblings added	
		χ^2 (t-value)	p-value	χ^2 (t-value)	p-value
2	rs4608580	0.86 (0.93)	0.35	0.008 (0.09)	0.92
2	rs10865016	7.18 (2.68)	0.007	7.61 (2.76)	0.0057
2	rs787151	9.42 (3.07)	0.002	11.49 (3.39)	0.0007
3	rs1599903	0.82 (0.91)	0.36	1.06 (1.03)	0.29
3	rs9824246	0.008 (0.09)	0.92	1.21 (1.10)	0.27
3	rs16860281	7.02 (-2.65)	0.008	6.20 (-2.49)	0.01
7	rs6960379	2.49 (-1.58)	0.11	1.16 (-1.08)	0.27
7	rs2237781	5.61 (2.37)	0.01	4.79 (2.19)	0.02
7	rs4725563	0.82 (-0.91)	0.36	0.64 (-0.80)	0.41
8	rs4509385	0.03 (-0.18)	0.85	0.16 (0.41)	0.67
10	rs10999845	1.08 (1.04)	0.29	0.79 (0.89)	0.37
10	rs3949478	12.74 (-3.57)	0.0004	20.16 (-4.49)	7.27E-06
10	rs1856801	0.88 (0.94)	0.34	0.64 (0.80)	0.42
10	rs7082195	0.36 (0.60)	0.54	0.13 (0.37)	0.70
11	rs17477949	4.45 (2.11)	0.03	4.66 (2.16)	0.03
11	rs12797615	4.92 (2.22)	0.02	5.95 (2.44)	0.01
12	rs7313149	2.01 (-1.42)	0.15	1.82 (-1.35)	0.17
14	rs8009082	0.46 (-0.68)	0.49	0.94 (-0.97)	0.32
14	rs8019291	1.04 (-1.02)	0.30	0.92 (-0.96)	0.33
15	rs4774925	2.19 (1.48)	0.13	1.10 (1.05)	0.29

Figure 3.5: Chi-square as obtained in three samples: sample 1, consisting of siblings with complete phenotype and genotype data (N = 4000), sample 2, consisting of siblings with observed (N = ∼1600) and imputed genotypes (N = ∼2400), and sample 3, where missing genotypes were not imputed (N = ∼1600 observed genotypes). Results are shown for three effect sizes (100 simulated samples). The grey dotted lines show analyses where the chi-square as obtained in the three samples is monotonically decreasing, as expected. The black lines show results inconsistent with this expectation.



significantly predicts the probability of switching from never-smoking to smoking initiation, conditional on sex and age, after the Bonferroni correction has been applied ($\alpha = .01/20$).

3.8 Discussion

Results of the present study suggest the following conclusions and recommendations concerning the use of family-based genotype imputation in genomewide association studies (GWAS). First, we found the mixture model and the dosage approach accommodate equally well the uncertainty of the imputed genotypes. That is, adding imputed sibling genotypes – either by making full use of the distribution of the imputed genotypes or by using genotype dosages - produced unbiased estimates of the parameter of interest. Furthermore, the power of the two approaches was equal across the conditions which were considered. Our findings confirm the results of Visscher and Duffy [349], who carried out a small scale study of the mixture approach limited to 10 replications. They are also in accordance with the findings of Zheng et al. [380], who considered the mixture and the dosage approaches in the context of genotype imputation of single nucleotide polymorphism markers (Scheet and Stephens [291]), and found the difference to be small, except given large effects and poor imputation precision. The comparison was performed under an additive genetic model; though, we expect the two approaches would perform equally well also under a non-additive genetic model, as shown by Zheng et al. [380]. All things being equal, the dosage approach is arguably the model of choice in analyzing family data with missing genotypes, as it is computationally more convenient. However, the more demanding mixture approach might prove advantageous in certain circumstances. For instance, this approach could be used to carry out within-family tests of association, allowing one to tackle with stratification (Fulker et al. [131]; Abecasis et al. [5]).

Results of simulations confirmed that the inclusion in an association analysis of imputed sibling genotypes may increase the statistical power. Therefore, for phenotypes for which the siblings resemble each other either weakly (phenotypic correlation $< .4$) or strongly (phenotypic correlation $> .6$) one should consider the inclusion into analysis of imputed genotypes as this approach may increase the power up to a factor of 1.3 relative to the "no imputation analysis". These gains will be greater if the imputation is informed by observed genotypes in more family members and at more loci - in which case the identical-by-descent information can be exploited to impute siblings with higher accuracy, as demonstrated by Chen and Abecasis [66] and by Burdick et al. [52].

Li et al. [205] noted the advantage of imputation: "(...) imputing genotypes for known relatives of the individuals included in a GWAS of mostly unrelated individuals will always increase power (...) and should be considered whenever phenotyped relatives for the individuals to be genotyped in a scan are available" (page 391, emphasis in original). However, the computational effort is not always rewarded by significant gains in power. Specifically, as discussed by Visscher and Duffy (2006), we found the yield of this procedure to low if the phenotypic correlations among the siblings are between about .4 and .6.

As the gains in power also depend on the precision of imputation, the question arises which individuals, if genotyped, would provide maximum information about the missing genotypes in their relatives? The question can possibly be answered by considering the distance from the unconditional H-W genotype probabilities to the probabilities based on the observed genotypes in the relatives. Kinghorn's genetic probability index (GPI) can be used to express this distance (Kinghorn [187]; see also Percy and Kinghorn [267]), as it equals zero if the imputed probabilities equal the H-W probabilities, and 100 if any genotype probability equals 1. To illustrate this, we used the R library GeneticsPed (Gorjan et al. [144]) to calculate the GPI of the probabilities in Table 3.1. For instance, in the small example of Table 3.1, we find that the precision of the imputation is greatest given observed sib AA genotype and observed parent AA genotype (GPI = 86.67), and smallest given sib genotype aa and parent genotype Aa (GPI = 26.69). In contrast, a single observed AA sib confers more information that an Aa sib and an aa parent (GPI 49.33 vs. 41.38). Given that the GPI is approximately related to power, in principle this index provides a means to allocate genotyping resources (Kinghorn [188]). See also Chen and Abecassis [66] for discussion and illustration of efficient allocation of genotyping resources in multi-locus family based imputation.

Second, we investigated how statistical modeling of the background covariance matrix affected the power to detect a measured (imputed) genetic effect. For low to moderate background correlations, the likelihood ratio test in the linear mixed model appeared to perform correctly when the residual structure was misspecified. Yet, the validity of this conclusion should be considered as confined to the settings of the simulation studies: the analysis was restricted to sibling data, a small effect size of 1% explained phenotypic variance, heritabilities of .15 and .45. How robust the test is in circumstances different from those considered here (i.e., in larger pedigrees or given larger effect sizes) is subject to further study. Careful specification of the residual structure, however, is required when the trait of interest is highly heritable, as in this circumstance, the misspecification will give more false positives than expected.

Finally, concerning the empirical results we note the following. The imputation will change the distribution of the test statistics under the alternative hypothesis (effect is present), such that the power increases. How much the power increases depends on the background phenotypic correlation among siblings, the number of additional imputed cases, and on the quality of the imputation in terms of the GPI. We note that the actual observed test statistic following imputation need not necessarily be larger than the value of the test statistic observed prior to imputation. As a single realization of the distribution of the test statistic it is likely to be larger if the imputation greatly increases the power. Conversely, if the power benefit is small, that the change in distribution of the test statistic under the alternative is relatively small, the probability is greater of obtaining a smaller value. As the genetic effects are typically hypothesized to be small, in practice, the decision on whether or not family-based imputation should be used as a means to increase power should be informed by prior power calculations and by the consideration of the background correlation.

Chapter 4

Sandwich Corrected Standard Errors in Family-Based Genomewide Association Studies

Abstract

Given the availability of genotype and phenotype data collected in family members, the question arises which estimator ensures the most optimal use of such data in genomewide scans. Using simulations we compared the Unweighted Least Squares (ULS) and Maximum Likelihood (ML) procedures. The former is implemented in Plink and uses a sandwich correction to correct the standard errors for model misspecification of ignoring the clustering. The latter is implemented by fast linear mixed procedures and models explicitly the familial resemblance. However, as it commits to a background model limited to additive genetic and unshared environmental effects, it employs a misspecified model for traits with a shared environmental component. We considered the performance of the two procedures in terms of type I and type II error rates, with correct and incorrect model specification in ML.

For traits characterized by moderate to large familial resemblance, using an ML procedure with a correctly specified model for the conditional familial covariance matrix should be the strategy of choice. The potential loss in power encountered by the sandwich corrected ULS procedure does not outweigh its computational convenience. Furthermore, the ML procedure was quite robust under model misspecification in the simulated settings and appreciably more powerful than the sandwich corrected ULS procedure. However, to correct for the effects of model misspecification in ML in circumstances other than those considered here we propose to use a sandwich correction. We show that the sandwich correction can be formulated in terms of the fast ML method.

4.1 Introduction

Given the availability of large datasets of genotyped and phenotyped family members, it is of interest to determine which statistical test is most efficient in genome-wide association studies (GWAS), where computational efficiency and statistical power are important. One option is to use Plink (Purcell, Neale et al. [279]), which employs the standard Unweighted Least Squares (ULS) estimator in combination with the ULS sandwich (Rogers [285], Williams [364]) to correct the standard errors for the model misspecification of ignoring the clustering. This approach is non-iterative, and produces unbiased estimates and correct standard errors, without the need to specify a background covariance model. However, given clustered data, ULS is not necessarily the most powerful estimator (Greene [147]). Maximum Likelihood (ML) is an important alternative, but is computationally more demanding. Fast algorithms have been developed, but these employ a model for the background covariance, which is limited to additive genetic and unshared environmental effects (Lippert, Listgarten et al. [209], Pirinen, Donnelly et al. [272]). We note that shared environmental effects are often found in lifestyle and psychiatric phenotypes, like substance use (van den Bree, Johnson et al. [330], Vink, Willemsen et al. [347], Kendler, Schmitt et al. [183], Thorgeirsson, Gudbjartsson et al. [321]). This raises a practical question: in conducting a family-based analysis, should one use the sandwich corrected ULS, which is fast, robust, and requires no model to be specified for the background covariance matrix, or should one use ML, which is efficient and fast, provided one commits to a background model limited to additive genetic and unshared environmental effects? In the latter case, one may ask whether discarding shared environmental effects, affects the results of the ML procedure (Litiere, Alonso et al. [212]).

The present aim is to compare the ULS procedure with the ML procedure using simulated data. We consider the performance in terms of type I and type II error rates, with correct and incorrect background specification in ML. To correct for the effects of this misspecification, we propose to use a sandwich correction (as in Plink Purcell, Neale et al. [279]). We show that the sandwich correction can be formulated in terms of the fast ML method of Lippert et al. [209].

4.2 Methods

4.2.1 Family based model for genetic association

Let y_{ij} be the vector of observed phenotypes, where subscript j stands for individual ($j = 1 \dots n_i$) and subscript i stands for family ($i = 1 \dots N$). Let g_{ij} be the vector of observed genetic markers coded as an additive genetic model, as 0 (aa), 1 (Aa) or 2 (AA) (Falconer and Mackay [116]). We test the statistical association between each observed genetic marker and the phenotype in an appropriate regression model:

$$y_{ij} = b_0 + b_1 \times g_{ij} + \epsilon_{ij} \tag{4.1}$$

where b_0 represents the intercept, b_1 is the regression coefficient and ϵ_{ij} is the residual term. Let k equal $\sum_i^N n_i$, $\mathbf{b}^t$ equal the vector $[\ b_0\ b_1\ ]$, and $\mathbf{X}$ equal the $k \times 2$ matrix with the first column the unit vector, and the second, the k vector $\mathbf{g}$ containing the genetic information. Other covariates may be included, if desired (e.g., age, sex). The k vector of residuals $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{Xb}$ is normally distributed with $k \times k$ background covariance matrix $\mathbf{V}$ (positive definite), i.e.,$\boldsymbol{\epsilon} \mid \mathbf{g} \sim N(0, \mathbf{V})$. We assume that $\mathbf{V}$ is block diagonal (but see Visscher, Benyamin et al. [351], Lippert, Listgarten et al. [209], Pirinen, Donnelly et al. [272]), with diagonal blocks, $\mathbf{V}_i$, representing the residual positive definite covariance matrix of each family. An advantage of retaining the full matrix $\mathbf{V}$ (and not reformulating the likelihood given the sparseness) is that the block diagonal structure can be relaxed to accommodate distant genetic relatedness (Lippert, Listgarten et al. [209], Pirinen, Donnelly et al. [272], Zaitlen, Kraft et al. [376]). This makes the linear mixed approach very flexible. We assume that the elements in the diagonal blocks in $\mathbf{V}$ parameter vector $\boldsymbol{\theta}$ contains the estimated elements of the conditional covariance matrix. Given MZ and DZ families, the covariance matrix $\mathbf{V}_i$ may be calculated conditional on zygosity, but otherwise unstructured and homoskedastic. We denote this the unstructured estimate of $\mathbf{V}(\boldsymbol{\theta})$. Alternatively, $\mathbf{V}$ may be parameterized, i.e., $\mathbf{V}(\boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ may contain shared (C) and unshared (E) environmental variance components (σ_C^2, σ_E^2), and additive (A) and dominance (D) variance components (σ_A^2, σ_D^2) (Eaves [108], Martin and Eaves [234]). In this case, MZ and DZ relatedness is expressed in terms of these genetic variance components.

4.2.2 Estimation

We compare tests of b_1 based on maximum likelihood estimation and unweighted least squares estimation, with regular and sandwich corrected standard errors. The log-likelihood function is:

$$LogL(\boldsymbol{\theta}, \mathbf{b}) = \log \left[(2\pi)^{-\frac{1}{k}} |\mathbf{V}(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mathbf{y} - \mathbf{Xb})^t \mathbf{V}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{Xb})\} \right] \tag{4.2}$$

where $\mathbf{b}$ represents the fixed effects, and $\boldsymbol{\theta}$, the random effects (Pinheiro and Bates [271]). Maximization of the log-likelihood function subject to the correct specification of the background structure, yields the ML estimate of $\mathbf{b}$, $\widehat{\mathbf{b}}_{\mathrm{ML}}$, which can be tested by means of the Wald test (e.g., Dobson [102], Greene [147]). The parameterization of $\mathbf{V}(\boldsymbol{\theta})$ in the linear mixed model, given family data, is well known (van den Oord [331], Guang GuoJianmin [149], Visscher, Benyamin et al. [351], McArdle and Prescott [239], Beem and Boomsma [28], Rabe-Hesketh, Skrondal et al. [281]).

The ML estimator of $\mathbf{b}$ is based on solving $\mathbf{b}$ in the first order derivative of the ML function with respect to $\mathbf{b}$:

$$\widehat{\mathbf{b}}_{\mathrm{ML}} = \left(\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{y} \tag{4.3}$$

If $\boldsymbol{\theta}$ is unknown, this requires iteration. Note that the covariance matrix $\mathbf{V}(\widehat{\boldsymbol{\theta}})$ can also be estimated once and then used as fixed in the Generalized Least Squares estimator (see, for example, Li, Basu et al. [204], Pirinen, Donnelly et al. [272]). The Wald test of $\mathbf{b}_{1\mathrm{ML}}$ is based on $var(\widehat{\mathbf{b}}_{\mathrm{ML}}) = \big(\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\big)^{-1}$. Unweighted Least Squares (ULS) is a special case with $\widehat{\boldsymbol{\theta}} = [\widehat{\sigma}_E^2]$, i.e., $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}_E^2\mathbf{I}$. The ULS estimator can be expressed as (Draper and Smith [105], Dobson [102], Greene [147]):

$$\widehat{\mathbf{b}}_{\mathrm{ULS}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} \tag{4.4}$$

with

$$var(\widehat{\mathbf{b}}_{\mathrm{ULS}}) = \widehat{\sigma}_E^2\big(\mathbf{X}^t\mathbf{X}\big)^{-1} \tag{4.5}$$

The ULS procedure involves misspecification in the case of family data, as $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \widehat{\sigma}_E^2\mathbf{I}$ is almost certainly incorrect. To correct the standard errors, we employ the sandwich correction of $var(\widehat{\mathbf{b}}_{\mathrm{ULS}})$ (Purcell, Neale et al. [279]),

$$var\big(\widehat{\mathbf{b}}_{\mathrm{R-ULS}}\big) = \big(\mathbf{X}^t\mathbf{X}\big)^{-1}\mathbf{X}^t(\mathbf{y}-\mathbf{X}\mathbf{b})(\mathbf{y}-\mathbf{X}\mathbf{b})^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \tag{4.6}$$

We note that the sandwich correction is equally applicable to ML, given misspecified $\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)$. For instance (e.g., Dobson [102]):

$$var\big(\widehat{\mathbf{b}}_{\mathrm{R-ML}}\big) = \big(\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})(\mathbf{y}-\mathbf{X}\mathbf{b})^t\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)^{-1}\mathbf{X}\big(\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)^{-1}\mathbf{X}\big)^{-1} \tag{4.7}$$

where we employ the subscript m to denote misspecification.

Below we consider various tests of b_1 in family data of two full sibs and MZ and DZ twins with and without parents (see below). Firstly, we compare the ULS and ML procedures given correct specification of the background in ML, i.e., $\boldsymbol{\theta} = [\sigma_A^2, \sigma_E^2]$. Specifically, we consider the standard ULS and ML procedures (i.e., based on the so-called naive variance, which incorporates the assumption that the background model is correctly specified). We also consider the sandwich corrected ULS procedure (as in Plink Purcell, Neale et al. [279]), and the sandwich corrected ML procedure with the background $\mathbf{V}(\boldsymbol{\theta})$ conditioned on zygosity, but otherwise unconstrained. That is, the family covariance matrix is freely estimated within the MZ and DZ families, which is consistent with the true model. We include the sandwich corrected ML procedure to investigate whether robustification does result in an overcorrection when the underlying model is in fact correct. Secondly, to assess the effects of misspecification, we consider standard

ML estimation, with the (true) background $\boldsymbol{\theta} = [\ \sigma^2_A, \sigma^2_C, \sigma^2_E\ ]$ misspecified as (a) $\widehat{\boldsymbol{\theta}}_m = [\ \widehat{\sigma}^2_A, \widehat{\sigma}^2_E\ ]$, or as (b) $\widehat{\boldsymbol{\theta}}_m = [\ \widehat{\sigma}^2_C, \widehat{\sigma}^2_E\ ]$. In addition, we use the misspecified $\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)$ with $\widehat{\boldsymbol{\theta}}_m = [\ \widehat{\sigma}^2_A, \widehat{\sigma}^2_E\ ]$ (and the misspecified $\mathbf{V}(\widehat{\boldsymbol{\theta}}_m)$ with $\widehat{\boldsymbol{\theta}}_m = [\ \widehat{\sigma}^2_C, \widehat{\sigma}^2_E\ ]$) - estimated with standard ML using the incorrect background model - in the sandwich corrected ML procedure. We also include the standard and the sandwich corrected ULS procedures. Finally we test b_1 using the standard ML procedure, with the background correctly parameterized (i.e., estimating the variance components of the true model). We consider both the type I and type II error rates.

4.2.3 Simulation details

We generated family data for MZ and DZ families consisting of 2 sibs and MZ and DZ twins, with and without parents. Each simulated sample had a size of 4000 individuals. We simulated a diallelic genetic variant (GV) in HWE, with a minor allele frequency (MAF) of .5, and explaining one percent (1%) of the phenotypic variance. We simulated the background covariance structure according to two models: (1) a model with additive (A) and unshared (E) environmental effects, i.e., an AE model, $\boldsymbol{\theta} = [\ \sigma^2_A, \sigma^2_E\ ]$, with $h^2 = \sigma^2_A/(\sigma^2_A + \sigma^2_E)$ equal to .3, .5, or .7); (2) a model with additive genetic, shared (C) and unshared environmental effects, i.e., an ACE model, $\boldsymbol{\theta} = [\sigma^2_A, \sigma^2_C, \sigma^2_E]$, with $h^2 = \sigma^2_A/\sigma^2_{ph} = .2$, $\sigma^2_C/\sigma^2_{ph} = .6$, and $\sigma^2_E/\sigma^2_{ph} = .2$. We also considered an ACE model, with $h^2 = \sigma^2_A/\sigma^2_{ph} = .6$, $\sigma^2_C/\sigma^2_{ph} = .2$, and $\sigma^2_E/\sigma^2_{ph} = .2$ (see Tables 2S and 3S, Supplementary material). These models were chosen to represent a range of complex phenotypes. For example, data generated based on the parameter values in the first cell of Table 4.1 are illustrative for family-based association studies of highly heritable traits such as height in adults (Silventoinen, Sammalisto et al. [298]), whereas the data generated based on the parameter values in Table 4.3 may inform genomewide analyses of ACE traits, such as initiation of substance use (e.g. Thorgeirsson, Gudbjartsson et al. [321]). We used the R package MASS (Venables and Ripley [340]) for data generation. We implemented the sandwich corrected ULS and the sandwich corrected ML procedures in R. We obtained the standard ML results using linear mixed modeling as implemented in the R-package nlme (Pinheiro, Bates et al. [270]). Observed power equals the proportion of datasets out of 10,000 replications, in which the p-value associated with the Wald test was smaller than our chosen alpha = 10^{-7}. Type I error rate was assessed at alpha = $.05, .01, .001$ and .0001, using 1,000,000 datasets, simulated under the null hypothesis of $b_1 = 0$. Otherwise, given $b_1 \neq 0$, we used 10,000 replications. Simulations were run on the Lisa Computer Cluster (`www.surfsara.nl`). The R script used to obtain the results is available at `http://cameliaminica.nl/scripts.php`.

Table 4.1: Power (alpha = 10^{-7}) and parameter estimates for the ML linear mixed (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures. We simulated a genetic marker having an effect of 1% explained phenotypic variance and a MAF = .5. The sample consisted of N = 4000 individuals. The trait was simulated according to an AE background model (the true model) given various heritabilities (h^2) (10,000 simulated samples for each cell). The background model in the ML procedure is correctly specified (true or saturated, i.e., unstructured).

(a) $h^2 = 70\%$

Family structure		ML standard true model	Sandwich corrected ULS	Sandwich corrected ML (unstructured)	ULS standard
2 parents & 4 sibs	mean (b_1)	-0.141	-0.141	-0.141	-0.141
	mean (st.err.)	0.025	0.031	0.025	0.022
	mean (t-value)	-5.60	-4.62	-5.67	-6.35
	power	60.3	24.4	62.6	76.8
4 sibs	mean (b_1)	-0.141	-0.141	-0.141	-0.141
	mean (st.err.)	0.025	0.029	0.025	0.022
	mean (t-value)	-5.70	-4.95	-5.73	-6.35
	power	63.5	35.1	64.9	78.9

(b) $h^2 = 50\%$

Family structure		ML standard true model	Sandwich corrected ULS	Sandwich corrected ML (unstructured)	ULS standard
2 parents & 4 sibs	mean (b_1)	-0.141	-0.141	-0.141	-0.141
	mean (st.err.)	0.025	0.028	0.025	0.022
	mean (t-value)	-5.56	-4.96	-5.62	-6.34
	power	59.1	36.4	61.5	78.4
4 sibs	mean (b_1)	-0.141	-0.141	-0.141	-0.141
	mean (st.err.)	0.025	0.027	0.025	0.022
	mean (t-value)	-5.68	-5.25	-5.71	-6.34
	power	63.1	46.6	65.0	80.0

(c) $h^2 = 30\%$

Family structure		ML standard true model	Sandwich corrected ULS	Sandwich corrected ML (unstructured)	ULS standard
2 parents & 4 sibs	mean (b_1)	-0.141	-0.141	-0.141	-0.141
	mean (st.err.)	0.025	0.026	0.025	0.022
	mean (t-value)	-5.68	-5.40	-5.74	-6.34
	power	64.0	53.2	66.0	80.8
4 sibs	mean (b_1)	-0.142	-0.142	-0.142	-0.142
	mean (st.err.)	0.024	0.025	0.024	0.022
	mean (t-value)	-5.81	-5.63	-5.84	-6.36
	power	67.8	61.3	69.2	81.4

Abbreviations: ML maximum likelihood, ULS unweighted least squares.

4.3 Results

4.3.1 Correctly specified background model: type I and type II rates

First we checked the distribution of the 4 Wald tests given $b_1 = 0$, and the correct specification of the AE background, i.e., $\boldsymbol{\theta} = [\sigma_A^2, \sigma_E^2]$ (except standard ULS which assumes independence). As expected, the null distributions of the ML-based Wald tests (standard and sandwich corrected) and of the sandwich corrected ULS-based Wald test were correct (see Table 1S, Supplementary material). In contrast, the standard ULS procedure (without a sandwich correction) produced an excess of false positives. For instance, in the 4 sibs condition and with a 70% heritable trait, the observed type I error rate was .0024 given an alpha of .0001. Given $b_1 = -.141$ (b_1 given the chosen effect size of 1%) and the correct specification of the AE background covariance matrix in ML (with $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$) equal to .3, .5, or .7) we obtained the results in Table 4.1 concerning the power to detect the GV effect.

Table 4.2: Type I error rates for the ML linear mixed (standard and sandwich-corrected) and the ULS (standard and sandwich corrected) procedures. The background model is (a) correctly specified (true) or (b) misspecified. Background covariance matrix was generated according to an ACE model ($h^2 = .2$, $c^2 = .6$). The samples comprised of 4000 individuals (1,000,000 simulated data sets/cell).

Family structure	alpha	ML standard ACE model (true)	ML standard AE model (false)	ML standard CE model (false)	Sandwich corrected ML (false: AE structured)	Sandwich corrected ML (false: CE structured)	ULS standard E model (false)	Sandwich corrected ULS E model (false)
2 parents & 4 sibs	0.05	0.049	0.049	0.06	0.05	0.049	0.2	0.051
	0.01	0.010	0.010	0.015	0.010	0.010	0.11	0.010
	0.001	0.0010	0.00097	0.0019	0.00097	0.0010	0.045	0.0011
	0.0001	0.0001	0.00009	0.0002	0.0001	0.00011	0.018	0.00012
4 sibs	0.05	0.05	0.05	0.057	0.05	0.05	0.18	0.05
	0.01	0.01	0.01	0.0127	0.01	0.01	0.08	0.01
	0.001	0.001	0.001	0.0014	0.001	0.001	0.025	0.001
	0.0001	0.0001	0.00012	0.00018	0.00012	0.00012	0.008	0.0001

Abbreviations: ML maximum likelihood, ULS unweighted least squares.

The mean parameter estimates as produced by ML and ULS are equal, across all conditions. This is expected as the estimators are all asymptotically unbiased and consistent (Greene [147]). The standard errors as produced by the

ML standard and by the sandwich corrected ML are identical. This is expected as both procedures are based on the correct background covariance structure, be it correctly structured (i.e., $\boldsymbol{\theta} = [\ \sigma_A^2, \sigma_E^2\ ]$) or unstructured (the sandwich corrected ML). Therefore, the use of the sandwich does not result in any overcorrection. The ULS procedures are consistent, but differ in terms of power. The power of the standard ULS procedure appears to be greatest, but this is due to the fact that the standard errors are underestimated, as mentioned above. The sandwich corrected ULS procedure comes at a relative cost in terms of power (compared to ML). The loss in power increases with the family clustering due to the heritability of the trait. For example, in the 4 sibs condition, with a 70% heritable trait, the power of the sandwich corrected ULS procedure is 35.1%, whereas the power of the ML procedures is about 64%. Besides the heritability of the trait, the size of the family cluster has a bearing on the power of ULS. For instance, given a 70% heritable trait the difference in power between the ML and ULS with a sandwich correction is ∼30% and ∼35% when the sample consists of size 4 sibships and when it consists of 2 parents and 4 sibs, respectively (see Table 4.1). Note also the difference in power between the two robust methods as well (the sandwich corrected ULS and ML), with the power of the sandwich corrected ML procedure being higher.

4.3.2 Misspecified background model

We evaluated consequences on type I and II error rates of misspecifying the background model, $\mathbf{V}(\boldsymbol{\theta})$. We employed a background model with additive genetic (σ_A^2), and shared and unshared variance components (σ_C^2 and σ_E^2), and discard the effects of σ_A^2 (ML with an incorrect CE structured background) or σ_C^2 (ML with an incorrect AE structured background), or discard both σ_A^2 and σ_C^2 (ULS with an incorrect E structured background). ML with a correctly specified background is also included. First we considered the type I error rates, given $b_1 = 0$. Table 4.2 contains the results.

Based on these results we conclude that the type I error rates of the ML procedure are not greatly affected by the misspecification. The misspecification $\boldsymbol{\theta}_m = [\ \sigma_C^2, \sigma_E^2\ ]$ is associated with a slight inflation (e.g., .0002 given alpha = .0001 in the 2 parents and 4 sibs cell), but the ML with the CE structured sandwich corrects this (.00011). The misspecification $\boldsymbol{\theta}_m = [\ \sigma_A^2, \sigma_E^2\ ]$ hardly affects type I error rates. As expected, the standard ULS procedure ($\boldsymbol{\theta}_m = [\sigma_E^2]$) produced incorrect type I error rates (e.g., .008, given alpha =.0001 in the 4 sibs cell). However, as above, the ULS sandwich correction yields correct type I rates. The ML with an ACE background is correctly specified and produces correct type I error rates.

Table 4.3 contains the results relating to the power given $b_1 \neq 0$ and misspecified background. As expected, all modeling approaches yielded similar mean estimates of b_1, regardless of the specification of the background structure. Given

Table 4.3: Power (given alpha = 10^{-7}) and parameter estimates for the ML (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures. The background model is (a) correctly specified (true) or (b) misspecified. Background covariance matrix was generated according to an ACE model ($h^2 = .2$, $c^2 = .6$). The genetic marker explained 1% phenotypic variance and had a MAF = .5. The samples consisted of N = 4000 individuals (10,000 simulated data sets per cell).

Family structure		ML standard ACE model (true)	ML standard AE model (false)	ML standard CE model (false)	Sandwich corrected ML (false: AE structured)	Sandwich corrected ML (false: CE structured)	ULS standard E model (false)	Sandwich corrected ULS E model (false)
2 parents and 4 sibs	mean(b_1)	-0.141	-0.141	-0.141	-0.141	-0.141	-0.141	-0.141
	mean(st.err.)	0.019	0.021	0.018	0.021	0.019	0.022	0.037
	mean(t-value)	-7.54	-6.59	-7.89	-6.6	-7.44	-6.33	-3.86
	power	98.6	89.4	99.2	89.4	98.1	73.0	7.5
4 sibs	mean(b_1)	-0.141	-0.141	-0.142	-0.141	-0.142	-0.141	-0.141
	mean(st.err.)	0.019	0.022	0.019	0.022	0.020	0.022	0.033
	mean(t-value)	-7.27	-6.49	-7.49	-6.50	-7.25	-6.36	-4.33
	power	97.4	88.1	98.2	88.0	97.1	75.9	16.4

Abbreviations: ML maximum likelihood, ULS unweighted least squares.

correct background specification ($\boldsymbol{\theta} = [\sigma_A^2, \sigma_C^2, \sigma_E^2]$) and sibships size 4, the power is about 97.4% (standard ML). The power of the standard ML procedure appears to increase to about 98.2%, when σ_A^2 is discarded ($\boldsymbol{\theta}_m = [\sigma_C^2, \sigma_E^2]$), but this is spurious as it is due to the effect of the misspecification on the type one error (see Table 4.2). This effect is likely to be more noticeable at more stringent alpha levels (see also Minică, Dolan et al. [244]). The ML with a CE structured sandwich, however, preserves the power equal to the power of the (true) ML ACE model, without inflating the type I error rate. Ignoring shared environmental effects, i.e., dropping σ_C^2 in a $\boldsymbol{\theta} = [\sigma_A^2, \sigma_C^2, \sigma_E^2]$ model results in a loss in power. For instance, in the 4 sibs condition, the power of the standard ML procedure drops to about 88.1%, when σ_C^2 is discarded ($\boldsymbol{\theta}_m = [\sigma_A^2, \sigma_E^2]$) (similar results were obtained when dropping σ_D^2 in a $\boldsymbol{\theta} = [\sigma_A^2, \sigma_D^2, \sigma_E^2]$ model, where D stands for dominance; see Table 4S Supplementary material). With an AE structured background, the standard errors as produced by the standard and the sandwich corrected ML are very similar, and so is the power. Given that the latter correctly reflects the parameter variance in presence of a misspecified model, this result indicates that in the conditions considered here this type of misspecification does not affect estimation (i.e., type I error rate is well controlled). However, this is not a general finding. Consider the extreme misspecification of the background employed by the ULS method. This has a clear effect, which is reflected in the notable discrepancy observed between the standard and the robust (correct) ULS standard errors (i.e., 0.022 vs. 0.033). Finally, although both are correct, we note that the sandwich corrected ML procedure is appreciably more powerful than the sandwich corrected ULS procedure (e.g., power of 88.1% for the sandwich corrected ML with a misspecified AE structured background vs. power of 16.4% for the sandwich corrected ULS procedure). Results follow similar trends in the samples consisting of 2 parents and 4 sibs.

Given these results pertain to averages over replications, we also looked at how often the ML t-values actually exceed the sandwich corrected ULS t-values, considering also the smaller effect sizes to be expected in GWAS. This might be of interest as it will provide an indication on how the two estimators are expected to perform in individual studies involving family data. Dots above the diagonal in Figure 4.1 show how often the ML-based Wald test is larger than the sandwich corrected ULS-based Wald test, given decline in the size of the genetic effect. Figure 4.1 left shows that the ML (true AE model) almost always produces a larger test statistic, when the effect size is relatively large (effect size of 1% explained phenotypic variance) and the sample is large enough to capture it. In the example, in just about 7.5% of the samples the sandwich corrected ULS test statistic was larger. However, as the effect size decreases, one can observe more and more sandwich corrected ULS-based Wald tests larger than those estimated by the ML procedure (as illustrated in Figure 4.1 center). It can be seen that under the null model (Figure 4.1, right) no differences occur between the two estimation methods, which is as expected provided both are correct.

4.3.3 FaST-LMM formulation of the ML sandwich correction

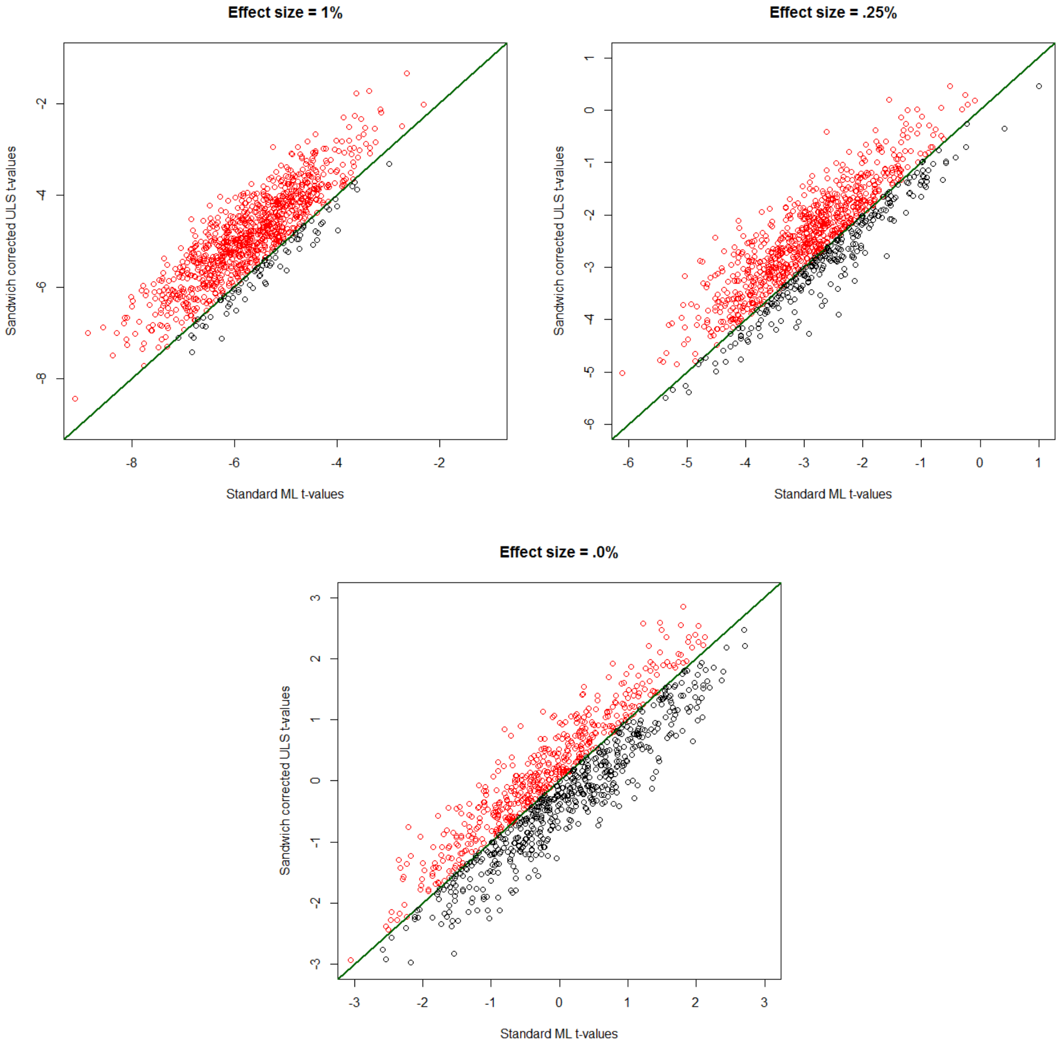
The sandwich correction is computationally relatively simple and quick in the standard formulation of the linear mixed model. We note that the fast full information maximum linear mixed procedures (Lippert, Listgarten et al. 2011 [209], Pirinen, Donnelly et al. [272]) are equally amenable to a sandwich correction. The ML sandwich can be presented as:

$$var\big(\widehat{\mathbf{b}}_{\mathrm{R-ML}}\big) = \big(\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\big)^{-1}\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})(\mathbf{y}-\mathbf{X}\mathbf{b})^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\big(\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}\big)^{-1} \tag{4.8}$$

Given random effects $\widehat{\boldsymbol{\theta}} = [\ \sigma_a^2, \sigma_e^2\ ]$, the background covariance matrix is reformulated as $\mathbf{V}(\boldsymbol{\theta}) = (\sigma_a^2 \times \mathbf{K} + \sigma_e^2 \times \mathbf{I}) = [\ \sigma_a^2(\mathbf{K} + \delta \times \mathbf{I})\ ]$ where $\mathbf{K}$ is the genetic relationship matrix (positive semi-definite), $\mathbf{I}$ is the identity matrix, and $\delta = \sigma_a^2/\sigma_e^2$. Lippert et al. ([209]; see also Pirinen, Donnelly et al. [272]) formulate the covariance matrix as follows:

$$\mathbf{V}(\boldsymbol{\theta}) = [\ \sigma_a^2 \times (\mathbf{USU}^t + \delta \times \mathbf{UIU}^t)\ ] = [\ \sigma_a^2 \times \mathbf{U}(\mathbf{S} + \delta \times \mathbf{I})\mathbf{U}^t\ ] \tag{4.9}$$

Figure 4.1: Wald tests produced by the sandwich corrected ULS procedure compared to the test statistic obtained based on full information maximum likelihood (standard ML) estimation method. We simulated 1000 data sets consisting of 500 MZ and 500 DZ 4-sib families, we varied the size of the genetic effect (1%, .25% and the null model). The heritability of the trait was $h^2 = 70\%$. The dots above the diagonal show the number of times the standard ML procedure produced a larger test statistic.



Effect size = 1%

Effect size = .25%

Effect size = .0%

where $K = \mathbf{USU}^t$ is the eigen value decomposition of $\mathbf{K}$, with $\mathbf{U}$, the eigenvectors, orthonormal, and $\mathbf{S}$ diagonal (eigenvalues). The matrix $\delta \times \mathbf{I}$, being diagonal and constant, can be written $\delta \times \mathbf{UIU}^t$. The inverse is:

$$\mathbf{V}(\boldsymbol{\theta})^{-1} = [\ \sigma_a^{-2} \times \mathbf{U}(\mathbf{S} + \delta \times \mathbf{I})^{-1}\mathbf{U}^t)\ ] \tag{4.10}$$

Note that the addition of off-diagonal terms in $\sigma_e^2 \times \mathbf{I}$, i.e., terms accommodating shared environmental effects, would render the method invalid as then the eigenvectors of the environmental covariance matrix cannot be chosen to equal $\mathbf{U}$. In terms of this treatment of the matrix $\mathbf{V}(\boldsymbol{\theta})$, the sandwich can be written

$$\begin{aligned} var\big(\widehat{\mathbf{b}}_{\mathrm{R-ML}}\big) = \sigma_a^2 \times \big[\, \mathbf{X}^t\mathbf{U}(\mathbf{S}+\delta\times\mathbf{I})^{-1}\mathbf{U}^t\mathbf{X}\,\big]^{-1}\sigma_a^{-2} \times \mathbf{X}^t\mathbf{U}(\mathbf{S}+\delta\times\mathbf{I})^{-1}\times \\ (\mathbf{U}^t\mathbf{y}-\mathbf{U}^t\mathbf{X}\mathbf{b})(\mathbf{U}^t\mathbf{y}-\mathbf{U}^t\mathbf{X}\mathbf{b})^t\times \\ \big[\,\sigma_a^{-2}\times\mathbf{X}^t\mathbf{U}(\mathbf{S}+\delta\times\mathbf{I})^{-1}\big]^t\times\sigma_a^2\times\big[\,\mathbf{X}^t\mathbf{U}(\mathbf{S}+\delta\times\mathbf{I})^{-1}\mathbf{U}^t\mathbf{X}\,\big]^{-1} \end{aligned} \quad (4.11)$$

In implementing this, the fact that $(\mathbf{S}+\delta\times\mathbf{I})^{-1}$ is diagonal may be exploited to increase computational efficiency.

4.4 Discussion

We compared the standard and sandwich corrected ULS and ML procedures, in the context of family-based association analysis of a normally distributed phenotype. Conditional on the correct specification of the background, the standard ML procedure is appreciably more powerful than the sandwich corrected ULS procedure. The actual difference in power depends on the magnitude of the residual correlations, but increases with greater family resemblance.

We also considered the sensitivity of ML to model misspecification. Model misspecification involves the mismatch between the true background covariance model (say, an ACE or ADE trait) and the background model used in the analyses (a CE or AE model).

This may occur in using fast ML procedures, which employ the background covariance matrix necessarily limited to additive genetic (A) and unshared environmental (E) effects, (e.g., Abecasis, Cherny et al. [4], Lippert, Listgarten et al. [209]). The standard ML procedure was quite robust under model misspecification in the simulated settings, and appreciably more powerful than the sandwich corrected ULS procedure. However, for circumstances other than those considered here, a sandwich correction is equally applicable to ML to correctly capture the parameter variance in presence of model misspecification. The sandwich corrected standard errors may also be employed as a means to get an indication of the effects of background misspecification on the type I error rate (i.e., the larger the discrepancy between the naive and sandwich corrected standard errors, the more likely the type I error rate of the procedure without a sandwich to be affected Chavance and Escolano [61]).

In the present paper, we considered a normally distributed phenotype. Our conclusions apply equally to generalized linear modeling of binary traits, such as disease status. To demonstrate this we included in the supplementary material (Tables 5S and 6S) results based on continuous and dichotomized (median - split) phenotypes. With respect to binary phenotypes, we note that a general (rather than generalized) linear model is often used in analyzing such variables (e.g., Zhou and Stephens [381]). Cogent arguments have been presented that the linear model may suffice in the analysis of binary phenotypes (Lippert, Listgarten et al. [209], Pirinen, Donnelly et al. [272]).

Although relatively simple to implement and more efficient than the sandwich corrected ULS in correcting for model misspecification, to our knowledge the ML sandwich correction has not been yet implemented by any of the current software for GWAS that can handle family data. With respect to implementation, we note that generalized estimating equations (GEE) procedure, as implemented in R (Carey [58]) has four useful aspects. First, it has a choice of background models, which includes the independence model and exchangeable model (the latter is equivalent to the CE model in linear mixed modeling). Second, it includes sandwich corrected standard errors of the parameters $\mathbf{b}$. Third, GEE covers generalized linear model. Fourth, as GEE is a library it can be accessed from Plink (Purcell, Neale et al. [279]) and so provides a computationally feasible strategy for running genomewide scans in family data. An annotated R script to do this is available at `http://cameliaminica.nl/scripts.php`.

In conclusion, for traits characterized by moderate to large familial resemblance, using ML with a correctly specified model for the familial covariance matrix should be the strategy of choice. For such traits, the potential loss in power encountered by the sandwich corrected ULS procedure does not outweigh its computational convenience. Using a fast ML algorithm that commits to a background model limited to additive and unshared environmental effects is acceptable even if shared environment has an influence on the phenotype of interest. That is, in the settings considered here, type I error rate of the standard ML was hardly affected by model misspecification. However, a sandwich correction is still of interest when employing ML in genomewide scans because: (a) it produces correct standard errors regardless of whether the model is correctly parameterized or misspecified; hence it should be useful for situations other than those considered here, (b) it does not result in any overcorrection when the background model is in fact correctly specified, (c) as shown above, it is computationally cheap and can easily be incorporated in the fast ML procedures, and (d) it is a useful diagnostic tool for assessing model misspecification (Chavance and Escolano [61]). Currently Plink often is the preferred software when consortia share GWA results for meta-analyses. When including data from cohorts that include relatives, one should realize that the corrected standard errors while in many circumstances larger than the ML standard errors, are accurate, and so therefore are its type I error rates. For ordinary GWAS (i.e., not family based), Plink is as good as FaST-LMM (as then ULS and ML are identical).

Supplementary information is available at the European Journal of Human Genetics's website.

Chapter 5

MZ Twin Pairs or MZ Singletons in Population Family-Based GWAS? More Power in Pairs.

Abstract

Occasionally in family-based GWAS, including monozygotic (MZ) twins, the data from one MZ twin are dropped, thus reducing the MZ pairs to singletons. Using simulations we show that the presence of MZ twin pairs does not affect the type I error rate, and reducing MZ pairs to singletons results in a loss of power. If the main interest is in the association, and not in the details of the conditional covariance matrix, adequate modeling of this matrix can be handled efficiently using GEE , with sandwich corrected standard errors.

5.1 MZ Twin Pairs or MZ Singletons?

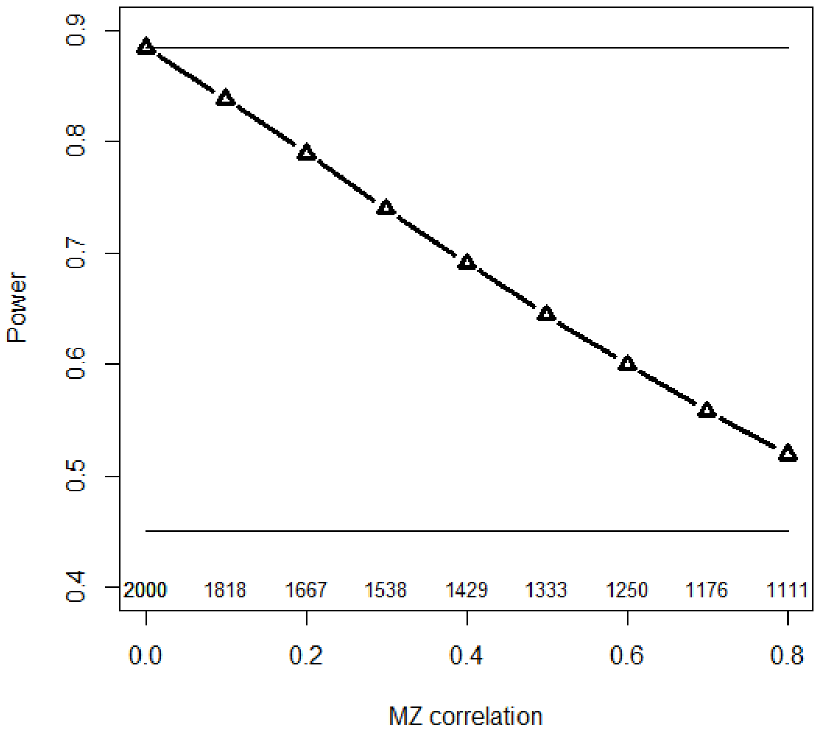
Family-based genome-wide association studies (GWAS) involve testing the genetic association of (many) genetic variants with the phenotype of interest, while taking into account the relatedness among family members. Occasionally in family-based GWAS, including monozygotic (MZ) twins, the data from one MZ twin are dropped, thus reducing the MZ pairs to singletons (e.g., Lowe, Maller et al. [219], Parsons, Lester et al. [263], Loukola, Wedenoja et al. [218], Psychosis Endophenotypes International Consortium, Wellcome Trust Case-Control Consortium et al. [76]). From a statistical power perspective, this practice is not optimal. To evaluate the issue of power, we consider the effective sample size (N_E), i.e., the number of independent cases that provides the same power as N MZ twin pairs. Given the MZ intraclass correlation of ρ, the effective sample size is calculated as $N_E = (2 \times N)/(1 + \rho)$, where N_E ranges from $N(\rho = 1)$ to $2 \times N(\rho = 0)$. For instance, given $N = 1000$ pairs discarding data from one MZ twin reduces the sample size to 1000 singletons, i.e., the N_E assuming $\rho = 1$. However, given $\rho = .2$ $(.4, .7)$, the N_E is 1667 (1429, 1176), so that 1000 twin pairs (2000 individuals) are equivalent – in terms of power – to 1667 (1429, 1176) unrelated individuals. To illustrate the loss in power, we consider a candidate gene explaining 1% of the variance, the power to detect the association in linear regression with $N = 1000$ MZ twin pairs ($\alpha = 0.001$). MZ singletons, i.e., 1000 unrelated subjects provide power of .450. Retaining data from both MZ twins (1000 pairs), the power varies with ρ as follows: .884 ($\rho = 0$), .789 ($\rho = .2$), .643 ($\rho = .5$), and .519 ($\rho = .8$). We refer to Figure 5.1 for more details.

Importantly, the gains associated with retaining MZ pairs involve no additional genotyping costs. That is, given the almost perfect concordance rate observed in MZ twins (>99%), genotyping one twin suffices in twins of confirmed monozygosity.

An important related question is whether retaining both MZ twins affects the type I error rate, i.e., does the empirical type I error rate equal the chosen α? We checked the type I error rate by means of simulations. Our results indicate that the empirical type I error rate is correct, i.e., invariably equals the chosen α (for details we refer to Table 5.1 and to Figure 5.2). Minică et al. [245] evaluated the type I error in samples involving MZ twins, full sibs and parents, and also found that the empirical α closely resembled the nominal α. We conclude that the presence of MZ twins alone, or MZ twins in combination with other family members, does not affect the type I error rate.

We note that many meta-analyses of GWASs rely heavily on twin registries. For example, the educational attainment GWAS (Rietveld, Medland et al. [283]) included more than 35% data from twin registries. Twin registries also contributed 13% cases and 9% controls to migraine meta-analysis (Anttila, Winsvold et al. [20]), 34% of the sample to telomere length meta-analysis (Codd, Nelson et al. [69]) and 31% cases and 19% controls to the meta-analysis of GWASs for

Figure 5.1: The power to detect a genetic effect (1%) in 1000 MZ twin pairs as a function of the MZ twin correlation ($\alpha = .001$). The effective sample size, shown above the x-axis, varies from 2000 (MZ correlation = .0) to 1111 (MZ correlation = .8). The top horizontal line indicates the power afforded by 2000 MZ individuals when MZ correlation equals 0. The bottom horizontal line indicates the power afforded by 1000 singletons.



major depressive disorder (Ripke, Wray et al. [284]). These registries are rich resources of phenotypic and genotypic twin data. Whereas the MZ data may be exploited fully in primary and in meta-analyses (e.g., the contribution of the Queensland Institute of Medical Research (QIMR) to Rietveld et al. [283]), consortia protocols often stipulate dropping MZ twins. Consider, for instance, the recent meta-analysis of GWASs for major depressive disorder (Ripke, Wray et al. [284]). Although genotypic data were available in ~1890, ~786, and ~300 MZ twin pairs at the Netherlands Twin Register, the QIMR and the TwinGene cohort, respectively, only 1 twin of a pair was selected for the analyses. Given an MZ correlation for depression of $\rho = .35$ these 2976 MZ twin pairs (5952 individuals) are equivalent in terms of power to $N_E = 4409$ unrelated subjects. By dropping 1 MZ twin, the equivalent of 4409 - 2976 = 1433 unrelated individuals was discarded from the meta-analysis. The corresponding loss in power is notable (i.e., from .823 power MZ twins would afford, to .395 power afforded by MZ singletons, given $\alpha = 10^{-8}$ and a genetic variant explaining 1% of the phenotypic variance).

