Towards solving the missing heritability in pharmacogenomics

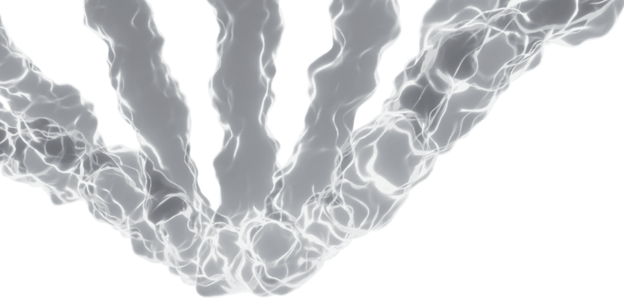
Lee, M. van der

Citation

Lee, M. van der. (2022, January 19). *Towards solving the missing heritability in pharmacogenomics*. Retrieved from https://hdl.handle.net/1887/3250514

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3250514

Note: To cite this publication please use the final published version (if applicable).



CHAPTER 3

Repurposing of diagnostic whole exome sequencing data of 1,583 individuals for clinical pharmacogenetics

Maaike van der Lee, William G. Allard, Sander Bollen, Gijs W.E. Santen,
Claudia A.L. Ruivenkamp, Mariëtte J.V. Hoffer, Marjolein Kriek, Henk-Jan Guchelaar,
Seyed Y. Anvar, Jesse J. Swen

Clinical Pharmacology and Therapeutics. 2020 Mar;107(3):617-627

Abstract

For approximately 80 drugs, widely recognized pharmacogenetics dosing guidelines are available. However, the use of these guidelines in clinical practice remains limited as only a fraction of patients is subjected to pharmacogenetic screening. We investigated the feasibility of repurposing whole exome sequencing (WES) data for a panel of 42 variants in 11 pharmacogenes to provide a pharmacogenomic profile. Existing diagnostic WES-data from child-parent trios totalling 1,583 individuals were used. Results were successfully extracted for 39 variants. No information could be extracted for three variants, located in *CYP2C19*, *UGT1A1* and *CYP3A5*, and for *CYP2D6* copy number. At least one actionable phenotype was present in 86% of the individuals. Haplotype phasing proved relevant for *CYP2B6* assignments as 1.5% of the phenotypes were corrected after phasing. In conclusion, repurposing WES-data can yield meaningful pharmacogenetic profiles for 7 out of 11 important pharmacogenes which can be used to guide drug treatment.

Introduction

Pharmacogenetics (PGx) aims to optimize drug treatment by preventing adverse drug reactions and by increasing drug efficacy through adjustments based on one's genetic profile. For approximately 80 drugs, there is convincing evidence that PGx testing prior to prescribing leads to improved patient outcome [1-3]. Hence, both the Clinical Pharmacogenetics Implementation Consortium (CPIC) and the Dutch Pharmacogenetics Working Group (DPWG) have developed widely recognized guidelines [3-5], which provide dose and drug adjustments based upon available PGx results [3-6]. In the Netherlands, for instance, these guidelines are integrated into the electronic prescribing and dispensing systems nationwide and are available at point of care [7].

Both the CPIC and DPWG PGx guidelines originally considered patients with a known genotype and were developed in anticipation of having clinical high-throughput and pre-emptive genotyping as the standard of care. However, many applications of PGx are still reactive, following a novel prescription or unexplained adverse drug reactions (ADRs). In contrast, pre-emptive testing for a panel of pharmacogenes can be used to prevent ADRs and improve treatment efficacy. The potential impact of such a pre-emptive panel-based approach is high [2]. Unfortunately, despite the established impact of PGx on the outcomes of drug treatment and the availability of guidelines, the number of patients with known pharmacogenetic genotypes remains limited [7].

Whole Exome (WES) and Whole Genome Sequencing (WGS) have rapidly become part of the diagnostics process in the field of clinical genetics to diagnose potential genetic disorders of unknown etiology. Their application has resulted in vast amounts of sequencing data being generated. These data hold the potential to retrieve genetic information for a panel of PGx genes which can then be repurposed to pre-emptively guide drug treatment. Next Generation Sequencing (NGS) approaches specifically designed for PGx have shown promising results, with high concordance (91–99%) with conventional PGx methods [8-10]. Moreover, Cousin et al. have shown that extracting information on three PGx genes from existing clinical WES data can be beneficial in terms of drug dose and response in a small cohort of 94 patients [11]. However, this study is limited by investigating a small number of PGx genes rather than extracting a full panel of genes with actionable recommendations in PGx guidelines as well as by the use of a limited sample size.

Given that PGx haplotype and phenotype assignments are currently performed based on linkage disequilibrium between known variants it is possible that having access to phasing information may change the haplotype and phenotype assignment and potentially the clinical recommendation. Thus, trio-based sequencing data is a particularly interesting source for the extraction of PGx variants.

In this study, we assessed the feasibility of repurposing diagnostic WES data to extract a meaningful PGx profile based on the panel used in the Ubiquitous Pharmacogenomics consortium (U-PGx, www.upgx.eu) that includes all actionable genes and variants in the DPWG guidelines [12,13]. Additionally, we investigated the added value of haplotype phasing.

Results

Study cohort

The entire cohort consists of a total of 1,583 individuals from two different sub-cohorts (Figure 3.1). Both sub-cohorts consist of patients suffering from an intellectual disability (ID) and/or multiple congenital anomalies and their parents, all of whom underwent diagnostic WES and received genetic counselling during the process at the Department of Clinical Genetics at the Leiden University Medical Centre. The first, prospective, sub-cohort consists of individuals who were offered the opportunity to receive their PGx profile in addition to their diagnostic WES results. Between August 2016 and April 2018, 168 individuals belonging to 57 families (55 full trios, one trio with the exclusion of the index patient, and one single parent) provided informed consent and had their WES data available at time of analysis. The second, retrospective, sub-cohort consists of individuals who underwent diagnostic WES prior to August 2016. This sub-cohort comprised of 1,415 individuals with fully anonymized data. The protocol was approved by the Institutional Review Board of the Leiden University Medical Centre.

Sequencing data and variant selection

Short reads were aligned to reference genome GRCh37, followed by variant calling and haplotype phasing (Figure 3.1). A panel of 42 genetic variants covering 11 actionable pharmacogenes (Supplementary Table S3.1), was composed. Variant and gene selection was based on the panel used in the Ubiquitous Pharmacogenomics consortium (U-PGx, www.upgx.eu) and includes all actionable genes and variants in the DPWG guidelines [12,13].

Genotype and diplotype calling

The majority of diplotypes (79.1%, N=13,768 out of 17,413 potential diplotype calls) could be extracted from the available WES data. For the remaining 20.9%, one or more Single Nucleotide Variants (SNVs) could not be determined. These sites either lacked sufficient coverage for reliable variant calling – *CYP2C19* (rs12248560; g.96521657C>T in 99% of individuals, N=1,568), *CYP3A5* (rs776746; g.99270539C>T in 26% of individuals, N=420)

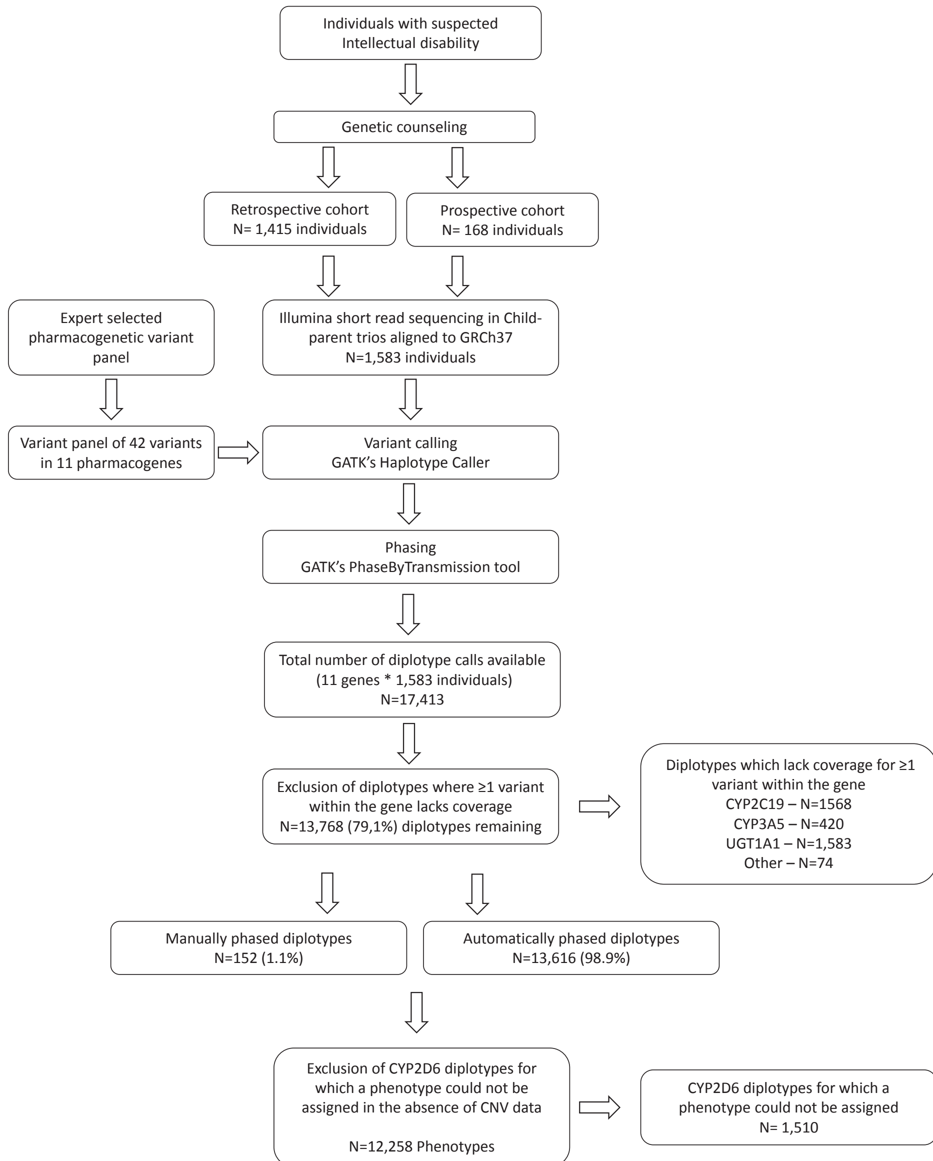


Figure 3.1: Study flowchart

Whole Exome Sequencing data from individuals sequenced for diagnostics was used to obtain a clinically relevant Pharmacogenetics profile. Retrospective cohort: individuals sequenced prior to August 2016, prospective cohort: individuals sequenced after august 2016 if they opted in for obtaining their pharmacogenetic (PGx) profile. The expert selected pharmacogenetic panel was obtained from the Ubiquitous Pharmacogenomics Consortium. Sufficient coverage was classified as haplotype quality of at least 20. Due to the absence of Copy Number variants (CNV), only *CYP2D6* diplotypes consisting of two null-alleles were included as CNVs would not change the phenotype assignment. Manual phasing and phenotype assignments were based on translation tables from the Ubiquitous Pharmacogenomics Consortium. GATK: Genome Analysis Tool Kit, CNV: Copy Number Variant.

and 74 calls divided over 6 genes – or could not be called using the GATK HaplotypeCaller – *UGT1A1* TATA-box repeat unit for all individuals (N=1,583) (Figure 3.2). Haplotypes were assigned based on U-PGx translation tables, using conventional *-alleles for Cytochrome P450 enzymes and DPWG nomenclature for the remaining haplotypes [14,15] (Supplementary Table S3.1).

