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. [386]), I provided evidence that the currently typed single nucleotide polymorphisms collectively explain \( \sim 25\% \) 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. [307]) 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. [304]). This finding reinforces the idea that synaptic plasticity, memory and learning are essential to the development of addiction behaviors (Uhl, Liu et al. [335], 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. [255]). 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. [255]). As suggested by Mota et al. (Mota, Araujo-Jnr et al. [255]), 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. [319], Uhl and Draganova [334]), disease traits known to be comorbid...
with cannabis use (see e.g., [86]). 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. [138]), alcohol (Yang, Kranzler et al. [383]), heroin dependence (Nelson, Lynskey et al. [261]), and comorbid alcohol and drug dependence (Yang, Kranzler et al. [384]). 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. [308]).

Thirdly, three association signals reached genomewide significance 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 [107], 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 development of addiction behaviors (Uhl, Liu et al. [335], 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 [79]. For instance, my analysis hit genomic regions previously associated with coronary artery disease (for which smoking is a known risk factor [74]), and the HLA locus previously implicated in e.g., schizophrenia (with which smoking shares genetic vulnerabilities [86]). 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 [333] and [355], respectively). I have also identified the MIR1323-MIR312-1-MIR312-2 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 \textit{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 [257], Benowitz [30], Tobacco and Genetics Consortium [76]). 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. [69]) for which smoking is a known risk factor; however, as hypothesized by Cho et al., (Cho, Castaldi et al. [69]) 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 [85], Kelley [182], Dani [84], 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 \(\gamma\)-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 accumbens). 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 behaviors, 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. [231]), as well as axon growth and synaptic plasticity (Li and Zhang [201]). 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 [151]) (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. [312]) and lung cancer (Hung, McKay et al. [170]) 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 [128]): "[…] cigarette smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype" (Fisher [128]). While the mediation study by VanderWeele et al. (VanderWeele, Asomaning et al. [348]) 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 [282], Lee, Forey et al. [197]), leukemia (e.g., see Firen尼斯, Merriam et al. [126]), heart disease (e.g., see Huxley and Woodward [172]), chronic bronchitis and emphysema (see e.g., Forey, Thornton et al. [129]), 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]), buspirone, diazepam or propranolol (designed to treat anxiety, see Hughes, Stead et al. [169]) 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. [66]) 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 [101]). 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. [69]) multiple COPD measures were collapsed into a dichotomous affection status, and in the GWAS of alcohol dependence by Edenberg et al (Edenberg, Koller et al. [111]), 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 [243]); TATES – Trait based Association Test that uses Extended Simes procedure (Van der Sluis, Posthuma et al. [346]). Note that the MultiPhen procedure (O’Reilly, Hoggart et al. [269]) is closely related to MANOVA.
12.2. Means of Improving Statistical Power in GWAS

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. [341]] 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. [347]] 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. [347], Van der Sluis, Posthuma et al. [346]], 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 [243]), 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 [243], Van Der Sluis, Verhage et al. [347], Van der Sluis, Posthuma et al. [343], Van der Sluis, Posthuma et al. [346], Xu, Gaysina et al. [381]). 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 [243]) – 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 [243]). 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. [343]). Dichotomization would be justified only if the variable is a true dichotomy, or if dichotomization increases measurement precision (see MacCallum, Zhang et al. [227]). 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. [227]).

Multivariate techniques such as MANOVA (MultiPhen), TATES or CMV are particularly powerful whenever the GV affects some but not all observed indica-
tors (see Chapter II; see also Ferreira and Purcell [124], Medland and Neale [243], Van der Sluis, Posthuma et al. [346], Van der Sluis, Dolan et al. [344]), 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 [243]) or when the data generating process can be well described by a network model (see (Van der Sluis, Posthuma et al. [346], Van der Sluis, Dolan et al. [344]) 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 [314]). Furthermore, as demonstrated by Van der Sluis et al. (Van der Sluis, Posthuma et al. [346]) and by Medland and Neale (Medland and Neale [243]), 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 [124], Medland and Neale [243] and for the increase in parameter estimation accuracy they afford Shriner [305]), many researchers feel that simpler statistical models are quite adequate in the GWAS context for their computational easiness (Sham and Purcell [303]) as well as for interpretational reasons (Stephens [313]). However, there is solid evidence from the recent literature that the interest in addressing the computational (e.g., Zhou and Stephens [393]) as well as the interpretational issues (Stephens [313]) 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. [346], Zhou and Stephens [393]). 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 [305] 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. [286]) 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. [346]) 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 [305]) for a recent review) to analyses focused on sets of SNPs, common (e.g., Van der Sluis, Dolan et al. [344]) or rare (Maity, Sullivan et al. [230]).
12.2. Means of Improving Statistical Power in GWAS

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 [124], Stephens [313]) in this circumstance an increase in phenotypic correlations enhances the power. Hence, as suggested by Morrison ( [254]) and Cole et al. (Cole, Maxwell et al. [72]) 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. [72], 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. [180] and Lippert, Listgarten et al. [210]) 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