Full modeling of data on families including MZs can be performed by using a mixed-effects variance components approach (e.g., using MERLIN and MERLIN-offline, see `http://genepi.qimr.edu.au/staff/sarahMe/merlin-offline.html`). If the families are highly variable in the number and composition of participating family members, retaining all data may pose a challenge as modeling the conditional (i.e., conditional on the genetic variant) covariance structure can be complicated and subject to misspecification. One tractable solution is to use generalized estimating equations (GEE) with a conditional covariance matrix containing equal covariances (i.e., 'exchangeable working correlation matrix' in GEE terms), in combination with a sandwich correction for the standard errors. The use of a sandwich correction is advisable as it produces correct type I error rates, regardless of misspecification. This method fares well in terms of power, in comparison to full (correct) modeling, while the computational burden is acceptable given typical GWAS requirements (Minică, Dolan et al. [245]). We note that GEE with the exchangeable option (as implemented in R (Carey, Lumley et al. [58])) can be conducted from the Plink platform (see `http://cameliaminica.nl/scripts.php`).

In conclusion, the presence of MZ twin pairs does not affect the type I error rate, and reducing MZ pairs to singletons results in a loss of power. If the main interest is in the association, and not in the details of the conditional covariance matrix, adequate modeling of this matrix can be handled efficiently using GEE, with sandwich corrected standard errors.

5.1.1 Supplementary Results: Type I Error Rate

The following results demonstrate that the type I error rate in a genetic association test in monozygotic (MZ) twin pairs is correct. We simulated, in 1000 MZ pairs, a normal phenotype and a single diallelic genetic variant with MAF = .5, of no effect. We varied the MZ correlation from .1 to .8, and ran one million replications for each value of the correlation. We tested the genetic variant effect using linear regression and, having dichotomized the continuous phenotype (probability of "affected" .05 and .20), using logit regression. We used generalized estimating equations (GEE) to accommodate the MZ dependence. The q-q plots of the observed and expected quantiles in Figure 5.2 and the results included in Table 5.1 show the type I error rates are correct.

Figure 5.2: The null distribution of the Wald test statistic (1,000,000 replications), given a continuous trait (Subfigures a and b) and a binary trait, based on the test in 1000 MZ pairs. The binary trait was obtained by dichotomizing the continuous trait into a binary 0/1 phenotype (probability of 1 is .20 in Subfigures c and d; .05 in Subfigures e and f).

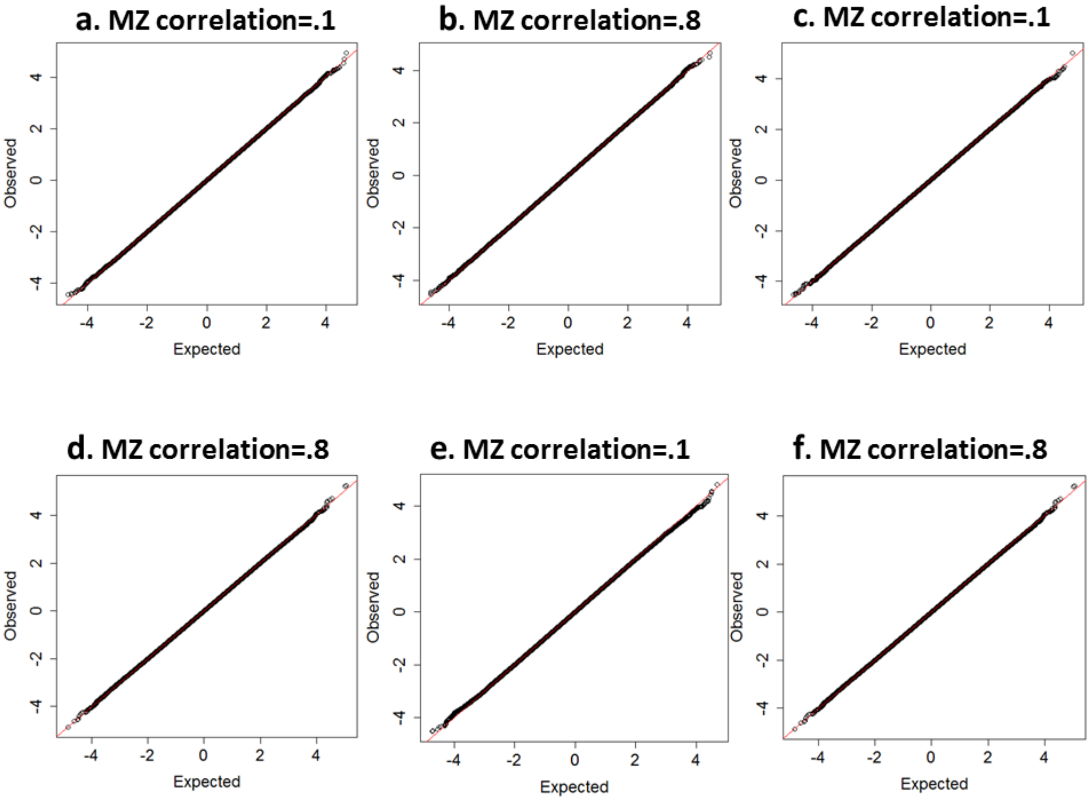


Table 5.1: Empirical type I error rates in a test of genetic association with continuous phenotypes (linear model) and with binary phenotypes (logit model) using only MZ individuals ($N = 1000$ MZ twin pairs). Within the square brackets we report the 99% confidence intervals (CI).

Trait	MZ correlation	alpha = .05 [99% CI]	alpha = .01 [99% CI]	alpha = .001 [99% CI]	alpha = .0001 [99% CI]
continuous	.1	0.0504 [0.04984, 0.05097]	0.0103 [0.01004, 0.01056]	0.00104 [0.00096, 0.00113]	0.000097 [0.00007, 0.00013]
	.2	0.0503 [0.04974, 0.05087]	0.0102 [0.00994, 0.01046]	0.00105 [0.00097, 0.00114]	0.000092 [0.00007, 0.00012]
	.4	0.0506 [0.05004, 0.05117]	0.0101 [0.00985, 0.01036]	0.00097 [0.00089, 0.00105]	0.000094 [0.00007, 0.00012]
	.6	0.0501 [0.04954, 0.05067]	0.0104 [0.01014, 0.01067]	0.00105 [0.00097, 0.00114]	0.00011 [0.00009, 0.00014]
	.8	0.0504 [0.04984, 0.05097]	0.0101 [0.00985, 0.01036]	0.00099 [0.00091, 0.00107]	0.000099 [0.00008, 0.00013]
binary (20% cases)	.1	0.0504 [0.04984, 0.05097]	0.0098 [0.00955, 0.01006]	0.00094 [0.00086, 0.00102]	0.000089 [0.00007, 0.00012]
	.8	0.050 [0.04944, 0.05056]	0.0099 [0.00965, 0.01016]	0.00094 [0.00086, 0.00102]	0.000092 [0.00007, 0.00012]
binary (5% cases)	.1	0.0492 [0.04865, 0.04976]	0.0094 [0.00915, 0.00965]	0.00082 [0.00075, 0.00090]	0.000077 [0.00006, 0.00010]
	.8	0.0494 [0.04884, 0.04996]	0.0096 [0.00935, 0.00985]	0.00085 [0.00078, 0.00093]	0.000071 [0.00005, 0.00010]

Chapter 6

The Weighting Is The Hardest Part: On The Behavior of the Likelihood Ratio Test and Score Test Under Data-Driven Weighting Scheme in Rare Variant Association Studies

Abstract

Rare variant association studies are at a critical inflexion point with the increasing availability of exome-sequencing data. A popular test of association is the sequence kernel association test (SKAT). Weights are embedded within SKAT to reflect the hypothesized contribution of the variants to the trait variance. Because the true weights are generally unknown, and so are subject to misspecification, we examined the efficiency of a data-driven weighting scheme.

We propose the use of a set of theoretically defensible weighting schemes, of which, we assume, the one that gives the largest test statistic is likely to capture best the allele frequency-functional effect relationship. As both the score test and the likelihood ratio test (LRT) may be used in this context, and may differ in power, we characterize the behavior of both tests in our procedure.

We found that the powers of the two tests is equivalent when the weights in the set included the correct ones. However, when the weights are all misspecified, the LRT is expected to show superior power (due to its robustness to weight misspecification). With this procedure and the LRT we detected significant enrichment of rare case mutations (MAF<5%; P-value=7E-04) of a set of constrained genes in the Swedish schizophrenia case-control cohort with exome-sequencing data.

The score test is currently preferred for its computational efficiency and power. Indeed, assuming correct specification, in some circumstances the score test is the most powerful test. However, LRT has the compelling qualities of being generally more powerful and more robust to misspecification. This is an important result given that, arguably, misspecified models are likely to be the rule rather than the exception in weighting-based approaches.

6.1 Introduction

With the availability of high-coverage exome/genome sequence data in increasingly large samples, rare variant association studies (RVAS) are gaining importance in human genetic research. One important test of association between a target set of rare variants (RVs) and a given phenotype is the sequence kernel association test (SKAT; [64,176,196,210,211,312,367]). SKAT is based on a random effects model, in which the effect sizes of the RVs are assumed to be drawn from a zero mean distribution and a given variance. That the effect sizes are characterized by a single variance is a strong assumption which is made plausible by plausible weighting of effect sizes. The required weights are typically assigned based on meta-information about the RVs, such as allele frequency and functional predictions [191,227,276,367], with rarer and functional variants expected to have larger effects. Allele frequency, in particular, is an important weighting factor, as the rarer the variant is, the stronger the average purifying selection coefficient [277,292]. Accordingly, the effect sizes for rare variants will tend to be larger than for more common variants.

The relationship between effect size, frequency and selection, however, rests on assumptions about the extent of direct selection on the phenotype in question and the demographic history of the population [115,276,383]. More specifically, there are several postulates that have to hold for the frequency to be genuinely informative on the functional effect that a genetic variant has on the trait, namely: (a) the population under study has not experienced recent severe bottlenecks; (b) the selection on the trait of interest is direct, (c) strong (i.e., selection coefficient $s \geq 10^{-2.5}$); and, (d) it acts uniformly across the associated genes. Yet, for the reasons detailed below, the circumstances in which these postulates are expected to hold are rather special. First, population genetics theory predicts that the frequency of deleterious variants will vary with the size of the effect the associated trait has on fitness. For instance, risk variants implicated in early-onset diseases (e.g., autism) will be mostly rare, i.e., kept at low frequencies by selection pressures because of the high impact these diseases have on reproductive fitness (Manolio, Collins et al. [233]). In contrast, variants associated with a trait having a negligible effect on fitness (e.g., Alzheimer disease), will likely escape selection and so may occur at relatively high frequencies in the population (Zuk, Schaffner et al. [383]). Second, it should be noted that even if the trait of interest is under strong selection pressures, variants across the whole frequency spectrum may jointly contribute to disease risk, as simulation studies (Price, Kryukov et al. [276]) and empirical results (e.g., Cohen, Boerwinkle et al. [70], Teslovich, Musunuru et al. [318]) have demonstrated. Thirdly, allele frequency distribution is expected to vary as a function of the demographic history of the population. Using population genetics simulations, Zuk et al. [383] showed that given the same selection coefficient s, the frequency of deleterious alleles influencing a trait will depend on whether the population under study has encountered recent severe bottlenecks, and on mutation rate. For example, given strong selection pressures (i.e., $s > 10^{-2.5}$) acting directly on the phenotype, the median frequency of the associated alleles may vary from as high as 0.0377 in recently bottlenecked populations (e.g., Finland), to as low as 9.36E-005 in a large population with simple exponential expansion. Finally, the strength of selection is expected to vary across genes, and so will do the allele frequency-functional effect relationship (Price, Kryukov et al. [276], Zuk, Schaffner et al. [383]). Genes under weak selection will harbor both common and rare variants, both with functional effects, whereas functional variants within genes under strong selective constraints will mainly be rare. The examples above indicate that testing genomic regions by relying on a weighting scheme which up-weighs rarer variants and puts low or zero weights on the more common ones is optimal only in specific circumstances.

Because the true weights are generally unknown, and so are subject to misspecification, we examined the efficiency of a data-driven weighting scheme. We propose the use of a set of theoretically defensible weighting schemes, of which, we assume, the one that gives the largest test statistic is likely to capture best the allele frequency-functional effect relationship. The use of alternative weighting schemes is intended to accommodate genomic regions where only very rare variants are likely to be functional, as well as regions under weak selection pressures, harboring both rare and common variants, both (possibly) related to the risk of the disease of interest. Family-wise error rate can be protected either by using a multiple testing correction (e.g., the Bonferroni or Benjamini and Hochberg [29] methods), or by permutations. We show the power benefits conferred by the use of such a variable data-driven weighting procedure both in simulated and in empirical data. As both the score test [367] and the likelihood ratio test [211] may be used in this context, and may differ in power [377], we characterize the behavior of both tests in our procedure.

Below we first formulate the model and briefly consider the likelihood ratio test and the score test. We then present and evaluate the use of a data-driven weighting scheme in simulated and empirical data. Specifically, we evaluate the efficiency of the two tests under (a) the variable weighting scheme, relative to their efficiency under (b) incorrect, and (c) correct weighting. Finally, we discuss the robustness of the two tests to misspecification, and the power advantages conferred by our proposed weighting procedure in SKAT.

6.2 Material and Methods

6.2.1 Model formulation

Let $\mathbf{y}$ be the n-dimensional vector of continuous phenotypes measured in a sample consisting of n individuals. Let $\mathbf{X}$ be the $n \times p$ design matrix containing the relevant covariates. Let $\mathbf{G}$ be the $n \times m$ matrix of genotype values, with the g_{ij}

element denoting the genotype value of the individual i ($i = 1 \ldots n$) at locus j ($j = 1 \ldots m$). Genotypes are coded as additive-codominant, i.e., $g_{ij} = (0, 1, 2)$. The association between the phenotype and the set of m variants is modeled within the linear mixed model framework as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{b} + \mathbf{e} \tag{6.1}$$

with $\boldsymbol{\beta}^t = (\beta_1, \ldots \beta_p)$ being the p-dimensional vector of fixed effects of covariates, $\mathbf{b}^t = (b_1, \ldots, b_m)$ being the $m \times 1$ vector of regression coefficients in the regression of the phenotype on the m genetic variants within the target set, and $\mathbf{e}$ being the n-dimensional vector of random residuals. The random vectors $\mathbf{b}$ and $\mathbf{e}$ are assumed to be normally distributed: $\mathbf{b} \sim N(0, \mathbf{I}\sigma_b^2)$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, with $\mathbf{I}$ being the identity matrix of appropriate dimension.

Let $\mathbf{W}$ be the $m \times m$ diagonal matrix containing the weights used to weigh the contribution to the test statistic of the variants in the set. The normally distributed phenotype $\mathbf{y}$ has expected mean $\mathbf{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix:

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{E}[(\mathbf{y} - \mathbf{E}(\mathbf{y}))(\mathbf{y} - \mathbf{E}(\mathbf{y}))^t] = \mathbf{GWG}^t \frac{\sigma_b^2}{m} + \mathbf{I}\sigma_e^2 \tag{6.2}$$

with $\mathbf{GWG}^t$ being the weighted kernel or genetic relationship matrix. As implemented in the SKAT [367], the diagonal elements of the matrix $\mathbf{W}$, diag$(w_1 \ldots, w_m)$, are related to the minor allele frequency of the j-th variant by means of the beta density distribution function (dbeta), which is characterized by two shape parameters. The specification of the two shape parameters is informed by the hypothesized relationship between the j-th variant effect and its minor allele frequency (MAF; see section on 'Weighting' below).

6.2.2 Tests of variance components

To test whether the parameter of interest σ_b^2 deviates significantly from zero, one can employ a likelihood ratio test (LRT) or a score test. The likelihood ratio test is computed as two times the difference between the log-likelihoods of the null model (σ_b^2 constrained to equal 0) and the alternative model (σ_b^2 estimated freely). Parameter estimation can be performed by restricted/residual maximum likelihood (REML):

$$LogL(\sigma_b^2, \sigma_e^2) = \frac{1}{2}\log|\boldsymbol{\Sigma}_y| - \frac{1}{2}\log\left|\mathbf{X}^t\boldsymbol{\Sigma}_y^{-1}\mathbf{X}\right| - \frac{1}{2}r^t\boldsymbol{\Sigma}_y^{-1}r - \frac{1}{2}(n-p)\log(2\pi) \tag{6.3}$$

where $r = \mathbf{y} - \mathbf{X}(\mathbf{X}^t\boldsymbol{\Sigma}_y^{-1}\mathbf{X})^{-}\mathbf{X}^t\boldsymbol{\Sigma}_y^{-1}\mathbf{y}$ with superscript '$-$' denoting a generalized inverse [27].

In evaluating the statistical significance of the restricted LRT, we note that the null distribution of the test statistic is an equally weighted .5 : .5 mixture of a χ_0^2 and a χ_1^2 distributions (see e.g., [308, 349, 366]). Alternatively, the null distribution can be constructed empirically by using a permutation-based approach (e.g., [211]), or a parametric bootstrap (e.g., [79]).

The score test is computed as:

$$Q_{SKAT} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t \mathbf{G}\mathbf{W}\mathbf{G}^t (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{6.4}$$

with its expected null distribution following a mixture of chi-square distribution and statistical significance assessed by means of the Davies exact method [88].

6.2.3 Data simulation

Phenotypes and genotypes were generated in samples of $n = 10,000$ unrelated individuals. Specifically, we simulated two m-dimensional random vectors of continuous variables representing alleles at m equidistant loci for each individual i from the sample. The vectors were drawn from a multivariate distribution with zero mean and $\boldsymbol{\Sigma}_{LD}$ correlation matrix. As rare variants are expected to be in linkage equilibrium (see e.g., [89]), we set $\boldsymbol{\Sigma}_{LD}$ to equal an identity matrix. The multivariate normally distributed variables were then discretized given chosen thresholds based on the MAF at each locus. We considered MAFs varying randomly between 0.005 and 0.05, sampled from a uniform distribution. Given the vectors of alleles, we then created the m vectors of genotypes, g_{ij}. Based on the genotypes, the $n \times 1$ vector of phenotypes, $\mathbf{y}$, was generated as:

$$y_i = \sum_{j=1}^{m} g_{ij} b_j \times \sqrt{\sigma_b^2} + e_i \times \sqrt{\sigma_e^2} \tag{6.5}$$

b_j, the regression weight of the variant at the j-th locus, was computed as a function of MAF_j and of its contribution to the standardized variance of the polygenic scores [237]. Namely, the regression weights varied with MAF, while their contribution to the genetic variance was equal. Simulating data in this fashion is equivalent to simulation according to dbeta(MAF, .5,.5) weights [367], with weights increasing with decreasing MAF. We also simulated data according to dbeta (1,1) weights (second simulation scenario), where variants had equal weights regardless of MAF. This scenario is illustrative for situations where the tested region harbors both common and rare variants, both having functional effects on the trait (i.e., where there is no relationship between allele frequency and effect size). The variance σ_b^2 equaled 0.01 across all scenarios we considered, and $\sigma_e^2 = 1 - \sigma_b^2$. The n-dimensional vector of environmental scores $\mathbf{e}$ was drawn from a standard normal distribution $N(0, 1)$.

6.2.4 Data-Driven search for optimal weights: exploring the misspecification space

Because the application of a single weighting scheme might not be accurate when testing thousands of genes scattered across the whole exome, we evaluated the efficiency of a data-driven search for the optimal weights. We carried out simulations to evaluate the efficiency of the two tests under (a) the variable weighting scheme relative to their efficiency under (b) incorrect, and (c) correct weighting.

The m-dimensional vector $\mathbf{w}$ of variant weights was computed using the beta density function, with the j-th element calculated as $w_j = dbeta(MAF_j; a_1, a_2)$ given the MAF of the j-th variant and the shape parameters a_1 and a_2. As described in the previous section, data were simulated according to: a) dbeta(.5,.5) weights (i.e., the true weights increase with decreasing MAF); and b) dbeta(1,1) weights (i.e., the variants have equal weights, regardless of MAF). Next, in computing the tests statistic we (mis)specified the weights as: a) dbeta(1,1); b) dbeta(.5,.5); c) dbeta(1,25). The first weighting scheme pertains to the hypothesis that there is no relationship between the regression weight and the frequency of the variant (hence, the more common variants contribute on average more to variation in the phenotype). In this scenario the association test is carried out with raw additive-codominant coding of the genotypes. The use of the second weighting scheme is equivalent to standardization of the genotypic values prior to the analysis. We considered the effect of this weighting scheme as this treatment of the genotypes is default in GCTA [375] and in FaST-LMM-set [211]. Standardization and assignment of weights dbeta(.5,.5) are equivalent weighting schemes [367] in which the contribution to the test of rarer variants is up-weighed relative to that of the more common ones [302], and hence the variants contribute on average equally to the variance in the phenotype, regardless of frequency. We also considered the effects of the third weighting scheme (dbeta(1,25)) as weights computed as such are the default weights in SKAT [367].

We performed association tests by using the set of 3 weighting schemes, i.e., a) dbeta(1,25); b) dbeta(1,1), and c) dbeta(.5,.5). The p-value for the gene equaled the minimum Bonferroni corrected p-value minP$_{LRT}$ (minP$_{score}$) out of the 3 p-values obtained given the genotypes transformed according to each of the weighting schemes enumerated earlier. We also report the power of the tests under misspecified weights, as it is of interest to assess whether our procedure confers power gains relative to a test which uses a single set of (misspecified) weights (i.e., 3 tests vs. 1 test). We assessed the behavior of the two tests under the variable weighting schemes by considering: a) target regions harboring solely functional variants with opposite effects on the phenotypic mean, and b) regions harboring a mixture of protective, deleterious and neutral effects.

6.2.5 Evaluating the type I error rates and power

We evaluated the type I error rates by generating 1,000,000 datasets under the null hypothesis of no phenotypic variance explained by the variants within the target set. The type I error rate was computed as the proportion of datasets in which the tests incorrectly rejected the null hypothesis and it was evaluated given $\alpha = 0.01$ and 0.001. Following Visscher ([349] but see also e.g., Self and Liang [294], Blangero, Diego et al. [35]) we used a .5 : .5 mixture of a χ^2_0 and a χ^2_1 distribution for computing the p-value. We used this approach as this is default in most statistical software (e.g., in GCTA, Yang, Lee et al. [375] or in FaST-LMM-set, Listgarten, Lippert et al. [211]). Because this asymptotic distribution of the LRT is expected to be conservative (Crainiceanu and Ruppert [79], Blangero, Diego et al. [35]), we also report the p-value given varying mixing proportions of χ^2_0 and χ^2_1 distributions by considering increasing proportions of test statistics of zero (from .5 to .6, by .1).

Power was assessed based on 1000 simulated datasets, an effect size of 1% explained phenotypic variance and 7 alpha thresholds. Given the 7 alpha thresholds, power equaled the proportion of datasets in which the effect was detected. As a validity check of our program, for all the scenarios considered we also report the power and the type I error rates of the true (i.e., correct) model.

6.2.6 Software

The R-package MASS [340] was used for data generation. Model fitting was performed in R-nlme [269], and SKAT [195]. We used the anova function in R to obtain the restricted likelihood ratio test, with the p-value computed by halving the supplied p-value [271]. To check our model fitting approach, we analyzed one simulated sample of 10,000 individuals by using 3 independent programs implementing genetic similarity/kernel-based variance component tests: the nlme R-package, the software Genome-wide Complex Trait Analysis (GCTA; [375]) and the software FaST-LMM-set [211]. The values for the restricted LRT and the estimate for the variance component obtained by the 3 programs were almost identical (see Table 6.6 Supplementary Material for details), indicating that these are equivalent approaches. Having established the equivalence, all the simulations were next conducted using the nlme program. Simulations were carried out on the Broad Institute Gold Compute cluster.

6.2.7 Empirical analysis: testing the constrained and the FMRP-Darnell gene sets for rare case mutations enrichment

We compared the performance of the likelihood ratio test and of the score test under our proposed data-driven weighting scheme in a real dataset. For this illustration we used the Swedish schizophrenia case-control cohort of 4940 individuals with exome-sequencing data from blood DNA. Cases had a clinical diagnosis of schizophrenia and at least two hospitalizations as determined by expert review based on the Hospital Discharge Register [82, 190]. Controls, without a diagnosis of schizophrenia or bipolar disorder, were randomly selected from population registries. Both cases and controls are of Scandinavian ancestry, aged 18 or older (see [280, 284] for a detailed description of the sample). There were 169 individuals with unreliable samples (i.e., duplicates, ethnic outliers or having a genotype missing rate higher than 10%) whom we removed from the analysis. This left for the analysis 2461 cases and 2479 controls. 2732 of these were males. Written informed consent was obtained from all participants (or legal guardian consent and subject assent). All procedures were approved by the ethical committees in Sweden and in the United States.

Exome-sequencing was performed in seven waves at the Broad Institute of MIT and Harvard. For samples in the first wave, hybrid capture was performed using the Agilent SureSelect Human All Exon Kit method. In this version, the method targets ∼28 million base-pairs partitioned in ∼160,000 regions. Sequencing was done using Illumina GAII instruments. For samples in the waves two to seven, hybrid capture was done by using the newer version of the Agilent SureSelect Human All Exon v.2 Kit method, which targets ∼32 million base-pairs partitioned in ∼190,000 regions. Sequencing was performed using the Illumina HiSeq 2000 and HiSeq 2500 instruments. We used BWA ALN version 0.5.9 [200] to align the reads to the GRCh37 human genome reference and we applied Picard/GATK to process the sequence data and to call variants [241]. Selected singletons were validated using Sanger sequencing (see [280] for details). Variants out of Hardy-Weinberg equilibrium (P-value $<$ 5E-8) and showing excess heterozygosity, or variants showing excessive correlation (P-value $<$ 5E-8) with the covariates (that could not be explained by principal components) were excluded from the analysis. In addition, we excluded variants that did not pass the GATK default filters [25, 96]. There were 892,306 variants with MAF $<$ 5% meeting all our quality control criteria.

For this empirical illustration we considered the gene-sets rather than the genes as the unit of analysis. We extended the targeted region because the current sample sizes afford insufficient power for gene-based tests (see Purcell et al., [280]) but are more adequate for gene-set enrichment analyses which consider jointly a larger number of weak effects. This type of analysis has the added benefit of reducing substantially the burden of multiple testing. By extending the targeted region, the number of tested variants is large, and hence the effects of (possible) weight misspecification are expected to be large. In addition, as we do not focus on a specific class of alleles but rather lump together all observed variants with frequency below specific thresholds, a large amount of variation contributing to the test statistic will possibly be neutral. This makes the example a near optimal situation for illustrating the difference in robustness to both model misspecification and neutral variation of the LRT and the score test.

We tested for enrichment of case mutations two partially overlapping gene-sets likely relevant to schizophrenia. The first set consisted of 899 genes which are part of the list identified by Samocha et al. [289] as highly constrained. These constrained genes were proposed as candidates in autism spectrum disorder (ASD) given their enrichment for de novo loss of function case mutations. Given evidence favouring the hypothesis that schizophrenia and ASD share genetic aetiology [76, 130], this set of genes is likely to be relevant also to schizophrenia. The second set consisted of 749 genes targeted by the Fragile-X mental retardation protein (FMRP). This set is part of the list of genes derived by Darnell et al. [86] from mouse brain as likely implicated in regulating synaptic plasticity. Genes targeted by FMRP were found to be enriched for de novo nonsynonimous case mutations in both ASD [177] and schizophrenia [130]. Purcell et al. [280] also tested the FMRP set for enrichment of rare variants in the current sample, and their analysis yielded nominally significant results. Note that the strategy we adopted here is however, different. That is, rather than using gene-based statistic, our procedure tests for the joint effect (variance explained) of rare variants with MAF lower than 5% and 1% within the gene-set (note that the MAF thresholds are, however, arbitrary: variants defined as rare in one sample might feature as common in another sample).

We performed sequence-based kernel association analyses using the likelihood ratio and score tests with variable weights. For this empirical analysis we used the FaST-LMM-Set software [211]. To adjust for ancestry we included into analysis the first two principal components. Principal components were computed from genotypes at variants shared with the 1000 Genomes Project phase 1 dataset. To accommodate the scenario in which only very rare variants are likely to be functional, as well as the scenario in which the targeted region is under weak selection pressures, harboring both rare and more common variants, both (possibly) related to the risk of disease (regardless of frequency), we used three alternative weighting schemes: dbeta(1,25), dbeta(.5,.5) and dbeta(1,1).

For each tested pathway, we chose the Bonferroni corrected p-value corresponding to the weighting scheme that yields the largest test statistic. An alpha of 0.05 was used, corrected for multiple hypothesis testing of 2 gene-sets, 2 frequency thresholds and 3 weighting schemes (note that the Benjamini and Hochberg method (Benjamini and Hochberg [29] is a less conservative alternative method to correct for multiple comparisons, and so it can be alternatively employed). For computational ease we used a linear model [211]. The linear LRT (and the linear score test) shows good control of the type I error rate and has performed as well as a generalized linear model in case-control samples (see [210]).

6.3 Results

6.3.1 Type I error

Tables 6.1 and 6.3 contain the results pertaining to the type I error rates of the two tests, given correct and incorrect model specification. Across all conditions evaluated here, the score test shows good control of the type I error rate.

The .5 : .5 mixture of a χ^2_0 and a χ^2_1 asymptotic distribution of the likelihood-ratio test is slightly conservative, regardless of whether the weights are correctly specified or misspecified; that is, the .5 weight on the component underestimates to proportion of test statistics of zero, yielding a conservative test. A similar result was reported by (Crainiceanu and Ruppert [79]) and by Listgarten et al. (Listgarten, Lippert et al. [211]). We used this approach in the simulations as this is default in most statistical software (e.g., in GCTA, Yang, Lee et al. [375] and also in FaST-LMM-set, Listgarten, Lippert et al. [211]). Alternatively, similar to (Blangero, Diego et al. [35]) we note that when using a .57 : .43 mixture of a χ^2_0 and a χ^2_1, the type I error rate follows the expectation.

Accurate estimation of the weights for the mixing proportions is desirable, although such approaches rely on permutations, bootstrap or simulations, and hence, are more intensive computationally (see e.g., Greven, Crainiceanu et al. [148], Blangero, Diego et al. [35], Listgarten, Lippert et al. [211]). Listgarten et al. [211] proposed a permutation based approach to construct the null distribution of the test statistic, approach that maintains the type I error rate of the restricted LRT closer to the expectation. This is the approach we used in the empirical analysis.

6.3.2 Power

Figure 6.1 and Figure 6.2 display the results relating to power. Five important conclusions follow from our simulation results. First, the restricted LRT and the score test have equal power under correct weight specification. This is expected, as the two tests are asymptotically equivalent when the model is true, i.e., correctly specified (e.g., [147]). The powers of the two tests – displayed in grey in the power figures – are indistinguishable when the assigned weights correspond to the true weights.

Second, misspecification of weights always reduces power. This is shown in Figure 6.1 and in Figure 6.2, as the departure of the power under model misspecification (the colored lines) from the power of the true models (the grey lines). The exact loss in power depends on the degree of weight misspecification and on the statistical test employed. We note that the power loss is relatively small given mild misspecification of weights. This result is illustrated in Figure 6.1A (LTR) and 1B (score), where the assigned weights dbeta(1,25) resemble the true weights dbeta(.5,.5). In this circumstance, it is mainly the presence of neutral

Table 6.1: Type I error for the restricted likelihood ratio test (LRT) and the score test, given genotypic data simulated under the null model of no association between the target region and the phenotype. The sample consisted of 10,000 individuals with genotypes at 50 variants having minor allele frequencies (MAFs) sampled from the uniform distribution and ranging from .5% to 5%. The restricted LRT and the score tests were computed for three sets of weights beta in each of the 1,000,000 simulated samples. Type I error equals the proportion of datasets in which the null hypothesis has been incorrectly rejected given the three significance thresholds. For the LRT we report the results given varying mixing proportions of χ^2_0 and χ^2_1 distributions.

	weights dbeta	alpha=0.01	alpha=0.001
Score	(.5,.5)	0.0099	0.0009
	(1,1)	0.0099	0.0009
	(1,25)	0.0098	0.0009
LRT χ^2_0 and χ^2_1 mixing proportions			
.6:.4	(.5,.5)	0.010660	0.001032
	(1,1)	0.010347	0.001007
	(1,25)	0.010575	0.001007
.59:.41	(.5,.5)	0.010408	0.001012
	(1,1)	0.010106	0.000989
	(1,25)	0.010313	0.000978
.58:.42	(.5,.5)	0.010167	0.000989
	(1,1)	0.009884	0.000963
	(1,25)	0.010084	0.000958
.57:.43	(.5,.5)	0.009934	0.000973
	(1,1)	0.009668	0.000934
	(1,25)	0.009858	0.000937
.56:.44	(.5,.5)	0.009695	0.000945
	(1,1)	0.009458	0.000916
	(1,25)	0.009620	0.000913
.55:.45	(.5,.5)	0.009467	0.000918
	(1,1)	0.009278	0.000889
	(1,25)	0.009412	0.000900

Continued in Table 6.2

Table 6.2: *Continued from Table 6.1*

	weights dbeta	alpha=0.01	alpha=0.001
.54:.46	(.5,.5)	0.009229	0.000893
	(1,1)	0.009061	0.000874
	(1,25)	0.009215	0.000884
.53:.47	(.5,.5)	0.009020	0.000876
	(1,1)	0.008871	0.000853
	(1,25)	0.009002	0.000865
.52:.48	(.5,.5)	0.008847	0.000857
	(1,1)	0.008700	0.000842
	(1,25)	0.008825	0.000851
.51:.49	(.5,.5)	0.008676	0.000840
	(1,1)	0.008527	0.000824
	(1,25)	0.008652	0.000835
.5:.5	(.5,.5)	0.008492	0.000828
	(1,1)	0.008342	0.000803
	(1,25)	0.008472	0.000821

variants in the target that dilutes the power (see Figure 6.1C and Figure 6.1D). However, the power may suffer dramatically with increasing misspecification. For instance, when data were simulated according to the dbeta(.5,.5) weights, using a dbeta (1,1) weighting scheme (equal weights assigned to all variants) results in a loss in power of up to ∼5% and ∼30% for the restricted LRT and for the score test, respectively (see Figures 6.1). This result is informative for RVASs in which the raw genotypes (unweighted) are used in the test of association. A more dramatic power loss is illustrated in Figure 6.2 where we consider the reverse situation: weights dbeta (.5,.5) are assigned to variants simulated under flat weights. That is, in this scenario, the allele frequency is incorrectly used to inform on the weights assignment. With this misspecification the drop in power relative to the true model is ∼17% and ∼80% for the restricted LRT and for the score test, respectively.

Third, the inclusion of neutral variants dilutes the power of both tests. In our examples, with 40% neutral variants the power drops are in the range of ∼10%-∼17% relative to the power of the true model, regardless of the degree of weight misspecification. Clearly, discarding neutral variation present within the target is beneficial to improve power to detect significant association.

Forth, relative to the score test, we note that the restricted LRT is consistently

Table 6.3: Type I error for the restricted likelihood ratio test (LRT) and the score test, given genotypic data simulated under the null model of no association between the target region and the phenotype. The sample consisted of 10,000 individuals with genotypes at 50 variants having equal beta weights and minor allele frequencies (MAFs) sampled from the uniform distribution and ranging from .5% to 5%. The LRT and the score tests were computed for three sets of weights beta in each of the 1,000,000 simulated samples. The type I error equals the percentage of datasets for which the null hypothesis has been incorrectly rejected, given the three significance thresholds. For the LRT we report the results given varying mixing proportions of χ_0^2 and χ_1^2 distributions.

	weights dbeta	alpha=0.01	alpha=0.001
Score	(.5,.5)	0.0098	0.0009
	(1,1)	0.0098	0.0009
	(1,25)	0.0099	0.0009
LRT χ_0^2 and χ_1^2 mixing proportions			
.6:.4	(.5,.5)	0.010590	0.001024
	(1,1)	0.010587	0.001023
	(1,25)	0.010220	0.000942
.59:.41	(.5,.5)	0.010315	0.000999
	(1,1)	0.010319	0.001001
	(1,25)	0.009964	0.000911
.58:.42	(.5,.5)	0.010039	0.000979
	(1,1)	0.010072	0.000976
	(1,25)	0.009733	0.000892
.57:.43	(.5,.5)	0.009809	0.000954
	(1,1)	0.009789	0.000961
	(1,25)	0.009523	0.000869
.56:.44	(.5,.5)	0.009565	0.000935
	(1,1)	0.009557	0.000938
	(1,25)	0.009304	0.000845
.55:.45	(.5,.5)	0.009356	0.000902
	(1,1)	0.009331	0.000911
	(1,25)	0.009112	0.000825

Continued in Table 6.4

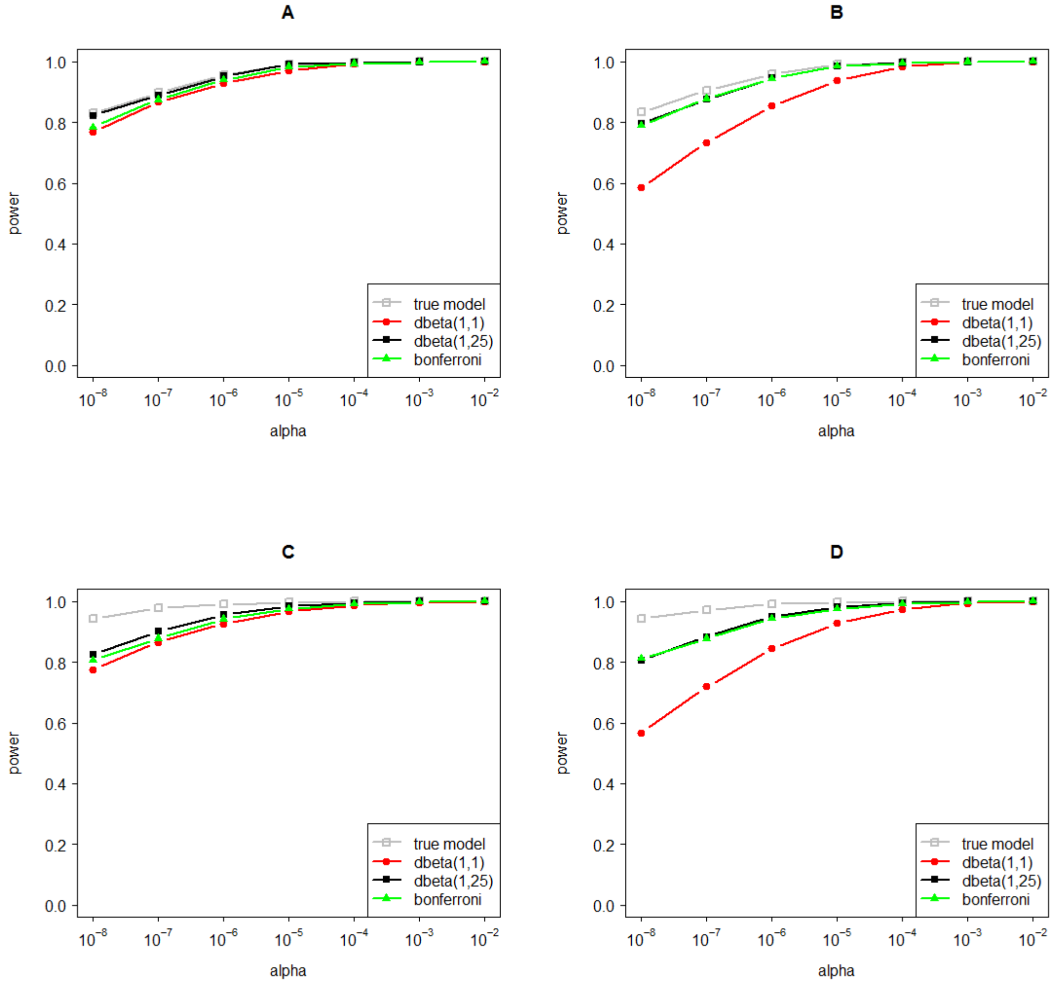
Table 6.4: *Continued from Table 6.3*

	weights dbeta	alpha=0.01	alpha=0.001
.54:.46	(.5,.5)	0.009161	0.000882
	(1,1)	0.009143	0.000885
	(1,25)	0.008918	0.000807
.53:.47	(.5,.5)	0.008955	0.000858
	(1,1)	0.008949	0.000860
	(1,25)	0.008734	0.000794
.52:.48	(.5,.5)	0.008778	0.000843
	(1,1)	0.008767	0.000844
	(1,25)	0.008548	0.000778
.51:.49	(.5,.5)	0.008610	0.000828
	(1,1)	0.008594	0.000821
	(1,25)	0.008347	0.000758
.5:.5	(.5,.5)	0.008445	0.000804
	(1,1)	0.008442	0.000801
	(1,25)	0.008172	0.000741

more robust, both to weight misspecification and to the presence of neutral variation in the target region. These results are consistent with those reported by Zeng et al. [377] and by Lippert et al. [210], who found their proposed LRT to be generally more powerful than the score test across their simulated settings. Although Lippert et al. did not consider the behavior of the two tests under misspecified weights, they reported the same pattern of results in real data analysis, where the LRT yielded consistently more associations than the score test. As the real weights are in all likelihood not known, the superior power of the restricted LRT in real data might be explained as well by its robustness to weight misspecification and to the inclusion of weighed neutral variation in the computation of the test statistic.

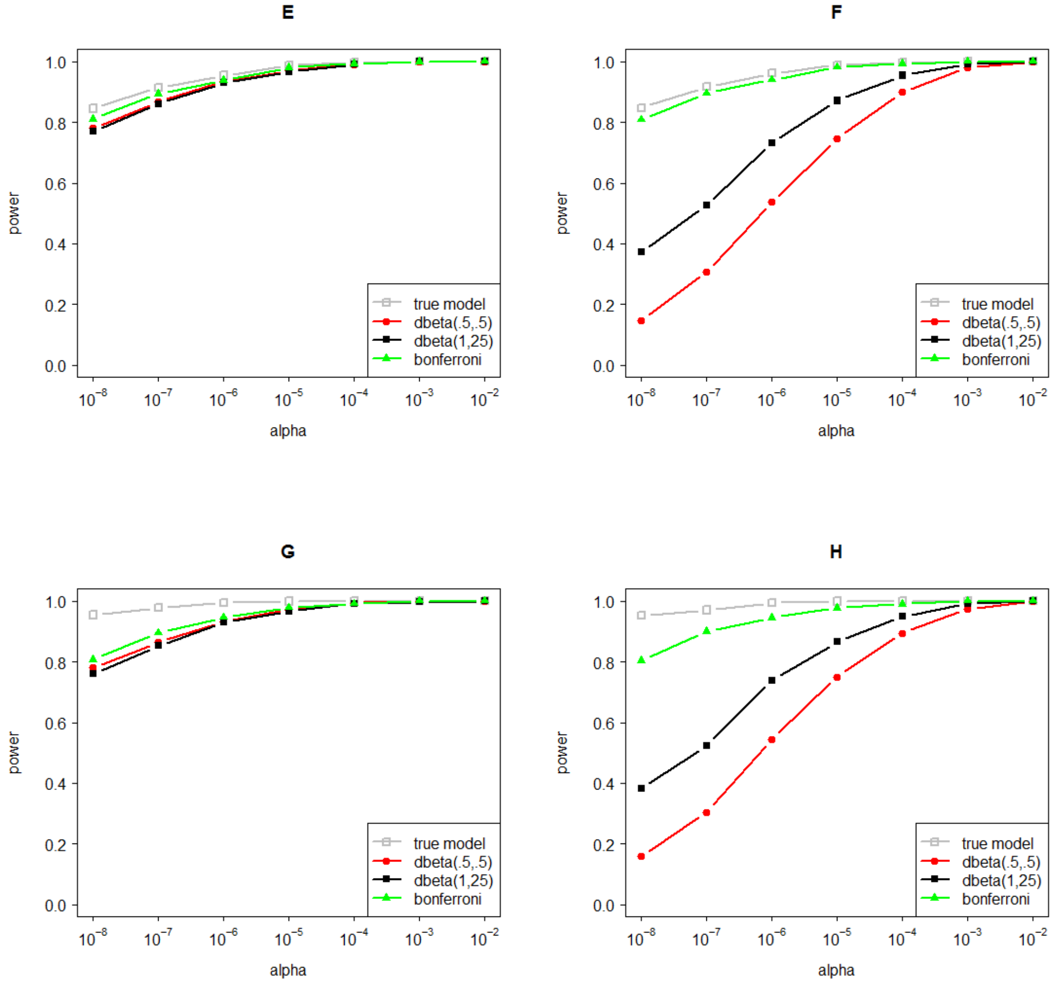
We note that both tests appear to benefit from the use of variable weights. The data-driven search for optimal weights confers power advantages over a model that uses misspecified weights, and maintains the power close to that afforded by a correctly specified model. It should be noted, however, that there is a price to pay in terms of power by using this data-driven weighting scheme in contrast to correct weighting. The price is largest for regions containing mixtures of functional and neutral variants (e.g., the power of both tests decreases from ∼94% given correct weights, to about ∼80% with the data-driven weighting approach;

Figure 6.1: The power of the likelihood ratio test (LRT; A and C) and the score test (B and D) to detect a gene harboring 50 low-frequency variants: all functional (A and B) or a mixture of 30 functional and 20 neutral variants (C and D). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(.5,.5). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases.



see scenarios displayed in Figures 6.1C and 6.1D), and relatively small for the (less realistic) scenarios in which the target set contains only functional variants (i.e., with both the LRT and the score and the data-driven weighting scheme, the

Figure 6.2: The power of the likelihood ratio test (LRT; E and G) and the score test (F and H) to detect a gene harboring 50 low-frequency variants: all functional (E and F) or a mixture of 30 functional and neutral variants (G and H). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(1,1). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases.



power drops about 4%). The two tests have equal powers with the Bonferroni corrected data-driven weighting procedure; it should be noted, however, that this is due to the fact that the correct weights were included in the procedure. Had the procedure included only misspecified weights, the power of the score test would have decreased dramatically relative to that of the LRT (which appears to be robust to misspecification). As typically the true weights are unknown, conjecturing the correct ones by employing alternative weights and using the likelihood ratio test appears to be the strategy likely to maintain the power close to that of the true model. This strategy appears to be advantageous also when the target region contains neutral variants.

6.3.3 Empirical analysis: testing the constrained and the FMRP-Darnell gene sets for rare case mutations enrichment

We also looked at the behavior of the score test and of the likelihood ratio test [211] under variable weights in the empirical dataset. Table 6.5 displays results pertaining to the enrichment tests in the gene-set-based analyses.

From Table 6.5 we note that the likelihood ratio test appears more powerful than the score test across all conditions evaluated here. It is likely the combination of weight misspecification coupled with the presence of neutral variation in the target set that yielded the difference in power between the two tests. With the current sample and the likelihood ratio test with weights dbeta(1,1), the set of constrained genes showed significant enrichment for disruptive case mutations with MAF below 5% (i.e., Bonferroni corrected P-value = 0.0084; see Table 6.5A). The score test under flat weights (i.e., dbeta(1,1)) with its associated p-value also passed the significance threshold, providing support for enrichment for disruptive rare case mutations of the constrained gene-set, although the evidence was weaker (Bonferroni corrected P-value = 0.037).

Note the difference in the strength of association of the two tests under variant weighting schemes. For instance, in the 5% MAF threshold analyses, the enrichment signal in the constrained gene-set was rendered non-significant when the dbeta(1,25) weights were used with the score test (Bonferroni corrected P-value = 0.397), and yet it reached statistical significance when the likelihood ratio test was employed instead (Bonferroni corrected P-value = 0.044). Had one relied on the score test and a default weighting scheme, the association signals in this pathway would have been missed.

The FMRP-Darnell gene-set showed no significant enrichment for rare case mutations, regardless of the test, MAF threshold and weighting schemes used. This result does not rule out the possibility that rarer variants (e.g., singletons) within the pathway play a role in the liability to schizophrenia phenotype. To implicate such variants, however, testing approaches other than those exploiting genetic similarity among the individuals are required.

The 1% MAF threshold yielded similar differences among the two tests (see Table 6.5B). Note that the signal in the constrained gene-set no longer reached

Table 6.5: Results of the gene-set enrichment analysis run in the Swedish sample (N = 4940; prevalence in the sample = 0.49). The 2 gene-sets included variants with MAF below 5% (A) or below 1% (B). *Bonferroni corrected P-values are given in italics.*

	Gene-set (autosome variants in set)	**weights** **dbeta**	**LRT**	**Score**
A.	constrained	(1,1)	**7E-04**	0.0031
	(63,492)		*(0.0084)*	*(0.037)*
		(.5,.5)	0.1240	0.3444
			(1)	*(1)*
		(1,25)	0.0037	0.0331
			(0.044)	*(0.397)*
	FMRP-Darnell	(1,1)	**0.0339**	0.0577
	(72,161)		*(0.406)*	*(0.692)*
		(.5,.5)	0.1062	0.3384
			(1)	*(1)*
		(1,25)	0.0434	0.1319
			(0.520)	*(1)*

	Gene-set (autosome variants in set)	**weights** **dbeta**	**LRT**	**Score**
B.	constrained	(1,1)	0.0373	0.1139
	(61,269)		*(0.447)*	*(1)*
		(.5,.5)	0.2341	0.3988
			(1)	*(1)*
		(1,25)	**0.0357**	0.1293
			(0.428)	*(1)*
	FMRP-Darnell	(1,1)	0.0723	0.1679
	(69,668)		*(0.867)*	*(1)*
		(.5,.5)	0.1467	0.3621
			(1)	*(1)*
		(1,25)	**0.0556**	0.1668
			(0.667)	*(1)*

statistical significance. This result suggests that imposing this threshold probably removed from the target causal variants and so, weakened the association signal.

Summarizing, the empirical analysis showed that the choice of the test and of the weighting scheme is no trivial matter. The LRT always yielded smaller p-values than the score test, probably due to the greater sensitivity the latter has to weighed neutral variation and to model misspecification (as we found in the simulated data). We also found that either thresholding or relying on default weights would trick one into missing association signals. We elaborate on these results in the Discussion.

6.4 Discussion

We considered the issue of optimizing weighting in association studies based on the rare variant sequence kernel test. Consistent with empirical [210] and simulation [377] results we found that the likelihood ratio test is generally more robust to weight misspecification, and more powerful than the score test in such a circumstance. The principal finding of this study is that using a weighting scheme that includes alternative weights is likely to boost the statistical power. Our results are of interest because weight assignment is embedded within any set-based test and the true weights of the variants within the target are generally unknown.

In the literature, weighting is mostly informed by allele frequency; frequency is taken as indicative of the strength of the purifying selection coefficient [191]. Accordingly, rarer variants are typically being assigned larger weights/contribution to the test statistic (e.g., [367]). This relationship between effect size, frequency and selection is not always straightforward, however, because it relies on assumptions about the extent of direct selection on the phenotype in question and the demographic history of the population [115, 276, 383]. Genes under weak selection may harbor rare as well as more common variants with disruptive effects [383]. Such variants with deleterious effects, escaping selection and occurring at relatively high frequencies in the population, are plausible also under strong purifying selection, as simulation studies have demonstrated [276]. Achieving maximal power when testing such regions requires adapting the weighting scheme to match the hypothesized selection. To this end, we proposed the use of a data-driven weighting approach. Our simulation results showed that such an approach maintains the power close to that of the true (i.e., correctly specified) model. When applied to real data, this approach allowed us to capture significant enrichment signal coming from variants with MAF below 5% within the constrained pathway [289]; Bonferroni corrected P-value = 0.0084), lending support to the conclusion that such a variable weighting approach is likely to boost statistical power. Such adaptive approaches were also recommended by Zuk et al. (2014) and by Price et al. (2010) as being optimal for gene-based tests (see also [206] and [197] for details on adaptive weighting schemes for burden tests). Deriving weights based on allele frequency is but one of the possible ways of prioritizing the contribution to the test statistic of the variants within the target set [367]. Alternative weighting schemes that incorporate probabilities of a variant being damaging (as estimated by annotation tools such as e.g., Polyphen-2 [6] or SIFT [258] may also be considered.

It should be emphasized that our data-driven weighting approach renders thresholding unnecessary. Thresholding (either based on counts or on allele frequency) has been initially used in burden tests (e.g., [199, 227, 276]; see also [129] for an overview on burden tests), but it has been employed also in sequence-based variance component tests (e.g., [217, 370]) for the purpose of removing neutral variation (see e.g., [191]). Yet, in our empirical analysis this practice was counterproductive: imposing the (arbitrarily chosen) 1% MAF threshold reduced the association signal in the constrained gene-set below the significance threshold. Considering common variants along with the rare ones in sequence-based kernel association tests appears to be justified for three main reasons. First, the use of variable weighting schemes is equivalent to applying variable frequency thresholds: the weights are removing from the test or favoring the contribution to the test statistic of the variants within the target set based on their frequency. Second, only the joint signal - coming from rare and more common variants - enabled us to detect significant enrichment. And third, importantly, with the current samples, our tests are mostly powered to locate regions under relatively weak selection pressures, and such regions are expected to harbour rare as well as common variants both with functional effects. To locate genes under stronger selection pressures, larger samples (see [383]) and the inclusion of more extreme weights (i.e., weights that overlook common variants and favour the rarer ones) will probably be required.