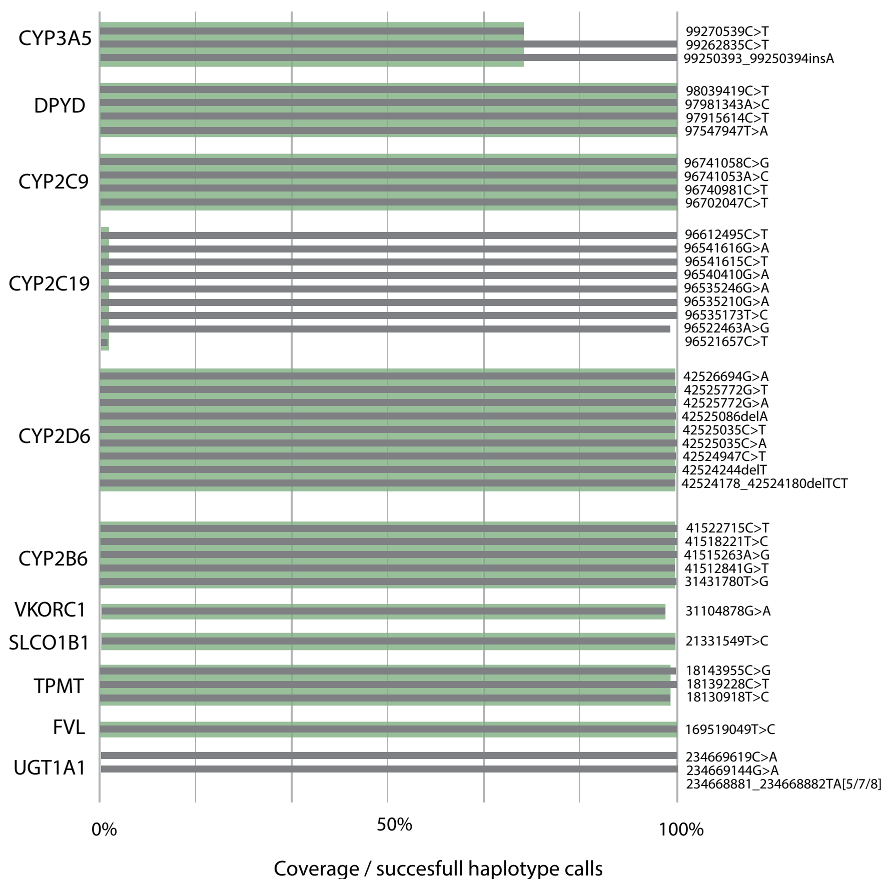


Figure 3.2: Call rate per gene
In grey: coverage per variant, in green: diplotypes available per gene. For 20.9% of all diplotype calls there was insufficient coverage (haplotype quality <20) for ≥1 variant. Copy number variants could not be determined, results are based on Single Nucleotide Variants only. In total 13,768 diplotype calls could be included.

Phasing

In total 13,768 diplotypes were called (Table 3.1 and Supplementary Table S3.2), leading to the majority of individuals (70%) with complete SNV data for at least 9 genes (Supplementary Figure S3.1). An automatically phased diplotype could be assigned to

Table 3.1: Haplotype frequencies
Frequencies based on all haplotypes, including manually phased haplotypes. Genes are included if there is sufficient coverage for all variants within that gene. Haplotype assignments are based on translation tables from the Ubiquitous Pharmacogenetics Consortium.

Gene	Haplotype assignment	Number of alleles	Frequency
CYP2B6	Total	3,154	
	*1	2,279	72.0%
	*18	3	0.09%
	*4	108	3.4%
	*6	733	23.2%
	*9	31	0.98%
CYP2C19	Total	30	
	*1	21	70.0%
	*17	5	16.7%
	*2	3	10.0%
	*4A/B	1	3.3%
CYP2C9	Total	3,166	
	*1	2,561	80.9%
	*11	4	0.13%
	*2	377	11.9%
	*3	222	7.0%
	*5	2	0.06%
CYP2D6[a]	Total	3,152	
	*1	1,966	62.4%
	*10	102	3.2%
	*17	21	0.7%
	*3	51	1.6%
	*4	554	17.6%
	*41	362	11.5%
	*6	28	0.9%
	*9	68	2.2%
CYP3A5	Total	2,326	
	*1	299	12.9%
	*3	2,002	86.1%
	*6	21	0.9%
	*7	4	0.17%
DPYD	Total	3,162	
	*1	3,046	96.3%
	*2A	21	0.66%
	1236G>A	75	2.2%
	2846A>T	20	0.63%

Table 3.1 continues on next page.

Table 3.1: ***Continued***

Gene	Haplotype assignment	Number of alleles	Frequency
FVL	Total	3,166	
	F5 positive	80	2.5%
	F5 negative	3,086	97.5%
SLCO1B1	Total	3,158	
	*5	443	14.0%
	wt	2,715	86.0%
TPMT	Total	3,124	
	*3A	122	3.9%
	*3C	20	0.29%
	*2	1	0.03%
	wt	2,981	95.8%
VKORC1	Total	3,098	
	1173T	1,247	40.3%
	wt	1,851	59.7%

[a] *CYP2D6* gene duplications and gene deletions could not be determined. F5: Factor V Leiden.

13,616 calls (98.9%). For 152 calls (1.1%) a phased diplotype call could not be automatically resolved and required manual curation based on the translation table from the U-PGx consortium. Of the 152 initially unresolved calls, 103 were in *CYP2B6*, for which the heterozygous presence of the g.41515263A>G and g.41512841G>T variant can lead to both a *1/*6 call as well as a *4/*9 call. Due to the high linkage disequilibrium between the g.41515263A>G and g.41512841G>T variants it is commonly assumed that these variants occur on the same allele. Therefore, individuals carrying both the g.41515263A>G and g.41512841G>T variants are generally genotyped as *CYP2B6**1/*6. This high linkage disequilibrium between g.41515263A>G and g.41512841G>T was also observed in our cohort (Figure 3.3A). Wherever automatic haplotype phasing was possible, we observed an improved accuracy in diplotype calls in the *CYP2B6* gene. Namely, of the heterozygous carriers of the g.41515263A>G and g.41512841G>T variants, 409 individuals with the *CYP2B6**1/*6 haplotype and 6 individuals with the *CYP2B6**4/*9 haplotype were identified (Figure 3.3B).

The remaining 49 diplotype calls that could not be phased automatically were distributed over *CYP2C9* (N=2), *CYP2D6* (N=28), *CYP3A5* (N=1) and *TPMT* (N=18) (Supplementary Table S3.3). Based on the final haplotype assignments there were no significant differences in the haplotype frequencies observed in the children compared to their parents, with the exception of *VKORC1* (Supplementary Table S3.2).

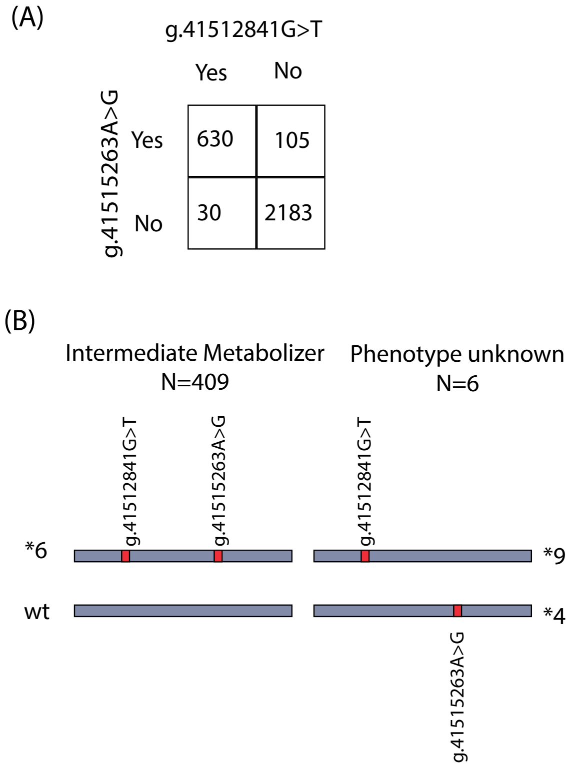


Figure 3.3: Phasing and linkage disequilibrium in *CYP2B6*
(a) High linkage disequilibrium is shown by the frequencies of the *CYP2B6**4 and *9 variants in all automatically phased haplotypes (N=2,948 alleles), when combined the haplotype is *6. χ^2, p<0.0001. **(b)** Possible configurations for the *CYP2B6* variants, leading to different phenotypes. Numbers are based on all individuals who carried both variants and could be phased automatically (N=415).

Phenotype calling

Diplotypes were translated into phenotypes based on the DPWG guidelines and U-PGx translation tables (Table 3.2). Due to the inability to call Copy Number Variants (CNVs) for *CYP2D6*, CYP2D6 phenotypes could only be called for individuals carrying two null-alleles (e.g. *CYP2D6**4/*4 or *CYP2D6**3/*6) and for these a poor metabolizer (PM) phenotype was assigned (N=66). For the remaining 1,510 individuals with sufficient coverage on all SNVs, no CYP2D6 phenotype could be assigned. In total 20.7% of assigned phenotype calls (2,534 out of 12,258) were actionable, with actionable defined as a phenotype which is mentioned in the DPWG guidelines with at least one actionable recommendation, e.g. a dose change or change of drug (Figure 3.4). The phenotype observed in children did not always match the phenotype observed in either one of the parents. For example, the child can be a CYP2C9 PM while the parents both have an IM phenotype (Supplementary

Table 3.2: Phenotype frequencies and actionability
Phenotypes are based on the U-PGx translation tables, actionability is based on the Dutch Pharmacogenetic Working Group guidelines, where actionable is defined as a phenotype accompanied by at least one dosing advise.

