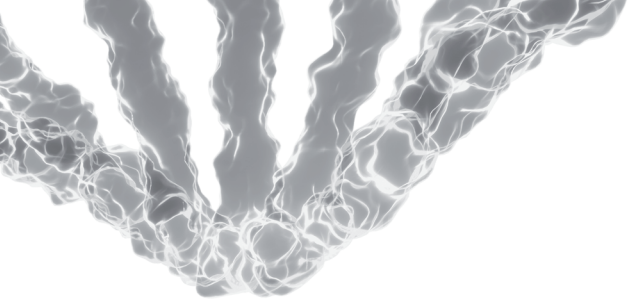
Towards solving the missing heritability in pharmacogenomics

Lee, M. van der



CHAPTER 9

Summary
Nederlandse samenvatting

Summary

Pharmacogenomics (PGx) is widely recognised as an important pillar of personalized medicine. By adjusting drug therapy based on genetic markers, adverse drug reactions might be prevented and efficacy can be optimized. However, with currently used clinical PGx tests we can only explain a part of the genetic variability of drug response, leading to substantial, so-called, missing heritability. Potential explanations for this missing heritability are; the contribution of rare variants (minor allele frequency (MAF) <1%), the inability to resolve complex PGx regions due to repetitive sections and pseudogene interference, haplotype phasing, the strict subtype characterization of PGx phenotypes instead of applying a continuous scale and the occurrence of substrate specificity.

Clinical PGx predominantly focusses on the use of single nucleotide variant (SNV) panels and copy number variants (CNVs) in combination with a categorical model to predict phenotypes. However, many more technologies to determine and interpret one's genetic make-up have become available. To resolve the missing heritability in PGx, these technologies provide an opportunity to improve the characterization and interpretation of PGx variants. Which can ultimately result in an increase in explained variability in drug response and better personalized therapy.

To resolve the missing heritability in PGx, it is of importance to accurately characterize the full genetic make-up of genes involved in drug response. For this, a wide range of technologies is available. In **chapter 2**, we have summarized these technologies and discussed their potential benefits and applications for PGx as well as their limitations. While most SNV panels used for clinical PGx offer rapid turnaround times at relatively low costs, they are limited in the amount of variants and type of variants they can detect. In recent years, advanced SNV panels containing millions of variants, including enrichment for variants in well-known PGx genes, have become available. These panels contain many more genes and variants than the routine diagnostic panels. However, they remain limited to the variants included on the array. Sequencing, on the other hand, is capable to detect all variants in the selected locus. Moreover, long-read sequencing is capable of resolving complex regions and can phase reads to their allele of origin. Ultimately, selecting the right technology is not a matter of fact but a matter of choosing the right technique for the right problem.

The potential of using existing clinical sequencing data for PGx was investigated in **chapter 3**. For 1,583 individuals we investigated if it was possible to extract a PGx profile from whole exome sequencing (WES) data generated for diagnostic purposes. All individuals were part of child-parent trios, allowing for pedigree based haplotyping phasing. The panel consisted of 42 variants in 11 genes with actionable recommendations in the DPWG

(Dutch pharmacogenetics working group) guidelines. Unfortunately, we were not able to infer copy number variants (CNV) based on WES data leading to an inability to assign the CYP2D6 phenotypes. Moreover, for two intronic variants (*CYP3A5**3 and *CYP2C19**17) the coverage was often insufficient which resulted in no calls for the CYP3A5 and CYP2C19 phenotypes in 26% and 99% of the individuals respectively. Overall, for 8 out of the 11 genes a PGx profile could be successfully generated. Of all patients, 85% had a genotypic profile which is related to at least one actionable phenotype for which a dose recommendation was available in the DPWG guidelines. We also showed that pedigree based haplotype phasing improved phenotype predictions.