The data-driven weighting approach rendered equal the powers of the two tests. Note, however, that this equivalence hinged upon the inclusion of the correct weights among the alternatives. The powers of the two tests will likely diverge when the weights in the set are all misspecified; in such a circumstance, the LRT is expected to show superior power (due to its robustness to weight misspecification). This is likely illustrated in the empirical analysis where the LRT has always yielded lower p-values. Both in the simulations and in the empirical analysis we chose to correct out alpha by using the Bonferroni method. Alternatively, the less conservative Benjamini and Hochberg (BH; [29]) method may be employed (we refer to the Supplemental Figures 6.3 and 6.4, which are based on the BH-corrected results). P-value correction for larger number of tests can be easily obtained using the p.adjust function implemented in the stats R-package [315]. Permutation may also be used to compute the p-value. However, the data-driven weighting approach based on permutations is prohibitively slow when the number of tested variants within the target set (or the number of genes) and the sample is large. The Bonferroni correction though easier computationally, comes at a price in terms of power: the more weighting schemes one tries, the more stringent the significance threshold correction. An optimization algorithm for an optimal search for the true weights (e.g., [253] or limiting the choice of weights based on knowledge on theorized selection on each gene [383] would decrease the burden of multiple testing, and further increase power.

The score test is currently widely used in sequence-based association studies (e.g., [81, 172, 265, 378] for both its computational efficiency and power [367]. Indeed, assuming correct specification, in some circumstances the score test is the most powerful test [210, 367]. However, the results provided herein showed that the likelihood ratio test has the compelling qualities of being generally more robust and more powerful under weight misspecification. This is an important result, given that, arguably, misspecified models are likely to be the rule rather than the exception in the weighting-based approaches.

6.5 Supplemental Table and Figures

Table 6.6: Results of a test of association between a gene harboring 10 active variants (with a minor allele frequency ranging between 5% and .05% and explaining 1% of the phenotypic variance) and a continuous phenotype, in a simulated sample of 10,000 individuals. Data were simulated in R using the MASS package. Analyses were performed in 3 independent programs: the R-nlme package, the software Genome-wide Complex Trait Analysis (GCTA), and the software FaST-LMM-set. We report the log restricted likelihood under the null model (LL_0), the log restricted likelihood under the alternative model (LL_1), the chi-square test with 1 degree of freedom ($\chi^2(1)$), the variance attributable to the 10 genetic variants (V(G)).

Software	**LL$_0$**	**LL$_1$**	**χ^2(1)**	**V(G)**	**V(G)/V$_{\text{phenotype}}$**
GCTA	-4905.36	-4868.93	72.86	0.00929	0.0094
R-nlme	-14090.14	-14053.70	72.86	0.00929	0.0094
FaST-LMM-set	-14089.22	-14052.79	72.86	-	0.0094

Figure 6.3: The power of the likelihood ratio test (LRT; A and C) and the score test (B and D) to detect a gene harboring 50 low-frequency variants: all functional (A and B) or a mixture of 30 functional and 20 neutral variants (C and D). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(.5,.5). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases. *Abbreviation: BH – Benjamini and Hochberg correction.*

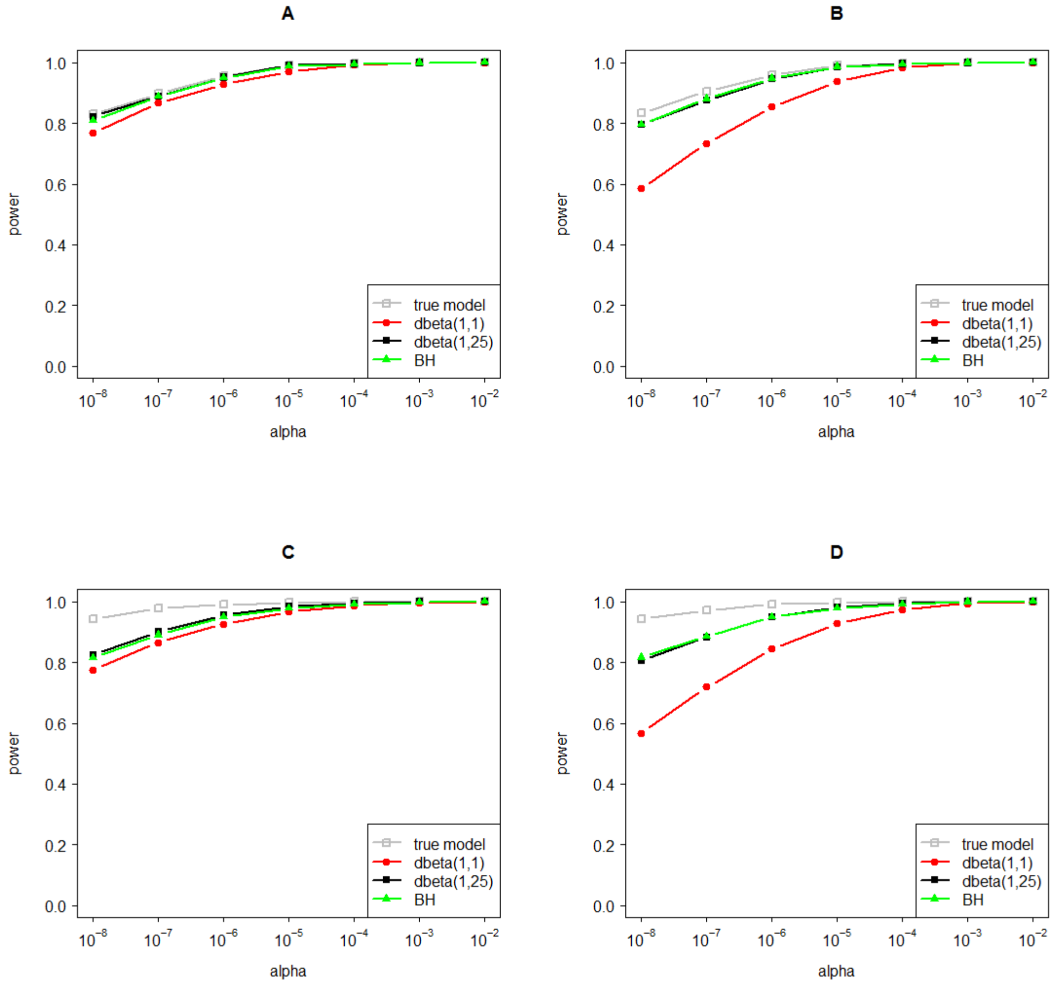
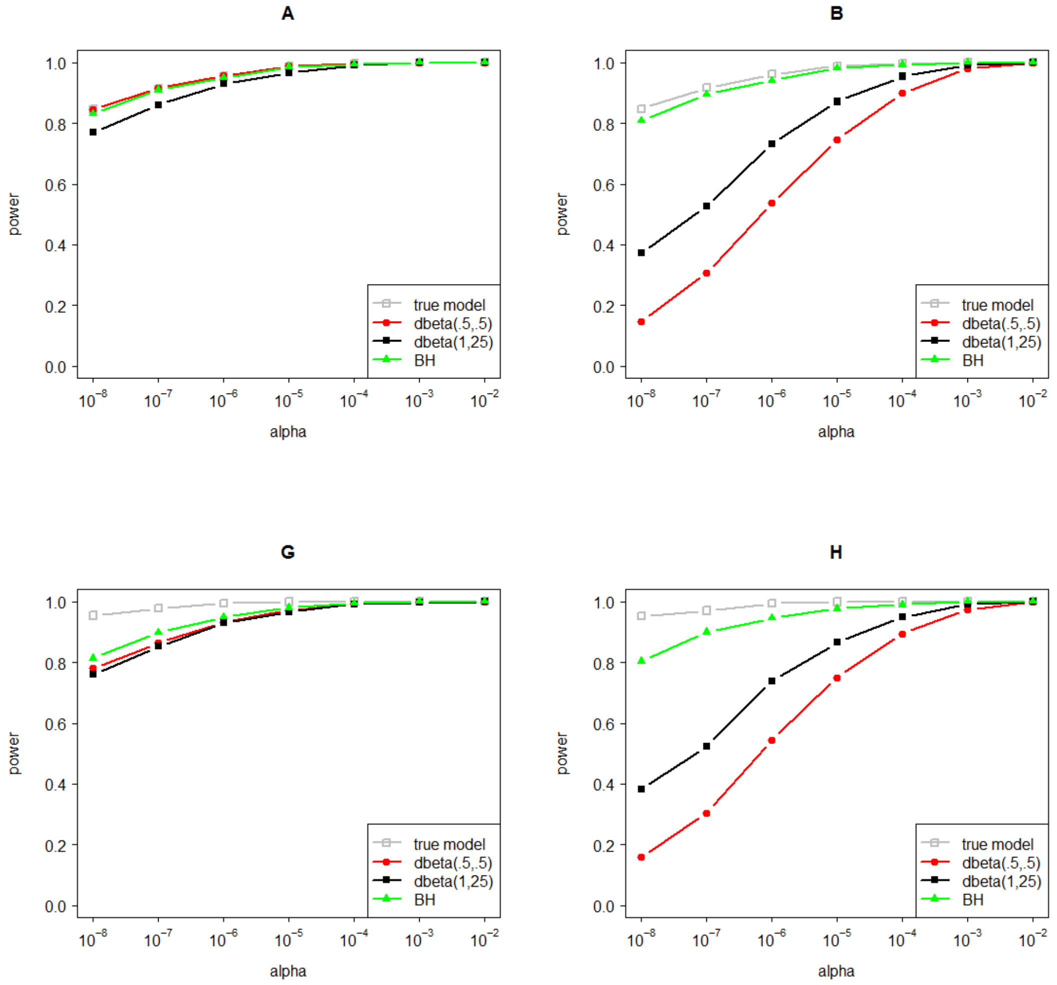


Figure 6.4: The power of the likelihood ratio test (LRT; E and G) and the score test (F and H) to detect a gene harboring 50 low-frequency variants: all functional (E and F) or a mixture of 30 functional and neutral variants (G and H). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(1,1). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases. *Abbreviation: BH – Benjamini and Hochberg correction.*



Part II

Applications

Chapter 7

Heritability, SNP- and Gene-Based Analyses of Cannabis Use Initiation and Age at Onset

Abstract

Prior searches for genetic variants implicated in initiation of cannabis use have been limited to common single nucleotide polymorphisms (SNP) typed in HapMap samples. Denser SNPs are now available with the completion of the 1000 Genomes and the Genome of the Netherlands projects. More densely distributed SNPs are expected to track the causal variants better. Therefore we extend the search for variants implicated in early stages of cannabis use to previously untagged common and low-frequency variants. We run heritability, SNP and gene-based analyses of initiation and age at onset. This is the first genome-wide study of age at onset to date.

Using GCTA and a sample of distantly related individuals from the Netherlands Twin Register, we estimated that the currently measured (and tagged) SNPs collectively explain 25% of the variance in initiation ($SE = 0.088$; $P = 0.0016$). Chromosomes 4 and 18, previously linked with cannabis use and other addiction phenotypes, account for the largest amount of variance in initiation (6.8%, SE = 0.025, P = 0.002 and 3.6%, SE = 0.01, P = 0.012, respectively). No individual SNP or gene-based test reached genomewide significance in the initiation or age at onset analyses.

Our study detected association signal in the currently measured SNPs. A comparison with prior SNP-heritability estimates suggests that at least part of the signal is likely coming from previously untyped common and low frequency variants. Our results do not rule out the contribution of rare variants of larger effect − a plausible source of the difference between the twin-based heritability estimate and that from GCTA. The causal variants are likely of very small effect (i.e., $< 1\%$ explained variance) and are uniformly distributed over the genome in proportion to chromosomes' length. Similar to other complex traits and diseases, detecting such small effects is to be expected in sufficiently large samples.

7.1 Introduction

Cannabis is among the drugs with the highest frequency of (ab)use. About 1 in 5 Europeans aged 15-64 reported to have experimented with cannabis. In the United States (US) the prevalence in ages 16-34 was estimated at 51.6% (European Monitoring Centre for Drugs and Drug Addiction, 2012). Regular cannabis use has been associated with health problems, including mood and anxiety disorders (e.g., Cheung et al. [67]) and chronic bronchitis (Hall [153]; Joshi et al. [178]). Early onset and regular use during adolescence has possible effects on cognitive functioning (e.g., Crean et al. [80]) and predicts diminished educational (Horwood et al. [163]; Lynskey and Hall [223]), and professional attainment (Fergusson and Boden [122]; Volkow et al. [356]). Furthermore, recent evidence suggests that high-potency cannabis use elevates the risk of developing psychotic disorders (Di Forti et al. [98]; Di Forti et al. [97]). Namely, the odds of showing psychotic symptoms in individuals who declared to have ever used high-potency cannabis are about three times larger than in individuals who declared to have never used cannabis during their lifetime. The risk of showing psychotic symptoms is further elevated if high-potency cannabis is used daily (i.e., $OR = 5.4$; $P = 0.002$; Di Forti et al. [97]). About 9% of those who initiate cannabis use progress to regular use and abuse (e.g. Volkow et al. [356]; Budney et al. [49]). Given the possible adverse effects on health and lifetime outcomes and given its possible role in triggering first-episode of psychosis, it is important to understand the causes of individual differences in the liability to initiate cannabis use.

Twin and family studies have shown that both genetic and environmental factors (both shared by, and specific to, family members) have an important role in the initiation of cannabis use (Kendler and Prescott [182]; van den Bree et al. [330]). A meta-analysis of twin studies (Verweij et al. [343]) showed that additive genetic factors explain nearly half the variance in liability to initiate cannabis use (i.e., 48% and 40% of the variance, in females and males, respectively), while the remaining variance is accounted for - almost equally - by shared and unshared environmental factors (both about 30%).

Among the several attempts to identify genes that explain the heritability of initiation, a linkage study (Agrawal et al. [14]) failed to identify statistically significant associated genomic regions, although it did identify several suggestive regions on chromosomes 18 and 1. Likewise, a meta-analysis by Verweij et al. (Verweij et al. [342]) combining the results of two genomewide association studies (GWAS) comprising about 10,000 individuals failed to detect common single nucleotide polymorphisms (SNPs) associated with initiation. It should be noted, however, that the association analysis by Verweij and colleagues was limited to common (i.e., minor allele frequency (MAF) > 5%) HapMap SNPs [75]. With the recent completion of large sequencing projects such as the 1000 Genomes (1000G) [77] and the Genome of the Netherlands (Boomsma et al. [44]; The Genome of the Netherlands [319]), more detailed genotypic information has become available in large GWAS samples. Given the availability of denser SNPs, which are expected to be in high linkage disequilibrium (LD) with the causal variants, we aim to extend the search for genetic variants (GVs) implicated in initiation to previously untagged common GVs, and to other (than common) GVs, such as low-frequency variants (1% < MAF < 5%). Such low frequency variants have not typically passed the quality control checks. However, the quality of imputation has been improved by recent advances in imputation techniques (Howie et al. [165]). This opens the door to including such genetic variants into a genome-wide association study (GWAS).

Furthermore, to date, the approach for finding genes underlying the heritability of cannabis initiation was to focus on the 'ever/never used' dichotomy at the expense of the age at which one initiates (i.e., age at onset). Yet, age at onset is a complex trait (Visscher et al. [355]), subject to the influences of both environmental and genetic factors (Lynskey et al. [224]), and may serve as an important proxy for heavy use. Initiation of cannabis use before age 18 is predictive of both experimentation with other drugs (Agrawal et al. [13]; Lynskey et al. [225]), and of escalated drug use (e.g., Lynskey et al. [224]). Among those initiating in adolescence the risk of progression to symptoms of abuse and dependence is higher relative to the general population (i.e., 17% vs. 9%, respectively; Volkow et al. [356]). Given its relevance as a predictor for escalated use, our second aim is to perform a genomewide search for GVs that give rise to individual differences in age at onset. To model age at onset as a function of genotype we will apply statistical methods based on survival analysis. This approach utilizes all available information on the age at onset among initiateds and takes into account the censored nature of the observations collected in those who did not initiate at the time they were last seen (i.e., they might initiate at a later time point). The approach is expected to show superior power relative to an analysis of the "ever-never" dichotomy or an analysis restricted to those who initiated (see e.g. Kiefer et al. [185]). To our knowledge, a genomewide survival analysis of age at onset of cannabis use has not yet been reported.

The outline of the paper is as follows. First, we estimate the amount of variance in initiation of cannabis use explained collectively by the currently measured SNPs. The purpose of such analysis is to obtain an indication of the total signal in the measured (and tagged) SNPs without identifying individual SNPs. Second, we conduct SNP-based association analyses of initiation and age at onset. Our primary focus is on identifying genes tagged by the SNPs, relevant to our traits. Therefore, next, we incorporate these SNP-based results in two gene-based analyses. These analyses are exploratory, i.e., conducted genomewide.

All analyses are performed in a sample of Dutch families from the Netherlands Twin Register (NTR). To maximize statistical power, imputation of genotypes in the NTR sample was based on two alternative reference panels: the 1000G Phase 1 project reference panel [77] and the reference panel generated by the Genome of the Netherlands (GoNL) project (Boomsma et al. [44]; The Genome of the Netherlands [319]). The GoNL reference panel was derived by sequencing the whole genome of 250 trio-Dutch families and matches therefore the ancestral background of our sample. The GoNL panel is expected to facilitate imputation of variants which are specific to the Dutch population (Boomsma et al. [44]). Furthermore, the use of the GoNL panel is expected to result in higher imputation accuracy relative to the 1000G panel, especially for low frequency GVs (MAF < 5%) (The Genome of the Netherlands [319]). Such increased accuracy is expected to increase the statistical power to capture the signal in the measured GVs.

7.2 Materials And Methods

7.2.1 Phenotypes

The phenotypic data were obtained in the longitudinal surveys on lifestyle, health, and personality of the NTR (e.g., Boomsma et al. [43]; Boomsma et al. [40]). The study protocols were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU Medical Center, Amsterdam. All participants provided informed consent. The study in young twins was approved also by the Central Committee on Research Involving Human Subjects. More details regarding the phenotyping in the NTR study can be found elsewhere (van Beijsterveldt et al. [329]; Willemsen et al. [362]).

7.2.2 Initiation of cannabis use ('ever/never')

Initiation was assessed by a multiple choice question (i.e., "At which age did you experiment with cannabis for the first time?") in the NTR surveys 1993, 1995, 2000, and by an open-ended question ("Have you ever tried hashish or cannabis? If yes, at which age?") in survey 2009. These surveys were sent to all adult twin families and were returned by 23,597 individuals. In addition, data collection in adolescent twins and sibs which took place since 1987 in age-specific surveys (around age 14 and age 16), included a multiple choice question ("Have you ever used soft drugs such as hashish or cannabis?") assessing frequency of use (on an eight-category scale ranging from 'never' to 'more than 40 times') in the whole life, in the last 12 months and in the last 4 weeks. This question was completed by 16,556 participants. The phenotypic data obtained from subjects who reported at more than one time point were checked for consistency, and unreliable measures were discarded. Due to inconsistencies, 284 self-reported measures were dropped. Next, the measurements were collapsed into a dichotomous phenotype (i.e., ever/never used cannabis). Furthermore, we included in the analysis only family members for whom both phenotypes and genotypes were available, i.e., N = 6744 participants. Of these, 5387 individuals reported never to have used cannabis, whereas the remaining 1357 individuals had initiated cannabis use. The age at the time of the last survey ranged from 10.5 to 94 years (mean age = 39.09, SD = 17.45). The participants were clustered within 3479 families varying in size from 1 to 9 family members (i.e., parents, siblings, spouses). More than half of the sample (60.9%) consisted of females.

7.2.3 Age at onset

A subset of the genotyped NTR sample (N = 5148) had declared never to have used cannabis, or declared an age at onset older than 10 years of age in survey 2009 (which included an open ended question on age at onset, see above). Among them, 852 (16.6%) had initiated cannabis use, whereas 4296 observations had not initiated at the time of data collection (i.e., censored observations). The participants were clustered within 2992 families of sizes varying from 1 to 8 members. Females represented 62.3% of the sample and the age ranged between 16 and 99 years (mean age = 46.93, SD = 17.54).

7.2.4 Genotypes

Genotyping was performed based on buccal or blood DNA samples collected in different research projects (see e.g., Willemsen et al. [361]). Imputation was performed based on the 1000G GIANT phase 1 panel as a first reference set, and on the GoNL version 4 as a second reference set (see Supplementary Methods for details). As best guess genotypes (computed using Beagle, Browning and Yu [48]) were used in the analyses, we applied stringent post imputation quality thresholds on the imputation quality measure (i.e., we retained only SNPs with an imputation quality score above 0.8) and for the Hardy-Weinberg equilibrium test ($\alpha = 10^{-4}$). Both the imputation quality and Hardy-Weinberg equilibrium (i.e., based on the summed genotype probability counts) were assessed in the phenotyped sample using SNPTEST (Marchini, 2007). The GoNL- and the 1000G-based imputed datasets contained ~6 million well-imputed SNPs (i.e., with a mean imputation quality score above 0.96 in both datasets). The association and survival analyses were carried-out by varying the reference panel used for imputation, while including the same phenotyped sample (i.e., 6744 and 5148 participants, respectively). The analyses included no monozygotic twin pairs, because genotypic data were available for only 1 twin of a pair in the GoNL dataset.

7.3 Statistical analyses

7.3.1 Estimating the heritability of initiation

We used the Genome-wide Complex Trait Analysis (GCTA) software (Yang et al. [374]) to estimate the amount of variance in initiation explained collectively by the SNPs. The aim of this analysis is to obtain an indication of the total signal in the SNPs, without identifying individual SNPs. Genetic similarity among the phenotyped individuals was computed based on best guess genotypes at 5,928,887 loci observed or imputed using the GoNL reference panel. The analyzed SNPs had a MAF larger than 1%, imputation quality greater than 0.8 and showed no significant deviation from Hardy-Weinberg equilibrium given $\alpha = 10^{-4}$. The sample with observed initiation status (N = 6744 related individuals of Dutch ancestry) and the relevant covariates included in the genomewide SNP-based analysis (see below) were also used in the GCTA analysis. Furthermore, one of a pair of closely genetically related individuals (i.e., with an estimated genetic relatedness larger than 0.025) was dropped, which left for the analysis 3616 distantly related individuals. We specified the prevalence as equal to 22%, value chosen in line with the prevalence of cannabis use estimated in Europeans (European Monitoring Centre for Drugs and Drug Addiction, 2012). Heritability of age at onset was not estimated as GCTA cannot handle survival data. We also investigated the relationship between chromosome length and the amount of variance explained in the trait. Consistent with the model of a polygenic trait, we expect – on average – the longer chromosomes to explain a larger amount of the variance. We tested this in a linear regression (one-tailed test) where we regressed the estimated proportion of variance explained by each chromosome on the chromosome length.

7.3.2 Power analysis

We performed a Monte Carlo power analysis to obtain an indication on the size of the genetic effects detectable in our sample. To this end, we simulated 10,000 samples consisting of 3690 families of various configurations reflecting the unbalanced structure of families included in the analyses, i.e., families consisting of singletons, two parents or families comprising sibships sizes 1 to 6 with 0, 1 or 2 parents. Genotypes in Hardy-Weinberg equilibrium were generated at a locus with a MAF of 0.5 and explaining 1.5% and 1% variance in the phenotype. The normally distributed phenotype was simulated conditional on the locus and then dichotomized using a cut-off point corresponding to a z-score of 0.85 to mimic the 20% prevalence of initiation observed in the NTR sample. The correlations between spouses, full siblings and parent-offspring estimated in our sample equaled 0.39, 0.35 and 0.15, respectively. An $\alpha = 10^{-8}$ was used to assess the power to detect association. To model association we used a Generalized Equations Estimation (GEE) procedure with an exchangeable working correlation matrix and a sandwich correction to correct the standard errors for misspecification of the background model (Minică et al. [245]). Empirical power analysis showed that our sample affords 45.3% and 87.4% power to detect GVs explaining 1% and 1.5% phenotypic variance, respectively (genomewide $\alpha = 10^{-8}$). These power computations are informative also for the age at onset phenotype given the large overlap among the samples included in the two analyses.

7.3.3 SNP-based association analysis of initiation

To test association, initiation was regressed on the best guess genotype and covariates. The covariates were sex, age at the last survey, the birth cohort (i.e., two birth cohorts containing individuals born between 1951-1970 and 1971-1999, respectively, and the 1915-1950 birth cohort as the reference category), 3 principal components to correct for Dutch population substructure (Abdellaoui et al. [1]), and sample specific covariates to account for batch and for chip effects. A GEE (Carey et al. [58]) logistic model was employed. To model the familial relatedness, we used an exchangeable working correlation matrix. This accounts for the familial correlations by means of a single correlation among the family members. The effect of possible misspecification of the familial covariances on the standard errors was corrected by means of a sandwich correction (Minică et al. [245]; Dobson [102]). The sandwich-corrected GEE approach was implemented by using the R-package gee (Carey et al. [58]), accessed from Plink (Purcell et al. [279]) which communicates with R [317] via the Rserve package (Urbanek [328]).

7.3.4 SNP-based survival analysis of age at onset

A Cox proportional hazards regression model was employed to model age at onset as a function of genotype and – as above – of other relevant covariates (i.e., birth cohort, sex, 3 PCs and study specific covariates). We included this approach as it utilizes all available information on the age of initiation among those who have initiated. It is expected to show superior power relative to an analysis of the "ever-never" dichotomy or an analysis restricted to those who initiated (see e.g. Kiefer et al. [185]). The Cox proportional hazard regression analysis was performed genomewide by accessing the survival R-package (Therneau [320]) from Plink. In fitting the model, we used the cluster option to get sandwich corrected standard errors that are robust to misspecification of the familial covariance matrix.

7.3.5 Gene-based analyses of initiation and age at onset

Gene-based tests of association with initiation and age at onset were carried out by using the gene-based association test that employs the extended Simes procedure (GATES) implemented in the Knowledge Based Mining System for Genome-wide Genetic Studies software (Li et al., 2011). Specifically, the Simes test extension was employed to combine the P-values of SNPs belonging to the same gene. SNPs were assigned to genes (or to genes' vicinity, i.e., within a region extended 5 kb at both the $5'$ and at the $3'$ ends) according to the Human Genome version 19 references. The linkage disequilibrium structure was derived based on the GoNL haplotypes and incorporated into the gene-based test as to account for the correlatedness among SNPs within a gene. Lacking prior significant genetic association information related to the cannabis use phenotypes, SNPs were assigned equal weights in the estimation process and the gene-based tests were conducted genomewide for both phenotypes. There were 22,764 genes tested for association with our phenotypes, hence for the gene-based tests the chosen alpha level equaled 0.01/22,746 (i.e., $\sim 4.3 \cdot 10^{-7}$).

7.4 Results

7.4.1 Estimating heritability based on genetic relatedness

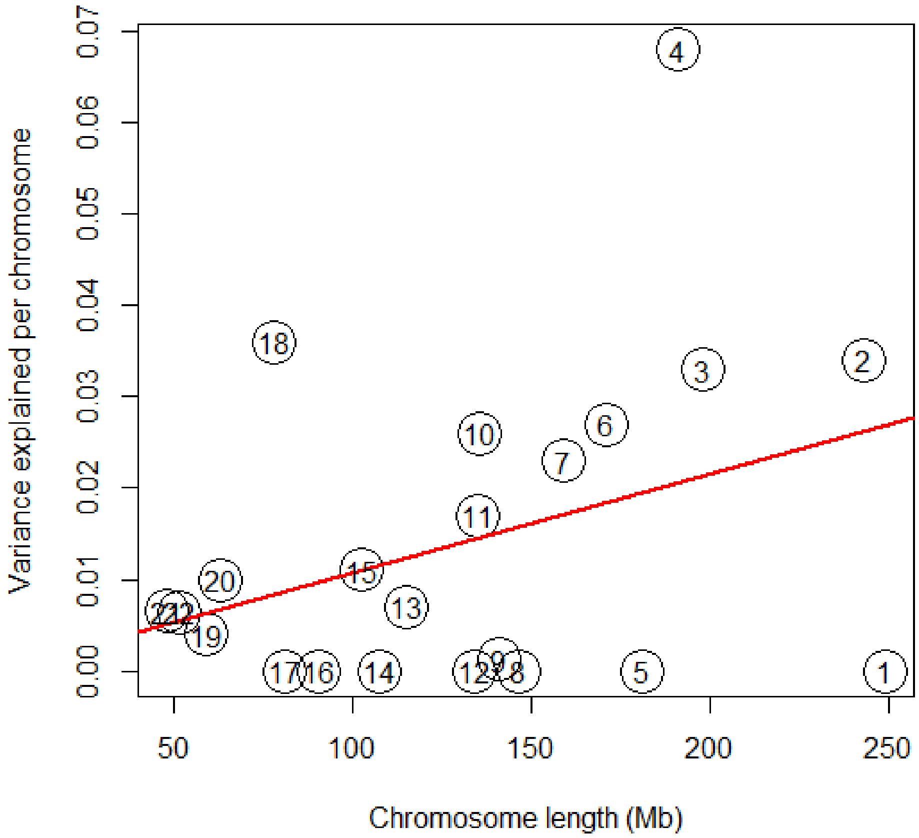
Results indicate that 25% (standard error (SE) = 0.088) of the variance on the observed scale in initiation is explained by the SNPs. This amount of variance explained collectively by the SNPs is significantly greater than zero (i.e., likelihood ratio test (LRT) (degrees of freedom = 1) = 8.60, P = 0.0016). The chromosome-by-chromosome heritability analysis indicated that the largest amount of variance in the trait is explained by chromosome 4 (i.e., the estimate on the observed scale equaled 6.8%, SE = 0.025, LRT(1) = 7.93, P = 0.002). Chromosome 18 accounted for about 3.6% (SE = 0.01, LRT(1) = 4.99, P = 0.012) of the variance on the observed scale in initiation. We also investigated the relationship between chromosome length and the amount of variance explained (see Supplemental Table 6.6 for details). We found that chromosome length is significantly associated with proportion of variance explained (one-tailed t-test(20) = 1.731, $P < 0.05$). On average longer chromosomes explain a larger percent of variance (Figure 8.1).

As shown in Figure 8.1, the linear trend is present, notwithstanding the low power to detect variance components attributable to individual chromosomes. The figure demonstrates a trend that is likely to be stronger with increasing sample size. Some parameter estimates hit the lower bound of zero, but this is due to sampling fluctuation (as we illustrate in a small simulation study described in the Supplementary notes). Similar results were reported for other complex traits like intelligence (see e.g., Davies et al. [87]).

7.4.2 SNP- and gene-based analyses of initiation

SNP-based P-values were obtained in two association analyses of initiation conducted in a sample comprising of 6744 participants. Two alternative reference panels – the 1000G and the GoNL, respectively – were used to impute geno-

Figure 7.1: Percent of variance in initiation of cannabis use explained per chromosome relative to chromosome length. The chromosome number is shown in circles.



types in our sample. Owing to a better imputation quality (The Genome of the Netherlands [319]), the association signals in the GoNL imputed genotype data were slightly stronger than those obtained based on the 1000G imputed SNPs. Consequently we took forward these results for the gene-based tests. The P-values for the 5,896,100 GoNL SNPs showed no inflation i.e., the lambda inflation factor equaled 1.019, where a value of 1 indicates no deviation from the expectation of the observed test statistic due to effects of population stratification. The quantile-quantile plot is given in Supplemental Figure S2. The most strongly associated SNP was the low frequency GoNL SNP rs35917943 (MAF $< 5\%$; P $= 1.6 \cdot 10^{-7}$). The region harboring this SNP is displayed in Supplemental Figure S3 (Pruim et al. [278]). Supplemental Table S2 contains the top SNPs associated with initiation at P $< 10^{-5}$. Table 7.1 contains the five genes showing the strongest association signal with initiation along with their functions (according to Gene

Ontology (GO) annotations, Ashburner et al. [23]).

Table 7.1: Top five genes showing the strongest association with initiation of cannabis use.

Gene Name (Gene Id)	Chr	Start Position	# SNPs Assigned to Gene	Key SNPs Position (rs number)	Gene Feature	Key SNPs P-Value	Gene P-Value	Molecular Function according to Gene Ontology Annotation
Zinc Finger Protein 181 (*ZNF181*)	19	35225479	2	35221228 (35487050)	upstream	$1.6\cdot10^{-7}$	$3.7\cdot10^{-6}$	nucleic acid binding; metal ion binding;
microRNA 643 (*MIR643*)	19	52785049	10	52787471 (2434422)	intronic	$3.7\cdot10^{-6}$	$3\cdot10^{-5}$	-
				52788044 (321908)	intronic	$8.5\cdot10^{-6}$	-	
Zinc Finger Protein 766 (*ZNF766*)	19	52772823	41	52787471 (2434422)	intronic	$3.7\cdot10^{-6}$	$1.1\cdot10^{-4}$	nucleic acid binding; metal ion binding;
				52788044 (321908)	intronic	$8.5\cdot10^{-6}$	-	
				52770905 (57523152)	upstream	$3.3\cdot10^{-5}$	-	
				52790542 (139570481)	intronic	$2.3\cdot10^{-4}$	-	
				52792311 (147711278)	intronic	$3.4\cdot10^{-4}$	-	
				52775301 (2089275)	intronic	$1\cdot10^{-2}$	-	

Continued in Table 7.2.

None of these genes had an association P-value below our chosen genomewide level of significance of α=$4.3\cdot10^{-7}$. The three genes with the lowest P-values are Zinc Finger Protein 181 (*ZNF181*, P = $3.7\cdot10^{-6}$), the non-coding RNA - microRNA 643 (*MIR643*, P = $3\cdot10^{-5}$) and the Zinc Finger Protein 766 gene (*ZNF766*, $1.1\cdot10^{-4}$), all located on chromosome 19.

Table 7.2: *Continued from Table 7.1*

Gene Name (Gene Id)	Chr	Start Position	# SNPs Assigned to Gene	Key SNPs Position (rs number)	Gene Feature	Key SNPs P-Value	Gene P-Value	Molecular Function according to Gene Ontology Annotation
Phosphatidylinositol-specific Phospholipase C, X Domain containing 2 (*PLCXD2*)	3	111393522	60	111416310 (1355767)	intronic	$1.1{\cdot}10^{-6}$	$1.1{\cdot}10^{-4}$	phosphoric diester hydrolase activity;
				111399209 (7651713)	intronic	$1.2{\cdot}10^{-6}$	-	
				111460129 (7651713)	intronic	$1.3{\cdot}10^{-2}$	-	
				111430969 (16858448)	intronic	$1.5{\cdot}10^{-2}$	-	
				111438443 (12637233)	intronic	$1.5{\cdot}10^{-2}$	-	
				111479048 (7643067)	intronic	$1.6{\cdot}10^{-2}$	-	
				111470751 (74571144)	intronic	$1.6{\cdot}10^{-2}$	-	
				111463864 (75923425)	intronic	$1.6{\cdot}10^{-2}$	-	
				111453629 (4682300)	intronic	$1.8{\cdot}10^{-2}$	-	
				111530499 (138770435)	intronic	$2.7{\cdot}10^{-2}$	-	
				111482694 (139568104)	intronic	$3{\cdot}10^{-2}$	-	
				111443003 (9854875)	intronic	$3.2{\cdot}10^{-2}$	-	
				111449944 (7624162)	intronic	$3.2{\cdot}10^{-2}$	-	
				111514564 (11715999)	intronic	$4{\cdot}10^{-2}$	-	

Continued in Table 7.3.

Table 7.3: *Continued from Table 7.2*

Gene Name (Gene Id)	Chr	Start Position	# SNPs Assigned to Gene	Key SNPs Position (rs number)	Gene Feature	Key SNPs P-Value	Gene P-Value	Molecular Function according to Gene Ontology Annotation
Prefoldin-like chaperone (*URI1*)	19	30433145	15	30511638 (57192507)	downstream	$2.2{\cdot}10^{-5}$	$1.8{\cdot}10^{-4}$	unfolded protein binding;
				30465196 (7249169)	intronic	$2.7{\cdot}10^{-5}$	-	
				30509036 (73924148)	downstream	$2.7{\cdot}10^{-5}$	-	
				30442432 (77858500)	intronic	$3.1{\cdot}10^{-5}$	-	
				30432202 (58563661)	intronic	$1.1{\cdot}10^{-4}$	-	
				30418009 (61340893)	intronic	$2.9{\cdot}10^{-2}$	-	

7.4.3 SNP- and gene-based analyses of age at onset

We conducted two genomewide survival analyses of age at onset in a sample comprising 5148 participants. Similar to the previous analysis, the association signals attained with the genotypes imputed based on the GoNL reference panel were used as input for the gene-based analysis, as these signals were stronger relative to those observed in the 1000G imputed sample (see for a comparison the Manhattan plots, Supplemental Figure S4). As we observed a slight inflation, we corrected the SNP-based P-values (genomic control $\lambda_{GC} = 1.1171$) to prevent potential false positives. Supplemental Figure S5 contains the lambda corrected quantile-quantile plots. The SNP with the strongest association signal was the low-frequency rs142324060 ($\lambda_{corrected}$ P = $7.6{\cdot}10^{-8}$; MAF < 5%). The region around the top SNP associated with initiation – rs142324060 on chromosome 5 is displayed in Supplemental Figure S6. The Supplemental Table S3 contains the top SNPs associated with age at onset (P < 10^{-5}). Table 7.4 includes the top five genes with the lowest P-values obtained in the gene-based analysis along with their functions (according to GO annotations).

In our exploratory gene-based analysis none of the genes reached the genomewide significance threshold of $\alpha = 4.3{\cdot}10^{-7}$. The genes showing the strongest association with our phenotype were Gem (nuclear organelle) associated protein 5 (*GEMIN5*) on chromosome 5 (P = $4.7{\cdot}10^{-4}$) and the uncharacterized *LOC101927911*

Table 7.4: Top five genes showing the strongest association with age at onset of cannabis use.

Gene Name (Gene Id)	Chr	Start Position	# SNPs Assigned to Gene	Key SNPs Position (rs number)	Gene Feature	Key SNPs P-Value (lambda adjusted)	Gene P-Value	Molecular Function according to Gene Ontology Annotation
Gem (nuclear organelle) associated protein 5 (*GEMIN5*)	5	154266975	3	154289310 (148816132)	intronic	$1.4 \cdot 10^{-5}$	$4.7 \cdot 10^{-4}$	protein binding; snRNA binding.
				154272889 (816735)	intronic	0.038	-	
Uncharacterized (*LOC101927911*)	17	2865540	9	2871545 (4790396)	intronic	$1.6 \cdot 10^{-4}$	$4.7 \cdot 10^{-4}$	-
Metallothionein 4 (*MT4*)	16	56598960	13	56598707 (141262031)	upstream	$1.9 \cdot 10^{-5}$	$5.2 \cdot 10^{-4}$	copper ion binding; zinc ion binding.
				56605477 (4784686)	downstream	0.001	-	
				56596812 (71387120)	upstream	0.003	-	
Kinesin family member 4B (*KIF4B*)	5	154393259	1	154401490 (115299630)	downstream	$3.9 \cdot 10^{-5}$	$5.3 \cdot 10^{-4}$	nucleotide binding; DNA binding; microtubule motor activity; ATP binding; microtubule binding;

Continued in Table 7.5.

on chromosome 17 ($P = 4.7 \cdot 10^{-4}$), followed by the Metallothionein 4 (*MT4*) on chromosome 16 ($P = 5.2 \cdot 10^{-4}$). The SNP with the strongest association signal - the rs142324060 ($\lambda_{corrected}$ $P = 7.6 \cdot 10^{-8}$) was not assigned to a gene in the GATES analysis.

Table 7.5: *Continued from Table 7.4*

Gene Name (Gene Id)	Chr	Start Position	# SNPs Assigned to Gene	Key SNPs Position (rs number)	Gene Feature	Key SNPs P-Value (lambda adjusted)	Gene P-Value	Molecular Function according to Gene Ontology Annotation
Peptidylprolyl isomerase G (cyclophilin G) (*PPIG*)	2	170440849	53	170439011 (118138006)	upstream	$3.5{\cdot}10^{-5}$	$5.8{\cdot}10^{-4}$	peptidylprolyl cis-trans isomerase activity; isomerase activity.
				170444201 (78740435)	intronic	$5.7{\cdot}10^{-5}$	-	
				170437115 (12618592)	upstream	$1{\cdot}10^{-4}$	-	
				170497179 (3731675)	downstream	$1.4{\cdot}10^{-4}$	-	
				170480402 (12612841)	intronic	$6.5{\cdot}10^{-4}$	-	
				170471270 (115697204)	intronic	$6.5{\cdot}10^{-4}$	-	
				170466028 (75173877)	intronic	$6.7{\cdot}10^{-4}$	-	
				170461257 (7421113)	intronic	0.001	-	
				170477394 (75968631)	intronic	0.001	-	

7.5 Discussion

The aim of the study was to explore the contribution of GVs to initiation of cannabis use and age at onset. Using GCTA and a sample of distantly related individuals from the NTR, we estimated that the genomewide SNPs collectively explain 25% (SE = 0.088; P = 0.0016) of the variance in initiation. Although lower than the twin-based heritability estimate, our estimate provides an indication of the total signal in the currently measured (and tagged) SNPs, confirming that initiation of cannabis use is a heritable trait. The remaining variance (up to about

40% as estimated by twin studies) may, in part, be attributable to rare SNPs, weakly correlated with the measured SNPs (Visscher et al. [354]). Our estimate is larger than that reported by Verweij and colleagues, namely 6% (95% CI [0%, 26%], P-value = ns). A possible reason for this difference is that we use more densely distributed SNPs. In addition to the common SNPs overlapping with the HapMap SNPs used by Verweij and colleagues (about 2.4 million common SNPs with MAF $> 5\%$), we included into analysis previously untagged common GVs, and other (than common) GVs, such as low-frequency variants (about 6 million SNPs having MAF $> 1\%$). More densely distributed SNPs are expected to be in higher LD with the causal variants, and so, to provide a more accurate heritability estimate (Visscher et al. [354]).

The chromosome-by-chromosome analyses showed that, on average, longer chromosomes account for a larger amount of variance in initiation. This result lends support to the conclusion that initiation is highly polygenic. The largest amount of variance is explained by chromosome 4 (6.8%; $P = 0.002$), followed by chromosome 18 (3.6%; $P = 0.012$). Regions on both chromosome 4 and 18 have been reported to play a role in cannabis use and other addiction phenotypes. For instance, regions on chromosome 4 harboring the *GABRA* cluster of genes were identified in a linkage study by Agrawal et al. (Agrawal et al. [14]) as plausibly associated with a cannabis abuse and dependence phenotype. Another linkage study (Prescott et al. [274]) provided strong evidence for a large region on chromosome 4 to be involved in alcohol dependence ($P = 2.1{\cdot}10^{-6}$), the same region being also reported by Uhl et al. to be associated with illicit drug abuse (Uhl et al. [327]). Regions on chromosome 18 were suggested to harbor GVs potentially associated with initiation of cannabis use (Agrawal et al. [14]), methamphetamine abuse (Lee et al. [195]) and alcohol dependence (Prescott et al. [274]). However, when tested individually, none of the GVs achieved an association P-value less than the adapted (i.e., for multiple testing) alpha of 10^{-8}.

We further explored how our results compare with previously published ones. Using the SNP effect concordance method (Nyholt [259]) and the NTR as a replication sample, we checked whether there is an excess of SNPs showing concordant effects in the meta-analysis by Verweij et al. (2012) and in our analysis. Of the 2,110,385 HapMap SNPs tested in both samples, we selected for the comparison 25,204 independent HapMap SNPs ($r2 > 0.1$) that showed the most significant association P-values in the meta-analysis sample. Although we compare summary results for the same phenotype (cannabis initiation) such an analysis is similar in scope to a search for significant pleiotropic effects (genetic overlap): we aimed to single out sets of SNPs showing concordant effects in the two samples beyond what is expected by chance. Concordance of effects was assessed by exact binomial tests. We observed no significant excess of SNPs with concordant effects in the two datasets. It is possible that the effects of the causal variants are too small to be accurately captured by the two samples. It is also likely that the causal GVs were imperfectly tagged by the selected SNPs (e.g., because they have a lower MAF than the selected SNPs), and this further decreased the estimation precision in both samples.

None of the tested genes achieved genomewide significance (P < ∼$4.3 \cdot 10^{-7}$). However, our results have pinpointed several possible candidate genomic regions, likely to have a bearing on the early stage of cannabis use. To name a few, the *ZNF181* and the *ZNF766* genes, both located on chromosome 19, yielded the strongest association signal in the gene-based analysis of initiation (i.e., P = $3.7 \cdot 10^{-6}$, $1.1 \cdot 10^{-4}$, respectively). According to the GO annotations, *ZNF181* and *ZNF766* are functional genes belonging to the zinc finger family of genes, being involved in nucleic acid binding and metal ion binding. The most strongly associated genes with age at onset were the protein coding genes *GEMIN5* (P = $4.7 \cdot 10^{-4}$) on chromosome 5 and *MT4* on chromosome 16 (P=$5.2 \cdot 10^{-4}$). *GEMIN5* plays a role in protein binding and snRNA binding, whereas *MT4* is involved in copper ion and zinc ion binding. The role these genes play in initiation and age at onset has yet to be clarified, as none have been previously reported to be associated with cannabis use or other addiction phenotypes.

To our knowledge this is the first genomewide survival analysis of age at onset of cannabis use to date. The survival modeling approach appears to be appropriate and computationally tractable given the detailed genotypic data currently available (an example dataset and annotated scripts for conducting such an analysis can be found at `http://cameliaminica.nl/research.php`). Clearly, further research on the genetic basis of age at onset would be of interest as the trait may serve as a proxy for both heavy use and experimentation with other drugs. Our study detected association signal in the measured SNPs. A comparison with prior SNP-heritability estimates suggests that at least part of the signal is likely coming from previously untyped common and from low frequency variants. The lack of genomewide significant results for the single variant and gene-based association tests suggests that initiation is a polygenic trait characterized by variants of very small effect (i.e., < 1% explained phenotypic variance). The causal variants are likely distributed over much of the genome, in proportion to the chromosomes' length. Our results do not rule out the contribution of rare variants of larger effect imperfectly tracked by the measured SNPs – a plausible source of the difference between the twin-based heritability estimate and that from GCTA. Powerful analytic strategies and very large samples combined with considering the contribution of rare variants (MAF < 1%) will allow one to further understand the causes of individual differences in the liability to initiate cannabis use.

Chapter 8

Genome-Wide Association Study of Cannabis Initiation Based on a Large Meta -Analytic Sample of 32,330 Subjects from The International Cannabis Consortium

Abstract

Initiation of cannabis use is a heritable trait, yet previous studies had limited success in identifying genetic risk variants. The International Cannabis Consortium was created with the aim of identifying genetic risk variants of cannabis use by conducting meta-analyses of genome-wide association data.

Here we report on the meta-analysis of cannabis use initiation in 13 cohorts (N = 32,330) and two independent replication samples (N = 2,998). The meta-analysis results were followed-up with gene-based tests of association, an estimate of SNP-based heritability, as well as an estimate of the genetic correlation between initiation of cannabis use and nicotine use. We showed that the SNPs on the chip explain 20% of the variance in initiation of cannabis use. While none of the individual SNPs reached genome-wide significance, by using a gene-based test (GATES) we identified four genes significantly associated with initiation of cannabis use: *NCAM1*, *CADM2*, *SCOC*, and *KCNT2*. Finally, we observed a very strong genetic correlation (rg = 0.83) between cannabis initiation and smoking initiation.

In conclusion, we performed the largest meta-analysis of GWAS investigating cannabis use phenotypes to date. Future studies should investigate the impact of the identified genes on the biological mechanisms that lead to cannabis use initiation.

8.1 Introduction

Cannabis is the most widely produced and consumed drug worldwide (United Nations Office on Drugs and Crime 2015) and its use is illicit in most countries. Occasional cannabis use can progress to frequent use, abuse and dependence. About 1 in 10 occasional users becomes dependent and cannabis abuse and dependency is associated with physical, psychological and social consequences (Hall and Solowij [152], Hall and Babor [151]). Despite the increasing use of cannabis for medicinal purposes (Aggarwal, Carter et al. [7], Lucas [221]), association with adverse health effects are reported (Hall [153], Volkow, Compton et al. [356], Wilkinson and D'Souza [360]). Cannabis use has been reported to be associated with increased risk for psychiatric disorders; several studies reported an association between cannabis use and psychosis, schizotypal personality disorder, and mania (Ferdinand, Sondeijker et al. [118], Gibbs, Winsper et al. [138], Radhakrishnan, Wilkinson et al. [282]). In a recently published genetic risk prediction study, Power et al. [273] showed that genes predisposing to schizophrenia predict use of cannabis. The strength of the association between cannabis exposure and those outcomes, the direction of causation, as well as the importance of cannabis as a key modifiable risk factor however remains uncertain (Gage, Zammit et al. [133]).

The probability of cannabis use initiation varies among individuals. Previous studies have shown that individual differences in cannabis use can be partly explained by genetic differences between individuals; a meta-analysis of twin studies reported significant heritability estimates of cannabis use initiation of 48% for males and 40% for females (Verweij, Zietsch et al. [343]). Shared environmental factors, such as cannabis availability and parental monitoring (Gillespie, Neale et al. [140], Gillespie, Lubke et al. [139]), also play a role, accounting for 25% and 39% of the risk for males and females, respectively (Verweij, Zietsch et al. [343]). Moreover, there is substantial overlap in the genetic risks underlying initiation of cannabis use versus cannabis use disorder (Agrawal, Neale et al. [9]).

Numerous studies have aimed to identify the specific genetic risk factors associated with cannabis use phenotypes. Genome-wide linkage studies have revealed suggestive evidence for linkage across many chromosomes (Hopfer, Lessem et al. [161], Agrawal, Hinrichs et al. [14], Agrawal, Morley et al. [15], Agrawal, Pergadia et al. [11], Ehlers, Gilder et al. [111], Ehlers, Gizer et al. [112]). Nearly all findings failed to meet genome-wide significance, with very little consistency across studies. Only one study has examined initiation of cannabis use (Agrawal, Morley et al. [15]) reporting a non-significant linkage locus on chromosome 18.

Candidate gene studies have been more successful in identifying variants associated with cannabis use. Candidate genes of interest include for example *CNR1*, *GABRA2*, *FAAH*, and *ABCB1* (see Agrawal and Lynskey [7] for a review). Again, replication has been inconsistent (Haughey, Marshall et al. [156], Lind, Macgregor et al. [207], Verweij, Zietsch et al. [341]). Based on a sample of 7,452 Caucasian individuals, Verweij et al [341] did not find significant association between lifetime frequency of cannabis use and the ten candidate genes proposed by Agrawal and Lynskey [7]. Overall, the results of candidate gene studies are inconclusive. Some associations have been replicated a few times, but failed to replicate in other studies. These findings may be further distorted due to publication bias favouring positive results (as shown for candidate gene studies in other traits (Farrell, Werge et al. [117]).

As an alternative to the candidate-gene approach, a genome-wide association study (GWAS) is a hypothesis-free method that aims to detect novel genetic variants involved in complex traits. To date, only three genome-wide association studies of cannabis use phenotypes have been published. In the first one, Agrawal et al. [8] performed a GWAS of cannabis dependence based on 708 cannabis-dependent individuals and 2,346 controls. Cannabis initiation was examined in a meta-analysis of two studies with a combined sample size of 10,091 individuals (40.7% cases) (Verweij, Vinkhuyzen et al. [342]) and, recently, in a GWAS sample of 6,744 individuals (Minică, Dolan et al. [246]). Neither study identified any genome-wide significant association.

The lack of genome-wide significant associations may be attributable to the small effect sizes typical of common variants underpinning highly polygenic traits (Manolio, Collins et al. [233], Vrieze, McGue et al. [357]), hence indicating that larger samples are required to ensure sufficient power of detection. In this context, the success of larger GWAS and international consortia examining a variety of complex traits is encouraging (see Sullivan, Daly et al. [309]). For example, multiple large meta-analyses of smoking behaviors have independently identified associations on chromosome 15q25 spanning the $\alpha5$, $\alpha3$, and $\beta4$ nicotinic receptor subunit gene clusters (*CHRNA5*, *CHRNA3*, *CHRNB4*) for the number of cigarettes smoked per day (Furberg, Kim et al. [132], Liu, Tozzi et al. [215], Thorgeirsson, Gudbjartsson et al. [321]).

The International Cannabis Consortium (ICC) was initiated to combine results of multiple GWASs in order to increase the power to detect genetic variants underlying individual differences in cannabis use phenotypes. Currently, the combined sample size of cannabis initiation data within the consortium is 32,330 individuals from 15 cohorts from Europe, the US, and Australia. This sample size is considerably larger than the sample sizes of the previous GWASs investigating the initiation of cannabis use, thereby providing greater power to identify genetic variants of small effect size. The aim of the current study is to meta-analyze the GWAS results from all contributing ICC samples in order to identify genetic variants associated with initiation of cannabis use. The meta-analysis results were first used to get an estimate of SNP-based heritability. Next, we performed an LD-score regression analysis to assess the genetic overlap between initiation of cannabis use and initiation of cigarette smoking, and finally, we carried-out gene-based tests of association.