Gene	Phenotype	Number of subjects	Frequency	Actionable
CYP2B6		1,577		
	PM	105	6.7%	Yes
	IM	528	33.5%	Yes
	EM	944	59.9%	
CYP2C19		15		
	PM	-		Yes
	IM	4	26.7%	Yes
	EM	11	73.3%	
	UM	-		Yes
CYP2C9		1,583		
	PM	59	3.7%	Yes
	IM	487	30.2%	Yes
	EM	1,037	65.5%	
CYP2D6		1,576		
	PM*	66	4.2%	Yes
	Not assigned	1,510	95.8%	
CYP3A5		1,163		
	PM	882	75.8%	
	IM	263	22.6%	Yes
	EM	18	1.5%	Yes
DPYD		1,581		
	AS: 0	-		Yes
	AS: 0.5	-		Yes
	AS: 1	21	1.3%	Yes
	AS: 1.5	95	6.0%	Yes
	AS: 2	1,465	92.7%	
F5L		1,583		
	F5 Absent	1,504	95.0%	
	F5 Heterozygous	78	4.9%	Yes
	F5 Homozygous	1	0.06%	Yes
SLCO1B1		1,579		
	Normal function	1,172	74.2%	
	Decreased function	371	23.5%	Yes
	Poor function	36	2.3%	Yes

Table 3.1 continues on next page.

Table 3.2: ***Continued***

Gene	Phenotype	Number of subjects	Frequency	Actionable
TPMT		1,562		
	PM	2	0.1%	Yes
	IM	139	8.9%	Yes
	EM	1,421	91.0%	
VKORC1		1,549		
	Normal function (1173CC)	564	36.4%	
	Decreased function (1173CT)	723	46.7%	
	Poor function (1173TT)	262	16.9%	Yes

PM: Poor metabolizer, IM: Intermediate Metabolizer, EM: Extensive Metabolizer, UM: ultra-rapid Metabolizer, AS: Activity Score, F5: Factor V Leiden.
* Poor metabolizer phenotype assigned based on diplotype consisting of 2 null-alleles. For all other diplotypes no CYP2D6 phenotype could be assigned as copy number variants could not be determined.

Figure S3.2). The majority of individuals (N=1,360; 85.9%) carried at least one gene with an actionable phenotype (Figure 3.4D).

Diplotype to phenotype translations based on CPIC guidelines yield similar results (Supplementary Table S3.4, Supplementary Figure S3.3), with 85.1% (N=1,347) of the population carrying at least one gene with an actionable phenotype and 2,459 actionable phenotypes.

Comparison of genotyping

To assess the correctness of assigned phenotypes based on WES data, 7 trios (21 individuals) were randomly selected for orthogonal genotyping on a commercial platform (The pharmacoscan from Thermo Fisher Scientific, Waltham, Massachusetts, USA). Due to an inability to call CNV for *CYP2D6* based on WES data, no CYP2D6 phenotype was assigned for the majority of individuals. Nonetheless, a comparison of SNVs identified in the WES pipeline with the SNVs identified on the commercial platform was possible. Of the diplotypes that could be called on both platforms (N=161), the concordance was 96.9% (N=156). Due to insufficient genotype quality, 49 diplotype calls had to be excluded from the WES data. These calls were located in the *UGT1A1*, *CYP2C19*, *CYP3A5* and *VKORC1* genes. On the commercial platform, calls were available for all individuals and all genes with the exception of Factor V Leiden which was not present on the array.

Of the five discordant calls, one was due to a *DPYD* variant (1236G>A) which was not included in the commercial platform. Two additional calls could not be resolved by the commercial platform due to the absence of phasing information, both of which were *TPMT* *1/*3A. The fourth discrepant call concerned a gene duplication of *CYP2D6* that

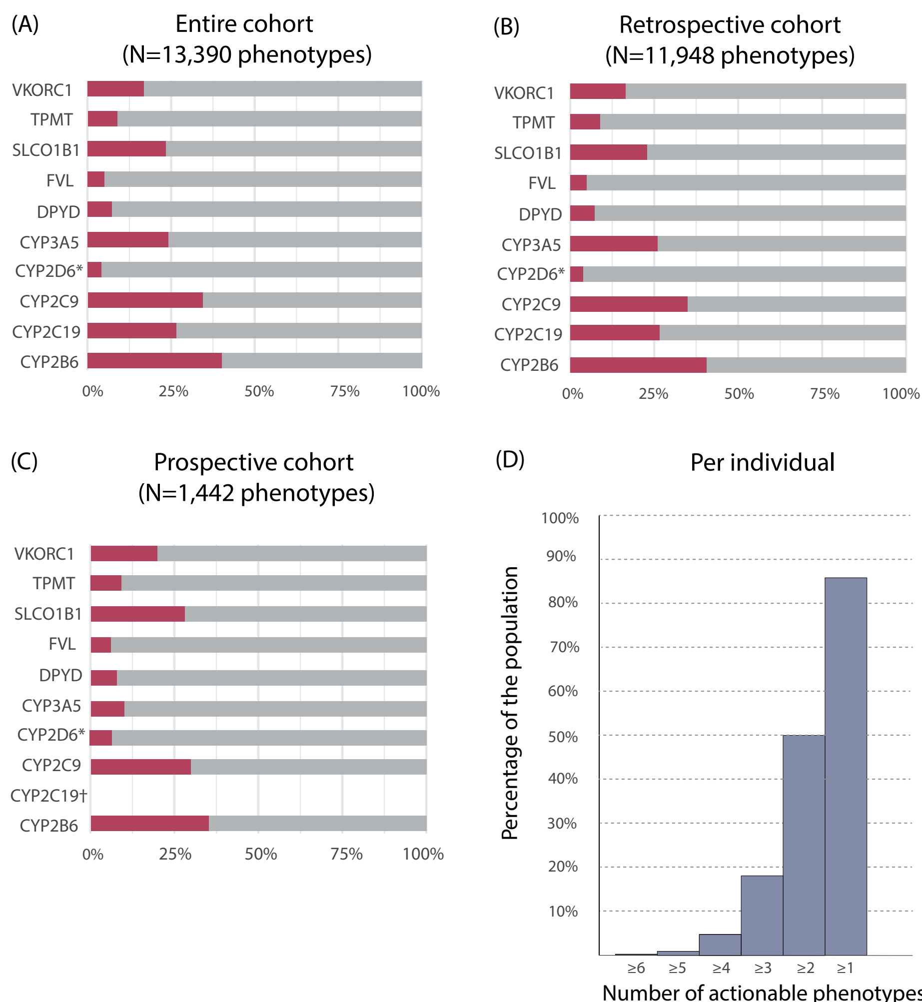


Figure 3.4: Actionable phenotypes
(a) Actionable phenotypes of the entire cohort, **(b)** the retrospective cohort and **(c)** the prospective cohort. Red: actionable, grey: not actionable **(d)** The total number of actionable phenotypes per individual. Actionable is classified as: any phenotype with a dosing advice available in the DPWG guidelines. An unknown phenotype is categorized as not actionable. Results are based on all genotypes with sufficient coverage (haplotype quality >20). * Due to an inability to call copy number variants, only *CYP2D6* diplotypes consisting of 2 null-alleles were assigned a phenotype (PM), no phenotype was assigned for other phenotypes classifying them as unknown. † CYP2C19 phenotypes could not be determined for any of the individuals in the prospective cohort due to a lack of coverage for one of the variants.

could not be identified in the WES data due to limitations in CNV calling. Given the observed diplotype of this last call (*CYP2D6**1/*4), the predicted phenotype does not change in the presence of a duplication based on the U-PGx translation tables and is classified as intermediate metabolizer both with and without duplication. Lastly, there

was a discrepant *CYP2D6* diplotype call. The WES pipeline called a *CYP2D6**4/*10 while the commercial platform called a *CYP2D6**4/*4 without duplication. Looking closer at the WES data revealed that this individual was homozygous for the *CYP2D6**10 (g.42526694G>A) variant. This variant is also part of the *CYP2D6**4 haplotype. The *CYP2D6**4 variant (g.42524947C>T) was found in 87% of all *CYP2D6* reads, indicating heterozygosity (Supplementary Figure S3.4). Given the presence of multiple wildtype calls for the *CYP2D6**4 (g.42524947C>T) variant in the WES data it not likely that this individual is homozygous for this variant, and the *CYP2D6**4/*10 assignment appears most probable.

Exploratory analysis of the heterozygosity ratio to assess CYP2D6 deletions

Due to limited consent, only genotypes for the selected SNVs in the *CYP2D6* locus (Supplementary Table S3.1) were available and it was not possible to use microsatellites to determine *CYP2D6* deletions [16]. Therefore we explored if it is possible to use the *CYP2D6* heterozygosity ratio to assess potential *CYP2D6* deletions. For each individual with complete SNV data (N=1,576), a heterozygosity ratio was calculated. A high proportion of heterozygous variants indicates the presence of two different *CYP2D6* alleles and therefore a low probability of a *CYP2D6* deletion. However, a low ratio does not confirm the presence of a deletion, as an individual can be homozygous for all variants in the *CYP2D6* locus. The distribution of the heterozygosity ratio (Figure 3.5) shows that approximately half of

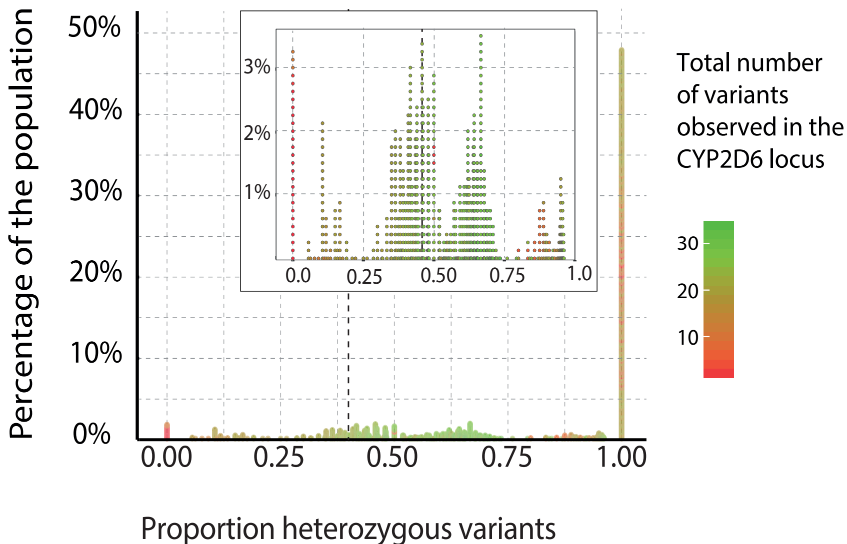


Figure 3.5: Heterozygosity ratios in *CYP2D6*
All variants in the *CYP2D6* locus were included. The higher the proportion of heterozygous variants the less likely a gene deletion is.

the individuals (47.4%) is heterozygous for all variants observed, thereby excluding the possibility of a *CYP2D6* deletion. A heterozygosity ratio cut-off set at <0.25 resulted in 78 (5.3%) individuals for whom a *CYP2D6* deletion could not be ruled out. Moreover, of these 78 individuals 7.7% were genotyped as heterozygous for at least one the SNVs in our panel (Supplementary Table S3.1). As these variants are all located in the exons of *CYP2D6*, a full gene deletion is highly unlikely for these individuals. To decrease the risk of false negative results a more conservative heterozygosity ratio cut-off was set at 0.4 resulting in 173 (11%) individuals for whom a *CYP2D6* deletion could not be ruled out.

