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. 

2.5 Continued from Table 2.4.

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.

2.7 Continued from Table 2.6.

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 $\beta_A$ and $\beta_E$ equal 1 and .7, respectively. The residual variances equal $\sigma^2_{\epsilon_A} = 1$ and $\sigma^2_{\epsilon_E} = .204$.

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.

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.
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 ($\rho_{bg}$). 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 sibship 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 $\chi^2$ 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 \( \chi^2_0 \) and \( \chi^2_1 \) 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 \( \chi^2_0 \) and \( \chi^2_1 \) 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.

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 (χ²(1)), the variance attributable to the 10 genetic variants (V(G)).

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

7.2 Continued from Table 7.1

7.3 Continued from Table 7.2

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

7.5 Continued from Table 7.4

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).

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 (R² < 0.1; and for SNPs with R² ≥ 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, EGCU1, EGCU2, 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.

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).
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; A1 - allele 1; A2 - allele 2; SE - standard error; N - sample size; EAF - effect allele frequency; Info - imputation quality; \( \lambda_{GEC} \) - lambda corrected standard error; \( \lambda_{GCP} \) - 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 [369]) .... 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 [369]) ......................................................... 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 [369]) ......................................................... 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 [369]) ......................................................... 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
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 $\alpha$ (grey). $\beta$ (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 b1 and b2). 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 $b_{A2,A1}$, and so also affects the indicators $y_3, y_5$. The value of the parameter $b_{A2,A1}$ was varied. Parameters are not shown to avoid clutter.

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.

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).

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.

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.

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.
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 MAPs 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: BII – 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
List of Figures

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 [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. 198