University of Groningen

The value of haplotypes

de Vries, Anne René

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
de Vries, A. R. (2009). *The value of haplotypes*. [s.n.].

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 12-10-2022

NEDERLANDSE SAMENVATTING

NEDERLANDSE SAMENVATTING

In het kort

Chronische ziekten zoals astma, reuma en bepaalde soorten kanker hebben deels een genetische achtergrond. Dit betekent dat de kans op de aandoening is verhoogd door bepaalde varianten in het DNA. Het opsporen van deze varianten kan sinds enkele jaren gedaan worden door het volledige genoom in één keer te scannen op verschillen tussen patiënten en gezonde mensen. Dit scannen van het genoom is echter niet volledig: slechts 1 op de circa 6000 meest informatieve posities in het DNA wordt geanalyseerd. De op te sporen genetische variatie kan ook in de overige posities aanwezig zijn en daarom gemist worden omdat informatie daarover onvolledig of ontbrekend is.

Wij hebben haplotype sharing statistieken ontwikkeld om de ontbrekende informatie beter te kunnen bekijken. Deze methoden maken gebruik van prehistorische familiestructuren die nog altijd in het DNA terug te vinden zijn. Dit is een andere benadering dan conventionele methoden. De haplotype sharing methoden kunnen gecombineerd toegepast worden met bestaande methoden om de kans te verhogen om genetische varianten te vinden die geassocieerd zijn met ziekte. Onze haplotype sharing statistieken hebben we toegepast op data van reumatische artritis, de ziekte van Hirschsprung en coeliakie.

Haplotypen kunnen daarnaast waardevol zijn in andere gebieden van genetisch onderzoek, met name in populatiegenetica. Zo hebben we laten zien dat de historisch oorsponkelijke versie van een genetische variant goed geschat kan worden door naar haplotypen te kijken.

1 Inleiding

Complexe ziekten zoals astma, reuma en bepaalde soorten kanker hebben deels een genetische achtergrond, wat betekent dat de kans op de aandoening is verhoogd door bepaalde genetische varianten in het DNA. Deze ziekten worden complex genoemd omdat meerdere genen en omgevingsfactoren betrokken zijn bij het ontstaan van de ziekte. Het is dus niet één enkel gen dat een ziekte bepaalt (zoals bijvoorbeeld bij sikkelcelanemie of taaislijmziekte), maar het is een set van tien, twintig of misschien wel meer genen die allemaal een bijdrage leveren aan het totale ziekterisico.

Het kennen van de belangrijkste genetische factoren voor een ziekte levert informatie op over het mechanisme voor het ontstaan van de ziekte en kan van belang zijn voor het ontwikkelen van nieuwe behandelmethoden. Tot op heden zijn er reeds diverse genen gekoppeld aan bepaalde complexe ziekten, waaronder de ziekte van Crohn, borstkanker en diabetes type I

en II. Echter, slechts een bescheiden deel van de ziektefrequentie wordt nog maar verklaard door deze genen, wat waarschijnlijk betekent dat er nog veel genen op te sporen zijn. Veel onderzoek daarnaar richt zich op het uitvoeren van steeds grotere studies, met steeds meer proefpersonen. Dit is zeer kostenintensief en heeft bovendien de laatste paar jaar, ondanks het, met wisselende zekerheid, vinden van allerlei genen, nauwelijks geleid tot een veel betere verklaring van de totale ziektefrequentie.

De meest gebruikte methode om genen op te sporen is ook de eenvoudigste: een simpele vergelijking van de allelfrequentie tussen patiënten en controles. Een allel is een variant van een gen of een nucleotide (in het laatste geval ook wel single nucleotide polymorphism (SNP) genoemd). Een waardevolle toevoeging aan de single-marker associatie-methoden zijn methoden die gebaseerd zijn op haplotype-informatie. Haplotypen zijn opeenvolgende genetische markers (of SNPs) op een chromosoom. Ieder mens heeft 23 chromosomen van de moeder en van de vader. We hebben elk chromosoom dus twee keer (met diverse verschillen) en bij elkaar vormt dat ons genotype. Bij het overerven zijn wel een paar kleine veranderingen opgetreden, te weten mutaties en recombinaties. Een mutatie kan de simpele verandering van een nucleotide in een andere zijn. Een recombinatie is de uitwisseling van gedeelten van twee homologe chromosomen (zie figuur 2 in hoofdstuk 1).