In **chapter 4**, we investigate the ability of long-read sequencing to resolve complex PGx regions by using publicly available whole genome long-read sequencing data of a single individual from the genome in a bottle (GIAB) consortium. For these GIAB samples high quality DNA sequencing data and ground truth variant benchmarking sets are available. With short-read sequencing, the reads are not always of sufficient length to span an entire complex region, leading to an inability to characterize this region. Moreover, with shorter reads it is more difficult to determine if a read originates from an pharmacogene or from a highly similar pseudogene.

For this study, data regarding 100 pharmacogenes were extracted. These genes included both clinically relevant pharmacogenes, similar to the panel in chapter 3, as well as a set of complex genes which have been associated with drug response. For 73 out of the 100 pharmacogenes, the entire locus could be resolved in a single haploblock. This means that for these genes both the paternal and maternal alleles were completely resolved. Of the genes characterized as being 100% complex (defined as repetitive sections and pseudogenes), 9 out of 15 were completely resolved. Finally, variant calling based on the long-read sequencing data was compared to the high quality benchmark data from genome in a bottle. This resulted in high precision and recall (>98%) of all variants in the pharmacogenomic regions, even in the highly complex regions. Resolving these complex regions can help to improve the characterization of pharmacogenes and thereby reduce the missing heritability.

In **chapter 5**, the use of long-read sequencing was combined with artificial intelligence to predict CYP2D6 drug metabolizer phenotypes on a continuous scale. For this, three patient cohorts were included. Two cohorts of individuals who used the CYP2D6 substrate tamoxifen (n=561 and n=167) and one venlafaxine treated cohort (n=69). For each substrate, the most CYP2D6 specific metabolic step was used as a proxy for CYP2D6 enzyme activity. With the use of amplicon based long-read sequencing of the entire *CYP2D6* locus, all genetic variants could be characterized and phased into haploblocks.

Subsequently, a neural network was trained on the data of the largest cohort (n=561). The neural network used 77 different variants to predict the metabolic capacity of CYP2D6 on a continuous scale. In the training cohort, this combination of full characterisation of the *CYP2D6* locus and a neural network resulted in an increase in explained variability to 79% compared to 54% with conventional phenotype predictions. These results were replicated in two independent cohorts, 66% for the neural network approach compared to 35% with the conventional approach for the second tamoxifen cohort and 64% compared to 55% for the venlafaxine cohort. This smaller increase in explained variability for the replication cohorts might be explained by the smaller sample size of these cohorts. Moreover, for the venlafaxine cohort substrate specific effects might play a role. Finally, the model was able to predict the impact of individual variants and alleles. For five of these variants, in vitro experiments with HEK cells and bufuralol were performed, this resulted in similar activities as predicted by the neural network model. This neural network approach showed a substantial increase in explained variability of CYP2D6 enzyme activity, indicating that a continuous scale for phenotype predictions might be one of the key components of solving the missing heritability.

Not every CYP2D6 substrate is affected equally by a particular variant within the *CYP2D6* locus, also known as substrate specificity. This generalization can result a mismatch between the predicted phenotype and observed phenotype. Formultiple CYP2D6 substrates, substrate specific effects have been reported in literature. However, a systematic comparison of the available data on substrate specificity in CYP2D6 metabolism is lacking. In **chapter 6**, we present such a systematic analysis of in vitro findings regarding substrate specificity. For *CYP2D6**17 it was observed that the measured activity was substantially higher for debrisoquine compared to other substrates. Which might be explained by the smaller molecule size of debrisoquine compared to the other substrates, leading to the drug being affected to a lesser extent by the changes in the size of the binding pocket caused by *CYP2D6*17*. For *CYP2D6**2 the activities varied greatly even within the same substrate, from clear decrease of function to even gain of function. Nonetheless, while in vitro assays can give an indication of variant impact, they might not always reflect the in vivo situation accurately. In the in vivo setting additional factors, such as genetic modifiers, might play a role as well. By accounting for substrate specific effect, the explained variability in drug response could be improved.