Discussion

We have shown that repurposing existing diagnostic WES data for pharmacogenetics yields successful results for a large proportion of diplotype calls (13,768 out of 17,413 potential calls, 76.9%). Unfortunately, inherent to the use of WES data, several phenotypes could not be accurately determined due to a lack of coverage and missing CNV information, resulting in 12,258 reliable phenotype calls (70.4%). Our data also show that 86% of the studied population carried at least one actionable phenotype. This number is lower compared to the 91–99% indicated in previous studies [17,18]. Since frequencies of the haplotypes that could be identified in this study were comparable to frequencies reported in literature this is not expected to be the cause of the lower number of individuals with an actionable phenotype [19]. However, the lack of coverage for several genes and inability to identify *CYP2D6* CNVs, can be the cause of this discrepancy. These limitations in coverage and CNV calling were also reported previously when using WES data for PGx [8,10].

A novelty of our study is the use of haplotype phasing for resolving pharmacogenetic genotypes and phenotypes. For 6 individuals the haplotype phasing proved to be valuable in genotyping *CYP2B6*. For *CYP2B6* it is known that the *4 and *9 variants can occur separately even though they are in strong linkage disequilibrium. Conventionally, haplotype assignments use linkage disequilibrium in the assignment. For *CYP2B6* this means that most laboratories assume the *4 (g.41515263A>G) and *9 (g.41512841G>T) variants to be located on the same allele, resulting in a *6 assignment. We have now shown that in 1.5% of the observations (N=6 out of 415), where an individual is heterozygous for both the *4 and the *9 variant, the variants are located in trans-conformation (*CYP2B6**4/*9). Animal and tissue studies have shown conflicting results in regards to the impact of these individual variants on enzyme function, resulting in uncertainty as to what the effect is on enzymatic function and therefore what phenotype should be assigned [20-24]. In the DPWG guidelines, *CYP2B6**6 is designated as non-functional. However, this assignment is

only for the combination of the two SNVs (*CYP2B6**6) and not for the individual variants (*CYP2B6**4 and *CYP2B6**9) [3].

While we applied haplotype phasing based on child-parent trios, alternative methods are available. Haplotype phasing methods can be roughly categorised into direct and inferential approaches. Direct methods include single-cell sequencing and paired read sequencing, which are accurate but also costly [25]. Inferential methods are either pedigree or population-based. Population based methods use estimated probabilities based on population frequencies and pedigree based methods use the shared alleles between two or more individuals [26].

Assessing CNVs from WES data is challenging, which is particularly limiting for *CYP2D6*. A previous study used eXome Hidden-Markov Model (XHMM) as a tool to identify CNVs in WES data, leading to discrepancies in CNV in 7% of *CYP2D6* calls, compared to orthogonal testing [8]. Furthermore, multiple approaches determining CNVs based on sequencing depth have been developed over the past years. Unfortunately, these methods have been shown to be unreliable for individual patient level CNV calling, leading to a limited value of these algorithms in the diagnostic setting [27]. Another widely recognized approach is the use of microsatellite markers with a high degree of heterozygosity [16]. However, due to consent limited to the predefined SNVs (Supplementary Table S3.1) we were not able to use this method. Therefore, we explored the use of the heterozygosity ratio of the entire *CYP2D6* locus to asses *CYP2D6* deletions. A heterozygosity ratio cut-off set at 0.25 (25% heterozygous variants) resulted in 5.3% of the individuals for whom a *CYP2D6* deletion could not be ruled out. This frequency appears concordant with the 1–7% *CYP2D6* gene deletions reported in literature for the Dutch population [28]. However, 6 individuals in this group were genotyped heterozygous (e.g. *CYP2D6**1/*4) for at least one of the selected important SNVs in *CYP2D6* (Supplementary Table S3.1), all of which are exonic variants. The presence of heterozygous variants in the *CYP2D6* exons indicates that a *CYP2D6* gene deletion is highly unlikely indicating that a low ratio of heterozygosity does not necessarily confirm a gene deletion. Additionally, the presence of heterozygous variants in the upstream and downstream regions captured in our aggregated data could result in a high ratio even in the presence of a gene deletion. A more conservative cut-off of was set at 0.4, to decrease the change of excluding individuals with false high ratios With this cut-off a *CYP2D6* deletion could not be ruled out for 11% of the individuals, indicating a lower risk for false negative results. In case of limited consent an approach based on the heterozygosity ratio could potentially be used to indicate individuals for whom a gene deletion is highly unlikely. Obtaining reliable data in regards to CNVs in *CYP2D6* is of importance for clinical practice, as the presence of a deletion or duplication can both increase and decrease CYP2D6 enzyme function. Only for individuals with two

null-alleles based on SNV data (e.g. *3/*3) a CNV will not change the assigned phenotype (PM). In our cohort this led to reliable phenotype calls for 4.2% of the individuals, for the remainder of the cohort it is expected that 2–9% of the individuals will carry either a gene deletion or duplication which will affect the phenotype assigned [19]. Therefore, we argue that due to the inability to accurately phenotype for the majority of the population, phenotypes based on our pipeline should not be used in clinical practice until there are more accurate methods for CNV calling available.

Since difficulties with coverage and CNV calling are inherent limitations to the use of WES, it is difficult to solve these problems without resorting to other technologies like WGS or Array-based techniques. These techniques have been suggested by other groups as more suitable for PGx profiling [8,29]. However, both these technologies have not been routinely implemented as extensively as WES in a diagnostic setting. Applying these technologies will, therefore, lead to additional testing costs while repurposing existing WES data does not. Moreover, a relatively simple solution to increase the performance of WES for PGx is to expand the WES capture kit used with relevant intronic sites. One downside to this, compared to WGS, is that the need might arise for additional intronic regions which are not yet included in the capture kits. Additionally, increasing the capture kit will still not resolve CNV calling problems and a new capture kit will need to be tested thoroughly. Nonetheless, the main limitations of WES are gene-specific and do not apply to all pharmacogenes. Therefore, we argue that despite the inherent limitations of WES data for PGx, a reliable profile can be extracted from WES data for the majority of clinically relevant pharmacogenes.

Currently multiple efforts are ongoing to develop tools that can assist in extracting PGx profiles from NGS data. One such approach is the Stargazer tool [30,31]. Stargazer incorporates haplotype calling for 28 pharmacogenes and CNV calling for *CYP2D6* based on NGS data in a user-friendly algorithm. Stargazer's ability to call *CYP2D6* diplotypes was evaluated on a sample consisting of WGS data of 32 trios, showing a 99.0% concordance with conventional SNV-typing. While results are very promising, validation in a larger sample is needed. Moreover, any approach with WES data is still restricted to the limitations inherent to the use of WES, as described above.

The potential impact of implementing a pre-emptive PGx panel-based test for a substantial number of individuals, as can be done by repurposing diagnostic sequencing data, is large. In the Netherlands there are 3,628,597 incident prescriptions per year (1 in every 19) for drugs that interact with the genes included in our study (excluding only *FVL*). Based on simulations it was estimated that 23.6% of these prescriptions would lead to an actionable gene-drug interaction [32]. By testing pre-emptively, these drug-gene interactions can be managed in a timely manner, potentially reducing the number of

ADRs. More specifically, the individuals included in this study were originally sequenced to diagnose the cause of Intellectual Disability (ID) in the children. Several studies have shown that polypharmacy is more common amongst ID patients [33-35]. More importantly, ID patients more often use antipsychotics, anticonvulsants and/or antidepressants that frequently result in gene-drug interactions making the value of a PGx profile for this population even more meaningful. Unfortunately, many of these drugs are metabolized by either CYP2D6 or CYP2C19, both of which have shown to be difficult to determine based on existing WES data in our study.

The results of our actionability analysis are based on a panel of 11 genes designed by the U-PGx consortium to cover all actionable pharmacogenes in the DPWG guidelines [12]. However, many groups implementing PGx also use CPIC guidelines [36]. The CPIC guidelines currently provide recommendations for genetic variants in 19 genes [5]. Due to limited consent, covering only the DPWG genes, it was not possible to determine the phenotypes for all genes covered by the CPIC guidelines. However, for the 11 genes for which SNV data were available, translations based on CPIC guidelines showed that the number of individuals with at least one actionable phenotype was comparable to the DPWG guidelines (85.1% and 85.9% respectively) [37]. The slight difference was due to the Factor V Leiden gene, which is not included in the CPIC guidelines and therefore not actionable.

Conclusion

Despite the inherent limitations in regards to coverage of intronic variants and CNVs, this study shows that it is possible to repurpose existing diagnostic WES data to extract a PGx profile for 7 out of 11 clinically actionable PGx genes. Additionally, the availability of trio data with phased haplotype information allows more accurate phenotype predictions, particularly for CYP2B6.

Methods

At the Leiden University Medical Center, WES for diagnostic purposes has been used since 2013. When possible, the index patient and both of the parents are sequenced to allow for haplotype phasing and the discovery of de novo variants. From August 2016 onwards, individuals were asked if they wanted to retrieve their PGx profile from their WES data. Individuals who consented were included in the prospective sub-cohort of this study (Figure 3.1). The retrospective sub-cohort consisted of individuals sequenced prior to August 2016 who were assigned anonymous study IDs before inclusion (N=1,415). All individuals received genetic counselling during the diagnostic WES process. The study was approved by the Institutional Review Board of the Leiden University Medical Center.

Variant selection

The gene and variant panel used was based on the panel designed for the U-PGx consortium's PREPARE study (version June 2017), with the exclusion of the *HLA* genes due to their high complexity and the lack of tagging SNPs in the Caucasian population [13,38]. In brief, variants were selected based on the availability of a corresponding DPWG guideline, the effect of the variant on protein function and the frequency of the variant [12]. The final panel consisted of 42 variants located in 11 pharmacogenes.