Haplotypen weerspiegelen prehistorische familiestruc-turen vanwege de continue overerving van DNA, soms verknipt door recombinaties. Bijvoorbeeld, wanneer het DNA van twee neven met elkaar wordt vergeleken, zal veel overeenkomst worden gevonden vanwege het behouden blijven van lange stukken DNA. Oftewel, deze twee personen delen lange haplotypen (haplotypesharing is groot). Dit is geïllustreerd in de figuur op de voorzijde van dit proefschrift.

Ook de kortere haplotypen zijn informatief. Stel dat er op een gegeven moment in het verleden een mutatie is opgetreden (zie figuur 4 van hoofdstuk 1). Dan is dit op een bepaald haplotype gebeurd. Wanneer de mutatie geen directe hinder voor het individu geeft of zelfs een selectievoordeel geeft, kan het worden doorgegeven aan het nageslacht. Het haplotype waarop de mutatie is ontstaan wordt steeds doorgegeven, maar wordt steeds korter naar mate er vaker recombinaties optreden. Dit is zo met het toenemen van het aantal generaties.

Als de mutatie met de ziekte te maken heeft, zal dus niet alleen de allelfrequentie van de ziekte-SNP geassocieerd zijn met de ziektestatus, maar ook het haplotype waar de SNP op ligt. Kortom, als we gaan kijken of er binnen een patiëntengroep haplotypen zijn die hetzelfde zijn, meer dan wat je gemiddeld zou verwachten, dan kunnen we gebieden identificeren die met de ziekte te maken hebben.

Zoals reeds gezegd kunnen we ook gewoon kijken naar het verschil tussen patiënten en controles in allelfrequentie van een ziekte-SNP, maar het kunnen vinden van een ziekte-geassocieerde SNP kan alleen positief zijn als die SNP ook in de dataset met gemeten SNPs aanwezig is. Als dat niet zo is (en dat is waarschijnlijk vaak het geval), kan de correlatie

(linkage disequilibrium, LD) tussen de gemeten SNPs en de niet-gemeten ziekte-SNP toch nog zorgen voor voldoende informatie. Naburige SNPs zijn vaak gecorreleerd, wat betekent dat een resultaat op een gemeten SNP iets zegt over een aantal omliggende (niet-gemeten) SNPs. Om optimaal informatie uit een studie te kunnen krijgen dienen de SNPs goed gekozen te worden, zodat een zo hoog mogelijke dekking van alle SNPs wordt bereikt. Een dergelijk goed gekozen SNP wordt ook wel tagging SNP genoemt.

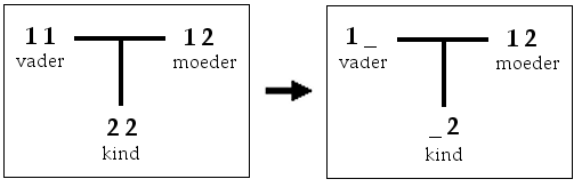
Helaas geven de tagging SNPs niet altijd voldoende informatie over de ziekte-SNP. In dat geval wordt de ziekte-SNP gemist. Echter, als je haplotypen zou analyseren, zou je vaak nog wel iets kunnen zien, mits de mutatie niet te oud is en de haplotypen niet al te kort zijn geworden.

In de hoofdstuken 2 tot en met 7 komen verschillende aspecten naar voren die te maken hebben met data van zogenaamde whole-genome studies. In dergelijke studies is informatie verkregen van patiënten en controles op minimaal 300.000 SNPs over het hele genoom. In de meeste hoofstukken zijn haplotypen gebruikt bij de analyse van dergelijke data.

2 Haplotype sharing analyse van nulallelen op 7,650 loci in a 474 trio genoombrede analyse laat een recente oorsprong zien van mutaties in de primer site

De primair gemeten genetische informatie (met behulp van Illumina apparatuur) bestaat per SNP uit een rood signaal voor het ene allel en een groen signaal voor het andere allel. Wanneer beide signalen evenveel aanwezig zijn, is de SNP dus heterozygoot. Wij hebben methoden ontwikkeld om in twijfelgevallen (als de observatie ligt tussen homozygoot en heterozygoot) een betere scheiding te verkrijgen en het genotype beter te bepalen door te corrigeren voor samplevariatie in het signaal. De call rate ging daardoor omhoog tot bijna 100%.