8.2 Materials and methods

8.2.1 Cohorts

We performed a meta-analysis of GWAS summary results from 13 discovery samples from Europe, USA and Australia including a total of 32,330 individuals of European ancestry. The size of the samples ranged from 721 to 6,778 individuals. The age of the participants varied across the samples from 16 to 87 years. Females represented 53% of the sample (weighted average). The percentage of lifetime users (i.e., never/ever used cannabis) varied from 1% to 92% (weighted average of 44.5%).

Table 8.1: Discovery and replication sample characteristics. Abbreviations: sample size (N), percentage of users that ever used cannabis (% users), percentage of females (% female), and number of SNPs used for the meta-analysis (N SNPs).

Discovery						
Sample	Country	N	% Users	% Female	Mean age (range)	N SNPs
ALSPAC	UK	2,976	42	56	18 (17-19)	5,182,231
BLTS	Australia	721	60	57	26 (18-33)	4,558,509
CADD	US	853	79	30	25 (18-36)	4,972,726
EGCUT1	Estonia	2,765	1.3	55	34 (18-66)	6,048,479
EGCUT2	Estonia	970	4.8	51	31 (18-50)	5,171,164
FinnTwin	Finland	1,029	27	52	23 (20-29)	4,364,135
HUVH	Spain	981	20	30	36 (17-87)	4,971,170
MCTFR	US	6241	59	54	37 (18-71)	6,304,767
NTR	Netherlands	4,653	27	66	37 (18-60)	4,644,238
QIMR	Australia	6,778	51	54	45 (18-85)	5,901,727
TRAILS	Netherlands	1,226	51	47	19 (18-21)	5,336,901
Utrecht	Netherlands	1,173	54	54	21 (18-37)	4,831,885
Yale Penn EA	US	1,964	92	40	38 (16-76)	5,856,902
Replication						
Sample	Country	N	% Users	% Female	Mean age (range)	N SNPs
Radar	Dutch	338	59	44	20 (17-22)	10
Yale Penn AA	US	2,660	82	46	42 (16-76)	10

Two additional independent samples with a total of 2,998 subjects were used for replication. One sample (N = 2,660) consisted of African-American subjects, whereas the second sample (N = 338) included subjects of European ancestry. We refer to Table 8.1 for characteristics of the individual samples. The procedures for data collection per sample are described in the Supplemental Information S1.

8.3 Phenotype and covariates

Subjects were asked whether they have ever used cannabis (yes (1)/ no (0)). Covariates included age, sex, and birth cohort which spanned 20 year-periods indicated by dummy variables, with the lowest birth cohort (i.e., oldest age group) used as the reference group. Details on phenotyping and individual sample characteristics for both the discovery samples and the replication samples are included in the Supplemental Information S1 and the Supplemental Table S1.

8.3.1 Genotype imputation

Genotype imputation was based on the 1000 Genomes phase 1 reference panel (Abecasis, Auton et al. [2]). To take account of genotype uncertainties, we used the allelic dosage in the association analysis. We refer to the Supplemental Table S2 for details on the genotyping platform, imputation program, and quality control thresholds used by each group.

8.4 Statistical analyses

8.4.1 GWAS in each discovery cohort

The GWASs were performed by each group according to a standardized protocol. Association between the binary phenotype and the genotypes was tested genome-wide with a logistic regression model. Age at the time of phenotypic assessment, sex, birth cohort and the first four principal components were included as covariates. For family-based samples, familial relatedness was taken into account by using a sandwich correction as implemented in PLINK (Purcell, Neale et al. [279]). Supplemental Table S2 includes information on the program used by each group to perform the analysis.

8.4.2 Meta-analysis of GWAS results

Before performing the meta-analysis, we applied a set of filters to each GWAS results set independently. First, we removed insertions and deletions, ensuring that all base pair positions were unique and referred to the same genetic variant (i.e. SNP). Second, we removed genotyped SNPs that were not in Hardy-Weinberg equilibrium ($p \leq 10^{-5}$). Third, we removed SNPs with MAF$< \sqrt{5}/N$, which corresponds to less than 5 estimated individuals in the least frequent genotype group, under the assumption of Hardy-Weinberg equilibrium (HWE). In the EGCUT1 sample, due to very low prevalence of cannabis initiation (1.3%), we excluded SNPs with MAF < 0.2. Fourth, we excluded SNPs with an imputation quality score below 0.6. This filter was applied regardless of the quality score type used (INFO, Proper info, or R^2hat). Finally, SNPs present in only one sample, and

SNPs with invalid alleles, or allele frequencies inconsistent with the 1000 Genomes phase I European reference panel (absolute MAF difference>0.15) were removed.

We performed a fixed effects meta-analysis based on the cohorts effect sizes and standard errors using METAL (Willer, Li et al. [363]). Our meta-analysis combined association summary statistics for 6,444,471 SNPs that passed all filters. We applied the conventional threshold of 5×10^{-8} as an indication of genome-wide significance (see (Sham and Purcell [295])).

8.4.3 Estimation of SNP-based heritability and genetic overlap with nicotine use

The proportion of phenotypic variance explained by the SNPs was estimated using the density estimation (DE) method developed by So et al. [299]. The DE method estimates the genome-wide distribution of effect sizes based on the difference between the observed distribution of test statistics in the meta-analysis and the corresponding null distribution. Prior to estimation, SNPs present in at least 25% of the meta-analysis samples were pruned for LD. We used the $r^2 = 0.15$ pruning level as the primary result for consistency with other applications of this method. More details are reported in the Supplemental information S2. LD Score regression (Bulik-Sullivan, Finucane et al. [50], Bulik-Sullivan, Loh et al. [51]) was used as an alternative method to estimate the SNP based heritability as well as to estimate the degree of genetic overlap between initiation of cannabis use and smoking (Furberg, Kim et al. [132]) (see Supplemental information S2).

8.4.4 Gene-based tests

The GWAS meta-analysis results were then used to perform a gene-based test of association using the Knowledge-based mining system for Genome-wide Genetic studies (KGG) software Version 3.5 (Li, Gui et al. [201], Li, Kwan et al. [202]). This approach employs an extended Simes procedure that obtains an overall association p-value by taking account of the linkage disequilibrium structure among the SNPs within a gene. 24,576 genes were tested for association with initiation. We set the genomewide significance threshold equal to 0.05, and we used the Benjamini and Hochberg (BH) method (Benjamini and Hochberg [29]) to correct the gene-based p-values for multiple testing.

8.5 Results

8.5.1 SNP-based heritability and genetic overlap with smoking initiation

Genome-wide SNPs explained 20% of the variance in the liability to initiate cannabis use (p <0.001). The estimate of variance explained is robust across pruned sets with similar r^2 thresholds (see Supplemental Table S6). Stricter LD pruning (i.e. $r^2 = 0.05$), or restricting to SNPs present in all studies, substantially decreased the estimate of variance explained. This is likely due to the reduced number of SNPs included, which may tag imperfectly the causal variants. An alternative estimation based on the LD Score regression method yielded similar results ($h^2 = 0.13$, SE = 0.02, z = 5.56, p = 1.4×10^{-7}). The genetic correlation between initiation of cannabis use and initiation of nicotine use was estimated at 0.83 (SE = 0.15, z = 5.64, p = 1.85E-08).

8.5.2 Meta-analysis

The strongest association signal was yielded by the rs4984460 SNP located on chromosome 15 (p-value of 4.6×10^{-7}; see Figure S5 for the forest plot). The SNP is located in an intergenic region between LOC400456/LOC145820 and NR2F2 and MIR1469 genes. However, the association signal at this SNP has not passed the genomewide significance threshold (see Manhattan plot, Figure S1a).

We identified suggestive signals at 153 SNPs on 11 chromosomes with SNP p-values< 10^{-5} (see supplemental Table S4), as illustrated in the QQ plot (Figure S1b). The Manhattan and the QQ plots for each sample are included in the Supplemental Figures S2a-m and S3a-m. Table 8.2 contains information on the ten independent SNPs (r^2 <0.1) which yielded the lowest p-values in the SNP-based association analysis.

None of these 10 SNPs were replicated in either of the two independent replication samples (Supplemental Table S3). Also, in a meta-analysis of the ten top SNPs perfomed in a sample combining the discovery and the replication samples, none of the SNPs reached statistical significance. Local plots of the most strongly associated regions, including neighboring genes, are provided in Supplemental Figures S4a-j.

Table 8.2: Top 10 SNPs with meta-analysis results of discovery samples, and results of combined discovery and replication samples. SNPs are displayed when not in linkage disequilibrium ($R^2 < 0.1$; and for SNPs with $R^2 \geq 0.1$ only the most significant SNP is shown in the top 10). * Direction per sample: allele A1 increases (+) or decreases (-) liability for cannabis use, or sample did not contribute to this SNP (?). Order of samples: ALSPAC, BLTS, CADD, EGCUT1, EGCUT2, FinnTwin, HUVH, MCTFR, NTR, QIMR, TRAILS, Utrecht, Yale Penn EA. Sample information can be found in Table 8.1. Abbreviations: Chromosome (Chr), location in base pairs in human genome version 19 (BP (hg19)), allele 1 (A1), allele 2 (A2), Frequency of allele 1 (Freq A1), standard error (s.e.). $ The combined sample contains the discovery, the Radar replication sample and the African Americans replication sample.

SNP	Chr	BP (hg19)	A1	A2	Freq A1	beta (s.e.)	p-value	Discovery*
						Discovery		
rs4984460	15	96424399	T	G	0.75	-.11 (.023)	4.6×10^{-7}	+- - ++- - - - - - -+
rs2099149	12	30479358	T	G	0.81	-.16 (.032)	9.8×10^{-7}	- - -?-??-?- - +-
rs7675351	4	141218757	A	C	0.86	-.15 (.031)	1.4×10^{-6}	- - -?+- - -?- - - -
rs4471463	11	112983595	T	C	0.55	-.09 (.020)	1.5×10^{-6}	- - - - +-+- - - - +-
rs7107977	11	915764	A	G	0.60	.27 (.058)	1.9×10^{-6}	??+++?+???+?+
rs58691539	2	52753909	T	G	0.91	-.29 (.062)	2.1×10^{-6}	-????-?-????-
rs2033867	2	175188281	A	G	0.06	.24 (.051)	2.6×10^{-6}	+??????+++??+
rs35053471	3	47124761	A	T	0.38	-.10 (.022)	2.7×10^{-6}	- - - - -?+- - - - - -
rs12518098	5	60864467	C	G	0.68	.10 (.022)	3.0×10^{-6}	++++-++++++++
rs73067624	1	196333461	T	C	0.90	-.18 (.039)	3.1×10^{-6}	-?-?- - - - ?- - - -
						Combined$		
rs4984460	15	96424399	T	G	0.75	-.09 (.022)	5.5×10^{-6}	
rs2099149	12	30479358	T	G	0.81	-. 13(.03)	5.3×10^{-6}	
rs7675351	4	141218757	A	C	0.86	-.14 (.029)	3.4×10^{-7}	
rs4471463	11	112983595	T	C	0.55	-.09 (.019)	2.1×10^{-6}	
rs7107977	11	915764	A	G	0.60	. 21(.045)	2.5×10^{-6}	
rs58691539	2	52753909	T	G	0.91	-.23 (.050)	3.7×10^{-6}	
rs2033867	2	175188281	A	G	0.06	.23 (.049)	2.4×10^{-6}	
rs35053471	3	47124761	A	T	0.38	-.08 (.021)	4.06×10^{-5}	
rs12518098	5	60864467	C	G	0.68	.10 (.021)	4.2×10^{-6}	
rs73067624	1	196333461	T	C	0.90	-.16 (.036)	7.3×10^{-6}	

8.5.3 Gene-based tests

We tested for association with initiation 24,576 genes. Figure 8.1 shows the Manhattan and the QQ plots of the gene-based association tests. Results for the top 100 genes can be found in Supplemental Table S5.

Four genes and one intergenic non-coding RNA region were significantly associated with lifetime cannabis use (BH corrected p < 0.05): Neural Cell Adhesion Molecule 1 (*NCAM1* on 11q23), Cell Adhesion Molecule 2 (*CADM2*, on

Figure 8.1: The Manhattan (A) and the QQ plot (B) based on results of the gene-based analysis performed in the discovery sample using HYST (Hybrid Set-based Test).

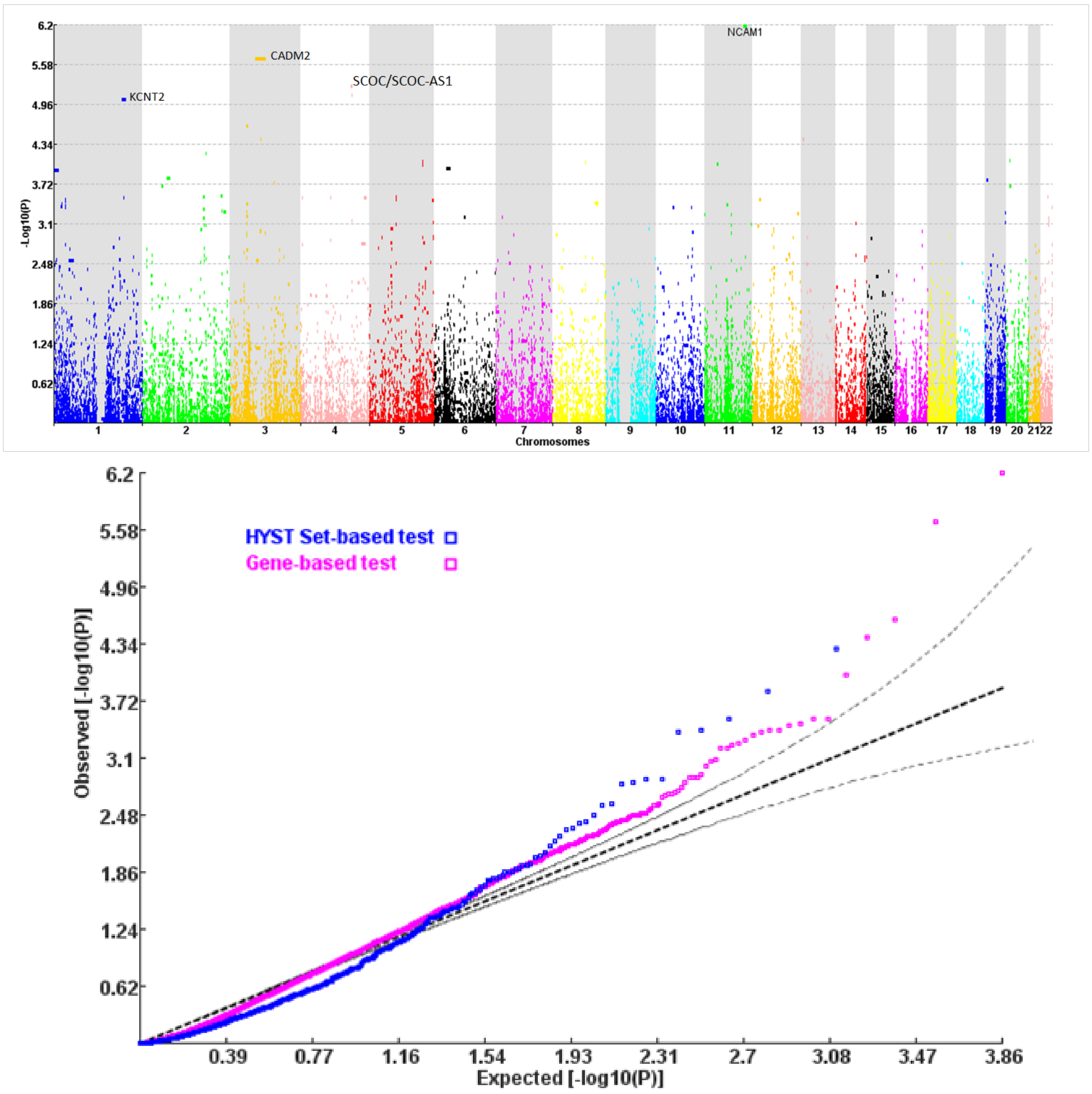
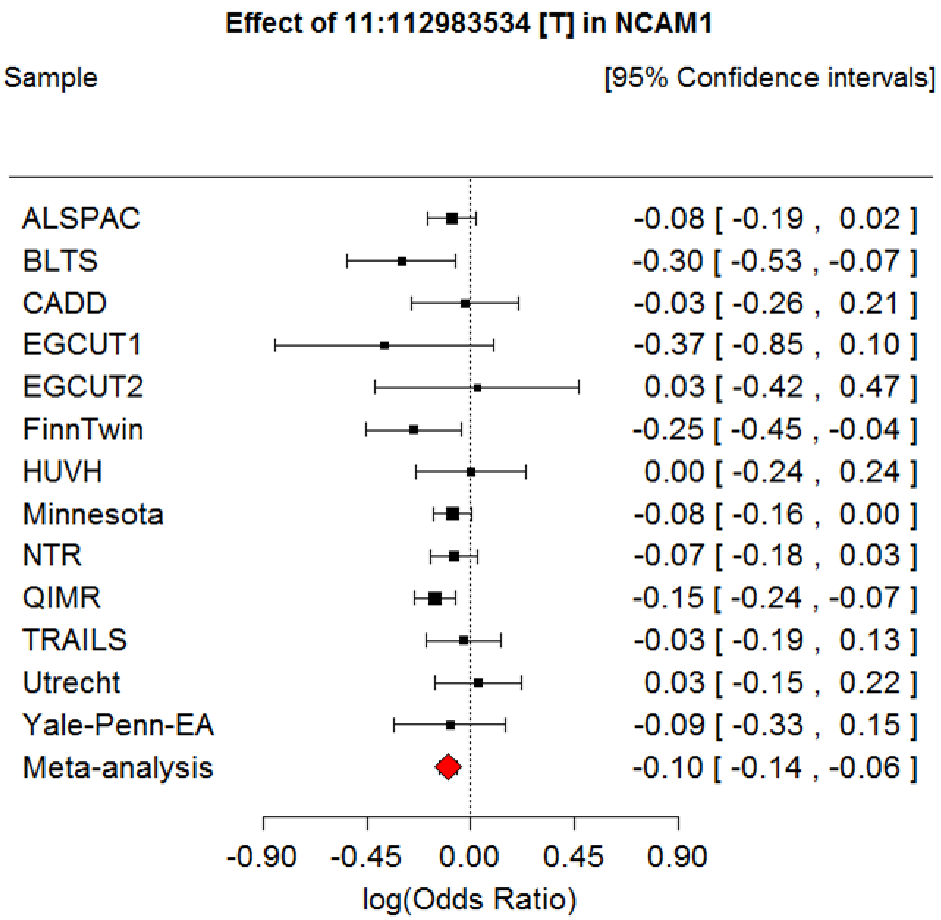


Figure 8.2: Forest plot for the top-SNP rs4471463 in the *NCAM1* gene on chromosome 11.



3p12), Short Coiled-Coil Protein (*SCOC*) and *SCOC* antisense RNA1 (*SCOC-AS1*) (both located on 4q31), and Potasium Channel, Subfamily T, Member 2 (*KCNT2*, on 1q31), see Table 8.3. Regional plots (Viechtbauer [344]) of these top genes can be found in Supplemental Figure S6.

The smallest gene-based p-value was found for the *NCAM1* gene. The effect of the key SNP in the individual samples is shown in the forest plot (see Figure 8.2). Most study samples show an effect in the same direction. The forest plot for two key SNPs with lowest p-values in the other significant gene regions can be found in Supplemental Figure S5.

Table 8.3: Top 5 genes from the gene-based tests of association with corrected *p*-values (Benjamini & Hochberg) based on the meta-analytic discovery and replication samples. Abbreviations: Human genome version 19 (hg19), base pair length (BP length), and number of SNPs used for the meta-analysis (N SNPs).

Gene	Chr	Start Position (hg19)	BP length	No SNPS	Nominal p-values Discovery	Corrected p-values Discovery	Nominal p-values Replication Radar sample	Nominal p-values Replication African-Americans
NCAM1	1	112831968	303952	400	6.26×10^{-7}	0.015	0.329	0.302
CADM2	3	85008132	1115448	978	2.13×10^{-6}	0.026	0.009	0.112
SCOC-AS1	4	141204879	89668	81	5.76×10^{-6}	0.046	0.864	0.044
SCOC	4	141264614	39097	111	7.85×10^{-6}	0.046	0.433	0.027
KCNT2	1	196194909	382653	237	9.38×10^{-6}	0.046	0.815	0.201

We replicated the association with *CADM2* in the Radar sample (p = 0.009), but not in the African American replication sample (p = 0.11; see Table 8.3). Furthermore, in the African American replication sample we found suggestive association with *SCOC-AS1* (p = 0.044) and *SCOC* (p = 0.027).

8.6 Discussion

We meta-analyzed GWAS results of a large discovery sample including 32,330 individuals from 13 cohorts, and two replication samples including 2,998 subjects. The heritability analyses revealed that about 20% of the variation in initiation of cannabis use is explained by common SNPs ($p < 0.001$). Our estimate is in between previous SNP-based heritability estimates. Verweij et al. [342] estimated that 6% of the variance in initiation is explained by the aggregated common SNPs (MAF > 0.05), while Minică et al. [246] found an estimate of 25% (p = 0.0016). Our estimate is smaller than the twin-based heritability estimate of 40-50% (Verweij, Zietsch et al. [343]). Several sources of variation may explain this difference. For example, age-related genetic differences, non-additive genetic variance, interactions between genetic variants and environmental risk factors, epistasis, rare mutations or shared environmental influences may play a role.

Commensurate with twin study findings, we observed a high genetic correlation between our measure of cannabis use and lifetime cigarette use when based on the full SNP panel. Agrawal et al. reported biometrical genetic correlations between lifetime cannabis use and nicotine use ranging between 0.59-0.74 (Agrawal, Silberg et al. [12]). Kendler et al. [183] also reported significant biometrical correlations between levels of cannabis, nicotine, and alcohol use, which were increasingly influenced by common genetic risks beginning early adulthood.

The difficulty of identifying individual SNPs implicated in cannabis initiation may be attributable to several reasons (Manolio, Collins et al. [233]). First, complex traits are known to be influenced by many variants, each with very small effect sizes. Although power calculations revealed suitable power (96%) to detect odds ratios of 1.15 based on common SNPs (MAF = 0.2), the power to detect smaller effect sizes is lower. For example, there is only 28% power to detect effect sizes with odds ratio of 1.1 and MAF = 0.2. Therefore, our data suggest that the effect sizes of single variants contributing to cannabis initiation are likely to be smaller than 1.15. While the statistical power to detect individual variants may still be inadequate, combining variants within larger units (i.e., genes) did reveal four significant genes associated with cannabis initiation implying that these genes are appropriate targets for future functional studies of cannabis use.

The gene-based tests of association identified four protein-coding genes and one intergenic region significantly associated with initiation including *NCAM1*. The role of *NCAM1* is to regulate pituitary growth hormone secretion as a membrane-bound glycoprotein that mediates cell-cell contact by hemophilic interactions (Rubinek, Yu et al. [286]). *NCAM1* is part of the *NCAM1-TTC12-ANKK1-DRD2 (NTAD)* gene cluster, which is related to neurogenesis and dopaminergic neurotransmission. Importantly, the *NTAD* cluster has been reported to be associated with smoking behavior and nicotine dependence (Munafo, Clark et al. [250], Gelernter, Yu et al. [137], Gelernter, Panhuysen et al. [135], Saccone, Hinrichs et al. [288], Laucht, Becker et al. [193], Bergen, Conti et al. [31], Ducci, Kaakinen et al. [107], Bidwell, McGeary et al. [32]), alcohol dependence (Yang, Kranzler et al. [372], Yang, Kranzler et al. [373]), heroin dependence (Nelson, Lynskey et al. [255]), as well as other substance use disorders (Yang, Kranzler et al. [373]). While it is plausible that *NCAM1* is capturing pleiotropic risks underpinning the liability to illicit substance use in general, we note that the gene was not identified to associate with smoking behavior phenotypes in the GWAS and very large GWAS meta-analyses for smoking behavior (Furberg, Kim et al. [132], Liu, Tozzi et al. [214], Thorgeirsson, Gudbjartsson et al. [321]).

Our second significant gene, *CADM2*, is a synaptic cell adhesion molecule (SynCAM family) belonging to the immunoglobulin (Ig) superfamily. Variants in the *CADM2* gene have been previously associated with body mass index (Speliotes, Willer et al. [303]), processing speed (Ibrahim-Verbaas, Bressler et al. [174]) and autism disorders (Casey, Magalhaes et al. [59]). Interestingly, these phenotypes were also associated with cannabis use in previous studies (Kelleher, Stough et al. [180], Hayatbakhsh, O'Callaghan et al. [157], De Alwis, Agrawal et al. [90]).

Our third significant gene, *SCOC*, encodes a short coiled-coiled domain-containing protein that localizes to the Golgi apparatus. Many coiled coil-type proteins are involved in important biological functions such as the regulation of gene expression through the regulation of transcription factor binding (Mason and Arndt [236]). The function of *SCOC* is largely unknown, and no previous association studies have linked *SCOC* to cannabis or to other substance use phenotypes. The *SCOC* antisense RNA1 gene is located in the same chromosomal region.

Finally, *KCNT2* encodes a potassium voltage-gated channel (subfamily S, member 2). The sodium-activated potassium channels Slack and Slick are encoded by *KCNT1* (Potassium Channel, Subfamily T, Member 1) and *KCNT2*, respectively, which are found in neurons throughout the brain. Suggestive association for SNPs near *KCNT2* have previously been found for cocaine dependence and for early-onset, highly comorbid, heavy opioid use (Gelernter, Kranzler et al. [134], Gelernter, Sherva et al. [136]). This suggests that potassium signalling may play a role in addiction.

We replicated the association with *CADM2* in the Radar sample (p = 0.009), and we found suggestive association with *SCOC-AS1* and *SCOC* in the African American sample (both p-values < 0.05). Finally, we showed that the genetic liability of cannabis initiation overlaps to a great extent (r = 0.83) with the genetic liability of smoking initiation. Our results are consistent with the hypothesis that cannabis initiation is a highly polygenic trait, comprising many SNPs, each with small effects contributing to risk, while part of the genetic risk overlaps with other substance use phenotypes, particularly with initiation of nicotine smoking.

Our findings must be interpreted in the context of four limitations. First, our study was underpowered to detect very small effect sizes of individual variants. The sample size should be increased with approximately two-fold to detect SNPs with effect sizes with an odds ratio of 1.1. Second, lifetime cannabis use is a dichotomous measure combining single lifetime, regular and chronic users, meaning that our sample comprises heterogenous patterns of use. Phenotypic heterogeneity among users has the potential to reduce the power to detect association (see e.g., [231]). Third, prevalences of lifetime cannabis use varied between 1% (EGCUT1) and 92% (Yale Penn EA), partly due to differences in sample characteristics, recruitment strategies, and policy differences across countries. However, despite these differences, the forest plots of key SNPs (see Figure 2; see also the Supplemental Figure S5) reveal that the 95% confidence intervals surrounding the effect estimate typically include the estimated meta-analytic effect, and the CI tend to overlap among studies. This indicates that the input samples are representative of the same population of users. Finally, the average age of the participants varied between 18 (ALSPAC) amd 45 (QIMR) years. The average age of each sample did not correlate significantly with sample prevalences ($r = -0.04$, p = 0.91). Moreover, the fact that younger participants may be prematurely classified as 'never users' is expected to decrease power, but does not invalidate our results.

Based on our observations, the following recommendations for future studies can be made. We have identified four genes significantly associated with cannabis use. These genes should be followed-up in future functional studies. Especially, the role of our top gene *NCAM1* should be carefully examined to understand its functional role, possibly in combination with the other genes in the same gene cluster (*NCAM1-TTC12-ANKK1-DRD2*). Obviously, we should aim to increase the statistical power by increasing the sample size and by focussing on continuous phenotypes and phenotypes indicative of more severe forms of cannabis use. The next goal of the International Cannabis Consortium is to perform a meta-analysis on GWASs of age at first cannabis use. Our rationale is based on the observation that early initiation of cannabis use is also associated with rapid progression towards cannabis abuse and dependence, poly-substance use, and other substance use disorders (Agrawal, Grant et al. [13], Lynskey, Vink et al. [225], Agrawal, Lynskey et al. [14], Grant, Lynskey et al. [145]). Methods such as rare variant association analyses may also be used to reveal the biological pathways of cannabis use. Environmental risk factors may be incorporated to investigate gene by environment interactions. Hopefully, the combination of advanced technologies and novel statistical approaches with larger samples will further contribute to our understanding of the genetic architecture of cannabis use.

8.7 Conclusion

We have performed the largest meta-analysis to date of GWAS investigating cannabis use phenotypes. With a sample of over 32,000 individuals, our results suggest the involvement of four genes: *NCAM1, CADM2, SCOC*, and *KCNT2*. The association with *CADM2* was confirmed within one independent replication sample. Future studies should investigate the impact of the identified genes on the biological mechanisms that lead to initiation. Our results further confirm that initiation is under the influence of many common genetic variants. The measured SNPs together explain about 20% of the phenotypic variation and show a high degree of genetic overlap (r = 0.83) with smoking initiation.

Chapter 9

Survival Meta-Analysis of Age at Onset of Cannabis Use

Abstract

Cannabis is one of the most commonly used substances among adolescents and young adults. Research shows that the age at first cannabis use is decreasing. This is probably due to lower risk perception and increased availability due to medicalization and decriminalisation. In this study, we aim to identify genetic variants underlying age of onset, a risk factor for multi-substance use and subsequent dependence.

We performed the largest molecular genetic study to date of age of onset of cannabis use in a sample consisting of 24,222 individuals from nine cohorts from Europe, United States, and Australia. Five SNPs located on chromosome 16 within the Calcium-transporting ATPase gene (*ATP2C2*) passed the genome-wide significance threshold in the SNP-based analysis. The five SNPs are in high LD ($r^2 > 0.8$), and thus may represent a single independent signal. The most significant association was with the intronic variant rs1574587 (P = 4.067E-09). Following the single variant analysis, we performed a genome-wide gene-based analysis. The gene-based tests also identified the *ATP2C2* gene on 16q24.1 (P = 1.54E-06), along with two additional genes: *ECT2L* on 6q24.1 (P = 6.59E-08) and *RAD51B* on 14q24.1 (P = 5.22E-06).

Our findings have the potential to deepen our understanding of the biological mechanisms underlying addiction. Especially the association at *ATP2C2* provides further support for the hypothesized link between the calcium signalling genes and addiction behaviors, and is consistent with the reported associations between early onset of cannabis use and multi-substance use as well as with subsequent dependence. A thorough investigation of the functional consequences of variation in these genes is warranted.

9.1 Introduction

Cannabis is one of the most commonly used substances among adolescents and young adults (Australian Institute of Health and Welfare 2013, U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration and Quality 2013). In the US, the average age at first cannabis use (among individuals who initiated between 2002 and 2013) was 18 years, with 57% of initiateds being under 18 years old (U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration and Quality 2013).

In 2013, the mean age at initiation was 16 years among individuals who started cannabis use prior to the age of 21 (U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration and Quality 2013). Globally, younger cohorts are more likely to use all types of drugs including cannabis. In addition, the male-female gap, which is commonly observed in older cohorts, has been found to be closing in more recent cohorts (Degenhardt, Chiu et al. [94], Butterworth, Slade et al. [54]). Furthermore, research shows there is a trend towards decreasing age at first use (Degenhardt, Lynskey et al. [95]), probably due to lower risk perception, especially among young people (UNODC, 2014), and increased availability due to medicalization and decriminalisation.

Following initiation, chronic cannabis use has been associated with various adverse physical, psychological, and social consequences. Previous studies have shown that people who initiate cannabis use at a younger age are at increased risk of detrimental outcomes. Early age of onset has been linked to educational under-achievement (Grant, Scherrer et al. [146], Verweij, Huizink et al. [342], Stiby, Hickman et al. [307]), greater family adversity and negative life events (Hyman and Sinha [173]), psychosis and psychopathology (Fergusson, Lynskey et al. [120], Fergusson, Lynskey et al. [121], Fergusson and Horwood [119], Arseneault, Cannon et al. [21]), progression to abuse-dependence and multi-substance use (Agrawal, Neale et al. [10], Lynskey, Vink et al. [225], King and Chassin [186], Agrawal, Lynskey et al. [14], Chen, Storr et al. [62], Grant, Lynskey et al. [145], Lynskey, Agrawal et al. [222], Bracken, Rodolico et al. [46]), and cognitive decline (Tamm, Epstein et al. [314]).

Given the widespread use and its associations with adverse life outcomes, it is important to identify and investigate the determinants of cannabis use initiation; identifying and quantifying the genetic risks associated with age of onset of cannabis use is therefore a public health concern. Based on dichotomized (early vs. late) or ordinal measures of age of onset, twin studies have revealed significant heritability accounting for individual differences in age of onset. Specifically, using a sample of ever users and a dichotomized measure of age of onset ('early', i.e., before the age of 16 years, versus 'late'), Lynskey et al. [222] estimated the heritability of age of onset to be about 80%.

In another study, in which age of initiation was categorized as 'never', 'late' (17 years or older) or 'early' (16 years or younger), Sartor et al. [290] estimated the heritability to be about 52%. Interestingly, measured shared environmental influences, including parenting styles, parental monitoring, neighbourhood the twins grew up in, and drug availability, had little bearing on the age of onset (Sartor et al. [290]).

Minică et al. [246] conducted a genome-wide survival analysis to identify the genetic variants that underlie individual differences in age of onset of cannabis use. They performed a genome-wide survival analysis in a sample of 5148 individuals from 2992 independent families from the Netherlands Twin Register (including 852 individuals who had initiated cannabis use). No individual SNPs or genes significantly associated with age of onset of cannabis use were detected, which may be due to a lack of statistical power.

However, the survival-based approach suggested by Minică et al. [246] is suitable for the analysis of this type of trait in the genome-wide context, and it is expected to be superior in terms of power to an analysis limited to initiated individuals, and to a logistic regression model (van der Net, Janssens et al. [333], Kiefer, Tung et al. [185]). Hence, this approach will be implemented also in the current study to detect genetic variants that significantly predict age of onset.

The International Cannabis Consortium (ICC) was established to identify genetic variants underlying cannabis use by combining data from various cohorts. The first meta-analysis focused on cannabis use initiation in 13 cohorts (N = 32,330; Stringer et al, under review) and identified four genes significantly associated with initiation of cannabis use: *NCAM1*, *CADM2*, *SCOC*, and *KCNT2*, two of which (*NCAM1* and *KCNT2*) were previously reported to be associated with other substance use (heroine and nicotine, and cocaine dependence, respectively).

In this second study, we extend our search to identify genetic variants underlying age of onset, an important risk factor for escalated use. We conducted a fixed-effects meta-analysis of genome-wide survival analyses in a sample consisting of nine cohorts with available data on age at initiation. Following Minică et al. [246], the phenotype was modelled in each participating cohort as a function of SNP and relevant covariates (i.e., sex, birth-cohort, principal components and study specific covariates) using a Cox proportional hazards regression model (Dobson [102]).

The SNP-based meta-analysis results were subsequently used in a gene-based analysis. Taking the gene as the unit of analysis is expected to increase the statistical power relative to the single SNP analysis. Specifically, a gene-based approach allows one to interrogate jointly all the SNPs within a gene and thus reduces the multiple testing burden (Neale and Sham [252], Li, Gui et al. [201], Li, Kwan et al. [202]).

9.2 Materials And Methods

9.2.1 Sample

The meta-analysis combined genome-wide survival analysis summary statistics obtained from cohorts from the Europe, United States, and Australia. The summary statistics were based on a total of 24,222 individuals with the mean age ranging from 17.31 to 46.93 years. Females represented 53.41% of the sample, and 43.1% of the observations were uncensored (i.e., 43.1% of the individuals declared to have initiated cannabis use). Table 9.1 contains descriptive information about the participating cohorts (see also the Supplemental Table S1 for more details).

Table 9.1: Descriptive information on the nine participating cohorts

Cohort	Size (or range/SNP)	% Females	%Uncensored Observations	Mean age (SD)	Number of SNPs
ALSPAC	6147	51.9	38.4	17.3(1.7)	6,284,747
BLTS	721	57.1	59.5	26(3.3)	4,093,835
HUVH	580	29.8	20.2	35.6	4,318,727
FinnTwin	978-1029	51.7	27.4	22.8(1.3)	4,362,100
NTR	5148	62.3	16.6	46.9(17.5)	4,773,834
QIMR	6758	53.8	51.3	45.3(10.9)	5,953,917
TRAILS	229-478	53.8	61.7	20.0(1.6)	4,109,101
Utrecht	1173	53.5	54.1	20.6(2.3)	4,260,457
Yale-Penn	2188	41.0	92.6	38.0(10.5)	5,732,659

9.2.2 Phenotyping

Age of onset was assessed by means of clinical interview or via questionnaire (see the Supplementary file S1 for information on the exact phrasing of the question used and for more information about the sample collection). Depending on the initiation status, subjects were coded as uncensored (i.e., if they initiated) or censored (i.e., if they did not initiate at the time of the last survey). To maximize the sample size we included all available data, i.e., censored and uncensored observations, without imposing any age restriction.

9.2.3 Genotyping

Genotyping was performed by each participating cohort (see Supplementary Table S2 for details on genotyping performed by each cohort). Following genotyping, each participating group employed quality checks. These resulted in removing single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) below 1%, call rate lower than 95%, p-values in the test of Hardy Weinberg equilibrium (HWE) below 1E-04. At the subject level, quality checks involved removal of individuals showing low overall call rates ($< 95\%$), gender conflicts, or excess autosomal heterozygosity (indicative of genotyping errors). Furthermore, duplicate samples and unintended 1st or 2nd degree relatives (in case of a sample of unrelated individuals) were also removed.

9.2.4 Imputation

The analysis protocol required all participating cohorts to perform genotype imputation using the 1000 Genomes release March 2012 as a reference sample. We refer to the Supplementary Table S2 for details on imputation performed by each participating cohort. The set of quality checks which were performed after imputation involved filtering out SNPs with poor imputation quality (R^2hat, Proper Info or Info less than 0.4) and SNPs whose allele frequencies mismatched those reported in the 1000 Genomes data by more than $|0.2|$. We used best-guess genotypes (given requirements of the software used for the genome-wide analysis) and we restricted the analysis to SNPs on autosomal chromosomes.

9.2.5 Quality checks prior to the meta-analysis

Prior to the meta-analysis, each input file underwent a set of quality checks pertaining to imputation quality, minor allele frequency and Hardy-Weinberg equilibrium. As we used best guess genotypes, we selected for the meta-analysis only SNPs with high imputation quality (R^2hat/Proper Info/Info above 0.8). With this threshold, the average imputation quality across the SNPs ranged from 0.95 to 0.99 across the nine cohorts. Second, we retained only those SNPs with minor allele frequency larger than ($\sqrt{5}/N$). The frequency-based filter was applied by taking into account the sample size, and ensured there were at least 5 estimated individuals in the least frequent genotype group. Third, genotyped SNPs were retained if Hardy-Weinberg equilibrium was not violated (p-value $>$ 1E-04). Lastly, SNPs were retained if their allele frequencies matched those reported in the 1000 Genomes data (i.e., the allele frequency difference has not exceeded $|0.2|$). The meta-analysis included 6,158,982 unique biallelic SNPs which passed quality control checks (see Table 9.1 for number of SNPs in each input file meeting all our quality control criteria).

9.2.6 Statistical analysis

Per-sample analysis was carried-out based on a standardized analysis protocol (see Supplementary file S2 for details). Each cohort performed a Cox proportional hazards regression analysis in which age of onset (or age at the last survey, for censored observations) was regressed on the SNP (coded additively co-dominant as 0, 1, 2) and on the following covariates: sex, birth-cohort (to correct for generation effects), the first four principal components (to correct for possible population stratification), and study specific covariates (to correct for chip and/or batch effects). To account for family relatedness we used the 'cluster' option (which assumes an independence working correlation matrix) as implemented in the R-package survival [320]. This is the only option implemented in the survival package that ensures the standard errors are robust to misspecification of the familial covariance matrix (while full, correct modelling of the background is prohibitively slow in the genome-wide context). The survival package was accessed either directly in R [317] or called from Plink (Purcell, Neale et al. [279]) via the Rserve package [328].

9.2.7 Meta-analysis

The meta-analysis was carried out in Metal (Willer, Li et al. [363]) using a fixed-effects model and the SCHEME STDERR option which weighs the beta coefficients by the inverse of their associated standard errors. To ensure that the bulk of the test statistic distribution follows the expectation under a theoretical null model, we applied genomic control to each input file prior to the meta-analysis. This ensured that none of the input samples contribute disproportionately to the meta-analysis results (De Bakker, Neale et al. [91]). Similar to e.g. Furberg et al. [75] and Allen, Estrada et al. [16], we computed the standard error (and the corresponding p-value) by multiplying the variance of the beta by the lambdaGC (genomic control) estimated for each sample (see Supplemental Table S2). The Supplemental Figures include the per-sample lambda corrected Manhattan and quantile-quantile plots. The meta-analysis was based on 6,158,983 SNPs present in at least two samples. As proposed by Pe'er et al., [264], Sham and Purcell [295], an alpha of 1E-08 was used as the genome-wide significance threshold. Statistical analyses were performed on the lisa Genetic Cluster Computer (`http://www.geneticcluster.org`).

9.2.8 Gene-based tests

Results from the genome-wide meta-analysis were then used in tests of gene-based association using the Gene-based Association Test Using the Extended Simes procedure (GATES) implemented in the Knowledge-based mining system for Genome-wide Genetic studies (KGG) software Version 3.5 (Li, Gui et al. [201], Li, Kwan et al. [202]). GATES combines the p-values of the SNPs within a gene by taking account of the linkage disequilibrium among the SNPs. The SNPs were mapped onto 24,404 genes (or within 5 kilobase pairs of each gene) based on NCBI gene coordinates. Linkage disequilibrium structure was inferred based on the 1000 genomes haplotypes ALL (version March, 2012). For this analysis, False Discovery Rate (FDR) of 0.05 Benjamini and Hochberg [29] was used as the genome-wide significance threshold. We opted for FDR to adjust the p-values for multiple significance testing rather than for the Bonferroni correction as the latter reduces the statistical power when the tests are correlated (see [29]).

9.3 Results

9.3.1 Meta-analysis

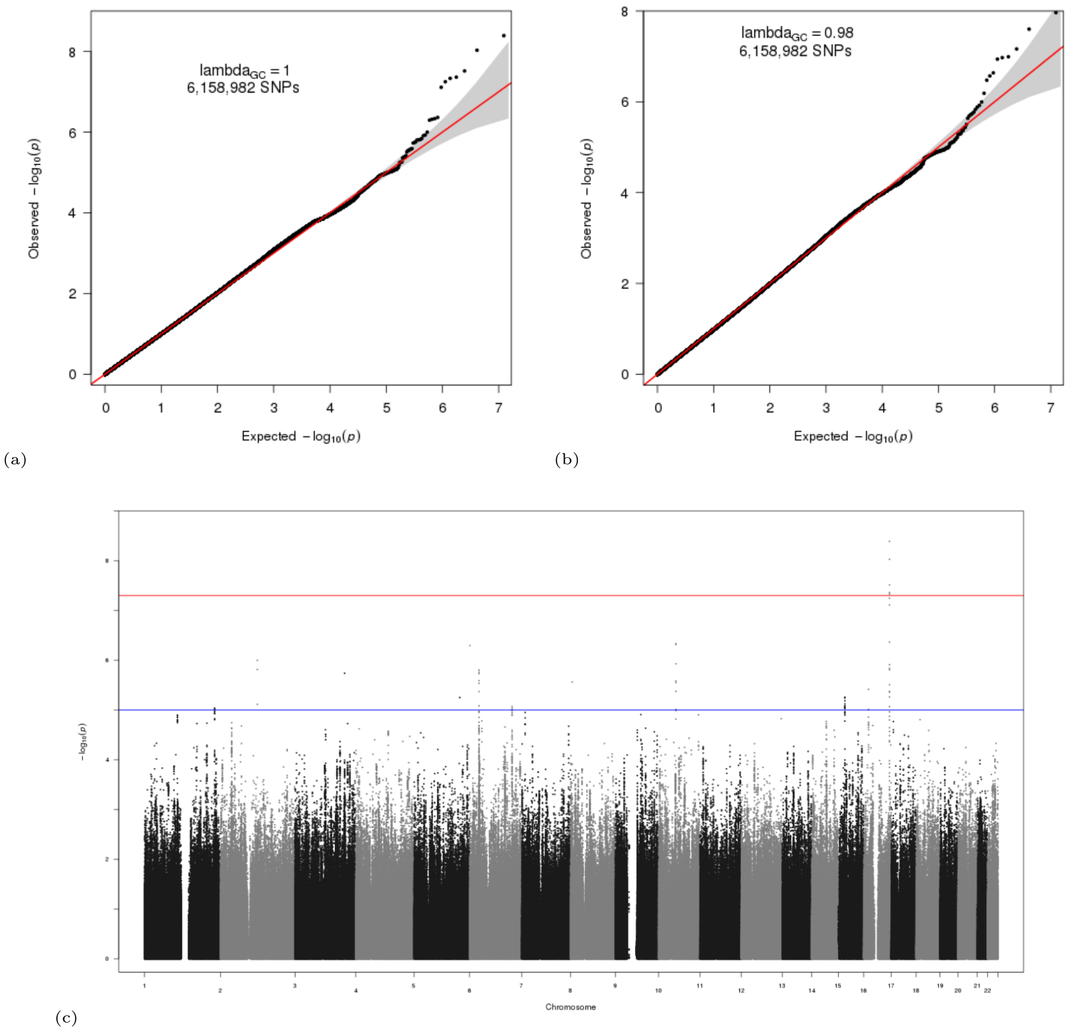
We conducted a fixed effects genome-wide meta-analysis for age of onset in a sample of 24,222 individuals of European ancestry. The quantile-quantile plot in Figure 9.1a indicates that the bulk of the test statistic distribution follows the expectation under a null hypothesis of no association ($\lambda_{GC} = 1$). It is important to note that the test statistic behaved similarly when genomic control (GC) was not applied (Figure 9.1b; $\lambda_{GC} = 0.98$). Taken together these results indicate that the meta-analysis results are robust to the slight deviations from the theoretical null model observed in some of the participating cohorts.

The Manhattan plot in Figure 9.1b displays the genome-wide association results, with a region on chromosome 16 passing the significance threshold of P $<$ 5E-08, and other suggestive signals on chromosomes 6 and 10 (with rs2249437 and rs4935127, respectively, as the top signals, both mapped to intergenic regions). Table 9.2 includes the association results and details on the SNPs that showed (suggestive) associations with p-vales below 5E-05 (displayed above the blue line in the Manhattan plot).

The genome-wide significant signals come from a set of five strongly correlated SNPs ($r^2 > 0.8$; see the regional plot around the top SNP rs1574587 in Figure 9.4a) located within the Calcium-transporting ATPase (*ATP2C2*) gene on chromosome 16. The most significant predictor of age of onset is the rs1574587 SNP (P = 4.067E-09), a SNP with a MAF ranging between 0.105 and 0.185 across the input samples (N = 23,611) and imputation quality above 0.89 (see Table 9.4 for more details on this SNP).

We note that GC had a bearing on the significance. Without GC the SNP still reaches genome-wide significance (P = 1.082E-08). The I^2 statistic for the top SNP equalled 32.6 ($\chi^2(7) = 10.391$, P = 0.17) indicating that there is no evidence of between samples heterogeneity. Given the size of the input samples, the I^2 statistic is sufficiently powerful to detect heterogeneity due to systematic differences among the studies. Furthermore, the top SNP showed the same direction of the effect in all but one of the participating cohorts. The 95% confidence intervals for the effect all include the meta-analytic estimate (and exclude zero in five samples), as illustrated in Figure 9.2.

Figure 9.1: The quantile-quantile plot based on lambdaGC corrected (a) and on lambdaGC uncorrected input files (b) and the Manhattan (c) plot of the meta-analysis results. In the Manhattan plot, the y-axis shows the strength of association (-log10(P)) and the x-axis indicates the chromosomal position. The blue line indicates suggestive significance level (P < 1E-05) while the red line indicates genome-wide significance level (P < 5E-08).