In **chapter 7**, we investigated the variability within CYP2D6 drug metabolizer phenotypes. We aimed to quantify the variability in enzyme activity within each phenotype category and find potential explanations and solutions for this variability. Data from three patient cohorts was used. Two cohort of individuals who used the CYP2D6 substrate tamoxifen

(n=561 and n=167) and one venlafaxine treated cohort (n=69). For the first cohort (n=561) there was a significant difference in variability between the NM (normal metabolizer) and IM (intermediate metabolizer) groups (1.26 fold, p=0.008), for the other cohorts the variability in the IM was higher or equal to that of the NM group. In all three cohorts, there were substantial differences between the GAS combinations 0.5+0.5 (decreased + decreased) and 1+0 (active + null-allele), indicative of a lower activity in the first group. Overall, these findings suggest that the variability is largest in the IM group, potentially caused by high variability in activity of the decreased activity alleles. Moreover, the fact that two decreased activity alleles do not show the same overall activity as one active and one non-active alleles is suggestive of an activity of less than 50% for the decreased activity alleles or of a non-additive model in regards to the gene activity scores.

In **chapter 8**, we discuss the findings of this thesis and potential future directions. The origin of missing heritability in pharmacogenomics is diverse and complex. Several key contributors to this missing heritability are; rare variants, genetic complexity and haplotype phasing, phenotype summarization and substrate specificity. We have shown that by accounting for these factors it is possible to resolve a substantial part of the missing heritability. This is most prominent for the combination of long-read sequencing with artificial intelligence (AI). Currently the application of long read sequencing combined with AI is labor- and cost intensive. However, in the near future the costs of sequencing will decrease and the acceptance of AI will increase, making the broad applications of AI based models a reality. These type of models can also include additional (non)-genetic factors which influence drug response, such as co-medication and comorbidities, making them truly unifying models to predict drug response. By combining high quality genetic data with all relevant patient characteristics and advanced AI models, missing heritability in PGx can be decreased and true personalized medicine becomes an achievable goal.

Nederlandse samenvatting

Farmacogenetica is een belangrijk onderdeel van *personalized medicine*. Door de keuze van geneesmiddel en/of dosis aan te passen op basis van genetische variatie is het mogelijk om bijwerkingen te voorkomen en de effectiviteit te verbeteren. Echter, de huidige farmacogenetische testen kunnen slechts een klein deel van de genetische variatie in geneesmiddelrespons verklaren. Dit leidt tot aanzienlijke *missing heritability*, wat gedefinieerd wordt als het deel van de variatie in enzymactiviteit waarvan bekend is dat de oorzaak genetisch is, maar dat nog niet verklaard kan worden met de genetische testuitslag. Mogelijke verklaringen voor deze *missing heritability* zijn: de rol van zeldzame varianten (met een allelfrequentie <1%), de complexiteit van de farmacogenetische loci door repetitieve secties en verstoring van pseudogenen, haplotype-fasering, substraatspecificiteit en het gebruik maken van categorieën in plaats van een continue schaal bij de voorspelling van metabole fenotypes.

De klinische farmacogenetica is voornamelijk gericht op het gebruikt van *single nucleotide variant* (SNV) panels en *copy number variants* (CNV's) in combinatie met een categoriaal model om fenotypes te voorspellen. Echter, in de afgelopen jaren zijn er nieuwe technieken om iemands genetische opmaak in kaart te brengen beschikbaar gekomen. Deze technieken bieden een kans om de *missing heritability* in farmacogenetica aan te pakken en om het effect van de reeds bekende farmacogenetische varianten beter te interpreteren . Uiteindelijk kan dit leiden tot een toename in de verklaarde variabiliteit in enzymactiviteit en daarmee tot betere behandeluitkomsten.