Enkele van onze whole genome studies zijn gebaseerd op genetische informatie van vader-moeder-kind trios. Trio data geeft informatie over de overerving van allelen. Het bleek uit onze data dat de overerving niet altijd klopt: we vinden onverwacht veel Mendeliaanse overervingsfouten. Eén van de ouders blijkt op een SNP bijvoorbeeld homozygoot 1 te zijn, terwijl het kind homozygoot 2 is (zie figuur 1). Dit kan een genotyperingsfout zijn, maar onze analyse gaf aan dat het waarschijnlijker is dat er een nul-allel is overgeërfd. (Waarschijnlijk gaat het in de meeste gevallen om een slechtere pimeraffiniteit, wat betekent dat er zeer nabij nog een extra SNP aanwezig is.) Ter bevestiging van



Figuur 1: *Deleties kunnen worden geïdentificeerd door een juiste interpretatie van overervingsfouten*

deze conclusie bleek dat de aanwezigheid van een nul-allel geassocieerd is met een lagere genotyperingsintensiteit (groen/rood signaal).

De gevallen waar een overervingsfout gevonden werd, zijn niet de enige gevallen waar een nul-allel aanwezig is. Er zijn vermoedelijk veel meer trios met dezelfde nul-allel, welke niet detecteerbaar zijn als een overervingsfout. Door gebruik te maken van haplotypesharing van het haplotype waarop de deletie ligt, konden meer trios geïdentifieerd worden met dezelfde deletie. Het gebruiken van deze informatie en het toekennen van het nul-allel verhoogt de juistheid van het genotyperingsalgoritme (bij trios).

3 Bepaling van het oorspronkelijke allel door gebruik te maken van haplotypesharing, gevalideerd door de 4-gamete regel

Haplotype-informatie kan voor meerdere doeleinden gebruikt worden. Niet alleen voor het identificeren van een ziektelocus, maar ook voor het bepalen van het oudste allel van een SNP. Zoals in de inleiding is weergegeven, is een overeenkomstig haplotype langer naarmate de verwantschap groter is. In andere woorden, de tijd tot het punt waar de twee takken van de stamboom samenkomen is korter. Dit heet de coalescentietijd. Een SNP kent ook een coalescentietijd, wat vaak (behalve bij verstoringen door genetische drift) overeen komt met de tijd tot het ontstaan van de SNP, oftewel het moment van mutatie.

Wij laten zien dat we in veel gevallen het oorspronkelijke allel kunnen bepalen aan de hand van de gemiddelde lengte van haplotypesharing nabij de SNP. Wanneer de haplotypesharing rondom haplotypes met allel 2 op een bepaalde SNP groter is dan de sharing voor haplotypes met allel 1, dan is allel 2 waarschijnlijk het jongere allel en allel 1 het oorspronkelijke allel. Immers, als allel 2 jonger is, is er minder tijd geweest om het omringende haplotype korter te laten worden door recombinaties, terwijl het oorspronkelijke allel 1 al veel langer bestaan heeft en het omringende haplotype al veel langer aan recombinaties onderhevig was. Je zou ook kunnen zeggen dat er op het moment van de mutatie verschillende haplotypen (en mensen) bestonden, maar dat de mutatie maar één van hen heeft getroffen (zie figuur 5 van hoofdstuk 1). Als daarna de overeenkomstige haplotypen korter worden vanwege recombinaties, zullen vanwege het kleinere aantal in het begin, de haplotypen waarop de mutatie zit meer op elkaar lijken dan de andere haplotypen. Zelfs wanneer het nieuwe allel sterk in frequentie zou toenemen.

Wij bepaalden het allel met de kortste sharing van het omringende haplotype en testten aan de hand van de chimpansee sequentie of dit inderdaad het oudste allel was. Immers, het chimpansee allel kunnen we beschouwen als het oorspronkelijke allel omdat het zeer waarschijnlijk sinds de splitsing van de mens onveranderd is gebleven in de chimpansee, terwijl het menselijke DNA sindsdien de mutatie heeft ondergaan. Wij vonden dat we in 80% van de gevallen een juiste inschatting van het oudste allel maakten, gebruik makend van

data van 120 haplotypen van Kaukasische afkomst uit de HapMap studie.

