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 $\sim$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.