Development of bioinformatic tools and application of novel statistical methods in genome-wide analysis

van der Most, Peter Johannes

Samenvatting

In dit proefschrift onderzochten wij methodes om de kwaliteit van de resultaten van genoom-brede associatie analyses te verbeteren. Wij streefden naar de ontwikkeling en toepassing van nieuwe methodes om zo resultaten van hogere kwaliteit te verkrijgen. Het eerste deel van dit proefschrift beschrijft drie computerprogramma's voor de (epi-)genoom-brede associatie analyses. Het tweede deel omvat de toepassing van nieuwe computerprogramma's en statistische methodes in de analyse van genoom-brede datasets.

Hoofdstukken 2 en 3 beschreven QCGWAS en QCEWAS, twee software pakketten op basis van het R platform. QCGWAS en QCEWAS kunnen een geautomatiseerde kwaliteitscontrole uitvoeren over de resultaten van, respectievelijk, genoom- en epigenoom-brede associatie studies (Engelse afkortingen: GWAS en EWAS). Het gebruik van deze programma's biedt de volgende voordelen: ze voeren een grondige en volledige kwaliteitscontrole uit (in minder tijd dan zelfs een eenvoudige, handmatige controle zou kosten), ze zijn flexibel in gebruik en kunnen voor verschillende toepassingen aangepast worden, ze kunnen testen of de resultaten voldoende compatibel zijn voor gebruik in een meta-analyse, en ze maken kant-en-klare bestanden voor zo'n analyse aan.

Hoofdstuk 4 beschreef lodGWAS, een R software pakket voor genoom-brede associatie analyse van fenotypes zoals biomarkers met een detectie limiet (in het Engels "limit of detection", LOD). Er zijn reeds bestaande methodes voor de analyse van biomarkers met een LOD, maar deze werken meestal door waardes buiten het LOD te verwijderen of op een vaste waarde te zetten. Door het gebruik van een survival analyse, en door deze metingen als gecensureerde data te behandelen, kan lodGWAS ze accuraat modelleren, zonder het verlies van informatie dat plaatsvind in de bestaande methodes. lodGWAS is een flexibel en eenvoudig te gebruiken programma met een simpele en elegante methode voor het uitvoeren van GWAS analyse van biomarkers met een LOD.

In **hoofdstuk 5** gebruikten we genetische risico scores (GRS) en genomic restricted maximum likelihood (GREML) methodes om te schatten welke proportie van de totale erfelijkheid kan worden toegeschreven aan reeds-bekende genetische markers, en hoeveel er nog onverklaard is, voor 32 complexe ziekte-gerelateerde fenotypes in het Noord-Nederlandse Lifelines cohort. De GRS is een optelsom van het effect van alle reeds-bekende enkel-nucleotide polymorfismen voor dat fenotype. (Enkel-nucleotide polymorfisme is de naam voor een genetische variatie op één enkele letter (nucleotide) in het DNA. In het Engels heet het "single nucleotide polymorphism", afgekort tot SNP.) De GREML methode schat de zogeheten "brede common-SNP erfelijkheid", d.w.z. de proportie van de erfelijkheid die kan worden toegeschreven aan alle (bekende en onbekende) niet-zeldzame SNPs. Zoals verwacht was het merendeel (mediaan = 75%) van de reeds-bekende SNPs significant geassocieerd met hun respectievelijke fenotype. Zowel de GRS als de brede common-SNP erfelijkheidsschattingen waren significant voor alle fenotypes. De schatting van de GRS was echter gemiddeld slechts 10,7% van de schatting van brede common-SNP erfelijkheid. Dit toont aan dat de GRS van deze complexe ziekte-gerelateerde fenotypes nog niet nauwkeurig genoeg zijn om gebruikt te worden voor persoonlijke medisch-genetische voorspellingen. Voor dit onderzoek ontwikkelden wij

ook een innovatieve methode om het effect van een specifiek cohort te verwijderen uit een effectgrootte berekend in een meta-analyse die datzelfde cohort omvatte. Indien we deze effectgrootte zonder correctie in onze GRS gebruikt zouden hebben, dan zou dat tot een overschatting hebben geleid, aangezien we de effectgrootte dan valideren in (een deel van) dezelfde data waarmee het in eerste instantie geschat was. Onze methode staat ons toe om gecorrigeerde effectgroottes te gebruiken in de GRS, en deze in Lifelines te analyseren zonder dat de verklaarde variantie overschat wordt.