Variant calling

Sequencing was performed on Illumina HiSeq4000 (Illumina, San Diego, California, USA) using 150bp reads, from 2015 onward. Samples analysed prior to 2015 were sequenced on HiSeq2500 (100bp reads) or HiSeq2000 (100bp reads). Paired-end sequencing technology was used. Agilent sureselect V5 was used for enrichment. Short reads were aligned to reference genome GRCh37, using the bwa tool with the BWA-MEM algorithm, [39], followed by variant calling using the GATK's HaplotypeCaller [40]. In order to accurately phase the reads, the analysis was performed in child-parent trios. Data from individuals who did not consent to retrieve their PGx results was used for phasing and disregarded thereafter. VCF files were phased using the GATK's PhaseByTransmission tool [41] resulting in two fully phased alleles for each individual in the trio. Variants that could not be phased were reported separately. For each locus of interest, fasta sequences were generated for each allele of each individual, by applying the variants in each locus to the reference sequence by using Mutalyzer [42,43]. VCF files were then used to create a coverage track in BED format for each individual for each locus of interest. A haplotype quality (GQ) of at least 20 is required to be considered 'covered'. The phased VCF file per individual is

additionally used to enumerate the total number and heterozygous number of variants per locus per individual.

Genotyping and phenotyping

The fasta sequences were used for genotype assignments. Haplotype assignments were done according to U-PGx translation tables [12,13]. A 'No call' was assigned when at least one variant in the gene lacked coverage. If an unphased variant was present and no other variants were observed, the individual was haplotyped as being heterozygous for the unphased variant. In the case of multiple variants in the gene of which at least one unphased, genotype calling was done manually based on linkage and the most likely combination of variants by using the U-PGx assignments for these variant combinations. Phenotypes were assigned according to the U-PGx translation tables based on the DPWG guidelines [12,13]. For *CYP2D6*, only when the assigned diplotype consisted of two null-alleles (e.g. *CYP2D6**4/*4 or *CYP2D6**3/*6) a PM phenotype was assigned, as a duplication or deletion would not change the assignment. All other *CYP2D6* diplotypes were excluded from further analysis as the presence of a CNV could change the phenotype. A phenotype was considered actionable when it was described in the DPWG guidelines with advice in regards to a dose adjustment, drug change or intensive monitoring. Additionally, genotype to phenotype translations were also performed based on CPIC guidelines for all 11 genes in our panel [5].

Comparison

For comparison 21 samples, retrieved from 7 trios, were randomly selected from the prospective cohort and genotyped on a commercial platform (The pharmacoscan from Thermo Fisher Scientific, Waltham, Massachusetts, USA) [44].

This platform identifies 4,627 variants in 1,191 pharmacogenes, amongst which are all genes from the panel used in this study with the exception of Factor V Leiden. Genotype calls from the WES-pipeline were compared to the results obtained with the commercial platform.

Regions of heterozygosity in CYP2D6

For all genes in the panel, consent did not extend to the entire gene locus but only to the specific pharmacogenetic variants in the selected genes (Supplementary Table S3.1). Additionally, per individual per gene locus aggregated data containing the number of all heterozygous and homozygous variants was available. To assess possible deletions

in *CYP2D6*, a heterozygosity ratio was calculated for each individual. The number of heterozygous variants in the *CYP2D6* locus was divided by the total number of variants within this locus, resulting in the heterozygosity ratio. A high proportion of heterozygous variants indicates the presence of two different alleles and therefore a low to non-existing chance of a deletion. As this locus also includes upstream and downstream sequences which are not included in a *CYP2D6* gene deletion, the proportion of heterozygous variants can be higher than 0 even in the presence of a deletion due to variants in these upstream and downstream regions. As exact locations of the variants could not be obtained the impact of variants in these regions could not be determined.

Based on the distribution of the ratio of heterozygosity, both a strict (0.25) and conservative (0.4) cut-off in the ratio were examined. The conservative cut-off decreases the change of falsely excluding the presence of a deletion based on a high ratio of heterozygosity due to upstream and downstream variants. As all the SNVs selected in this study (Supplementary Table S3.1) are located in *CYP2D6* exons, which are part of the deletion region, heterozygosity for these variants would automatically rule out the presence of a deletion. Genotypes of individuals below either cut-off where assessed to determine the heterozygosity for the selected SNVs. Heterozygous calls for any of these SNVs will indicate the presence of two *CYP2D6* alleles and therefore the absence of a *CYP2D6* deletion despite a low ratio of heterozygosity, providing an estimate of the number of falsely low ratios of heterozygosity. This approach is exploratory, as there is no detailed information available regarding all variant locations definite answers regarding the presence of a *CYP2D6* gene deletion cannot be provided.

Highlights

What is the current knowledge on the topic?

Whole Exome Sequencing (WES) data is generated abundantly in clinical diagnostics and can potentially be repurposed for pharmacogenetics.

What question did this study address?

Can a clinically relevant pharmacogenetics profile be extracted from existing diagnostic WES? And what is the added value of haplotype phasing?

What does this study add to our knowledge?

- Out of 42 variants in 11 pharmacogenes for 1,583 individuals, genotypes and phenotypes based on existing WES data could be assigned to 70.4% of all potential genotype calls.
- Due to a lack of coverage and Copy Number Variant calling, genotypes could not be assigned *UGT1A1*, *CYP2C19* and *CYP2D6*.
- Eighty-six percent of all individuals carried at least one actionable phenotype.
- Haplotype phasing resulted in clinically relevant differences in phenotyping results compared to conventional, linkage based, assumptions for *CYP2B6* genotype assignments.

How might this change clinical pharmacology or translational science?

Repurposing existing WES data can yield a meaningful pharmacogenetic profile, which can be used in combination with existing guidelines, without the need for additional genetic testing.

References

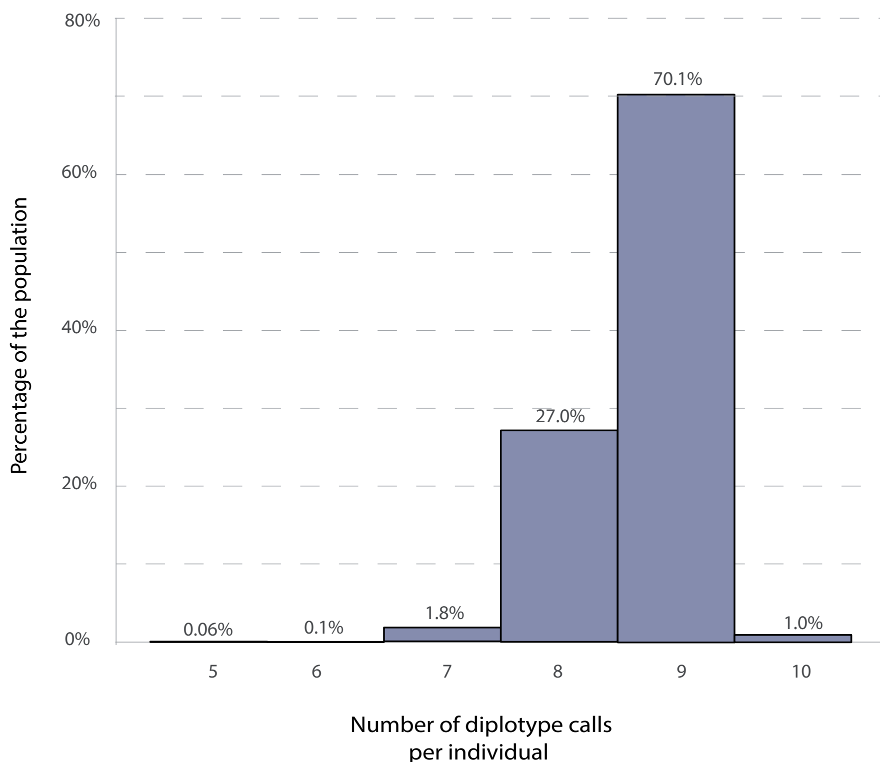
1. Relling, M.V.; Evans, W.E. Pharmacogenomics in the clinic. *Nature* **2015**, *526*, 343-350, doi:10.1038/nature15817.
2. Schildcrout, J.S.; Denny, J.C.; Bowton, E.; Gregg, W.; Pulley, J.M.; Basford, M.A.; Cowan, J.D.; Xu, H.; Ramirez, A.H.; Crawford, D.C., et al. Optimizing drug outcomes through pharmacogenetics: a case for preemptive genotyping. *Clinical pharmacology and therapeutics* **2012**, *92*, 235-242, doi:10.1038/clpt.2012.66.
3. Royal Dutch Pharmacy association. Pharmacogenetics. Availabe online: www.knmp.nl/farmacogenetica
4. Pharmacogenomics Knowledge base DPWG: Dutch Pharmacogenetics Working Group. Availabe online: https://www.pharmgkb.org/page/dpwg
5. Clinical Pharmacogenetics Implementation Consortium. CPIC-guidelines. Availabe online: https://cpicpgx.org/
6. Whirl-Carrillo, M.; McDonagh, E.M.; Hebert, J.M.; Gong, L.; Sangkuhl, K.; Thorn, C.F.; Altman, R.B.; Klein, T.E. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics* **2012**, *92*, 414-417, doi:10.1038/clpt.2012.96.
7. Bank, P.C.D.; Swen, J.J.; Guchelaar, H.J. Implementation of Pharmacogenomics in Everyday Clinical Settings. *Advances in pharmacology (San Diego, Calif.)* **2018**, *83*, 219-246, doi:10.1016/bs.apha.2018.04.003.
8. Yang, W.; Wu, G.; Broeckel, U.; Smith, C.A.; Turner, V.; Haidar, C.E.; Wang, S.; Carter, R.; Karol, S.E.; Neale, G., et al. Comparison of genome sequencing and clinical genotyping for pharmacogenes. *Clinical pharmacology and therapeutics* **2016**, *100*, 380-388, doi:10.1002/cpt.411.
9. Bielinski, S.J.; Olson, J.E.; Pathak, J.; Weinshilboum, R.M.; Wang, L.; Lyke, K.J.; Ryu, E.; Targonski, P.V.; Van Norstrand, M.D.; Hathcock, M.A., et al. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. *Mayo Clinic proceedings* **2014**, *89*, 25-33, doi:10.1016/j.mayocp.2013.10.021.
10. Londin, E.R.; Clark, P.; Sponziello, M.; Kricka, L.J.; Fortina, P.; Park, J.Y. Performance of exome sequencing for pharmacogenomics. *Personalized medicine* **2014**, *12*, 109-115, doi:10.2217/pme.14.77.
11. Cousin, M.A.; Matey, E.T.; Blackburn, P.R.; Boczek, N.J.; McAllister, T.M.; Kruisselbrink, T.M.; Babovic-Vuksanovic, D.; Lazaridis, K.N.; Klee, E.W. Pharmacogenomic findings from clinical whole exome sequencing of diagnostic odyssey patients. *Molecular genetics & genomic medicine* **2017**, *5*, 269-279, doi:10.1002/mgg3.283.
12. van der Wouden, C.H.; van Rhenen, M.H.; Jama, W.O.M.; Ingelman-Sundberg, M.; Lauschke, V.M.; Konta, L.; Schwab, M.; Swen, J.J.; Guchelaar, H.J. Development of the PGx-Passport: A Panel of Actionable Germline Genetic Variants for Pre-Emptive Pharmacogenetic Testing. *Clinical pharmacology and therapeutics* **2019**, 10.1002/cpt.1489, doi:10.1002/cpt.1489.
13. van der Wouden, C.H.; Cambon-Thomsen, A.; Cecchin, E.; Cheung, K.C.; Davila-Fajardo, C.L.; Deneer, V.H.; Dolzan, V.; Ingelman-Sundberg, M.; Jonsson, S.; Karlsson, M.O., et al. Implementing Pharmacogenomics in Europe: Design and Implementation Strategy of the Ubiquitous Pharmacogenomics Consortium. *Clinical pharmacology and therapeutics* **2017**, *101*, 341-358, doi:10.1002/cpt.602.
14. Dutch Pharmacogenetics Working Group. Transparency document pharmacogenetics. Availabe online: https://www.knmp.nl/downloads/g-standaard/transparantiedocumenten/transparantiedocument-farmacogenetica.pdf