(a)

(b)

(c)

Table 9.2: SNPs showing associations above the suggestive line in the Manhattan plot (p-values < 1E-05 in the meta-analysis). *Abbreviations: RSID - rs number; Chr - chromosome; BP - base pair position; A1 - allele 1; A2 - allele 2; Freq1 - frequency of allele 1; MinFreq - minimum allele frequency; MaxFreq - maximum allele frequency; SE-standard error.*

RSID	Chr	BP	A1	A2	Freq1	Min Freq	Max Freq	BETA	SE	P
1574587	16	84453056	T	C	0.1415	0.1054	0.1853	0.0980	0.0167	4.067E-09
12922606	16	84453352	A	G	0.8585	0.8132	0.8948	-0.0952	0.0166	9.345E-09
11644628	16	84452597	T	C	0.1431	0.1145	0.1898	0.0940	0.0170	3.054E-08
11644673	16	84452771	A	G	0.8626	0.8196	0.8919	-0.0956	0.0174	4.347E-08
11644663	16	84452541	A	G	0.1465	0.1385	0.1903	0.0938	0.0172	4.667E-08
12922477	16	84453332	A	C	0.8598	0.8140	0.8948	-0.0935	0.0172	5.652E-08
79927873	16	84452497	A	C	0.1392	0.1308	0.1818	0.0943	0.0176	7.748E-08
1008994	16	84450857	C	G	0.1454	0.1020	0.1843	0.0845	0.0167	4.327E-07
4935127	10	56654986	C	G	0.7741	0.7081	0.8168	-0.0684	0.0136	4.637E-07
1733786	10	56681617	A	G	0.7742	0.6892	0.8241	-0.0685	0.0136	4.818E-07
2249437	6	1595216	T	C	0.4595	0.3977	0.4759	0.0707	0.0141	5.055E-07
62156986	2	120072326	T	G	0.9349	0.9322	0.9361	0.1925	0.0393	1.001E-06
1349893	10	56701951	T	C	0.7658	0.6856	0.8155	-0.0656	0.0135	1.177E-06
11643072	16	84451155	A	G	0.1475	0.1014	0.1812	0.0808	0.0167	1.228E-06
3943846	16	84455781	A	G	0.8146	0.7844	0.8793	-0.0762	0.0158	1.446E-06
62159383	2	120045513	T	C	0.9347	0.9342	0.9349	0.1889	0.0393	1.53E-06
2163036	16	84455766	T	C	0.1730	0.1159	0.2108	0.0785	0.0163	1.562E-06
9266245	6	31325702	A	G	0.2655	0.1537	0.2962	-0.0728	0.0152	1.568E-06
9266262	6	31325932	A	G	0.7251	0.6912	0.7835	0.0735	0.0154	1.735E-06
115259011	3	161789904	T	G	0.9563	0.9346	0.9632	-0.1446	0.0303	1.822E-06
9266244	6	31325692	A	G	0.7345	0.7038	0.8455	0.0723	0.0152	1.864E-06
141294240	6	31325822	A	G	0.7296	0.7015	0.8466	0.0709	0.0151	2.562E-06
1733762	10	56697898	A	G	0.2167	0.1664	0.3051	0.0666	0.0142	2.612E-06
28622199	8	5392103	T	C	0.8012	0.7836	0.8162	0.0712	0.0152	2.744E-06
1670812	10	56689178	T	C	0.2164	0.1664	0.3052	0.0664	0.0142	2.8E-06
2523578	6	31328542	A	G	0.7333	0.7068	0.7868	0.0718	0.0154	2.901E-06
8045313	16	84455540	T	G	0.8158	0.7819	0.8796	-0.0740	0.0159	3.091E-06
215069	16	16091237	T	C	0.0685	0.0639	0.0850	-0.1192	0.0258	3.841E-06
1733763	10	56697536	A	C	0.7839	0.6950	0.8336	-0.0653	0.0142	4.202E-06

Continued in Table 9.3

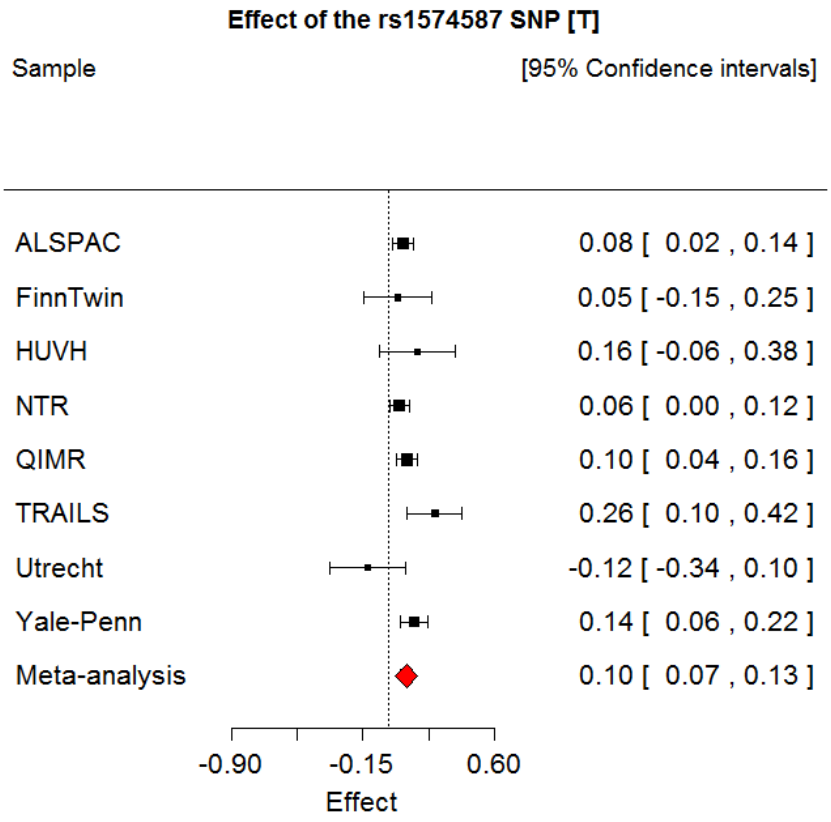
Table 9.3: *Continued from Table 9.2*

RSID	Chr	BP	A1	A2	Freq1	Min Freq	Max Freq	BETA	SE	P
55966520	16	84454043	A	G	0.1636	0.1168	0.1940	0.0766	0.0167	4.226E-06
2523582	6	31328092	A	G	0.2632	0.1581	0.2934	-0.0694	0.0151	4.253E-06
59006942	16	84454029	A	G	0.1632	0.1327	0.1979	0.0768	0.0167	4.368E-06
71386833	16	84454170	A	G	0.1636	0.1168	0.1937	0.0757	0.0167	5.471E-06
4924506	15	41129467	A	C	0.7318	0.7082	0.7827	0.0608	0.0134	5.513E-06
34659052	5	148816223	T	C	0.7351	0.7312	0.7446	0.0974	0.0214	5.592E-06
689589	15	41139250	T	G	0.2580	0.2137	0.2793	-0.0608	0.0134	5.734E-06
647930	15	41141459	A	G	0.7218	0.6899	0.7754	0.0601	0.0133	6.526E-06
2412569	15	41140159	A	G	0.2540	0.2135	0.2709	-0.0603	0.0135	7.537E-06
114529675	2	120136433	T	C	0.0614	0.0595	0.0658	-0.1771	0.0396	7.659E-06
2395475	6	31326920	A	G	0.6563	0.6112	0.7328	0.0643	0.0144	8.217E-06
2917953	15	41131916	T	C	0.2558	0.2136	0.2740	-0.0599	0.0134	8.278E-06
690660	15	41139165	T	C	0.2540	0.2134	0.2708	-0.0599	0.0134	8.306E-06
7773177	6	139143088	A	G	0.7383	0.6823	0.7564	-0.0613	0.0138	8.492E-06
2326270	16	84461051	A	C	0.0994	0.0907	0.1184	0.0896	0.0201	8.52E-06
11639292	15	41129528	A	G	0.2601	0.2140	0.2817	-0.0595	0.0134	8.781E-06
668750	15	41135827	A	T	0.7415	0.7196	0.7864	0.0595	0.0134	8.91E-06
11589605	1	230084670	A	T	0.9599	0.9513	0.9663	-0.2056	0.0463	9.133E-06
689618	15	41133008	T	C	0.2558	0.2136	0.2740	-0.0596	0.0134	9.14E-06
12193938	6	139142855	C	G	0.7383	0.6823	0.7564	-0.0611	0.0138	9.279E-06
2412570	15	41140168	T	C	0.2555	0.2136	0.2737	-0.0595	0.0134	9.304E-06
16850641	1	230097368	A	G	0.9596	0.9506	0.9663	-0.2047	0.0462	9.498E-06
7528099	1	230097883	C	G	0.9596	0.9506	0.9663	-0.2047	0.0462	9.523E-06
60064513	1	230098302	T	G	0.0404	0.0337	0.0494	0.2047	0.0463	9.624E-06
12413522	10	56642064	T	C	0.7838	0.7157	0.8267	-0.0609	0.0138	9.71E-06
73113155	1	230087856	A	G	0.0401	0.0337	0.0487	0.2049	0.0463	9.729E-06
452277	16	16079948	T	C	0.9385	0.9183	0.9465	0.1266	0.0286	9.768E-06
435683	16	16079952	T	C	0.9385	0.9183	0.9465	0.1266	0.0286	9.807E-06

Table 9.4: Association results and descriptive information for the top SNP rs1574587 in the participating cohorts. *Abbreviations: Chr - chromosome; BP - base pair position; A_1 - allele 1; A_2 - allele 2; SE - standard error; N - sample size; EAF -effect allele frequency; Info - imputation quality; λ_{GC}SE - lambda corrected standard error; λ_{GC}P- lambda corrected P-value;*

Sample	Chr	BP	A_1 A_2	BETA	SE	P	N	EAF	Info	λ_{GC}SE	λ_{GC}P
ALSPAC	16	84453056	T C	0.0858	0.0323	0.0079	6147	0.139	0.988	0.0317	0.0067
FinnTwin	16	84453056	T C	0.0535	0.0994	0.5898	1022	0.105	0.993	0.1076	0.6187
HUVH	16	84453056	T C	0.167	0.1057	0.114	580	0.139	0.967	0.1116	0.1348
NTR	16	84453056	T C	0.0697	0.0311	0.0248	5148	0.145	0.977	0.0324	0.0314
QIMR	16	84453056	T C	0.1084	0.0404	0.0072	6758	0.134	0.987	0.0362	0.0027
TRAILS	16	84453056	T C	0.2630	0.0917	0.0041	421	0.185	0.897	0.0856	0.0021
Utrecht	16	84453056	T C	-0.1270	0.1129	0.2605	1173	0.156	0.976	0.1127	0.2597
Yale-Penn	16	84453056	T C	0.1454	0.0435	0.0008	2362	0.139	0.975	0.0441	0.0009

Figure 9.2: Forest plot of the top SNP



9.3.2 Gene-based tests

The results of the SNP-based meta-analysis were subsequently used in a genome-wide gene-based analysis. We tested 24,404 genes for association with age of onset of cannabis use. Figure 9.3 gives an overview of the genome-wide results. The quantile-quantile plot (Figure 9.3a) shows that the bulk of the test statistic distribution follows the expectation under the null model, and that several genomic regions are enriched for small p-values. Note also that merely genic regions (rather than noncoding regions) are enriched for SNPs that yielded strong association signals in the single variant analysis (see Figure 9.3a).

As shown in the Manhattan plot (Figure 9.3b), three genes reached the False Discovery Rate threshold of 0.05, namely the Epithelial cell-transforming sequence 2 oncogene-like (*ECT2L*) on chromosome 6, the Calcium-transporting ATPase (*ATP2C2*) gene on chromosome 16 and the DNA repair protein RAD51 homolog 2 (*RAD51B*) gene on chromosome 14 (see Supplemental Table S4 for the SNPs assigned to each of these 3 significant genes along with information on other genes showing suggestive associations, i.e., P < 1E-04). Figure 9.4 zooms into the three significant regions.

ECT2L gene had the strongest association (P = 6.59E-08); *ECT2L* is a protein coding gene located on chromosome 6q24.1 (Figure 9.4a). The SNP with lowest p-value in the *ECT2L* gene is rs7773177 (P = 8.492E-06). The *ATP2C2* gene on 16q24.1 had the second strongest association (P = 1.54E-06; Figure 4b). The association signal yielded by the *ATP2C2* gene was also tagged in the SNP-based analysis, as the top SNP rs1574587 is located within this gene. The *RAD51B* gene (protein coding gene on 14q24.1) yielded the third strongest association signal passing the genome-wide FDR threshold (P = 5.22E-06).

As displayed in Figure 9.4c, *RAD51B* is a large gene (comprising 910,440 bases) that harbours several SNPs in low LD ($r^2 < 0.2$). The top SNP within the gene is rs17193049 (P = 3.97E-05). Table 9.5 includes descriptive information on the top associated genes along with their functions according to the Gene Ontology (GO) annotations (Ashburner, Ball et al. [23], [74]).

Figure 9.3: The quantile-quantile (a) and the Manhattan (b) plots for the gene-based tests.

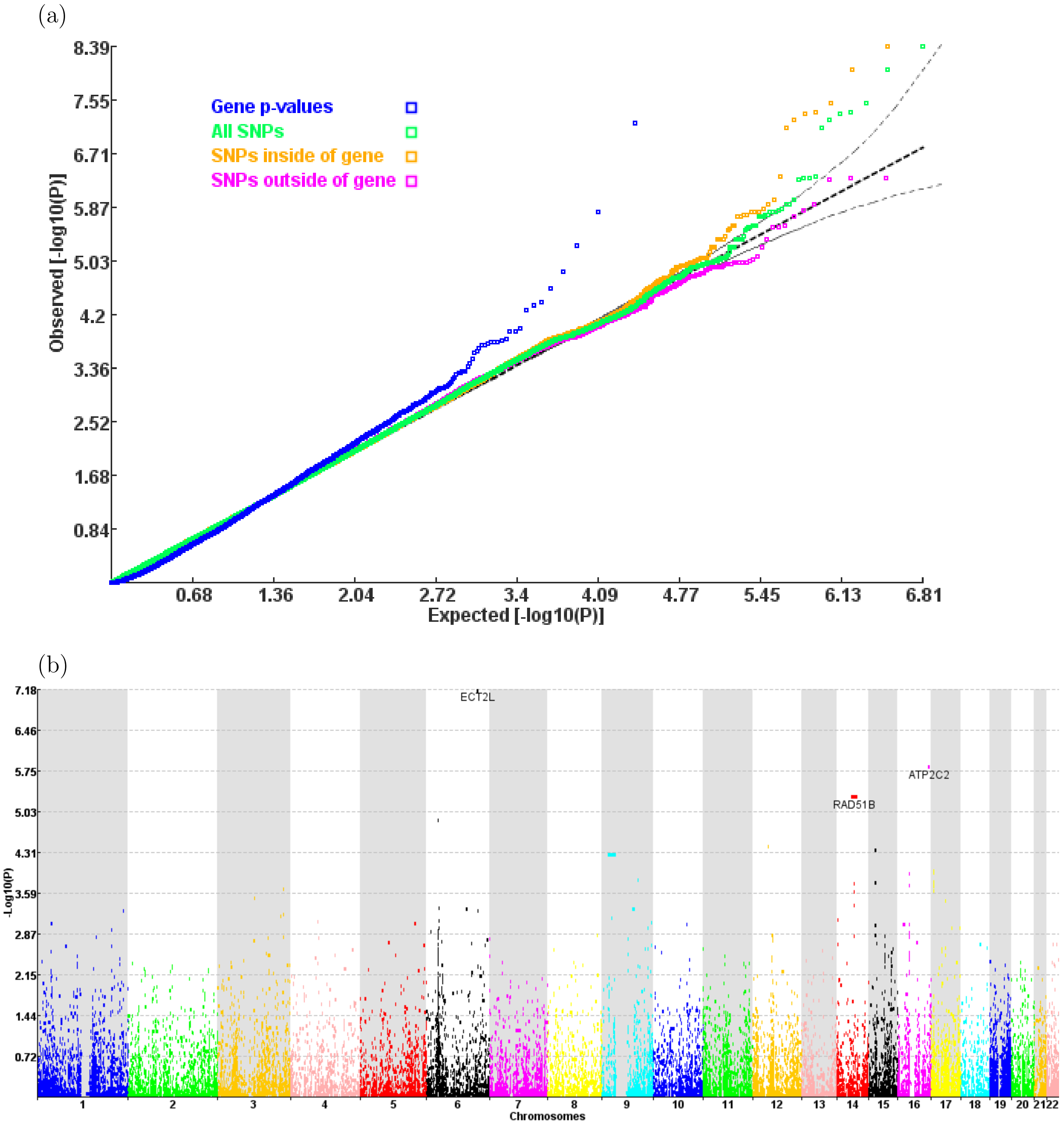
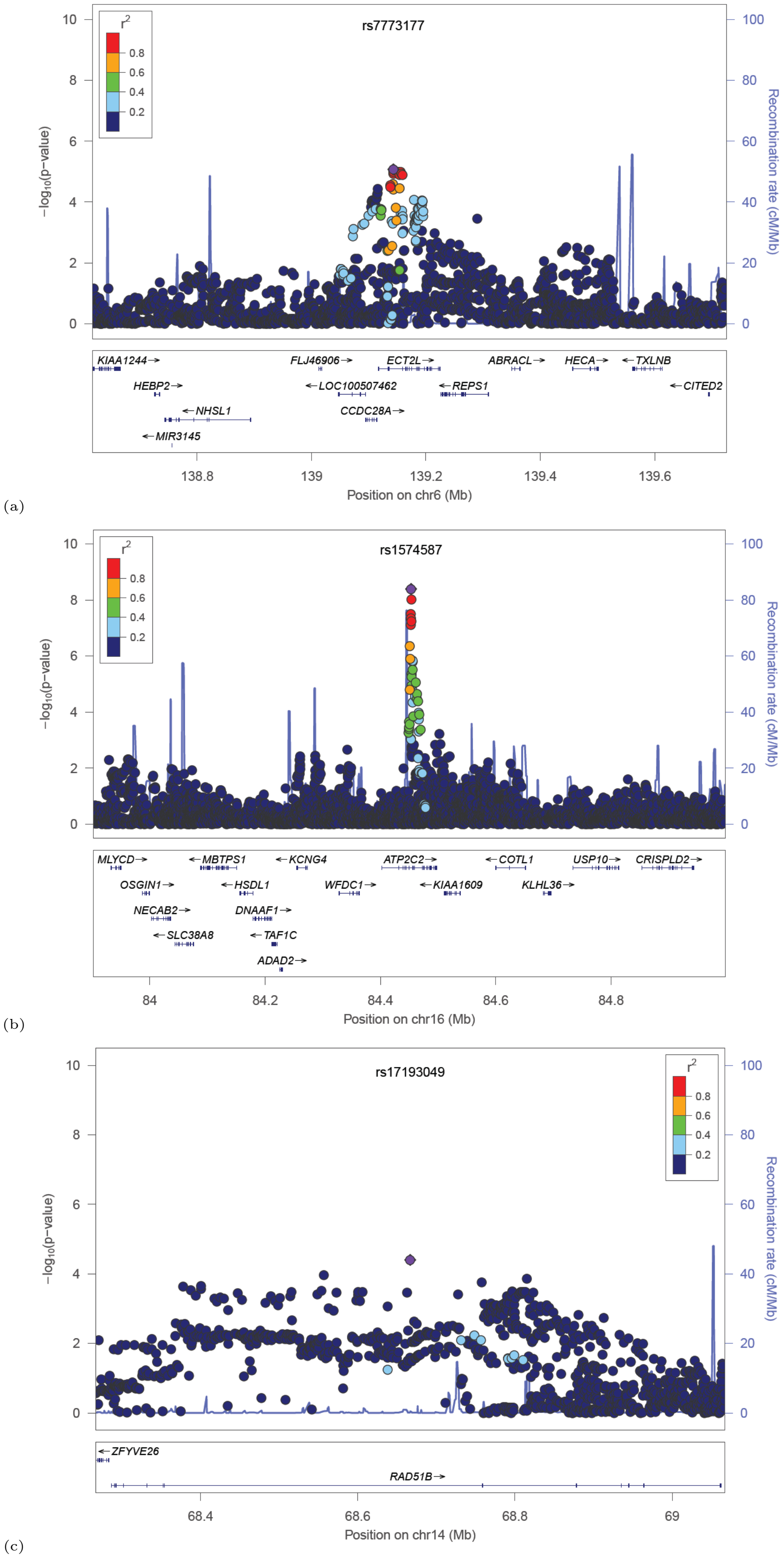


Table 9.5: Genes significantly associated with age at onset of cannabis use. Reported below are the nominal p-values and the corrected p-values based on the Benjamini and Hochberg method [29].

Gene Symbol (name)	Nominal P	Corrected P	Chr	Start Position	Group	Gene function according to GO annotations
ECT2L (Epithelial cell-transforming sequence 2 oncogene-like)	6.59E-08	1.61E-03	6	139117247	protein-coding gene	positive regulation of GTPase activity, regulation of Rho protein signal transduction, Rho guanyl-nucleotide exchange factor activity
ATP2C2 (Calcium-transporting ATPase)	1.54E-06	1.88E-02	16	84440193	protein-coding gene	calcium-transporting ATPase activity, metabolic process, calcium ion transmembrane transport, integral component of membrane, ATP binding, metal ion binding
RAD51B (DNA repair protein RAD51 homolog 2 or RAD51-like 1)	5.22E-06	4.24E-02	14	68286495	protein-coding gene	double-strand break repair via homologous recombination, ATP binding, DNA binding, Rad51B-Rad51C-Rad51D-XRCC2 complex, DNA-dependent ATPase activity

Figure 9.4: Regional plots around the significantly associated genes (a) the *ECT2L* gene (b) the *ATP2C2* gene (c) the *RAD51B* gene



(a)

(b)

(c)

9.4 Discussion

We performed the largest molecular genetic study of age of onset to date in a sample consisting of 24,222 individuals from nine cohorts. The analysis revealed genome-wide significant association with five SNPs located on chromosome 16 within the Calcium-transporting ATPase gene (*ATP2C2*). The five SNPs are in high LD ($r^2 > 0.8$), and thus are likely to represent a single independent signal. The strongest association was with an intronic variant rs1574587 (P = 4.067E-09). The gene-based tests also implicated the *ATP2C2* gene and identified two additional genes: the Epithelial cell-transforming sequence 2 oncogene-like gene (*ECT2L*) and the DNA repair protein RAD51 homolog 2 gene (*RAD51B*) located on chromosomes 6 and 14, respectively.

The *ATP2C2* gene (16q24.1) is expressed in the brain (Xiang, Mohamalawari et al. [369]) and is involved in calcium homeostasis (Newbury, Winchester et al. [256]), which in turn regulates processes including synaptic plasticity, memory and learning (Zheng and Poo [379]). Importantly, in a recent study by Gelernter, Sherva et al. [134] the *ATP2C2* gene together with the ATPase, Ca2+ -transporting, plasma membrane gene (*ATP2B2*) yielded strong association signals that implicated the calcium transport pathway in cocaine dependence (P = 0.002). Taken together these analyses suggest that the effects of *ATP2C2* are likely general rather than substance specific (noteworthy is that the calcium signalling pathway was also implicated in opioid dependence by Gelernter, Kranzler et al. [134]). Furthermore, our findings are consistent with the observed associations between early onset of cannabis use and experimentation with other drugs (Lynskey, Vink et al. [225]) and progression to escalated use/dependence (Lynskey, Agrawal et al. [222]). In other words, it is plausible that some of the same genetic factors increase both the probability to initiate early substance use, and to progress to abuse and dependence.

The *ECT2L* gene (6q24.1) displayed the strongest association signal in the gene-based analysis. This gene is involved in positive regulation of GTPase activity, i.e., the activity of heterotrimeric guanine nucleotide binding proteins (G proteins), i.e., proteins which are crucial in signal transduction across the cell membrane. Rat and in vitro addiction models hinted at the role disruptions in G proteins signaling play in the etiology of cocaine, alcohol and heroin dependence (Cami and Farre [57], Bowers [45]). Our results provide genetic support for this hypothesis (assuming the same genetic factors influence both age of onset of cannabis use and the probability to experiment and to develop dependence/abuse other drugs). The gene-based test also identified the *RAD51B* gene on 14q24.1. *RAD51B* gene (also known as *RAD51L1*) belongs to the RAD51 paralogue family, and is involved in double-strand break repair via homologous recombination, DNA and ATP binding. This gene has been previously advanced as a plausible candidate gene for nicotine dependence (Drgon, Montoya et al. [106]).

Several factors have contributed to the success of this second analysis of the

International Cannabis Consortium. First, with nine participating cohorts we gathered a sample of more than 24,000 individuals, the largest sample to date used in a genome-wide study of age of onset of cannabis use. Second, the success of both the SNP- and the gene-based analyses is likely attributable to the survival-based method that we used.

To our knowledge this is the first large meta-analysis to date that employed survival-based methods to establish genetic association with addiction phenotypes. This approach allowed us to exploit all the available information in the participating samples, and to correctly take account of the censored nature of the observations. The approach has been shown to be superior in terms of power to a logistic regression model (van der Net, Janssens et al. [333]) and to analyses limited to initiated individuals (Kiefer, Tung et al. [185]). Third, the single SNP analysis was complemented by a gene-based analysis. Taking the gene as the unit of analysis is expected to increase the statistical power relative to the single SNP analysis as this approach allows one to interrogate jointly all the SNPs within a gene, and reduces the multiple testing burden (Neale and Sham [252], Huang, Chanda et al. [166], Li, Gui et al. [202], Li, Kwan et al. [202]).

In conclusion, we have performed the largest meta-analysis to date of genome-wide studies investigating age of onset of cannabis use. With a sample of over 24,222 individuals, our results suggest the involvement of multiple correlated genome-wide significant SNPs in *ATP2C2*. The gene-based tests also identified *ATP2C2* as a significant predictor of age of onset, and in addition, implicated the *ECT2L* and the *RAD51B* genes. A thorough investigation of the functional consequences of mutations in these genes is warranted.

Chapter 10

Pathways to Smoking: Biological Insights from the Tobacco and Genetics Consortium Meta-Analysis

Abstract

By running gene and pathway analyses in the Tobacco and Genetics Consortium (TAG) sample of 74,053 individuals, we implicated twenty-one genes and forty biological pathways in several smoking behaviors. Thirteen genes are novel and were missed with the SNP-based approach in the original TAG analysis. For quantity smoked, fourteen genes passed the corrected for multiple testing false discovery rate of 0.05, and the strongest association signal was with the *IREB2* gene (P = 1.57E-37). Three genomic loci were significantly associated with ever smoking. The lowest p-value was yielded by the noncoding antisense RNA transcript *BDNF-AS* (P = 6.25E-07) on 11p14.1. The *SLC25A21* gene (P = 2.09E-08) yielded the top association signal with smoking cessation, and the signal at the 19q13.42 noncoding RNA locus exceeded genome-wide significance in the age at initiation analysis (P = 1.33E-06). The pathway analyses revealed that mutations in the *Neuronal system* pathways were the strongest predictors of quantity smoked. Especially enriched was the *Highly calcium permeable postsynaptic nicotinic acetylcholine receptors* pathway (P = 4.90E-42). Additionally, pathways belonging to 'a subway map of cancer pathways' which control appropriate mitotic DNA replication, axon growth and synaptic plasticity were enriched for mutations in smokers, and also predicted quantity smoked. The strongest association with ever smoking was yielded by the *Conversion from* APC^{Cdc20} *to* APC^{Cdh1} *in late anaphase* pathway (P = 1.61E-07), while in the quantity smoked analysis, strong enrichment signal came from the *Autodegradation of Cdh1 by Cdh1:APC/C* pathway (P = 4.28E-17). Our results shed light on the world's leading cause of preventable death and open a path to potential therapeutic targets for smoking cessation. These results are informative in decoding the biological bases of other disease traits such as cancers with which smoking shares genetic vulnerabilities.

10.1 Introduction

Tobacco smoking kills almost 6 million people each year (World Health Organization, [261]). Despite smoking prevalence decreasing in the past 30 years, there has been a steady increase in the absolute number of smokers, i.e., from 721 million in 1980 to 967 million in 2012 due to accelerated population growth (Ng, Freeman et al. [257]). It is an intriguing question why almost one billion of the world's population take up smoking while the remaining ones do not.

Genetic factors are implicated in all stages of smoking (Amos, Spitz et al. [19]), from experimentation (Vink, Willemsen et al. [347]) to dependence (Vink, Willemsen et al. [347], Lubke, Hottenga et al. [220]) and cessation (Xian, Scherrer et al. [368]). Although genome-wide association studies (GWASs; see (Wang and Li [358], Loukola, Wedenoja et al. [218], Bhler, Gin et al. [55]) for a recent overview) and especially meta-analyses conducted by large consortia (Liu, Tozzi et al. [215], Thorgeirsson, Gudbjartsson et al. [321], Tobacco and Genetics Consortium [75]) have provided several informative insights into the biological bases of smoking, the progress has been slow due to the small effects of the single nucleotide polymorphisms (SNPs), and to the multiple testing burden. Identifying such small effects largely depends on the sample size but also on the approach to analyze the genotype-phenotype relation (Li, Gui et al. [201], Sham and Purcell [295]).

Most consortia start by focusing on individual SNPs showing the strongest evidence for association. The TAG consortium with the largest sample yet of 74,053 individuals located 130 SNPs (tagging the 15q25 locus) that passed the genome-wide threshold of 5×10^{-8} in the quantity smoked analysis and next, focused on 1025 SNPs that passed the significance threshold of 10^{-4} to be included in the follow-up SNP-based analyses in the combined TAG, ENGAGE (Thorgeirsson, Gudbjartsson et al. [321]) and the Oxford-GlaxoSmithKline (Liu, Tozzi et al. [214]) sample. We note that the remaining data (summary statistics for up to 2.5 million SNPs in each of their four analyses) have remained largely unexploited.

Set-based tests (with a gene or a biological pathway as the unit of analysis) are an important alternative power-wise (Li, Gui et al. [201], Li, Kwan et al. [202]) as they consider jointly the weak effects of multiple SNPs within a gene or biological pathway. Furthermore, by targeting genomic regions rather than individual SNPs the number of tests drops from millions to thousands, thus alleviating the multiple testing burden.

Here we take the next step and further mine the publicly available TAG meta-analysis results by conducting pathway analyses and gene-based tests. Our aim is to identify biological pathways implicated in smoking behaviors (rather than to explain variance). Our analysis is not limited to the most significant loci pinpointed by the TAG meta-analysis, and serves to provide additional insights into the biological bases of smoking behaviors.

10.2 Materials and Methods

10.2.1 Sample description

The TAG Consortium investigated four phenotypes relating to different stages of smoking behaviour, i.e., ever/never smoked regularly, age at initiation, quantity smoked, and smoking cessation. The published TAG meta-analysis results (downloaded from the Psychiatric Genomics Consortium website `https://www.med.unc.edu/pgc/downloads`) combined the SNP-based summary statistics of 16 participating cohorts (of which, seven population-based cohort studies, and nine case-control studies). Across the participating studies, the mean age ranged from 39.6 to 72.3 years. More than half of the meta-analytic sample (64.37%) were females. Data on smoking status (ever/never smoked regularly) were available in all participating studies. The percentage of ever smokers varied from 37.7% to 75.2% (weighted average 56.58%), where ever smokers declared to have smoked ≥ 100 cigarettes. Data on age at initiation (i.e., age at first cigarette or age when started to smoke regularly) was available in thirteen cohorts, with averages falling within the 17 – 32.3 years range. Thirteen cohorts also had observed data on quantity smoked (i.e., the average, or the maximum number of cigarettes smoked per day); across these cohorts, the mean quantity smoked varied between 13.1 and 23.4 cigarettes per day. For more details on the characteristics of the participating samples, we refer to Table 1 (see `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2914600/table/T1` in [322]).

10.2.2 Statistical analyses

We lifted the SNP positions from HapMap 2 NCBI build 36, Human Genome references 18 to the Human Genome references 19 positions based on dbSNP build 138. Following conversion, there were meta-analysis results available for 2,342,540 SNPs; 2,339,474 SNPs; 2,340,171 SNPs and 2,341,140 SNPs obtained in the fixed-effect meta-analysis of quantity smoked, ever smoking , smoking cessation , and age at initiation, respectively. The SNPs passed the TAG consortium's quality criteria (see Supplementary Table 3 in (Tobacco and Genetics Consortium [322])).

Analyses were carried out using the HYbrid Set-based Test (HYST) as implemented in the Knowledge-based mining system for Genome-wide Genetic studies software (Li, Gui et al. [201], Li, Kwan et al. [202]). In the HYST approach, the SNPs within the tested genes/pathways are firstly grouped in blocks (sets) based on linkage disequilibrium (LD) information, such that the resulting blocks are weakly correlated. Next, for each of these blocks a p-value is obtained with the Gene-based Association Test that uses the Extended Simes procedure (Li, Gui et al. [201]) by taking account of the correlatednes among SNPs. Lastly, HYST employs the scaled chi-square test to combine these LD-block-based p-values in order to obtain an overall p-value for the gene/pathway. The LD blocks were inferred based on the 1000 Genomes Haplotypes Phase 3 reference panel ALL (downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`); including this information to construct the gene-based tests is expected to show good control of the type I error rate (Li, Gui et al. [201]). We opted for the 1000 Genomes as this is the highest resolution genetic map to date. Both the univariate gene-based tests and the pathway analyses were conducted genome-wide and utilized the unweighted form of HYST. Variants missing LD information were discarded.

10.3 Results

10.3.1 Gene-based tests

The gene-based analyses revealed twenty-one genes significantly associated with smoking behaviours, including nine genes in the 15q25 region. Fifteen genes of these were not reported previously as significantly associated with any smoking or addiction behavior in the GWAS catalogue (as of 2015–10–23, see `www.genome.gov/gwastudies` [359]), and were missed with the SNP-based approach in the original TAG analysis. Results are displayed in Tables 10.1 to 10.4 .

Table 10.1: Genes implicated in quantity smoked . Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23; see `www.genome.gov/gwastudies` [359])

Symbol	NominalP	CorrectedP	Chromosome	Start_Position	Group
IREB2	1.57E-37	3.79E-33	15	78730517	protein-coding gene
CHRNA3	6.45E-34	7.78E-30	15	78887646	protein-coding gene
CHRNA5	3.17E-32	2.55E-28	15	78857861	protein-coding gene
HYKK	2.14E-30	1.29E-26	15	78799905	protein-coding gene
CHRNB4	3.87E-28	1.87E-24	15	78916635	protein-coding gene
PSMA4	6.47E-23	2.60E-19	15	78832746	protein-coding gene
ADAMTS7	4.48E-17	1.55E-13	15	79051544	protein-coding gene
MORF4L1	1.18E-06	3.55E-03	15	79165330	protein-coding gene
SHISA4	1.52E-06	4.07E-03	1	201857804	protein-coding gene
IPO9	5.45E-06	1.32E-02	1	201798287	protein-coding gene
HLA-DPB1	7.65E-06	1.68E-02	6	33043702	protein-coding gene
HLA-DPA1	9.52E-06	1.92E-02	6	33032345	protein-coding gene
SLC25A21	1.41E-05	2.61E-02	14	37147125	protein-coding gene
LOC646938	1.66E-05	2.86E-02	15	79044378	Unknown

Table 10.2: Genes implicated in ever smoking . Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23 [359])

Symbol	NominalP	CorrectedP	Chromosome	Start_Position	Group
BDNF-AS	6.25E-07	1.51E-02	11	27528398	non-coding RNA
APBB2	2.43E-06	2.44E-02	4	40812043	protein-coding gene
CDC27	3.04E-06	2.44E-02	17	45195062	protein-coding gene

Table 10.3: Genes implicated in smoking cessation. Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23 [359])

Symbol	NominalP	CorrectedP	Chromosome	Start_Position	Group
SLC25A21	2.09E-08	5.04E-04	14	37147125	protein-coding gene
SEMA6D	1.64E-06	1.99E-02	15	48010685	protein-coding gene

Fourteen genes passed the corrected for multiple testing False Discovery Rate (FDR) of 0.05 [29] in the gene-based analysis of quantity smoked (see Table 10.1), with the *IREB2* gene on 15q25.1 yielding the lowest p-value (P = 1.57E-37). Three genomic loci exceeded genome-wide significance in the ever smoking analysis (Table 10.2). The strongest signal was yielded by the noncoding antisense RNA transcript *BDNF-AS* (P = 6.25E-07) on 11p14.1. Note also that the *CDC27* gene, harbouring the rs16941640 SNP with the lowest p-value in the initial TAG analysis (P = 2.2E-07), achieved significance in our gene-based analysis (P = 3.04E-06). Two genes were significantly associated with smoking cessation (see Table 10.3), with the *SLC25A21* gene on 14q13.3 yielding the strongest signal (P = 2.09E-08). The gene-based tests revealed the non-coding RNA *MIR1323-MIR512-1-MIR512-2* locus on 19q13.42 as significantly associated with age at initiation (P = 1.33E-06; Table 10.4). Note that no individual SNP reached significance in the original TAG analysis of age at initiation. We refer to the Supplementary Table S1 for details on gene functions and to the Tables S2-S5 for details on SNPs assigned to genes.

10.3.2 Pathway-based tests

Results for the statistically significant pathways are summarized in Tables 10.5 to 10.12. Details on genes assigned to the pathways are given in Supplementary Tables S5 and S6. Note that some of the genes are assigned to more than one pathway, indicating that they play roles in multiple biological processes under-

Table 10.4: Genes implicated in age at initiation . Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23 [359])

Symbol	NominalP	CorrectedP	Chromosome	Start_Position	Group
MIR1323	1.33E-06	1.21E-02	19	54175221	non-coding RNA
MIR512-1	1.50E-06	1.21E-02	19	54172416	non-coding RNA
MIR512-2	1.50E-06	1.21E-02	19	54172410	non-coding RNA

Table 10.5: Biological pathways implicated in quantity smoked. The Neuronal system chain.

Pathway	P (HYST)	Total genes
Highly calcium permeable postsynaptic nicotinic acetylcholine receptors	4.90E-42	11
Acetylcholine binding and downstream events	3.11E-41	14
Presynaptic nicotinic acetylcholine receptors	8.42E-32	12
Neuroactive ligand receptor interaction	2.66E-22	249
Transmission across chemical synapses	2.13E-20	174
Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell	2.87E-20	127
Neuronal system	1.43E-19	263

lying the smoking behaviours. The reported p-values passed the corrected for multiple testing FDR of 0.05.

As indicated in Tables 10.5 to 10.12, note that the pathways reaching statistical significance form chains of pathways, rather than being isolated hits spread across the whole genome. We identified thirty-five biological pathways significantly associated with quantity smoked (see Table 10.5). Highly enriched are the *Neuronal system* pathways harbouring the nicotinic acetylcholine receptor genes expressing the α (*CHRNA* 1-9), β (*CHRNB* 1-4), γ, δ and ϵ subunits; the strongest association was with the *Highly calcium permeable postsynaptic nicotinic acetylcholine receptors* pathway (P = 4.90E-42). Furthermore, quantity smoked was statistically associated with pathways regulating the immune system (with the *Cross presentation of soluble exogenous antigens endosomes* pathway showing the strongest association; P = 4.38E-17), metabolism (where the *Regulation of ornithine decarboxylase ODC* pathway yielded the strongest signal; P = 6.71E-16),

Table 10.6: Biological pathways implicated in quantity smoked. The cell-cycle chain.

Pathway	P (HYST)	Total genes
P53-dependent G1 DNA damage response	7.24E-20	52
SCF-Skp2 mediated degradation of p27/p21	7.59E-19	52
Cyclin E associated events during G1/S transition	2.35E-18	61
Autodegradation of the E3 ubiquitin ligase COP1	3.37E-18	46
Autodegradation of Cdh1 by Cdh1:APC/C	4.28E-17	53
CDK-mediated phosphorylation and removal of Cdc6	5.70E-17	45
APC/C^{Cdh1} *mediated degradation of Cdc20 and other APC/C*^{Cdh1} *targeted proteins in late mitosis/early G1*	1.99E-16	61
P53-independent G1/S DNA damage checkpoint	2.02E-16	47
CDT1 association with the Cdc6:ORC:origin complex	2.03E-16	53
SCF (beta-TrCP) mediated degradation of EMI1	6.24E-16	48
APC/C Cdc20 mediated degradation of mitotic proteins	4.53E-14	62
Assembly of the pre-replicative complex	6.46E-17	61
Proteasome	5.95E-18	44
ORC1 removal from chromatin	3.31E-16	63
Regulation of apoptosis	1.84E-09	54

Table 10.7: Biological pathways implicated in quantity smoked. The immune system chain.

Pathway	P (HYST)	Total genes
Cross presentation of soluble exogenous antigens endosomes	4.38E-17	44
Activation of NF-kB in B cells	6.72E-14	60
Translocation of ZAP-70 to immunological synapse	3.90E-04	13
Phosphorylation of CD3 and TCR zeta chains	1.64E-03	15
Allograft rejection	7.50E-04	34
ER phagosome pathway	7.12E-14	57
Interferon signalling	1.70E-03	148

Table 10.8: Biological pathways implicated in quantity smoked. Metabolism.

Pathway	P (HYST)	Total genes
Regulation of ornithine decarboxylase ODC	6.71E-16	47
Metabolism of amino acids and derivatives	1.80E-05	187

Table 10.9: Biological pathways implicated in quantity smoked. Disease.

Pathway	P (HYST)	Total genes
Vif-mediated degradation of APOBEC3G	9.55E-18	48

Table 10.10: Biological pathways implicated in quantity smoked. Signal transduction.

Pathway	P (HYST)	Total genes
Signaling by WNT	1.18E-17	61

Table 10.11: Biological pathways implicated in quantity smoked. Gene expression.

Pathway	P (HYST)	Total genes
Destabilization of mRNA by AUF1 (hnRNP D0)	2.21E-15	49
Asthma	3.10E-04	27

Table 10.12: Biological pathways implicated in ever smoking. Regulation of the mitotic cell-cycle chain.

Pathway	P (HYST)	Total genes
Conversion from APC^{Cdc20} *to* APC^{Cdh1} *in late anaphase*	1.61E-07	14
Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components	2.19E-07	16
Phosphorylation of the APC/C	4.54E-07	15
APC^{Cdc20} *mediated degradation of Cyclin B*	1.13E-06	17
APC^{Cdc20} *mediated degradation of Nek2A*	2.30E-06	19

signal transduction (with the *Signaling by WNT* pathway exceeding the significance threshold; P = 1.18E-17) and *Asthma* pathways (P = 3.10E-04).

Additionally, particularly enriched are the cell-cycle checkpoints pathways governing the transition of the new cell from one stage to another and the programmed cell-death, with strong association signals coming from the *P53-dependent G1 DNA damage response* pathway (P = 7.24E-20) and from the pathway regulating apoptosis (P = 1.84E-09). Note that the cell-cycle pathways are implicated also in ever smoking (see Tables 10.5 to 10.12). These pathways relate to the *Anaphase-Promoting Complex/Cyclosome-mediated degradation of Cdc20 and other APC/*C^{Cdh1} *targeted proteins.* Although none of these pathways is associated with both smoking phenotypes, they are inter-connected, forming chains of pathways governing different stages of cell division.

Especially enriched in ever smokers is the *Conversion from* APC^{Cdc20} *to* APC^{Cdh1} *in late anaphase* pathway (P = 1.61E-07), while in quantity smoked, enrichment signals come from the *Autodegradation of Cdh1 by Cdh1:APC/C* pathway (P = 4.28E-17) and from the *APC/*C^{Cdh1}*-mediated degradation of Cdc20 and other APC/*C^{Cdh1}*-targeted proteins in late mitosis/early G1* pathway (P = 1.99E-16).

10.4 Discussion

We have extended the original TAG analysis and the previous work that has implicated SNPs and genes in smoking behaviours. Our results may guide future genome-wide analyses as they demonstrate that complementing the SNP-based tests by gene- and pathway-based analyses can lead to considerable gains in statistical power and can yield important insights into the biological mechanisms underlying the trait of interest.

The gene-based analysis revealed twenty-one genes implicated in smoking behaviours. Of these, thirteen are novel and were missed with the SNP-based approach in the original TAG analysis. Aside from the known cluster of genes (*IREB2-CHRNA3-CHRNA5-CHRNB4-HYKK-PSMA4*), we identified a cluster of three loci on the same chromosome 15 – *ADAMTS7* (P = 4.48E-17), *MORF4L1* (P = 1.18E-06) and *LOC646938* (P = 1.66E-05). The *ADAMTS7-MORF4L1* locus has been previously associated with e.g., coronary artery disease [73] for which smoking is a known risk factor. Based on a joint-analysis, SNPs within *ADAMTS7* and *MORF4L1* were recently listed as candidate signals for smoking behaviour independent of those yielded by the known loci rs16969968 or rs588765 in *CHRNA5* (SchwantesAn, Culverhouse et al. [293]). Regarding the two significant associations on chromosome 1 (i.e., *SHISA4* and *IPO9* genes, P = 1.52E-06 and 5.45E-06, respectively), we note that neither of these have been associated with any addiction behaviour. We also located statistically significant signal coming from the known *HLA* locus on chromosome 6, result suggesting a link between smoking and the immune system. Loci in the *HLA* region have been previously implicated in e.g., schizophrenia [76], with which smoking dependence shares genetic vulnerabilities.

Three genomic loci were significantly associated with ever smoking. The strongest signal came from the noncoding antisense RNA transcript *BDNF-AS* locus (P = 6.25E-07) on chromosome 11. *BDNF-AS* downregulates the *BDNF* gene (Lipovich, Dachet et al. [208]) which in turn has a key role in regulating neuronal growth and synaptic plasticity (Lipovich, Dachet et al. [208], Modarresi, Faghihi et al. [247]). The *BDNF* gene did not reach genome-wide significance in our gene-based analysis (P = 0.00013), result indicating that the eight significant SNPs reported in the combined TAG analysis were tagging better the non-coding RNA *BDNF-AS* locus.

Association signals at two genes - at the *SLC25A21* gene on chromosome 14 and at the *SEMA6D* gene on chromosome 15 (P = 1.64E-06) exceeded genome-wide significance with smoking cessation. The role these genes play in smoking cessation and their potential as therapeutic targets has yet to be elucidated. *SEMA6D* expression levels were recently shown to significantly predict survival in patients with breast cancer (Chen, Li et al. [63]); also the gene was listed among potential targets in cancer therapy owing to its role in regulation of the immune response (Tamagnone [313]). The *SLC25A21* gene was previously proposed as a plausible candidate for smoking cessation (Uhl, Drgon et al. [324]) and current smoking (Vink, Smit et al. [346]). Notice that the *SLC25A21* gene was the only one significantly associated with two smoking behaviours, namely with quantity smoked (P = 1.41E-05) and smoking cessation (P = 2.09E-08).

The 19q13.42 *MIR1323-MIR512-1-MIR512-2* noncoding RNA locus exceeded genome-wide significance in the age at initiation analysis (P = 1.33E-06). Importantly, loci in the 19.q13 region were previously implicated in quantity smoked by the TAG Consortium in their combined analysis, and also by the ENGAGE (Liu, Tozzi et al. [214]) analysis ; the region has been proposed as a plausible candidate for further investigation in relation to smoking quantity given its role in nicotine metabolism (see e.g., (Tobacco and Genetics Consortium [322])). We now add age at initiation to the list of smoking phenotypes to be further investigated in relation to this biologically plausible region.

By extending the tested genomic region from genes to pathways – where a pathway is a 'meta-gene' comprising multiple genes having the same biological function – we identified chains of pathways statistically associated with ever smoking and quantity smoked. The strongest association with quantity smoked was yielded by the *Highly calcium permeable postsynaptic nicotinic acetylcholine receptors* pathway (P = 4.90E-42) harboring the nicotinic acetylcholine receptors. The results are consistent with the hypothesis that mechanisms underlying smoking dependence involve the mesocorticolimbic dopamine system (Benowitz [30], Wang and Li [358]). We implicated in quantity smoked and ever smoking pathways relating to the *Anaphase-Promoting Complex/Cyclosome mediated degradation of Cdc20 and other APC/C^{Cdh1}-targeted proteins.* Although none of these pathways was associated with both smoking phenotypes, they are inter-connected, forming chains of pathways governing different stages of cell division. These pathways control not only the mitotic regulators of DNA replication (i.e., APC^{Cdc20}) but also axon growth and synaptic plasticity (i.e., APC^{Cdh1}) (Li and Zhang [203]); hence our results lend support to the idea that neuronal plasticity and learning play a paramount role in the development of nicotine addiction (Benowitz [30]). Because the cell-cycle pathways are known to belong to 'a subway map of cancer pathways' (Hahn and Weinberg [150]) given their role in cancer development (a disease of unregulated cell proliferation), our results suggest that (as first conjectured by Fisher (Fisher [126])) some of the same biological mechanisms underlie both smoking and cancer. Finally, these results suggest that targeting cell-cycle regulators (as novel cancer therapies do (Diaz-Moralli, Tarrado-Castellarnau et al. 2013 [99])) might work in smoking cessation therapy.

As detailed in the Supplementary Tables S6 and S7, our results provide clues on many other genes sharing the relevant pathways, genes that act in concert to give rise to individual differences in smoking behaviours. The genes and pathways reported herein worth further investigation in relation to other addiction phenotypes and disease traits such as schizophrenia or cancers with which smoking shares genetic factors (de Viron, Morr et al. [93]).

As it is based on gene-level statistic obtained in a meta-analytic sample, our study overcomes the limitations of previous exploratory pathway studies (Wang and Li [358], de Viron, Morr et al. [93], Liu, Fan et al. [216]) such as e.g., the literature selection bias, the heterogeneity problem or the bias arising from the use of incongruent analysis protocols across the selected studies. More importantly, we use a hypothesis-free/unbiased genome-wide approach in deriving the list of genes to be included in the pathway enrichment analyses. In so doing, our study surmounts the 'circularity' bias of previous pathway studies (Wang and Li [358], Liu, Fan et al. [216]) built on lists of input genes mostly derived from published confirmatory/candidate gene studies.

Furthermore, by being based on HYST (Li, Gui et al. [201]) – which employs the scaled chi-square test to combine the LD-block-based p-values calculated with the GATES procedure – the study presented herein overcomes the limitations of other pathway-based tests such as e.g., overrepresentation tests which fail to take into account the correlation structure among the genes within pathways and among the SNPs within genes. Yet, in large samples as the one we used, results obtained based on alternative approaches are expected to converge. To check this, we re-run the pathway-based analyses with the PANTHER classification system (`http://www.pantherdb.org`), using an overrepresentation test [243]. We provided as input lists of genes selected based on a p-value threshold of 10^{-3}. Consistent with the HYST analysis, results revealed that pathways belonging to the Neuronal system chain were the top ranked pathways in the quantity smoked analysis, whilst pathways belonging to the cell-cycle chain were significantly enriched in the ever smoking analysis (see Supplementary Tables S8 and S9).

Our results shed light on the world's leading cause of preventable death and open a path to potential targets for therapeutics. Using the largest sample amassed yet in a GWAS of smoking we found that aside from the nicotinic acetylcholine receptors – known for the rewarding role they play in nicotine dependence – the cell-cycle regulators are possible targets in smoking cessation therapy. These results are informative in decoding the biological bases of other addiction phenotypes and disease traits such as schizophrenia and cancers with which smoking shares risk loci and biological pathways.

Chapter 11

Summary

Multivariate data may confer power advantages in GWAS, yet multivariate data require modeling choices. **Chapter II** compared the efficiency (in terms of power) of several analytic strategies to detect a genetic variant in multivariate phenotypic data. Twin data were simulated to fit exactly the following five models: 1) single common genetic factor, 2) a correlated genetic common factors model, 3) a latent regression model, 4) a hybrid simplex (AE) factor (C) model, and 5) a stationary double simplex (AE) model. The effect of the genetic variant on all or a subset of the phenotypes was mediated by the common genetic factor(s). In twin 1 data the following analytic strategies were considered: a) univariate tests in which each phenotype was regressed on the genetic variant (single phenotype ANOVA); b) univariate tests based on sum scores (ANOVA); c) exploratory factor analysis (EFA), in which the common factors were regressed on the genetic variant; c) multivariate tests based on MANOVA, in which all phenotypes were concurrently regressed on the genetic variant. Power calculations were based on the non-centrality parameter (NCP). Results demonstrated that: a) the sum scores ANOVA and the exploratory factor analysis were the most powerful strategies when the genetic effect was general, i.e., propagated in all phenotypic indicators, while MANOVA was the least powerful in this circumstance; b) MANOVA and EFA were particularly powerful when the genetic variant was propagated in a subset of phenotypes, and their power increased with increasing phenotypic correlations; c) the NCPs of MANOVA and EFA were equal across all scenarios indicating that the differences in power between the two strategies arisen from the differences in degrees of freedom.

Family-based genotype imputation was proposed as a means of increasing power in GWAS, as it allows for the inclusion into association analysis of individuals with observed phenotypes but missing genotypes. **Chapter III** considered factors affecting the power to detect genetic association following family-based genotype imputation. The study focused on sibships of sizes 2 to 4, where imputation was informed by 1 sibling, or by 1 sibling and 1 parent. Monte Carlo simulations were used to compare the power of the mixture approach (involving the full distribution of the imputed genotypes) with the power of the dosage approach (where the mean of the conditional distribution featured as the imputed genotype). Furthermore, the effect on power and type I error rates of misspecification of the familial covariance matrix was considered given low, moderate and highly heritable traits. Misspecification pertained to the use of an exchangeable model which accounts for the sibling correlations by means of a single correlation (a model of interest also for computational reasons). Finally, the simulation results were verified in two empirical datasets. Results showed that: a) the power differences among the dosage and the mixture approaches are quite small and recommend the use of the dosage approach because it is computationally easier; b) correct model specification is desirable particularly when the trait is highly heritable in order to yield correct type I error rates; c) lastly, it was showed that family-based imputation yields considerable power gains only in specific circumstances.

Full, correct modeling of the conditional familial covariance matrix confers power advantages and yields correct type I error rates. Yet, correct modeling can be complicated and subject to misspecification when families are variable in size and composition. Model misspecification - as discussed in chapter III - is also of interest for computational reasons. **Chapter IV** focused on the effect on power of misspecification of the familial covariance matrix and considered several sandwich corrections of the standard errors to ensure correct type I error rates in family based GWASs. Specifically, the performance of the unweighted least squares (ULS) and of the maximum likelihood estimators (ML) was compared given: a) AE and ACE traits simulated in families comprising 4 siblings (2 MZ/DZ twins and 2 siblings), with and without parents, and b) various background correlations. Results demonstrated that the extreme misspecification employed by the sandwich corrected ULS procedure implemented in Plink leads to a dramatic loss in power given moderate to large background correlations. Furthermore, it was shown that the fast ML procedure is equally amenable to a sandwich correction. To analyze A(C)E traits in samples consisting of families varying in size and composition (when full, correct modeling is complicated and subject to misspecification), a misspecified CE/AE linear mixed model in combination with a sandwich correction is likely to maintain the power close to that of a correctly specified (yet, computationally more demanding) background model.

Monozygotic twin pairs represent a considerable part of the samples collected at the twin registries. **Chapter V** evaluated in terms of power and type I error rates the practice of dropping one individual of an MZ twin pair from family-based genome-wide association analyses. Simulation results demonstrated that including both MZ twins of a pair in GWASs yields calibrated type I error rates and increases the effective sample size and so, it increases power. It was illustrated how the power gain varies as a function of the phenotypic correlation. Finally, several modeling alternatives suitable for family-based samples including MZ twin pairs were discussed.

Rare variants are hypothesized to explain an important proportion of the variance in complex psychiatric traits. **Chapter VI** focused on tests of association with rare variants. Monte Carlo simulations were used to assess the effect of weight misspecification on the type I and type II error rates of the likelihood ratio test and of the sequence kernel association test (SKAT). Results showed that the LRT is generally robust to weight misspecification, while there are specific circumstances in which the power of the score test is far from adequate to begin with. To optimize the power of detection, a weighting procedure was proposed, and its power benefits were evaluated in simulated and empirical data. The power studies conducted herein informed the application studies aimed at identifying genes and biological pathways implicated in cannabis use initiation and smoking behaviors. **Chapter VII** aimed to estimate the heritability of cannabis initiation based on recently developed methods. Next, the chapter focused on locating genes underlying the heritability of cannabis use initiation and age at onset. This is among the first studies in the literature that used genotypes imputed based on a population specific reference panel (i.e., the Genome of the Netherlands reference panel). The study demonstrated that there is significant association signal coming from the currently measured (and imputed) SNPs. Furthermore, the study showed that cannabis use initiation is a polygenic trait, subject to the influences of many genetic variants of small effect, uniformly distributed over the genome.

Chapter VIII continued the searches for genes associated with cannabis use initiation in a meta-analytic sample of 32,330 individuals from 13 cohorts from Europe, United States and Australia. This GWAS is the first in the literature to locate genomic loci that significantly predict initiation of cannabis use.

Chapter IX employed survival-based methods to identify genetic variants that predict age at onset of cannabis use in the International Cannabis Consortium meta-analytic sample of 24,222 individuals from 9 cohorts.

Chapter X was based on the observation that although the SNP-based tests are still underpowered to detect the small genetic effects in the current samples, the largest to date meta-analysis of smoking behaviors conducted by the Tobacco and Genetics Consortium focused exclusively on testing individual SNPs (i.e., on 1052 SNPs). The unexploited TAG results (up to ∼2.5 million SNPs for four smoking behaviors) were further mined by using set-based tests. This powerful approach located twenty-one genes and forty biological pathways statistically associated with quantity smoked, smoking initiation, age at initiation and smoking cessation. Results showed that: a) pathways harbouring genes regulating neuronal plasticity and learning play an important role in the development of smoking dependence; b) the cell-cycle regulators, metabolism and the immune system are also implicated in smoking dependence; c) some of the same biological mechanisms underlie both smoking and cancer (as first conjectured by Fisher in 1959). This is the first study based on an unbiased/hypothesis free testing approach that reports biological pathways statistically associated with smoking behaviours.