In **hoofdstuk 6** gebruikten we GREML om in het Estse Biobank cohort (EGCUT) en het Noord-Nederlandse Lifelines cohort te onderzoeken welke proportie van de erfelijkheid van neuroticisme toegeschreven kan worden aan niet-zeldzame SNPs. Wij keken hierbij niet alleen naar de totale neuroticisme score, maar ook naar facetscores van neuroticisme en naar residuale scores. Verder vergeleken we neuroticisme scores gerapporteerd door het individu zelf met scores gerapporteerd over het individu door een echtgenoot, verwante of bekende. Als laatste keken wij ook naar verschillen tussen de twee cohorten door middel van een gen-omgeving interactie analyse, waar het cohort als omgevingsfactor diende. In EGCUT vonden we erfelijkheidsproporties voor neuroticisme scores (zowel op domein als facet niveau) tussen 0,08 en 0,20, maar enkel het facet impulsiviteit ($h^2 = 0,20$) was statistisch significant. In Lifelines waren alle erfelijkheidsproporties (0,07-0,16) significant, met uitzondering van impulsiviteit ($h^2 = 0,03$). Erfelijkheidsschattingen voor residuale scores en neuroticisme scores gerapporteerd door echtgenoot, verwante of bekende (deze scores waren enkel beschikbaar in het EGCUT cohort) waren allen niet statistisch significant. Het gebrek aan significante bevindingen kan tenminste deels worden toegeschreven aan de beperkte grootte van de EGCUT dataset, waardoor de analyse niet voldoende onderscheidend vermogen had. Als laatste gebruikte wij een nieuwe methode in GREML om een gen-omgeving interactie analyse uit te voeren, waarmee we onderzochten of er verschillen waren tussen de twee cohorten. We vonden geen significante interacties, behalve voor impulsiviteit, waar de toevoeging van een interactie de cohort-onafhankelijke erfelijkheidsproportie tot nagenoeg nul reduceerde ($h^2 = 0,001$, n.s.).

Hoofdstuk 7 beschrijft een grootschalige meta-GWAS van nierfunctie, gekwantificeerd als de geschatte glomerulaire filtratiesnelheid (in het Engels: estimated glomerular filtration rate, eGFR), waarbij gebruik werd gemaakt van een nieuw referentie genoom, het 1000 Genomes Project, voor de genetische imputatie. De analyse bevestigde 39 eerder-gevonden loci, en identificeerde 10 nieuwe loci, waarvan er zes niet door het oudere referentie genoom gevonden zouden kunnen worden. Dit toont de voordelen van het gebruik van een gedetailleerder referentie genoom aan. Daarnaast berekenden we polygene risico scores om te bepalen of er geassocieerde loci zijn die niet in deze analyse geïdentificeerd konden worden (als gevolg van een beperkt statistisch onderscheidend vermogen voor varianten met kleine effecten). De polygene risico score werd uitgevoerd in de TRAILS en NESDA cohorten en verklaarde een maximum van 2,2% variantie in de eGFR.

In **hoofdstuk 8** combineerden wij een traditionele survival analyse met GWAS, en voerden wij een meta-analyse over zulke GWAS uit. Wij zochten naar genetische markers geassocieerd met de leeftijd waarop men begint met het gebruik van cannabis. Een enkele locus met vijf significante SNPs werd gevonden in het Calcium-transporting ATPase (*ATP2C2*) gen. Een gene-based associatie test identificeerde dit gen

ook, plus nog twee anderen (*ECT2L* en *RAD51B*). Helaas kon de meest significante SNP in *ATP2C2* niet gerepliceerd worden in ons kleine replicatie sample (n = 4.478), maar het was nog steeds significant in de gecombineerde analyse. Een SNP-gebaseerde erfelijkheidstest en een polygene risico score vonden geen significante resultaten.

Hoofdstuk 9 is een algemene discussie van de nieuwe programma's en methodes beschreven in dit proefschrift. Deze methodes maken het mogelijk om een kwalitatief betere genoom-brede analyse uit te voeren, in plaats van het simpelweg vergroten van de datasets.

We verwachten dat er in de toekomst twee grote uitdagingen zijn voor de Genetische Epidemiologie. De eerste is om de overgebleven ontbrekende erfelijkheid te vinden. In afwachting van de beschikbaarheid van volledige exoom of genoom-sequencing datasets zal dit waarschijnlijk plaatsvinden middels de huidige array-gebaseerde GWAS methodes, met grotere sample sizes en geïmputeerd tegen gedetailleerdere genoom referentie datasets. De tweede uitdaging is hoe we de enorme hoeveelheid aan reeds-beschikbare omics data effectief gebruiken om de mechanismen van reeds-geïdentificeerde genetische effecten te bepalen. Helaas vormt de grootte van genoom-brede datasets een probleem op zich, zowel in het verzamelen als in het analyseren van de data. Er zijn echter ook hier oplossingen, zoals bijvoorbeeld het gebruik van de uitkomsten van (epi-) genoom-brede analyses (de zogeheten "summary statistics") als proxy voor individuele genetische data. Dit is ook weer een goed voorbeeld van de toepassing van geavanceerdere methodes, in plaats van het gebruik van grotere datasets, om genetisch-epidemiologisch onderzoek te doen.