15. Gaedigk, A.; Ingelman-Sundberg, M.; Miller, N.A.; Leeder, J.S.; Whirl-Carrillo, M.; Klein, T.E. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clinical pharmacology and therapeutics* **2018**, *103*, 399-401, doi:10.1002/cpt.910.
16. Dezentje, V.O.; Opdam, F.L.; Gelderblom, H.; Hartigh den, J.; Van der Straaten, T.; Vree, R.; Maartense, E.; Smorenburg, C.H.; Putter, H.; Dieudonne, A.S., et al. CYP2D6 genotype- and endoxifen-guided tamoxifen dose escalation increases endoxifen serum concentrations without increasing side effects. *Breast cancer research and treatment* **2015**, *153*, 583-590, doi:10.1007/s10549-015-3562-5.
17. Dunnenberger, H.M.; Crews, K.R.; Hoffman, J.M.; Caudle, K.E.; Broeckel, U.; Howard, S.C.; Hunkler, R.J.; Klein, T.E.; Evans, W.E.; Relling, M.V. Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers. *Annual review of pharmacology and toxicology* **2015**, *55*, 89-106, doi:10.1146/annurev-pharmtox-010814-124835.
18. Van Driest, S.L.; Shi, Y.; Bowton, E.A.; Schildcrout, J.S.; Peterson, J.F.; Pulley, J.; Denny, J.C.; Roden, D.M. Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. *Clinical pharmacology and therapeutics* **2014**, *95*, 423-431, doi:10.1038/clpt.2013.229.
19. Zhou, Y.; Ingelman-Sundberg, M.; Lauschke, V.M. Worldwide distribution of cytochrome P450 alleles: A meta-analysis of population-scale sequencing projects. *Clinical pharmacology and therapeutics* **2017**, *102*, 688-700, doi:10.1002/cpt.690.
20. Lang, T.; Klein, K.; Fischer, J.; Nussler, A.K.; Neuhaus, P.; Hofmann, U.; Eichelbaum, M.; Schwab, M.; Zanger, U.M. Extensive genetic polymorphism in the human CYP2B6 gene with impact on expression and function in human liver. *Pharmacogenetics* **2001**, *11*, 399-415.
21. Kirchheiner, J.; Klein, C.; Meineke, I.; Sasse, J.; Zanger, U.M.; Murdter, T.E.; Roots, I.; Brockmoller, J. Bupropion and 4-OH-bupropion pharmacokinetics in relation to genetic polymorphisms in CYP2B6. *Pharmacogenetics* **2003**, *13*, 619-626, doi:10.1097/01.fpc.0000054125.14659.d0.
22. Tsuchiya, K.; Gatanaga, H.; Tachikawa, N.; Teruya, K.; Kikuchi, Y.; Yoshino, M.; Kuwahara, T.; Shirasaka, T.; Kimura, S.; Oka, S. Homozygous CYP2B6 *6 (Q172H and K262R) correlates with high plasma efavirenz concentrations in HIV-1 patients treated with standard efavirenz-containing regimens. *Biochemical and biophysical research communications* **2004**, *319*, 1322-1326, doi:10.1016/j.bbrc.2004.05.116.
23. Hofmann, M.H.; Blievernicht, J.K.; Klein, K.; Saussele, T.; Schaeffeler, E.; Schwab, M.; Zanger, U.M. Aberrant splicing caused by single nucleotide polymorphism c.516G>T [Q172H], a marker of CYP2B6*6, is responsible for decreased expression and activity of CYP2B6 in liver. *The Journal of pharmacology and experimental therapeutics* **2008**, *325*, 284-292, doi:10.1124/jpet.107.133306.
24. Watanabe, T.; Saito, T.; Rico, E.M.G.; Hishinuma, E.; Kumondai, M.; Maekawa, M.; Oda, A.; Saigusa, D.; Saito, S.; Yasuda, J., et al. Functional characterization of 40 CYP2B6 allelic variants by assessing efavirenz 8-hydroxylation. *Biochemical pharmacology* **2018**, *156*, 420-430, doi:10.1016/j.bcp.2018.09.010.
25. Snyder, M.W.; Adey, A.; Kitzman, J.O.; Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nature reviews. Genetics* **2015**, *16*, 344-358, doi:10.1038/nrg3903.
26. Browning, S.R.; Browning, B.L. Haplotype phasing: existing methods and new developments. *Nature reviews. Genetics* **2011**, *12*, 703-714, doi:10.1038/nrg3054.
27. Yao, R.; Zhang, C.; Yu, T.; Li, N.; Hu, X.; Wang, X.; Wang, J.; Shen, Y. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular cytogenetics* **2017**, *10*, 30, doi:10.1186/s13039-017-0333-5.

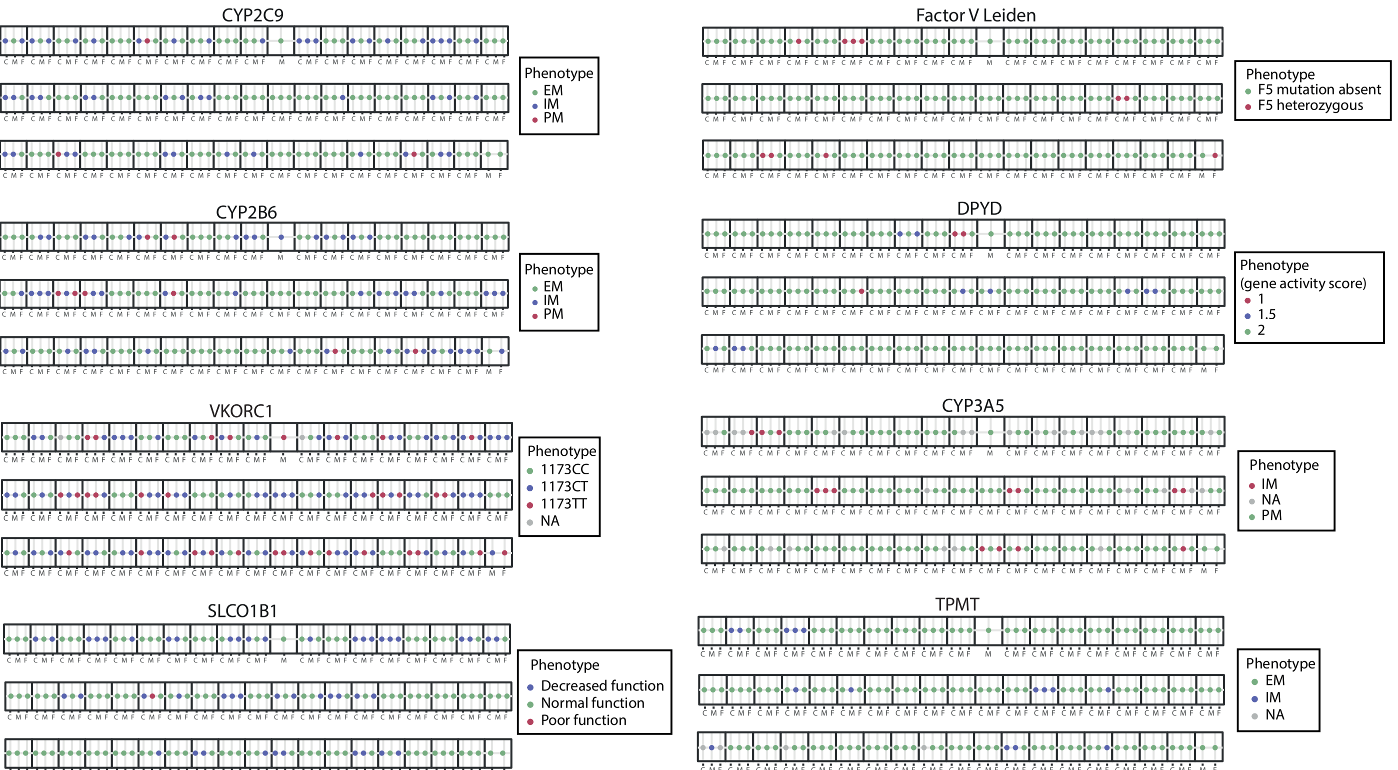
28. Bank, P.C.D.; Swen, J.J.; Schaap, R.D.; Klootwijk, D.B.; Baak-Pablo, R.; Guchelaar, H.J. A pilot study of the implementation of pharmacogenomic pharmacist initiated pre-emptive testing in primary care. *European journal of human genetics : EJHG* **2019**, 10.1038/s41431-019-0454-x, doi:10.1038/s41431-019-0454-x.
29. Reisberg, S.; Krebs, K.; Lepamets, M.; Kals, M.; Magi, R.; Metsalu, K.; Lauschke, V.M.; Vilo, J.; Milani, L. Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genetics in medicine : official journal of the American College of Medical Genetics* **2018**, 10.1038/s41436-018-0337-5, doi:10.1038/s41436-018-0337-5.
30. Lee, S.B.; Wheeler, M.M.; Patterson, K.; McGee, S.; Dalton, R.; Woodahl, E.L.; Gaedigk, A.; Thummel, K.E.; Nickerson, D.A. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genetics in medicine : official journal of the American College of Medical Genetics* **2019**, *21*, 361-372, doi:10.1038/s41436-018-0054-0.
31. Lee, S.B.; Wheeler, M.M.; Thummel, K.E.; Nickerson, D.A. Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences. *Clinical pharmacology and therapeutics* **2019**, 10.1002/cpt.1552, doi:10.1002/cpt.1552.
32. Bank, P.C., Swen J.J., Guchelaar H.J. The nationwide impact of implementing a preemptive pharmacogenetic panel approach to guide drug prescribing in primary care in the Netherlands. [Abstract]. In Proceedings of American Society for Clinical Pharmacology and Therapeutics, Orlando, Florida.
33. Peklar, J.; Kos, M.; O'Dwyer, M.; McCarron, M.; McCallion, P.; Kenny, R.A.; Henman, M.C. Medication and supplement use in older people with and without intellectual disability: An observational, cross-sectional study. *PloS one* **2017**, *12*, e0184390, doi:10.1371/journal.pone.0184390.
34. Salomon, C.; Britt, H.; Pollack, A.; Trollor, J. Primary care for people with an intellectual disability - what is prescribed? An analysis of medication recommendations from the BEACH dataset. *BJGP open* **2018**, *2*, bjgpopen18X101541, doi:10.3399/bjgpopen18X101541.
35. O'Dwyer, M.; McCallion, P.; McCarron, M.; Henman, M. Medication use and potentially inappropriate prescribing in older adults with intellectual disabilities: a neglected area of research. *Therapeutic advances in drug safety* **2018**, *9*, 535-557, doi:10.1177/2042098618782785.
36. Relling, M.V.; Klein, T.E. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical pharmacology and therapeutics* **2011**, *89*, 464-467, doi:10.1038/clpt.2010.279.
37. Bank, P.C.D.; Caudle, K.E.; Swen, J.J.; Gammal, R.S.; Whirl-Carrillo, M.; Klein, T.E.; Relling, M.V.; Guchelaar, H.J. Comparison of the Guidelines of the Clinical Pharmacogenetics Implementation Consortium and the Dutch Pharmacogenetics Working Group. *Clinical pharmacology and therapeutics* **2018**, *103*, 599-618, doi:10.1002/cpt.762.
38. Osanlou, O.; Pirmohamed, M.; Daly, A.K. Pharmacogenetics of Adverse Drug Reactions. *Advances in pharmacology (San Diego, Calif.)* **2018**, *83*, 155-190, doi:10.1016/bs.apha.2018.03.002.
39. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* **2013**.
40. Poplin, R.; Ruano-Rubio, V.; DePristo, M.A.; Fennell, T.J.; Carneiro, M.O.; Van der Auwera, G.A.; Kling, D.E.; Gauthier, L.D.; Levy-Moonshine, A.; Roazen, D. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* **2017**, 201178.
41. Broad Institute PhaseByTransmission. Availabe online: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_phasing_PhaseByTransmission.php