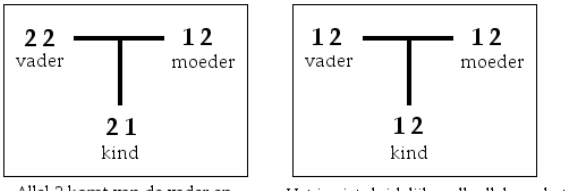
Onze bevinding dat er langere haplotypesharing is rondom jongere allelen konden we valideren aan de hand van de zogenaamde 4-gamete regel. Deze regel zegt dat van de vier mogelijke 2-SNP haplotypes er altijd drie door mutaties gevormd kunnen worden, terwijl de vierde haplotype slechts door een recombinatie kan ontstaan. Met name bij jonge mutaties is het daarom de verwachting dat het vierde haplotype vaak afwezig is vanwege de lage blootstelling aan recombinaties. Het ontbreken van het vierde haplotype bleek inderdaad geassocieerd te zijn met de leeftijdsschatting zoals we die voor SNPs maakten op basis van haplotypesharing.

Onze methode voor het bepalen van het oorsponkelijke allel kan van nut zijn voor andere populaties dan de humane en kan ook gebruikt worden om SNPs op leeftijd te sorteren.

4 Faseren van genotypedata met behulp van de 4-gamete regel

Haplotype-informatie kan zeer informatief zijn voor het opsporen van ziekteloci. Echter, alleen genotypen (beide homologe chromosomen tegelijk) worden experimenteel bepaald. Om de haplotypen te bepalen, dient specifieke software gebruikt te worden. Dit proces heet faseren. Het gaat er in feite om, om de haplotypen die ontvangen zijn van de vader en de moeder te onderscheiden. Voor heterozygote SNPs van losse (case-control) genotypen is het onduidelijk welk allel tot welk haplotype behoort. Bij vader-moeder-kind trios is meer informatie aanwezig zodat meer fasen (van wie welk allel afkomstig is) bepaald kunnen worden (zie figuur 2). Maar wanneer alle drie personen heterozygoot zijn, is ook daar de fase onbekend.

Wij hebben een algoritme ontwikkeld voor het faseren van genotypedata, gebaseerd op de 4-gamete regel. Volgens deze regel kan één van de vier mogelijke 2-SNP haplotypen alleen maar gevormd worden via recombinatie. In bepaalde gevallen is de vierde combinatie afwezig en dit is informatie die we kunnen gebruiken om te faseren. Wanneer bij een bepaald 2-SNP combinatie, voordat de data gefaseerd is, slechts drie verschillende haplotypen in de dataset aanwezig zijn, moet ook na het faseren het vierde haplotype nog steeds afwezig zijn. Dit is een restrictie die wordt opgelegd bij het invullen van de fasen. Informatie van alle 2-SNP combinaties wordt gebruikt om continu en iteratief een grote set restricties te vormen voor het invullen de fasen.



Figuur 2: *Faseren van een SNP van een trio.*

De resultaten van ons programma (TrioHap) hebben we vergeleken met

reeds bestaande software. Ons programma is één van de weinige die naast losse genotypen ook trio data kan faseren. De correctheid van het faseren van case-control genotypen werd berekend door uit trio data de genotypen van de ouders te faseren en de resulterende haplotypen te controleren aan de hand van het genotypen van het kind. Helaas bleek ons programma iets meer fouten te maken dan andere reeds bestaande programma's. In hoeverre dit voor het faseren van trio data ook geldt is onduidelijk omdat het foutpercentage daarvan niet is te berekenen.

5 Genoombrede associatieanalyse met behulp van haplotype-sharinglengte gebaseerde methoden

Wij hebben twee genetische associatietesten ontwikkeld op basis van de lengte van haplotypesharing. De berekening van de sharinglengte tussen twee haplotypen gaat zoals uitgelegd in figuur 3 van hoofdstuk 8. De sharinglengte is simpelweg de telling van het aantal aansluitende SNPs dat hetzelfde allel heeft.