Om de *missing heritability* aan te pakken is het van belang om de variatie in de genen die betrokken zijn bij het geneesmiddelmetabolisme volledig in kaart te brengen. Om dit te doen zijn er verschillende technieken beschikbaar. In **hoofdstuk 2** hebben we deze technieken samengevat en besproken wat hun voordelen, nadelen en de potentiële toepassingen binnen de farmacogenetica zijn. SNV panels worden het meest gebruikt in de klinische praktijk vanwege hun snelheid en de relatief lage kosten. Echter, ze zijn beperkt in het aantal varianten en het type varianten dat gedetecteerd kan worden. In de afgelopen jaren zijn er steeds grotere SNV panels ontwikkeld, waarvan sommige miljoenen varianten bevatten, inclusief een verrijking van bekende farmacogenetische varianten. Desalniettemin zijn zij nog steeds beperkt tot de vooraf gedefinieerde varianten die op het panel zitten. Sequencing is daarentegen wel in staat om alle varianten in het stuk van het genoom dat gesequenced wordt te detecteren. Een van de meeste geavanceerde sequencingtechnieken is *long-read* sequencing wat in staat is om complexe regio's goed in kaart te brengen en de DNA-fragmenten te faseren naar het allel van oorsprong. Dit gaat echter wel gepaard met hogere kosten en een langere verwerkingstijd. Uiteindelijk is de keuze voor een genoty-

peringstechniek gebaseerd op de keuze voor de juiste techniek voor het juiste probleem en is er niet per se één techniek het beste.

Het hergebruiken van bestaande genetische data voor een farmacogenetische toepassing hebben we onderzocht in **hoofdstuk 3**. Voor 1.583 individuen hebben we gekeken of het mogelijk was om een farmacogenetisch profiel te extraheren uit *whole exome sequencing* (WES) data die waren gegenereerd voor diagnostische doeleinden. Alle individuen maakten deel uit van kind-ouder trio's, wat het mogelijk maakte om de haplotypes te faseren gebaseerd op familiaire relaties. Het panel bestond uit 42 varianten in 11 genen waarvoor een dosis advies beschikbaar is in de DPWG (*Dutch Pharmacogenetics Working Group*) richtlijnen. Helaas waren we niet in staat om CNV's van *CYP2D6* te bepalen in de WES-data. Dit heeft ertoe geleid dat er geen CYP2D6 fenotypes toegekend konden worden. Daarnaast was de dekking vaak onvoldoende voor 2 intronische varianten (*CYP3A5**3 en *CYP2C19**17) waardoor er geen fenotypes toegekend konden woorden voor CYP2C19 en CYP3A5 voor respectievelijk 99% en 26% van de individuen. Uiteindelijk kon voor 8 van de 11 genen een farmacogenetisch profiel gegenereerd worden. Van de gehele populatie had 85% een profiel waarbij tenminste 1 variant aanwezig was welke in de Nederlandse farmacogenetica-richtlijnen is opgenomen. Tevens bleek dat haplotype-fasering zorgt voor een verbetering in het voorspellen van fenotypes.

In **hoofdstuk 4** hebben we gekeken naar de mogelijkheden van *long-read sequencing* om de complexiteit van farmacogenen op te lossen. Hiervoor hebben we gebruik gemaakt van publiekelijk beschikbare *long-read sequencing* data van een individu van het *genome in a bottle* (GIAB) consortium. Voor dit GIAB sample is hoge kwaliteit *long-* en *short-read sequencing* data en benchmarking data beschikbaar. Met *short-read sequencing* zijn de *reads* niet altijd lang genoeg om de gehele complexe regio te dekken, wat weer leidt tot problemen met het accuraat in kaart brengen van deze genetische regio's. Daarnaast is het met kortere *reads* lastig om te achterhalen of een *read* afkomstig is van het coderende gen zelf of van een pseudogen dat voor een groot deel dezelfde sequentie heeft. Voor deze studie hebben we data voor 100 farmacogenen geëxtraheerd uit de *long-read sequencing* data. Deze genen omvatten zowel klinisch relevante genen, vergelijkbaar met het panel uit hoofdstuk 3, alsook een set met complexe genen welke geassocieerd zijn met geneesmiddelrespons. Voor 73 van de 100 farmacogenen was het mogelijk om de gehele locus in haploblok op te lossen. Dit betekent dat voor deze genen beide allelen volledig in kaart gebracht konden worden. Van de 15 genen die voor 100% als complex gedefinieerd worden (pseudogenen of repetitieve secties), konden er 9 volledig in kaart gebracht wordem. Als laatste hebben we de variantidentificatie gebaseerd op *long-read sequencing* data vergeleken met de *benchmarking* data van GIAB. Dit resulteerde in een hoge precisie en *recall* (>98%)

