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 gem-

puteerd 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.