Chapter 12

Discussion

Abstract

The aim of the present thesis was two-fold: first, to study and select from the pool of available statistical methods the most powerful and computationally efficient ones for conducting common and rare variant association studies; second, using powerful methods to identify genes and biological pathways associated with early stages of cannabis use and smoking behaviors. Below I first discuss the main empirical findings I have contributed to the field, and next I elaborate on the implications my power analyses carry over to future rare and common variant association studies. Each of the two parts ends with a conclusion.

12.1 Empirical Findings: Genes and Biological Pathways Implicated in Smoking Behaviors and Initiation of Cannabis Use

The empirical analyses revealed several important insights into the biological mechanisms underlying cannabis use and smoking behaviors. First, based on a sample of distantly related individuals from the Netherlands Twin Register and the Genome-wide Complex Trait Analysis method (Yang, Lee et al. [375]), I provided evidence that the currently typed single nucleotide polymorphisms collectively explain ∼25% of the variance in cannabis use initiation (95%CI[7.7,42.2]; Chapter VII). This finding reaffirmed that initiation of cannabis use is a heritable trait as established by previous twin studies. The result was next confirmed using the So et al. method (So, Li et al. [299]) which yielded the close heritability estimate of 20% ($P < 0.001$) based on the large meta-analytic sample of the International Cannabis Consortium with subjects from Europe, United States and Australia. These results motivated the continued searches for genes and biological pathways underlying the trait heritability.

The first five genetic loci that significantly predict initiation of cannabis use were identified based on the large sample of the International Cannabis Consortium (Chapter VIII). The strongest association was with the Neuronal Cell Adhesion Molecule 1 (*NCAM1*) gene on 11q23.1, followed by the Cell Adhesion Molecule 2 (*CADM2*) on 3p12.1 (2.13E-06), two loci on chromosome 4 – the Short Coiled-Coil Protein (*SCOC*) and the non-coding *SCOC* Antisense RNA 1 (*SCOC-AS1*) on 4q31.1, and lastly, the Potassium Channel, Sodium Activated Subfamily T, Member 2 (KCNT2) gene on 1q31.3 (P = 7.85E-06, 5.76E-06, and 9.38E-06, respectively). The top association – the *NCAM1* gene – is a cell-adhesion molecule implicated in regulation of synaptic plasticity and axonal regeneration, as well as in regulating memory formation (Sheng, Leshchyns'ka et al. [296]). This finding reinforces the idea that synaptic plasticity, memory and learning are essential to the development of addiction behaviors (Uhl, Liu et al. [326], Benowitz [30]). The *NCAM1* gene belongs to the NTAD cluster spanning 521 kb in the 11q22-23 region which includes in addition the *TTC12*, the *ANKK1* and the *DRD2* genes (Mota, Araujo-Jnr et al. [249]). Comparative analyses indicate that the cluster has been highly conserved for more than 400 million years likely because of its essential role in dopaminergic transmission and in the development of the central nervous system (Mota, Araujo-Jnr et al. [249]). As suggested by Mota et al. (Mota, Araujo-Jnr et al. [249]), extending the research focus to the surrounding region is probably required in order to grasp a complete characterization of the role the locus plays in psychiatric traits such as substance use. An interesting observation is that SNPs in the *NCAM1* gene were also implicated in bipolar disorder (Atz, Rollins et al. [24]) and schizophrenia (Atz, Rollins et al. [24], Sullivan, Keefe et al. [311], Uhl and Drgonova [325]), disease traits known to be comorbid with cannabis use (see e.g., [85]). In addition, loci at the haplotype harboring the *NCAM1* gene were previously associated with several other addiction phenotypes such as nicotine (particularly loci tagging the neighboring genes *ANKK1* and *TTC12* genes, see (Gelernter, Yu et al. [137]), alcohol (Yang, Kranzler et al. [372]), heroin dependence (Nelson, Lynskey et al. [255]), and comorbid alcohol and drug dependence (Yang, Kranzler et al. [373]). Taken together, these cross-phenotype effects indicate that the *NCAM1* gene is likely to display strong pleiotropic effects, although more research is needed in order to disentangle true biological pleiotropy from mediated and spurious pleiotropic effects (Solovieff, Cotsapas et al. [300]).

Thirdly, three association signals reached genomewide signficance in the meta-analysis of age at onset of initiation of cannabis use, namely the Epithelial cell-transforming sequence 2 oncogene-like (*ECT2L*) on chromosome 6, the Calcium-transporting ATPase (*ATP2C2*) gene on chromosome 16, and the DNA repair protein *RAD51* homolog 2 (*RAD51B*) gene on chromosome 14 (Chapter IX). Both *ECT2L* and *RAD51B* are plausible predictors of age at onset, as supported by previous nicotine dependence association studies [106], and by in vitro and rat addiction models [57], respectively. Especially interesting is the association with *ATP2C2*, given its involvement in calcium homeostasis, which in turn is essential for regulating processes like synaptic plasticity, learning and formation of new memories. This result provides further support for the idea that synaptic plasticity, memory and learning contribute to the developement of addiction behaviors (Uhl, Liu et al. [326], Benowitz [30]). These results also point to interesting candidate genes for later stages of substance addiction (i.e., abuse/dependence), given that previous studies demonstrated that age at initiation may serve as a relevant proxy for the liability to heavy use.

Finally, the empirical analyses yielded important biological clues to several smoking behaviors (Chapter X). Using the largest meta-analytic sample to date in a GWAS of smoking, and powerful set-based tests, I reported twenty-one genes associated with quantity smoked, smoking initiation, smoking cessation and age at initiation. Fifteen genes are novel (i.e., have not yet been significantly associated with smoking behaviors according to the GWAS catalogue release $2015 - 23 - 10$) and were missed in the original SNP-based analysis [78]. For instance, my analysis hit genomic regions previously associated with coronary artery disease (for which smoking is a known risk factor [73]), and the *HLA* locus previously implicated in e.g., schizophrenia (with which smoking shares genetic vulnerabilities [85]). Moreover, the analysis implicated two novel loci in smoking cessation, one of which (the *SEMA6D* gene) was suggested as plausible candidate for smoking cessation and current smoking by previous (insufficiently powered) studies (see [324] and [346], respectively). I have also identified the *MIR1323-MIR512-1-MIR512-2* locus within the 19q13 region significantly associated with age at initiation. The 19q13 region has been proposed as a plausible candidate for further investigation in relation to smoking quantity as it harbours the *CYP2A6* gene, a hepatic enzyme involved in nicotine metabolism. Variants in the *CYP2A6* gene were previously associated with variation in the rate of nicotine metabolism and predicted quantity smoked (i.e., the quantity smoked per day increases with increasing nicotine metabolism rate; see Mwenifumbo and Tyndale [251], Benowitz [30], Tobacco and Genetics Consortium [75]). With this result I added age at smoking initiation to the list of smoking phenotypes to be further investigated in relation to this biologically plausible region. Interestingly, loci in the same 19q13 region were also reported as associated with chronic obstructive pulmonary disease (Cho, Castaldi et al. [68]) for which smoking is a known risk factor; however, as hypothesized by Cho et al., (Cho, Castaldi et al. [68]) this association likely represents a mediated (by smoking dependence) rather than a true biological pleiotropic effect.

The pathway analysis is the first reported in the literature to provide evidence for significant association between several biological pathways and smoking behaviors based on an unbiased/hypothesis free approach (Chapter X). The analysis of quantity smoked revealed strong enrichment signal coming from the neuronal system pathways, which harbor the nicotinic acetylcholine receptors. This finding is consistent with the hypothesis that mechanisms underlying smoking dependence involve the mesocorticolimbic dopamine system (Dani and De Biasi [84], Kelley [181], Dani [83], Benowitz [30]). In short, as Benowitz described (Benowitz [30]), the biological mechanisms underlying nicotine addiction involve nicotine binding to the nicotinic acetylcholine receptors which, in turn, release several neurotransmitters (dopamine, glutamate and γ-aminobutyric acid) in regions of the brain known to be involved in the perception of pleasure and reward (i.e., in the ventral tegmental area and the shell of the nucleus acumbens). Following repeated exposure, the $\alpha4\beta2$ nicotinic acetylcholine receptors adapt to nicotine and become unresponsive. It is hypothesized that the reactivation of these closed receptors following abstinence/cessation gives rise to symptoms of craving and withdrawal which, in turn, reinforce continuing smoking/relapse. Quantity smoked was also statistically associated with pathways regulating cell-cycle checkpoints and apoptosis, pathways regulating the immune system, metabolism, signal transduction, as well as with asthma pathways.

Furthermore I identified several pathways regulating the mitotic cell-cycle chain that are significantly enriched for mutations in the ever smoking analysis and in the quantity smoked analysis. While no pathway was shared by the two smoking behaviours, these pathways form chains of pathways regulating different stages of cell division and sharing biological functions. These pathways control appropriate DNA replication by degrading regulatory proteins throughout anaphase, throughout exit from mitosis and during the G1 phase (Castro, Bernis et al. [60], Manchado, Eguren et al. [230]), as well as axon growth and synaptic plasticity (Li and Zhang [200]). As alluded to earlier, this finding also emphasizes and provides further support for the idea that synaptic plasticity and learning have a strong bearing on the development of addiction behaviors. Because the cell-cycle pathways are also known to belong to 'a subway map of cancer pathways' (Hahn and Weinberg [150]) (given their role in cancer development), this result suggest that some of the same biological mechanisms underlie both smoking and cancer. The results of GWASs of smoking dependence (Spitz, Amos et al. [304]) and lung cancer (Hung, McKay et al. [169]) are consistent with this finding. Both GWASs identified the same *CHRNA5-A3* genomic region on chromosome 15, suggesting that cancer and smoking share genetic vulnerabilities – as first conjectured by Fisher in 1959 (Fisher [127]): "[. . .] *cigarette smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype*" (Fisher [127]). While the mediation study by VanderWeele et al. (VanderWeele, Asomaning et al. [339]) demonstrated that variants at the 15q25.1 locus have a direct effect on both smoking and lung cancer, it is of interest to determine whether their conclusion generalizes at the pathway level as suggested by my results.

These findings have important implications for reducing the disease burden associated with smoking. Smoking is a known risk factor for various disease traits such as lung cancer (see [275], Lee, Forey et al. [196]), leukemia (e.g., see Fircanis, Merriam et al. [125]), heart disease (e.g., see Huxley and Woodward [171]), chronic bronchitis and emphysema (see e.g., Forey, Thornton et al. [128]), and it is well recognized as the world's leading cause of preventable disease and death. Currently there are several pharmacological treatments, including bupropion and nortriptyline (designed to treat depression, see Cahill, Stevens et al. [56]), buspiron, diazepam or propranolol (designed to treat anxiety, see Hughes, Stead et al. [168]) and nicotine replacement therapy (Cahill, Stevens et al. [56]). However, the mechanisms underlying some of these treatments are still yet to be known (Cahill, Stevens et al. [56]). For instance, it is yet unknown why bupropion might work in some individuals (Chen, Bloom et al. [65]) while it is associated with side effects such as increased risk of seizures in others (see e.g., Cahill, Stevens et al. [56]). The empirical findings reported herein open a path to potential targets for therapeutics. Aside from the nicotinic acetylcholine receptors, known for their rewarding role in nicotine dependence, the cell-cycle regulators are possible targets in smoking cessation therapy as proposed for novel cancer therapies.

12.1.1 Conclusion

The findings reported herein emphasize and lend further support for the idea that synaptic plasticity and learning have a strong bearing on the development of addiction behaviours. These results are informative in decoding the biological bases of other addiction phenotypes and disease traits such as schizophrenia and cancers with which smoking shares risk loci and biological pathways.

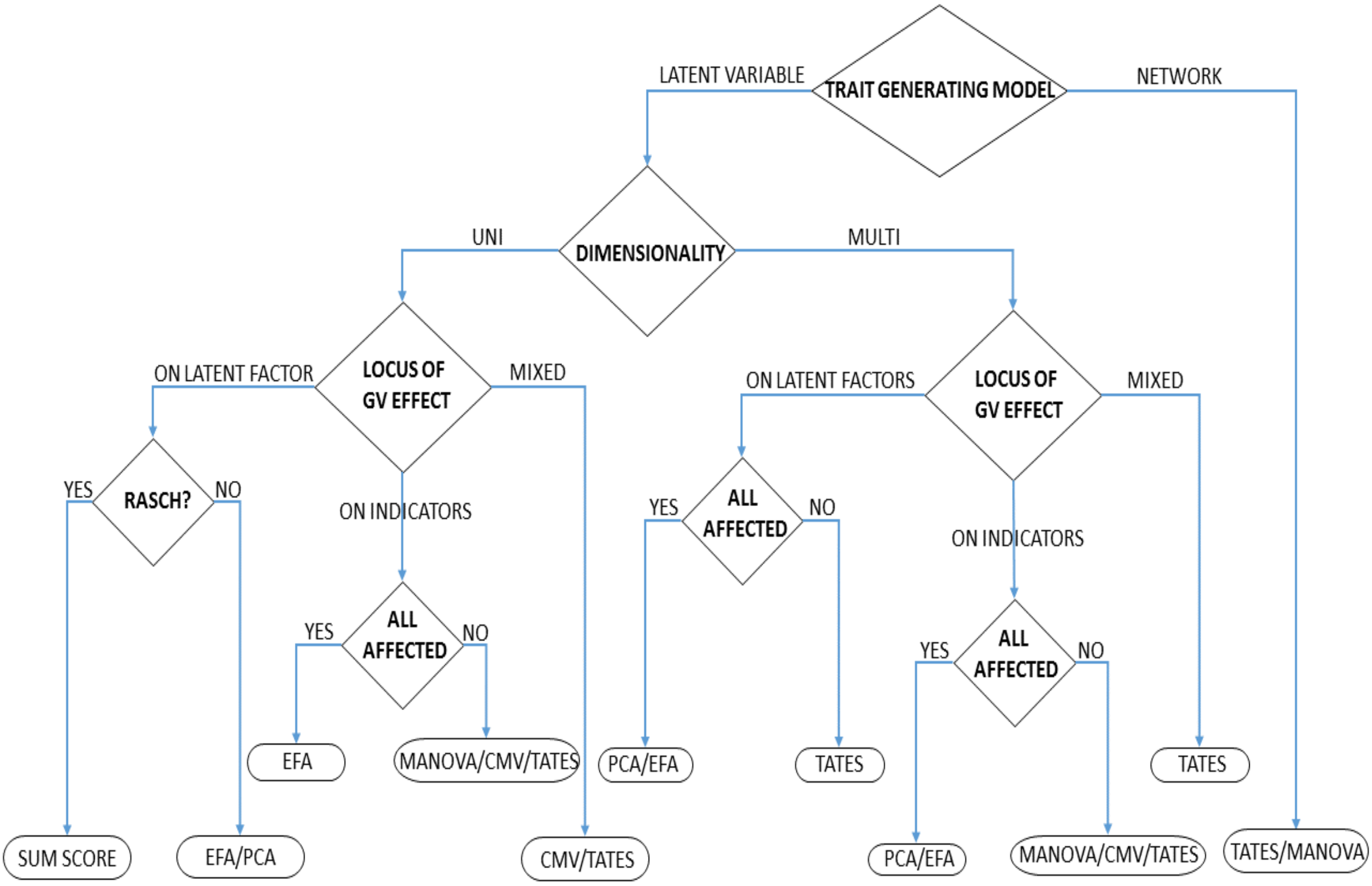
12.2 Means of Improving Statistical Power in GWAS

Underpowered genome-wide association studies are more likely to capture chance characteristics of the data, than true genetic effects. The past ten years of GWAS taught us that large samples are required to reliably identify individual SNPs associated with complex psychiatric traits. This is mainly due to the small SNP effects – each accounting individually for less than 0.1% of the phenotypic variance – and to the multiple testing burden. Yet, the success of GWAS also hinges upon the definition of the phenotype, the informativeness of markers (usually SNPs), and the approach to analyze the genotype-phenotype relation. I have considered each of these determinants of statistical power in some detail in the first part of this thesis. Below I tie together the recommendations stemming from the results of my power studies into an overall strategy for improving statistical power in GWAS interrogating the contribution to disease of common as well as rare variants.

12.2.1 On the definition of the phenotype

Complex traits are often multivariate in nature, that is, the phenotype comprises several correlated, but distinct components. For instance, consider the items relating to behavioral and physiological symptoms in the substance use disorder (DSM-IV), or the multiple correlated measures relating to forced expiratory volume, forced vital capacity, total lung capacity, functional residual capacity, residual volume and inspiratory capacity in the chronic obstructive pulmonary disease (COPD; see e.g., Dirksen [100]). Yet, to date, most association studies involved univariate phenotypes obtained by collapsing multivariate measures to create a sum score or an affection status dichotomy. For example, in the GWAS by Cho et al. (Cho, Castaldi et al. [68]) multiple COPD measures were collapsed into a dichotomous affection status, and in the GWAS of alcohol dependence by Edenberg at al (Edenberg, Koller et al. [110]), DSM-IV symptoms were used to create a case control dichotomy. However, reducing phenotypic dimensionality – by collapsing the multivariate measures into a sum score (which may in turn be dichotomized) – will increase the power only in certain situations. The flowchart in Figure 12.1 shows when such an approach is to be preferred over a multivariate one by considering several trait-generating models.

Figure 12.1: Provisional flowchart for selecting an analytic technique based on the hypothesized trait generating model. The GV effect on the observed indicators is assumed to be consistent (i.e., in the same direction). The flowchart covers many but not all possibilities (as the best test in the case of a network, may depend on the characteristics of the network). *Abbreviations: GV – genetic variant; EFA – exploratory factor analysis; CMV – combined multivariate approach* (Medland and Neale [242]); *TATES – Trait based Association Test that uses Extended Simes procedure* (Van der Sluis, Posthuma et al. [337]). Note that the MultiPhen procedure (O'Reilly, Hoggart et al. [262]) is closely related to MANOVA.



As depicted in Figure 12.1, the choice of the analytic technique depends on: (a) assumptions concerning the data generating model (e.g., conditional independence in latent variable models, or mutualism (Van Der Maas, Dolan et al. [332]) in the network model); (b) the dimensionality of the model; (c) the exact locus of the genetic effect and, related to this, (d) on whether the genetic variant impinges on all or some indicators, i.e., on how the variables are connected and where the GV exerts its effects. Figure 12.1 shows that the use of a sum score would be justified when the trait can be well described by a Rasch model. This draws heavily on the tenability of strong IRT psychometric assumptions such as unidimensionality, conditional independence and measurement invariance with respect to the genetic variant (i.e., the genetic effect is on the latent trait, see (Van Der Sluis, Verhage et al. [338]) for more details). This amounts to a highly idealized situation, as data on complex traits rarely fit the Rasch model perfectly.

Furthermore, whenever the model is multidimensional and the GV affects some of the latent factors (but not all; see Chapter II; see also (Van Der Sluis, Verhage et al. [338], Van der Sluis, Posthuma et al. [337]), or when the effect is specific to some of the observed indicators (i.e., not general, propagated via the latent trait in all indicators; see e.g., Medland and Neale [242]), collapsing the measurements on multiple phenotypes into a sum score typically leads to a loss in information and this in turn reduces power (see Chapter II; see also Medland and Neale [242], Van Der Sluis, Verhage et al. [338], Van der Sluis, Posthuma et al. [334], Van der Sluis, Posthuma et al. [337], Xu, Gaysina et al. [371]). Similarly, as discussed in Chapter II, transforming the phenotypes to factor scores or principal components, and resorting on univariate analyses would be justified only if the phenotypes are psychometric indicators which can be described well by a common pathway model or a single common genetic factor with relatively small genetic residuals. If the trait is multidimensional, this approach is likely to be powerful only if all indicators are affected by the GV in the same direction (Medland and Neale [242]) – either directly or via the common factors. In the latter scenario, the power of detection is expected to vary with the magnitude of the factor loadings, i.e., to be larger for higher factor loadings (see Medland and Neale [242]). Note that dichotomization (i.e., collapsing the sum score into a case-control dichotomy) has been omitted from the flowchart because resorting on this technique is almost never recommended given the associated reduction in power (see Van der Sluis, Posthuma et al. [334]). Dichotomization would be justified only if the variable is a true dichotomy, or if dichotomization increases measurement precision (see MacCallum, Zhang et al. [226]). In circumstances other than these, dichotomization (either by mean/median split or based on a clinical threshold) discards information about individual differences and so, is likely to result in misclassification of some individuals, which reduces the statistical power (MacCallum, Zhang et al. [226]).

Multivariate techniques such as MANOVA (MultiPhen), TATES or CMV are particularly powerful whenever the GV affects some but not all observed indicators (see Chapter II; see also Ferreira and Purcell [123], Medland and Neale [242], Van der Sluis, Posthuma et al. [337], Van der Sluis, Dolan et al. [335]), when the genetic effects are mixed (i.e., the effect is on the latent trait as well as specific to some of the observed indicators), when the GV displays contrasting effects (see Medland and Neale [242]) or when the data generating process can be well described by a network model (see (Van der Sluis, Posthuma et al. [337], Van der Sluis, Dolan et al. [335]) for more details). Multivariate analyses have the ability to capture and exploit the additional information on the correlations between the variables, or the ability to assess the separation among the genotype groups along a set of underlying dimensions (i.e., variates) by considering jointly the set of phenotypes (Stevens [306]). Furthermore, as demonstrated by Van der Sluis et al. (Van der Sluis, Posthuma et al. [337]) and by Medland and Neale(Medland and Neale [242]), the multivariate techniques perform well also in the scenarios in which data are missing completely at random, being particularly robust when the GV effect is on the latent trait.

Although the multivariate techniques have merit (i.e., for the power advantages they confer; see Chapter II and also see e.g., Ferreira and Purcell [123], Medland and Neale [242] and for the increase in parameter estimation accuracy they afford Shriner [297]), many researchers feel that simpler statistical models are quite adequate in the GWAS context for their computational easiness (Sham and Purcell [295]) as well as for interpretational reasons (Stephens [305]). However, there is solid evidence from the recent literature that the interest in addressing the computational (e.g., Zhou and Stephens [382]) as well as the interpretational issues (Stephens [305]) has intensified over the last years. Applying multivariate techniques genome-wide is now greatly facilitated by recently developed GWAS dedicated software (Van der Sluis, Posthuma et al. [337], Zhou and Stephens [382]). In addition to these, R–packages like gee (Carey, Lumley et al. [58]) and mmm (Asar and Ilk [22]); see also Table 1 in Shriner [297] implement multivariate models suitable for the analysis of traits that follow distributions other than Gaussian (i.e., binomial, Poisson, Gamma). Applying these methods genome-wide is feasible given the genotype data are typically chunked in manageable slices and hence the chunk-based analyses can be parallelized provided access to a cluster. This procedure can be accessed from Plink (Purcell, Neale et al. [279]) which comes with the advantage of being efficient in handling large datasets, thus speeding-up the analysis considerably. Following-up univariate gee-Plink analyses on each phenotype with the TATES procedure (Van der Sluis, Posthuma et al. [337]) is an option worth to consider especially for the analysis of family-based samples. Importantly, the advantages conferred by multivariate techniques began to be appreciated and extended from analyses focused on individual SNPs (see (Shriner [297]) for a recent review) to analyses focused on sets of SNPs, common (e.g., Van der Sluis, Dolan et al. [335]) or rare (Maity, Sullivan et al. [229]).

A point to bear in mind before embarking in multivariate analyses is that the multivariate techniques are particularly powerful if only a subset of the phenotypes (not all phenotypes) are affected by the genetic variant. As demonstrated by the power analyses carried-out in Chapter II and by others (e.g., Ferreira and Purcell [123], Stephens [305]) in this circumstance an increase in phenotypic correlations enhances the power. Hence, as suggested by Morrison ([248]) and Cole et al. (Cole, Maxwell et al. [71]) whenever one avails oneself of multivariate techniques it might be prudent to include variables correlated with the trait of interest yet not affected by the genetic variant to improve the power sharply: "*Thus, the counterintuitive possibility arises that greater power might result from the inclusion of weak variables (for which the effect size is zero) in the dependent variable system (as long as they are highly correlated with the outcome variables)*" (Cole, Maxwell et al. [71], page 466).

12.2.2 In time of test, family is best

As highlighted in Chapters III-V, over the past ten years of GWAS, family-based samples collected at the twin and family registries have contributed substantially to the discovery of genetic variants implicated in complex traits and diseases. Regarding the occasional practice of limiting the analyses to unrelated individuals, the power studies conducted in Chapters III-V demonstrate clearly and unambiguously that this practice is counterproductive, that is, discarding family members generally reduces the effective sample size and, correspondingly, it reduces the power. To get an indication of the power loss incurred in such a case, take the results displayed in Figure 3.1a (Chapter III): with a sample of 500 families comprising sibships size 4 and given a genetic variant with a MAF = 0.2 and explaining 1% of the variance, the power 'bounces around' 90% across the whole range of the phenotypic correlations, whilst limiting the sample to singletons reduces the power to as low as 37%. Arguments pertaining to computational tractability or to the effects of model misspecification that could justify this power loss ought to be reconsidered in the light of recent software developments. The fast algorithms developed recently (e.g., see Kang, Sul et al. [179] and Lippert, Listgarten et al. [209]) reduced dramatically the computational load associated to the family-based analysis. Actually, over the past five years there has been a plethora of papers concerned with developing efficient and fast algorithms tailored to handle clustered data – be this due to familial or to population stratification – and generally these were implemented in software programs that are made freely available. Several examples are listed in Table 12.1.

Table 12.1: Software freely available for family-based genome-wide association analysis

Software	Regression model	Model for the background correlation matrix	Specification of the Cluster variable (sandwich correction?)	URL
Plink [279] + gee [58] R package	Binomial Gaussian Gamma Inverse Gaussian Poisson Quasi Quasibinomial Quasipoisson	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\sigma}}_E^2\mathbf{I}$ (Independence) $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_C^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (Exchangeable) $\mathbf{V}(\widehat{\boldsymbol{\theta}}_f) = [\ \widehat{\boldsymbol{\sigma}}_A^2, \widehat{\boldsymbol{\sigma}}_C^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (Fixed, e.g., ACE background) $\mathbf{V}(\widehat{\boldsymbol{\theta}}_f)$ (Unstructured)	'id' (yes)	gee[1] Plink documentation[2]
Plink	Gaussian Binomial	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\sigma}}_E^2\mathbf{I}$ (Independence)	'family' (yes)	Plink[3]
Plink + survival [320] R package	Cox proportional hazards	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\sigma}}_E^2\mathbf{I}$ (Independence) $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_C^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (Exchangeable)	'cluster' (yes) 'frailty' (no)	survival[4] documentation[5]
Plink + nlme [269, 271] R package	Linear mixed	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\sigma}}_E^2\mathbf{I}$ (Independence) $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_C^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (Exchangeable) $\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_A^2, \widehat{\boldsymbol{\sigma}}_C^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (ACE background) $\mathbf{V}(\widehat{\boldsymbol{\theta}}_f) = [\ \widehat{\boldsymbol{\sigma}}_A^2, \widehat{\boldsymbol{\sigma}}_C^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (Fixed, e.g., ACE background) $\mathbf{V}(\widehat{\boldsymbol{\theta}}_f)$ (Unstructured)	'random' (yes, see documentation[6])	nlme[7] documentation[8]
GCTA [375]	Linear mixed	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_A^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (AE model)	Observed Genetic Relationship Matrix (no)	GCTA[9]
Merlin [3]	Linear mixed	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_A^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (AE model)	Expected Genetic Relationship Matrix (no)	Merlin[10] documentation[11]
FaST-LMM [209]	Linear mixed	$\mathbf{V}(\widehat{\boldsymbol{\theta}}) = [\ \widehat{\boldsymbol{\sigma}}_A^2, \widehat{\boldsymbol{\sigma}}_E^2\ ]$ (AE model)	Observed Genetic Relationship Matrix(no, but see [245])	FaST-LMM[12]

Although the list in Table 12.1 is not exhaustive, it shows that there are multiple modeling strategies readily available which can handle a variety of traits (e.g., continuous, binary, counts, time to event). Conveniently, practically any statistical model implemented in an R package can be accessed from Plink via Rserve (Urbanek [328]) and applied genome-wide. As mentioned above, the R-Rserve-Plink procedure is feasible given the genotype data are typically chunked in manageable slices and hence the chunk-based analyses can be parallelized provided access to a cluster.

It is important to realize that the power of the analytic strategy depends heavily on the choice of the model for the correlation matrix (conditional on the fixed regressors), i.e. the matrix which accommodates the dependency in the data due to family clustering. This choice should be directed by the theoretical and empirical knowledge of the covariance structure at hand (i.e., derived either based on genetically informative samples or based on the literature). For instance, given ACE traits (i.e., subject to Additive (A) genetic, common (C), and unshared (E) environmental effects) characterized by moderate to large familial resemblance arising both from shared genetic factors and common environment, maximum likelihood estimator with a correctly specified background model, i.e., a model that includes information regarding genetic relatedness and relatedness due to common environment, should be the strategy of choice (e.g., use Plink + nlme).

A complication arises when the samples consist of families highly variable in number and composition, as full detailed modeling of the background might be complex and subject to misspecification. How to arrive at correct standard errors given the background is (possibly) misspecified? In this situation, a sandwich correction can be applied to capture correctly the variance of the parameters of interest. It is important to note that there are many 'flavors' of sandwiches, i.e., a sandwich correction by itself can include any background model. Hence, once again, the choice of the model for the working correlation matrix becomes an important consideration. Simpler models might be preferable for computational reasons, but they are likely to exact price in terms of power which depends merely on the degree of misspecification. Returning to our ACE trait example: a quick and simple alternative would be to (incorrectly) assume an E model for the background and use the ULS sandwich (i.e., using Plink), but with this modeling choice the price in power increases sharply with increasing background correlations. However, using a maximum likelihood sandwich procedure with the background misspecified as a CE model (e.g., employing Plink + robust GEE with an exchangeable correlation matrix – or using nlme (see Table 12.1) instead[a] to fit a random intercept model – and getting the robust standard errors) will likely maintain the power close to that of the true model (see Table 3, Chapter IV).

Fitting a misspecified AE model for the background is an alternative (i.e., using a linear mixed model as implemented in e.g., FaST-LMM or GCTA), which has

[a] Note, these two methods are equivalent, conditional on the treatment of $\mathbf{V}(\widehat{\boldsymbol{\theta}})$

the added benefit that the block diagonal structure of the background correlation matrix can be relaxed to accommodate distant relatedness. However, although a sandwich is quick and simple to incorporate in the fast maximum likelihood procedure (see Chapter IV) currently none of these programs implement a sandwich to correct the standard errors for misspecifying of the background (i.e., by ignoring the shared environmental effects).

It should be highlighted that although the focus was on selecting from the pool of available methods, the most efficient ones to conduct family-based genome-wide association studies, the analytic strategies discussed in Chapters V-VI are regression based approaches, hence relevant to any analysis involving family data. That is, the predictor can be a genetic variant, a polygenic score or any other covariate one may be interested in.

12.2.3 Set-based analyses: expedient in a genome-wide scan

Improving the power of SNP-based tests by fully exploiting the phenotypic information and the sample at hand improves the power of downstream analyses – such as meta-analyses and set-based tests – that rely on the SNP-based p-values. Simulations and empirical data analyses (Liu, Mcrae et al. [215], Li, Gui et al. [201], Li, Kwan et al. [202], Listgarten, Lippert et al. [211]) including the applications reported in Chapters VII-IX demonstrate that following-up the SNP-based analyses with set-based tests generally boosts the power of detection and leads to additional insights into the biology of complex traits and diseases. The increase in power has two main sources: first, set-based tests consider jointly the weak effects of SNPs within the target region – be it the gene, the biological pathway (Liu, Mcrae et al. [215], Li, Gui et al. [201], Li, Kwan et al. [202], Listgarten, Lippert et al. [211]) (i.e., the set of genes having the same biological function) or the whole genome (Yang, Benyamin et al. [374], Yang, Lee et al. [375]); second, by targeting sets rather than SNPs the number of tests drops from millions to thousands and this mitigates considerably the multiple testing problem.

This boost in power afforded by set-based tests is nicely illustrated by the results reported in Chapter X. Whilst the collaborative analyses conducted by the TAG consortium (which combined three meta-analytic samples: the TAG, the ENGAGE and the Oxford-GlaxoKline samples, comprising 140,000 individuals) followed up loci that passed the 10^{-4} threshold located 14 SNPs significantly associated with smoking behaviors, the set-based tests (Li, Gui et al. [201], Li, Kwan et al. [202]) in the initial TAG sample (N = 74,053 individuals) afforded sufficient power to implicate 15 new genes and 40 biological pathways. This is just one example (but see also Chapters VII and VIII, and the applications on Chron's disease in (Li, Gui et al. [201], Li, Kwan et al. [202]) and Type 2 Diabetes in (Li, Gui et al. [201]) for additional examples) that illustrates the power advantages conferred by a set-based approach: all these hints concerning the biological mechanisms underlying smoking behaviors were missed in the SNP-based approach of GWAS.

Interestingly, although the tests focused on individual SNPs are still often underpowered with the current sample sizes, this standard SNP-based approach was proposed also for rare variant detection in sequencing studies (e.g., see Kinnamon, Hershberger et al. [189]). This observation is cause for concern, given that single variant tests are not only underpowered (Li and Leal [199], Madsen and Browning [227], Sham and Purcell [295]) but they are likely biased in their asymptotics (Bigdeli, Neale et al. [33]) with a small number of counts (whatever the sample size). Clearly, for the reasons emphasized above (and discussed extensively in the literature Li and Leal [199], Madsen and Browning [227], Price, Kryukov et al. [276], Wu, Lee et al. [367], Lee, Wu et al. [196], Chen, Meigs et al. [64], Ionita-Laza, Lee et al. [176], Listgarten, Lippert et al. [211], Lippert, Xiang et al. [210], Svishcheva, Belonogova et al. [312]), set-based tests are to be the preferred tool also for rare variant discovery. As there are several rare variant tests, their robustness to model misspecification could be the criterion of preferring one over the others. In this regard, in Chapter VI I have considered two of the most widely used test statistics in rare variant association studies – the score and the likelihood ratio tests – and argued in favor of the later, because of its greater robustness both to weight misspecification and to the inclusion in the target set of weighted neutral variation.

The availability of sequence data in increasingly large samples opens up the possibility to interrogate the contribution to disease of both common and rare variants. It is important to note that rare variant tests such as the sequence kernel association test (SKAT) allows for testing the combined effects of rare and common variants, whose contribution to the test statistic may be easily prioritized by assignment of weights. Although running separate tests for rare and common variants is the prevailing approach in the literature, results of my empirical analyses in Chapter VI question this practice. Considering common variants along with the rare ones in sequence-based kernel association tests appears to be justified for three main reasons. First, the use of variable weighting schemes is equivalent to applying variable frequency thresholds: the weights are removing from the test or favoring the contribution to the test statistic of the variants within the target set based on their frequency. Second, only the joint signal – coming from rare and more common variants – increased power to detect significant enrichment. And third, importantly, with the current samples, our tests are mostly powered to locate regions under relatively weak selection pressures, and such regions are expected to harbour rare as well as common variants both with functional effects. To locate pathways and genes under stronger selection pressures, larger samples (see Zuk, Schaffner et al. [383]) and the inclusion of more extreme weights (i.e., weights that overlook common variants and favour the rarer ones) will probably be required.

12.2.4 Conclusion

The past ten years of GWAS have taught us that we need large samples to reliably identify individual SNPs associated with complex psychiatric traits. This is mainly due to the small SNP effects – each accounting individually for less than 0.1% of the phenotypic variance – and to the multiple testing burden. Yet, as I demonstrate in this thesis, the success of GWAS hinges also upon the phenotype definition and the approach to analyze the genotype-phenotype relation. Opting for multivariate analyses rather than relying on dimension reduction techniques, exploiting at the fullest the rich resources collected at the twin registries, and complementing SNP-based analyses with set-based tests are key components of the strategy for improving statistical power in GWAS. This strategy is to be highly relevant to future genetic association studies facilitated by full exome and genome sequencing technologies.

Notes

[1] `https://cran.r-project.org/web/packages/gee/index.html`

[2] `http://cameliaminica.nl/scripts.php`

[3] `http://pngu.mgh.harvard.edu/ purcell/plink/`

[4] `https://cran.r-project.org/web/packages/survival/index.html`

[5] `http://cameliaminica.nl/research.php`

[6] `http://cameliaminica.nl/scripts.php`

[7] `https://cran.r-project.org/web/packages/nlme/index.html`

[8] `http://cameliaminica.nl/scripts.php`

[9] `http://cnsgenomics.com/software/gcta/mlmassoc.html`

[10] `http://csg.sph.umich.edu/abecasis/Merlin/`

[11] `https://genepi.qimr.edu.au/staff/sarahMe/merlin-offline.html`

[12] `http://research.microsoft.com:8082/en-us/um/redmond/projects/MSCompBio/Fastlmm/`

Chapter 13

Chapters' Origin

Chapter 2:

Genetic Association In Multivariate Phenotypic Data

Based on: **Camelia C. Minică**, Dorret I. Boomsma, Sophie van der Sluis, Conor V. Dolan: *Genetic Association In Multivariate Phenotypic Data: Power In Five Models*

Twin Research and Human Genetics 13(6):523-543, 2010,

DOI: 10.1375/twin.13.6.525.

Chapter 3:

The Use Of Imputed Sibling Genotypes In Sibship-Based Association Analysis: On Modeling Alternatives, Power And Model Misspecification

Based on: **Camelia C. Minică**, Conor V. Dolan, Jouke-Jan Hottenga, Gonneke Willemsen, Jacqueline M. Vink, Dorret I. Boomsma: *The Use Of Imputed Sibling Genotypes In Sibship-Based Association Analysis: On Modeling Alternatives, Power And Model Misspecification*

Behavior Genetics 43:254-266, 2013.

Chapter 4:

Sandwich Corrected Standard Errors In Family-Based GWAS

Based on: **Camelia C. Minică**, Conor V. Dolan, Maarten M.D. Kampert, Dorret I. Boomsma, Jacqueline M. Vink: *Sandwich Corrected Standard Errors In Family-Based Genomewide Association Studies*

European Journal of Human Genetics, 11 June 2014,

DOI: 10.1038/ejhg.2014.94.

Chapter 5:

MZ Twin Pairs or MZ Singletons in Population Family-Based GWAS? More Power in Pairs.

Based on: **Camelia C. Minică**, Dorret I. Boomsma, Jacqueline M. Vink, Conor V. Dolan: *MZ Twin Pairs or MZ Singletons in Population Family-Based GWAS? More Power in Pairs*

Molecular Psychiatry, 30 September 2014,

DOI: 10.1038/mp.2014.121

Chapter 6:

The Weighting Is The Hardest Part: On The Behavior Of The Likelihood Ratio Test And Score Test Under a Data-Driven weighting Scheme In Rare Variant Association Studies

Based on the manuscript (in preparation): **Camelia C. Minică**, Giulio Genovese, Christina M. Hultman, Rene Pool, Jacqueline M. Vink, **Conor V. Dolan**$^{\#}$**, Benjamin M. Neale**$^{\#}$: *The Weighting Is The Hardest Part: On The Behavior Of The Likelihood Ratio Test And Score Test Under a Data-Driven weighting Scheme In Rare Variant Association Studies.*

Chapter 7:

Heritability, SNP- And Gene-Based Analyses Of Cannabis Use Initiation And Age At Onset

Based on: **Camelia C. Minică**, Conor V. Dolan, Jouke-Jan Hottenga, Rene Pool, The Genome of the Netherlands Consortium, Iryna Fedko, Hamdi Mbarek, Charlotte Huppertz, Meike Bartels, Dorret I. Boomsma and Jacqueline M. Vink: *Heritability, SNP- and Gene-Based Analyses of Cannabis Use Initiation and Age at Onset*

Behav Genet. 2015 Sep;45(5):503-13. doi: 10.1007/s10519-015-9723-9. Epub 2015 May 19.

Chapter 8:

Genome-Wide Association Study of Cannabis Initiation Based on a Large Meta-Analytic Sample of 32,330 Subjects from The International Cannabis Consortium.

Based on the manuscript (submitted): *Genome-Wide Association Study of Cannabis Initiation Based on a Large Meta-Analytic Sample of 32,330 Subjects from The International Cannabis Consortium*

Sven Stringer*, **Camelia C. Minică***, **Karin J.H. Verweij***, Hamdi Mbarek, Jaime Derringer, Kristel R. van Eijk, Joshua D. Isen, Anu Loukola, Dominique F. Maciejewski, Evelin Mihailov, Peter J. van der Most, Cristina Sanchez-Mora, Richard Sherva, Raymond Walters, Jennifer J. Ware, Abdel Abdellaoui, Timothy B. Bigdeli, Susan J.T. Branje, Sandra A. Brown, Marcel Bruinenberg, Miguel Casas, Tonu Esko, Iris Garcia-Martinez, Scott D. Gordon, Catharina A. Hartman, Anjali K. Henders, Andrew C. Heath, Ian B. Hickie, Matthew Hickman, Christian J. Hopfer, Jouke-Jan Hottenga, Anja C. Huizink, Daniel E. Irons, Rene S. Kahn, Tellervo Korhonen, Henry R. Kranzler, Ken Krauter, Pol A.C. van Lier, Gitta H. Lubke, Pamela A.F. Madden, Reedik Magi, Matt K. McGue, Sarah E. Medland, Wim H.J. Meeus, Michael B. Miller, Grant W. Montgomery, Michel G. Nivard, Ilja M. Nolte, Albertine J. Oldehinkel, Beenish Qaiser, Josep A. Ramos-Quiroga, Vanesa Richarte, Richard J. Rose, Michael C. Stallings, Alex I. Stiby, Tamara L. Wall, Margaret J. Wright, Hans M. Koot, John K. Hewitt, Marta Ribases, Jaakko Kaprio, Marco P. Boks, Harold Snieder, Marcus R. Munafo, Andres Metspalu, Joel Gelernter, Dorret I. Boomsma, William G. Iacono , Nicholas G. Martin, **Nathan A. Gillespie$^{\#}$, Eske M. Derks$^{\#}$, & Jacqueline M. Vink$^{\#}$**

Chapter 9:

Survival Meta-Analysis of Age at Onset of Cannabis Use

Based on the manuscript (in preparation): *Survival Meta-Analysis of Age at Onset of Cannabis Use Based on The International Cannabis Consortium Sample*

Camelia C. Minică*, Karin J.H. Verweij*, Peter van der Moost*, Hamdi Mbarek,+others (to be added before printing the thesis), Dorret I. Boomsma, **Jacqueline M. Vink$^{\#}$, Nathan A. Gillespie$^{\#}$, Eske M. Derks$^{\#}$**

Chapter 10:

Pathways to Smoking: Biological Insights from the Tobacco and Genetics Consortium Meta-Analysis

Based on the article (under revision): **Camelia C. Minică**, Hamdi Mbarek, Conor V. Dolan, **Dorret I. Boomsma$^{\#}$, Jacqueline M. Vink$^{\#}$**

Pathways to smoking behaviors: biological insights from the TAG meta-analysis

Chapter 14

Samenvatting

Multivariate fenotypische data kan in Genoom-wijde associatie analyse (GWAS) het onderscheidend vermogen (oftewel power) om een genetisch variant (GV) te detecteren verhogen. Echter, multivariate data kunnen op verschillende manieren geanalyseerd worden. **Hoofdstuk II** betreft een vergelijking van verschillende strategien om multivariate fenotypische data te modelleren als men een GV wil detecteren. Wij simuleerden multivariate fenotypische data volgens de vijf modellen: (1) een n-factor model; (2) een model met meerdere correleerde genetische factoren; (3) een latente regressie model; (4) een hybride model bestaande uit een n-factor model voor de gedeelde omgevingsinvloeden (C), en autoregressieve modellen voor de additief genetische en niet-gedeelde omgevingsinvloeden (A and E), en 5) een stationair AE autoregressief model. In deze modellen introduceerden we het effect van de GV, als onderdeel van de additief genetisch invloeden, op verschillende manieren. Zodoende was het effect soms aanwezig in alle fenotypes, en soms beperkt tot een of een subset van de fenotypes. We vergeleken vervolgens de power van de volgende analyses om de GV te detecteren: (a) univariate regressie analyse, waarbij ieder fenotype apart op de GV werd geregresseerd (ANOVA); (b) univariate regressie analyse, waarbij de som van de fenotypes scores werd geregresseerd op de GV (ANOVA); (c) een exploratieve factor analyses (EFA), waarbij de factoren werden geregresseerd op de GV; en (d) multivariate regressie analyse waarbij de fenotypes tegelijkertijd werden geregresseerd op de GV (MANOVA). Power werd berekend aan de hand van de non-centraliteitsparameters (NCP) van de likelihood ratio test. Uit de resultaten bleek dat het gebruik van een somscores en de factor scores relatief hoge power hadden als de GV een effect had op alle fenotypes; MANOVA had in deze situatie relatief lage power. Voorts bleek dat MANOVA en EFA relatief hoge power hadden als de GV een effect had om sommige maar niet alle fenotypes, waarbij de power toenam met toenemende correlatie tussen de fenotypes. Ook bleken de NCPs van de MANOVA en EFA gelijk, hetgeen betekent dat verschillen in power volledig toe te schrijven zijn aan het verschil in vrijheidsgraden van de tests.

Imputatie van genotypes in families kan de power in GWAS verhogen omdat het de mogelijkheid creert om extra familieleden op te nemen in de analyses voor wie wel fenotypische, maar geen genetische, data beschikbaar zijn. De genotypes van subjecten kunnen gemputeerd worden op grond van de gemeten genotypes in haar of zijn familieleden. In **hoofdstuk III** is de invloed van verschillende factoren op de power om een GV te detecteren onderzocht in de context van dit type imputatie. Hierbij is gekeken naar families met 2 of 4 kinderen, waarbij de imputatie van de genotype data van een kind gebaseerd was op 1 broer (of zus), of 1 broer (of zus) en 1 ouder. Monte Carlo data simulaties zijn gebruikt om twee schattingsprocedures te vergelijken die verschillen in de behandeling van de inherente onzekerheid van gemputeerde data. De mixture aanpak houdt rekening met het gegeven dat de mogelijke gemputeerde waarden gekenmerkt worden door een waarschijnlijkheid (idealiter liggen deze dicht bij 1, zodat de correcte imputatie een hoge waarschijnlijkheid heeft). De dosage aanpak gebruikt een gewogen genotype-waarde die gebaseerd is op deze waarschijnlijkheden. In dit onderzoek is, geven fenotypes met verschillende erfelijkheidscofficinten (laag, middel, hoog), gekeken naar de power en kans op een type I fout (oftewel de conclusie dat een GV zonder effect toch effect heeft), en de rol van misspecificatie van de covariantie structuur tussen de families leden worden gemodelleerd,. Tenslotte zijn ter illustratie twee echte dataset geanalyseerd. Uit de resultaten bleek dat het verschil in power tussen de twee schattingsprocedures klein was. Op grond van de betere computationele efficintie verdient de dosage aanpak echter de voorkeur. De correcte specificatie van de covariantie structuur bleek wel belangrijk: met name bij hoog erfelijke fenotypen leidt misspecificatie tot een verlaagde kans op een type I fout. Tenslotte, bleek dat dit type imputatie onder specifieke omstandigheden kan resulteren in aanzienlijke toename in power.

Als de regressie van een fenotype op een voorspeller (bijv. een GV) wordt uitgevoerd in familie data dient rekening gehouden te worden met de covariantie structuur van de fenotypische scores van de familieleden. Het correct modelleren van deze structuur is belangrijk voor de power om de regressie relatie aan te tonen. Echter, dit is vaak gecompliceerd met name als de families binnen een studie verschillen in grootte en samenstelling. Model misspecificatie is dan moeilijk te vermijden. In **hoofdstuk IV** is gekeken naar het effect van model misspecificatie op de power om een regressie relatie (zoals een associatie met een GV) te detecteren. Hierbij is, in de context van een GWAS, gekeken naar de rol van misspecificatie in de ULS (unweighted least squares) en de ML (maximum likelihood) schattingsprocedures en de efficintie van zogenaamde sandwich correcties, die de toets van de regressierelatie corrigeren voor de misspecificatie. De rol van misspecificatie is onderzocht in families van twee of vier kinderen (monozygote (MZ) of dizygote (DZ) tweelingen, met of zonder 2 broers of zussen), met en zonder de ouders. De covariantie structuur was gebaseerd op additive genetische (A) en ongedeelde (E) omgevingsinvloeden (een AE model), of op een model met ook gedeelde (C) omgevingsinvloeden (een ACE model). Uit de resultaten bleek dat de sandwich correctie van de ULS en de ML resultaten leidde tot een correcte kans op een type I fout (d.w.z. de kans op vals positieve bevindingen was correct en gelijk voor beide methodes). Echter, de power van de gecorrigeerde ML toets bleek hoger dan die van de gecorrigeerde ULS toets. Het verschil in power bleek af te hangen van de correlatie tussen de familieleden: hoe hoger de correlatie gecreerd door gedeelde genen en/of gedeelde omgevingsinvloeden, hoe groter de winst in power die geboekt kan worden met ML. De power van de gecorrigeerde ML test lag niet veel lager dan de power in een correct gespecificeerd model. Voor regressie analyse uitgevoerd in familieleden raadden wij daarom de ML schattingsprocedure met sandwich correctie aan, als de hoofdvraag de regressie relatie betreft en niet de covariantie structuur van de familieleden.

Monozygote tweelingen maken een belangrijk deel uit van de populatie van participanten in tweeling registers bij wie data wordt verzameld. In GWAS, waarbij fenotypische scores worden geregresseerd op een GV, wordt vaak van de MZ tweelingenparen de data van n MZ tweeling weggelaten. Uit **hoofdstuk V** blijkt dat het simultaan analyseren van de data van beide tweelingen in een paar geen invloed heeft op de kans op een type I fout: de kans op vals positieve bevindingen blijft onveranderd. Voorts blijkt dat het behouden van beide tweelingen leidt tot hogere power, waarbij de winst in power afhangt van de fenotypische correlaties tussen de tweelingen (hoe lager de correlatie, hoe groter de winst in power als data van beide tweelingen geanalyseerd wordt). Het effectief modeleren van familie data (inclusie van alle MZ data) wordt besproken in het licht van de resultaten van **hoofdstuk IV**. De conclusie is dat de hogere power een goede reden is data van beide MZ tweelingen te behouden in GWAS.

Van zeldzame GVs (allele frequentie $< .01$) wordt aangenomen dat zij aanzienlijk kunnen bijdragen tot de genetische variantie van complexe fenotypes. Toetsen van zeldzame GVs zijn veelal gebaseerd om het gezamenlijk effect van meerdere zeldzame GVs. In zogenaamde Sequence Kernel Association Tests (SKAT) wordt het gewogen effect van ieder GV geacht een realisaties te zijn van een normaal verdeling met een gemiddelde van nul en een gegeven (te schatten) variantie. De gewichten zijn een functie van de (minor) allel frequenties van de individuele GVs, waarbij een GV met een lagere allel frequentie verondersteld wordt een groter effect te hebben op een fenotype. Echter, de ware waardes van de gewichten zijn onbekend. In **hoofdstuk VI** is gekeken naar de rol van misspecificatie van deze gewichten op de power en op de kans op een type I fout. Hierbij is gekeken naar zowel de score test en de likelihood ratio test van de associatie test. De likelihood ratio test blijkt robuuster dan de score test voor misspecificatie van de gewichten. Voorts is onderzocht of het gebruik van meerdere gewichten leidt tot een efficinte toets die minder afhankelijk is van de keuze van de gewichten.

In **hoofdstuk VII** is de erfelijkheid van initiatie van cannabis gebruik en rook gedragingen onderzocht aan de hand van recente methoden. Voorts zijn genomewijde analyses uitgevoerd om genen te identificeren die bijdragen tot individuele verschillen in cannabis initiatie en leeftijd van initiatie. Hierbij zijn SNPs gemputeerd op grond van het Genome of the Netherlands referentie paneel. Uit de resultaten bleek dat de gemeten en gemputeerde SNPs gezamenlijk een significant deel (25%; P = 0.0016) van de fenotypische variantie verklaarden. Cannabis gebruik blijkt een polygenetisch fenotype, waarvan de genetische variantie toe te schrijven is aan een groot aantal GVs, die verspreid liggen over het genoom.

In **hoofdstukken VIII** en **XI** worden de resultaten van genoom-wijde analyses van cannabis initiatie en leeftijd van initiatie gepresenteerd. Deze analyses zijn gebaseerd op de resultaten van meerdere GWAS analyses uitgevoerd in Europa, de US, en Australia onder leiding van het Internationale Cannabis Consortium. De studies in deze hoofstukken zijn de eerste die de associatie aantonen tussen GVs en zowel cannabis gebruik als leeftijd (NCAM1, CADM2, SCOC, SCOC, SCOC-AS1, and KCNT2) van initiatie van cannabis gebruik (ATP2C2, ECT2L, and RAD51B).