42. Wildeman, M.; van Ophuizen, E.; den Dunnen, J.T.; Taschner, P.E. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human mutation* **2008**, *29*, 6-13, doi:10.1002/humu.20654.
43. Vis, J.K.; Vermaat, M.; Taschner, P.E.; Kok, J.N.; Laros, J.F. An efficient algorithm for the extraction of HGVS variant descriptions from sequences. *Bioinformatics (Oxford, England)* **2015**, *31*, 3751-3757, doi:10.1093/bioinformatics/btv443.
44. ThermoFisher Scientific. Pharmacoscan Assay Availabe online: https://www.thermofisher.com/order/catalog/product/903010TS

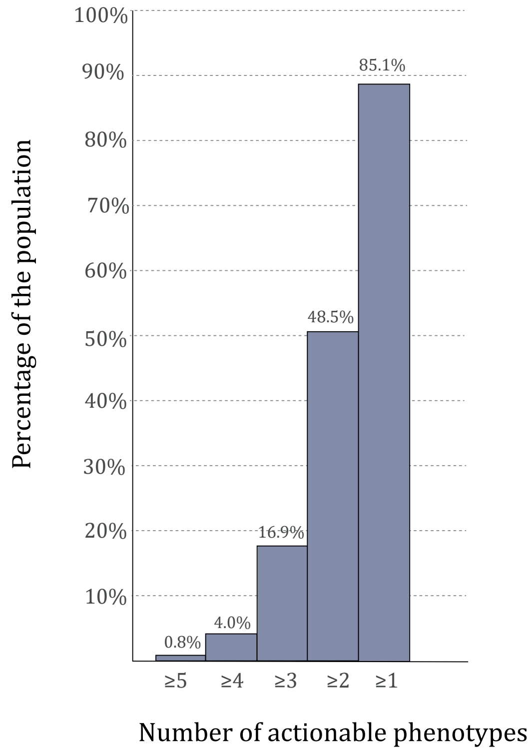
Supplementary material



Supplementary Figure S3.1: Number of diplotype calls based on SNV data per individual
11 genes were screened in 1,583 individuals, when there was a lack of coverage for at least one variant a diplotype was not assigned. Copy Number Variants could not be determined and are not included in the diplotype assigment. The majority of individuals had diplotype data for 9 out of 11 genes. No genotype calls for UGT1A1 were available, only 1% of the diplotypes for CYP2C19 could be called.

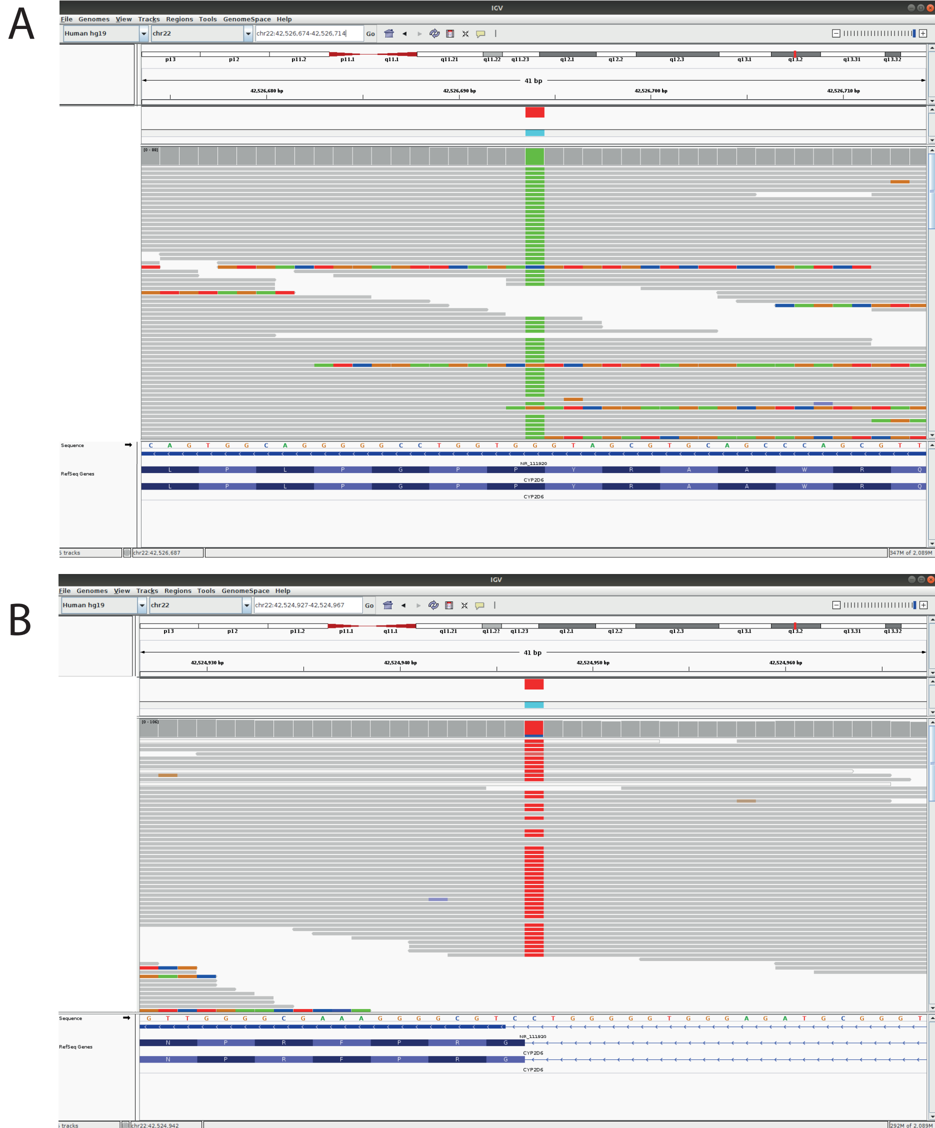


Supplementary Figure S3.2: Phenotypes per gene split out per family
Phenotypes are based on the variant panel from the Ubiquitous pharmacogenomics consortium, 42 variants in 11 genes were investigated. No results for UGT1A1 and for CYP2C19 could be obtained. CYP2D6 was excluded from this analysis due to the inability to call phenotypes for all individuals. In the absence of copy number data, only diplotypes consisting of two null-alleles were assigned a phenotype (N=11) leading to incomplete data for the child-parent trios. Each cluster represents one child-parent trio and their phenotypes. C: Child, M: Mother, F: father, PM: Poor Metabolizer, IM: Intermediate Metabolizer, EM: Extensive Metabolizer, F5: factor V Leiden.



Supplementary Figure S3.3: Total number of actionable phenotypes per individual based on CPIC guidelines

Actionable is classified as: any phenotype with a dosing advice available in the CPIC guidelines. An unknown phenotype is categorized as not actionable. Results are based on all genotypes with sufficient coverage (haplotype quality > 20). Due to an inability to call copy number variants, only CYP2D6 diplotypes consisting of 2 null-alleles were phenotyped (PM), no phenotype was assigned for other phenotypes classifying them as unknown.



Supplementary Figure S3.4: Readpile of individual for whom the CYP2D6 call is discrepant between Pharmacoscan (*4/*4) and WES data (*4/*10)
(A) g.42526694G>A, which is homozygous mutant, is found in both the *CYP2D6**4 and the *CYP2D6**10 allele **(B)** g.42524947C>T for which the mutation is observed in 87% of all reads, is only found in the CYP2D6*4 allele.

Supplementary Table S3.1: Genes and variants included in the panel
Inclusion is based on the panel developed by the U-PGx consortium.