van alle varianten in de farmacogeneticaregio's, zelfs voor de zeer complexe regio's. Het oplossen van deze complexe farmacogenetische regio's kan helpen om de genen beter te karakteriseren en daarbij de *missing heritability* te reduceren.

In **hoofdstuk 5** hebben we het gebruik van long-read sequencing gecombineerd met kunstmatige intelligentie (KI) om zo CYP2D6 fenotypes te voorspellen op een continue schaal. Er werden drie cohorten geïncludeerd. Twee van deze cohorten bestonden uit individuen die het CYP2D6 substraat tamoxifen gebruikten (n=561 en n=167) en een cohort van gebruikers van het CYP2D6 substraat venlafaxine (n=69). Voor elk substraat werd de meest CYP2D6 specifieke stap in het metabolisme gebruikt als een maat voor de CYP2D6 enzymactiviteit. Door gebruik te maken van gerichte *long-read sequencing* van de gehele CYP2D6 locus, was het mogelijk om alle genetische varianten te karakteriseren en te faseren. Daarna werd een neuraal netwerk getraind aan de hand van de data van het grootste cohort (n=561). Het neurale netwerk gebruikte 77 varianten om de metabole capaciteit van CYP2D6 te voorspellen op een continue schaal. In het trainingscohort resulteerde deze combinatie van *long-read sequencing* met het neurale netwerk in een toename van verklaarde variabiliteit van 54% met de conventionele benadering tot 79% met de nieuwe benadering. Deze resultaten werden gerepliceerd in twee onafhankelijke cohorten, met 66% voor het neurale netwerk en 35% voor de conventionele benadering in het tweede tamoxifencohort en 64% ten opzichte van 55% voor het venlafaxinecohort. Deze beperktere toenames in de replicatiecohorten ten opzichte van het trainingscohort kunnen mogelijk verklaard worden door de kleinere omvang van de replicatiecohorten. Daarnaast kan substraatspecificiteit een rol spelen in het venlafaxinecohort. Als laatste hebben we in vitro experimenten uitgevoerd voor vijf varianten. Voor deze vijf varianten is het metabolisme van bufuralol onderzocht in HEK cellen. Dit resulteerde in een vergelijkbare activiteit in de HEK cellen als was voorspeld door het neurale netwerk. Deze KI-benadering vertoonde een substantiële verbetering in de verklaarde variabiliteit in CYP2D6 enzymactiviteit. Dit toont aan dat een continue schaal voor fenotype voorspellingen mogelijk een van de kernpunten is in het oplossen van de *missing heritability*.

Niet elk CYP2D6 substraat wordt in dezelfde mate beïnvloed door genetische varianten in het *CYP2D6*-gen, dit heet ook wel substraatspecificiteit. Op dit moment wordt substraatspecificiteit niet meegenomen in de klinische farmacogenetica. Voor elk substraat wordt hetzelfde effect van een variant gebruikt. Deze generalisatie kan leiden tot een discrepantie tussen de voorspelde en de daadwerkelijke fenotypes. Voor meerdere CYP2D6-substraten zijn er substraatspecifieke effecten vermeld in de literatuur. Echter, een systematische vergelijking van deze data ontbreekt nog. In **hoofdstuk 6** hebben we een systematische analyse van in vitro substraatspecificiteitdata uitgevoerd. Voor *CYP2D6**17 werd gezien