De eerste associatietest is de Haplotype Sharing Statistiek (HSS), welke de gemiddelde haplotypesharing van patiënten (pat-pat) vergelijkt met die van controles (ctr-ctr). In het geval er een niet al te oude mutatie geassocieerd is met de ziekte, dan zal de haplotypesharing rondom de mutatie waarschijnlijk groter zijn dan voor haplotypen zonder de mutatie. De hypothese is daarom dat de gemiddelde haplotypesharing bij patiënten langer zal zijn dan bij controles in het gebied nabij de mutatie.

De tweede test die wij ontwikkeld hebben is de CROSS test, welke steeds één patiënthaplotype vergelijkt met één controle (pat-ctr) en van alle pat-ctr combinaties de gemiddelde haplotypesharing bepaalt. De gemiddelde cross-sharing wordt vervolgens statistisch vergeleken met de gemiddelde sharing over alle mogelijk haplotypeparen (pat-pat + ctr-ctr + pat-ctr). De CROSS test geeft dus aan wanneer haplotypen met het ziekte-allel minder op "gezonde" haplotypen lijken dan gemiddeld.

Wij hebben beide methoden toegepast op een echte dataset van 2300 SNPs op bijna 10 kB van chromosoom 18q. De dataset bevatte genotypedata van 460 controles en 460 patiënten met reumatische artritis. We konden met redelijke zekerheid een locus op 20 kB nauwkeurig identificeren, waarbij de CROSS het sterkste resultaat gaf, gevolgd door de HSS en tenslotte de veel gebruikte single marker allelische χ^2-test. Wanneer alleen de laatste test gebruikt zou worden, zou het locus niet zijn gevonden omdat de score niet boven de ruis uitkwam. Dit in tegenstelling tot beide haplotypesharing methoden.

6 Hoe stabiel zijn extreme P waarden: een empirische illustratie

Genoombrede associatiestudies focussen meestal op de kleinste P waarden die gevonden worden voor de >300.000 single marker allelische χ^2-testen. Een set van 100 à 1000 SNPs met de kleinste P waarden is vaak onderwerp van vervolgstudie. In dit hoofdstuk wordt aangetoond dat er veel variatie is in de P waarden van de primaire studie, afhankelijk van welke samples er in de controlegroep aanwezig zijn. Hetzelfde geldt overigens voor de patiëntengroep.

Uit een dataset van 1417 controles, trokken we steeds willekeurig 771 andere controles als de controlegroep en testten we deze met 771 coeliakie-patiënten. Voor drie SNPs die uit een eerdere studie als met coeliakie geassocieerd naar voren waren gekomen, berekenden we steeds de P waarde van de χ^2-test. De resultaten van 10.000 trekkingen staan weergegeven in figuur 1 van hoofdstuk 6. Er is een brede distributie aan P waarden voor alle drie SNPs.

Onze resultaten geven aan dat extreme P waarden lage reproduceerbaarheid kunnen hebben voor typisch middelgrote controlegroepen (en patiëntgroepen). Om de betrouwbaarheid van P waarden te vergroten, zijn grotere samplegroepen nodig, maar grotere studies ondervinden potentieel powerverlies vanwege genetische heterogeniteit.

7 Haplotypesharing-testen verhogen de power om ziekteloci te detecteren in genoombrede associatiestudies

Zoals in de inleiding is beschreven, kan associatie gemist worden wanneer de ziekte-SNP niet in de studie zit en de correlatie (LD) tussen de gemeten SNPs en de ziekte-SNP laag is. Het is onze hypothese dat een haplotypesharing statistiek gevoeliger is voor dergelijke situaties in vergelijking met single marker associatietesten.

In deze studie analyseerden we, gebruik makend van uit echte genotyperingsdata vervaardigde simulatiedata, het verband tussen de genoemde correlatie tussen SNPs en de resultaten die een test geeft. We vergeleken de CROSS test met de single marker χ^2-test. Met name wanneer de correlatie laag is, de CROSS test nog goed het locus kan vinden terwijl de single marker χ^2-test dat niet kan.

Daarnaast toont de studie aan dat het gecombineerd testen van beide statistieken over het geheel winst aan power oplevert ten opzichte van het gebruiken van alleen de χ^2-test. Haplotypesharing verhogen dus de power om associate te detecteren en zouden in huidige genoombrede associatiestudies altijd toegepast moeten worden in combinatie met de single marker χ^2-test.