From correlation to causation: Data-driven exploration of transcriptional regulation using population genomics

Luijk, R.

NEDERLANDSE SAMENVATTING

Moleculaire epidemiologie en transcriptionele regulatie

Epidemiologie is een wetenschappelijke discipline die zich bezig houdt met het bestuderen van veelvoorkomende ziekten en de oorzaken die daaraan ten grondslag liggen. Moleculaire epidemiologie is een deeldiscipline die specifiek ingaat op de moleculaire veranderingen die een rol spelen bij de verschillende aspecten omtrent ziekten. De vergaarde moleculaire informatie betreft vaak verscheidene aspecten van het genoom die een rol spelen bij transcriptionele regulatie. Dit proces bepaalt de mate waarin een gen geactiveerd wordt, wanneer dit gebeurt, en in welk weefsel.

Voor verschillende ziekten is al ontdekt dat misregulatie van genen gecorreleerd is met de ontwikkeling van deze ziekten. Deze correlaties zijn gevonden in grote genoombrede associatiestudies, die de verbanden tussen miljoenen genetische varianten en een ziekte onderzoeken, en zo genen vinden die mogelijk bij de ontwikkeling van deze ziekte betrokken zijn. Het gaat hierbij veelal om complexe ziekten die ontstaan door een samenspel van verschillende genen. Mede doordat deze ziekten niet eenduidige oorzaken hebben, geven de verbanden tussen genetische varianten en de aanwezigheid of afwezigheid van een ziekte nog niet direct een volledig beeld van welke genen er precies bij betrokken zijn, of hoe deze zich tot elkaar verhouden.

Andere informatie omtrent transcriptionele regulatie kan hierbij helpen, zoals het transcriptoom (expressieniveaus van verschillende genen), en het epigenoom. Epigenetische modificaties zijn moleculaire dimmers op het DNA. Door het veranderen van de toegangelijkheid van het DNA zelf, beïnvloeden ze de mate waarin genen afgeschreven kunnen worden, en dus ook hun expressieniveaus. Een belangrijke epigenetische dimmer die in dit proefschrift bestudeerd wordt, is DNA methylatie, waarbij een bepaald molecuul (een methylgroep) op het DNA geplaatst wordt. Hoewel de relatie tussen methylatieniveaus en expressieniveaus complex is, kan het over algemeen worden gezegd dat hogere methylatieniveaus leidt tot lagere activiteit van het gen. Het relateren van veranderingen in expressie-, en methylatieniveaus aan elkaar en aan

genetische markers van ziekte, kunnen een beter beeld geven welke netwerken van genen betrokken zijn bij de ontwikkeling van ziekten.

Deze aanpak is echter ook niet vrij van beperkingen. Een verband tussen een ziekte en de expressieniveaus van een gen betekent niet direct dat er ook een oorzakelijk verband is. In dit proefschrift proberen we deze oorzakelijk verbanden wel te leggen, om zo beter te begrijpen hoe transcriptionele regulatie werkt, en mogelijk tot ziekte kan leiden.

Van correlatie tot causatie

Oorzakelijke verbanden in de regulatie van genen onderling is helaas erg lastig aan te tonen. Experimentele manipulatie van het expressieniveau van genen is vaak noodzakelijk om te bewijzen dat er een causaal verband bestaat. Dit wordt vaak gebruikt om eerdere correlaties die bij genoombrede associatiestudies zijn gevonden te bevestigen in een laboratorium. Het is echter niet altijd haalbaar om alle mogelijke aanwijzingen uit deze studies op deze wijze te onderzoeken. Soms is het aantal te onderzoeken aanwijzingen simpelweg te groot, is het ethisch onverantwoord om deze experimenten in mensen uit te voeren, of heeft een experimentele manipulatie onverwachte neveneffecten.

Gelukkig is het mogelijk om, gebruik makende van observationele data, tot op zekere hoogte uitspraken te doen over causaliteit, en de noodzakelijke, tijdrovende experimenten beter te prioriteren. Quantitative trait loci (QTL) mapping kan hierbij een eerste stap in de goede richting zijn. Hierbij worden verbanden tussen genetische varianten, verspreid over het gehele genoom (genoombreed) en de methylatieniveaus van CpG-dinucleotiden gezocht. CpG-dinucleotiden zijn locaties op het genoom waarbij de twee basen (of “letters”) C en G elkaar opvolgen. Een vaak onderzochte vorm van DNA methylatie komt voor op deze locaties. Dergelijke verbanden helpen om gevonden associaties tussen specifieke genetische varianten en ziekten verder te onderzoeken.

Een volgende stap is het leggen van causale verbanden in de regulatie van verschillende genen onderling (gen-gen interacties). Dit kan met het gebruik van zogeheten Mendelian Randomization technieken, waardoor we met redelijke zekerheid oorzakelijke verbanden tussen de expressieniveaus van genen in mensen kunnen aanwijzen, zonder hierbij experimenten uit te voeren. Dit soort analyses vormen een belangrijke extra stap in de interpretatie van de resultaten van genoombrede associatiestudies.

In dit proefschrift pogen we middels deze technieken een eerste stap te nemen richting het vinden van bewijs voor causaliteit met betrekking tot transcriptionele regulatie. Het uiteindelijke doel is om hierdoor verder te komen dan het simpelweg duiden van correlaties, en in plaats daarvan causale hypothesen te kunnen opstellen over transcriptionele netwerken.