Het Tobacco and Genetics (TAG) Consortium heeft de relatie onderzocht tussen 2.5 miljoen SNPs en roken. Hierbij zijn 1052 SNPs gevonden die geassocieerd zijn met roken (bij een alfa van 10E-4). In **hoofdstuk X** zijn deze resultaten 2.5m tests gebruikt om set-based associatie toetsen uit te voeren. Hierbij worden de effecten van individuele SNP die een gen vormen samengevoegd in gene-based tests, en worden de effecten van individuele genen samengevoegd tot pathway-based tests (oftewel een test per groep van genen in plaats van per gen). Het aantal uit te voeren tests is dan aanzienlijk kleiner, waardoor de correctie van de alfa voor het aantal uitgevoerde tests ook minder extreem is. De power om effecten te detecteren is derhalve groter dan in de SNP-based (2.5m) tests. Op grond van deze analyses zijn 21 gene-based associaties en 40 pathway-based associaties gedentificeerd die samenhangen met initiatie van roken, hoeveelheid (roken), leeftijd van initiatie, en het stoppen met roken. De paden, die geassocieerd zijn met afhankelijkheid, bevatten genen die betrekking hebben op neuronale plasticiteit, leren, cel-cyclus regulatie, metabolisme, en het immuun systeem. Voorts hebben sommige pathways betrekking op zowel roken als kanker (in overeenstemming met Fisher's vermoeden uit 1959). Dit is de eerste studie die op grond van exploratieve gen-based en pathway-based tests, de associatie tussen biologische pathways en aan roken gerelateerd gedrag heeft aangetoond.

List of Tables

2.1 Variance components in the 4 scenarios that were used to generate the data. The total variance of each phenotype, conditional on the GV, equaled one. We provide only 4 parameter values in each scenario, as we did not vary these parameter values over the phenotypes. For instance in scenario S3, conditional on the GV, 4 (or 8) tests loaded on the common A factor with loadings equal to $\sqrt{.5}$, the genetic residual is .1. The loadings on the common shared environmental factor equaled $\sqrt{.2}$. The unshared environmental residuals equaled .2. Therefore in scenario S3, the decomposition of phenotypic variance conditional on GV is $h^2 = .6$, $c^2 = .2$, and $e^2 = .2$. 22

2.2 The power, non-centrality parameter, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 1. In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/4$ (.0025; 4 phenotypes) or $\alpha = .01/8$ (.00125; 8 phenotypes). The power for the corrected alpha is displayed in italics. 24

2.3 Variance components, conditional on GV, used to simulate data in study 2 (correlated genetic factors) and study 3 (regression of genetic factor 2 on genetic factor 1). The within (between) set phenotypic correlation is among phenotypes that load on the same (different) genetic factor (factors). For instance, in the 3 indicator S22 scenario, the phenotypes y_1 to y_3 (y_4 to y_6) loaded $\sqrt{.3}$ on the common A1 (A2) factor, each phenotype loading $\sqrt{.1}$ on the common C factor. The residual variance of each phenotype equaled .6 (.5 due to specific environment; .1 due to specific genes). So in scenario S22, the decomposition of phenotypic variance conditional on GV is $h^2 = .4$, $c^2 = .1$, and $e^2 = .5$. * Within set correlation is among phenotypes that load on the same genetic factor (y_1-y_2), the between set correlation is among phenotypes that load on the different genetic factors (y_1-y_6). 27

2.4 Power, non-centrality parameter, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 2. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which the 2 genetic factors are regressed on the GV (2 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/6$ (.0016; 6 phenotypes) and $\alpha = .01/10$ (.001; 10 phenotypes). The power for the corrected alpha is displayed in italics. 29

2.5 *Continued from Table 2.4.* 30

2.6 Power, non-centrality parameter, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 3. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which the 2 genetic factors are regressed on the GV (2 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/6$ (.0016; 6 phenotypes) and $\alpha = .01/10$ (.001; 10 phenotypes). The power for the corrected alpha is displayed in italics. 33

2.7 *Continued from Table 2.6.* 34

2.8 Variance components in study 4 at the 4 occasions, and the implied phenotypic covariance matrix, conditional on GV. The model comprises a simplex for A and E, and a common C factor. The factor loadings on the common C factor are $\sqrt{.3}, \sqrt{.2}, \sqrt{.1}$, and 0. The C factor loading decreases, the additive genetic variance increases, and the unshared environmental variance remains constant. Consequently the total phenotypic variance, conditional on the GV, remains 1 at each occasion. The autoregressive parameters β_A and β_E equal 1 and .7, respectively. The residual variances equal $\sigma^2_{\zeta_A} = .1$ and $\sigma^2_{\zeta_E} = .204$. 37

2.9 Power, non-centrality parameters, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 4. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which all 4 genetic factors are regressed on the GV (4 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/4$ (.0025; 4 phenotypes). The power for the corrected alpha is displayed in italics. 38

2.10 Power, non-centrality parameters, and degrees of freedom (in parentheses) of univariate and multivariate tests of association given $\alpha = .01$ in study 5. The power in the true model is included for the likelihood ratio test of the correctly specified GV (1 DF test) and for the omnibus test, in which all 4 genetic factors are regressed on the GV (4 DF test). In the case of the single phenotype ANOVA, power is reported for $\alpha = .01$ and $\alpha = .01/4$ (.0025; 4 phenotypes). The power for the corrected alpha is displayed in italics. 40

3.1 Posterior probabilities of the sibling 2 (s2) genotype AA, Aa, or aa, conditional on the observed genotype in a single sib (s1) or in a single sib and a single parent (p1), and given MAF = .2. The Hardy Weinberg (H-W) probabilities are the unconditional probabilities. The GPI is Kinghorn's genetic probability index, a distance measure (ranging from 0 to 100) of the imputed probabilities from the H-W probabilities. 52

3.2 Power in the full information model given an effect size of 1%, $\alpha = .01$ and $N = 500$ families. Power is shown as a function of the sibship size (nsib) and background correlation (ρ_{bg}). In the case of a singleton (nsib = 1), the background correlation is not relevant. 58

3.3 Average estimates of the genetic effect b_1 and the associated standard deviations (in parenthesis) for the Mixture models, for MAF = .2. The true parameter value is $b_1 = .1767$ (1000 replicates) 60

3.4 Average estimates of the genetic effect b_1 and the associated standard deviations (in parenthesis) for the Dosage models, for MAF = .2. The true parameter value is $b_1 = .1767$ (1000 replicates) 61

3.5 Type I error rates in the Full information and in the Dosage settings, in the correctly specified model (AE background) and in the misspecified model (CE background, results displayed in italics). We simulated sibling phenotypes for 500 monozygotic and 500 dizygotic families and a SNP having a MAF = .5 and explaining 1% of the phenotypic variance. We varied the sibship size and the magnitude of the MZ background correlations (10,000 simulations/cell for the cells $\alpha = 10^{-2}$ and $\alpha = 10^{-3}$; 100,000 replicates for the $\alpha = 10^{-4}$ cell). 62

3.6 Increase in χ^2 obtained in a family-based association analysis that includes 2410 observed and 3500 imputed sibling genotypes, relative to an association analysis limited to the observed genotypes. The first 4 SNPs are hits at $\alpha = .01$, the SNP rs1351394 is a Bonferroni significant result. 65

3.7 Results of 20 tests of genetic association with smoking initiation, ran in the 'complete data' sample (N = 2458) and in the sample that includes additionally imputed siblings genotypes (N = 5981). Sibling data only were included into analysis. The background covariance matrix was modeled by an AE model. The model was fitted by means of quasi-likelihood and provided Wald-type tests of effects, which, for consistency, were converted to chi-square values. 67

4.1 Power (alpha = 10^{-7}) and parameter estimates for the ML linear mixed (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures. We simulated a genetic marker having an effect of 1% explained phenotypic variance and a MAF = .5. The sample consisted of N = 4000 individuals. The trait was simulated according to an AE background model (the true model) given various heritabilities (h^2) (10,000 simulated samples for each cell). The background model in the ML procedure is correctly specified (true or saturated, i.e., unstructured). 79

4.2 Type I error rates for the ML linear mixed (standard and sandwich-corrected) and the ULS (standard and sandwich corrected) procedures. The background model is (a) correctly specified (true) or (b) misspecified. Background covariance matrix was generated according to an ACE model ($h^2 = .2,\ c^2 = .6$). The samples comprised of 4000 individuals (1,000,000 simulated data sets/cell). 80

4.3 Power (given alpha $= 10^{-7}$) and parameter estimates for the ML (standard and sandwich corrected) and the ULS (standard and sandwich corrected) procedures. The background model is (a) correctly specified (true) or (b) misspecified. Background covariance matrix was generated according to an ACE model ($h^2 = .2,\ c^2 = .6$). The genetic marker explained 1% phenotypic variance and had a MAF = .5. The samples consisted of N = 4000 individuals (10,000 simulated data sets per cell). 82

5.1 Empirical type I error rates in a test of genetic association with continuous phenotypes (linear model) and with binary phenotypes (logit model) using only MZ individuals ($N = 1000$ MZ twin pairs). Within the square brackets we report the 99% confidence intervals (CI). 93

6.1 Type I error for the restricted likelihood ratio test (LRT) and the score test, given genotypic data simulated under the null model of no association between the target region and the phenotype. The sample consisted of 10,000 individuals with genotypes at 50 variants having minor allele frequencies (MAFs) sampled from the uniform distribution and ranging from .5% to 5%. The restricted LRT and the score tests were computed for three sets of weights beta in each of the 1,000,000 simulated samples. Type I error equals the proportion of datasets in which the null hypothesis has been incorrectly rejected given the three significance thresholds. For the LRT we report the results given varying mixing proportions of χ^2_0 and χ^2_1 distributions. 106

6.2 *Continued from Table 6.1* 107

6.3 Type I error for the restricted likelihood ratio test (LRT) and the score test, given genotypic data simulated under the null model of no association between the target region and the phenotype. The sample consisted of 10,000 individuals with genotypes at 50 variants having equal beta weights and minor allele frequencies (MAFs) sampled from the uniform distribution and ranging from .5% to 5%. The LRT and the score tests were computed for three sets of weights beta in each of the 1,000,000 simulated samples. The type I error equals the percentage of datasets for which the null hypothesis has been incorrectly rejected, given the three significance thresholds. For the LRT we report the results given varying mixing proportions of χ^2_0 and χ^2_1 distributions. 108

6.4 *Continued from Table 6.3* 109

6.5 Results of the gene-set enrichment analysis run in the Swedish sample (N = 4940; prevalence in the sample = 0.49). The 2 gene-sets included variants with MAF below 5% (A) or below 1% (B). *Bonferroni corrected P-values are given in italics.* 113

6.6 Results of a test of association between a gene harboring 10 active variants (with a minor allele frequency ranging between 5% and .05% and explaining 1% of the phenotypic variance) and a continuous phenotype, in a simulated sample of 10,000 individuals. Data were simulated in R using the MASS package. Analyses were performed in 3 independent programs: the R-nlme package, the software Genome-wide Complex Trait Analysis (GCTA), and the software FaST-LMM-set. We report the log restricted likelihood under the null model (LL_0), the log restricted likelihood under the alternative model (LL_1), the chi-square test with 1 degree of freedom ($\chi^2(1)$), the variance attributable to the 10 genetic variants (V(G)). 116

7.1 Top five genes showing the strongest association with initiation of cannabis use. 131

7.2 *Continued from Table 7.1* 132

7.3 *Continued from Table 7.2* 133

7.4 Top five genes showing the strongest association with age at onset of cannabis use. 134

7.5 *Continued from Table 7.4* 135

8.1 Discovery and replication sample characteristics. Abbreviations: sample size (N), percentage of users that ever used cannabis (% users), percentage of females (% female), and number of SNPs used for the meta-analysis (N SNPs). 143

8.2 Top 10 SNPs with meta-analysis results of discovery samples, and results of combined discovery and replication samples. SNPs are displayed when not in linkage disequilibrium ($R^2 < 0.1$; and for SNPs with $R^2 \geq 0.1$ only the most significant SNP is shown in the top 10). * Direction per sample: allele A1 increases (+) or decreases (-) liability for cannabis use, or sample did not contribute to this SNP (?). Order of samples: ALSPAC, BLTS, CADD, EGCUT1, EGCUT2, FinnTwin, HUVH, MCTFR, NTR, QIMR, TRAILS, Utrecht, Yale Penn EA. Sample information can be found in Table 8.1. Abbreviations: Chromosome (Chr), location in base pairs in human genome version 19 (BP (hg19)), allele 1 (A1), allele 2 (A2), Frequency of allele 1 (Freq A1), standard error (s.e.). $ The combined sample contains the discovery, the Radar replication sample and the African Americans replication sample. 147

8.3 Top 5 genes from the gene-based tests of association with corrected *p*-values (Benjamini & Hochberg) based on the meta-analytic discovery and replication samples. Abbreviations: Human genome version 19 (hg19), base pair length (BP length), and number of SNPs used for the meta-analysis (N SNPs). 150

9.1 Descriptive information on the nine participating cohorts 159

9.2 SNPs showing associations above the suggestive line in the Manhattan plot (p-values < 1E-05 in the meta-analysis). *Abbreviations: RSID - rs number; Chr - chromosome; BP - base pair position; A1 - allele 1; A2 - allele 2; Freq1 - frequency of allele 1; MinFreq - minimum allele frequency; MaxFreq - maximum allele frequency; SE-standard error.* 164

9.3 *Continued from Table 9.2* 165

9.4 Association results and descriptive information for the top SNP rs1574587 in the participating cohorts. *Abbreviations: Chr - chromosome; BP - base pair position; A_1 - allele 1; A_2 - allele 2; SE - standard error; N - sample size; EAF -effect allele frequency; Info - imputation quality; λ_{GC}SE - lambda corrected standard error; λ_{GC}P- lambda corrected P-value;* 166

9.5 Genes significantly associated with age at onset of cannabis use. Reported below are the nominal p-values and the corrected p-values based on the Benjamini and Hochberg method [29]. 169

10.1 Genes implicated in quantity smoked . Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23; see `www.genome.gov/gwastudies` [359]) . . 177

10.2 Genes implicated in ever smoking . Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23 [359]) 178

10.3 Genes implicated in smoking cessation. Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23 [359]) 178

10.4 Genes implicated in age at initiation . Highlighted are genes that were not previously associated with smoking behaviours according to the GWAS catalogue (as of 2015–10–23 [359]) 179

10.5 Biological pathways implicated in quantity smoked. The Neuronal system chain. 179

10.6 Biological pathways implicated in quantity smoked. The cell-cycle chain. 180

10.7 Biological pathways implicated in quantity smoked. The immune system chain. 181

10.8 Biological pathways implicated in quantity smoked. Metabolism. 181

10.9 Biological pathways implicated in quantity smoked. Disease. 181

10.10 Biological pathways implicated in quantity smoked. Signal transduction. 181

10.11 Biological pathways implicated in quantity smoked. Gene expression. . 182

10.12 Biological pathways implicated in ever smoking. Regulation of the mitotic cell-cycle chain. 182

12.1 Software freely available for family-based genome-wide association analysis 202

List of Figures

1.1 The sampling distributions of a test statistic under the null (H_0) and under the alternative hypothesis (H_0 false). For this illustration, we assumed the sampling distribution of the test statistic under H_0 is standard normal, and we set the critical value at $T_{critical}$ corresponding to a type I error of α (grey). β (red) represents the probability of type II error, $1-\beta$ represents the statistical power, and $1-\alpha$ represents the probability of correctly rejecting the null hypothesis when the tested variant has no effect on the trait. 6

2.1 Path diagram for the common factor model with 4 phenotypes. The triangles represent fixed regressors (i.e., the GV and the unit vector). The parameters t1 to t4 are intercepts, the parameter **b** is the effect of the GV on the common genetic factor. The GV enters the model via common genetic factor A and affects the indicators y_1 to y_4. 23

2.2 Path diagram of the oblique two common factor model (three indicator model). The triangle represents the GV as a fixed regressor. The unit vector, which is used to estimate intercepts is not included to avoid clutter (see Figure 2.1). The parameter **b** represents the effect of the GV. Note that the GV contributes to the variance of the first latent genetic factor A1 and affects its indicators (y_1-y_3), but does not affect the second common factor A2, or its indicators (y_4-y_6). The value of the correlation between A1 and A2 was varied. Parameters are not shown to avoid clutter. 26

2.3 Exploratory (oblique) two common factor model as used in studies 2 and 3. Two factor loadings are fixed to zero (as depicted) to achieve rotational determinacy. The common factors are denoted F1 and F2, r1 to r6 represent the residuals. The triangles represent fixed regressors. The regression on the unit vector serves to estimate the intercepts, the regression on the GV estimates the effect of the GV (i.e., the parameters **b**1 and **b**2). Other parameters are not shown to avoid clutter. 28

2.4 Path diagram for the latent genetic regression model (3 indicator model). The triangle represents the GV as fixed regressor. The unit vector used to estimate intercepts is not included to avoid clutter (see Figure 2.1). The parameter **b** represents the effect of the GV. Note that the GV contributes to the variance of the first latent genetic factor A1 and affects its indicators (y_1-y_3). The GV contributes to A2 via the regression coefficient $\mathbf{b}_{A2A1}$, and so also affects the indicators y_4-y_6. The value of the parameter $\mathbf{b}_{A2A1}$ was varied. Parameters are not shown to avoid clutter. 32

2.5 Path diagram for the hybrid simplex-factor model. The triangle represents the GV as fixed regressor. The unit vector used to estimate intercepts is not included to avoid clutter (see Figure 2.1). In this model the GV enters at occasion 1. We also considered the cases in which the GV enters at occasions 2, 3, or 4. 36

3.1 The expected power in the full information setting for various background correlations, given $\alpha = .01$, MAF = .2 and an effect size of 1%. Left: 500 families with 1, 2, 3 and 4 siblings. Right: 500 genotyped siblings regardless of sibship size (i.e., 500 singletons, 250 size 2 sibships, 166 size 3 sibships, and 125 size 4 sibships). 57

3.2 The empirical power of the Dosage model (top) and the Mixture model (bottom), relative to the expected power afforded by 500 singletons (the black bolded line), given $\alpha = .01$. The grey lines: the empirical power afforded by sibships sizes 2, 3 and 4 when imputation is based on 1 genotyped sibling. The black lines: the empirical power afforded by sibships sizes 2, 3 and 4 when imputation is based on 1 sibling and 1 parental genotypes. Power calculations are based on 1000 datasets comprising 500 families, each dataset with a simulated genetic variant explaining 1% of the phenotypic variance, regardless of MAF. 59

3.3 The empirical power to detect a genetic variant with a MAF = .5, that explains 1% of the trait variance in the correctly specified AE linear mixed model (the grey line) and in the misspecified CE linear mixed model (the black dashed line). In the correct model the background covariances among identical twins were specified as twice larger than in fraternal twins. In the incorrect model the background covariance matrix was estimated subject to equal covariances. The empirical power was computed for 10,000 datasets (100,000 datasets for the 10^{-7} cell) consisting of 500 MZ and 500 DZ families with sibships of size 2 and 4. 63

3.4 Chi-square values obtained in the analysis that incorporates 3500 imputed sibling genotypes along with the 2410 observed genotypes relative to the chi-square values obtained in the "no imputation analysis". In the latter analysis the sample is limited to the 2410 observed sibling genotypes. 112 SNPs were tested for association with height. Shown in black are the 9 hits at $\alpha = .01$ based on the observed data. Points below the diagonal are due to drop in test statistic following imputation. 64

3.5 Chi-square as obtained in three samples: sample 1, consisting of siblings with complete phenotype and genotype data (N = 4000), sample 2, consisting of siblings with observed (N = ∼1600) and imputed genotypes (N = ∼2400), and sample 3, where missing genotypes were not imputed (N = ∼1600 observed genotypes). Results are shown for three effect sizes (100 simulated samples). The grey dotted lines show analyses where the chi-square as obtained in the three samples is monotonically decreasing, as expected. The black lines show results inconsistent with this expectation. 68

4.1 Wald tests produced by the sandwich corrected ULS procedure compared to the test statistic obtained based on full information maximum likelihood (standard ML) estimation method. We simulated 1000 data sets consisting of 500 MZ and 500 DZ 4-sib families, we varied the size of the genetic effect (1%, .25% and the null model). The heritability of the trait was $h^2 = 70\%$. The dots above the diagonal show the number of times the standard ML procedure produced a larger test statistic. . . 84

5.1 The power to detect a genetic effect (1%) in 1000 MZ twin pairs as a function of the MZ twin correlation ($\alpha = .001$). The effective sample size, shown above the x-axis, varies from 2000 (MZ correlation = .0) to 1111 (MZ correlation = .8). The top horizontal line indicates the power afforded by 2000 MZ individuals when MZ correlation equals 0. The bottom horizontal line indicates the power afforded by 1000 singletons. 90

5.2 The null distribution of the Wald test statistic (1,000,000 replications), given a continuous trait (Subfigures a and b) and a binary trait, based on the test in 1000 MZ pairs. The binary trait was obtained by dichotomizing the continuous trait into a binary 0/1 phenotype (probability of 1 is .20 in Subfigures c and d; .05 in Subfigures e and f). 92

6.1 The power of the likelihood ratio test (LRT; A and C) and the score test (B and D) to detect a gene harboring 50 low-frequency variants: all functional (A and B) or a mixture of 30 functional and 20 neutral variants (C and D). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(.5,.5). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases. 110

6.2 The power of the likelihood ratio test (LRT; E and G) and the score test (F and H) to detect a gene harboring 50 low-frequency variants: all functional (E and F) or a mixture of 30 functional and neutral variants (G and H). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(1,1). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases. 111

6.3 The power of the likelihood ratio test (LRT; A and C) and the score test (B and D) to detect a gene harboring 50 low-frequency variants: all functional (A and B) or a mixture of 30 functional and 20 neutral variants (C and D). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(.5,.5). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases. *Abbreviation: BH – Benjamini and Hochberg correction.* 117

6.4 The power of the likelihood ratio test (LRT; E and G) and the score test (F and H) to detect a gene harboring 50 low-frequency variants: all functional (E and F) or a mixture of 30 functional and neutral variants (G and H). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(1,1). Power was evaluated in 1000 datasets consisting of 10,000 individuals. Note that while the variants-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual variant weights increase as the proportion of functional variants in the set decreases. *Abbreviation: BH – Benjamini and Hochberg correction.* 118

7.1 Percent of variance in initiation of cannabis use explained per chromosome relative to chromosome length. The chromosome number is shown in circles. 130

8.1 The Manhattan (A) and the QQ plot (B) based on results of the gene-based analysis performed in the discovery sample using HYST (Hybrid Set-based Test). 148

8.2 Forest plot for the top-SNP rs4471463 in the *NCAM1* gene on chromosome 11. 149

9.1 The quantile-quantile plot based on lambdaGC corrected (a) and on lambdaGC uncorrected input files (b) and the Manhattan (c) plot of the meta-analysis results. In the Manhattan plot, the y-axis shows the strength of association (-log10(P)) and the x-axis indicates the chromosomal position. The blue line indicates suggestive significance level (P < 1E-05) while the red line indicates genome-wide significance level (P < 5E-08). 163

9.2 Forest plot of the top SNP 166

9.3 The quantile-quantile (a) and the Manhattan (b) plots for the gene-based tests. 168

9.4 Regional plots around the significantly associated genes (a) the *ECT2L* gene (b) the *ATP2C2* gene (c) the *RAD51B* gene 170

12.1 Provisional flowchart for selecting an analytic technique based on the hypothesized trait generating model. The GV effect on the observed indicators is assumed to be consistent (i.e., in the same direction). The flowchart covers many but not all possibilities (as the best test in the case of a network, may depend on the characteristics of the network). *Abbreviations: GV – genetic variant; EFA – exploratory factor analysis; CMV – combined multivariate approach* (Medland and Neale [242]); *TATES – Trait based Association Test that uses Extended Simes procedure* (Van der Sluis, Posthuma et al. [337]). Note that the MultiPhen procedure (O'Reilly, Hoggart et al. [262]) is closely related to MANOVA. 198

Bibliography

[1] Abdel Abdellaoui, Jouke-Jan Hottenga, Peter de Knijff, Michel G Nivard, Xiangjun Xiao, Paul Scheet, Andrew Brooks, Erik A Ehli, Yueshan Hu, and Gareth E Davies. Population structure, migration, and diversifying selection in the netherlands. *European Journal of Human Genetics*, 21(11):1277–1285, 2013.

[2] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[3] Gonçalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlinrapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97–101, 2002.

[4] Gonalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlinrapid analysis of dense genetic maps using sparse gene flow trees. *Nature genetics*, 30(1):97–101, 2002.

[5] GR Abecasis, LR Cardon, and WOC Cookson. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66(1):279–292, 2000.

[6] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[7] A. Agrawal and M. T. Lynskey. Candidate genes for cannabis use disorders: findings, challenges and directions. *Addiction*, 104(4):518–532, 2009.

[8] A. Agrawal, M. T. Lynskey, A. Hinrichs, R. Grucza, S. F. Saccone, R. Krueger, R. Neuman, W. Howells, S. Fisher, L. Fox, R. Cloninger, D. M. Dick, K. F. Doheny, H. J. Edenberg, A. M. Goate, V. Hesselbrock, E. Johnson, J. Kramer, S. Kuperman, J. I. Nurnberger, E. Pugh, M. Schuckit, J. Tischfield, J. P. Rice, K. K. Bucholz, L. J. Bierut, and Geneva Consortium. A genome-wide association study of dsm-iv cannabis dependence. *Addiction Biology*, 16(3):514–518, 2011.

[9] A. Agrawal, M. C. Neale, K. C. Jacobson, C. A. Prescott, and K. S. Kendler. Illicit drug use and abuse/dependence: modeling of two-stage variables using the ccc approach. *Addict Behav*, 30(5):1043–1048, 2005.

[10] A. Agrawal, M. C. Neale, C. A. Prescott, and K. S. Kendler. A twin study of early cannabis use and subsequent use and abuse/dependence of other illicit drugs. *Psychological Medicine*, 34(7):1227–1237, 2004.

[11] A. Agrawal, M. L. Pergadia, S. F. Saccone, M. T. Lynskey, J. C. Wang, N. G. Martin, D. Statham, A. Henders, M. Campbell, R. Garcia, U. Broms, R. D. Todd, A. M. Goate, J. Rice, J. Kaprio, A. C. Heath, G. W. Montgomery, and P. A. F. Madden. An autosomal linkage scan for cannabis use disorders in the nicotine addiction genetics project. *Arch Gen Psychiatry*, 65(6):713–722, 2008.

[12] A. Agrawal, J. L. Silberg, M. T. Lynskey, H. H. Maes, and L. J. Eaves. Mechanisms underlying the lifetime co-occurrence of tobacco and cannabis use in adolescent and young adult twins. *Drug Alcohol Depend*, 108(1-2):49–55, 2010.

[13] Arpana Agrawal, Julia D Grant, Mary Waldron, Alexis E Duncan, Jeffrey F Scherrer, Michael T Lynskey, Pamela AF Madden, Kathleen K Bucholz, and Andrew C Heath. Risk for initiation of substance use as a function of age of onset of cigarette, alcohol and cannabis use: findings in a midwestern female twin cohort. *Preventive medicine*, 43(2):125–128, 2006.

[14] Arpana Agrawal, Katherine I Morley, Narelle K Hansell, Michele L Pergadia, Grant W Montgomery, Dixie J Statham, Richard D Todd, Pamela AF Madden, Andrew C Heath, and John Whitfield. Autosomal linkage analysis for cannabis use behaviors in australian adults. *Drug and Alcohol Dependence*, 98(3):185–190, 2008.

[15] Arpana Agrawal, Michele L Pergadia, Scott F Saccone, Michael T Lynskey, Jen C Wang, Nicholas G Martin, Dixie Statham, Anjali Henders, Megan Campbell, and Robertino Garcia. An autosomal linkage scan for cannabis use disorders in the nicotine addiction genetics project. *Archives of general psychiatry*, 65(6):713–721, 2008.

[16] Hana Lango Allen, Karol Estrada, Guillaume Lettre, Sonja I Berndt, Michael N Weedon, Fernando Rivadeneira, Cristen J Willer, Anne U Jackson, Sailaja Vedantam, and Soumya Raychaudhuri. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[17] David B Allison, Bonnie Thiel, Pamela St Jean, Robert C Elston, Ming C Infante, and Nicholas J Schork. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *The American Journal of Human Genetics*, 63(4):1190–1201, 1998.

[18] Christopher I Amos, Mariza de Andrade, and Dakai K Zhu. Comparison of multivariate tests for genetic linkage. *Human Heredity*, 51(3):133–144, 2001.

[19] Christopher I. Amos, Margaret R. Spitz, and Paul Cinciripini. Chipping away at the genetics of smoking behavior. *Nat Genet*, 42(5):366–368, 2010.

[20] Verneri Anttila, Bendik S Winsvold, Padhraig Gormley, Tobias Kurth, Francesco Bettella, George McMahon, Mikko Kallela, Rainer Malik, Boukje de Vries, and Gisela Terwindt. Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nature genetics*, 45(8):912–917, 2013.

[21] L. Arseneault, M. Cannon, R. Poulton, R. Murray, A. Caspi, and T. E. Moffitt. Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. *Bmj*, 325(7374):1212–3, 2002.

[22] zgr Asar and zlem lk. mmm: an r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine*, 112(3):649–654, 2013.

[23] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, and Janan T Eppig. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[24] Mary E Atz, Brandi Rollins, and Marquis P Vawter. Ncam1 association study of bipolar disorder and schizophrenia: polymorphisms and alternatively spliced isoforms lead to similarities and differences. *Psychiatric genetics*, 17(2):55, 2007.

[25] Geraldine A Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami LevyMoonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, and Joel Thibault. From fastq data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, doi: 10.1002/0471250953.bi1110s43, 2013.

[26] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[27] A. Basilevsky. *Applied matrix algebra in the statistical sciences.* Elsevier Science Publishing, New York, 1983.

[28] A Leo Beem and Dorret I Boomsma. Implementation of a combined association-linkage model for quantitative traits in linear mixed model procedures of statistical packages. *Twin Research and Human Genetics*, 9(03):325–333, 2006.

[29] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[30] Neal L Benowitz. Nicotine addiction. *The New England journal of medicine*, 362(24):2295, 2010.

[31] A. W. Bergen, D. V. Conti, D. Van Den Berg, W. Lee, J. Liu, D. Li, N. Guo, H. Mi, P. D. Thomas, C. N. Lessov-Schlaggar, R. Krasnow, Y. He, D. Nishita, R. Jiang, J. B. McClure, E. Tildesley, H. Hops, R. F. Tyndale, N. L. Benowitz, C. Lerman, and G. E. Swan. Dopamine genes and nicotine dependence in treatment-seeking and community smokers. *Neuropsychopharmacology*, 34(10):2252–64, 2009.

[32] L. C. Bidwell, J. E. McGeary, J. C. Gray, R. H. Palmer, V. S. Knopik, and J. MacKillop. Ncam1-ttc12-ankk1-drd2 variants and smoking motives as intermediate phenotypes for nicotine dependence. *Psychopharmacology (Berl)*, 232(7):1177–86, 2015.

[33] T Bernard Bigdeli, Benjamin M Neale, and Michael C Neale. Statistical properties of single-marker tests for rare variants. *Twin Research and Human Genetics*, 17(03):143–150, 2014.

[34] AJ Birley, JB Whitfield, MC Neale, DL Duffy, AC Heath, DI Boomsma, and NG Martin. Genetic time-series analysis identifies a major qtl for in vivo alcohol metabolism not predicted by in vitro studies of structural protein polymorphism at the adh1b or adh1c loci. *Behavior genetics*, 35(5):509–524, 2005.

[35] John Blangero, Vincent P Diego, Thomas D Dyer, Marcio Almeida, Juan Peralta, Jack W Kent Jr, Jeff T Williams, Laura Almasy, and Harald HH Gring. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. *Advances in genetics*, 81:1, 2013.

[36] Steven Boker, Michael Neale, Hermine Maes, Michael Wilde, Michael Spiegel, Timothy Brick, Jeffrey Spies, Ryne Estabrook, Sarah Kenny, Timothy Bates, et al. Openmx: an open source extended structural equation modeling framework. *Psychometrika*, 76(2):306–317, 2011.

[37] DI Boomsma and CV Dolan. Multivariate qtl analysis using structural equation modeling: A look at power under simple conditions. 2000.

[38] Dorret Boomsma, Andreas Busjahn, and Leena Peltonen. Classical twin studies and beyond. *Nature reviews genetics*, 3(11):872–882, 2002.

[39] Dorret I Boomsma. Using multivariate genetic modeling to detect pleiotropic quantitative trait loci. *Behavior Genetics*, 26(2):161–166, 1996.

[40] Dorret I Boomsma, Eco JC De Geus, Jacqueline M Vink, Janine H Stubbe, Marijn A Distel, Jouke-Jan Hottenga, Danielle Posthuma, Toos CEM Van Beijsterveldt, James J Hudziak, and Meike Bartels. Netherlands twin register: from twins to twin families. *Twin Research and Human Genetics*, 9(06):849–857, 2006.

[41] Dorret I Boomsma and Conor V Dolan. A comparison of power to detect a qtl in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. *Behavior Genetics*, 28(5):329–340, 1998.

[42] Dorret I Boomsma and Peter CM Molenaar. The genetic analysis of repeated measures. i. simplex models. *Behavior Genetics*, 17(2):111–123, 1987.

[43] Dorret I Boomsma, Jacqueline M Vink, Toos CEM Van Beijsterveldt, Eco JC de Geus, A Leo Beem, Elles JCM Mulder, Eske M Derks, Harriette Riese, Gonneke AHM Willemsen, and Meike Bartels. Netherlands twin register: a focus on longitudinal research. *Twin Research*, 5(05):401–406, 2002.

[44] Dorret I Boomsma, Cisca Wijmenga, Eline P Slagboom, Morris A Swertz, Lennart C Karssen, Abdel Abdellaoui, Kai Ye, Victor Guryev, Martijn Vermaat, and Freerk van Dijk. The genome of the netherlands: design, and project goals. *European Journal of Human Genetics*, 22(2):221–227, 2014.

[45] M. Scott Bowers. Activators of g-protein signaling 3: A drug addiction molecular gateway. *Behavioural pharmacology*, 21(5-6):500–513, 2010.

[46] B. K. Bracken, J. Rodolico, and K. P. Hill. Sex, age, and progression of drug use in adolescents admitted for substance use disorder treatment in the northeastern united states: comparison with a national survey. *Subst Abus*, 34(3):263–72, 2013.

[47] Ulla Broms, Karri Silventoinen, Pamela AF Madden, Andrew C Heath, and Jaakko Kaprio. Genetic architecture of smoking behavior: a study of finnish adult twins. *Twin Research and Human Genetics*, 9(01):64–72, 2006.

[48] Brian L Browning and Zhaoxia Yu. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, 85(6):847–861, 2009.

[49] Alan J Budney, Roger Roffman, Robert S Stephens, and Denise Walker. Marijuana dependence and its treatment. *Addiction science & clinical practice*, 4(1):4, 2007.

[50] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, J. R. B. Perry, N. Patterson, E. Robinson, M. J. Daly, A. L. Price, and B. M. Neale. An atlas of genetic correlations across human diseases and traits. 2015.

[51] B. K. Bulik-Sullivan, P. R. Loh, H. K. Finucane, S. Ripke, and J. Yang. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. 47(3):291–5, 2015.

[52] Joshua T Burdick, Wei-Min Chen, Gonçalo R Abecasis, and Vivian G Cheung. In silico method for inferring genotypes in pedigrees. *Nature genetics*, 38(9):1002–1004, 2006.

[53] Margit Burmeister, Melvin G McInnis, and Sebastian Zllner. Psychiatric genetics: progress amid controversy. *Nature Reviews Genetics*, 9(7):527–540, 2008.

[54] P. Butterworth, T. Slade, and L. Degenhardt. Factors associated with the timing and onset of cannabis use and cannabis use disorder: results from the 2007 australian national survey of mental health and well-being. *Drug Alcohol Rev*, 33(5):555–64, 2014.

[55] KoraMareen Bhler, Elena Gin, Victor EcheverryAlzate, Javier CallejaConde, Fernando Rodriguez de Fonseca, and Jose Antonio LpezMoreno. Common single nucleotide variants underlying drug addiction: more than a decade of research. *Addiction Biology*, 2015.

[56] Kate Cahill, Sarah Stevens, Rafael Perera, and Tim Lancaster. Pharmacological interventions for smoking cessation: an overview and network meta-analysis. *Cochrane Database Syst Rev*, 5(5), 2013.

[57] Jordi Cami and Mag Farr. Drug addiction. *New England Journal of Medicine*, 349(10):975–986, 2003.

[58] V. J. Carey, T. Lumley, and B. Ripley. gee: Generalized estimation equation solver, r package version 4.13-18, http://cran.r-project.org/package=gee., 2012.

[59] J. P. Casey, T. Magalhaes, J. M. Conroy, R. Regan, N. Shah, R. Anney, D. C. Shields, B. S. Abrahams, J. Almeida, E. Bacchelli, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, P. F. Bolton, T. Bourgeron, S. Brennan, P. Cali, C. Correia, C. Corsello, M. Coutanche, G. Dawson, M. de Jonge, R. Delorme, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, S. Foley, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Green, S. J. Guter, H. Hakonarson, R. Holt, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, J. A. Lamb, M. Leboyer, A. Le Couteur, B. L. Leventhal, C. Lord, S. C. Lund, E. Maestrini, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, J. Miller, F. Minopoli, G. K. Mirza, J. Munson, S. F. Nelson, G. Nygren, G. Oliveira, A. T. Pagnamenta, K. Papanikolaou, J. R. Parr, B. Parrini, A. Pickles, D. Pinto, J. Piven, D. J. Posey, A. Poustka, F. Poustka, J. Ragoussis, B. Roge, M. L. Rutter, A. F. Sequeira, L. Soorya, I. Sousa, N. Sykes, V. Stoppioni, R. Tancredi, M. Tauber, A. P. Thompson, S. Thomson, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, J. A. Vorstman, S. Wallace, K. Wang, T. H. Wassink, K. White, K. Wing, et al. A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum Genet*, 131(4):565–79, 2012.

[60] Anna Castro, Cyril Bernis, Suzanne Vigneron, Jean-Claude Labb, and Thierry Lorca. The anaphase-promoting complex: a key factor in the regulation of cell cycle. *Oncogene*, 24(3):314–325, 2005.

[61] M Chavance and S Escolano. Misspecification of the covariance structure in generalized linear mixed models. *Statistical methods in medical research*, page 0962280212462859, 2012.

[62] C. Y. Chen, C. L. Storr, and J. C. Anthony. Early-onset drug use and risk for drug dependence problems. *Addict Behav*, 34(3):319–22, 2009.

[63] Dongquan Chen, Yufeng Li, Lizhong Wang, and Kai Jiao. Sema6d expression and patient survival in breast invasive carcinoma. *International Journal of Breast Cancer*, 2015:10, 2015.

[64] Han Chen, James B Meigs, and Jose Dupuis. Sequence kernel association test for quantitative traits in family samples. *Genetic epidemiology*, 37(2):196–204, 2013.

[65] LiShiun Chen, A Joseph Bloom, Timothy B Baker, Stevens S Smith, Megan E Piper, Maribel Martinez, Nancy Saccone, Dorothy Hatsukami, Alison Goate, and Laura Bierut. Pharmacotherapy effects on smoking cessation vary with nicotine metabolism gene (cyp2a6). *Addiction*, 109(1):128–137, 2014.

[66] Wei-Min Chen and Gonçalo R Abecasis. Family-based association tests for genomewide association scans. *The American Journal of Human Genetics*, 81(5):913–926, 2007.

[67] Joyce TW Cheung, Robert E Mann, Anca Ialomiteanu, Gina Stoduto, Vincy Chan, Kari Ala-Leppilampi, and Jrgen Rehm. Anxiety and mood disorders and cannabis use. *The American journal of drug and alcohol abuse*, 36(2):118–122, 2010.

[68] Michael H Cho, Peter J Castaldi, Emily S Wan, Mateusz Siedlinski, Craig P Hersh, Dawn L Demeo, Blanca E Himes, Jody S Sylvia, Barbara J Klanderman, and John P Ziniti. A genome-wide association study of copd identifies a susceptibility locus on chromosome 19q13. *Human molecular genetics*, page ddr524, 2011.

[69] Veryan Codd, Christopher P Nelson, Eva Albrecht, Massimo Mangino, Joris Deelen, Jessica L Buxton, Jouke Jan Hottenga, Krista Fischer, Tnu Esko, and Ida Surakka. Identification of seven loci affecting mean telomere length and their association with disease. *Nature genetics*, 45(4):422–427, 2013.

[70] Jonathan C Cohen, Eric Boerwinkle, Thomas H Mosley Jr, and Helen H Hobbs. Sequence variations in pcsk9, low ldl, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272, 2006.

[71] David A Cole, Scott E Maxwell, Richard Arvey, and Eduardo Salas. How the power of manova can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin*, 115(3):465, 1994.

[72] Psychiatric GWAS Consortium Coordinating Committee. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *The American journal of psychiatry*, 166(5):540–556, 2009.

[73] Coronary Artery Disease Genetics Consortium. A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nature genetics*, 43(4):339–344, 2011.

[74] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2015.

[75] International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

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

[77] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[78] Tobacco Consortium and Genetics. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, 42(5):441–447, 2010.

[79] Ciprian M Crainiceanu and David Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):165–185, 2004.

[80] Rebecca D Crean, Natania A Crane, and Barbara J Mason. An evidence based review of acute and long-term effects of cannabis use on executive cognitive functions. *Journal of addiction medicine*, 5(1):1, 2011.

[81] Carlos Cruchaga, Celeste M Karch, Sheng Chih Jin, Bruno A Benitez, Yefei Cai, Rita Guerreiro, Oscar Harari, Joanne Norton, John Budde, and Sarah Bertelsen. Rare coding variants in the phospholipase d3 gene confer risk for alzheimer/'s disease. *Nature*, 505(7484):550–554, 2014.

[82] CH Dalman, J Broms, J Cullberg, and P Allebeck. Young cases of schizophrenia identified in a national inpatient register. *Social psychiatry and psychiatric epidemiology*, 37(11):527–531, 2002.

[83] JA Dani. Roles of dopamine signaling in nicotine addiction. *Mol Psychiatry*, 2003.

[84] John A Dani and Mariella De Biasi. Cellular mechanisms of nicotine addiction. *Pharmacology Biochemistry and Behavior*, 70(4):439–446, 2001.

[85] John A Dani and R Adron Harris. Nicotine addiction and comorbidity with alcohol abuse and mental illness. *Nature neuroscience*, 8(11):1465–1470, 2005.

[86] Jennifer C Darnell, Sarah J Van Driesche, Chaolin Zhang, Ka Ying Sharon Hung, Aldo Mele, Claire E Fraser, Elizabeth F Stone, Cynthia Chen, John J Fak, and Sung Wook Chi. Fmrp stalls ribosomal translocation on mrnas linked to synaptic function and autism. *Cell*, 146(2):247–261, 2011.

[87] Gail Davies, Albert Tenesa, Antony Payton, Jian Yang, Sarah E Harris, David Liewald, Xiayi Ke, Stephanie Le Hellard, Andrea Christoforou, and Michelle Luciano. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular psychiatry*, 16(10):996–1005, 2011.

[88] R. Davies. The distribution of a linear combination of chi-square random variables. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 29:323333, 1980.

[89] Z John Daye, Hongzhe Li, and Zhi Wei. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic acids research*, 40(8):e60–e60, 2012.

[90] D. De Alwis, A. Agrawal, A. M. Reiersen, J. N. Constantino, A. Henders, N. G. Martin, and M. T. Lynskey. Adhd symptoms, autistic traits, and substance use and misuse in adult australian twins. *J Stud Alcohol Drugs*, 75(2):211–21, 2014.

[91] Paul IW De Bakker, Benjamin M Neale, and Mark J Daly. 10 meta-analysis of genome-wide association studies. 2009.

[92] Eco JC De Geus and DI Boomsma. A genetic neuroscience approach to human cognition. *European Psychologist*, 6(4):241, 2001.

[93] Sylviane de Viron, Servaas A Morr, Herman Van Oyen, Angela Brand, and Sander Ouburg. Genetic similarities between tobacco use disorder and related comorbidities: an exploratory study. *BMC medical genetics*, 15(1):85, 2014.

[94] L. Degenhardt, W. T. Chiu, N. Sampson, R. C. Kessler, J. C. Anthony, M. Angermeyer, R. Bruffaerts, G. de Girolamo, O. Gureje, Y. Huang, A. Karam, S. Kostyuchenko, J. P. Lepine, M. E. Mora, Y. Neumark, J. H. Ormel, A. Pinto-Meza, J. Posada-Villa, D. J. Stein, T. Takeshima, and J. E. Wells. Toward a global view of alcohol, tobacco, cannabis, and cocaine use: findings from the who world mental health surveys. *PLoS Med*, 5(7):e141, 2008.

[95] L. Degenhardt, M. Lynskey, and W. Hall. Cohort trends in the age of initiation of drug use in australia. *Aust N Z J Public Health*, 24(4):421–6, 2000.

[96] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, and Matt Hanna. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[97] Marta Di Forti, Arianna Marconi, Elena Carra, Sara Fraietta, Antonella Trotta, Matteo Bonomo, Francesca Bianconi, Poonam Gardner-Sood, Jennifer OConnor, and Manuela Russo. Proportion of patients in south london with first-episode psychosis attributable to use of high potency cannabis: a case-control study. *The Lancet Psychiatry*, 2015.

[98] Marta Di Forti, Hannah Sallis, Fabio Allegri, Antonella Trotta, Laura Ferraro, Simona A. Stilo, Arianna Marconi, Caterina La Cascia, Tiago Reis Marques, Carmine Pariante, Paola Dazzan, Valeria Mondelli, Alessandra Paparelli, Anna Kolliakou, Diana Prata, Fiona Gaughran, Anthony S. David, Craig Morgan, Daniel Stahl, Mizanur Khondoker, James H. MacCabe, and Robin M. Murray. Daily use, especially of high-potency cannabis, drives the earlier onset of psychosis in cannabis users. *Schizophrenia Bulletin*, 40(6):1509–1517, 2014.

[99] Santiago Diaz-Moralli, Mriam Tarrado-Castellarnau, Anibal Miranda, and Marta Cascante. Targeting cell cycle regulation in cancer therapy. *Pharmacology & therapeutics*, 138(2):255–271, 2013.

[100] A Dirksen. Outcome measures in chronic obstructive pulmonary disease (copd). *Thorax*, 58(12):1007–1008, 2003.

[101] Marijn A. Distel, Jacqueline M. Vink, Meike Bartels, Catharina E. M. van Beijsterveldt, Michael C. Neale, and Dorret I. Boomsma. Age moderates non-genetic influences on the initiation of cannabis use: a twin-sibling study in dutch adolescents and young adults. *Addiction*, 106(9):1658–1666, 2011.

[102] A. Dobson. *An introduction to generalized linear models.* Texts in Statistical Science Series. Chapman & Hall/CRC, London, 2002.

[103] Conor Dolan and Stephanie van den Berg. *Statistical power*, book section Statistical power, pages 61–86. Taylor & Francis Group, New York, 2008.

[104] Conor V Dolan, Frans J Oort, Reinoud D Stoel, and Jelte M Wicherts. Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling*, 16(2):295–314, 2009.

[105] Norman R Draper and Harry Smith. *Applied regression analysis.* John Wiley & Sons, 2014.

[106] Tomas Drgon, Ivan Montoya, Catherine Johnson, Qing-Rong Liu, Donna Walther, Dean Hamer, and George R Uhl. Genome-wide association for nicotine dependence and smoking cessation success in nih research volunteers. *Molecular Medicine*, 15(1-2):21, 2009.

[107] F. Ducci, M. Kaakinen, A. Pouta, A. L. Hartikainen, J. Veijola, M. Isohanni, P. Charoen, L. Coin, C. Hoggart, J. Ekelund, L. Peltonen, N. Freimer, P. Elliott, G. Schumann, and M. R. Jarvelin. Ttc12-ankk1-drd2 and chrna5-chrna3-chrnb4 influence different pathways leading to smoking behavior from adolescence to mid-adulthood. *Biol Psychiatry*, 69(7):650–60, 2011.

[108] Lindon J Eaves. Inferring the causes of human variation. *Journal of the Royal Statistical Society. Series A (General)*, pages 324–355, 1977.

[109] Lindon J Eaves, J Long, and Andrew C Heath. A theory of developmental change in quantitative phenotypes applied to cognitive development. *Behavior genetics*, 16(1):143–162, 1986.

[110] Howard J Edenberg, Daniel L Koller, Xiaoling Xuei, Leah Wetherill, Jeanette N McClintick, Laura Almasy, Laura J Bierut, Kathleen K Bucholz, Alison Goate, and Fazil Aliev. Genomewide association study of alcohol dependence implicates a region on chromosome 11. *Alcoholism: Clinical and Experimental Research*, 34(5):840–852, 2010.

[111] C. L. Ehlers, D. A. Gilder, I. R. Gizer, and K. C. Wilhelmsen. Heritability and a genome-wide linkage analysis of a type ii/b cluster construct for cannabis dependence in an american indian community. *Addiction Biology*, 14(3):338–348, 2009.

[112] C. L. Ehlers, I. R. Gizer, C. Vieten, and K. C. Wilhelmsen. Linkage analyses of cannabis dependence, craving, and withdrawal in the san francisco family study. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics*, 153B(3):802–811, 2010.

[113] David M Evans and David L Duffy. A simulation study concerning the effect of varying the residual phenotypic correlation on the power of bivariate quantitative trait loci linkage analysis. *Behavior genetics*, 34(2):135–141, 2004.

[114] DM Evans. Factors affecting power and type one error in association. *Statistical genetics: gene mapping through linkage and association. Taylor & Francis, UK*, pages 487–533, 2008.

[115] Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, 2007.

[116] Douglas S Falconer, Trudy FC Mackay, and Richard Frankham. Introduction to quantitative genetics (4th edn). *Trends in Genetics*, 12(7):280, 1996.

[117] M. S. Farrell, T. Werge, P. Sklar, M. J. Owen, R. A. Ophoff, M. C. O'Donovan, A. Corvin, S. Cichon, and P. F. Sullivan. Evaluating historical candidate genes for schizophrenia. *Mol Psychiatry*, 20(5):555–62, 2015.

[118] R. F. Ferdinand, F. Sondeijker, J. van der Ende, J. P. Selten, A. Huizink, and F. C. Verhulst. Cannabis use predicts future psychotic symptoms, and vice versa. *Addiction*, 100(5):612–8, 2005.

[119] D. M. Fergusson and L. J. Horwood. Early onset cannabis use and psychosocial adjustment in young adults. *Addiction*, 92(3):279–296, 1997.

[120] D. M. Fergusson, M. T. Lynskey, and L. J. Horwood. Conduct problems and attention deficit behaviour in middle childhood and cannabis use by age 15. *Aust N Z J Psychiatry*, 27(4):673–82, 1993.

[121] D. M. Fergusson, M. T. Lynskey, and L. J. Horwood. The short-term consequences of early onset cannabis use. *Journal of Abnormal Child Psychology*, 24(4):499–512, 1996.

[122] David M Fergusson and Joseph M Boden. Cannabis use and later life outcomes. *Addiction*, 103(6):969–976, 2008.

[123] Manuel AR Ferreira and Shaun M Purcell. A multivariate test of association. *Bioinformatics*, 25(1):132–133, 2009.

[124] Manuel AR Ferreira, Peter M Visscher, Nicholas G Martin, and David L Duffy. A simple method to localise pleiotropic susceptibility loci using univariate linkage analyses of correlated traits. *European journal of human genetics*, 14(8):953–962, 2006.

[125] Sophia Fircanis, Priscilla Merriam, Naushaba Khan, and Jorge J. Castillo. The relation between cigarette smoking and risk of acute myeloid leukemia: An updated meta-analysis of epidemiological studies. *American Journal of Hematology*, 89(8):E125–E132, 2014.

[126] Ronald Fisher. Cigarettes, cancer, and statistics. *Centennial Review of Arts & Science*, pages 151–166, 1958.

[127] Sir Ronald Aylmer Fisher. *Smoking: the cancer controversy: some attempts to assess the evidence*. Oliver and Boyd Edinburgh, 1959.

[128] Barbara A Forey, Alison J Thornton, and Peter N Lee. Systematic review with meta-analysis of the epidemiological evidence relating smoking to copd, chronic bronchitis and emphysema. *BMC pulmonary medicine*, 11(1):36, 2011.

[129] Sanja Frani, Conor V Dolan, John Broxholme, Hao Hu, Tomasz Zemojtel, Garreth E Davies, Kelly A Nelson, Erik A Ehli, Ren Pool, and Jouke-Jan Hottenga. Mendelian and polygenic inheritance of intelligence: A common set of causal genes? using next-generation sequencing to examine the effects of 168 intellectual disability genes on normal-range intelligence. *Intelligence*, 49:10–22, 2015.

[130] Menachem Fromer, Andrew J Pocklington, David H Kavanagh, Hywel J Williams, Sarah Dwyer, Padhraig Gormley, Lyudmila Georgieva, Elliott Rees, Priit Palta, and Douglas M Ruderfer. De novo mutations in schizophrenia implicate synaptic networks. *Nature*, 506(7487):179–184, 2014.

[131] DW Fulker, SS Cherny, PC Sham, and JK Hewitt. Combined linkage and association sib-pair analysis for quantitative traits. *The American Journal of Human Genetics*, 64(1):259–267, 1999.