dat de gemeten activiteit hoger was in debrisoquine vergeleken met andere substraten. Wel was de activiteit voor alle substraten verminderd ten opzichte van het wildtype. Dit betekent dat debrisoquine minder beïnvloed wordt door de verandering in de substraat *binding pocket* die door *CYP2D6**17 veroorzaakt wordt. Bij *CYP2D6**2 werden grote variaties in de enzymactiviteit gezien, ook voor hetzelfde substraat. Dit ging van een duidelijke verminderde activiteit tot een toegenomen activiteit, ten opzichte van *1. Hoewel in vitro assays een indicatie kunnen geven van de impact van een variant, geven zij niet altijd de in vivo situatie goed weer. In een in vivo situatie zijn er meerdere additionele factoren die een rol kunnen spelen. Desondanks, door substraatspecificiteit mee te nemen in fenotypevoorspellingen, kan de verklaarde variabiliteit in geneesmiddelrespons waarschijnlijk in vele situaties wel verbeterd worden.

In **hoofdstuk 7** hebben we gekeken naar de variabiliteit binnen de CYP2D6 metabole fenotypegroepen. Hierbij was het doel om de variabiliteit binnen elke fenotypecategorie te kwantificeren en om mogelijk verklaringen en oplossingen voor deze variabiliteit te vinden. We hebben hiervoor data van drie patiëntcohorten gebruikt: twee cohorten met data van individuen die het CYP2D6 substraat tamoxifen (n=561 en n=167) hebben gebruikt en een cohort individuen die met venlafaxine waren behandeld (n=69). Bij het eerste cohort (n=561) was er een significant verschil in de variabiliteit tussen de NM (normal metabolizers) en IM (intermediated metabolizers) groepen (1,26 keer hoger in de IMs, p=0,008). Ook bij de andere cohorten was de variabiliteit in de IM-groep groter of gelijk aan dat van de NM-groep (p>0,05). In alle groepen waren er substantiële verschillen tussen de genactivititeitscore combinaties 0,5+0,5 (verminderd + verminderd) en 1+0 (actief + inactief), indicatief voor een lagere activiteit in de eerste groep. Deze bevindingen suggereren dat de variabiliteit het grootst is in de IM-groep, mogelijk veroorzaakt door de grote variabiliteit in de verminderd actieve allelen (genactiviteitscore 0.5). Het feit dat twee verminderd actieve allelen niet dezelfde enzymactiviteit vertonen als een volledig actief en een volledig inactief allel kan potentieel verklaard worden door een activiteit van minder dan 50% voor verminderd actieve allelen. Tevens kan dit een indicatie zijn voor een model waarbij de activiteit van beide allelen niet additief is.

In **hoofdstuk 8** bespreken we de bevindingen van dit proefschrift en de potentiële toekomstige onderzoeken. De oorsprong van *missing heritability* in farmacogenetica is divers en complex. Factoren die een cruciale rol spelen in de *missing heritability* zijn: zeldzame varianten, genetische complexiteit en haplotype-fasering, categorisatie van fenotypes en substraatspecificiteit. Wij hebben laten zien dat het meenemen van deze factoren het mogelijk maakt om een substantieel deel van de *missing heritability* te verklaren. Dit effect is het duidelijkst door toepassing van *long-read sequencing* gecombineerd met kunstmatige intel-

ligentie (KI). Op dit moment is de toepassingen van *long-read sequencing* met KI arbeids- en kostenintensief. Echter, in de nabije toekomst zullen de kosten van sequencing dalen en zal de acceptatie van KI stijgen. Dit maakt de brede toepassing van farmacogenetica-KI toepassingen mogelijk. Neurale netwerken kunnen ook aanvullende (niet-)genetische factoren die de geneesmiddelrespons beïnvloeden, zoals comedicatie en comorbiditeiten, meenemen. Door het combineren van hoge kwaliteit genetische data met alle relevante patiëntkarakteristieken en KI-modellen kan de *missing heritabilitlity* in farmacogenetica verkleind worden en kunnen geneesmiddelbehandelingen nog beter geoptimaliseerd en gepersonaliseerd worden.