Gene	Star-allele	RS-number	HGVS (GRCh37)
CYP2B6	*4	rs2279343	NC_000019.9:g.41515263A>G
	*18	rs28399499	NC_000019.9:g.41518221T>C
	*9	rs3745274	NC_000019.9:g.41512841G>T
	*16	rs2279343; rs28399499	NC_000019.9:g.41515263A>G; NC_000019.9:g.41518221T>C
	*6	rs2279343; rs3745274	NC_000019.9:g.41515263A>G; NC_000019.9:g.41512841G>T
CYP2C19	*17	rs12248560	NC_000010.10:g.96521657C>T
	*9	rs17884712	NC_000010.10:g.96535246G>A
	*4A/B	rs28399504	NC_000010.10:g.96522463A>G
	*8	rs41291556	NC_000010.10:g.96535173T>C
	*2	rs4244285	NC_000010.10:g.96541616G>A
	*3	rs4986893	NC_000010.10:g.96540410G>A
	*5	rs56337013	NC_000010.10:g.96612495C>T
	*10	rs6413438	NC_000010.10:g.96541615C>T
	*6	rs72552267	NC_000010.10:g.96535210G>A
CYP2C9	*3	rs1057910	NC_000010.10:g.96741053A>C
	*2	rs1799853	NC_000010.10:g.96702047C>T
	*11	rs28371685	NC_000010.10:g.96740981C>T
	*5	rs28371686	NC_000010.10:g.96741058C>G
CYP2D6	*10	rs1065852	NC_000022.10:g.42526694G>A
	*17	rs28371706	NC_000022.10:g.42525772G>A
	*41	rs28371725	NC_000022.10:g.42523805C>T
	*3	rs35742686	NC_000022.10:g.42524244delT
	*4	rs3892097; rs1065852	NC_000022.10:g.42524947C>T; NC_000022.10:g.42526694G>A
	*6	rs5030655	NC_000022.10:g.42525086delA
	*9	rs5030656	NC_000022.10:g.42524176_42524178delCTT
	*8	rs5030865	NC_000022.10:g.42525035C>A
	*14B	rs5030865	NC_000022.10:g.42525035C>T
	*14A	rs5030865; rs1065852	NC_000022.10:g.42525035C>T; NC_000022.10:g.42526694G>A
CYP3A5	*6	rs10264272	NC_000007.13:g.99262835C>T
	*7	rs41303343	NC_000007.13:g.99250393_99250394insA
	*3	rs776746	NC_000007.13:g.99270539C>T
DPYD	*2	rs3918290	NC_000001.10:g.97915614C>T
	*13	rs55886062	NC_000001.10:g.97981343A>C
	1236G>A	rs56038477	NC_000001.10:g.98039419C>T
	2846A>T	rs67376798	NC_000001.10:g.97547947T>A
F5L	FvL	rs6025	NC_000001.10:g.169519049T>C

Supplementary Table S3.1 continues on next page.

Supplementary Table S3.1: ***Continued***

Gene	Star-allele	RS-number	HGVS (GRCh37)
SLCO1B1	*5	rs4149056	NC_000012.11:g.21331549T>C
TPMT	*3A	rs1142345; rs1800460	NC_000006.11:g.18130918T>C; NC_000006.11:g.18139228C>T
	*3C	rs1142345	NC_000006.11:g.18130918T>C
	*3B	rs1800460	NC_000006.11:g.18139228C>T
	*2	rs1800462	NC_000006.11:g.18143955C>G
UGT1A1	*27	rs35350960	NC_000002.11:g.234669619C>A
	*6	rs4148323	NC_000002.11:g.234669144G>A
	*36	rs8175347[5]	NC_000002.11:g.234668881_234668882TA[5]
	*28	rs8175347[7]	NC_000002.11:g.234668881_234668882TA[7]
	*37	rs8175347[8]	NC_000002.11:g.234668881_234668882TA[8]
VKORC1	1173	rs9934438	NC_000016.9:g.31104878G>A

Supplementary Table S3.2: Haplotype frequencies of the prospective cohort
There were no significant differences in haplotype frequencies between the parents and children (χ^2, $p<0.001$) with the exception of *VKORC1* ($p=0.48$).

		Parents only (prospective)		Children only (prospective)	
Gene	Haplotype assignment	N	Frequency	N	Frequency
CYP2B6		226		110	
	*1	165	73.0%	86	78.2%
	*4	13	5.8%	3	2.7%
	*6	46	20.4%	21	19.1%
	*9	2	0.88%		
CYP2C19		0		0	
CYP2C9		226		110	
	*1	189	83.6%	94	85.5%
	*2	25	11.1%	11	10%
	*3	12	5.3%	5	4.5%
CYP2D6		224		110	
	*1	129	57.6%	57	51.8%
	*10	10	4.5%	5	4.5%
	*3	8	3.6%	2	1.8%
	*4	41	18.3%	26	23.6%
	*41	23	10.3%	12	10.9%
	*6	2	0.9%	1	0.9%
	*9	11	4.9%	7	6.4%
CYP3A5		192		88	
	*1	9	4.7%	5	5.7%
	*3	181	94.3%	82	93.2%
	*6	2	1.0%	1	1.1%
DPYD		226		110	
	*1	217	96.0%	106	96.4%
	*2A	2	0.88%	1	0.91%
	1236G>A	6	2.7%	3	2.7%
	2846A>T	1	0.44%	0	
FVL		226		110	
	F5 positive	7	3.1%	3	2.7%
	F5 negative	219	96.9%	107	97.3%
SLCO1B1		226		110	
	*5	31	13.7%	18	16.4%
	wt	195	86.3%	92	83.6%

Supplementary Table S3.2 continues on next page.

Supplementary Table S3.2: *Continued*

		Parents only (prospective)		Children only (prospective)	
Gene	Haplotype assignment	N	Frequency	N	Frequency
TPMT		224		106	
	*3A	11	4.9%	5	4.7%
	wt			1	0.94%
		213	95.1%	100	94.3%
VKORC1		226		106	
	1173T	93	41.2%	48	45.3%
	wt	133	58.8%	58	54.7%

Supplementary Table S3.3: Manually assigned diplotype for calls that could not be phased automatically
Diplotype assignments are based on linkage and adapted from the Ubiquitous pharmacogenomics consortium translation tables.

Gene	Assigned diplotype	Variant on allele A	Variant on allele B	Unphased variants	Frequency
CYP2B6	*1/*6	-	-	41515263A>G; 41512841G>T	81
	*1/*6	-	41515263A>G	41512841G>T	3
	*1/*6	-	41512841G>T	41515263A>G	7
	*1/*6	41515263A>G	-	41512841G>T	8
	*4/*6	41515263A>G	41515263A>G	41512841G>T	3
	*6/*9	41512841G>T	41512841G>T	41515263A>G	1
CYP2C9	*2/*3	-	96741053A>C	96702047C>T	2
CYP2D6	*4/*1	42526694G>A	-	42524947C>T	2
	*4/*1	42524947C>T	-	42526694G>A	2
	*4/*3	42524947C>T; 42526694G>A	-	42524244delT	1
	*3/*4	-	42526694G>A; 42524947C>T	42524244delT	1
	*9/*4	42524178_ 42524180delTCT	-	42526694G>A; 42524947C>T	1
	*41/*4	42523805C>T	-	42526694G>A; 42524947C>T	2
	*41/*4	-	42526694G>A; 42524947C>T	42523805C>T	1
	*1/*4	-	42524947C>T	42526694G>A	2
	*1/*4		42526694G>A	42524947C>T	1
	*41/*4	42523805C>T	42524947C>T	42526694G>A	1
	*4/*41	-	-	42526694G>A; 42523805C>T; 42524947C>T	2
	*4/*41	-	42523805C>T	42526694G>A; 42524947C>T	3
	*10/*41	-	42523805C>T	42526694G>A	1
	*10/*41	42526694G>A	-	42523805C>T	2
	*4/*41	-	-	42523805C>T; 42526694G>A; 42524947C>T	1
	*9/*41	42524178_ 2524180delTCT	-	42523805C>T	1
	*4/*9	-	42524178_ 42524180delTCT	42526694G>A; 42524947C>T	1
	*17/*41	42523805C>T	-	42525772G>A	1
	*17/*41	-	42523805C>T	42525772G>A	1
	*41/*9	-	42524178_ 42524180delTCT	42523805C>T	1

Supplementary Table S3.3 continues on next page.

Supplementary Table S3.3: *Continued*

Gene	Assigned diplotype	Variant on allele A	Variant on allele B	Unphased variants	Frequency
CYP3A5	*3/*7	-	99250394_ 99250395insA	99270539C>T	1
TPMT	wt/*3A	-	-	18130918T>C; 18139228C>T	18

Supplementary Table S3.4: Phenotype frequencies and actionability based on CPIC guidelines
Phenotypes are based on the U-PGx translation tables, actionability is based on the Clinical Pharmacogenomics implementation consortium (CPIC) guidelines, where actionable is defined as a phenotype accompanied by at least one dosing advise. Guidelines for Factor V Leiden are not available.

Gene	Phenotype	Number of subjects	Frequency	Actionable
CYP2B6		1,577		
	PM	105	6.7%	Yes
	IM	528	33.5%	Yes
	EM	944	59.9%	
CYP2C19		15		
	PM	-		Yes
	IM	4	26.7%	Yes
	EM	7	46.7%	
	RM	4	26.7%	Yes
	UM	-		Yes
CYP2C9		1,583		
	PM	59	3.7%	Yes
	IM	487	30.2%	Yes
	EM	1,037	65.5%	
CYP2D6		1,576		
	PM*	66	4.2%	Yes
	Not assigned	1,510	95.8%	
CYP3A5		1,163		
	PM	882	75.8%	
	IM	263	22.6%	Yes
	EM	18	1.5%	Yes
DPYD		1,581		
	AS: 0	-		Yes
	AS: 0.5	-		Yes
	AS: 1	21	1.3%	Yes
	AS: 1.5	95	6.0%	Yes
	AS: 2	1,465	92.7%	
SLCO1B1		1,579		
	Normal function	1,172	74.2%	
	Decreased function	371	23.5%	Yes
	Poor function	36	2.3%	Yes
TPMT		1,562		
	PM	1	0.1%	Yes
	IM	139	8.9%	Yes
	EM	1,421	91.0%	

Supplementary Table S3.4 continues on next page.

Supplementary Table S3.4: ***Continued***

Gene	Phenotype	Number of subjects	Frequency	Actionable
VKORC1		1,549		
	Normal function (1173CC)	564	36.4%	
	Decreased function (1173CT)	723	46.7%	
	Poor function (1173TT)	262	16.9%	Yes

PM: Poor Metabolizer, IM: Intermediate Metabolizer, EM: Extensive Metabolizer, RM: Rapid Metabolizer, UM: Ultra-rapid Metabolizer, AS: Activity Score, * Based on Single Nucleotide Variants (2 null-alleles), as Copy Number variants could not be determined.