We beginnen in **hoofdstuk 2** met een methodologische bijdrage aan het zoeken naar verbanden tussen genoombrede patronen van genetische varianten en CpG-dinucleotiden die bij elkaar in de buurt op het genoom liggen, methylatie-QTLs

genoemd, waarbij we gebruik maken van een kleine openbare dataset. We laten zien dat een veel gebruikte multiple testing-strategie een hoog aantal CpG-dinucleotiden foutief aanduidt als beïnvloed door lokale (*cis*) genetische variatie, en ontwikkelen een nieuwe methode die dit voorkomt. Verder concluderen we dat *cis*-meQTLs nog lokaler zijn dan voorheen gedacht, en voor het vinden lokale effecten doorgaans niet verder gekeken hoeft te worden dan 50kb.

In **hoofdstuk 3** maken we gebruik van een grotere dataset van 3.841 Nederlandse individuen om de effecten van 6.111 geselecteerde genetische varianten te relateren aan alle verder weggelegen CpG-dinucleotiden (*trans*). De varianten werden geselecteerd omdat zij in genoombrede associatestudies (GWAS) geassocieerd bleken met één of meerdere ziekten of andere complexe fenotypen. Van de 6.111 varianten kunnen we er 1.907 relateren aan meerdere verder weg gelegen CpG-dinucleotiden. Veel van deze varianten bleken daarbij ook de expressie van nabijgelegen transcriptiefactoren te beinvloeden en de CpG-dinucleotiden die in een bindingssite van de desbetreffende transcriptiefactor liggen. Dit leidt tot onze eerste hypothese: genetische varianten brengen veranderingen teweeg in de expressieniveaus van verder weggelegen genen door de expressieniveaus van nabijgelegen transcriptiefactoren te beïnvloeden. Tot slot stellen we dat een derde van alle onderzochte CpG-dinucleotiden onder invloed staat van nabijgelegen (*cis*) genetische varianten, veel meer dan aanvankelijk gedacht.

In **hoofdstuk 4** onderzoeken we welke diverse rollen epigenetische regulatie speelt bij X-chromosomale inactivatie (XCI), het proces waarbij één van de twee X-chromosomen in vrouwelijke zoogdieren wordt "uitgezet". We veronderstellen dat genetische varianten die alleen in vrouwen de X-chromosomale methylatie beïnvloeden betrokken moeten zijn bij DNA methylatie en XCI. Een drietal van zulke varianten worden geïdentificeerd en gerepliceerd, en zijn dus vermoedelijk betrokken bij XCI. De aangedane CpG-dinucleotiden liggen voornamelijk in de buurt van X-chromosomale genen die veelal ontsnappen aan XCI, wat suggereert dat er een genetische basis is voor dit verschijnsel. Vervolgens onderzoeken we de effecten van de genetische varianten op de expressieniveaus van nabijgelegen genen en wijzen verschillende genen toe aan iedere variant die derhalve gedacht worden verantwoordelijk te zijn voor de veranderingen op het X-chromosoom. Twee van de drie aangewezen genen zijn nog niet eerder geïmpliceerd in XCI en kunnen dus nieuwe aanwijzingen zijn voor het reguleren van XCI via DNA methylatie of aanverwante epigenomische veranderingen.

Tot slot proberen we in **hoofdstuk 5** uitspraken te doen over welke genen veranderingen teweegbrengen in de expressieniveaus van andere genen. Hiermee gaan we met het gebruik van een gemodificeerde Mendelian Randomization-analyse voorbij aan de QTL mapping, hoewel we nog steeds genetische variatie als causaal anker gebruiken om deze hypotheses te kunnen opstellen. Hierbij proberen we de correlaties tussen de verschillende genetische instrumenten, alsook pleiotropische effecten tegen te gaan om zo één gen aan te kunnen wijzen als causale driver. Net als in **hoofdstuk 3** blijkt ook hier dat transcriptiefactoren vaker dan verwacht verantwoordelijk lijken voor veranderde expressieniveaus van andere *in cis* en *trans* gelegen genen. De resulterende catalogus van

gen-geninteracties leverden reeds nieuwe biologische inzichten op en zouden daarnaast de basis kunnen vormen voor vervolgonderzoek omtrent de causale drivers.

Conclusie

Vele ziekten worden veroorzaakt door een verstoring van transcriptionele regulatie. Bij het ontstaan van veelvoorkomende, complexe aandoeningen zijn vaak grote aantallen genen betrokken. Het is van groot belang om de transcriptionele regulatie tussen genen onderling te onderzoeken en met name waar het genen betreft waarvan een relatie met ziekte al is aangetoond. Door verschillende beperkingen is het echter lastig om causale relaties te leggen tussen bijvoorbeeld de expressieniveaus van verschillende genen. Tezamen zijn de verschillende hoofdstukken in dit proefschrift voorbeelden van hoe genetische variatie gebruikt kan worden om met observationele data toch uitspraken te kunnen doen over deze oorzakelijke verbanden. De resultaten uit dit proefschrift helpen hierbij door een beter begrip van transcriptionele (dis)regulatie te geven, terwijl de gebruikte methoden in het algemeen gebruikt kunnen worden om ook met andere type databronnen soortgelijke analyses uit te voeren.