<table>
<thead>
<tr>
<th>Software</th>
<th>Regression model</th>
<th>Model for the background correlation matrix</th>
<th>Specification of the Cluster variable (sandwich correction?)</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plink [286] + gee [58] R package</td>
<td>Binomial Gaussian Gamma Inverse Gaussian Poisson Quasi-Linear mixed</td>
<td>( \mathbf{V}(\hat{\theta}) = \sigma^2_F \mathbf{I} ) (Independence) ( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_C, \sigma^2_E] ) (Exchangeable) ( \mathbf{V}(\hat{\theta}_f) = [\sigma^2_A, \sigma^2_C, \sigma^2_E] ) (Fixed, e.g., ACE background) ( \mathbf{V}(\hat{\theta}_f) ) (Unstructured)</td>
<td>'id' (yes)</td>
<td>gee(^1) Plink documentation(^2)</td>
</tr>
<tr>
<td>Plink</td>
<td>Gaussian Binomial Cox proportional hazards</td>
<td>( \mathbf{V}(\hat{\theta}) = \sigma^2_F \mathbf{I} ) (Independence) ( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_C, \sigma^2_E] ) (Exchangeable)</td>
<td>'family' (yes) 'cluster' (yes) 'frailty' (no)</td>
<td>Plink(^3) survival(^4) documentation(^5)</td>
</tr>
<tr>
<td>Plink + survival [329] R package</td>
<td>Linear mixed</td>
<td>( \mathbf{V}(\hat{\theta}) = \sigma^2_F \mathbf{I} ) (Independence) ( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_C, \sigma^2_E] ) (Exchangeable) ( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_C, \sigma^2_E] ) (ACE background) ( \mathbf{V}(\hat{\theta}_f) ) (Unstructured)</td>
<td>'random' (yes, see documentation(^6))</td>
<td>nlme(^7) documentation(^8)</td>
</tr>
<tr>
<td>GCTA [386]</td>
<td>Linear mixed</td>
<td>( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_E] ) (AE model)</td>
<td>Observed Genetic Relationship Matrix (no)</td>
<td>GCTA(^9)</td>
</tr>
<tr>
<td>Merlin [3]</td>
<td>Linear mixed</td>
<td>( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_E] ) (AE model)</td>
<td>Expected Genetic Relationship Matrix (no)</td>
<td>Merlin(^10) documentation(^11)</td>
</tr>
<tr>
<td>FaST-LMM [210]</td>
<td>Linear mixed</td>
<td>( \mathbf{V}(\hat{\theta}) = [\sigma^2_A, \sigma^2_E] ) (AE model)</td>
<td>Observed Genetic Relationship Matrix (no, but see [246])</td>
<td>FaST-LMM(^12)</td>
</tr>
</tbody>
</table>
12.2. Means of Improving Statistical Power in GWAS

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 [337]) 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\(^*\) 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

\(^*\)Note, these two methods are equivalent, conditional on the treatment of \(V(\hat{\theta})\)
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, Mcrea et al. [216], Li, Gui et al. [202], Li, Kwan et al. [203], Listgarten, Lippert et al. [212]) 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, Mcrea et al. [216], Li, Gui et al. [202], Li, Kwan et al. [203], Listgarten, Lippert et al. [212]) (i.e., the set of genes having the same biological function) or the whole genome (Yang, Benyamin et al. [385], Yang, Lee et al. [386]); 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. [202], Li, Kwan et al. [203]) 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. [202], Li, Kwan et al. [203]) and Type 2 Diabetes in (Li, Gui et al. [202]) for additional examples) that illustrates the
12.2. Means of Improving Statistical Power in GWAS

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 Kin-
namom, Hershberger et al. [190]). This observation is cause for concern, given
that single variant tests are not only underpowered (Li and Leal [200], Madsen
and Browning [228], Sham and Purcell [303]) 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 ex-
tensively in the literature Li and Leal [200], Madsen and Browning [228], Price,
Kryukov et al. [283], Wu, Lee et al. [377], Lee, Wu et al. [197], Chen, Meigs et
al. [64], Ionita-Laza, Lee et al. [177], Listgarten, Lippert et al. [212], Lippert,
Xiang et al. [211], Svishtova, Belonogova et al. [320]), 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 prefer-
ring 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 analy-
ses 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. [394]) 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

1https://cran.r-project.org/web/packages/gee/index.html
2http://cameliaminica.nl/scripts.php
3http://pngu.mgh.harvard.edu/~purcell/plink/
4https://cran.r-project.org/web/packages/survival/index.html
5http://cameliaminica.nl/research.php
6http://cameliaminica.nl/scripts.php
7https://cran.r-project.org/web/packages/nlme/index.html
8http://cameliaminica.nl/scripts.php
9http://cnsagenomics.com/software/gcta/mlmassoc.html
10http://csg.sph.umich.edu/abecasis/Merlin/