[132] H. Furberg, Y. Kim, J. Dackor, E. Boerwinkle, N. Franceschini, D. Ardissino, L. Bernardinelli, P. M. Mannucci, F. Mauri, P. A. Merlini, D. Absher, T. L. Assimes, S. P. Fortmann, C. Iribarren, J. W. Knowles, T. Quertermous, L. Ferrucci, T. Tanaka, J. C. Bis, C. D. Furberg, T. Haritunians, B. McKnight, B. M. Psaty, K. D. Taylor, E. L. Thacker, P. Almgren, L. Groop, C. Ladenvall, M. Boehnke, A. U. Jackson, K. L. Mohlke, H. M. Stringham, J. Tuomilehto, E. J. Benjamin, S. J. Hwang, D. Levy, S. R. Preis, R. S. Vasan, J. Duan, P. V. Gejman, D. F. Levinson, A. R. Sanders, J. X. Shi, E. H. Lips, J. D. McKay, A. Agudo, L. Barzan, V. Bencko, S. Benhamou, X. Castellsague, C. Canova, D. I. Conway, E. Fabianova, L. Foretova, V. Janout, C. M. Healy, I. Holcatova, K. Kjaerheim, P. Lagiou, J. Lissowska, R. Lowry, T. V. Macfarlane, D. Mates, L. Richiardi, P. Rudnai, N. Szeszenia-Dabrowska, D. Zaridze, A. Znaor, M. Lathrop, P. Brennan, S. Bandinelli, T. M. Frayling, J. M. Guralnik, Y. Milaneschi, J. R. B. Perry, D. Altshuler, R. Elosua, S. Kathiresan, G. Lucas, O. Melander, C. J. O'Donnell, V. Salomaa, S. M. Schwartz, B. F. Voight, B. W. Penninx, J. H. Smit, N. Vogelzangs, D. I. Boomsma, E. J. C. de Geus, J. M. Vink, G. Willemsen, S. J. Chanock, F. Y. Gu, S. E. Hankinson, D. J. Hunter, A. Hofman, H. Tiemeier, A. G. Uitterlinden, C. M. van Duijn, S. Walter, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*, 42(5):441–U134, 2010.

[133] S. H. Gage, S. Zammit, and M. Hickman. Stronger evidence is needed before accepting that cannabis plays an important role in the aetiology of schizophrenia in the population. *F1000 Med Rep*, 5:2, 2013.

[134] J. Gelernter, H. R. Kranzler, R. Sherva, R. Koesterer, L. Almasy, H. Zhao, and L. A. Farrer. Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. *Biol Psychiatry*, 76(1):66–74, 2014.

[135] J. Gelernter, C. Panhuysen, R. Weiss, K. Brady, J. Poling, M. Krauthammer, L. Farrer, and H. R. Kranzler. Genomewide linkage scan for nicotine dependence: identification of a chromosome 5 risk locus. *Biol Psychiatry*, 61(1):119–26, 2007.

[136] J. Gelernter, R. Sherva, R. Koesterer, L. Almasy, H. Zhao, H. R. Kranzler, and L. Farrer. Genome-wide association study of cocaine dependence and related traits: Fam53b identified as a risk gene. *Mol Psychiatry*, 19(6):717–23, 2014.

[137] Joel Gelernter, Yi Yu, Roger Weiss, Kathleen Brady, Carolien Panhuysen, Bao-zhu Yang, Henry R Kranzler, and Lindsay Farrer. Haplotype spanning ttc12 and ankk1, flanked by the drd2 and ncam1 loci, is strongly associated to nicotine dependence in two distinct american populations. *Human molecular genetics*, 15(24):3498–3507, 2006.

[138] M. Gibbs, C. Winsper, S. Marwaha, E. Gilbert, M. Broome, and S. P. Singh. Cannabis use and mania symptoms: A systematic review and meta-analysis. *J Affect Disord*, 171c:39–47, 2014.

[139] N. A. Gillespie, G. H. Lubke, C. O. Gardner, M. C. Neale, and K. S. Kendler. Two-part random effects growth modeling to identify risks associated with alcohol and cannabis initiation, initial average use and changes in drug consumption in a sample of adult, male twins. *Drug Alcohol Depend*, 123(1-3):220–8, 2012.

[140] N. A. Gillespie, M. C. Neale, and K. S. Kendler. Pathways to cannabis abuse: a multistage model from cannabis availability, cannabis initiation and progression to abuse. *Addiction*, 104(3):430–438, 2009.

[141] Nathan A Gillespie, David E Evans, Margie M Wright, and Nicholas G Martin. Genetic simplex modeling of eysenck's dimensions of personality in a sample of young australian twins. *Twin Research*, 7(06):637–648, 2004.

[142] Nathan A Gillespie, Anjali K Henders, Tracy A Davenport, Daniel F Hermens, Margie J Wright, Nicholas G Martin, and Ian B Hickie. The brisbane longitudinal twin study: pathways to cannabis use, abuse, and dependence projectcurrent status, preliminary results, and future directions. *Twin Research and Human Genetics*, 16(01):21–33, 2013.

[143] Derek Gordon and Stephen J Finch. Factors affecting statistical power in the detection of genetic association. *Journal of Clinical Investigation*, 115(6):1408, 2005.

[144] Gregor Gorjanc, David A Henderson, and Maintainer David Henderson. Package geneticsped.

[145] J. D. Grant, M. T. Lynskey, J. F. Scherrer, A. Agrawal, A. C. Heath, and K. K. Bucholz. A cotwin-control analysis of drug use and abuse/dependence risk associated with early-onset cannabis use. *Addictive Behaviors*, 35(1):35–41, 2010.

[146] J. D. Grant, J. F. Scherrer, M. T. Lynskey, A. Agrawal, A. E. Duncan, J. R. Haber, A. C. Heath, and K. K. Bucholz. Associations of alcohol, nicotine, cannabis, and drug use/dependence with educational attainment: evidence from cotwin-control analyses. *Alcohol Clin Exp Res*, 36(8):1412–20, 2012.

[147] William H. Greene. Econometric analysis. *New Jersey: Prentice Hall*, 2003.

[148] Sonja Greven, Ciprian M Crainiceanu, Helmut Kchenhoff, and Annette Peters. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4):870–891, 2008.

[149] Guang Guo and Jianmin Wang. The mixed or multilevel model for behavior genetic analysis. *Behavior genetics*, 32(1):37–49, 2002.

[150] William C Hahn and Robert Allan Weinberg. *A subway map of cancer pathways*. Nature Publishing Group, 2002.

[151] W. Hall and T. F. Babor. Cannabis use and public health: assessing the burden. *Addiction*, 95(4):485–490, 2000.

[152] W. Hall and N. Solowij. Adverse effects of cannabis. *Lancet*, 352(9140):1611–1616, 1998.

[153] Wayne Hall. What has research over the past two decades revealed about the adverse health effects of recreational cannabis use? *Addiction*, 110(1):19–35, 2015.

[154] Christie A Hartman, Christian J Hopfer, Brett Haberstick, Soo Hyun Rhee, Thomas J Crowley, Robin P Corley, John K Hewitt, and Marissa A Ehringer. The association between cannabinoid receptor 1 gene (cnr1) and cannabis dependence symptoms in adolescents and young adults. *Drug Alcohol Depend*, 104(1):11–16, 2009.

[155] Bent Harvald, Gudrun Hauge, Kirsten Ohm Kyvik, Kaare Christensen, Axel Skytthe, and Niels V Holm. The danish twin registry: past and present. *Twin research*, 7(04):318–335, 2004.

[156] H. M. Haughey, E. Marshall, J. P. Schacht, A. Louis, and K. E. Hutchison. Marijuana withdrawal and craving: influence of the cannabinoid receptor 1 (cnr1) and fatty acid amide hydrolase (faah) genes. *Addiction*, 103(10):1678–1686, 2008.

[157] M. R. Hayatbakhsh, M. J. O'Callaghan, A. A. Mamun, G. M. Williams, A. Clavarino, and J. M. Najman. Cannabis use and obesity and young adults. *Am J Drug Alcohol Abuse*, 36(6):350–6, 2010.

[158] Aryeh I Herman, Henry R Kranzler, Joseph F Cubells, Joel Gelernter, and Jonathan Covault. Association study of the cnr1 gene exon 3 alternative promoter region polymorphisms and substance dependence. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141(5):499–503, 2006.

[159] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[160] Rosa A Hoekstra, Meike Bartels, and Dorret I Boomsma. Longitudinal genetic study of verbal and nonverbal iq from early childhood to young adulthood. *Learning and Individual Differences*, 17(2):97–114, 2007.

[161] C. J. Hopfer, J. M. Lessem, C. A. Hartman, M. C. Stallings, S. S. Cherny, R. P. Corley, J. K. Hewitt, K. S. Krauter, S. K. Mikulich-Gilbertson, S. H. Rhee, A. Smolen, S. E. Young, and T. J. Crowley. A genome-wide scan for loci influencing adolescent cannabis dependence symptoms: Evidence for linkage on chromosomes 3 and 9. *Drug and Alcohol Dependence*, 89(1):34–41, 2007.

[162] Christian J Hopfer, Susan E Young, Shaun Purcell, Thomas J Crowley, Michael C Stallings, Robin P Corley, Soo Hyun Rhee, Andrew Smolen, Ken Krauter, and John K Hewitt. Cannabis receptor haplotype associated with fewer cannabis dependence symptoms in adolescents. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141(8):895–901, 2006.

[163] L John Horwood, David M Fergusson, Mohammad R Hayatbakhsh, Jake M Najman, Carolyn Coffey, George C Patton, Edmund Silins, and Delyse M Hutchinson. Cannabis use and educational achievement: Findings from three australasian cohort studies. *Drug and Alcohol Dependence*, 110(3):247–253, 2010.

[164] Jouke-Jan Hottenga and Dorret I Boomsma. Qtl detection in multivariate data from sibling pairs. *Statistical Genetics*, page 239, 2012.

[165] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Goncalo R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 44(8):955–959, 2012.

[166] Hailiang Huang, Pritam Chanda, Alvaro Alonso, Joel S Bader, and Dan E Arking. Gene-based tests of association. *PLoS Genet*, 7(7):e1002177, 2011.

[167] Carl J Huberty and John D Morris. Multivariate analysis versus multiple univariate analyses. *Psychological bulletin*, 105(2):302, 1989.

[168] John R Hughes, Lindsay F Stead, and Tim Lancaster. Anxiolytics for smoking cessation. *The Cochrane Library*, 2000.

[169] Rayjean J Hung, James D McKay, Valerie Gaborieau, Paolo Boffetta, Mia Hashibe, David Zaridze, Anush Mukeria, Neonilia Szeszenia-Dabrowska, Jolanta Lissowska, and Peter Rudnai. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, 452(7187):633–637, 2008.

[170] Yoon-Mi Hur and Jeffrey M Craig. Twin registries worldwide: an important resource for scientific research. *Twin Research and Human Genetics*, 16(01):1–12, 2013.

[171] Rachel R Huxley and Mark Woodward. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *The Lancet*, 378(9799):1297–1305, 2011.

[172] Jeroen R Huyghe, Anne U Jackson, Marie P Fogarty, Martin L Buchkovich, Alena Stankov, Heather M Stringham, Xueling Sim, Lingyao Yang, Christian Fuchsberger, and Henna Cederberg. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature genetics*, 45(2):197–201, 2013.

[173] S. M. Hyman and R. Sinha. Stress-related factors in cannabis use and misuse: implications for prevention and treatment. *J Subst Abuse Treat*, 36(4):400–13, 2009.

[174] C. A. Ibrahim-Verbaas, J. Bressler, S. Debette, M. Schuur, A. V. Smith, J. C. Bis, G. Davies, S. Trompet, J. A. Smith, C. Wolf, L. B. Chibnik, Y. Liu, V. Vitart, M. Kirin, K. Petrovic, O. Polasek, L. Zgaga, C. Fawns-Ritchie, P. Hoffmann, J. Karjalainen, J. Lahti, D. J. Llewellyn, C. O. Schmidt, K. A. Mather, V. Chouraki, Q. Sun, S. M. Resnick, L. M. Rose, C. Oldmeadow, M. Stewart, B. H. Smith, V. Gudnason, Q. Yang, S. S. Mirza, J. W. Jukema, P. L. deJager, T. B. Harris, D. C. Liewald, N. Amin, L. H. Coker, O. Stegle, O. L. Lopez, R. Schmidt, and A.I Teumer. Gwas for executive function and processing speed suggests involvement of the cadm2 gene. 2015.

[175] SAS Institute Inc. SAS version 9.3 for Linux Electronic Software Delivery (ESD), Cary, NC, USA. 2011.

[176] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D Buxbaum, and Xihong Lin. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, 2013.

[177] Ivan Iossifov, Michael Ronemus, Dan Levy, Zihua Wang, Inessa Hakker, Julie Rosenbaum, Boris Yamrom, Yoon-ha Lee, Giuseppe Narzisi, and Anthony Leotta. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–299, 2012.

[178] Manish Joshi, Anita Joshi, and Thaddeus Bartter. Marijuana and lung diseases. *Current opinion in pulmonary medicine*, 20(2):173–179, 2014.

[179] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[180] L. M. Kelleher, C. Stough, A. A. Sergejew, and T. Rolfe. The effects of cannabis on information-processing speed. *Addict Behav*, 29(6):1213–9, 2004.

[181] Ann E Kelley. Nicotinic receptors: addiction's smoking gun? *Nature medicine*, 8(5):447–449, 2002.

[182] Kenneth S Kendler and Carol A Prescott. Cannabis use, abuse, and dependence in a population-based sample of female twins. *American Journal of Psychiatry*, 155(8):1016–1022, 1998.

[183] Kenneth S Kendler, Eric Schmitt, Steven H Aggen, and Carol A Prescott. Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. *Archives of general psychiatry*, 65(6):674–682, 2008.

[184] J Kettunen, M Perola, NG Martin, BK Cornes, SG Wilson, GW Montgomery, B Benyamin, JR Harris, D Boomsma, and G Willemsen. Multicenter dizygotic twin cohort study confirms two linkage susceptibility loci for body mass index at 3q29 and 7q36 and identifies three further potential novel loci. *International journal of obesity*, 33(11):1235–1242, 2009.

[185] Amy K Kiefer, Joyce Y Tung, Chuong B Do, David A Hinds, Joanna L Mountain, Uta Francke, and Nicholas Eriksson. Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS genetics*, 9(2):e1003299, 2013.

[186] K. M. King and L. Chassin. A prospective study of the effects of age of initiation of alcohol and drug use on young adult substance dependence. *J Stud Alcohol Drugs*, 68(2):256–65, 2007.

[187] BP Kinghorn. An index of information content for genotype probabilities derived from segregation analysis. *Genetics*, 145(2):479–483, 1997.

[188] By BP Kinghorn. Use of segregation analysis to reduce genotyping costs. *Journal of Animal Breeding and Genetics*, 116(3):175–180, 1999.

[189] Daniel D Kinnamon, Ray E Hershberger, and Eden R Martin. Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. *PLoS One*, 7(2):e30238, 2012.

[190] Einar Kristjansson, Peter Allebeck, and Brje Wistedt. Validity of the diagnosis schizophrenia in a psychiatric inpatient register: a retrospective application of dsm-iii criteria on icd-8 diagnoses in stockholm county. *Nordic Journal of Psychiatry*, 41(3):229–234, 1987.

[191] Gregory V Kryukov, Len A Pennacchio, and Shamil R Sunyaev. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, 80(4):727–739, 2007.

[192] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010.

[193] M. Laucht, K. Becker, J. Frank, M. H. Schmidt, G. Esser, J. Treutlein, M. H. Skowronek, and G. Schumann. Genetic variation in dopamine pathways differentially associated with smoking progression in adolescence. *J Am Acad Child Adolesc Psychiatry*, 47(6):673–81, 2008.

[194] Derrick Norman Lawley and Albert Ernest Maxwell. *Factor analysis as a statistical method*. Butterworths, London, 1971.

[195] Byung Dae Lee, Je Min Park, Young Min Lee, Eun Soo Moon, Hee Jeong Jeong, Young In Chung, and Hyo Deog Rim. A pilot study for discovering candidate genes of chromosome 18q21 in methamphetamine abusers: Case-control association study. *Clinical Psychopharmacology and Neuroscience*, 12(1):54–64, 2014.

[196] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012.

[197] Seunggeun Shawn Lee and Maintainer Seunggeun Shawn Lee. Package skat, 2014.

[198] Christina N Lessov, Nicholas G Martin, Dixie J Statham, Alexandre A Todorov, Wendy S Slutske, Kathleen K Bucholz, Andrew C Heath, and Pamela AF Madden. Defining nicotine dependence for genetic research: evidence from australian twins. *Psychological medicine*, 34(05):865–879, 2004.

[199] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[200] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[201] M. X. Li, H. S. Gui, J. S. Kwan, and P. C. Sham. Gates: a rapid and powerful gene-based association test using extended simes procedure. *Am J Hum Genet*, 88(3):283–93, 2011.

[202] Miao-Xin Li, Johnny SH Kwan, and Pak C Sham. Hyst: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. *The American Journal of Human Genetics*, 91(3):478–488, 2012.

[203] Min Li and Pumin Zhang. The function of apc/ccdh1 in cell cycle and beyond. *Cell division*, 4(1):2, 2009.

[204] Xiang Li, Saonli Basu, Michael B Miller, William G Iacono, and Matt McGue. A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. *Human heredity*, 71(1):67–82, 2011.

[205] Yun Li, Cristen Willer, Serena Sanna, and Gonçalo Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387, 2009.

[206] Dan-Yu Lin and Zheng-Zheng Tang. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367, 2011.

[207] P. A. Lind, S. Macgregor, A. Agrawal, G. W. Montgomery, A. C. Heath, N. G. Martin, and J. B. Whitfield. The role of gabra2 in alcohol dependence, smoking, and illicit drug use in an australian population sample. *Alcoholism-Clinical and Experimental Research*, 32(10):1721–1731, 2008.

[208] Leonard Lipovich, Fabien Dachet, Juan Cai, Shruti Bagla, Karina Balan, Hui Jia, and Jeffrey A Loeb. Activity-dependent human brain coding/noncoding gene regulatory networks. *Genetics*, 192(3):1133–1148, 2012.

[209] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

[210] Christoph Lippert, Jing Xiang, Danilo Horta, Christian Widmer, Carl Kadie, David Heckerman, and Jennifer Listgarten. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics*, 30(22):32063214, 2014.

[211] Jennifer Listgarten, Christoph Lippert, Eun Yong Kang, Jing Xiang, Carl M Kadie, and David Heckerman. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, 29(12):1526–1533, 2013.

[212] Saskia Litière, Ariel Alonso, and Geert Molenberghs. Type i and type ii error under random-effects misspecification in generalized linear mixed models. *Biometrics*, 63(4):1038–1044, 2007.

[213] Dajiang J Liu, Gina M Peloso, Xiaowei Zhan, Oddgeir L Holmen, Matthew Zawistowski, Shuang Feng, Majid Nikpay, Paul L Auer, Anuj Goel, and He Zhang. Meta-analysis of gene-level tests for rare variant association. *Nature genetics*, 46(2):200–204, 2014.

[214] Jason Z Liu, Federica Tozzi, Dawn M Waterworth, Sreekumar G Pillai, Pierandrea Muglia, Lefkos Middleton, Wade Berrettini, Christopher W Knouff, Xin Yuan, and Grard Waeber. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature genetics*, 42(5):436–440, 2010.

[215] Jimmy Z Liu, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, Nicholas K Hayward, Grant W Montgomery, Peter M Visscher, and Nicholas G Martin. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.

[216] Meng Liu, Rui Fan, Xinhua Liu, Feng Cheng, and Ju Wang. Pathways and networks-based analysis of candidate genes associated with nicotine addiction. 2015.

[217] Kirk E Lohmueller, Thomas Spars, Qibin Li, Ehm Andersson, Thorfinn Korneliussen, Anders Albrechtsen, Karina Banasik, Niels Grarup, Ingileif Hallgrimsdottir, and Kristoffer Kiil. Whole-exome sequencing of 2,000 danish individuals and the role of rare coding variants in type 2 diabetes. *The American Journal of Human Genetics*, 93(6):1072–1086, 2013.

[218] Anu Loukola, Juho Wedenoja, Kaisu Keskitalo-Vuokko, Ulla Broms, Tellervo Korhonen, Samuli Ripatti, Antti-Pekka Sarin, Janne Pitkniemi, Liang He, and Anja Hppl. Genome-wide association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample. *Mol Psychiatry*, 19(5):615–624, 2014.

[219] Jennifer K Lowe, Julian B Maller, Benjamin M Neale, Jacqueline Salit, Eimear E Kenny, Jessica L Shea, Ralph Burkhardt, J Gustav Smith, Weizhen Ji, et al. Genome-wide association studies in an isolated founder population from the pacific island of kosrae.

[220] Gitta H Lubke, Jouke Jan Hottenga, Raymond Walters, Charles Laurin, Eco JC de Geus, Gonneke Willemsen, Jan H Smit, Christel M Middeldorp, Brenda WJH Penninx, and Jacqueline M Vink. Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol Psychiatry*, 72(8):707–709, 2012.

[221] P. Lucas. Cannabis as an adjunct to or substitute for opiates in the treatment of chronic pain. *J Psychoactive Drugs*, 44(2):125–33, 2012.

[222] M. T. Lynskey, A. Agrawal, A. Henders, E. C. Nelson, P. A. Madden, and N. G. Martin. An australian twin study of cannabis and other illicit drug use and misuse, and other psychopathology. *Twin Res Hum Genet*, 15(5):631–41, 2012.

[223] Michael Lynskey and Wayne Hall. The effects of adolescent cannabis use on educational attainment: a review. *Addiction*, 95(11):1621–1630, 2000.

[224] Michael T Lynskey, Andrew C Heath, Kathleen K Bucholz, Wendy S Slutske, Pamela AF Madden, Elliot C Nelson, Dixie J Statham, and Nicholas G Martin. Escalation of drug use in early-onset cannabis users vs co-twin controls. *Jama*, 289(4):427–433, 2003.

[225] Michael T Lynskey, Jacqueline M Vink, and Dorret I Boomsma. Early onset cannabis use and progression to other drug use in a sample of dutch twins. *Behavior Genetics*, 36(2):195–200, 2006.

[226] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. On the practice of dichotomization of quantitative variables. *Psychological methods*, 7(1):19, 2002.

[227] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384, 2009.

[228] Patrik KE Magnusson, Catarina Almqvist, Iffat Rahman, Andrea Ganna, Alexander Viktorin, Hasse Walum, Linda Halldner, Sebastian Lundstrm, Fredrik Ulln, and Niklas Lngstrm. The swedish twin registry: establishment of a biobank and other recent developments. *Twin Research and Human Genetics*, 16(01):317–329, 2013.

[229] Arnab Maity, Patrick F Sullivan, and Juning Tzeng. Multivariate phenotype association analysis by markerset kernel machine regression. *Genetic epidemiology*, 36(7):686–695, 2012.

[230] Eusebio Manchado, Manuel Eguren, and Marcos Malumbres. The anaphase-promoting complex/cyclosome (apc/c): cell-cycle-dependent and-independent functions. *Biochemical Society Transactions*, 38(1):65, 2010.

[231] Mirko Manchia, Jeffrey Cullis, Gustavo Turecki, Guy A Rouleau, Rudolf Uher, and Martin Alda. The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PloS one*, 8(10):e76295, 2013.

[232] Frantisek Mandys, Conor V Dolan, and Peter CM Molenaar. Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics*, 19(3):201–215, 1994.

[233] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[234] NG Martin, LJ Eaves, et al. The genetical analysis of covariance structure. *Heredity*, 38(1):79–95, 1977.

[235] Nicholas Martin, Dorret Boomsma, and Geoffrey Machin. A twin-pronged attack on complex traits. *Nature genetics*, 17(4):387–392, 1997.

[236] J. M. Mason and K. M. Arndt. Coiled coil domains: stability, specificity, and biological implications. *Chembiochem*, 5(2):170–6, 2004.

[237] K. Mather and J.L. Jinks. Introduction to biometrical genetics, ithaca, ny: Cornell university press, 1977.

[238] JJ McArdle and HH Goldsmith. Structural equation modeling applied to the twin design - comparative multivariate models of the wais. In *Behavior Genetics*, volume 14. PLENUM PUBL CORP 233 SPRING ST, NEW YORK, NY 10013, 1984.

[239] John J McArdle and Carol A Prescott. Mixed-effects variance components models for biometric family analyses. *Behavior genetics*, 35(5):631–652, 2005.

[240] RP McDonald. *Test theory: A unied treatment.* Mahwah, NJ: LEA. 1999.

[241] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, and Mark Daly. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[242] Sarah Elizabeth Medland and Michael Churton Neale. An integrated phenomic approach to multivariate allelic association. *European Journal of Human Genetics*, 18(2):233–239, 2010.

[243] Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551–1566, 2013.

[244] Camelia C Minică, Conor V Dolan, Jouke-Jan Hottenga, Gonneke Willemsen, Jacqueline M Vink, and Dorret I Boomsma. The use of imputed sibling genotypes in sibship-based association analysis: on modeling alternatives, power and model misspecification. *Behavior genetics*, 43(3):254–266, 2013.

[245] Camelia C Minica, Conor V Dolan, Maarten MD Kampert, Dorret I Boomsma, and Jacqueline M Vink. Sandwich corrected standard errors in family-based genome-wide association studies. *European Journal of Human Genetics 2014; e-pub ahead of print 11 June 2014; DOI: 10.1038/ejhg.2014.94.*, 2014.

[246] Camelia C Minic, Conor V Dolan, Jouke-Jan Hottenga, Ren Pool, Iryna O Fedko, Hamdi Mbarek, Charlotte Huppertz, Meike Bartels, Dorret I Boomsma, and Jacqueline M Vink. Heritability, snp-and gene-based analyses of cannabis use initiation and age at onset. *Behavior Genetics*, pages 1–11, 2015.

[247] Farzaneh Modarresi, Mohammad Ali Faghihi, Miguel A Lopez-Toledano, Roya Pedram Fatemi, Marco Magistri, Shaun P Brothers, Marcel P van der Brug, and Claes Wahlestedt. Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nature biotechnology*, 30(5):453–459, 2012.

[248] Donald F Morrison. Multivariate statistical methods. 3. *New York, NY. Mc*, 1990.

[249] Nina Roth Mota, Eli Vieira Araujo-Jnr, Vanessa Rodrigues Paixo-Crtes, Maria Ctira Bortolini, and Claiton Henrique Dotto Bau. Linking dopamine neurotransmission and neurogenesis: the evolutionary history of the ntad (ncam1-ttc12-ankk1-drd2) gene cluster. *Genetics and molecular biology*, 35(4):912–918, 2012.

[250] M. Munafo, T. Clark, E. Johnstone, M. Murphy, and R. Walton. The genetic basis for smoking behavior: a systematic review and meta-analysis. *Nicotine Tob Res*, 6(4):583–97, 2004.

[251] Jill C Mwenifumbo and Rachel F Tyndale. Genetic variability in cyp2a6 and the pharmacokinetics of nicotine. *Pharmacogenomics*, 8(10):1385–1402, 2007.

[252] Benjamin M Neale and Pak C Sham. The future of association studies: gene-based analysis and replication. *The American Journal of Human Genetics*, 75(3):353–362, 2004.

[253] Michael Neale and Lon Cardon. *Methodology for genetic studies of twins and families.* Springer Science & Business Media, 1992.

[254] Michael C Neale, Steven M Boker, Gary Xie, and H Mx Maes. *Mx: Statistical modeling.* Richmond, Virginia: Department of Psychiatry. 1999.

[255] Elliot C Nelson, Michael T Lynskey, Andrew C Heath, Naomi Wray, Arpana Agrawal, Fiona L Shand, Anjali K Henders, Leanne Wallace, Alexandre A Todorov, and Andrew J Schrage. Ankk1, ttc12, and ncam1 polymorphisms and heroin dependence: importance of considering drug exposure. *JAMA psychiatry*, 70(3):325–333, 2013.

[256] Dianne F Newbury, Laura Winchester, Laura Addis, Silvia Paracchini, Lyn-Louise Buckingham, Ann Clark, Wendy Cohen, Hilary Cowie, Katharina Dworzynski, and Andrea Everitt. Cmip and atp2c2 modulate phonological short-term memory in language impairment. *The American Journal of Human Genetics*, 85(2):264–272, 2009.

[257] Marie Ng, Michael K Freeman, Thomas D Fleming, Margaret Robinson, Laura Dwyer-Lindgren, Blake Thomson, Alexandra Wollum, Ella Sanman, Sarah Wulf, and Alan D Lopez. Smoking prevalence and cigarette consumption in 187 countries, 1980-2012. *Jama*, 311(2):183–192, 2014.

[258] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[259] Dale R Nyholt. Seca: Snp effect concordance analysis using genome-wide association summary results. *Bioinformatics*, 30(14):20862088, 2014.

[260] Jane M Olson, John S Witte, and Robert C Elston. Tutorial in biostatistics genetic mapping of complex traits. *Statist. Med*, 18:2961–2981, 1999.

[261] World Health Organisation. *http://www.who.int/tobacco/health_priority/en/.* Thesis, 2015.

[262] Paul F OReilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and LJ Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

[263] Michael J Parsons, Kathryn J Lester, Nicola L Barclay, Patrick M Nolan, Thalia C Eley, and Alice M Gregory. Replication of genome-wide association studies (gwas) loci for sleep in the british g1219 cohort. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 162(5):431–438, 2013.

[264] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, 32(4):381–385, 2008.

[265] Gina M Peloso, Paul L Auer, Joshua C Bis, Arend Voorman, Alanna C Morrison, Nathan O Stitziel, Jennifer A Brody, Sumeet A Khetarpal, Jacy R Crosby, and Myriam Fornage. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics*, 94(2):223–232, 2014.

[266] Leena Peltonen. Genomeutwin: a strategy to identify genetic influences on health and disease. *Twin Research*, 6(05):354–360, 2003.

[267] Andrew Percy and BP Kinghorn. A genotype probability index for multiple alleles and haplotypes. *Journal of Animal Breeding and Genetics*, 122(6):387–392, 2005.

[268] Markus Perola, Sampo Sammalisto, Tero Hiekkalinna, Nick G Martin, Peter M Visscher, Grant W Montgomery, Beben Benyamin, Jennifer R Harris, Dorret Boomsma, and Gonneke Willemsen. Combined genome scans for body stature in 6,602 european twins: evidence for common caucasian loci. *PLoS genetics*, 3(6):e97, 2007.

[269] J Pinheiro, D Bates, S DebRoy, D Sarkar, and R Core Team. nlme: Linear and nonlinear mixed effects models. *R package version*, 3:118, 2014.

[270] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. nlme: Linear and nonlinear mixed effects models. *R package version*, 3:103, 2013.

[271] Jos C Pinheiro and Douglas M Bates. *Mixed-effects models in S and S-PLUS.* Springer Science & Business Media, 2000.

[272] Matti Pirinen, Peter Donnelly, Chris CA Spencer, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics*, 7(1):369–390, 2013.

[273] R. A. Power, K. J. Verweij, M. Zuhair, G. W. Montgomery, A. K. Henders, A. C. Heath, P. A. Madden, S. E. Medland, N. R. Wray, and N. G. Martin. Genetic predisposition to schizophrenia associated with increased use of cannabis. *Mol Psychiatry*, 2014.

[274] CA Prescott, PF Sullivan, PH Kuo, BT Webb, J Vittum, DG Patterson, DL Thiselton, JM Myers, M Devitt, and LJ Halberstadt. Genomewide linkage study in the irish affected sib pair study of alcohol dependence: evidence for a susceptibility region for symptoms of alcohol dependence on chromosome 4. *Molecular psychiatry*, 11(6):603–611, 2006.

[275] Centers for Disease Control and Prevention. Smoking-attributable mortality, years of potential life lost, and productivity lossesunited states, 20002004. *MMWR Morb Mortal Wkly Rep*, 57:12261228, 2008.

[276] Alkes L Price, Gregory V Kryukov, Paul IW de Bakker, Shaun M Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010.

[277] Jonathan K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69(1):124–137, 2001.

[278] Randall J Pruim, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Gonalo R Abecasis, and Cristen J Willer. Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337, 2010.

[279] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, and Mark J Daly. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[280] Shaun M Purcell, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm ODushlaine, Kimberly Chambert, Sarah E Bergen, and Anna Khler. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–190, 2014.

[281] Sophia Rabe-Hesketh, Anders Skrondal, and HK Gjessing. Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, 64(1):280–288, 2008.

[282] R. Radhakrishnan, S. T. Wilkinson, and D. C. D'Souza. Gone to pot - a review of the association between cannabis and psychosis. *Front Psychiatry*, 5:54, 2014.

[283] Cornelius A Rietveld, Sarah E Medland, Jaime Derringer, Jian Yang, Tonu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, and Arpana Agrawal. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, 340(6139):1467–1471, 2013.

[284] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L Moran, Anna K Khler, Susanne Akterin, Sarah E Bergen, Ann L Collins, James J Crowley, and Menachem Fromer. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10):1150–1159, 2013.

[285] WH Rogers. Regression standard errors in clustered samples. *Stata Technical Bulletin*, 13:19–23, 1993.

[286] T. Rubinek, R. Yu, M. Hadani, G. Barkai, D. Nass, S. Melmed, and I. Shimon. The cell adhesion molecules n-cadherin and neural cell adhesion molecule regulate human growth hormone: a novel mechanism for regulating pituitary hormone secretion. *J Clin Endocrinol Metab*, 88(8):3724–30, 2003.

[287] Nancy L Saccone, Jen C Wang, Naomi Breslau, Eric O Johnson, Dorothy Hatsukami, Scott F Saccone, Richard A Grucza, Lingwei Sun, Weimin Duan, and John Budde. The chrna5-chrna3-chrnb4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in african-americans and in european-americans. *Cancer research*, 69(17):6848–6856, 2009.

[288] Scott F Saccone, Anthony L Hinrichs, Nancy L Saccone, Gary A Chase, Karel Konvicka, Pamela AF Madden, Naomi Breslau, Eric O Johnson, Dorothy Hatsukami, and Ovide Pomerleau. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 snps. *Human molecular genetics*, 16(1):36–49, 2007.

[289] Kaitlin E Samocha, Elise B Robinson, Stephan J Sanders, Christine Stevens, Aniko Sabo, Lauren M McGrath, Jack A Kosmicki, Karola Rehnstrm, Swapan Mallick, and Andrew Kirby. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–950, 2014.

[290] Carolyn E Sartor, Arpana Agrawal, Michael T Lynskey, Kathleen K Bucholz, Pamela AF Madden, and Andrew C Heath. Common genetic influences on the timing of first use for alcohol, cigarettes, and cannabis in young african-american women. *Drug and alcohol dependence*, 102(1):49–55, 2009.

[291] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[292] Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–219, 2009.

[293] TaeHwi Schwantes-An, Robert Culverhouse, Weimin Duan, Shelina Ramnarine, John P Rice, and Nancy L Saccone. Interpreting joint snp analysis results: when are two distinct signals really two distinct signals? *Genet Epidemiol*, 37(3):301–309, 2013.

[294] Steven G Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[295] Pak C Sham and Shaun M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, 2014.

[296] Lifu Sheng, Iryna Leshchyns' ka, and Vladimir Sytnyk. Cell adhesion and intracellular calcium signaling in neurons. *Cell Commun Signal*, 11:94, 2013.

[297] Daniel Shriner. Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Frontiers in genetics*, 3:1, 2012.

[298] Karri Silventoinen, Sampo Sammalisto, Markus Perola, Dorret I Boomsma, Belinda K Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R Harris, Jacob VB Hjelmborg, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin research*, 6(05):399–408, 2003.

[299] HonCheong So, Miaoxin Li, and Pak C Sham. Uncovering the total heritability explained by all true susceptibility variants in a genomewide association study. *Genetic epidemiology*, 35(6):447–456, 2011.

[300] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013.

[301] Tim D Spector and Frances MK Williams. The uk adult twin registry (twinsuk). *Twin Research and Human Genetics*, 9(06):899–906, 2006.

[302] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[303] E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. Lango Allen, C. M. Lindgren, J. Luan, R. Magi, J. C. Randall, S. Vedantam, T. W. Winkler, L. Qi, T. Workalemahu, I. M. Heid, V. Steinthorsdottir, H. M. Stringham, M. N. Weedon, E. Wheeler, A. R. Wood, T. Ferreira, R. J. Weyant, A. V. Segre, K. Estrada, L. Liang, J. Nemesh, J. H. Park, S. Gustafsson, T. O. Kilpelainen, J. Yang, N. Bouatia-Naji, T. Esko, M. F. Feitosa, Z. Kutalik, M. Mangino, S. Raychaudhuri, A. Scherag, A. V. Smith, R. Welch, J. H. Zhao, K. K. Aben, D. M. Absher, N. Amin, A. L. Dixon, E. Fisher, N. L. Glazer, M. E. Goddard, N. L. Heard-Costa, V. Hoesel, J. J. Hottenga, A. Johansson, T. Johnson, S. Ketkar, C. Lamina, S. Li, M. F. Moffatt, R. H. Myers, N. Narisu, J. R. Perry, M. J. Peters, M. Preuss, S. Ripatti, F. Rivadeneira, C. Sandholt, L. J. Scott, N. J. Timpson, J. P. Tyrer, S. van Wingerden, R. M. Watanabe, C. C. White, F. Wiklund, C. Barlassina, D. I. Chasman, M. N. Cooper, J. O. Jansson, R. W. Lawrence, N. Pellikka, I. Prokopenko, J. Shi, E. Thiering, H. Alavere, M. T. Alibrandi, P. Almgren, A. M. Arnold, T. Aspelund, L. D. Atwood, B. Balkau, A. J. Balmforth, A. J. Bennett, Y. Ben-Shlomo, R. N. Bergman, S. Bergmann, H. Biebermann, A. I. Blakemore, T. Boes, L. L. Bonnycastle, S. R. Bornstein, M. J. Brown, T. A. Buchanan, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–48, 2010.

[304] Margaret R Spitz, Christopher I Amos, Qiong Dong, Jie Lin, and Xifeng Wu. The chrna5-a3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *Journal of the National Cancer Institute*, 100(21):1552–1556, 2008.

[305] Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8(7):e65245, 2013.

[306] James P Stevens. *Applied multivariate statistics for the social sciences.* Routledge, 2012.

[307] A. I. Stiby, M. Hickman, M. R. Munafo, J. Heron, V. L. Yip, and J. Macleod. Adolescent cannabis and tobacco use and educational outcomes at age 16: birth cohort study. *Addiction*, 110(4):658–68, 2015.

[308] Reinoud D Stoel, Francisca Galindo Garre, Conor Dolan, and Godfried Van Den Wittenboer. On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, 11(4):439, 2006.

[309] P. F. Sullivan, M. J. Daly, and M. O'Donovan. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet*, 13(8):537–51, 2012.

[310] Patrick F Sullivan. The psychiatric gwas consortium: big science comes to psychiatry. *Neuron*, 68(2):182–186, 2010.

[311] Patrick F Sullivan, Richard SE Keefe, Leslie A Lange, Ethan M Lange, T Scott Stroup, Jeffrey Lieberman, and Patricia F Maness. Ncam1 and neurocognition in schizophrenia. *Biological psychiatry*, 61(7):902–910, 2007.

[312] Gulnara R Svishcheva, Nadezhda M Belonogova, and Tatiana I Axenovich. Ffbskat: fast family-based sequence kernel association test. *PloS one*, 9(6):e99407, 2014.

[313] Luca Tamagnone. Emerging role of semaphorins as major regulatory signals and potential therapeutic targets in cancer. *Cancer Cell*, 22(2):145–152, 2012.

[314] L. Tamm, J. N. Epstein, K. M. Lisdahl, B. Molina, S. Tapert, S. P. Hinshaw, L. E. Arnold, K. Velanova, H. Abikoff, and J. M. Swanson. Impact of adhd and cannabis use on executive functioning in young adults. *Drug Alcohol Depend*, 133(2):607–14, 2013.

[315] R Core Team. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria., 2014.

[316] RDevelopment CORE TEAM. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. Report, ISBN 3900051070, URL: http://www. R-project. org, 2010.

[317] RDevelopment Core Team. R: A language and environment for statistical computing. *R foundation for Statistical Computing*, 2013.

[318] Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, and Cristen J Willer. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[319] Consortium The Genome of the Netherlands. Whole-genome sequence variation, population structure and demographic history of the dutch population. *Nat Genet*, 46(8):818–825, 2014.

[320] Terry M. Therneau. A package for survival analysis in s, r package version 2.37-7, http://cran.r-project.org/package=survival. 2014.

[321] Thorgeir E Thorgeirsson, Daniel F Gudbjartsson, Ida Surakka, Jacqueline M Vink, Najaf Amin, Frank Geller, Patrick Sulem, Thorunn Rafnar, Tnu Esko, and Stefan Walter. Sequence variants at chrnb3-chrna6 and cyp2a6 affect smoking behavior. *Nature genetics*, 42(5):448–453, 2010.

[322] Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, 42(5):441–447, 2010.

[323] Rachel F Tyndale, Jennifer I Payne, Alexandra L Gerber, and Jack C Sipe. The fatty acid amide hydrolase c385a (p129t) missense variant in cannabis users: studies of drug use and dependence in caucasians. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 144(5):660–666, 2007.

[324] George R Uhl, Tomas Drgon, Catherine Johnson, Marco F Ramoni, Frederique M Behm, and Jed E Rose. Genome-wide association for smoking cessation success in a trial of precessation nicotine replacement. *Molecular medicine*, 16(11-12):513, 2010.

[325] George R Uhl and Jana Drgonova. Cell adhesion molecules: druggable targets for modulating the connectome and brain disorders? *Neuropsychopharmacology*, 39(1):235, 2014.

[326] George R Uhl, Qing-Rong Liu, Tomas Drgon, Catherine Johnson, Donna Walther, and Jed E Rose. Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 snps. *BMC genetics*, 8(1):10, 2007.

[327] George R Uhl, Qing-Rong Liu, and Daniel Naiman. Substance abuse vulnerability loci: converging genome scanning data. *TRENDS in Genetics*, 18(8):420–425, 2002.

[328] Simon Urbanek. Rserve: Binary r server, r package version 1.7-3, http://cran.r-project.org/package=rserve. 2013.

[329] Catharina EM Van Beijsterveldt, Maria Groen-Blokhuis, Jouke Jan Hottenga, Sanja Frani, James J Hudziak, Diane Lamb, Charlotte Huppertz, Eveline de Zeeuw, Michel Nivard, and Nienke Schutte. The young netherlands twin register (yntr): longitudinal twin and family studies in over 70,000 children. *Twin Research & Human Genetics*, 16(1):252–267, 2013.

[330] Marianne B. M. Van den Bree, Eric O. Johnson, Michael C. Neale, and Roy W. Pickens. Genetic and environmental influences on drug use and abuse/dependence in male and female twins. *Drug and Alcohol Dependence*, 52(3):231–241, 1998.

[331] Edwin JCG van den Oord. Estimating effects of latent and measured genotypes in multilevel models. *Statistical Methods in Medical Research*, 10(6):393–407, 2001.

[332] Han LJ Van Der Maas, Conor V Dolan, Raoul PPP Grasman, Jelte M Wicherts, Hilde M Huizenga, and Maartje EJ Raijmakers. A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, 113(4):842, 2006.

[333] Jeroen B van der Net, A Cecile JW Janssens, Marinus JC Eijkemans, John JP Kastelein, Eric JG Sijbrands, and Ewout W Steyerberg. Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics*, 16(9):1111–1116, 2008.

[334] S Van der Sluis, D Posthuma, MG Nivard, M Verhage, and CV Dolan. Power in gwas: lifting the curse of the clinical cut-off. *Molecular psychiatry*, 18:2–3, 2013.

[335] Sophie Van der Sluis, Conor V Dolan, Jiang Li, Youqiang Song, Pak Sham, Danielle Posthuma, and Miao-Xin Li. Mgas: a powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics*, 31(7):1007–1015, 2015.

[336] Sophie Van Der Sluis, Conor V Dolan, Michael C Neale, and Danielle Posthuma. Power calculations using exact data simulation: a useful tool for genetic study designs. *Behavior genetics*, 38(2):202–211, 2008.

[337] Sophie Van der Sluis, Danielle Posthuma, and Conor V Dolan. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genetics*, 9(1):e1003235, 2013.

[338] Sophie Van Der Sluis, Matthijs Verhage, Danielle Posthuma, and Conor V Dolan. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. 2010.

[339] Tyler J VanderWeele, Kofi Asomaning, Eric J Tchetgen Tchetgen, Younghun Han, Margaret R Spitz, Sanjay Shete, Xifeng Wu, Valerie Gaborieau, Ying Wang, and John McLaughlin. Genetic variants on 15q25. 1, smoking, and lung cancer: an assessment of mediation and interaction. *American journal of epidemiology*, 175(10):1013–1020, 2012.

[340] William N Venables and Brian D Ripley. *Modern applied statistics with S*. Springer Science & Business Media, 2002.

[341] K. J. H. Verweij, B. P. Zietsch, J. Z. Liu, S. E. Medland, M. T. Lynskey, P. A. F. Madden, A. Agrawal, G. W. Montgomery, A. C. Heath, and N. G. Martin. No association of candidate genes with cannabis use in a large sample of australian twin families. *Addiction Biology*, 17(3):687–690, 2012.

[342] Karin JH Verweij, Anna AE Vinkhuyzen, Beben Benyamin, Michael T Lynskey, Lydia Quaye, Arpana Agrawal, Scott D Gordon, Grant W Montgomery, Pamela Madden, and Andrew C Heath. The genetic aetiology of cannabis use initiation: a metaanalysis of genomewide association studies and a snpbased heritability estimation. *Addiction biology*, 18(5):846–850, 2013.

[343] Karin JH Verweij, Brendan P Zietsch, Michael T Lynskey, Sarah E Medland, Michael C Neale, Nicholas G Martin, Dorret I Boomsma, and Jacqueline M Vink. Genetic and environmental influences on cannabis use initiation and problematic use: a metaanalysis of twin studies. *Addiction*, 105(3):417–430, 2010.

[344] Wolfgang Viechtbauer. Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010.

[345] Jacqueline M Vink and DI Boomsma. Gene finding strategies. *Biological Psychology*, 61(1):53–71, 2002.

[346] Jacqueline M Vink, August B Smit, Eco JC de Geus, Patrick Sullivan, Gonneke Willemsen, Jouke-Jan Hottenga, Johannes H Smit, Witte J Hoogendijk, Frans G Zitman, and Leena Peltonen. Genome-wide association study of smoking initiation and current smoking. *The American Journal of Human Genetics*, 84(3):367–379, 2009.

[347] Jacqueline M Vink, Gonneke Willemsen, and Dorret I Boomsma. Heritability of smoking initiation and nicotine dependence. *Behavior Genetics*, 35(4):397–406, 2005.

[348] Jacqueline M Vink, Liselot MC Wolters, Michael C Neale, and Dorret I Boomsma. Heritability of cannabis initiation in dutch adult twins. *Addictive behaviors*, 35(2):172–174, 2010.

[349] Peter M Visscher. A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin research and human genetics*, 9(04):490–495, 2006.

[350] Peter M Visscher, Toby Andrew, and Dale R Nyholt. Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *European Journal of Human Genetics*, 16(3):387–390, 2008.

[351] Peter M Visscher, Beben Benyamin, and Ian White. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin Research*, 7(06):670–674, 2004.

[352] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[353] Peter M Visscher, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G Hill, Jouke-Jan Hottenga, Gonneke Willemsen, Dorret I Boomsma, et al. Genome partitioning of genetic variation for height from 11,214 sibling pairs. *The American Journal of Human Genetics*, 81(5):1104–1110, 2007.

[354] Peter M Visscher, Jian Yang, and Michael E Goddard. A commentary on common snps explain a large proportion of the heritability for human heightby yang et al.(2010). *Twin Research and Human Genetics*, 13(06):517–524, 2010.

[355] Peter M Visscher, M Hossein Yazdi, Alan D Jackson, Martin Schalling, Kerstin Lindblad, Qui-Ping Yuan, David Porteous, Walter J Muir, and Douglas HR Blackwood. Genetic survival analysis of age-at-onset of bipolar disorder: evidence for anticipation or cohort effect in families. *Psychiatric genetics*, 11(3):129–137, 2001.

[356] Nora D Volkow, Ruben D Baler, Wilson M Compton, and Susan RB Weiss. Adverse health effects of marijuana use. *New England Journal of Medicine*, 370(23):2219–2227, 2014.

[357] S. I. Vrieze, M. McGue, M. B. Miller, B. M. Hicks, and W. G. Iacono. 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. *Behavior Genetics*, 43(2):97–107, 2013.

[358] Ju Wang and Ming D Li. Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation. *Neuropsychopharmacology*, 35(3):702–719, 2010.

[359] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

[360] S. T. Wilkinson and D. C. D'Souza. Problems with the medicalization of marijuana. *Jama*, 311(23):2377–8, 2014.

[361] Gonneke Willemsen, Eco JC De Geus, Meike Bartels, CEM Van Beijsterveldt, Andy I Brooks, G Frederique Estourgie-van Burk, Douglas A Fugman, Chantal Hoekstra, Jouke-Jan Hottenga, and Kees Kluft. The netherlands twin register biobank: a resource for genetic epidemiological studies. *Twin Research and Human Genetics*, 13(03):231–245, 2010.

[362] Gonneke Willemsen, Jacqueline M Vink, Abdel Abdellaoui, Anouk den Braber, Jenny HDA van Beek, Harmen HM Draisma, Jenny van Dongen, Dennis vant Ent, Lot M Geels, and Rene van Lien. The adult netherlands twin register: twenty-five years of survey and biological data collection. *Twin Research and Human Genetics*, 16(01):271–281, 2013.

[363] C. J. Willer, Y. Li, and G. R. Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–1, 2010.

[364] Rick L Williams. A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56(2):645–646, 2000.

[365] Margaret J Wright and Nicholas G Martin. Brisbane adolescent twin study: outline of study methods and research projects. *Australian Journal of Psychology*, 56(2):65–78, 2004.

[366] Hao Wu and Michael C Neale. On the likelihood ratio tests in bivariate acde models. *Psychometrika*, 78(3):441–463, 2013.

[367] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[368] Hong Xian, Jeffrey F Scherrer, Pamela AF Madden, Michael J Lyons, Ming Tsuang, William R True, and Seth A Eisen. The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit. *Nicotine & Tobacco Research*, 5(2):245–254, 2003.

[369] Minghui Xiang, Deepti Mohamalawari, and Rajini Rao. A novel isoform of the secretory pathway ca2+, mn2+-atpase, hspca2, has unusual properties and is expressed in the brain. *Journal of Biological Chemistry*, 280(12):11608–11614, 2005.

[370] ChangJiang Xu, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, and Celia MT Greenwood. Estimating genome-wide significance for whole-genome sequencing studies. *Genetic epidemiology*, 38(4):281–290, 2014.

[371] Man K Xu, Darya Gaysina, Jennifer H Barnett, Linda Scoriels, LN van de Lagemaat, Andrew Wong, Marcus Richards, Tim J Croudace, and Peter B Jones. Psychometric precision in phenotype definition is a useful step in molecular genetic investigation of psychiatric disorders. *Translational psychiatry*, 5(6):e593, 2015.

[372] Bao-Zhu Yang, Henry R Kranzler, Hongyu Zhao, Jeffrey R Gruen, Xingguang Luo, and Joel Gelernter. Association of haplotypic variants in drd2, ankk1, ttc12 and ncam1 to alcohol dependence in independent casecontrol and family samples. *Human molecular genetics*, 16(23):2844–2853, 2007.

[373] BaoZhu Yang, Henry R Kranzler, Hongyu Zhao, Jeffrey R Gruen, Xingguang Luo, and Joel Gelernter. Haplotypic variants in drd2, ankk1, ttc12, and ncam1 are associated with comorbid alcohol and drug dependence. *Alcoholism: Clinical and Experimental Research*, 32(12):2117–2127, 2008.

[374] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, and Grant W Montgomery. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[375] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[376] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. 2013.

[377] Ping Zeng, Yang Zhao, Jin Liu, Liya Liu, Liwei Zhang, Ting Wang, Shuiping Huang, and Feng Chen. Likelihood ratio tests in rare variant detection for continuous phenotypes. *Annals of Human Genetics*, 78(5):320–332, 2014.

[378] Xiaowei Zhan, David E Larson, Chaolong Wang, Daniel C Koboldt, Yuri V Sergeev, Robert S Fulton, Lucinda L Fulton, Catrina C Fronick, Kari E Branham, and Jennifer Bragg-Gresham. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nature genetics*, 45(11):1375–1379, 2013.

[379] James Q Zheng and Mu-ming Poo. Calcium signaling in neuronal motility. *Annu. Rev. Cell Dev. Biol.*, 23:375–404, 2007.

[380] Jin Zheng, Yun Li, Gonçalo R Abecasis, and Paul Scheet. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic epidemiology*, 35(2):102–110, 2011.

[381] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[382] Xiang Zhou and Matthew Stephens. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *arXiv preprint arXiv:1305.4366*, 2013.

[383] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.